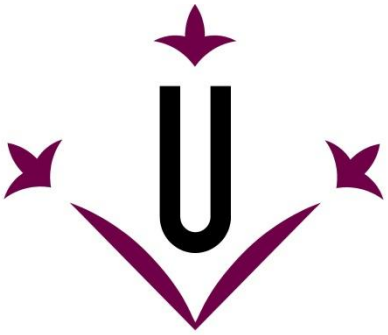
Development and application of computational methodologies for Integrated Molecular Systems Biology

Hiren Mahendrabhai Karathia

Dipòsit Legal: L. 280-2013

http://hdl.handle.net/10803/110518



Universitat de Lleida

DEVELOPMENT AND APPLICATION OF COMPUTATIONAL METHODOLOGIES FOR INTEGRATED MOLECULAR SYSTEMS BIOLOGY

HIREN MAHENDRABHAI KARATHIA

DOCTOR OF PHILOSOPHY

SUPERVISOR: RUI ALVES

GROUP OF BIOMATHEMATICS AND BIOSTATISTICS,

DEPARTMENT DE CIÈNCIES MÈDIQUES BÀSIQUES

UNIVERSITY OF LLEIDA – IRB LLEIDA,

LLEIDA (CATALUÑA)

SPAIN

2012

શ્રી ૧।

॥श्री गणेशाय नमः॥

શ્રી ગોપલો જયતે શ્રી લાલજી મહારાજ જયતે શ્રી કાળકા માં જયતે

या कुन्देन्दु तुषार हार धवला
या शुभ्रवस्त्रावृता ।
या वीणावरदंड मंडितकरा
या श्वेतपद्मासना ॥

Declaration

As required under the University of Lleida regulations, I hereby declare that this thesis work is not substantially the same as any that I have submitted or will be submitting for a degree or diploma or other qualification at this or any other university. Furthermore, this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specified explicitly in the text.

Lleida, 4th September, 2012.

Signature

Hiren M. Karathia

I, Rui Carlos Vaqueiro de Castro Alves, Associate Professor in the Basic Medical Sciences Department of the University of Lleida, certify that the present study, entitled "Development and Application of Computational Methodologies for Integrated Molecular Systems Biology" and presented by Hiren Karathia for the award of the degree of Doctor, has been carried out under my supervision in the department where I am assigned.

Lleida, 4th September, 2012.

Signature

Rui Alves

Acknowledgements

This thesis dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost, I dedicate this thesis to my *"Guru"* (supervisor), **Prof. Rui Alves**, who has supported me throughout my thesis with his patience and knowledge whilst allowing me the room to work in my own way. I attribute the level of my knowledge to his encouragement and effort. Without him this thesis work would not have been completed or written. One simply could not wish for a better or friendlier supervisor.

I would like to thank **Prof. Albert Sorribas**, who has always provided me friendly and familial atmosphere. I would also like to thank **Ester Villaprinyó**, who has provided friendly and social atmosphere inside as well as outside the department. I would like to thank **Montserrat Rue, Montserrat Martinez, Carles Forné, Anabel Usie, Alexandra** for being so much cooperative colleagues. Without them, I would have been lonely in the lab.

I cannot forget support of my friends in Lleida. I would always keep great memories of time that I spent with **Jordi Rosse, Ernesto, Paul, Jose, Marta, Jordi, Uxue, Bupesh, Arindam, Sarvanam** and many more. I thank to all of them for always providing me with joyful support. I would also thankful to our collaborators, **Francesc Solsona**, **Ivan Texidó**, who provided technical supports and help in establishing back end server of Homol-MetReS.

I would also like to thank to **AGAUR**, which has been source of my fellowship during the thesis and Ministry of Education and Science, Spain for providing support of partial fundings for development of Homol-MetReS.

I am so lucky that **Falguni Karathia** is my wife and she has been with me during my stay in Lleida and during my Ph.D. program. She has always been cheering me up and stood by me through the good times and bad. My research would not have been possible without her help. I also devote achievements to my **Parents**, **grandmother**, **Harshit, Nilesh, Megha, Sanjay Kumar, parents-in-law, brothers-in-law**, the dearest sister, **Arti** and my nieces (**Devangi and Smeet**). They were always supporting me and encouraging me with their best wishes.

Last but not the least, one above all of us, I devote all these to the ever present **Gopal lalji** and **Lalaji maharaj** for answering my prayers and always being with me to provide strength for finishing this thesis.

Publications, Posters, and Congress Presentations

Published Papers:

Hiren Karathia, Ester Vilaprinyo, Albert Sorribas and Rui Alves, 2011, "*Saccharomyces cerevisiae as a model organism: a comparative study*." PLoS One 6(2): e16015 **(Chapter 2 of this thesis)**.

Papers included in this thesis that have been submitted or are in preparation:

Hiren Karathia, Anabel Usie, Ivan Teixido, Ester Vilaprinyo, Albert Sorribas, Francesc Solsona and Rui Alves, "*A human centric comparison of eukaryotic proteomes: Implications for the study of human biology*."

Hiren Karathia, Ivan Teixidó, Anabel Usie, Ester Vilaprinyo, Albert Sorribas, Francesc Solsona and Rui Alves, "*Homol-MetReS: An integrated framework tool to study evolutionary molecular systems biology*."

Posters and oral talks in meetings:

Hiren Karathia, Anabel Usie, Ester Vilaprinyo, Francesc Solsona and Rui Alves, 2012, "*Homol-MetReS: A web application for integration between molecular systems biology and evolutionary biology*", Oral talk in Student symposium at XI Jornadas de Bioinformática, Barcelona.

Hiren Karathia, Anabel Usie, Ester Vilaprinyo, Albert Sorribas., Francesc Solsona and Rui Alves, 2011, "*Homol-MetReS: Network Reconstruction Based on Whole Proteome Comparisons*", Poster presentation at ICMSB 2011 -XIIth International Congress on Molecular Systems Biology, Lleida, Spain.

Hiren Karathia and Rui Alves, 2010, "*A tool for comparison of complete proteomes between pairs of organisms*", Poster presentation at 9th European Conference on Computational Biology, Ghent, Belgium.

Hiren Karathia and Rui Alves, 2010, "*Saccharomyces cerevisiae as a model organism: Strengths & drawbacks*", Poster presentation at European Molecular Biology Organization (EMBO), Barcelona, Spain.

Published collaborations that led to papers to be included in theses by other people:

Baldiri Salvado, Ester Vilaprinyo, Hiren Karathia, Albert Sorribas and Rui Alves, 2012. "*Two component systems: physiological effect of a third component*." PLoS One 7(2): e31095.

Baldiri Salvado, Hiren Karathia, Anabel Usie Chimenos, Ester Vilaprinyo, Stig Omholtc, Albert Sorribas and Rui Alves, 2011, "*Methods for and results from the study of design principles in molecular systems.*" Math Biosci 231(1): 3-18.

Anabel Usié, Hiren Karathia, Ivan Teixidó, Joan Valls, Xavier Faus, Rui Alves, and Francesc Solsona1, 2011, "*Biblio-MetReS: a bibliometric network reconstruction application and server.*" BMC Bioinformatics 12: 387.

Oussama Abdelli, Anabel Usié, Hiren Karathia, Jordi Vilaplana, Francesc Solsona and Rui Alves, "*Parallelizing Biblio-MetReS, a data mining tool*." XXII Jornadas del paralelismo, La Laguna 7-9 Sep, 2011.

Abbreviations

2D-PAGE	2 Dimensional Polyacrylamide Gel Electrophoresis
A	Class of Absent Proteins
AIDS	Acquired Immune Deficiency Syndrome
AL	Alignment Length
AP	Absent Protein
BLAST	Basic Local Alignment Sequence Tool
CaMK	Calcium/Calmodulin-dependent protein Kinase
CCA	Common Component Architecture
CDC	Cell Division Cycle
CEMP	Cementum Protein
CoA	Coenzyme A
D	Class of Domain Ortholog Proteins
DEFB4	Defensin Beta 4
DOCK	Dedicator of Cytokinesis
DP	Domain ortholog Protein
E.C	Enzyme Commission
EGFR	Epidermal Growth Factor Receptor
FO	Class of Functional Orthologous proteins
FOXP	Forkhead box P
GHRL	Ghrelin/obestatin prepropeptide
GO	Gene Ontology
GPI	Glycosylphosphatidylinositol
GST	General Systems Theory
H	Homologues Class

HD	Hamming Distance
HMHB1	Histocompatibility (minor) HB-1
HP	Homologous Protein
HPRD	Human Protein Reference Database
HTTP	Hypertext Transfer Protocol
IR	Illegitimate Recombination
IVF	In-Vitro Fertilization
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KISS	KISSpeptin
MAPK	Mitogen Activated Protein Kinase
M-M	Many-to-Many
M-O	Many-to-One
MSML	Molecular Systems Mark-up Language
NCBI	National Centre for Biotechnology Information
NHD	Normalized Hamming Distance
NHEJ	Non-Homologous End-Joining
NMR	Numerical Magnetic Resonance
O	Class of Functional Orthologous proteins
Og	Class of General Orthologous proteins
O-M	One-to-Many
O-O	One-to-One
OP	Orthologous Protein
ORM	Object Relation Mapping
PAEP	Progestagen-Associated Endometrial Protein
PERL	Practical Extraction and Report Language
PL	Protein Length

PLN PhosphoLambaN

RSK Ribosomal S6 Kinase

RXR Retinoid X Receptor

SBML Systems Biology Mark-up Language

ScCAGs *Saccharomyces cerevisiae* Clusters of Absent Genes

ScCHGs *Saccharomyces cerevisiae* Clusters of Homologues

ScCOGs *Saccharomyces cerevisiae* Clusters of Orthologs

SCOP Structural Classification of Protein

SGD *Saccharomyces Genome* Database

SLN Sarcolipin

SNARE Soluble N-ethylmaleimide-sensitive factor Attachment Protein Receptor

SP Significant Protein

SPRR Small Proline-Rich protein

STD Sexually Transmitted Disease

TAT Trans-Activator of Transcription

TCA Tri Carboxylic Acid

THMB Thymosin Beta

TIMM Translocase of Inner Mitochondrial Membrane

TNRFRSF Tumour Necrosis factor Receptor Superfamily

TOP1 Topoisomerase 1

TOR Target of Rapamycin

UQCR UbiQuinol-Cytochrome c Reductase

ZODB Zope Object Database

Glossary

Glossary

Alternative Organism - any non-classical model organisms from any biological kingdom from which new biological knowledge can be extracted with minimum constraints.

Biological System - a group of common sub-systems that interact to perform and regulate a certain biological task.

Cluster of Absent Proteins - collection of query proteins with no homologous sequences in a target proteome.

Cluster of Homologues - collection of proteins from multiple proteomes that meet sequence homology criteria.

Cluster of Orthologs - collection of proteins from multiple proteomes that meet sequence orthology criteria.

Complete Proteome - entire set of proteins coded in a genome.

Cybernetics - a trans-disciplinary approach for exploring regulatory systems, their structures, constraints, and possibilities.

Dynamical Systems Theory - an area of mathematics used to describe the behavior of complex dynamical systems, usually by employing differential or difference equations.

E-value or Expect value (E) - a parameter that describes the number of similar sequence hits one can "expect" to see by chance when searching a sequence database of a particular size. The lower the E-value, the more "significant" the match is between two analyzed sequences.

F-score - a composite score proposed to evaluate most likely ortholog pairs.

Functional Genome - complete set of genes that are required to build a functional organism.

Functional Module - sets of molecules that are involved in a given biological process.

Genome Project - a scientific research project designed to study and identify all of the genes in an organism's genome, to determine the base-pair sequences in human DNA, and to store this information in computer databases.

Hamming Distance - distance between two vectors of equal length in the number of positions at which the corresponding numerical symbols are different.

Homologous Protein – proteins that diverged from a common ancestor.

Model Organism - a representative organism that can be used to study a given process whose results can be extended to other organisms.

Molecular Organization - sets which comprise one or more molecular entity and that assemble in order to form cellular phenomena.

Omics - informally referred to a field of study in biology ending in -omics, such as genomics, proteomics or metabolomics.

Orthologous Proteins - proteins that have diverged from a common ancestor and have the same function in different species.

Predictive Biology - an inter-disciplinary area of biological predictions based on available genomics and proteomics data.

Protein Annotation - an act or process of furnishing critical commentary or explanatory notes to describe a protein X in terms of topic Y.

Protein Ontology – a classification of proteins for their nature of being, existence, or reality, as well as the basic categories of being and their relations.

Proteome Hierarchical Information - collection of proteins arrangements in a way in which each protein is represented as being "above," "below," or "at the same level as" one another at sequence diversity level.

Proteome Spatial Information - collection of information for expression, localization, synthesis, degradation, and turnover rates of endogenously expressed, untagged proteins in different subcellular compartments.

Proteome Functional Information - collection of information for interaction mapping, interaction network and cellular functions in signaling pathway/networks.

Proteome Structural Information - complete information of entire proteins in a proteome in context of the proteins amino acid sequences.

Proteomics - is the large-scale study of proteins, particularly their structures and functions.

Sequence Alignment - a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

Sequence Identity or Similarity - criteria used for measuring how equivalent two sequences are over the span of an alignment of proteins, DNA or RNA.

Systems Biology - a biology-based inter-disciplinary field of study that focuses on complex interactions within biological systems, using a more holistic perspective (holism instead of the more traditional reductionism) approach to biological and biomedical research.

Theoretical Biology - a scientific research field with a range of applications in biology, medicine and biotechnology. The field may be referred to as mathematical biology or biomathematics to stress the mathematical side, or as theoretical biology to stress the biological side

Summary

English

The aim of the work presented in this thesis was the development and application of computational methodologies that integrate sequence, functional, and genomic information to provide tools for the reconstruction, annotation and organization of complete proteomes in such a way that the results can be compared between any number of organisms with fully sequenced genomes.

Methodologically, I focused on identifying molecular organization within a complete proteome of a reference organism, linking each protein in that proteome to proteins of other organisms in such a way that anyone can compare the two proteomes at spatial, structural, functional, cellular tissue, development or physiology levels. Such linkage between proteomes is based on estimations of sequence similarity between the proteins in the reference proteome and the proteins in the alternative proteome. The similarities are used to link functional information between proteomes and identify both, the most likely functional orthologs and the proteins that are absent in either of the organisms with respect to the other. This methodology was implemented in a pipeline that integrates a central database with independent modules for computation, annotation, analysis, and visualization of results.

The methodology was applied to address the issue of identifying appropriate model organisms to study different biological phenomena. To do so we made and partially tested the hypothesis that "*similarity between the set of proteins that comprise the network responsible for a given biological phenomenon in two organisms is a reasonable proxy for similarity in the dynamics and adaptive responses of those networks".* This was done by comparing the protein sets involved in different biological phenomena in *Saccharomyces cerevisiae* and *Homo sapiens* to corresponding sets in other organisms with fully sequenced genomes to find that, whenever experimental data was available, our hypothesis was consistent with the data. Furthermore, our analysis could explain differences in phenotypes between similar species of organisms.

This thesis concludes by presenting a web server, Homol-MetReS, on which the methodology is implemented. It provides an open source environment to the scientific community on which they can perform multi-level comparison and analysis of proteomes.

Español

El objetivo del trabajo presentado en esta tesis fue el desarrollo y la aplicación de metodologías computacionales que integran el análisis de la secuencia y de la información funcional y genómica, con el objetivo de reconstruir, anotar y organizar proteomas completos, de tal manera que estos proteomas se puedan comparar entre cualquier número de organismos con genomas completamente secuenciados.

Metodológicamente, me he centrado en la identificación de la organización molecular dentro de un proteoma completo de un organismo de referencia, vinculando cada proteína del proteoma a las proteínas de otros organismos, de tal manera que cualquiera pudiera comparar las siguientes características entre proteomas: distribución de las proteínas a nivel espacial, tejidular, funcional, fisiológico y de desarrollo del organismo . Tal conexión entre proteomas se basa en estimaciones de similitud de secuencia entre las proteínas en el proteoma de referencia y las proteínas en el proteoma alternativo. Las similitudes se utilizan para transferir información funcional entre proteomas e identificar tanto los ortólogos más similares en funcionalidad como las proteínas que están ausentes en cualquier organismo con respecto a algún otro. Esta metodología se aplicó en un “pipeline” que integra una base de datos central con módulos independientes para el cálculo, anotación, análisis y visualización de los resultados.

La metodología se aplicó para abordar la cuestión de la identificación de organismos modelo adecuados para estudiar diferentes fenómenos biológicos. Para ello hemos realizado, y parcialmente validado, la hipótesis de que "la similitud entre el conjunto de proteínas que conforman la red responsable de un determinado fenómeno biológico en dos organismos se corresponde de forma razonable con una similitud en la dinámica y las respuestas adaptativas de las redes". Esto se hizo comparando conjuntos de proteínas involucradas en diferentes fenómenos biológicos en *Saccharomyces cerevisiae* y *Homo sapiens* con los conjuntos correspondientes de otros organismos con genomas completamente secuenciados. Se observó que nuestra hipótesis era consistente con los datos en los casos en que existe información experimental. Además, nuestro análisis podría explicar las diferencias en los fenotipos similares entre las especies de organismos.

La tesis concluye con la presentación de un servidor web, Homol-MetReS, en el que se implementa la metodología. Homol-MetReS proporciona un entorno de código abierto a la comunidad científica en la que se pueden realizar múltiples niveles de comparación y análisis de proteomas.

Català

L'objectiu del treball presentat en aquesta tesi va ser el desenvolupament i l'aplicació de metodologies computacionals que integren l'anàlisis de informació sobre seqüències proteiques, informació funcional i genòmica per a la reconstrucció, anotació i organització de proteomes complets, de manera que els resultats es poden comparar entre qualsevol nombre d'organismes amb genomes completament seqüenciats.

Metodològicament, m'he centrat en la identificació de l'organització molecular dins d'un proteoma complet d'un organisme de referència, associant cada proteïna del proteoma a les proteïnes funcionalment corresponents en altres organismes, de manera que qualsevols dos proteomes es poden comparar respecte a la distribució de les seves proteïnes en las següents dimensions: espacial, estructural, funcional, teixidular, el desenvolupament o els nivells de la fisiologia. Tal associació entre proteomes es basa en estimacions de similitud de seqüència entre les proteïnes en el proteoma de referència i les proteïnes en el proteoma alternatiu. Les similituds s'utilitzen per transferir informació funcional entre proteomes i identificar tant, els orthologs més similars funcionalment com les proteïnes que estan absents en qualsevol dels organismes pel que fa a l'altre. Aquesta metodologia es va aplicar en un pipeline que integra una base de dades central amb mòduls independents per al càlcul, anotacions, anàlisi i visualització dels resultats.

La metodologia es va aplicar per abordar la qüestió de la identificació de organismes model adequats per a estudiar diferents fenòmens biològics. Per això hem realitzat, i parcialment testat, la hipòtesi que "la similitud entre el conjunt de proteïnes que conformen la xarxa responsable d'un determinat fenomen biològic en dos organismes es correspon de forma raonable amb una similitud en la dinàmica i les respostes adaptatives de les xarxes". Això es va fer mitjançant la comparació d'un conjunt de proteines involucrades en diferents fenòmens biològics en *Saccharomyces cerevisiae* i *Homo sapiens* amb els conjunts corresponents d'altres organismes amb genomes completament seqüenciats. Vam trobar quela nostra hipòtesi era consistent amb les dades experimentals disponibles en la literatura. A més, el nostre anàlisi podria explicar les diferències en els fenotips similars entre les espècies d'organismes.

La tesi conclou amb la presentació d'un servidor web, Homol-MetReS, en què s'implementa la metodologia. Homol-MetReS proporciona un entorn de codi obert a la comunitat científica en què es poden realitzar múltiples nivells de comparació i anàlisi de proteomes.

Gujarati

આ સ્નાતક ઉપાધી મેળવવા માટે લખેલા આ મહાનિબંધ કાર્ય નો મૂળ હેતુ ગણકયંત્ર (કમ્પ્યુટર) દ્વારા પદ્ધતિ અને તેનો વ્યવસ્થિત ઉપયોગ એવી રીતે પ્રસ્થાપિત કરવાનો હતો કે જેના દ્વારા કોઈ પણ ઔજસદ્રવ્ય (પ્રોટેઈન) માં રહેલા બહુધા અણુસમુદાયો નો ક્રમ, તેના દ્વારા થતું જૈવિક કાર્ય, અને આનુંવાન્શિક ને લગતી માહિતી આપતું એક વ્યવસ્થિત માળખાકીય યંત્ર વૈજ્ઞાનિક સમાજ ને આપી શકાય. આ યંત્ર માંથી બહાર નીકળતી માહિતીઓનો અભ્યાસ વૈજ્ઞાનિકો એવી રીતે કરી શકે કે જેના દ્વારા કોઈ પણ સજીવ કોષ માં રહેલા સંપૂર્ણ જૈવિક બંધારણ અને તેની તુલાનાકીય માહિતી, તેઓનું નામકરણ, તેની અંદર રહેલી માહિતી આપલે ની સંપૂર્ણ માળખાકીય વ્યવસ્થા ની તુલના, બીજા કોઈ સજીવ નાં કોષો માં રહેલા માળખાકીય વ્યવસ્થા સાથે કરી શકાય.

મારા અભ્યાસ નો આરંભિક ઉદ્દેશ બહુ પ્રચલિત જીવો નો અભ્યાસ, કે જેનો અભ્યાસ વૈજ્ઞાનિક સૃષ્ટી માં બહુ જ થાય છે, તેઓનાં વિશ્લેષણો ને કેન્દ્ર સ્થાને રાખીને કરવાનો છેઅમોએ . સૌ પ્રથમ આ જીવ માં રહેલા બધા જ પ્રોટેઈન માં રહેલા ક્રમિક અણુસમુદાયો ને એક પછી એક બીજા જીવ માં રહેલા પ્રોટેઈનો નાં અણુસમુદાયો સાથે એવી રીતે સરખાવામાં માં આવેલ છે કે દરેક બે પ્રોટેઈનો ની એક જોડી બની રહે. ત્યારબાદ, દરેક પ્રોટેઈન ની જોડીઓ ને એક સમાનતા ધરાવતા સમૂહો માં જમા કરવામાં આવેલ છે કે જેની અંદર દરેક પ્રોટેઈન ની જોડીઓ માં રહેલા અણુસમુદાયો નાં ક્રમ નું પૃત્થ્કરણ કોષો માં રહેલા કોઈ એક કાર્ય ના સંદર્ભ માં કરી શકાય. આવા સમૂહો મા રહેલ દરેક પ્રોટેઈન ની જોડીઓ નો ઉપયોગ બાદમાં તેના જે તે માતૃત્વ કોષ સાથે નીશ્બત ધરાવતા સિદ્ધાંતો, જેવા કે, અવકાશિકતા, કાર્યશીલતા, ભૌતિકતા અને એક કોષ માંથી બીજા કોષ નાં ઉત્પાદન કાર્ય ક્ષમતા વગેરે નાં પૃત્થ્કરણ માટે કરવામાં આવેલ છે. આ પ્રકારની વિધેયાત્મક માહિતી દ્વારા મોટે ભાગે એવું જાણી શકાય છે કે ક્યા કોષો નાં સીદ્ધાંતો સાથે સંકળાયેલા પ્રોટીનો ક્યા બીજા સજીવો ના જૈવિક કોષો માં હાજર છે અને ક્યા સજીવો ના જૈવિક કોષો માં ગેર હાજર છે. આ પદ્ધતિમાં શ્રુન્ખ્લાકીય ગણતરી દ્વારા પ્રાપ્ત કરેલી માહિતીઓ, જેવી કે, પ્રોટેઈનો માટે નાં માપાંકો, માનદંડો, નામકરણ, વિશ્લેષણ, અને તેના પરિણામો નાં દ્રશ્યો વગેરે ને લગતી માહિતીઓ એક સાથે અને એક સમૂહ માં કેન્દ્રીય કરી કમ્પ્યુટર નાં ડેટાબેઝ સાથે સલગ્ન કરવામાં આવેલી છે.

આ પદ્ધતિ દ્વારા સજીવ ની જૈવિક પ્રતિકૃતિ ઓળખવા માટે અને તેઓમાં રહેલી વિવિધ જૈવિક ઘટના નાં અભ્યાસ ના મુદ્દાને ઉદાહરણ સ્વરૂપે લાગુ કરવામાં આવેલ છે. આમ કરવા માટે, અમે એક આંશિક પૂર્વધારણા ને ધ્યાન માં લીધી હતી કે "*જે પ્રોટીનો નો સમૂહ આપેલ બે સજીવો નાં જૈવિક ઘટના માટે જવાબદાર હોય અને તેઓના બહુધા અણુસમુદાયો સમાનતા ધરાવતા હોય, તેઓ આ જૈવિક ઘટનાઓ ની ગતિશીલતા અને તે માળખા નાં અનુકૂલનશીલતા ની સામ્યતા માટે પણ એક પણ એક બીજા ની પ્રતિકૃતિ સમાન છે*". આ પૂર્વધારણા નાં પ્રયોગિક પુરાવા માટે અમે બે સજીવો ને પ્રતિકૃતી ની સાબિતી માટે ઉદાહરણ સ્વરૂપે લીધા છે અને તેને સૃષ્ટિ ની દરેક જાતી નાં સજીવો નાં આનુંવાન્શિકતા અને કોષો ની કાર્યક્શીલતા અને જૈવિક ઘટના માટે જવાબદાર સમૂહો વચ્ચે સમાનતા ધરાવતો આભ્યાસ હાથ ધર્યો છે. આ પૂર્વધારણા નાં પ્રયોગિક પુરાવા માટે અમે એવું પણ સાબિત કરી શક્યા છીએ કે જ્યારે પણ પ્રાયોગિક માહિતી ઉપલબ્ધ હોય, અમારા વિશ્લેષણ ની માહિતી આ માહિતીઓ સાથે સુસંગત મળતી આવે છે. આ ઉપરથી અમારું તારણ એ નીકળે છે કે અમારી પદ્ધતિ અને તેના વિશ્લેષણ દ્વારા કોઈ પણ સજીવ ના જૈવિક કાર્ય ના સંશોધન બીજા સજીવ ની સરખામણી નાં આધારે કરવા માટે ઉપયોગ માં લઇ શકાય છે.

આ મહાનિબંધ નું અંતિમ તારણ અમે આ સંપૂર્ણ અભ્યાસ પદ્ધતિ ને એક ગણ યંત્ર ઉપર વ્યવસ્થિત રીતે સંસ્થાપન કરીને ને સંપૂર્ણ કરેલ છે. જેનું નામ "હોમોલ-મેતરેસ" છે. આ ગણક યંત્ર પદ્ધતિ દ્વારા કોઈ પણ વૈજ્ઞાનિક કોઈ પણ જૈવિક સમુદાય ની પ્રતિકૃતિ નો અભ્યાસ બીજા જૈવિક સમુદાય ના ઉદાહરણ સ્વરૂપે કરી શકે છે કે જેના દ્વારા વૈજ્ઞાનિક સમાજ ને એક નવાજ પ્રકાર નો વૈજ્ઞાનિક અભિગમ પ્રાકટ્ય થઇ શકે અને તેના વિષેની માહિતીઓ ની આપલે પણ થઇ શકે.

Table of Contents

List of Figures

List of Supplementary Figures

List of Tables

List of Supplementary Tables

Objectives

There are three main objectives for this thesis.

1. To develop an integrative methodology that systematically:

    a) Identifies analogous proteins between different organisms and proposes relationships between the analogues at functional and/or spatial organization levels.
    b) Identifies clusters of functionally similar proteins from different organisms, based on levels of sequence similarity.
    c) Facilitates functional (re)annotation and mapping of proteins to alternative functional categories and transfer of that annotation between organisms.
    d) Applies a method to integrate the information in such a way that any two proteins can be mapped at any level of the functional and/or evolutionary organization proposed in this thesis.

2. Applying the methodology to the study of well characterized organisms and comparing the proteomes of these organisms to the proteomes of other living beings with fully sequenced genomes.

3. The third and final objective is to implement the methodology developed in objective **1** on a web server for free use by the community.

Chapter 1. Introduction

"ॐ पूर्णमदः पूर्णमिदं पूर्णात्पुर्णमुदच्यते
पूर्णश्य पूर्णमादाय पूर्णमेवावशिष्यते ॥"

"Om purnam adah, purnam idam, purnat purnam udachyate,
Purnasya purnam adaya, purnam evavasisyate."

"That is whole, this is whole;
From that whole, this whole came;
From that whole, this whole removed or added,
What remains is whole."
(Vedas)

These are the two philosophical/conceptual lines in Sanskrit quoted from *"PURNAMADAH"*, by Swami Dayanand Saraswati. Their interpretation differs between individuals.

From a theoretical biologist's point of view, the first two lines could be interpreted as suggesting that an organism is regulated by its cells, each of which is itself a whole organizational unit of life, and descended from another cell. These two cells are distinct but similar entities and biological organization principles appear to exist at the different levels at which they can be studied. Investigating even the simplest object within the whole that are those cells could contribute to decoding some of the general principles that impinge upon how they work.

From a system biologist's point of view, the last two lines of the above saying can be interpreted as describing the molecular organization of the cell, in which each molecule establish various relationships at spatial, functional, conditional or temporal level with others, and contribute to functional modules within that cell. Each event generated by one such set of relationships leads to the emergence of a given phenotype. Each change in the relationships or in the components can lead to a change in that phenotype. Such changes are caused by regulatory events at the different levels of organization and the interacting whole runs the life of the cell. Investigating the changes and their timing provides information about the past and present of a system. Sometimes, the information is enough to predict future behavior.

Inspired by these beautiful four lines and the philosophy contained in them, the work presented in this *Ph. D.* thesis deals with the problem of designing, standardizing and applying a methodology to,

a) identify molecular organization within a complete sequenced proteome from any organism, and
b) compare and annotate complete proteomes of any two organisms at spatial, structural, functional, cellular and tissue level.

This is an important issue, given that it is impossible to study all organisms in detail. By performing the types of comparisons described in a) and b) we may be able to identify groups of organisms that are similar and dissimilar regarding different aspects of their biology. Such identification permits the rational choosing of the "average" organism in that group to study a given process and extrapolate its functioning to the other organisms of the group. This can reduce the number of organisms that one has to study in order to understand the molecular biology of a given process across the tree of life. We also aimed at implementing the methodology in web tool that is accessible to be used by other researchers.

Developing such a methodology requires considering the general properties of molecular organization in a cell and the functional information of its components. Integrating this information will allow extrapolating behavior between organisms, through comparison of the similarities and differences between sequences of the proteomes. This can be more easily achieved by comparing the molecular modules involved in a given biological response and assuming that similarity between the modules is correlated to similarity between the responses. With this in mind the work done in this thesis was planned to:

a) Develop the basic methodology and apply it to test the assumption that similarity between molecular module components is positively correlated to similarity between responses. To do so, we applied the method to compare the proteome of *Saccharomyces cerevisiae* to that of other fully sequence organisms, followed by a limited phenotypic comparison between that yeast and a few selected organisms;

b) Apply the validated methodology to compare the human proteome to that of other eukaryotes and identify both, what makes us unique at the protein level and those eukaryotes that could serve as good models to study different aspects of human physiology.

c) Implement the methodology in a web tool and make it available for other researchers to use in their research.

With these goals in mind, in this chapter we will provide a short introduction to the biological and methodological considerations that directly led to this work as well as to the organization of the remainder of the thesis. We start by a brief description of what is a biological system. We then move forward to shortly discuss systems biology and limit that discussion to the aspects we find strictly relevant for the work presented in the thesis. We then zoom in on the effects of the various genome projects on the amount of data that made this thesis possible and follow with a briefly categorical analysis of proteomics, the science that focuses on the analysis of the proteome, which is our subject matter. We conclude the chapter with a schematic description of the goals for the work, and of the remaining chapters of the thesis and their organization.

1.1. Biological Systems

A biological system (from Latin systēma, in turn from Greek σύστημα systēma, "whole compounded of several parts or members, system", literary "composition") is a group of common sub-systems that interact to perform and regulate a certain task. Generally, a cell, a tissue, an organ, an organism can be considered as living components of the biological system that works at various levels of coordination to perform certain functions that make its system alive. Thus, living systems are organized in modular fashion. This modularity appears to exist all the way down to the molecular level (**Figure 1.1**).

As the modules interact at different levels, they create a system that reacts and adjusts over time to environmental stimuli, allowing the organisms survive, reproduce, and evolve. One can only understand these processes if one are willing to study biological systems as a whole. This notion was proposed by Alexander Bogdanov. It was later given a more serious scientific framework by Bertalanffy in his General Systems Theory (GST) and Cybernetics

Figure 1.1

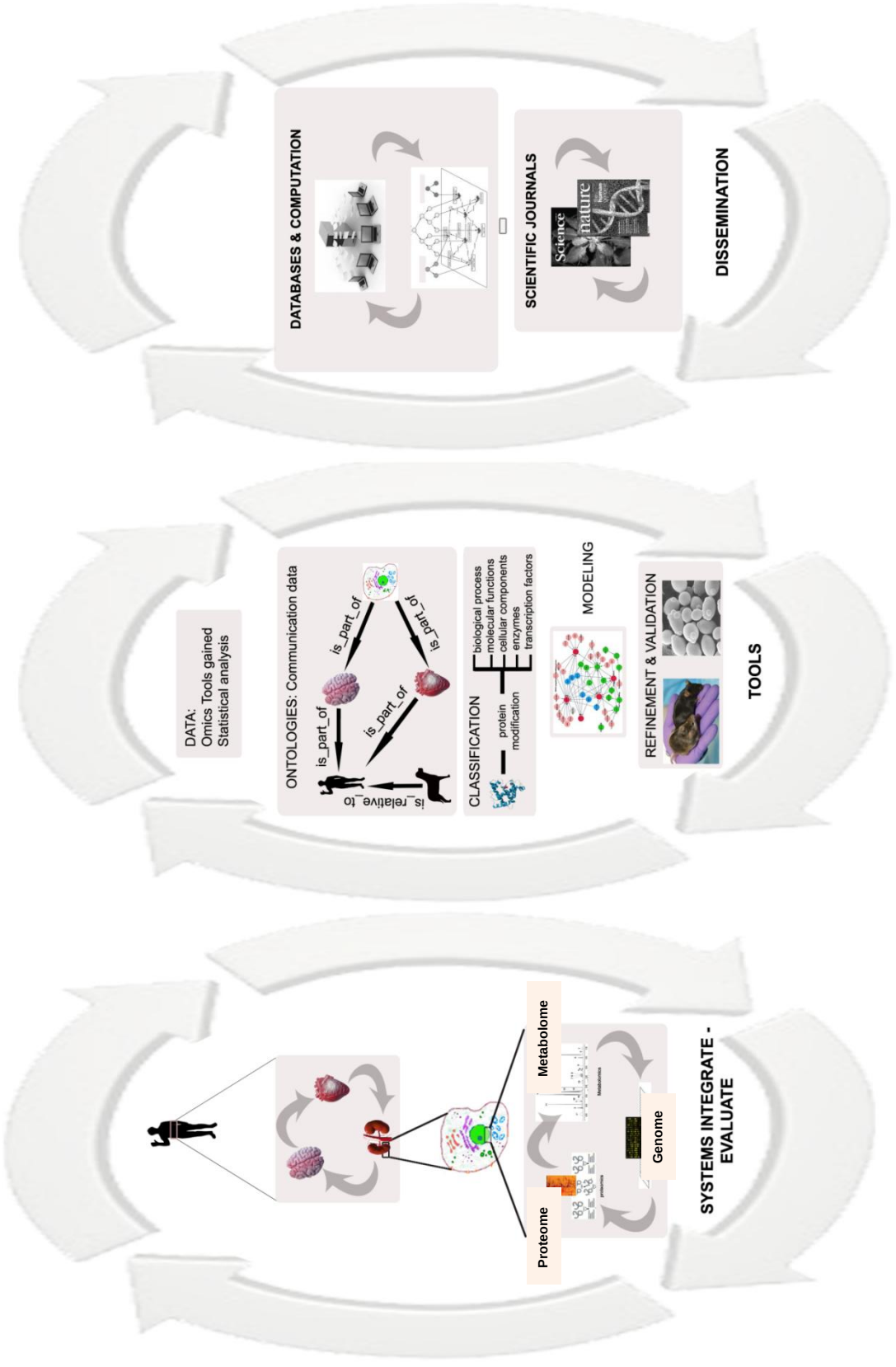


Figure 1.1 Modularity and Systems Biology in living beings

1.2. Systems Biology

Current systems biology directly evolved from the ideas of General Systems Theory. The term describes a rapidly evolving interdisciplinary field that endeavors to understand the detailed coordinated workings of a large set of components of cells and organisms [1, 2]. Often, these workings are used to distinguish between healthy and pathological states [3]. Currently, typical systems biology studies can be of different types:

a) Those emphasizing the use of high-throughput *"omics"* technologies to measure the changes in the complete set of individual molecular components of a certain type in the whole system [4-7]. After analyzing the whole set of data, the components of the bigger system are often grouped into smaller subsystems that are then analyzed, in a *"top-down"* approach.

b) Those emphasizing a *"bottom-up"* approach that begins by analyzing the behavior of individual (sets of) components of the bigger system, from molecules to functional modules. The different analyses are then integrated and used to understand the behavior of whole system [8-12].

c) Those emphasizing systems biology as a *"New physiology"*, that complements the reductionist molecular biology with integrative approaches [13]. Such approach has been adopted in pioneering work on heart models, in a *"middle-out"* strategy, starting from tissue models (*"middle level"*), incrementally extending to the organ and *"higher"* levels as well as *"down"* to molecular detail [14, 15].

d) Those focusing on dynamic systems theory. Such studies integrate approaches from "*dynamic system theory*" and use those approaches to describe the behavior of complex dynamic systems, usually by employing differential equations or difference equations [16-19].

e) Those that focus on specific processes and responses [20]. Examples are the work that tries to understand how bacteria regulated their chemo-tactile behavior [21-24] or nitrogen assimilation through two component systems [25].

f) Those that use systems approaches for doing *"Predictive Biology"* or *"Quantitative Biology"* [26]. Datasets are collected and used to characterize a given biological system [27]. This characterization is then used to predict how the system will work under different circumstances [28]. Doing so typically requires the use of various software packages for modeling, analysis, visualization, and general data manipulation [29-32]. These packages run on different platforms and communicate to use each other's capabilities via a fast binary encoded-message system like markup languages [33-36].

In all the cases, however, appropriately identified individual components of the systems with high quality functional annotation and inter-dependent relationships are required to make the most out of systems biology approaches. The work described in the current thesis addresses that issue, as it enables a better identification and functional annotation of components in new genomes. It also allows for the systemic reconstruction of the molecular modules and circuits that are responsible for different types of biological processes and events. Finally it allows for a clear comparison between the molecular components of corresponding functional modules in different proteomes.

1.3. Fruits of Genome Projects

The identification of the complete list of individual components of proteomes is a direct consequence of the number of genome projects that have been finished or are undergoing [37, 38]. Genome projects sequence the complete DNA content of organisms and identify the regulatory elements, genes and proteins coded in that sequence. These results have wide implications and applications [39-44]. Some of the later are described in **Table 1.1**.

Table 1.1 Applications resulting from genome projects

Genome Projects	Area of Applications	Specific Applications
Human Genome	**Molecular Medicine**	Improved diagnosis of disease
		Earlier detection of genetic predispositions to disease
		Rational drug design
		Gene therapy and control systems for drugs
		Pharmacogenomics "custom drugs"
	Risk Assessment	Assess health damage and risks caused by radiation exposure, including low-dose exposures
		Assess health damage and risks caused by exposure to mutagenic chemicals and cancer-causing toxins
		Reduce the likelihood of heritable mutations
Animal Genome	**Bio-archaeology, Anthropology, Evolution, and Human Migration**	Study evolution through germ-line mutations in lineages
		Study migration of different population groups based on female genetic inheritance
		Study mutations on the Y chromosome to trace lineage and migration of males
		Compare breakpoints in the evolution of mutations with ages of populations and historical events
Plants and Animal Genomes	**Agriculture, Livestock Breeding, and Bioprocessing**	Disease-, insect-, and drought-resistant crops
		More nutritious produce
		Bio-pesticides
		Edible vaccines incorporated into food products
		New environmental clean-up uses for plants like tobacco
Microbial Genome	**Energy and Environmental Applications**	Create new energy sources (biofuels)
		Develop environmental monitoring techniques to detect pollutants

The immense amounts of data generated by these projects create several problems. It is not enough to simply annotate the genes and proteins in the genomes. It is also important to connect this annotation to

a) the functional information about the individual proteins and the modules and circuits that they create through their interactions, physical or otherwise,

b) the information about the hierarchical relationships among the modules at different scales of biological scale and organization, and

c) the information about temporal and spatial behaviour of the different components of the systems, ranging from gene expression, to protein abundance and activity, to metabolic fluxes and concentrations.

The work described in this thesis focuses on complete proteomes and provides ways to integrate and transfer structural, functional, hierarchical and spatial information between proteins of different organisms. The scheme for the functional organization of the proteome that we use for enabling such transfer is summarized in **Figure 1.2**.

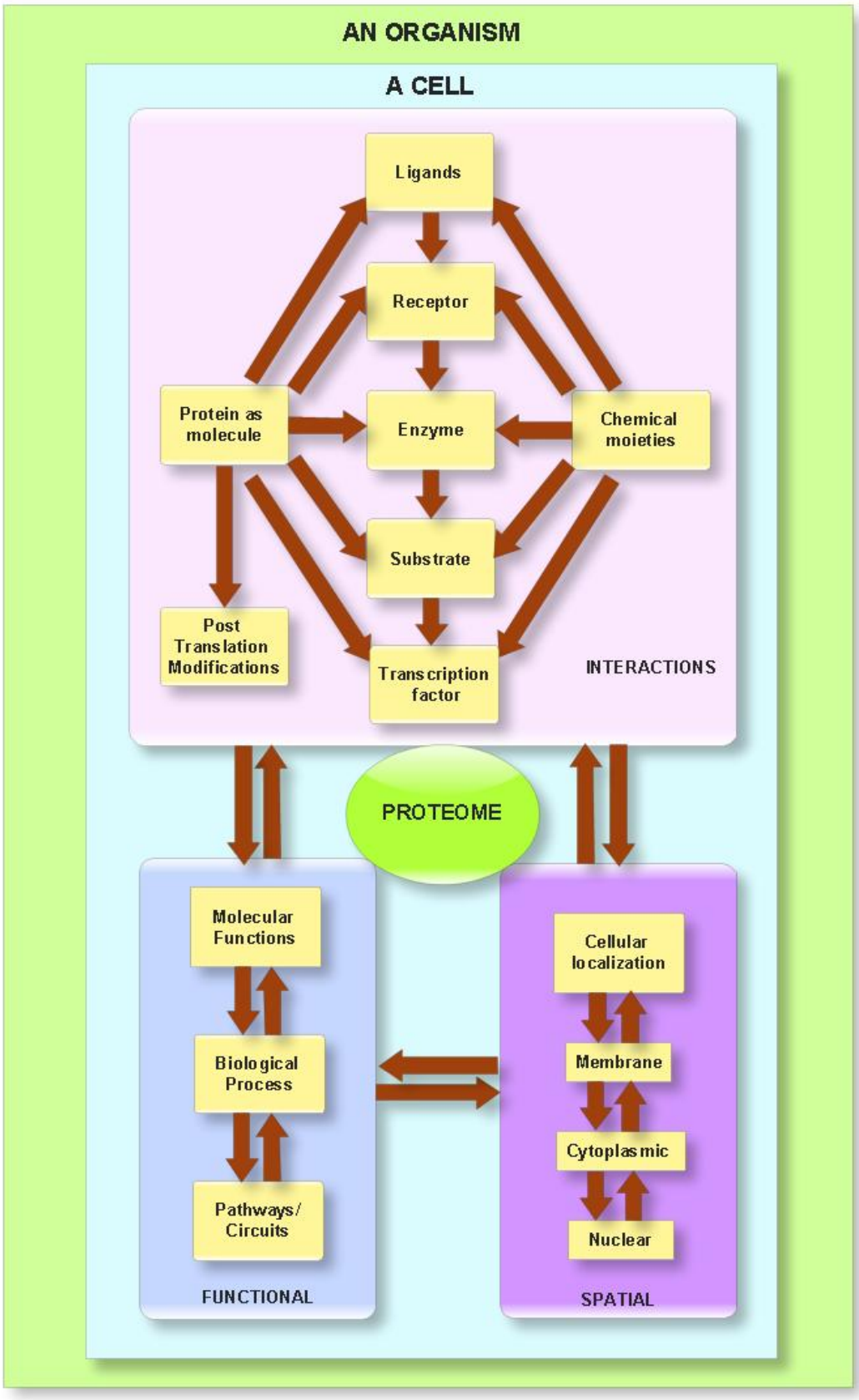


Figure 1.2 Relational scheme used to integrate protein information of cells

1.4. Proteome

An organism's genome contains the complete set of genes that are required to build a functional organism. The complete proteome is the entire set of proteins coded in those genes. These proteins are differentially expressed by the genome at any given time, depending on environmental conditions and cell types, among other factors. Proteomics is the science that studies the proteome [**45**]. Until recently, protein function analyses mainly focused on single proteins. As high-throughput technologies improved it became possible to study large fractions of a cell's proteome. Generally proteomic analysis results from three broad types of experiments:

a) **Structural proteomics, the large-scale analysis of protein structures:** Protein structure comparisons can help to identify the functions of newly discovered genes [**46**]. Structural analysis can also show where drugs bind to proteins and where proteins interact to each other. This is achieved using technologies such as X-ray crystallography and NMR spectroscopy, and, more recently, protein modeling. The protein oriented information provide background of feature predictions for secondary structure, solvent accessibility, trans-membrane helices, globular regions, coiled-coil regions, structural switch regions, B-values, disorder regions, intra-residue contacts, protein-protein and protein-DNA binding sites, sub-cellular localization, domain boundaries, beta-barrels, cysteine bonds, metal binding sites and disulphide bridges.

b) **Expression proteomics, the large-scale analysis of protein expression:** Measurements of protein abundance and activity identified by the main proteins found in a particular sample and proteins differentially expressed in related samples, such as diseased vs. healthy tissue. A protein found only in a diseased sample may represent a useful drug target or diagnostic marker. Proteins with similar expression profiles may also be functionally related. Technologies such as 2D-PAGE and mass spectrometry are used here. These and other technologies allow for protein identification [**47**], measurement of protein abundance [**48**] and processing [**49**], including post translation modifications, determination of protein interactions [**50**], compartmentalization [**51**], turnover time [**52**], etc. Proteome signatures that are specific to a given cell type, phenotype, or adaptive response can be identified through the qualitative and quantitative comparison of proteomes measured under the alternative relevant conditions.

c) **Interaction proteomics, the large-scale analysis of protein interactions:** The characterization of protein-protein interactions helps to determine protein functions and can also show how proteins assemble in larger complexes. Technologies such as affinity purification, mass spectrometry and the yeast two-hybrid system are particularly useful.

While proteomic evaluation has improved research output in a variety of disciplines, a number of distinct classes of proteins can be identified within a proteomic dataset. Such classes are defined in **Box 1.1**. They were used to organize the database that underlies the tool described in Chapter 4 and facilitate functional analysis of large protein sets.

Box 1.1 Classifiers of proteome into various dimensions

Complete proteome - The proteome is the entire set of proteins coded in a completely sequenced genome, including alternative products such as splice variants for those species in which these may occur.

Cellular proteome - Subset of a complete proteome containing all proteins expressed in specific types of cells under induction of particular sets of environmental stimuli.

Functional proteome - Subset of a complete proteome containing all proteins for which functional information is available.

Enzymatic proteome - Subset of a complete proteome containing all proteins that have an associated enzyme activity.

Receptor proteome - Subset of a complete proteome containing all proteins that are known to be integral membrane proteins and are involved in recognizing and binding to signals in order to initiate signal transduction.

Ligand proteome - Subset of a complete proteome containing all proteins that bind to receptors and lead to signal transduction.

Localized proteome - Subset of a complete proteome containing all proteins that are specifically associated to the different subcellular compartments and components.

Gene Regulatory proteome - Subset of a complete proteome containing all proteins that are involved in regulation of gene expression through direct interaction with genetic elements.

Box 1.1 [continued…]

Post-translational modified proteome - Subset of a complete proteome containing all proteins that are known to suffer post translational chemical modifications in the side-chains of their amino acids.

Interacting proteome - Subset of a complete proteome, containing all sets of proteins that are known to physically interact with each other.

Biological process proteome - Subset of a complete proteome containing all sets of proteins associated to known biological processes

Pathways proteome - Subset of a complete proteome containing all sets of proteins associated to known biological pathways and circuits.

1.5. Thesis Organization

The remaining chapters of this thesis are organized as follows:

a) The second chapter focuses on describing the development of a methodology for the functional comparison of proteomes and applying that to the well characterized model organism *Saccharomyces cerevisiae*. A large scale functional comparison between its proteome and that of other organisms with fully sequenced genomes is made.

b) The third chapter focuses on applying the methodology developed in Chapter 2 to the comparative study of the complete human proteome to that of other eukaryotes with fully sequenced genomes. We identify the protein(function)s and modules that are specific to humans.

c) The fourth chapter describes the implementation of the methodology in a web server that will be made available to the community.

d) The fifth chapter presents a general discussion of the work, together with perspectives for future developments in the area.

e) The two remaining chapters present the conclusions and bibliography of the thesis.

Chapter 2. *Saccharomyces cerevisiae* as a Model Organism: A Comparative Study[1]

[1] This chapter was published as Karathia, H., et al., *Saccharomyces cerevisiae as a model organism: a comparative study. PLoS One, 2011. 6(2): p. e16015*.

2.1. Abstract

2.1.1. Background

Model organisms are used for research because they provide a framework on which to develop and optimize methods that facilitate and standardize analysis. Such organisms should be representative of the living beings for which they are to serve as proxy. However, in practice, a model organism is often selected *ad hoc*, and without considering its representativeness, because a systematic and rational method to include this consideration in the selection process is still lacking.

2.1.2. Methodology/Principal Findings

In this work we propose such a method and apply it in a pilot study of strengths and limitations of *Saccharomyces cerevisiae* as a model organism. The method relies on the functional classification of proteins into different biological pathways and processes and on full proteome comparisons between the putative model organism and other organisms for which we would like to extrapolate results. Here we compare *S. cerevisiae* to 704 other organisms from various phyla.

For each organism, our results identify the pathways and processes for which *S. cerevisiae* is predicted to be a good model to extrapolate from. We find that animals in general and *Homo sapiens* in particular are some of the non-fungal organisms for which *S. cerevisiae* is likely to be a good model in which to study a significant fraction of common biological processes. We validate our approach by correctly predicting which organisms are phenotypically more distant from *S. cerevisiae* with respect to several different biological processes.

2.1.3. Conclusions/Significance

The method we propose could be used to choose appropriate substitute model organisms for the study of biological processes in other species that are harder to study. For example, one could identify appropriate models to study either pathologies in humans or specific biological processes in species with a long development time, such as plants.

2.2. Introduction

The use of model organisms for research is a hallmark of scientific endeavour (e.g. [**53-59**]). Such organisms are used because a) they may help overcomes ethical and experimental constraints that hold for the target life form, b) they provide a framework on which to develop and optimize analytical methods that facilitate and standardize analysis, and c) they are thought to be representative of a larger class of living beings for whatever biological phenomenon or process the community is interested in. However, the choice of a model organism is often guided more by the first two considerations than by the last one. Nevertheless, selection of a model organism based on accumulated technical experience and on availability of experimental techniques does not guarantee representative results in other organisms. In fact, a gap exists in systematically establishing how close different organisms are with respect to a given process, before choosing one of them as a model for studying that process.

Such a choice should be informed by several considerations. First, the processes of interest for comparison must be clearly identified. Then, one should establish a qualitative or quantitative metric that measures similarity between the different organisms with respect to those processes. Finally, the processes of interest should be sufficiently well characterized in the alternative organisms so that the metric can be used for comparison. If rigorously performed, this final step defeats the purpose of using the model system as a tool to extrapolate from, because all organism would be rigorously characterized beforehand. In fact, this characterization (by proxy) is the purpose of using a model organism. Therefore, methods that rationally predict how similar different organisms might be with respect to biological processes of interest are needed.

The accumulation of fully sequenced genomes [**60**] and the advances in comparative genomics [**61**, **62**] and computational systems biology [**63**] allows us to develop such methods. This can be done by applying strategies that compare the protein or gene networks involved in the process of interest in order to establish a similarity ranking that can be used to predict, to a first approximation, the accuracy of extrapolating the behavior of specific processes between organisms. Testing this idea requires a thorough analysis of the molecular circuits in a well-known model organism and a comparison of these circuits to those in other living beings.

To do this we have chosen the yeast *Saccharomyces cerevisiae* (*S. cerevisiae*) to perform a pilot study. This yeast is one of the most widely used eukaryotic model organisms. It has been used as a model to study aging [**64**], regulation of gene expression [**65**], signal transduction [**66**], cell cycle [**67**], metabolism [**68**, **69**], apoptosis [**70**], neurodegenerative disorders [**71**], and many other biological processes. For example, up to **30%** of genes implicated in human disease may have orthologs in the yeast proteome [**72**].

We use the protein networks that are involved in specific biological processes to compare the differences between *S. cerevisiae* and 704 other organisms, and predict in which organisms the different processes should behave more similarly to the corresponding process in the yeast. We validate some of the predictions by comparing the dynamic behavior of a number of specific pathways in different organisms to that of the corresponding pathway in *S. cerevisiae*.

Our results suggest that the method proposed here is adequate for its purpose. Furthermore, they support the use of *S. cerevisiae* as a model organism to study different processes, while pinpointing specific biological phenomena from this yeast that may not be readily comparable to their analogous processes in other organisms. The method we propose here could be especially relevant to assist in the choice of appropriate model organisms for both, the study of human specific biological processes and the characterization of a specific biological phenomenon in a large class of organisms. It could also be useful in choosing appropriate models for processes in organisms, such as plants, that due to their long duplication times cannot be easily studied.

2.3. Results

2.3.1. Strategy for the comparison of different processes in different organisms

The strategy we use to establish how similar a given process is in two different organisms is as follows. First, we identify orthologs (i.e. genes in different species that evolved from a common ancestral gene by speciation) between the genome of the potential model organism and that of the target organism(s). Then, we attribute function to the different genes in the

organisms under comparison and assign each gene to specific biological processes, using biological ontologies [**73**]. Specifically, we use:

a) The Gene Ontology (GO) [**74**], which has been widely used for annotating function and localization of genes at a coarse level in many organisms [**75-79**], and

b) The pathways that regulate and execute the processes that one is interested in studying, as defined in KEGG [**80**] (one can also use MetaCYC [**81**]).

Finally, we compare the sets of genes responsible for the different processes that are present in each organism. Such an approach predicts if two organisms are likely to be comparable with respect to specific processes of interest, by establishing whether the elements that are a part of the molecular circuits executing the relevant processes are analogous between the organisms (see methods for further details).

Because this is a pilot study, we focus on an organism that is widely used and well characterized, *S. cerevisiae*. We have attributed function to each of the proteins in *S. cerevisiae*, according to the information derived from GO and KEGG. This allowed us to create a functional classification of the proteins with respect to the biological processes that they are involved in. Details about this classification are given in supplementary **Figures of S1** and **Supplementary Tables of S1** materials. With the functional classification of proteins in place, we can compare the different molecular circuits and processes of yeast to their analogues in 704 other organisms.

To compare these molecular circuits and biological processes between *S. cerevisiae* and other organisms, we created clusters of **orthologs** (ScCOGs: *S. cerevisiae* Clusters of Orthologs), **homologues** (ScCHGs: *S. cerevisiae* Clusters of Homologues) and **absent** proteins (ScCAGs: *S. cerevisiae* Clusters of Absent Genes) for each *S. cerevisiae* protein with respect to the translated genome of each of the other 704 organisms. Hereafter we only discuss the results for ScCOGs, because these are consistent with those for ScCHGs. The results for each organism are summarized in Supplementary Table S1. The detailed clusters are provided as supplementary files **ScCOGs.S1.txt** and **ScCHGs.S2.txt**. We are also preparing a server where these results can be further explored and the method can be applied to other organisms.

Each cluster was associated with the functional terms corresponding to its *S. cerevisiae* protein. To analyze the differences between *S. cerevisiae* and a specific organism with

respect to a given process, we compare the fraction of proteins that are annotated as functioning in that process in both organisms. We investigate if orthologs or homologues for each of these proteins are simultaneously present in both organisms or not. Then, we rank organisms with respect to the differences in the set of proteins responsible for each process, analysing for **ScCOGs**, **ScCHGs** and **ScCAGs** at the level of domain, kingdom and phyla for all the 704 organisms (summarized in Supplementary **Table S1.1**, **Table S1.2** and **Table S1. 3**).

2.3.2. Functional comparison of the full *S. cerevisiae* protein complement to that of archaea, bacteria and eukaryotes

We compared how well the proteins in the different ScCOGs, ScCHGs and ScCAGs are conserved between *S. cerevisiae* and various classes of organisms. This allowed us to predict if *S. cerevisiae* can be a good model for specific processes in different classes of organisms, rather than in individual species. The details of the analysis are presented in **Appendix S1**. No *S. cerevisiae* protein has orthologs in all 704 organisms. Furthermore, no *S. cerevisiae* protein has homologues in all the Prokaryotes (Archaea & Bacteria together). In addition, 2642 (45%) *S. cerevisiae* proteins are absent in all the Prokaryotes (for more details see Supplementary **Table S1.2** and **Table S1. 3**).

ARCHEA DOMAIN

We analyzed 48 species of *Archaea*. About 20% (1158) of all *S. cerevisiae* proteins generate ScCOGs that contain *Archaea* sequences. However, only 2% (103) of all yeast proteins generate ScCOGs that contain at least a sequence from each sequenced species of *Archaea*. An additional 18 (0.3%) *S. cerevisiae* proteins have homologues in all *Archaea*. 3672 (62%) *S. cerevisiae* proteins are absent in all *Archaea*. Most of these have unknown function. Overall, there is no group of organisms for which the networks of proteins responsible for a large fraction of biological processes in *S. cerevisiae* are similar to their counterparts in *Archaea*. However, some biological processes are predicted to be similar between *S. cerevisiae* and some *Archaea* (see below).

BACTERIA DOMAIN

We analysed 598 species of *Bacteria*. 1612 (27%) of all *S. cerevisiae* proteins generate ScCOGs that contain bacteria sequences. However, no **ScCOG** or **ScCHG** contains a sequence from each bacterial species. Furthermore, 2881 (49%) *S. cerevisiae* genes are absent from all *Bacteria*, a smaller percentage than that for *Archaea*. As was the case in archaea, overall, there is no group of organisms for which the networks of proteins responsible for a large fraction of biological processes in *S. cerevisiae* are similar to their counterparts in *Bacteria*. However, some biological processes are predicted to be similar between *S. cerevisiae* and some *Bacteria* (see below).

EUKARYOTA DOMAIN

Overall, there are 59 species of *Eukaryotes* in our dataset. About 4.5% (263) of all ScCOGs contain sequences from each of these organisms. Between 40% and 60% of all *S. cerevisiae* proteins involved in "MAPK signalling pathways", "Signal transduction" biological process, and "Helicase activity" molecular functions have orthologs in all 59 species. Furthermore, between 60% and 80% of all proteins involved in "Microtubule organizing centre" of *S. cerevisiae* are also found in all 59 sequenced eukaryotes. Overall, the networks of proteins responsible for a large fraction of biological processes in *S. cerevisiae* are similar to their counterparts in *Ascomycetes*. Furthermore, several biological processes are predicted to be similar between *S. cerevisiae* and other *Eukaryotes*.

2.3.3. Functional comparison of biological processes and pathways between *S. cerevisiae* and other organisms

After getting such a bird's eye view of the similarities and differences between *S. cerevisiae* and different clades of organisms with respect to different biological processes, we now focus on individual organisms. To obtain an approximate estimation of how close a given biological process is between *S. cerevisiae* and another organism we build a matrix of 704x5880 entries. In this matrix, a row represents an organism, while a column represents a ScCOG. The matrix entries are 0 if no sequence from the corresponding organism is found in the appropriate ScCOG and 1 otherwise.

Then, we build a secondary set of four additional matrices containing information about KEGG pathways, biological processes, molecular activity and cellular localization. In

each matrix, the rows represent the organisms and the columns represent the biological process, the cellular localization, the molecular function, or the KEGG pathway. Each entry in one of these matrices is a vector with a variable number of elements that is constant for each column of a matrix. The number of elements in the vector is equal to the number of different proteins that is associated to the specific biological process or pathway corresponding to the column (See methods for details).

Subsequently, we calculate the Normalized Hamming Distance (NHD) between the vector of proteins in one entry of the matrix and the corresponding vector for *S. cerevisiae* from that same column. This NHD is a metric based on the number of elements that are different between the two vectors. The smaller the NHD, the more similar the two vectors are and the more similar the set of proteins executing a specific process in both organisms is. Consequently, the more likely it is that *S. cerevisiae* is a good model to study the relevant process and generalize the results to the other organism. Using this metric we have clustered the organisms in the matrix according to growing overall NHD with respect to *S. cerevisiae*.

KEGG Pathways

Figure 2.1 summarizes the results for KEGG pathways (see Supplementary **Figure S1.2** for a complete analysis). "Benzoate degradation via hydroxylation", "Geraniol degradation", "Propanoate metabolism", "Valine, leucine and isoleucine biosynthesis", "Glycolysis/Gluconeogenesis", "methane metabolism", "Glycolysis/Gluconeogenesis" and "Aminoacyl-t-RNA biosynthesis" are pathways that appear to be similar to those of *S. cerevisiae* in a large fraction of organisms. Pathways such as *S. cerevisiae*'s "RNA polymerase" (29 genes), "Lysosome" (14 genes), "Endocytosis" (33 genes), "Oxidative phosphorylation" (76 genes), "Ribosome" (142 genes), "MAPK signaling pathway - yeast" (55 genes), "DNA replication" (30 genes), and "Ubiquitin mediated proteolysis" (44 genes) and "Nucleotide excision repair" (34 genes) are much more similar to those from other eukaryotes than to the corresponding prokaryotic pathways (when they exist). Among the pathways that are central for life, the one that appears to be more unique to *S. cerevisiae* and other Saccharomycetes is cell cycle (115 genes), because only a small fraction of its proteins have orthologs in other eukaryotes. Thus, these results suggest that extrapolating cell cycle studies in *S. cerevisiae* to other organisms outside of the Saccharomycetes clade should be done only at the level of basic principles, if at all (see for example [82, 83]).

Figure 2.1

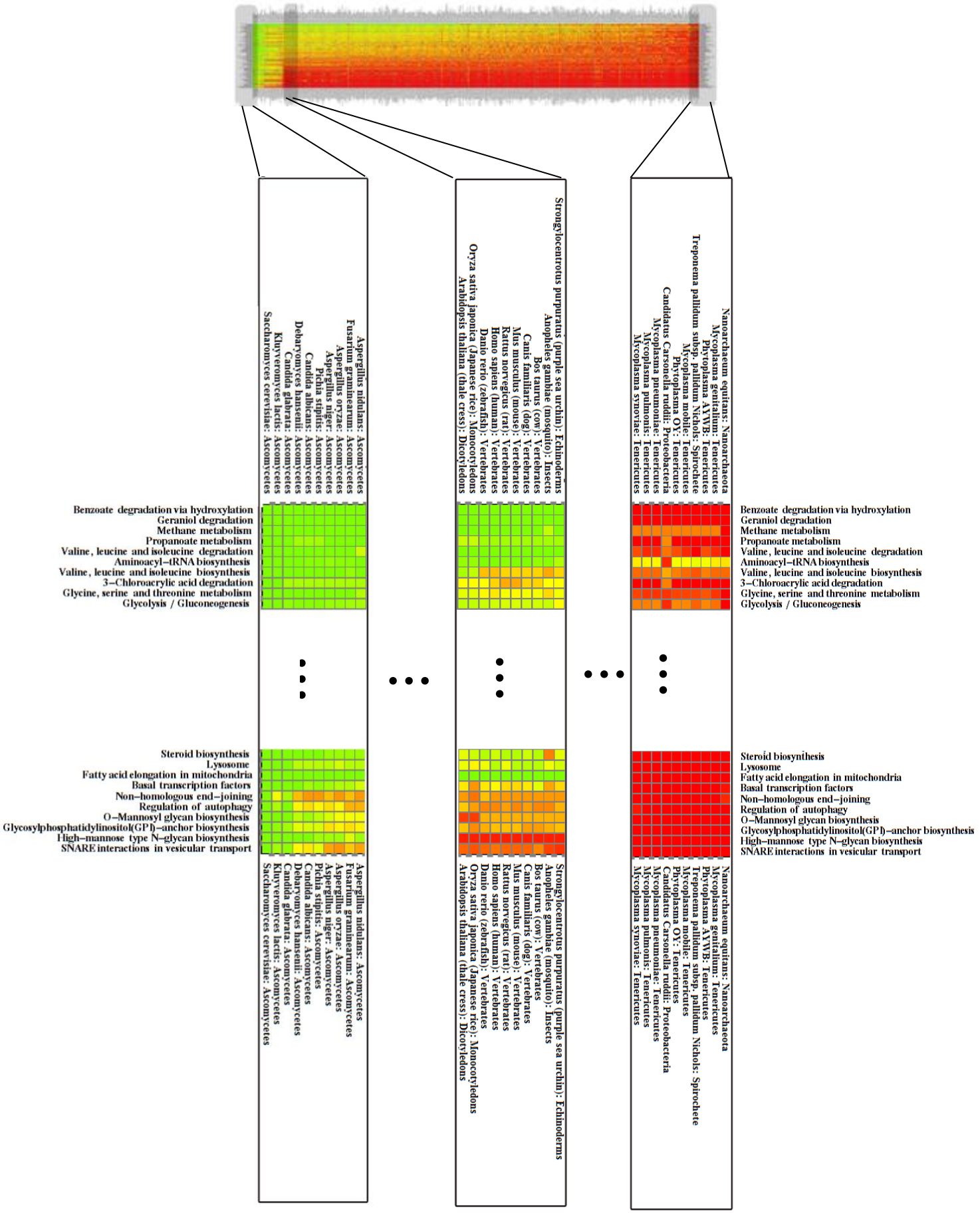

Nanoarchaeum equitans: Nanoarchaeota
Mycoplasma genitalium: Tenericutes
Phytoplasma AYWB: Tenericutes
Treponema pallidum subsp. pallidum Nichols: Spirochete
Mycoplasma mobile: Tenericutes
Phytoplasma OY: Tenericutes
Candidatus Carsonella ruddii: Proteobacteria
Mycoplasma pneumoniae: Tenericutes
Mycoplasma pulmonis: Tenericutes
Mycoplasma synoviae: Tenericutes
Strongylocentrotus purpuratus (purple sea urchin): Echinoderms
Anopheles gambiae (mosquito): Insects
Bos taurus (cow): Vertebrates
Canis familiaris (dog): Vertebrates
Mus musculus (mouse): Vertebrates
Rattus norvegicus (rat): Vertebrates
Homo sapiens (human): Vertebrates
Danio rerio (zebrafish): Vertebrates
Oryza sativa japonica (Japanese rice): Monocotyledons
Arabidopsis thaliana (thale cress): Dicotyledons
Aspergillus nidulans: Ascomycetes
Fusarium graminearum: Ascomycetes
Aspergillus oryzae: Ascomycetes
Aspergillus niger: Ascomycetes
Pichia stipitis: Ascomycetes
Candida albicans: Ascomycetes
Debaryomyces hansenii: Ascomycetes
Candida glabrata: Ascomycetes
Kluyveromyces lactis: Ascomycetes
Saccharomyces cerevisiae: Ascomycetes
Benzoate degradation via hydroxylation
Geraniol degradation
Methane metabolism
Propanoate metabolism
Valine, leucine and isoleucine degradation
Aminoacyl−tRNA biosynthesis
Valine, leucine and isoleucine biosynthesis
3−Chloroacrylic acid degradation
Glycine, serine and threonine metabolism
Glycolysis / Gluconeogenesis
Steroid biosynthesis
Lysosome
Fatty acid elongation in mitochondria
Basal transcription factors
Non−homologous end−joining
Regulation of autophagy
O−Mannosyl glycan biosynthesis
Glycosylphosphatidylinositol(GPI)−anchor biosynthesis
High−mannose type N−glycan biosynthesis
SNARE interactions in vesicular transport

Figure 2.1 Details of a heat-map representation showing how distant each organism is from *S. cerevisiae* with respect to each individual KEGG pathway. A green square indicates a high level of coincidence between the set of proteins involved in the specific pathway (column) in a given organism (row) and the set of proteins for the same pathway in *S. cerevisiae*. A red square indicates complete absence of the set of proteins involved in the specific pathway (column) in a given organism (row) with respect to the same pathway in *S. cerevisiae*. Intermediate colours indicate intermediate degrees of coincidence between the set of proteins in the target organism and that in *S. cerevisiae*. The complete heat-map can be seen in **Figure S1.2**.

An encouraging observation for the use of *S. cerevisiae* as a model organism for mammals is that most of the studied mammals (humans, dogs, mice, cows and rats) are among the non-fungal organisms that have biological processes with protein sets that are similar to the corresponding sets of *S. cerevisiae*. Specifically, the sets of *S. cerevisiae* proteins that are associated to "Mismatch repair" (18 genes), "Ubiquitin and other terpenoid-quinone biosynthesis" (5 genes), "Inositol phosphate metabolism" (15 genes), "Steroid biosynthesis" (15 genes), "Ubiquitin mediated proteolysis" (44 genes), "DNA replication" (30 genes), "Ribosome" (142 genes), "Proteasome" (35 genes), "Galactose metabolism" (23 genes), "One carbon pool by folate" (14 genes) and "Glycolysis/gluconeogenesis" (48 genes) are those that appear to be more similar to the corresponding sets of proteins in man. A more thorough analysis is given in the Supplementary **Appendix 1**.

GO Biological Processes, Cellular Component and Molecular Function

Figure 2.2, **Figure 2.3** and **Figure 2.4** summarize the results for the comparisons between *S. cerevisiae* and the other organisms using the GO categories classification. Details can be further analyzed in Supplementary **Figure S1.3**, **Figure S1.4** and **Figure S1.5**. The results are similar to those described for **Figure 2.1** (or those reported in Supplementary **Figure S1.2**), which suggests that these functional classifications are, to a large extent, equivalent, in spite of all problems that they might have (see discussion).

Figure 2.2

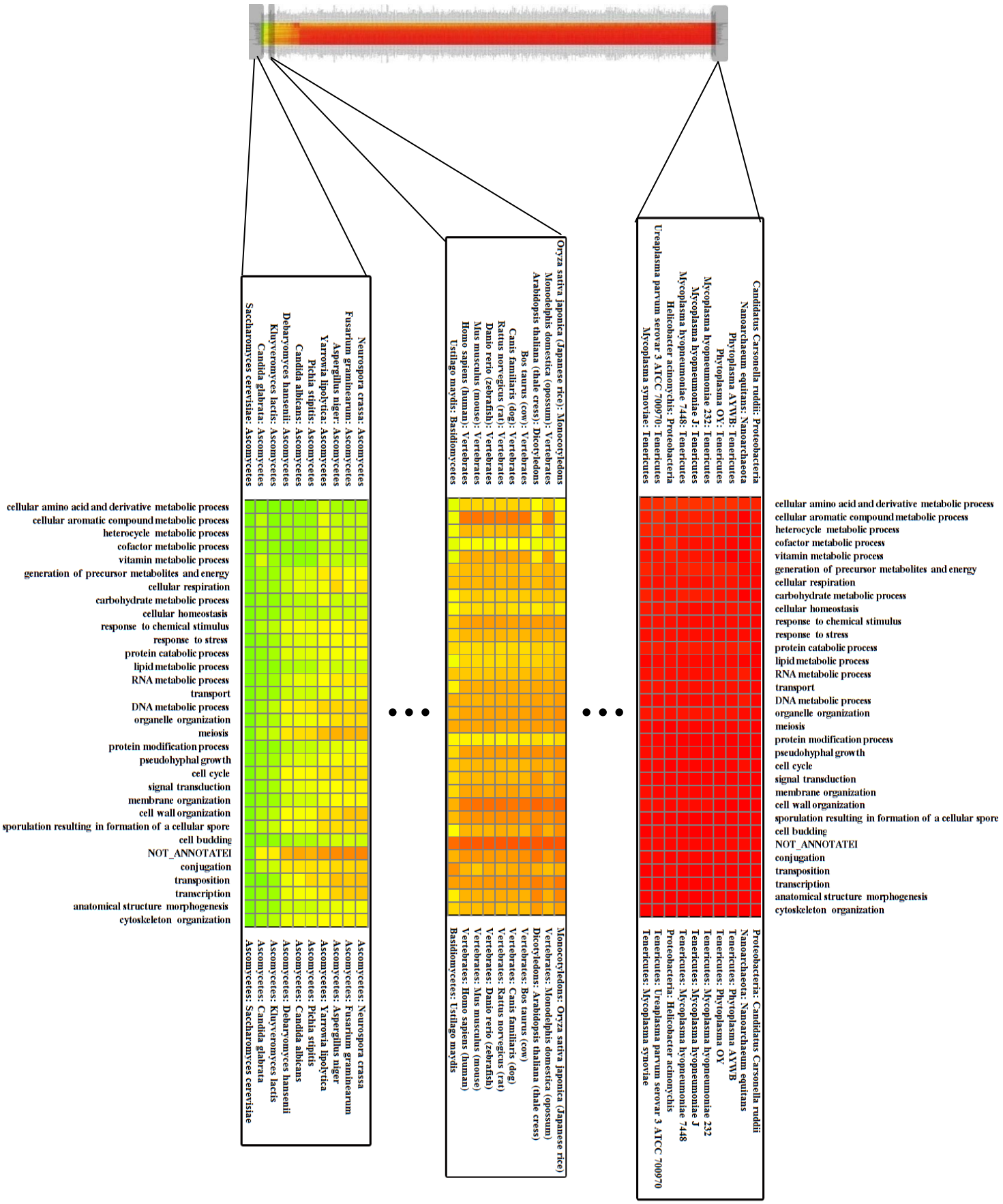

Candidatus Carsonella ruddii: Proteobacteria
Nanoarchaeum equitans: Nanoarchaeota
Phytoplasma AYWB: Tenericutes
Phytoplasma OY: Tenericutes
Mycoplasma hyopneumoniae 232: Tenericutes
Mycoplasma hyopneumoniae J: Tenericutes
Mycoplasma hyopneumoniae 7448: Tenericutes
Helicobacter acinonychis: Proteobacteria
Ureaplasma parvum serovar 3 ATCC 700970: Tenericutes
Mycoplasma synoviae: Tenericutes
Oryza sativa japonica (Japanese rice): Monocotyledons
Monodelphis domestica (opossum): Vertebrates
Arabidopsis thaliana (thale cress): Dicotyledons
Bos taurus (cow): Vertebrates
Canis familiaris (dog): Vertebrates
Rattus norvegicus (rat): Vertebrates
Danio rerio (zebrafish): Vertebrates
Mus musculus (mouse): Vertebrates
Homo sapiens (human): Vertebrates
Ustilago maydis: Basidiomycetes
Neurospora crassa: Ascomycetes
Fusarium graminearum: Ascomycetes
Aspergillus niger: Ascomycetes
Yarrowia lipolytica: Ascomycetes
Pichia stipitis: Ascomycetes
Candida albicans: Ascomycetes
Debaryomyces hansenii: Ascomycetes
Kluyveromyces lactis: Ascomycetes
Candida glabrata: Ascomycetes
Saccharomyces cerevisiae: Ascomycetes
cellular amino acid and derivative metabolic process
cellular aromatic compound metabolic process
heterocycle metabolic process
cofactor metabolic process
vitamin metabolic process
generation of precursor metabolites and energy
cellular respiration
carbohydrate metabolic process
cellular homeostasis
response to chemical stimulus
response to stress
protein catabolic process
lipid metabolic process
RNA metabolic process
transport
DNA metabolic process
organelle organization
meiosis
protein modification process
pseudohyphal growth
cell cycle
signal transduction
membrane organization
cell wall organization
sporulation resulting in formation of a cellular spore
cell budding
NOT_ANNOTATEI
conjugation
transposition
transcription
anatomical structure morphogenesis
cytoskeleton organization
Proteobacteria: Candidatus Carsonella ruddii
Nanoarchaeota: Nanoarchaeum equitans
Tenericutes: Phytoplasma AYWB
Tenericutes: Phytoplasma OY
Tenericutes: Mycoplasma hyopneumoniae 232
Tenericutes: Mycoplasma hyopneumoniae J
Tenericutes: Mycoplasma hyopneumoniae 7448
Proteobacteria: Helicobacter acinonychis
Tenericutes: Ureaplasma parvum serovar 3 ATCC 700970
Tenericutes: Mycoplasma synoviae
Monocotyledons: Oryza sativa japonica (Japanese rice)
Vertebrates: Monodelphis domestica (opossum)
Dicotyledons: Arabidopsis thaliana (thale cress)
Vertebrates: Bos taurus (cow)
Vertebrates: Canis familiaris (dog)
Vertebrates: Rattus norvegicus (rat)
Vertebrates: Danio rerio (zebrafish)
Vertebrates: Mus musculus (mouse)
Vertebrates: Homo sapiens (human)
Basidiomycetes: Ustilago maydis
Ascomycetes: Neurospora crassa
Ascomycetes: Fusarium graminearum
Ascomycetes: Aspergillus niger
Ascomycetes: Yarrowia lipolytica
Ascomycetes: Pichia stipitis
Ascomycetes: Candida albicans
Ascomycetes: Debaryomyces hansenii
Ascomycetes: Kluyveromyces lactis
Ascomycetes: Candida glabrata
Ascomycetes: Saccharomyces cerevisiae

Figure 2.2 Details of a heat-map representation showing how distant each organism is from *S. cerevisiae* with respect to each biological process from the GOSLIM classification. A green square indicates a high level of coincidence between the set of proteins involved in the specific biological process (column) in a given organism (row) and the set of proteins for the same pathway in *S. cerevisiae*. A red square indicates complete absence of the set of proteins involved in the specific pathway (column) in a given organism (row) with respect to the same biological process in *S. cerevisiae*. Intermediate colours indicate intermediate degrees of coincidence between the set of proteins in the target organism and that in *S. cerevisiae*. The complete heat-map can be seen in **Figure S1.2**.

S. cerevisiae metabolic activities like "Cellular amino acid and derivative metabolic process", "Cellular aromatic compound metabolic process", "Heterocycle metabolic process", "Cofactor metabolic process" and "Vitamin metabolic process" are the ones that are more conserved in all organisms. In contrast, "cytoskeleton organization", "Transcription", "Anatomical structure morphogenesis", "Transposition", "conjugation", "Cell budding", and "Protein modification process" appear to be conserved mostly in eukaryotes. Conservation of the "Cell wall organization" pathway is restricted to fungi. A more detailed analysis of these pathways and their similarity between *S. cerevisiae* and the other 704 organisms can be found in the appendix and in Supplementary **Figure S1.2**.

Figure 2.3

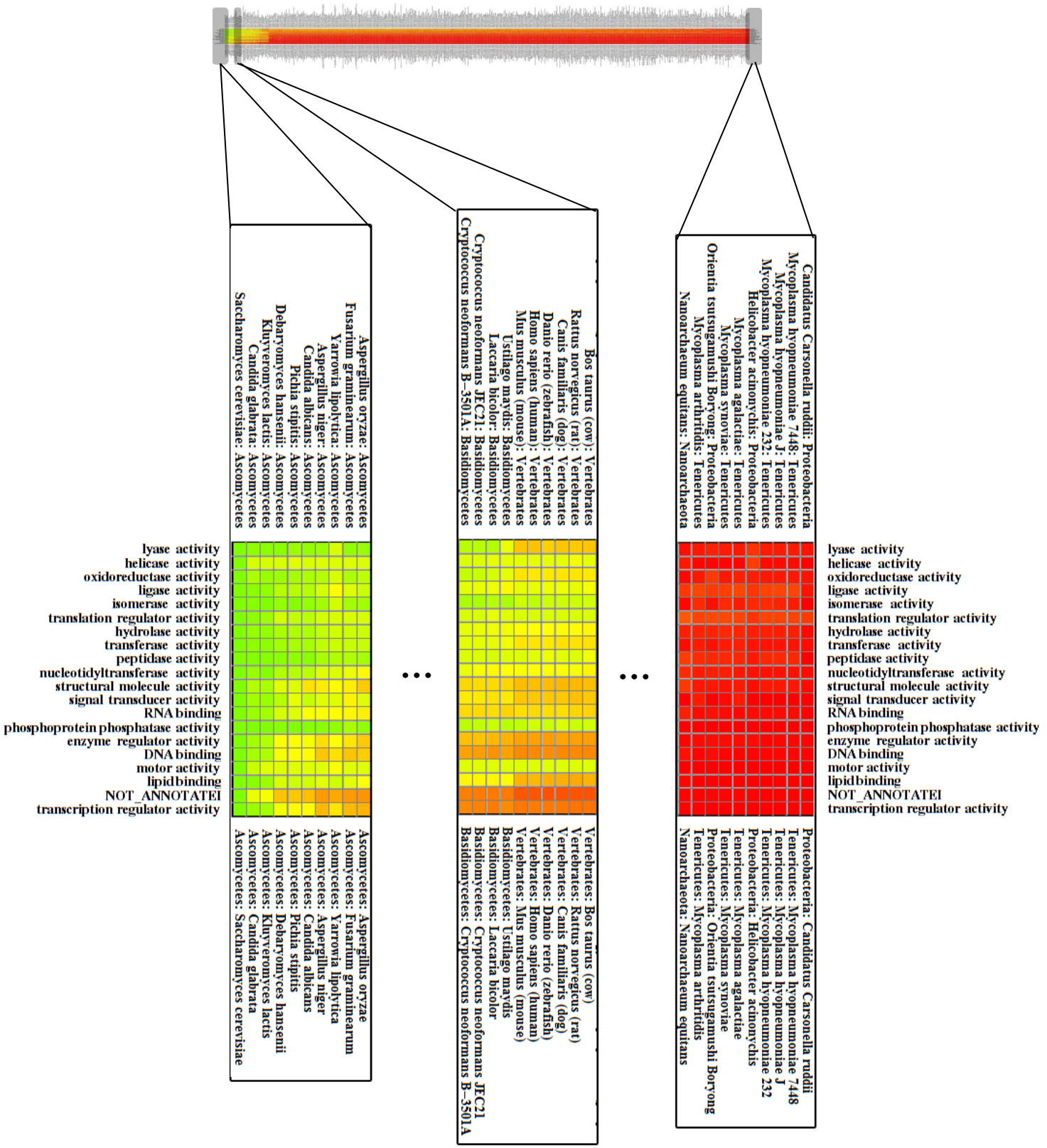

Candidatus Carsonella ruddii: Proteobacteria
Mycoplasma hyopneumoniae 7448: Tenericutes
Mycoplasma hyopneumoniae J: Tenericutes
Mycoplasma hyopneumoniae 232: Tenericutes
Helicobacter acinonychis: Proteobacteria
Mycoplasma agalactiae: Tenericutes
Mycoplasma synoviae: Tenericutes
Orientia tsutsugamushi Boryong: Proteobacteria
Mycoplasma arthritidis: Tenericutes
Nanoarchaeum equitans: Nanoarchaeota
Bos taurus (cow): Vertebrates
Rattus norvegicus (rat): Vertebrates
Canis familiaris (dog): Vertebrates
Danio rerio (zebrafish): Vertebrates
Homo sapiens (human): Vertebrates
Mus musculus (mouse): Vertebrates
Ustilago maydis: Basidiomycetes
Laccaria bicolor: Basidiomycetes
Cryptococcus neoformans JEC21: Basidiomycetes
Cryptococcus neoformans B-3501A: Basidiomycetes
Aspergillus oryzae: Ascomycetes
Fusarium graminearum: Ascomycetes
Yarrowia lipolytica: Ascomycetes
Aspergillus niger: Ascomycetes
Candida albicans: Ascomycetes
Pichia stipitis: Ascomycetes
Debaryomyces hansenii: Ascomycetes
Kluyveromyces lactis: Ascomycetes
Candida glabrata: Ascomycetes
Saccharomyces cerevisiae: Ascomycetes
lyase activity
helicase activity
oxidoreductase activity
ligase activity
isomerase activity
translation regulator activity
hydrolase activity
transferase activity
peptidase activity
nucleotidyltransferase activity
structural molecule activity
signal transducer activity
RNA binding
phosphoprotein phosphatase activity
enzyme regulator activity
DNA binding
motor activity
lipid binding
NOT_ANNOTATEI
transcription regulator activity
Proteobacteria: Candidatus Carsonella ruddii
Tenericutes: Mycoplasma hyopneumoniae 7448
Tenericutes: Mycoplasma hyopneumoniae J
Tenericutes: Mycoplasma hyopneumoniae 232
Proteobacteria: Helicobacter acinonychis
Tenericutes: Mycoplasma agalactiae
Tenericutes: Mycoplasma synoviae
Proteobacteria: Orientia tsutsugamushi Boryong
Tenericutes: Mycoplasma arthritidis
Nanoarchaeota: Nanoarchaeum equitans
Vertebrates: Bos taurus (cow)
Vertebrates: Rattus norvegicus (rat)
Vertebrates: Canis familiaris (dog)
Vertebrates: Danio rerio (zebrafish)
Vertebrates: Homo sapiens (human)
Vertebrates: Mus musculus (mouse)
Basidiomycetes: Ustilago maydis
Basidiomycetes: Laccaria bicolor
Basidiomycetes: Cryptococcus neoformans JEC21
Basidiomycetes: Cryptococcus neoformans B-3501A
Ascomycetes: Aspergillus oryzae
Ascomycetes: Fusarium graminearum
Ascomycetes: Yarrowia lipolytica
Ascomycetes: Aspergillus niger
Ascomycetes: Candida albicans
Ascomycetes: Pichia stipitis
Ascomycetes: Debaryomyces hansenii
Ascomycetes: Kluyveromyces lactis
Ascomycetes: Candida glabrata
Ascomycetes: Saccharomyces cerevisiae

Figure 2.3 Details of a heat-map representation showing how distant each organism is from *S. cerevisiae* with respect to each molecular function from the GOSLIM classification. A green square indicates a high level of coincidence between the set of proteins involved in the specific molecular function (column) in a given organism (row) and the set of proteins for the same pathway in *S. cerevisiae*. A red square indicates complete absence of the set of proteins involved in the specific pathway (column) in a given organism (row) with respect to the same molecular function in *S. cerevisiae*. Intermediate colours indicate intermediate degrees of coincidence between the set of proteins in the target organism and that in *S. cerevisiae*. The complete heat-map can be seen in **Figure S1.4**.

Figure 2.4

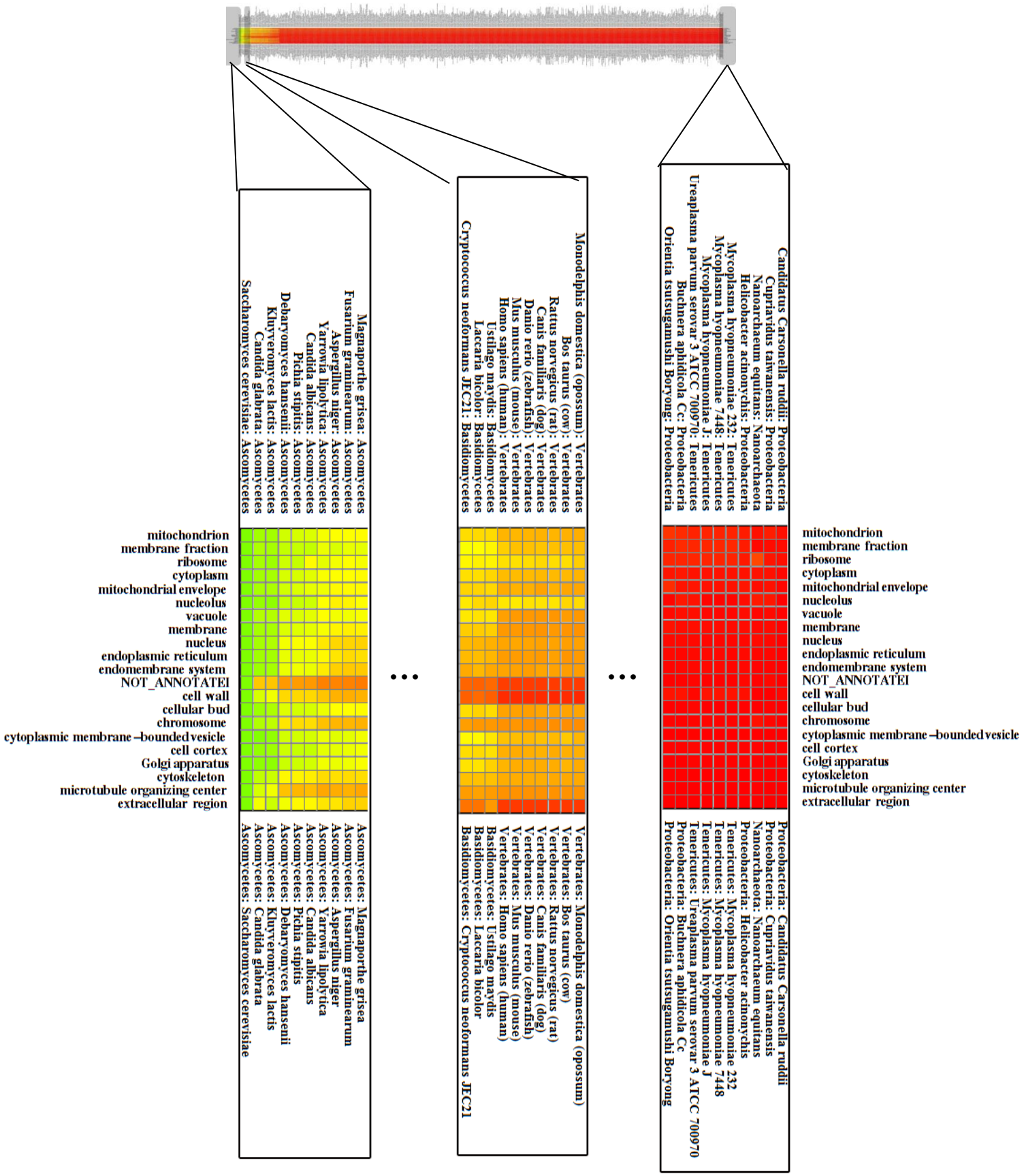

mitochondrion
membrane fraction
ribosome
cytoplasm
mitochondrial envelope
nucleolus
vacuole
membrane
nucleus
endoplasmic reticulum
endomembrane system
NOT_ANNOTATEI
cell wall
cellular bud
chromosome
cytoplasmic membrane–bounded vesicle
cell cortex
Golgi apparatus
cytoskeleton
microtubule organizing center
extracellular region
Candidatus Carsonella ruddii: Proteobacteria
Cupriavidus taiwanensis: Proteobacteria
Nanoarchaeum equitans: Nanoarchaeota
Helicobacter acinonychis: Proteobacteria
Mycoplasma hyopneumoniae 232: Tenericutes
Mycoplasma hyopneumoniae 7448: Tenericutes
Mycoplasma hyopneumoniae J: Tenericutes
Ureaplasma parvum serovar 3 ATCC 700970: Tenericutes
Buchnera aphidicola Cc: Proteobacteria
Orientia tsutsugamushi Boryong: Proteobacteria
Proteobacteria: Candidatus Carsonella ruddii
Proteobacteria: Cupriavidus taiwanensis
Nanoarchaeota: Nanoarchaeum equitans
Proteobacteria: Helicobacter acinonychis
Tenericutes: Mycoplasma hyopneumoniae 232
Tenericutes: Mycoplasma hyopneumoniae 7448
Tenericutes: Mycoplasma hyopneumoniae J
Tenericutes: Ureaplasma parvum serovar 3 ATCC 700970
Proteobacteria: Buchnera aphidicola Cc
Proteobacteria: Orientia tsutsugamushi Boryong
Monodelphis domestica (opossum): Vertebrates
Bos taurus (cow): Vertebrates
Rattus norvegicus (rat): Vertebrates
Canis familiaris (dog): Vertebrates
Danio rerio (zebrafish): Vertebrates
Mus musculus (mouse): Vertebrates
Homo sapiens (human): Vertebrates
Ustilago maydis: Basidiomycetes
Laccaria bicolor: Basidiomycetes
Cryptococcus neoformans JEC21: Basidiomycetes
Vertebrates: Monodelphis domestica (opossum)
Vertebrates: Bos taurus (cow)
Vertebrates: Rattus norvegicus (rat)
Vertebrates: Canis familiaris (dog)
Vertebrates: Danio rerio (zebrafish)
Vertebrates: Mus musculus (mouse)
Vertebrates: Homo sapiens (human)
Basidiomycetes: Ustilago maydis
Basidiomycetes: Laccaria bicolor
Basidiomycetes: Cryptococcus neoformans JEC21
Magnaporthe grisea: Ascomycetes
Fusarium graminearum: Ascomycetes
Aspergillus niger: Ascomycetes
Yarrowia lipolytica: Ascomycetes
Candida albicans: Ascomycetes
Pichia stipitis: Ascomycetes
Debaryomyces hansenii: Ascomycetes
Kluyveromyces lactis: Ascomycetes
Candida glabrata: Ascomycetes
Saccharomyces cerevisiae: Ascomycetes
Ascomycetes: Magnaporthe grisea
Ascomycetes: Fusarium graminearum
Ascomycetes: Aspergillus niger
Ascomycetes: Yarrowia lipolytica
Ascomycetes: Candida albicans
Ascomycetes: Pichia stipitis
Ascomycetes: Debaryomyces hansenii
Ascomycetes: Kluyveromyces lactis
Ascomycetes: Candida glabrata
Ascomycetes: Saccharomyces cerevisiae

Figure 2.4 Details of a heat-map representation showing how distant each organism is from *S. cerevisiae* with respect to each cellular component from the GOSLIM classification. A green square indicates a high level of coincidence between the set of proteins involved in the specific molecular function (column) in a given organism (row) and the set of proteins for the same pathway in *S. cerevisiae*. A red square indicates complete absence of the set of proteins involved in the specific pathway (column) in a given organism (row) with respect to the same molecular function in *S. cerevisiae*. Intermediate colours indicate intermediate degrees of coincidence between the set of proteins in the target organism and that in *S. cerevisiae*. The complete heat-map can be seen in **Figure S1.5**.

2.3.4. Validating the predictions

The analysis described above and the results given in **Figure 2.1-Figure 2.4** and in Supplementary **Figure S1.2, Figure S1.3, Figure S1.4 and Figure S1.5** ranks the difference between the proteins set responsible for a given biological process in each organism and the corresponding set in *S. cerevisiae*. If our earlier arguments are correct, one would expect that the similarity between the adaptive responses that involve a given process in other organisms and the same responses in *S. cerevisiae* is directly correlated to the similarity between the protein sets that regulate and execute that process.

In other words, we define a static metric of closeness of processes between organisms that is based solely on the similarity between the sets of proteins involved in those processes in both organisms. Can we assume that such a metric is also a good measure of closeness between physiological and adaptive responses of the pathways regulating the processes in the organisms being compared, even though it does not include any kinetic or regulatory information?

To answer this question we selected pathways for which dynamic, regulatory, and/or phenotypic information was available for *S. cerevisiae* and for a scope of different organisms. This selection was based on a careful analysis of Supplementary **Figure S1.1**. We systematically identified pathways or processes with more than 4 genes and then searched the literature for comparable studies of the dynamical and adaptive behaviour of these processes in different organisms that belong to our dataset. We were able to identify twelve cases that could be used to answer the question from the previous paragraph. The results are summarized in Supplementary **Table S1.4.**

They show that the phenotypic adaptations and dynamical behaviour of a given pathway is more similar to that of *S. cerevisiae* in organisms that are found to be closer to *S. cerevisiae* according to our analysis than in more distant organisms. Thus, even if the method

we propose is based on static information, the results of the analysis appear to be adequate for pinpointing an appropriate model organism from which to study and extrapolate the dynamical and adaptive behavior of specific biological processes.

2.4. Discussion

2.4.1. The rational choice of model organisms and its technical limitations

In this work we ask the question "How can one chose an appropriate model organism in which to study a specific biological process in such a way that the results may be extrapolated to another organism?" We propose a systematic way to answer this question that involves comparing the similarity between the set of proteins that participate in the biological process of interest in the organism to the equivalent set of proteins in the organism to which we want to extrapolate the results. The closer the set of proteins is between the two, the more likely it is that the results from one organism can be extrapolated to the other. To compare the sets of proteins between organisms, we propose a procedure that involves: a) associating a protein to a process or pathway, for example using GO categories or the KEGG pathways, and b) compare the sets of proteins associated to the process between the relevant organisms. This method offers a proxy for establishing probable equivalency of processes between organisms, but it has some drawbacks.

First, more often than not, there will be little functional information associated to the proteins of a given organism. To overcome such a problem, we propose choosing an initial subject organism that is well studied and functionally well characterized at the molecular level. As our method relies on ortholog identification and functional annotation, it requires that this annotation be continuously improved even in well studied organisms. By choosing *S. cerevisiae* as an example we use the eukaryotic organism that we believe has the best overall functional annotation. It must also be emphasized that, when comparing the set of proteins that participate in a given process in different organisms, one must consider the "super set" of proteins participating in that process and compare the differences. In other words, for example when comparing KEGG pathways, one can consider the pathway that includes all possible E.C. numbers and then compare the two organisms in this context. This was also

done here. Otherwise, one may find a situation where two organisms are predicted as being good models with respect to a given process when the proteins in one organism are a small subset of those in the other.

Second, using sequence similarity to establish functional orthology also has its drawbacks. On one hand, sometimes functional orthology exists even in the absence of sequence orthology and vice versa. Comparing the structures of proteins as well as their amino acid motifs and active centres provides some assistance in tackling this problem. However, at the current stage of development in bioinformatics, sequence comparison is still the most efficient and accurate way to make such predictions on the scale that we made them for this work. On the other hand, sometimes, due to gene duplication and domain shuffling, proteins that are unique in one organism may have several close sequence homologues in another. We address this problem by proposing a procedure that takes several similarity factors between sequences into account before deciding which of the homologues is the more likely to be orthologous to the query protein. These factors include e-value score, similarity of the sequences and the fraction of the two proteins that is comparable. Nevertheless, if one also analyzes homologues separately, as we also do here, one stands a better chance of controlling for false negative orthologs.

Third, by comparing only the set of proteins associated with a given biological process in different organisms, we are disregarding regulatory and dynamic information that could be important for the comparison. This shortcoming may not be problematic. On one hand our method is a good way to eliminate processes and organisms for which the reference organism is not a good model. If the sets of proteins that execute a given process are very dissimilar, then the dynamics are not even an issue because other model organisms need to be chosen. On the other hand, having a more similar set of proteins associated to a specific process makes it more likely that the adaptive and regulatory responses of the process be similar. This claim can be supported by comparing the physiological responses of different organisms to that of the model organism (see below).

Fourth, sometimes the logic used to define the proteins associated to specific biological pathways or processes is questionable. This is a very important factor and a successful general application of the method described here requires that the annotation of genomes and ontologies/pathways keeps on improving. Poorly characterized biological processes will lead to greater errors in the comparisons. There is little we can do with respect

to this limitation at this time. One of the actions that can be taken to minimize this problem is to choose as a model an organism that is one of the best annotated worldwide. We did so by choosing *S. cerevisiae* as a model for the study. This organism has the additional advantages of being well characterized at the molecular level and used to study many biological processes that are important in other organisms. To further ameliorate this problem we carefully curated both the KEGG and GO associations of yeast.

2.4.2. *S. cerevisiae* as a model organism

We apply our method to a pilot study of *S. cerevisiae* as a model organism, by comparing it to 704 other organisms. The results are presented in detail in Supplementary **Figure S1.2, Figure S1.3, Figure S1.4** and **Figure S1.5** and Supplementary **Table S1.1**, **Table S1.2**, and **Table S1. 3**. In *S. cerevisiae* 4571 proteins are not associated to any pathway in the KEGG database. Analyzing the approximately 1000 proteins that have such a functional association, we find that, as expected, in many cases evolutionary closeness goes on par with similarity between sets of proteins that are associated to a specific biological process.

As mentioned above, our inference of closeness between *S. cerevisiae* and the other organisms is based upon an analysis of similarity between the sets of proteins involved in a specific process in both organisms. This analysis does not include any information about the physiological responses and the dynamic or regulatory aspects of the biological processes and pathways being compared between organisms. To understand if this limitation is in general important we selected pathways for which dynamic, regulatory, and phenotypic information was available for *S. cerevisiae* and for a scope of other different organisms. We then compare the behaviour of those pathways in yeast and in the other organisms. In this comparison, organisms that are predicted to be closer for a specific pathway or process also have more similar adaptive responses (Supplementary **Table S1.4**). Furthermore, recent work that uses orthology between human genes and those in other organisms to find models for human diseases support these results [84-86]. Together, this suggests that our method is adequate both for eliminating unsuitable model organisms and for choosing an appropriate model organism from which to study and extrapolate the dynamical and adaptive behaviour of specific biological processes.

2.5. Conclusion

Our results support the use of *S. cerevisiae* as a model organism to study different biological processes and pathways in specific organisms, while pinpointing specific processes in this yeast that may not be readily generalizable to other organisms. We conclude that using a single proteome as a reference and applying a methodology such as the one suggested here, one can in general appropriately select model organisms to study the dynamic and adaptive responses of a given biological process, as long as the proteins that participate in that process are known.

2.6. Materials and Methods

2.6.1. Selection of genome sequences

The complete proteome of *Saccharomyces cerevisiae* (5880 proteins) was downloaded from NCBI (December 2009). The complete sequences for the full protein complement of 704 organisms with fully sequenced genomes was downloaded from the KEGG database (December 2009) and cross-referenced to that provided the NCBI database.

2.6.2. Homology analysis

We downloaded BLAST version 2.2.18 from NCBI and used it locally. All genome and protein sequences were formatted using FormatDB. A pipeline for selecting orthologous proteins, homologous proteins and proteins of the *S. cerevisiae* that are absent in each of the other organisms was developed and implemented in PERL.

2.6.3. Orthology analysis

The collection of all proteins in a target genome that blasted against a specific protein of *S. cerevisiae* with an $\boldsymbol{e-value \leq 10^{-10}}$ was analyzed. Manually and through the comparison of the *S. cerevisiae* proteome to that of two organisms from each class, we setup a cut-off value for separating orthologs from homologues. Pairs of proteins with e-value between $\mathbf{10^{-10}}$ and $\mathbf{10^{-36}}$ and identity score below 30% are considered as homologues. If the alignment spans over 85% of either sequence and either the e-value of the blast search is bellow $\mathbf{10^{-36}}$ or the identity score is higher than 30%, both proteins are considered as belonging to the same family of orthologs [**87**]. When more than one protein in a target genome meets these

conditions with respect to the same *S. cerevisiae* protein we calculate an orthology score function, ***F***. The protein with the highest ***F***-score function is considered to be the most likely ortholog with respect to the *S. cerevisiae* protein, while the remaining proteins are flagged as in-paralogs of that ortholog. ***F*** is defined as follows:

$$\boldsymbol{F = (F1 + F2) - F3} \qquad \textbf{Eq. 1}$$

Factor ***F1*** is calculated as follows.

$$\boldsymbol{F1 = 1 - (S - I)/S} \qquad \textbf{Eq. 2}$$

In Eq. 2, ***S*** represents the similarity score and ***I*** represent the identity score of the alignment. ***F1*** is always between 0 and 1. The more similar two sequences are, the closer to 1 will ***F1*** be.

Factor ***F2*** is calculated as follows.

$$\boldsymbol{F2 = AL/PL} \qquad \textbf{Eq. 3}$$

In Eq. 3, ***AL*** represents the length of the alignment, and ***PL*** is the total length of the query sequence. F2 is always between 0 and 1. The larger the fraction of the query sequence that aligns with the target sequence is, the more similar the two proteins will be and the closer to 1 will ***F2*** be.

Finally, factor ***F3*** is calculated as follows.

$$\boldsymbol{F3 = (G1/L1) + (G2/L2)} \qquad \textbf{Eq. 4}$$

In Eq. 4, ***G1*** represents the number of gaps within the aligned region of the query sequence, ***L1*** represents the length of the query sequence, ***G2*** represents the number of gaps within the aligned region of the target sequence, and ***L2*** represents the full length of the target sequence. The closer to zero ***F3*** is the more similar will the two sequences be.

Theoretically, $-\infty \leq \boldsymbol{F} \leq \boldsymbol{2}$. However, in practice, we found that ***F*** typically assumes values between 0 and 2. The higher ***F*** is, the more likely it is that the query and target sequence are orthologs. The whole process is summarized in **Figure 2.5.** At the end of the analysis we obtain clusters of orthologs (ScCOGs) and homologues (ScCHGs) for all the *S. cerevisiae* genes with respect to the other 704 organisms. We also obtain a third family of clusters (ScCAGs), that of proteins from *S. cerevisiae* that are absent from the target genomes.

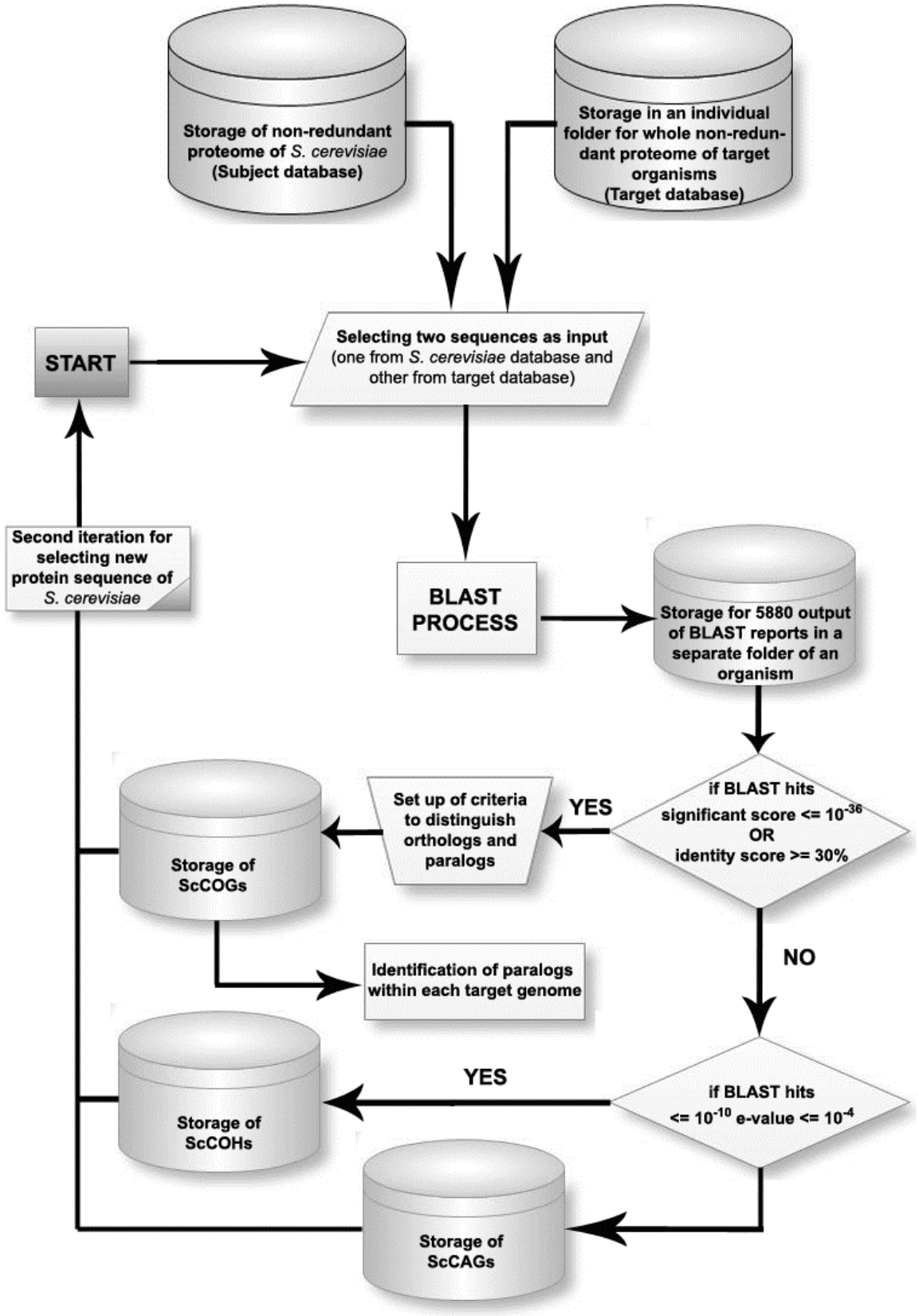


Figure 2.5 **Summary of the process used to build ScCOGs, ScCHGs and ScCAGs. The full proteome of *S. cerevisiae* was compared to the full proteome of each of 704 different organisms using BLAST. See methods for details.**

2.6.4. Classification of clusters according to pathways and biological processes

In order to attribute biological function to the ScCOGs, ScCHGs and ScCAGs, we implemented the following procedure. On one hand, we used the GOSLIM classification of gene function for *S. cerevisiae* from SGD [88, 89] to attribute biological function, molecular functions and cellular localization to each cluster. On the other, we downloaded data from KEGG that associates genes to KEGG metabolic circuits in fully sequenced genomes [90] and attribute pathways terms to each of the clusters.

2.6.5. Calculation of the Hamming distance

The Hamming Distance (HD) between the vector $\overrightarrow{V1}$ of protein functions associated to a specific process, localization or pathway in *S. cerevisiae* and the vector $\overrightarrow{V2}$ of corresponding protein functions in another organism gives a measure of how different the two vectors are. It is calculated using the formula $\boldsymbol{HD = \sum_{i=1}^{n}(1 - \delta_i)}$ where δ_i is the Kronecker delta. δ_i is 1 if the elements in position *i* of both vectors are orthologs and 0 otherwise. The smaller the distance, the more similar the two vectors are and the more similar is the set of genes executing a specific process in both organisms. HD can be normalized (NHD) by dividing it by the maximum HD between corresponding vectors of all organisms. Consequently, the smaller NHD is, the more likely that *S. cerevisiae* is a good model to study the relevant process or pathways and generalize the results for the other organism. The vectors we define for each pathway include all proteins that could participate in that pathway in all organisms in the KEGG database. This ensures that the comparison we are making accounts for differences between the pathway in *S. cerevisiae* and that in the other organism and vice-versa. All calculations were performed using Mathematica [91].

2.7. Supporting Materials

2.7.1. Appendix 1 - Detailed functional analysis of *S. cerevisiae* as a model organism

The 5880 non-redundant genes of *S. cerevisiae* are obtained from NCBI (http://ncbi.nlm.nih.gov) database (December 2009). These genes are grouped according to four different functional classifications. Three of these describe the biological function of the proteins according to Gene Ontology (GO) categories [molecular function, biological process, and cellular localization]. The fourth category describes the KEGG pathways in which the proteins are involved.

With respect to the GO classification, 1725 (29%) proteins are not annotated, 1646 (28%) proteins are associated with a single term and 2509 (43%) proteins are associated with more than one term for biological process. Further details about the functional GO classification of *S. cerevisiae* proteins can be found in **Figure S1.1**.

Mapping the total *S. cerevisiae* protein set to KEGG pathways terms shows that 78% (4571) of genes have not been associated with any pathway. The remaining 22% (1309) of proteins are associated to 105 terms for pathways. Almost all these proteins are associated with more than one pathway. The results are summarized in Supplementary Figure 1 and in Supplementary **Table S1.2** and **Table S1. 3**.

With this functional classification of *S. cerevisiae* proteins in place, we can compare the different molecular circuits and processes of yeast their analogues in other organisms. To do this we downloaded the fully sequenced genome of 704 organisms, distributed in three domains (*Eukaryota*, *Bacteria* and *Archaea*). A list of all organisms is given as Supplementary **Table S1.1**.

Functional comparison of the full *S. cerevisiae* protein complement to that of other organisms

To compare molecular circuits, biological process, molecular function and cellular localization between *S. cerevisiae* and the other organisms, we created clusters of **orthologs** (ScCOGs), **homologues** (ScCHGs) and **absent** genes (ScCAGs) for each *S. cerevisiae* protein with respect to the genome of each of the other 704 organisms. The results are

summarized in Supplementary **Table S1.1, Table S1.2** and **Table S1. 3** and the detailed clusters can be downloaded in the supplementary files. Each cluster was associated with the functional terms corresponding to its *S. cerevisiae* protein, as classified in the previous section. Supplementary **Figure S1.2, Figure S1.3, Figure S1.4** and **Figure S1.5** describe these results. Given the functional associations and the ScCOGs, we can now look for differences and similarities in specific pathways, functions, or processes between *S. cerevisiae* and any of the studied organisms (for more details see Supplementary **Table S1.1, Table S1.2** and **Table S1. 3**).

ARCHAEA DOMAIN

We analysed 48 species of *Archaea*. About 20% (1158) of all *S. cerevisiae* proteins generate ScCOGs that contain *Archaea* sequences. However, only 2% (103) of all yeast proteins generate ScCOGs that contain at least a sequence from each sequenced species of *Archaea*. An additional 18 (0.3%) *S. cerevisiae* proteins have homologues in all *Archaea*. 3672 (62%) *S. cerevisiae* proteins are absent in all *Archaea*. Most of these have unknown function. At the phyla level, *Crenarchaeota* commonly share orthologs to 164 (3%) *S. cerevisiae* proteins, while *Euryarchaeota* commonly share orthologs to 148 (3%) *S. cerevisiae* proteins. *Korarchaeota* and *Nanoarchaeota* are represented in our sampling with only one organism per phylum.

Globally, the biological pathways of *S. cerevisiae* that share the highest fraction of their protein complements with all *Crenarchaeota* are "Aminoacyl-tRNA biosynthesis" and "Proteasome". Even so, less than 60% of the *S. cerevisiae* proteins associated with these pathways have orthologs in the organisms of the phylum. 35% of the *S. cerevisiae* proteins associated with "RNA metabolic process" have orthologs in all the organisms of the phylum. Orthologs for the protein complements of the remaining *S. cerevisiae* proteins associated with other pathways and processes are mostly absent from the phylum.

There are 3672 (62%) *S. cerevisiae* genes those are totally absent in all *Archaea*. 88% of these genes have no associated function in the KEGG database. No significant homology is found in any Archaea with respect to all of proteins from *S. cerevisiae* that are involved in "SNARE interaction vesicular transport" pathways (23 genes). Homologues for these proteins are also absent from all *Bacteria*, which is consistent with the fact that the function is very specific to eukaryotes [**92**]. Homologues for more than 80% of all *S. cerevisiae* proteins that are involved in "Glycosylphosphatidylinositol (GPI)-anchored biosynthesis", "High-

mannose type N-glycan biosynthesis" and "Unsaturated fatty acid biosynthesis" are absent from *Archaea*.

Homologues for 4209 (72%) *S. cerevisiae* proteins are absent in all *Crenarchaeota*, while only 3807 (65%) are absent in all *Euryarchaeota*. More than 80% of all *S. cerevisiae* proteins involved in "Ubiquitin mediated proteolysis", "Endocytosis", "Fructose and mannose metabolism", "Mismatch repair", "sphingolipid metabolism", "High-mannose type N-glycan biosynthesis", "Biosynthesis of unsaturated fatty acid" are absent in all *Crenarchaeota*. In contrast, all *Euryarchaeota* have significant homologues for more than 40% of the *S. cerevisiae* proteins involved in "Mismatch repair" and "Sphingolipid metabolism". These results suggest that, for these pathways, *Euryarchaeota* are closer to *S. cerevisiae* than *Crenarchaeaota*.

BACTERIA DOMAIN

We analysed 598 species of bacteria. 1612 (27%) of all *S. cerevisiae* proteins generate ScCOGs that contain bacteria sequences. However, no ScCOG or ScCHG contains a sequence from each bacterial species. Furthermore, 2881 (49%) *S. cerevisiae* genes are absent from all *Bacteria*, a smaller percentage than that for *Archaea*.

Interestingly, a higher percentage of *S. cerevisiae* proteins that participate in the "RNA polymerase", "DNA replication", "Pyrimidine metabolism", and "Ribosome" pathways is absent from *Bacteria* than from *Archaea*. This suggests that, for these pathways, *S. cerevisiae* may be more similar to *Archaea* than to *Bacteria*. On the other hand, a higher percentage of genes that participate in the "Starch and sucrose metabolism", "O-Mannosyl glycan biosynthesis", "High-mannose type N-glycan biosynthesis", "Biosynthesis of unsaturated fatty acid" and "Androgen and estrogen metabolism" pathways in *S. cerevisiae* is absent from *Archaea* than from *Bacteria*.

Most bacteria with fully sequenced genomes are *Proteobacteria* (315 organisms) and *Firmicutes* (122 organisms). Only 11 (0.2%) *S. cerevisiae* proteins have orthologs in all *Proteobacteria*, while 96 (2%) *S. cerevisiae* proteins have orthologs in all *Firmicutes*. One additional protein (0.01%) has homologues in all *Proteobacteria*, while an additional 8 (0.1%) proteins have homologues in all *Firmicutes*. By and large, the sets of proteins from *S. cerevisiae* that associate with individual biological processes and pathways are closer to the corresponding set of proteins from *Proteobacteria* than to those from *Firmicutes*. The

exception to this rule is observed for the "Lysosome" pathway of *S. cerevisiae*, which has more absent genes in *Proteobacteria* than in *Firmicutes*.

Our dataset contains genomes for 48 *Actinobacteria* and 30 *Cyanobacteria*. *S. cerevisiae* has 180 proteins that are present in all *Actinobacteria* genomes [146 orthologs (3%) + 34 (0.6%) homologues], and 352 proteins that are present in all *Cyanobacteria* genomes [263 (4%) orthologs and 89 (2%) homologues]. The set of proteins associated with "Aminoacyl-tRNA biosynthesis", "Pentose phosphate pathways", "Valine, leucine and isoleucine biosynthesis", "Histidine metabolism" pathways, "Cellular amino acid and derivative metabolic process", "Generation of precursor metabolites and energy", "Cofactor metabolic process" and "Cellular respiration" in *S. cerevisiae* are more similar to the corresponding sets in *Cyanobacteria* than to those in *Actinobacteria*.

3889 (66%) of all *S. cerevisiae* proteins are absent in all sequenced *Actinobacteria* and 3940 (67%) proteins of *S. cerevisiae* are absent in all *Cyanobacteria*. In terms of biological function, the sets of proteins that lack a higher number of homologues in *Actinobacteria* than in *Cyanobacteria* are associated with "Proteasome", "Amino sugar and nucleotide sugar metabolism", "Galactose metabolism", "Pentose and glucuronate interconversions", and "Lipid metabolic process".

Other bacterial phyla have a smaller number of organisms with fully sequenced genomes [*Tenericutes* (19 organisms), *Spirochete* (11 organisms), *Bacterioides* (11 organisms), *Green nonsulfur bacteria* (8 organisms), *Chlamydia* (13 organisms), *Hyperthermophilic bacteria* (4 organisms), *Green sulfur bacteria* (7 organisms) and *Deinococcus-thermus* (4 organisms)]. The set of *S. cerevisiae* proteins that is associated with the "Aminoacyl-tRNA biosynthesis" pathway is that which is most conserved in all organisms from these phyla, with the exception of *Tenericutes*. In this phylum, only 3 genes associated to "Aminoacyl-tRNA biosynthesis" have homologues. Three out of these seven phyla have a similar number of organisms with fully sequenced genomes. Those phyla are *Spirochete*, *Bacterioides* and *Chlamydia*. The set of proteins involved in "Glycolysis/Gluconeogenesis" and "TCA cycle" pathways in *S. cerevisiae* is more similar to that of *Chlamydia* than to those of the two other phyla. The set of proteins associated with "One carbon pool by folate" in *S. cerevisiae* is more similar to that of *Bacterioides* than to those in the other phyla. All the three phyla have an equal level of similarity to *S. cerevisiae*

with respect to the "Proteosome" pathway. The genes involved in "SNARE interaction in vesicular transport" pathway are totally absent in all the three phyla.

Another interesting comparison is that between *Green nonsulfur bacteria* (8 organisms) and *Green sulfur bacteria* (7 organisms). The set of *S. cerevisiae* proteins involved in "Glycolysis/Gluconeogenesis", "Glycine, serine and threonine metabolism", "Histidine metabolism", "Riboflavin metabolism", "Limonene and pinene degradation" and "Thiamine metabolism" pathways are more similar to the corresponding sets of *Green sulfur bacteria* than to those of *Green nonsulfur bacteria*. Similar fractions of the sets of *S. cerevisiae* proteins involved in "Phenylalanine, tyrosine and tryptophan biosynthesis", "One carbon pool folate" and "Fatty acid biosynthesis" pathways are found in both phyla.

More than 80% of the *S. cerevisiae* proteins associated with "Endocytosis", "RNA polymerase", "Basal transcription factor", "Glycosylphosphatidylinositol(GPI)-anchored biosynthesis", "Porphyrin and chlorophyll metabolism", "Steroid biosynthesis", "Sulfur metabolism" and "High-mannose type N-glycan biosynthesis" are absent in both phyla.

EUKARYOTA DOMAIN

Overall, there are 59 species of eukaryotes in our dataset. About 4.5% (263) of all ScCOGs contain sequences from each of these organisms. Between 40% and 60% of all *S. cerevisiae* proteins involved in "MAPK signalling pathways", "signal transduction" biological process, and "helicase activity" molecular functions are present in all 59 species. Furthermore, between 60% and 80% of all proteins involved in "Microtubule organizing centre" of *S. cerevisiae* are also found in all 59 sequenced eukaryotes.

FUNGI DOMAIN

We analyze 19 fungal species. 781 (13%) of the ScCOGs contain sequences from all these species. More than 80% of the proteins of *S. cerevisiae* involved in "O-mannosyl glycan biosynthesis", "Synthesis and degradation of ketone bodies", "Microtubule organizing centre", "helicase activity" and "motor activity" are also present in all other *Fungi*. More than 60% of all *S. cerevisiae* proteins involved in "RNA metabolic process", "Organelle organization", "Protein modification process", "Cell cycle", "Response to stress", "DNA metabolic process", and "Response to chemical stimuli" are also present in all other fungi.

2310 (39%) ScCOGs contain sequences from *Basidiomycetes* (4 organisms), while only 2174 (36%) ScCOGs contain sequences from *Ascomycetes* (14 organisms) (*S. cerevisiae*'s phylum). 469 (8%) ScCHGs have sequences from all *Basidiomycetes*, while 1525 (26%) genes are absent in all sequenced *Basidiomycetes*.

ANIMAL KINGDOM

We analysed the genomes of 20 animal species, distributed throughout 4 phyla: Vertebrates (12 organisms), Insects (4 organisms), Nematodes (3 organisms), and Echinoderms (1 organism, *Strongylocentrotus purpuratus* [purple sea urchin]). 2737 (47%) ScCOGs contain animal sequences. 480 (8%) of the ScCOGs contain sequences from all animals. An additional 81 (1%) *S. cerevisiae* proteins also have homologues in all animals.

More than 60% of the *S. cerevisiae* proteins that are associated with "MAPK signalling pathways - yeast", "Fatty acid metabolism", "Limonene and pinene degradation pathways" are also present in all animals. Between 40% and 60% of the *S. cerevisiae* proteins associated with "Signal transduction" and between 60% and 80% of *S. cerevisiae* proteins associated with "Signal transducer activity" and "Cytoskeleton, cellular bud and "Microtubule organizing centre" are also found in all animals.

2028 (34%) yeast proteins are absent in all the animal genomes. Most of these proteins have unknown biological function. Between 40% and 60% of the proteins involved in "Cell wall organization", "Sporulation", and "Transcription regulator activity" in *S. cerevisiae* are absent from all animals. This is expected, given that animals do not have cell walls.

Globally, 573 (10%) ScCOGs have sequences from all sequenced Vertebrates. This is phylum that has the lowest number of proteins that are common to all its organisms and have orthologs in *S. cerevisiae*. Homologues for the proteins from *S. cerevisiae* associated with the following processes are mostly absent from Vertebrates : "MAPK signalling pathway yeast", "Protein modification process", "Response to chemical stimuli", "Signal transduction", "Meiosis", "Transposition"; molecular functions involved genes like "RNA binding", "Translation regulator activity" and "Signal transducer activity". This suggests that *S. cerevisiae* is not a good model to study these processes in vertebrates.

The sets of *S. cerevisiae* proteins involved in "Starch and sucrose metabolism", "Galactose metabolism", "GPI-anchored biosynthesis", "Porphyrin and chlorophyll

metabolism", "One carbon pool by folate", "O-Mannosyl glycan biosynthesis", "Gamma-Hexachlorocyclohexane degradation", "Protein modification process", "Carbohydrate metabolic process", "Cellular amino acid and derivative metabolic process", "Heterocycle metabolic process" are more similar to the analogous sets found in *Insects* than to those found in *Nematodes*. The sets of *S. cerevisiae* proteins involved in "DNA metabolic process", "Helicase activity" and "Lipid binding" are more similar to the analogous sets found in *Nematodes* than to those found in *Insects*. Both phyla have orthologs for a similar proportion of the *S. cerevisiae* proteins involved in: "Ribosomes", "Ubiquitin mediated proteolysis", "Aminoacyl-tRNA biosynthesis", "TCA cycle", "DNA replication", "Glutathione metabolism", "Phosphatidylinositol signalling system", "Valine, leucine and isoleucine degradation" and "beta-Alanine metabolism", "RNA metabolic process", "Response to chemical stimulus", "Transferase activity", "DNA binding" and "Enzyme regulatory activity". Both phyla have orthologs for a similar proportion of the *S. cerevisiae* proteins localized to: "Nucleolus", "membrane fraction", "Golgi apparatus", and "Cytoplasmic membrane bounded vesicle".

PLANTS KINGDOM

We analysed 5 plant organisms distributed throughout 4 phyla. The *Dicotyledons*, *Monocotyledons*, and Red algae have one fully sequence genome each, while two Green algae genomes have been fully sequenced. 1371 (23%) ScCOGs contain sequences from all plants. An additional 253 (4%) yeast proteins have homologues in all plants.

All plants have orthologs for more than 80% of the *S. cerevisiae* proteins associated with: "Ribosome", "Aminoacyl-tRNA biosynthesis", "Pentothenate and CoA biosynthesis", "Propeonate metabolism", "Valine, leucine and isoleucine degradation", "Limonene and pinene degradation", "Homologous recombination", "Selenoamino acid metabolism", "Mismatch repair", "Valine, leucine and isoleucine biosynthesis", "Lysosome" ,"Alpha-Linolenic acid metabolism", "Benzoate degradation via hydroxylation", "Meiosis", "Structure molecular activity", "Ligase activity", "Helicase activity", and "Isomerase activity". All plants have orthologs for more than 80% of the set of *S. cerevisiae* proteins localized at "Ribosomes".

Interestingly, in Dicotyledons only 282 (5%) *S. cerevisiae* proteins are absent, whereas in *Monocotyledons* and *Red algae* at least 43% of *S. cerevisiae* proteins are absent. It is not possible at this time to know if this difference is just a consequence of the very limited

sampling of plant genomes that is available for our analysis or if it reflects some fundamental difference between the phyla. Nevertheless, *Arabidopsis thaliana* has the protein complement that is closest to that of *S. cerevisiae* in the plant kingdom.

PROTISTS KINGDOM

We analysed 15 protists organisms distributed throughout 7 phyla. Cellular slime molds, *Choanoflagellates*, *Diplomonads*, *Entamoeba* and *Parabasalids* have only one fully sequenced genome each, while *Alveolates* has 7 and *Euglenozoa* has 3 fully sequenced genomes. 591 (10%) ScCOGs contain sequences from all protists. An additional 43 (0.7%) yeast proteins have homologues in all protists. Between 60% and 80% of all *S. cerevisiae* proteins associated with the "Proteasome" have orthologs in all sequenced organisms from the Protists kingdom. Between 40% and 60% of the proteins associated with the following pathways in *S. cerevisiae* have orthologs in all protists: "Ribosome", "MAPK signaling pathway - yeast", "Ubiquitin mediated proteolysis", "Aminoacyl-tRNA biosynthesis", "Nucleotide excision repair", "Endocytosis", "DNA replication", "Homologous recombination", "Mismatch repair", "Lysosome", and "Protein export".

839 (14%) *S. cerevisiae* proteins have orthologs and 194 (3%) *S. cerevisiae* proteins have homologues in *Alveolates*. Orthologs or homologues for more than 80% of the *S. cerevisiae* proteins involved in "Proteosome" pathways are also presents in Alveolates. Between 60% and 80% of the proteins involved in the following pathways of *S. cerevisiae* have orthologs and/or homologues in Alveolates: "Ribosome", "Aminoacyl-tRNA biosynthesis", "Pyrimidine metabolism", "MAPK signaling pathway - yeast", "Ubiquitin mediated proteolysis", "Nucleotide excision repair", "DNA replication", "Homologous recombination", "Mismatch repair", "Aminoacyl-tRNA biosynthesis", "Fatty acid metabolism", "Nitrogen metabolism", "Glyoxylate and dicarboxylate metabolism", and "CO2 fixation".

2740 (47%) *S. cerevisiae* proteins are absent in *Alveolates*. Homologues and orthologs for the *S. cerevisiae* proteins involved in the following pathways are absent from the genome of all sequenced *Alveolates*: "O-Mannosyl glycan biosynthesis", "Riboflavin metabolism" and "High-mannose type N-glycan biosynthesis".

In *Euglenozoa*, 1281 (22%) *S. cerevisiae* proteins have orthologs and 695 (12%) *S. cerevisiae* proteins have homologous. Orthologs and homologues for more than 80% of the *S.*

cerevisiae proteins involved in the following pathways are also found in *Euglenozoa*: "Ribosome", "Proteosome", "Citrate cycle (TCA cycle)", "Glutathione metabolism", "Homologous recombination", "Mismatch repair", "Inositol phosphate metabolism", "Phosphatidylinositol signalling system", "Lysosome".

3157 (54%) *S. cerevisiae* proteins are absent in *Euglenozoa*. Most of these are also absent in *Alveolates*.

Proteins that are specific to *S. cerevisiae*

There are 24 *S. cerevisiae* proteins that have no orthologs in any other organism. However, out of these, only ten have no homologues in any of the analyzed genomes. The NCBI references for these proteins are NP_010097, NP_010148 (ribosomal protein L47 of 60S subunit), NP_010496, NP_013364, NP_878067, NP_010319, NP_013978, NP_878042, NP_878075, NP_878108. These ten genes code for small peptides. A few of them may be miss-annotated as genes. However, some have been predicted based on microarray expression data, which strongly suggests that they are being expressed and may have a function that is specific to this yeast.

Functional comparison of biological processes and pathways between *S. cerevisiae* and other organisms

KEGG PATHWAYS

Figure 2.2 summarizes the results for KEGG pathways (for more detail analysis see Supplementary **Figure S1.2**). Here, we find that "Benzoate degradation via hydroxylation" (2 genes) is the biological pathway that is fully present in the highest fraction of organisms. Even so, this pathway is fully absent from all *Tenericutes organisms*. "Geraniol degradation" (1 gene), "Methane metabolism" (7 genes), "Propanoate metabolism" (11 genes), "Valine, leucine and isoleucine degradation" (18 genes), "Aminoacyl-t-RNA biosynthesis" (39 genes) and "Glycolysis/Gluconeogenesis" (48 genes) are also pathways that appear to be similar to those of *S. cerevisiae* in a large fraction of organisms. Pathways such as *S. cerevisiae*'s "RNA polymerase" (29 genes), "Lysosome" (14 genes), "Endocytosis" (33 genes), "Oxidative phosphorylation" (76 genes), "Ribosome" (142 genes), "MAPK signaling pathway - yeast" (55 genes), "DNA replication" (30 genes), and "Ubiquitin mediated proteolysis" (44 genes)

and "Nucleotide excision repair" (34 genes) are much more similar to those from other eukaryotes than to the corresponding prokaryotic pathways (when they exist).

The full "Valine, leucine and isoleucine" pathway (18 genes) is found in all eukaryotes. Of all pathways, this is the one that is closest to that of a largest fraction of *Proteobacteria*, *Actinobacteria* and *Firmicutes*. Other prokaryotic phyla only have orthologs for less than 40% of the proteins in the pathway.

The *S. cerevisiae* "DNA replication" pathway (30 genes) is similar to that of all other eukaryotes and *Archaea*. *Bacteria* have no orthologs to protein associated with the yeast pathway. However, sequence homologues for the pathway are present in the *Bacteria* domain. *S. cerevisiae* "MAPK signalling pathways" (55 genes) are also well conserved in *Fungi*, and partially conserved in Animals, Plants and Protists. The *S. cerevisiae* "Ubiquitin mediated proteolysis" pathway (44 genes) is similar to those of other *Fungi*, *Animal*, *Plants*, and *Protists*. Orthologs for proteins involved in this pathway are often absent in *Alveolates*. The "Proteasome" pathway (35 genes) is similar to that of most *Eukaryotes*, with *Diplomonads* being the exception. These organisms have orthologs for only a few genes of the pathway.

The "Glycolysis/Gluconeogenesis" pathway (48 genes) is very similar between *S. cerevisiae* and all *Eukaryotes* and most *Bacteria* and *Archaea*, although, all the *Proteobacteria* being the exception. The *S. cerevisiae* "Thiamine metabolism" pathway (5 genes) is most similar to the corresponding pathways in other *Fungi*, in *Plants*, and in some *Proteobacteria*. The pathway is absent in *Animals* and *Protists*. The "Steroid biosynthesis" pathway (15 genes) is fully present in *Fungi, Plants* and *Vertebrates*. In *Insects* and *Nematodes* the pathway is absent.

The *S. cerevisiae* "Basal transcription factor" (23 genes) and "High-mannose type N-glycan biosynthesis" (12 genes) are similar only to the corresponding pathways of other *Ascomycetes*. Nevertheless, a fraction of the proteins for the first pathway are present in human, dog, zebra fish and African clawed frog. The "SNARE interaction in vesicular transport" pathway (23 genes) is fully present only in *Kluyveromyces lactis, Candida glabrata* and *Pichia stipitis*. It is completely absent from other *Ascomycetes*, from *Basidiomycetes*, from *Animals*, and from *Plants*.

The pathway that executes *S. cerevisiae* cell cycle (115 genes) appears to be quite unique to Fungi, because only a small fraction of its proteins have orthologs in other eukaryotes. This suggests that extrapolating the results of studying cell cycle in *S. cerevisiae* to other organisms should be done only at the level of basic principles, if at all [see for example [82, 83].

As expected, the closest organisms to *S. cerevisiae* in our analysis are *Kluyveromyces lactis* and *Candida glabrata*. *A. thaliana* (Dicotyledons) and *Oryza sativa* (Monocotyledons) are the closest organisms to *S. cerevisiae,* outside of the *Fungi* clade. A curious observation is that, when clustering organisms with respect to *S. cerevisiae*, most of the mammals remain close to *S. cerevisiae*. Humans, dogs, mice, cows and rats are among the organisms that are closer to the yeast, when you disregard other fungi. *Dictyostelium discoideum* (Cellular slime molds) is the closest protist to *S. cerevisiae*, whereas, *Giardia lamblia* (Diplomonads) is the most distant protist. Interestingly, *E. cuniculi* (*Microsporidians*) is the eukaryotic organism that appears to be the most different from *S. cerevisiae*, even though it belongs to the *Fungi* kingdom. Only some proteins from a few of the pathways from *S. cerevisiae* have orthologs in *E. cuniculi*. These pathways are "Arachidonic acid metabolism", "Alpha linolenic acid metabolism", "Pentose & glucuronate interconversion", "Terpenoid backbone biosynthesis", "Mismatch repair", "Proteasome", "RNA polymerase" and "Base excision repair". This organism has what appears to be a vestigial mitochondrial organelle, the mitosome. Fe-S cluster biogenesis, which takes place in the *S. cerevisiae* mitochondria, is also initiated in the mitosome of *E. cuniculi*. The remaining *S. cerevisiae* pathways are absent in *E. cuniculi*. This is consistent with the evolutionary history of *Microsporideans* [93].

When it comes to human metabolism, *S. cerevisiae* is likely to be a reasonable model for the study of "mismatch repair" (18 genes), "Ubiquitin and other terpenoid-quinone biosynthesis" (5 genes), "Inositol phosphate metabolism" (15 genes), "Steroid biosynthesis" (15 genes), "Ubiquitin mediated proteolysis" (44 genes), "DNA replication" (30 genes), "Ribosome" (142 genes), "Proteasome" (35 genes), "Mismatch repair" (18 genes), "Galactose metabolism" (23 genes), "One carbon pool by folate" (14 genes) and "Glycolysis/gluconeogenesis" (48 genes). It might also be a moderately good model to study "basal transcription factor" pathways (17 genes), "N-Glycan biosynthesis" (28 genes), "RNA polymerase" (29 genes), "Glycine, serine and threonine metabolism" (30 genes) and "Glycerophospholipid metabolism" (16 genes). *S. cerevisiae* pathways that have orthologs in humans for only a small fraction of their proteins are: "androgen estrogen metabolism" (4

genes), "Cyanoamino acid metabolism" (9 genes), "Nitrogen metabolism" (16 genes), "SNARE interactions in vesicular transport" (23 genes), "Gamma-Hexachlorocyclohexane degradation" (10 genes), "Phenylalanine, tyrosine and tryptophan biosynthesis" (22 genes), "GPI-anchored biosynthesis" (22 genes) and "MAPK signaling pathway - yeast" (55 genes).

Most *Bacteria* are closer to *S. cerevisiae* than any *Archaea*. Specifically, *Proteobacteria* are the closest to *S. cerevisiae* and *Klebsiella pneumoniae* is the closest Proteobacteria. *Tenericutes* are the most distant bacterial phylum to *S. cerevisiae*, and *Mycoplasma genitalium* is the most distant organism. In *Archaea, Haloarcula marismortui* (*Euryarchaeota*) is the closest organism to *S. cerevisiae* and *Nanoarchaeum equitans* (*Nanoarchaeota*) is the most distant.

Even though *S. cerevisiae* is an *Ascomycetes,* the sets of *S. cerevisiae* proteins involved in "Basal transcription factor", "Glycerophospholipid metabolism", "Tyrosine metabolism", "High-mannose type N-glycan biosynthesis" and "Ether lipid metabolism" are more similar to the corresponding sets of *Basidiomycetes* than to those of other *Ascomycetes*. This is also true for the sets of proteins involved in the following biological processes: "transport", "lipid metabolic process", "cellular amino acid derivative metabolic process", "membrane organization", "generation of precursor metabolites and energy", "heterocycle metabolic process", "meiosis" and "Vitamin metabolic process", molecular functions like "Structural molecular activity", "RNA binding", "Oxidoreductase activity", "Nucleotidyltransferase activity" and "isomerase activity" ScCOGs sequence for cellular localization at cytoplasm, membrane and ribosomes. This suggests that for these pathways and biological processes *S. cerevisiae* might be closer to *Basidiomycetes* than to other organisms of its own phylum. The protein complements associated with the remaining pathways and biological processes in *S. cerevisiae* appear to be 80% similar to those of other *Ascomycetes*. Thus, as expected based on its evolutionary history, most *S. cerevisiae* biological processes are more similar to those of *Ascomycetes* than to those of *Basidiomycetes*.

Of all fungi, *Encephalitozoon cuniculi* is the organism with the lowest number of proteins that are similar to those of *S. cerevisiae.* Only 1764 *S. cerevisiae* proteins are also found in *E. cuniculi* [1015 (17%) orthologs and 749 (13%) homologues]. 4097 (70%) of the proteins from *S. cerevisiae* are absent from *E. cuniculi*. The sets of proteins associated with the following pathways and processes in *S. cerevisiae* are absent from *E. cuniculi*: "TCA

cycle", "Arginine and proline metabolism", "Cysteine and methionine metabolism", "N-glycan biosynthesis", "SNARE interaction in vesicular transport", "Nitrogen metabolism", "Steroid biosynthesis", "Sulfur metabolism", "1- and2- Methylnaphthalene degradation", "3-chloroacrylic acid degradation", "Cellular amino acid and derivative metabolic process", "Generation and precursor metabolites and energy", "Heterocycle metabolic process", "Cellular respiration", "Vitamin metabolic process", and "Cellular aromatic compound metabolic process". Supplementary Figures 2-5 detail which other functional groups of proteins differ the most between the two organisms.

GO Biological Processes, Cellular Component and Molecular Function

Figure 2.2, **Figure 2.3** and **Figure 2.4** summarize the results for the comparisons between *S. cerevisiae* and the other organisms using the GO categories classification. The results are quite similar to those described for **Figure 2.1**, which suggests that these functional classifications could be equivalent to a large extent, in spite of all problems that they might have (see discussion). *S. cerevisiae* metabolic activities like "Cellular amino acid and derivative metabolic process", "Cellular aromatic compound metabolic process", "Heterocycle metabolic process", "Cofactor metabolic process" and "Vitamin metabolic process" are the ones that are more conserved in all organisms. In contrast, "Motor activity", "Transcription", "Anatomical structure morphogenesis", "Transposition", "Conjugation", "Cell budding", and "Protein modification process" appear to be conserved mostly in eukaryotes. Conservation of "Cell wall organization" pathways is restricted to fungi.

Evolutionary aspects of this work

As one would predict beforehand, the organisms that have the highest fraction of processes associated to proteins sets that are similar to the corresponding proteins sets of *S. cerevisiae* are *Kluyveromyces lactis, Candida glabrata*, and other Ascomycetes. *A. thaliana* (Dicotyledons) and *Oryza sativa* (Monocotyledons) are the organisms with the largest fraction of processes with protein sets that are similar to those of *S. cerevisiae*, outside of the *Fungi* clade. In general, ranking the organisms with respect to the global similarity between their protein sets and the corresponding set in *S. cerevisiae* creates a clustering tree that mostly replicates phylogenetic trees built using ribosomal RNA (data not shown).

Interestingly, in that clustering tree, *Encephalitozoon cuniculi* (*Microsporidians*) is the eukaryotic organism that is the most distant from *S. cerevisiae*, even though it belongs to

the *Fungi* kingdom. Only some proteins from a few of the pathways of *S. cerevisiae* have orthologs in *E. cuniculi* (see supplementary appendix for details). This is consistent with the evolutionary history of *Microsporideans* as a specialized intracellular fungi that both, lost many of its biological functions and has a high rate of divergence from other eukaryotes [**93**].

Another interesting fact is that 138 (111 orthologs and 27 homologues) out of 352 *S. cerevisiae* proteins that have homologues in all *Cyanobacteria* are mitochondrial proteins. 39% of all *S. cerevisiae* genes with orthologs in all *Cyanobacteria* are mitochondrial. In contrast 13.5% of all *S. cerevisiae* genes are mitochondrial. Thus, there are 2.9 (±0.24) times more mitochondrial genes in the *Cyanobacteria* ortholog set than one would expect from change alone. Given that a) the mitochondrial ancestor is a *Rickettsia* genus and not a *Cyanobacteria*, and b) the ancestor of chloroplasts is a *Cyanobacteria*, this result puzzled us and we speculated that it could provide functional insight into the evolution of both organelles [**94-98**].

To understand if there were any genes with specific functions and strong homologues that were common to mitochondria and chloroplasts, we decided to compare the sets of genes that have strong homologues between *S. cerevisiae*, all *Cyanobacteria*, all *Rickettsia* and the plant *A. thaliana*. We discovered that, 98 out of the 111 *S. cerevisiae* mitochondrial genes that had orthologs in all *Cyanobacteria* also had orthologs localized to the chloroplast in *A. thaliana*. Out of these, 92 where predicted to have strong homologues both in mitochondria and chloroplast. The localization of genes in *A. thaliana* was determined by checking the GO annotation (cellular component) of the genes in the TAIR database [**99**]. We also found that 110 *S. cerevisiae* mitochondrial genes had orthologs in all *Rickettsia*. Out of these, 68 had orthologs also in all *Cyanobacteria* and in *A. thaliana*. Mitochondrial and/or chloroplast genes are 3 to 4 times more common in this data set than one would expect from the set of *S. cerevisiae* proteins. The biological processes that dominate these sets of genes according to the GO classification is “biological process unknown”. Furthermore, genes involved in energy production are also abundant. However, no specific biological process or molecular function was significantly enriched in these datasets when compared to all *A. thaliana* chloroplast and/or mitochondrial genes. Thus, further work that requires better functional classifications is needed in order to understand if these datasets have any functional implications in the evolution of energy producing organelles.

2.7.2. Supplementary Figures

Figure S1.1

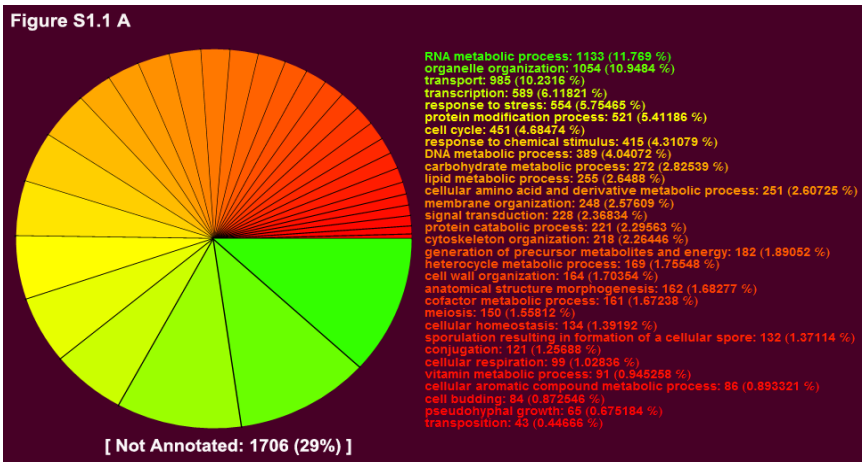


Figure S1.1.A. Frequency distribution of *S. cerevisiae* proteins according to GOSLIM biological process.

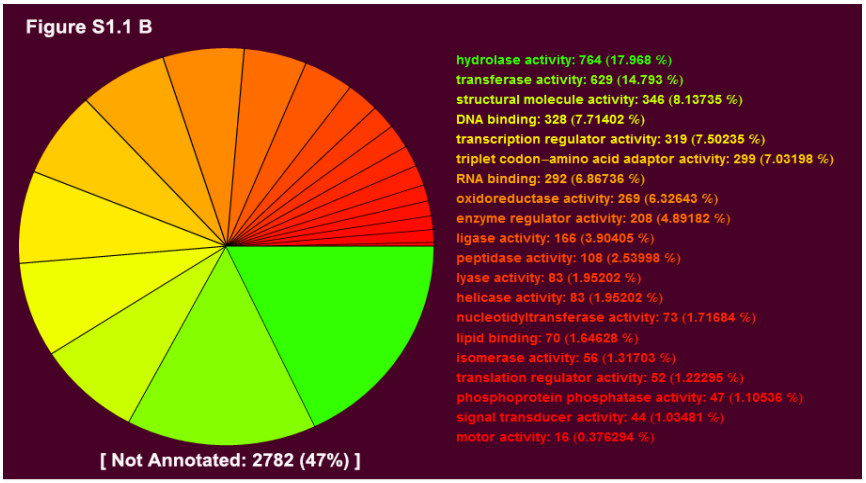


Figure S1.1.B. Frequency distribution of *S. cerevisiae* proteins according to GOSLIM molecular function.

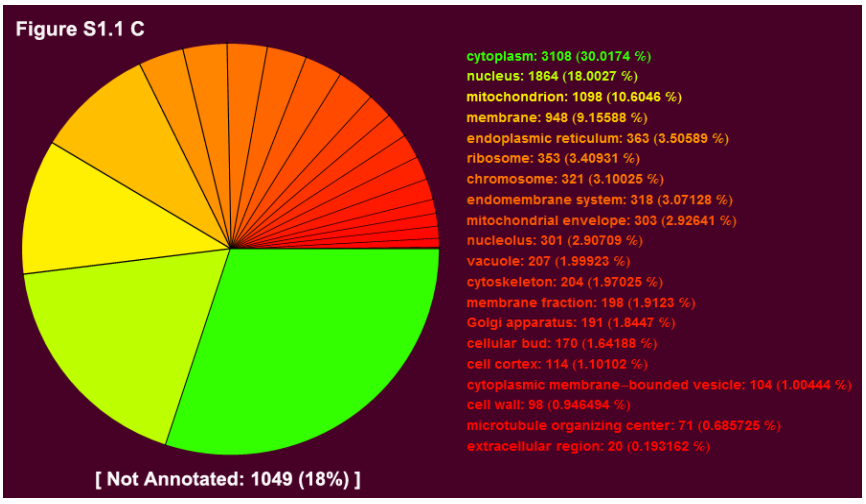


Figure S1.1.C. Frequency distribution of *S. cerevisiae* proteins according to GOSLIM cellular localization.

Figure S1.1 [continued…]

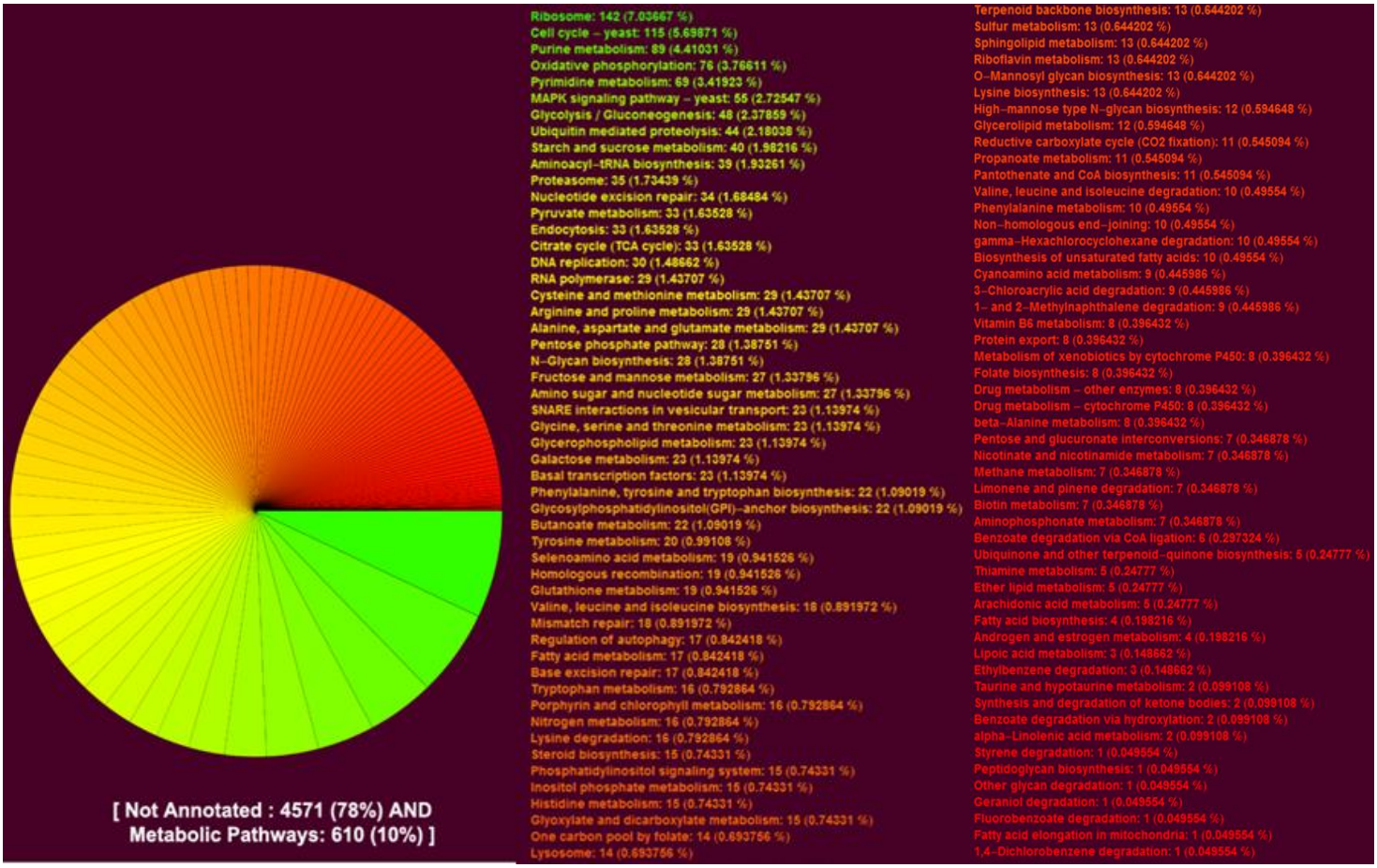


Figure S1.1.D. Frequency distribution of *S. cerevisiae* proteins according to KEGG pathways.

Figure S1.1 **Frequency distribution of *S. cerevisiae* proteins according to different functional classifications.**

Figure S1.2*

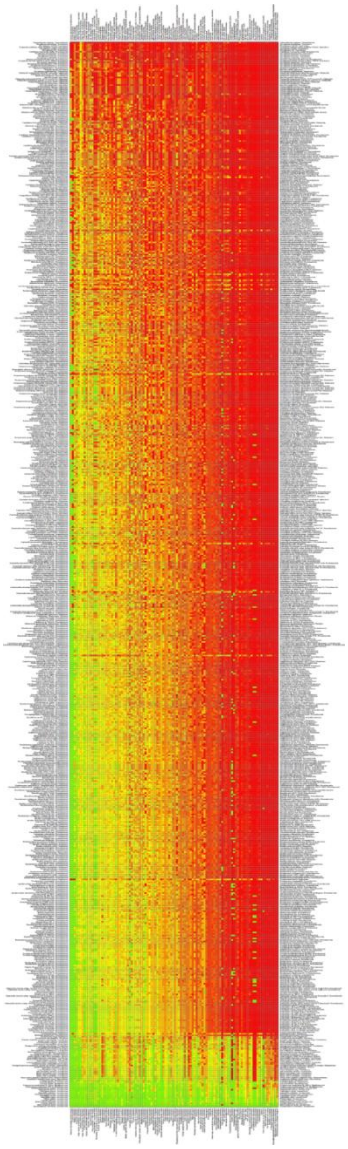


Figure S1.2 Full heat-map representation showing how distant each organism is from *S. cerevisiae* with respect to each individual KEGG pathway. A green square indicates a high level of coincidence between the set of proteins involved in the specific pathway (column) in a given organism (row) and the set of proteins for the same pathway in *S. cerevisiae*. A red square indicates complete absence of the set of proteins involved in the specific pathway (column) in a given organism (row) with respect to the same pathway in *S. cerevisiae*. Intermediate colours indicate intermediate degrees of coincidence between the set of proteins in the target organism and that in *S. cerevisiae*. ***Enlarged figure is available as Figure S1.2 in the CD that is provided with this thesis.**

Figure S1.3*

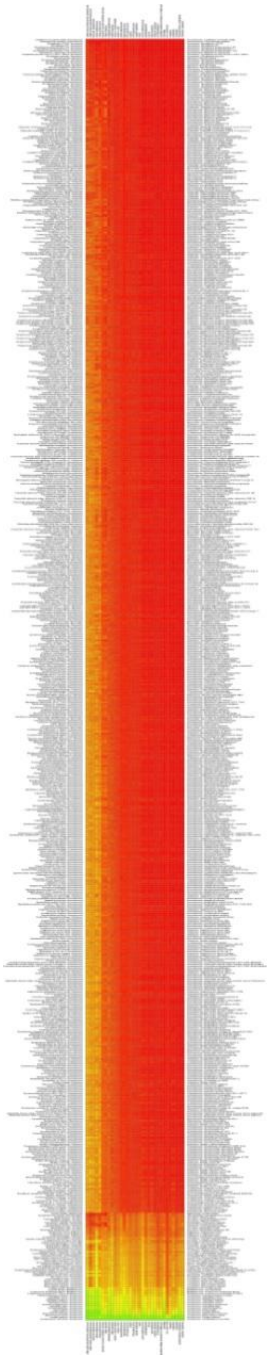


Figure S1.3 Full heat-map representation showing how distant each organism is from *S. cerevisiae* with respect to each biological process from the GOSLIM classification. A green square indicates a high level of coincidence between the set of proteins involved in the specific biological process (column) in a given organism (row) and the set of proteins for the same process in *S. cerevisiae*. A red square indicates complete absence of the set of proteins involved in the specific process (column) in a given organism (row) with respect to the same biological process in *S. cerevisiae*. Intermediate colours indicate intermediate degrees of coincidence between the set of proteins in the target organism and that in *S. cerevisiae*. ***Enlarged figure is available as Figure S1.3 in the CD that is provided with this thesis.**

Figure S1.4*

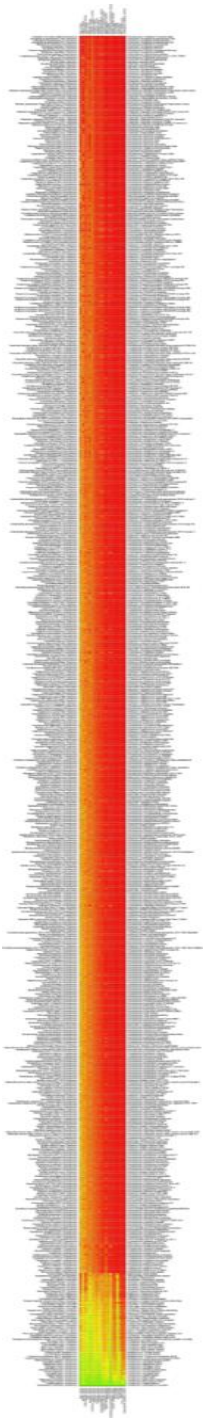


Figure S1.4 Full heat-map representation showing how distant each organism is from *S. cerevisiae* with respect to each molecular function from the GOSLIM classification. A green square indicates a high level of coincidence between the set of proteins involved in the specific molecular function (column) in a given organism (row) and the set of proteins for the same process in *S. cerevisiae*. A red square indicates complete absence of the set of proteins involved in the specific function (column) in a given organism (row) with respect to the same molecular function in *S. cerevisiae*. Intermediate colours indicate intermediate degrees of coincidence between the set of proteins in the target organism and that in *S. cerevisiae*. *Enlarged figure is available as Figure S1.4 in the CD that is provided with this thesis.

Figure S1.5*

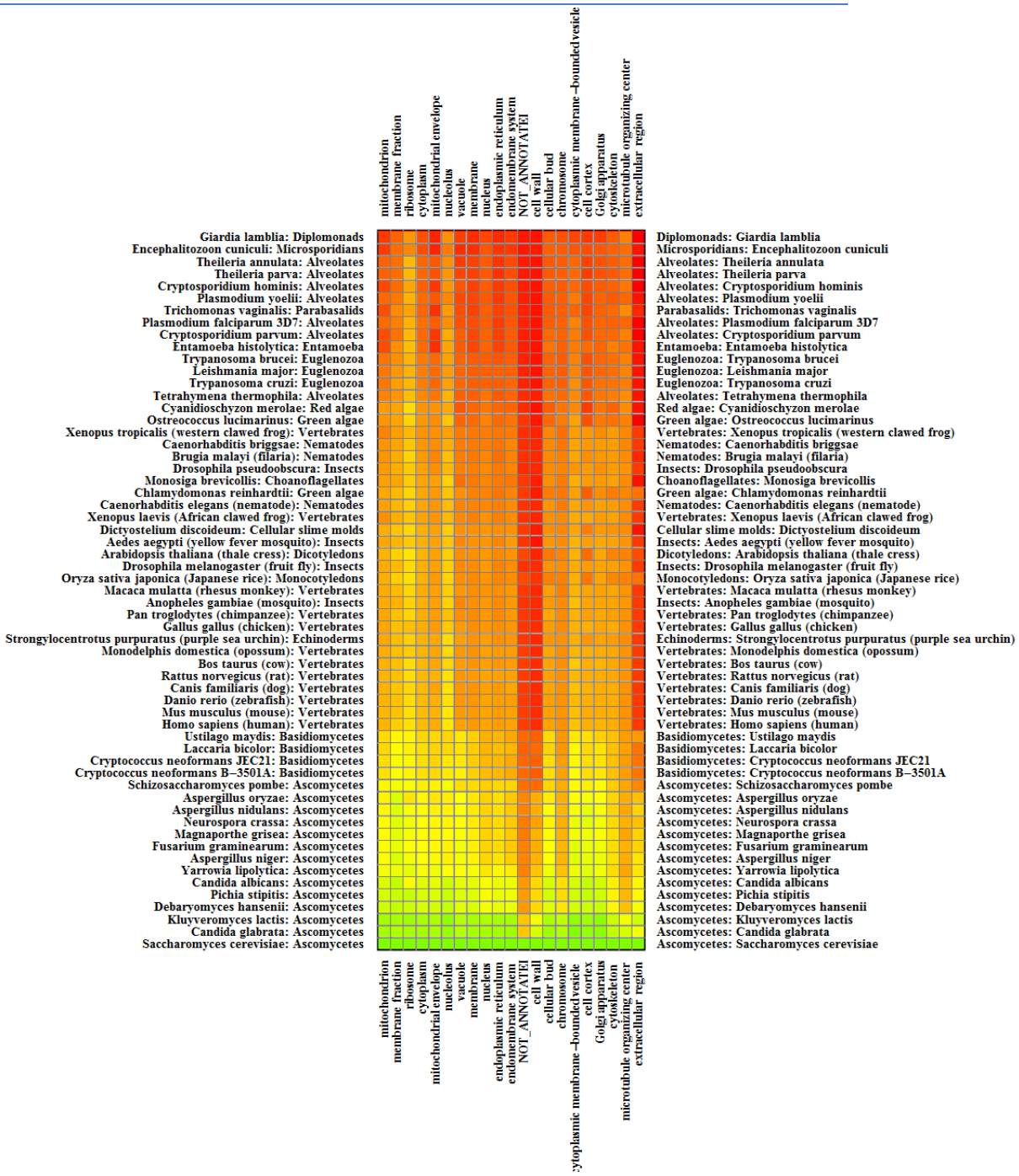


Figure S1.5 Full heat-map representation showing how distant each organism is from *S. cerevisiae* with respect to each cellular localization category from the GOSLIM classification. A green square indicates a high level of coincidence between the set of proteins assigned to a specific cellular localization (column) in a given organism (row) and the set of proteins for the localization in *S. cerevisiae*. A red square indicates complete absence of the set of proteins assigned to the specific cellular localization (column) in a given organism (row) with respect to the same localization in *S. cerevisiae*. Intermediate colours indicate intermediate degrees of coincidence between the set of proteins in the target organism and that in *S. cerevisiae*. *Enlarged figure is available as Figure S1.5 in the CD that is provided with this thesis.

2.7.3. Supplementary Tables

Table S1.1 Analyzed organisms and lumped homology with respect to the *S. cerevisiae* genome

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
EUKARYOTA	Fungi	Ascomycetes	1	*Saccharomyces cerevisiae*	5880	5856	0	0
			2	*Candida glabrata*	5191	4912	281	687
			3	*Kluyveromyces lactis*	5335	4879	416	585
			4	*Debaryomyces hansenii*	6324	4021	769	1090
			5	*Pichia stipitis*	5816	3988	760	1132
			6	*Candida albicans*	14105	3864	754	1262
			7	*Yarrowia lipolytica*	6543	3498	859	1523
			8	*Aspergillus niger*	14102	3410	957	1513
			9	*Fusarium graminearum*	11656	3364	982	1534
			10	*Aspergillus nidulans*	9541	3316	980	1584
			11	*Magnaporthe grisea*	14010	3309	983	1588
			12	*Neurospora crassa*	9824	3308	1009	1563
			13	*Aspergillus oryzae*	12074	3260	896	1724
			14	*Schizosaccharomyces pombe*	5003	3139	970	1771
		Basidiomycetes	15	*Cryptococcus neoformans B-3501A*	6500	2935	1084	1861
			16	*Cryptococcus neoformans JEC21*	6273	2899	1036	1945
			17	*Laccaria bicolor*	18215	2838	1075	1967
			18	*Ustilago maydis*	6538	2821	1074	1985
		Microsporidians	19	*Encephalitozoon cuniculi*	1996	1015	749	4116
	Animals	Vertebrates	20	*Homo sapiens (human)*	24305	2344	1087	2449
			21	*Mus musculus (mouse)*	29537	2342	1105	2433
			22	*Danio rerio (zebrafish)*	35022	2320	1108	2452
			23	*Rattus norvegicus (rat)*	26207	2304	1070	2506
			24	*Canis familiaris (dog)*	19797	2301	1128	2451
			25	*Bos taurus (cow)*	24127	2276	1086	2518
			26	*Monodelphis domestica (opossum)*	19113	2237	1059	2584
			27	*Pan troglodytes (chimpanzee)*	25163	2160	1074	2646
			28	*Macaca mulatta (rhesus monkey)*	23956	2142	1112	2626
			29	*Gallus gallus (chicken)*	18107	2139	1083	2658
			30	*Xenopus laevis (African clawed frog)*	10355	1987	1038	2855
			31	*Xenopus tropicalis (western clawed frog)*	7091	1795	1022	3063
		Nematodes	32	*Caenorhabditis elegans (nematode)*	20077	1999	1167	2714
			33	*Brugia malayi (filaria)*	11371	1896	1085	2899
			34	*Caenorhabditis briggsae*	16441	1842	1157	2881
		Insects	35	*Drosophila melanogaster (fruit fly)*	14144	2149	1075	2656
			36	*Anopheles gambiae (mosquito)*	12527	2143	1114	2623
			37	*Aedes aegypti (yellow fever mosquito)*	15432	2131	1079	2670
			38	*Drosophila pseudoobscura*	9869	1945	1052	2883
		Echinoderms	39	*Strongylocentrotus purpuratus (purple sea urchin)*	28881	2204	1075	2601

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
EUKARYOTA [continued...]	PLANTS	Dicotyledons	40	*Arabidopsis thaliana (thale cress)*	27216	2269	1147	2464
		Green algae	41	*Ostreococcus lucimarinus*	7603	1860	1057	2963
			42	*Chlamydomonas reinhardtii*	14416	2019	1090	2771
		Monocotyledons	43	*Oryza sativa japonica (Japanese rice)*	26937	2244	1091	2545
		Red algae	44	*Cyanidioschyzon merolae*	5013	1766	1051	3063
	PROTISTS	Alveolates	45	*Tetrahymena thermophila*	26052	1656	1169	3055
			46	*Plasmodium falciparum 3D7*	5261	1350	945	3585
			47	*Plasmodium yoelii*	7353	1294	916	3670
			48	*Cryptosporidium parvum*	3805	1293	917	3670
			49	*Theileria annulata*	3795	1230	848	3802
			50	*Theileria parva*	4061	1226	849	3805
			51	*Cryptosporidium hominis*	3885	1198	839	3843
		ellular slime mol	52	*Dictyostelium discoideum*	13437	2174	1156	2550
		Choanoflagellates	53	*Monosiga brevicollis*	9203	2007	1162	2711
		Diplomonads	54	*Giardia lamblia*	6500	959	752	4169
		Entamoeba	55	*Entamoeba histolytica*	11065	1300	945	3635
		Euglenozoa	56	*Trypanosoma cruzi*	19607	1491	1036	3353
			57	*Leishmania major*	8264	1470	1079	3331
			58	*Trypanosoma brucei*	8712	1427	1058	3395
		Parabasalids	59	*Trichomonas vaginalis*	59679	1257	1054	3569
BACTERIA	BACTERIA	Acidobacteria	60	*Solibacter usitatus*	7826	708	740	4432
		Actinobacteria	61	*Rhodococcus sp. RHA1*	9145	742	621	4517
			62	*Streptomyces avermitilis*	7673	702	684	4494
			63	*Streptomyces coelicolor*	8154	700	691	4489
			64	*Saccharopolyspora erythraea*	7197	699	697	4484
			65	*Arthrobacter sp. FB24*	4506	680	609	4591
			66	*Mycobacterium smegmatis*	6716	667	660	4553
			67	*Frankia sp. EAN1pec*	7191	660	608	4612
			68	*Mycobacterium sp. JLS*	5739	649	656	4575
			69	*Nocardia farcinica*	5936	649	674	4557
			70	*Mycobacterium sp. KMS*	5975	646	649	4585
			71	*Mycobacterium sp. MCS*	5615	645	646	4589
			72	*Mycobacterium vanbaalenii*	5979	639	646	4595
			73	*Frankia alni*	6711	638	687	4555
			74	*Mycobacterium gilvum*	5579	633	709	4538
			75	*Mycobacterium abscessus ATCC 19977T*	4941	627	649	4604
			76	*Thermobifida fusca*	3110	622	657	4601
			77	*Kineococcus radiotolerans*	4681	620	656	4604
			78	*Mycobacterium avium paratuberculosis*	4350	617	606	4657
			79	*Frankia sp. CcI3*	4499	617	593	4670
			80	*Mycobacterium bovis AF2122/97*	3920	616	608	4656

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Actinobacteria [continued...]	81	*Mycobacterium avium 104*	5120	613	607	4660
			82	*Mycobacterium bovis BCG Pasteur 1173P2*	3952	613	612	4655
			83	*Mycobacterium tuberculosis F11*	3941	610	614	4656
			84	*Mycobacterium tuberculosis H37Ra*	4034	609	614	4657
			85	*Mycobacterium tuberculosis H37Rv*	3989	609	614	4657
			86	*Mycobacterium tuberculosis CDC1551*	4189	609	610	4661
			87	*Corynebacterium glutamicum R*	3080	607	609	4664
			88	*Corynebacterium glutamicum ATCC 13032 (Kyowa Hakko)*	2993	606	598	4676
			89	*Corynebacterium glutamicum ATCC 13032 (Bielefeld)*	3057	603	599	4678
			90	*Nocardioides sp. JS614*	4909	603	655	4622
			91	*Salinispora tropica*	4536	603	629	4648
			92	*Mycobacterium ulcerans*	4160	600	624	4656
			93	*Kocuria rhizophila*	2357	586	586	4708
			94	*Clavibacter michiganensis subsp. michiganensis*	3079	581	628	4671
			95	*Corynebacterium efficiens*	2950	564	610	4706
			96	*Clavibacter michiganensis subsp. sepedonicus*	3117	563	631	4686
			97	*Rubrobacter xylanophilus*	3140	555	636	4689
			98	*Corynebacterium diphtheriae*	2272	547	609	4724
			99	*Corynebacterium jeikeium*	2120	547	567	4766
			100	*Corynebacterium urealyticum*	2024	541	573	4766
			101	*Bifidobacterium longum DJO10A*	2003	539	521	4820
			102	*Bifidobacterium longum NCC2705*	1729	532	527	4821
			103	*Leifsonia xyli xyli CTCB07*	2030	528	559	4793
			104	*Bifidobacterium adolescentis*	1631	516	507	4857
			105	*Propionibacterium acnes*	2297	514	624	4742
			106	*Mycobacterium leprae*	1605	503	531	4846
			107	*Tropheryma whipplei TW08/27*	783	301	454	5125
			108	*Tropheryma whipplei Twist*	808	300	457	5123
		Bacteroides	109	*Flavobacterium johnsoniae*	5017	600	607	4673
			110	*Bacteroides thetaiotaomicron*	4816	562	630	4688
			111	*Parabacteroides distasonis*	3850	559	621	4700
			112	*Cytophaga hutchinsonii*	3785	556	617	4707
			113	*Gramella forsetii*	3584	551	549	4780
			114	*Salinibacter ruber*	2833	550	673	4657
			115	*Bacteroides fragilis NCTC9343*	4231	542	540	4798
			116	*Bacteroides fragilis YCH46*	4625	539	558	4783
			117	*Bacteroides vulgatus*	4065	502	550	4828
			118	*Flavobacterium psychrophilum*	2412	482	503	4895
			119	*Porphyromonas gingivalis W83*	1909	377	471	5032

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Chlamydia	120	*Candidatus Protochlamydia amoebophila*	2031	441	579	4860
			121	*Chlamydophila felis*	1013	340	408	5132
			122	*Chlamydophila caviae*	1005	336	401	5143
			123	*Chlamydophila pneumoniae CWL029*	1052	324	410	5146
			124	*Chlamydophila pneumoniae TW183*	1113	323	411	5146
			125	*Chlamydophila pneumoniae J138*	1069	322	412	5146
			126	*Chlamydophila pneumoniae AR39*	1112	320	411	5149
			127	*Chlamydophila abortus*	932	317	428	5135
			128	*Chlamydia trachomatis A/HAR-13 (serovar A)*	919	303	409	5168
			129	*Chlamydia trachomatis 434/Bu*	874	302	410	5168
			130	*Chlamydia trachomatis L2b/UCH-1/proctitis*	874	302	413	5165
			131	*Chlamydia trachomatis D/UW-3/CX (serovar D)*	895	301	411	5168
			132	*Chlamydia muridarum*	911	292	448	5140
		Cyanobacteria	133	*Anabaena variabilis*	5661	675	768	4437
			134	*Anabaena sp. PCC7120*	6131	665	763	4452
			135	*Cyanothece sp. ATCC 51142*	5304	656	753	4471
			136	*Trichodesmium erythraeum*	4451	624	735	4521
			137	*Microcystis aeruginosa*	6312	615	719	4546
			138	*Gloeobacter violaceus*	4430	610	747	4523
			139	*Synechocystis sp. PCC6803*	3264	602	704	4574
			140	*Synechococcus sp. CC9311*	2892	520	539	4821
			141	*Cyanobacteria Yellowstone B-Prime*	2862	517	721	4642
			142	*Cyanobacteria Yellowstone A-Prime*	2760	513	692	4675
			143	*Synechococcus elongatus PCC7942*	2662	512	694	4674
			144	*Thermosynechococcus elongatus*	2475	510	750	4620
			145	*Synechococcus elongatus PCC6301*	2527	505	701	4674
			146	*Synechococcus sp. CC9902*	2307	499	504	4877
			147	*Prochlorococcus marinus MIT 9303*	2997	495	562	4823
			148	*Synechococcus sp. WH7803*	2533	495	533	4852
			149	*Synechococcus sp. RCC307*	2535	493	529	4858
			150	*Synechococcus sp. WH8102*	2519	490	499	4891
			151	*Synechococcus sp. CC9605*	2645	490	522	4868
			152	*Prochlorococcus marinus MIT9313*	2269	463	509	4908
			153	*Prochlorococcus marinus SS120*	1883	453	504	4923
			154	*Prochlorococcus marinus NATL1A*	2193	445	484	4951
			155	*Prochlorococcus marinus NATL2A*	2163	445	487	4948
			156	*Prochlorococcus marinus MIT 9515*	1906	436	465	4979
			157	*Prochlorococcus marinus MED4*	1717	425	485	4970
			158	*Prochlorococcus marinus MIT9312*	1810	423	497	4960
			159	*Prochlorococcus marinus MIT 9301*	1907	423	491	4966
			160	*Prochlorococcus marinus MIT 9215*	1983	421	492	4967
			161	*Prochlorococcus marinus AS9601*	1921	418	505	4957
			162	*Acaryochloris marina*	8383	347	1075	3912
		Deinococcus-Thermus	163	*Deinococcus geothermalis*	3062	608	658	4614
			164	*Deinococcus radiodurans*	3181	567	627	4686
			165	*Thermus thermophilus HB8*	2238	534	625	4721
			166	*Thermus thermophilus HB27*	2210	524	573	4783

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Firmicutes	167	*Bacillus cereus ATCC 10987*	5844	683	691	4506
			168	*Bacillus licheniformis DSM13*	4196	669	667	4544
			169	*Bacillus anthracis Sterne*	5287	669	682	4529
			170	*Bacillus subtilis*	4105	669	655	4556
			171	*Bacillus anthracis Ames*	5311	668	671	4541
			172	*Bacillus anthracis Ames 0581*	5617	667	670	4543
			173	*Bacillus thuringiensis Al Hakam*	4798	667	693	4520
			174	*Bacillus licheniformis ATCC 14580*	4178	666	664	4550
			175	*Bacillus thuringiensis 97-27*	5197	664	684	4532
			176	*Bacillus cereus ZK*	5641	663	709	4508
			177	*Bacillus anthracis A2012*	5852	661	654	4565
			178	*Clostridium beijerinckii*	5020	658	599	4623
			179	*Bacillus amyloliquefaciens*	3693	655	681	4544
			180	*Bacillus weihenstephanensis*	5653	653	686	4541
			181	*Bacillus pumilus*	3681	650	689	4541
			182	*Bacillus cereus ATCC 14579*	5255	648	706	4526
			183	*Geobacillus kaustophilus*	3540	645	650	4585
			184	*Oceanobacillus iheyensis*	3500	629	634	4617
			185	*Geobacillus thermodenitrificans*	3445	626	645	4609
			186	*Lysinibacillus sphaericus*	4771	622	655	4603
			187	*Bacillus cereus subsp. cytotoxis NVH 391-98*	3844	614	666	4600
			188	*Bacillus halodurans*	4066	609	681	4590
			189	*Bacillus clausii*	4096	606	646	4628
			190	*Clostridium acetobutylicum*	3848	600	620	4660
			191	*Staphylococcus saprophyticus*	2514	596	612	4672
			192	*Listeria monocytogenes EGD-e*	2846	595	584	4701
			193	*Listeria innocua*	3043	595	584	4701
			194	*Listeria monocytogenes F2365*	2821	595	588	4697
			195	*Listeria welshimeri SLCC5334*	2774	589	580	4711
			196	*Staphylococcus epidermidis ATCC 12228*	2485	588	608	4684
			197	*Staphylococcus epidermidis RP62A*	2526	582	625	4673
			198	*Staphylococcus aureus MSSA476*	2598	581	612	4687
			199	*Staphylococcus aureus MW2*	2632	580	614	4686
			200	*Staphylococcus aureus NCTC8325*	2892	577	612	4691
			201	*Desulfitobacterium hafniense*	5060	577	632	4671
			202	*Staphylococcus aureus Newman*	2614	577	605	4698
			203	*Staphylococcus aureus USA300*	2604	576	615	4689
			204	*Staphylococcus aureus N315*	2619	572	610	4698
			205	*Staphylococcus aureus Mu50*	2731	572	614	4694
			206	*Staphylococcus aureus MRSA252*	2656	571	635	4674
			207	*Staphylococcus haemolyticus*	2676	571	649	4660
			208	*Staphylococcus aureus JH1*	2780	571	625	4684
			209	*Staphylococcus aureus Mu3*	2698	571	611	4698
			210	*Staphylococcus aureus JH9*	2726	569	622	4689
			211	*Exiguobacterium sibiricum*	3015	563	658	4659
			212	*Clostridium botulinum B1 Okra*	3852	562	607	4711
			213	*Clostridium botulinum A3 Loch Maree*	3984	561	607	4712
			214	*Staphylococcus aureus RF122*	2509	561	615	4704
			215	*Clostridium botulinum F Langeland*	3659	561	597	4722
			216	*Clostridium botulinum A ATCC 3502*	3590	559	608	4713

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Firmicutes	217	*Clostridium kluyveri*	3913	558	633	4689
			218	*Clostridium botulinum A ATCC 19397*	3552	558	602	4720
			219	*Lactococcus lactis subsp. lactis IL1403*	2321	552	566	4762
			220	*Clostridium botulinum A Hall*	3404	548	609	4723
			221	*Lactobacillus plantarum*	3057	544	575	4761
			222	*Staphylococcus aureus COL*	2618	539	530	4811
			223	*Alkaliphilus metalliredigens*	4625	539	591	4750
			224	*Clostridium difficile*	3753	538	616	4726
			225	*Desulfotomaculum reducens*	3276	536	633	4711
			226	*Clostridium perfringens ATCC 13124*	2876	530	566	4784
			227	*Lactococcus lactis subsp. cremoris MG1363*	2434	527	571	4782
			228	*Clostridium phytofermentans*	3902	527	578	4775
			229	*Clostridium perfringens 13*	2723	523	560	4797
			230	*Leuconostoc mesenteroides*	2005	521	547	4812
			231	*Clostridium botulinum B Eklund 17B*	3527	518	610	4752
			232	*Carboxydothermus hydrogenoformans*	2620	516	553	4811
			233	*Lactococcus lactis subsp. cremoris SK11*	2504	515	567	4798
			234	*Streptococcus sanguinis*	2270	514	548	4818
			235	*Symbiobacterium thermophilum*	3338	513	657	4710
			236	*Lactobacillus casei BL23*	3044	513	576	4791
			237	*Clostridium thermocellum*	3189	508	596	4776
			238	*Heliobacterium modesticaldum*	3000	508	612	4760
			239	*Lactobacillus casei ATCC 334*	2771	506	579	4795
			240	*Clostridium novyi*	2315	504	558	4818
			241	*Lactobacillus fermentum*	1843	504	563	4813
			242	*Clostridium perfringens SM101*	2578	500	559	4821
			243	*Thermoanaerobacter tengcongensis*	2588	499	587	4794
			244	*Lactobacillus reuteri F275 (JGI)*	1900	493	573	4814
			245	*Pelotomaculum thermopropionicum*	2920	493	611	4776
			246	*Leuconostoc citreum*	1820	492	547	4841
			247	*Streptococcus thermophilus LMG18311*	1889	492	513	4875
			248	*Streptococcus suis 05ZYH33*	2186	490	514	4876
			249	*Streptococcus thermophilus CNRZ1066*	1915	489	516	4875
			250	*Streptococcus suis 98HAH33*	2185	489	525	4866
			251	*Enterococcus faecalis*	3265	487	613	4780
			252	*Streptococcus pneumoniae R6*	2043	486	532	4862
			253	*Streptococcus pneumoniae D39*	1914	485	526	4869
			254	*Streptococcus agalactiae 2603 (serotype V)*	2124	483	519	4878
			255	*Lactobacillus brevis*	2218	479	557	4844
			256	*Streptococcus agalactiae A909 (serotype Ia)*	1996	478	526	4876
			257	*Streptococcus thermophilus LMD-9*	1716	477	516	4887
			258	*Streptococcus mutans*	1960	476	557	4847
			259	*Streptococcus pneumoniae TIGR4*	2105	476	537	4867
			260	*Streptococcus agalactiae NEM316 (serotype III)*	2094	475	522	4883
			261	*Pediococcus pentosaceus*	1755	470	606	4804
			262	*Streptococcus gordonii*	2051	468	556	4856
			263	*Lactobacillus sakei*	1879	467	546	4867
			264	*Alkaliphilus oremlandii*	2836	460	552	4868
			265	*Moorella thermoacetica*	2465	456	664	4760
			266	*Oenococcus oeni*	1691	454	567	4859

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Firmicutes [continued...]	267	*Clostridium tetani E88*	2432	443	547	4890
			268	*Candidatus Desulforudis audaxviator*	2157	441	560	4879
			269	*Streptococcus pyogenes MGAS5005 (serotype M1)*	1865	439	505	4936
			270	*Streptococcus pyogenes MGAS10270 (serotype M3)*	1986	439	506	4935
			271	*Lactobacillus salivarius*	2013	438	483	4959
			272	*Streptococcus pyogenes MGAS10394 (serotype M6)*	1886	436	505	4939
			273	*Streptococcus pyogenes MGAS2096 (serotype M3)*	1898	436	510	4934
			274	*Streptococcus pyogenes MGAS9429 (serotype M3)*	1877	435	517	4928
			275	*Streptococcus pyogenes MGAS6180 (serotype M28)*	1894	435	505	4940
			276	*Streptococcus pyogenes Manfredo (serotype M5)*	1745	435	491	4954
			277	*Streptococcus pyogenes MGAS315 (serotype M3)*	1865	434	496	4950
			278	*Streptococcus pyogenes SF370 (serotype M1)*	1697	432	509	4939
			279	*Streptococcus pyogenes SSI-1 (serotype M3)*	1861	428	494	4958
			280	*Streptococcus pyogenes MGAS8232 (serotype M18)*	1839	423	514	4943
			281	*Streptococcus pyogenes MGAS10750 (serotype M3)*	1979	422	520	4938
			282	*Lactobacillus acidophilus*	1862	418	548	4914
			283	*Lactobacillus delbrueckii ATCC BAA-365*	1721	413	546	4921
			284	*Lactobacillus delbrueckii ATCC 11842*	1562	407	534	4939
			285	*Lactobacillus helveticus*	1610	403	534	4943
			286	*Lactobacillus johnsonii*	1821	401	549	4930
			287	*Lactobacillus gasseri*	1755	399	539	4942
			288	*Finegoldia magna*	1813	359	516	5005
		Fusobacteria	289	*Fusobacterium nucleatum*	2067	419	491	4970
		Green nonsulfur bacteria	290	*Roseiflexus sp. RS-1*	4517	652	680	4548
			291	*Roseiflexus castenholzii DSM13941*	4330	647	691	4542
			292	*Herpetosiphon aurantiacus*	5278	647	711	4522
			293	*Chloroflexus aurantiacus*	3853	637	733	4510
			294	*Dehalococcoides sp. CBDB1*	1458	366	473	5041
			295	*Dehalococcoides sp. BAV1*	1371	355	476	5049
			296	*Dehalococcoides ethenogenes*	1580	346	466	5068
		Green sulfur bacteria	297	*Chlorobium phaeobacteroides DSM 266*	2650	500	600	4780
			298	*Chloroherpeton thalassium*	2710	499	635	4746
			299	*Chlorobaculum parvum NCIB 8327*	2043	495	592	4793
			300	*Chlorobium phaeobacteroides BS1*	2469	489	613	4778
			301	*Pelodictyon luteolum*	2083	477	579	4824
			302	*Chlorobaculum tepidum*	2252	466	608	4806
			303	*Chlorobium chlorochromatii*	2002	446	642	4792
			304	*Prosthecochloris vibrioformis*	1753	444	590	4846
		Hyperthermophilic bacteria	305	*Thermotoga maritima*	1858	415	571	4894
			306	*Thermotoga petrophila*	1785	407	566	4907
			307	*Fervidobacterium nodosum*	1750	391	519	4970
			308	*Thermosipho melanesiensis*	1879	391	457	5032
		Planctomyces	309	*Rhodopirellula baltica*	7325	632	712	4536
		Proteobacteria	310	*Myxococcus xanthus*	7331	763	667	4450
			311	*Pseudomonas aeruginosa PAO1*	5568	715	695	4470
			312	*Pseudomonas aeruginosa PA7*	6286	712	679	4489
			313	*Pseudomonas aeruginosa UCBPP-PA14*	5892	711	690	4479
			314	*Burkholderia phymatum*	7496	708	720	4452
			315	*Burkholderia xenovorans*	8702	703	707	4470
			316	*Pseudomonas fluorescens Pf-5*	6138	700	743	4437

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Proteobacteria [continued...]	317	*Burkholderia cenocepacia MC0-3*	7008	696	695	4489
			318	*Klebsiella pneumoniae*	5187	694	556	4630
			319	*Burkholderia vietnamiensis*	7617	692	646	4542
			320	*Hahella chejuensis*	6778	691	694	4495
			321	*Pseudoalteromonas atlantica*	4281	690	702	4488
			322	*Pseudomonas putida F1*	5252	689	687	4504
			323	*Escherichia coli E24377A*	4997	688	635	4557
			324	*Ralstonia eutropha H16*	6626	687	700	4493
			325	*Ralstonia eutropha JMP134*	6446	687	709	4484
			326	*Burkholderia cenocepacia HI2424*	6919	686	688	4506
			327	*Pseudomonas putida KT2440*	5350	684	671	4525
			328	*Burkholderia ambifaria MC40-6*	6697	684	699	4497
			329	*Burkholderia cepacia*	6617	682	721	4477
			330	*Burkholderia cenocepacia AU1054*	6477	681	687	4512
			331	*Burkholderia sp. 383*	7717	681	705	4494
			332	*Serratia proteamaculans*	4942	679	604	4597
			333	*Pseudomonas fluorescens Pf0-1*	5736	678	748	4454
			334	*Citrobacter koseri ATCC BAA-895*	5008	676	526	4678
			335	*Pseudomonas entomophila*	5134	675	684	4521
			336	*Escherichia coli K-12 W3110*	4226	672	557	4651
			337	*Escherichia coli K-12 MG1655*	4132	669	557	4654
			338	*Escherichia coli ATCC 8739*	4200	667	556	4657
			339	*Escherichia coli HS*	4378	667	563	4650
			340	*Pseudomonas syringae pv. syringae B728a*	5089	666	697	4517
			341	*Escherichia coli SECEC*	4913	666	565	4649
			342	*Escherichia coli O157 Sakai (EHEC)*	5318	666	558	4656
			343	*Bradyrhizobium japonicum*	8317	666	770	4444
			344	*Mesorhizobium loti*	7272	664	717	4499
			345	*Burkholderia thailandensis*	5634	664	729	4487
			346	*Burkholderia pseudomallei 668*	7230	663	720	4497
			347	*Escherichia coli UTI89 (UPEC)*	5166	663	566	4651
			348	*Pseudomonas syringae pv. tomato DC3000*	5613	662	707	4511
			349	*Burkholderia pseudomallei K96243*	5728	659	725	4496
			350	*Enterobacter sp. 638*	4240	659	555	4666
			351	*Escherichia coli K-12 DH10B*	4126	659	534	4687
			352	*Burkholderia pseudomallei 1106a*	7183	658	735	4487
			353	*Escherichia coli 536 (UPEC)*	4620	658	567	4655
			354	*Enterobacter sakazakii*	4420	657	590	4633
			355	*Photobacterium profundum*	5489	656	716	4508
			356	*Shigella sonnei*	4475	655	559	4666
			357	*Salmonella enterica subsp. enterica serovar Typhi Ty2*	4318	655	628	4597
			358	*Salmonella enterica subsp. enterica serovar Typhi CT18*	4758	655	628	4597
			359	*Salmonella typhimurium LT2*	4527	654	553	4673
			360	*Yersinia enterocolitica*	4051	654	542	4684
			361	*Burkholderia pseudomallei 1710b*	6347	653	734	4493
			362	*Pseudomonas syringae pv. phaseolicola 1448A*	5171	652	703	4525
			363	*Escherichia coli APEC O1*	4851	650	564	4666
			364	*Salmonella enterica subsp. enterica serovar Paratyphi A*	4093	650	531	4699
			365	*Escherichia coli O157 EDL933 (EHEC)*	5397	649	669	4562
			366	*Rhizobium leguminosarum*	7143	648	657	4575
			367	*Colwellia psychrerythraea*	4910	648	675	4557

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Proteobacteria [continued...]	368	*Erwinia tasmaniensis*	3622	647	581	4652
			369	*Pseudomonas mendocina*	4594	647	654	4579
			370	*Salmonella enterica subsp. enterica serovar Choleraesuis*	4634	646	540	4694
			371	*Xanthomonas campestris pv. campestris ATCC 33913*	4181	642	586	4652
			372	*Burkholderia mallei NCTC 10247*	5852	642	716	4522
			373	*Xanthomonas campestris pv. campestris 8004*	4273	642	586	4652
			374	*Shigella dysenteriae*	4506	640	535	4705
			375	*Caulobacter sp. K31*	5438	640	688	4552
			376	*Aeromonas hydrophila*	4122	640	614	4626
			377	*Shigella flexneri 301 (serotype 2a)*	4440	640	548	4692
			378	*Burkholderia multivorans ATCC 17616 (JGI)*	6258	639	642	4599
			379	*Erwinia carotovora*	4472	639	525	4716
			380	*Ralstonia metallidurans*	6319	639	687	4554
			381	*Xanthomonas campestris pv. vesicatoria*	4726	639	697	4544
			382	*Yersinia pseudotuberculosis IP31758*	4324	638	576	4666
			383	*Burkholderia mallei SAVP1*	5189	638	701	4541
			384	*Shigella flexneri 2457T (serotype 2a)*	4061	637	552	4691
			385	*Shewanella frigidimarina*	4029	637	657	4586
			386	*Rhizobium etli CFN 42*	5963	636	671	4573
			387	*Bradyrhizobium sp. ORS278*	6717	636	697	4547
			388	*Shigella flexneri 8401 (serotype 5b)*	4115	636	553	4691
			389	*Xanthomonas axonopodis*	4427	636	676	4568
			390	*Pseudomonas stutzeri*	4128	635	680	4565
			391	*Delftia acidovorans*	6040	635	704	4541
			392	*Escherichia coli CFT073 (UPEC)*	5339	634	580	4666
			393	*Burkholderia mallei ATCC 23344*	5024	633	709	4538
			394	*Ralstonia solanacearum*	5116	633	678	4569
			395	*Chromobacterium violaceum*	4407	633	651	4596
			396	*Shewanella loihica*	3859	632	627	4621
			397	*Shigella boydii Sb227*	4285	632	556	4692
			398	*Leptothrix cholodnii*	4363	632	718	4530
			399	*Yersinia pestis Pestoides*	4069	631	557	4692
			400	*Yersinia pseudotuberculosis IP32953*	4038	630	564	4686
			401	*Bradyrhizobium sp. BTAi1*	7622	629	686	4565
			402	*Aeromonas salmonicida*	4437	628	581	4671
			403	*Yersinia pestis Antiqua*	4364	628	554	4698
			404	*Yersinia pestis Nepal516*	4094	628	546	4706
			405	*Agrobacterium tumefaciens C58 (UWash/Dupont)*	5335	625	609	4646
			406	*Azoarcus sp. BH72*	3989	623	612	4645
			407	*Yersinia pestis KIM*	4202	622	537	4721
			408	*Sinorhizobium medicae*	6213	622	598	4660
			409	*Azorhizobium caulinodans*	4717	622	662	4596
			410	*Yersinia pestis Mediaevails*	4142	622	548	4710
			411	*Yersinia pestis CO92*	4066	621	546	4713
			412	*Shewanella baltica OS155*	4489	620	661	4599
			413	*Photorhabdus luminescens*	4683	620	541	4719
			414	*Shewanella baltica OS185*	4394	620	599	4661
			415	*Sinorhizobium meliloti*	6201	620	600	4660
			416	*Rhodopseudomonas palustris CGA009*	4820	620	690	4570
			417	*Vibrio parahaemolyticus*	4832	619	719	4542
			418	*Ochrobactrum anthropi*	4799	619	563	4698

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Proteobacteria [continued...]	419	*Agrobacterium tumefaciens C58 (Cereon)*	5288	619	600	4661
			420	*Polaromonas sp. JS666*	5453	617	688	4575
			421	*Marinobacter aquaeolei*	4272	617	637	4626
			422	*Rhodoferax ferrireducens*	4418	617	653	4610
			423	*Shewanella sp. W3-18-1*	4044	615	636	4629
			424	*Shewanella sediminis*	4497	615	700	4565
			425	*Xanthobacter autotrophicus*	5035	611	645	4624
			426	*Shewanella sp. ANA-3*	4360	611	671	4598
			427	*Psychromonas ingrahamii*	3545	611	680	4589
			428	*Vibrio harveyi*	6055	607	693	4580
			429	*Dechloromonas aromatica*	4171	605	681	4594
			430	*Pseudoalteromonas haloplanktis*	3486	605	629	4646
			431	*Vibrio vulnificus CMCP6*	4484	603	677	4600
			432	*Beijerinckia indica*	3784	602	593	4685
			433	*Vibrio vulnificus YJ016*	5024	601	688	4591
			434	*Rhodopseudomonas palustris BisB18*	4886	600	672	4608
			435	*Rhodopseudomonas palustris BisA53*	4878	600	653	4627
			436	*Rhodopseudomonas palustris BisB5*	4397	600	671	4609
			437	*Azoarcus sp. EbN1*	4599	600	653	4627
			438	*Shewanella pealeana*	4241	599	707	4574
			439	*Shewanella sp. MR-7*	4014	598	678	4604
			440	*Shewanella sp. MR-4*	3924	597	673	4610
			441	*Shewanella putrefaciens*	3972	597	583	4700
			442	*Anaeromyxobacter sp. Fw109-5*	4466	597	719	4564
			443	*Chromohalobacter salexigens*	3298	596	634	4650
			444	*Cellvibrio japonicus*	3754	595	630	4655
			445	*Xanthomonas oryzae MAFF311018*	4372	595	683	4602
			446	*Bordetella petrii*	5027	594	677	4609
			447	*Gluconacetobacter diazotrophicus*	3778	593	608	4679
			448	*Minibacterium massiliensis*	3697	593	648	4639
			449	*Shewanella oneidensis*	4467	592	650	4638
			450	*Nitrobacter hamburgensis*	4326	591	544	4745
			451	*Brucella melitensis 16M*	3198	591	552	4737
			452	*Xanthomonas oryzae KACC10331*	4144	590	666	4624
			453	*Paracoccus denitrificans*	5077	590	650	4640
			454	*Mesorhizobium sp. BNC1*	4543	589	615	4676
			455	*Hyphomonas neptunium*	3505	589	607	4684
			456	*Polaromonas naphthalenivorans*	4929	589	700	4591
			457	*Shewanella denitrificans*	3754	589	678	4613
			458	*Saccharophagus degradans*	4008	587	698	4595
			459	*Bordetella bronchiseptica*	4994	586	578	4716
			460	*Rhodopseudomonas palustris HaA2*	4683	586	629	4665
			461	*Marinomonas sp. MWYL1*	4439	584	669	4627
			462	*Rhodobacter sphaeroides ATCC 17029*	4132	583	638	4659
			463	*Rhodobacter sphaeroides 2.4.1*	4242	582	625	4673
			464	*Acidiphilium cryptum JF-5*	3559	582	664	4634
			465	*Parvibaculum lavamentivorans*	3636	580	571	4729
			466	*Shewanella amazonensis*	3645	580	600	4700
			467	*Geobacter metallireducens*	3532	579	577	4724
			468	*Caulobacter crescentus*	3737	579	595	4706
			469	*Novosphingobium aromaticivorans*	3937	579	655	4646

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Proteobacteria [continued...]	470	*Rhodospirillum rubrum*	3841	579	615	4686
			471	*Brucella abortus S19*	3000	579	525	4776
			472	*Vibrio fischeri*	3802	573	689	4618
			473	*Brucella canis*	3251	572	554	4754
			474	*Methylibium petroleiphilum*	4449	572	705	4603
			475	*Geobacter uraniumreducens*	4357	572	538	4770
			476	*Brucella suis ATCC 23445*	3241	571	555	4754
			477	*Bordetella parapertussis*	4185	569	556	4755
			478	*Brucella ovis*	2890	569	554	4757
			479	*Brucella melitensis biovar Abortus*	3034	569	529	4782
			480	*Herminiimonas arsenicoxydans*	3325	567	628	4685
			481	*Geobacter lovleyi*	3685	567	609	4704
			482	*Brucella suis 1330*	3271	567	556	4757
			483	*Brucella abortus 9-941*	3085	566	536	4778
			484	*Vibrio cholerae O395*	3875	565	578	4737
			485	*Rhodobacter sphaeroides ATCC 17025*	4333	564	593	4723
			486	*Erythrobacter litoralis*	3011	563	612	4705
			487	*Vibrio cholerae O1*	3835	562	572	4746
			488	*Silicibacter pomeroyi*	4252	562	644	4674
			489	*Bordetella avium*	3381	560	569	4751
			490	*Verminephrobacter eiseniae*	4947	559	674	4647
			491	*Legionella pneumophila Corby*	3206	558	731	4591
			492	*Sphingomonas wittichii*	5345	558	583	4739
			493	*Psychrobacter cryohalolentis*	2511	558	638	4684
			494	*Legionella pneumophila Lens*	2934	557	693	4630
			495	*Methylococcus capsulatus*	2956	556	659	4665
			496	*Granulobacter bethesdensis*	2437	556	514	4810
			497	*Legionella pneumophila Paris*	3166	555	732	4593
			498	*Legionella pneumophila Philadelphia 1*	2942	554	723	4603
			499	*Roseobacter denitrificans*	4129	554	557	4769
			500	*Dinoroseobacter shibae*	4187	554	522	4804
			501	*Acidovorax sp. JS42*	4155	553	665	4662
			502	*Gluconobacter oxydans*	2664	548	583	4749
			503	*Bordetella pertussis*	3436	547	528	4805
			504	*Sphingopyxis alaskensis*	3195	547	552	4781
			505	*Magnetospirillum magneticum*	4559	545	631	4704
			506	*Silicibacter sp. TM1040*	3864	545	580	4755
			507	*Pelobacter propionicus*	3804	545	581	4754
			508	*Jannaschia sp. CCS1*	4283	544	631	4705
			509	*Geobacter sulfurreducens*	3446	543	543	4794
			510	*Nitrosococcus oceani*	3017	542	592	4746
			511	*Nitrobacter winogradskyi*	3122	541	560	4779
			512	*Francisella philomiragia*	1915	539	498	4843
			513	*Idiomarina loihiensis*	2628	537	579	4764
			514	*Syntrophobacter fumaroxidans*	4064	536	690	4654
			515	*Sodalis glossinidius*	2516	536	490	4854
			516	*Nitrosospira multiformis*	2805	532	504	4844
			517	*Pelobacter carbinolicus*	3352	530	523	4827
			518	*Methylobacillus flagellatus*	2753	529	626	4725
			519	*Psychrobacter sp. PRwf-1*	2385	528	632	4720
			520	*Nitrosomonas eutropha*	2551	527	535	4818

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Proteobacteria [continued...]	521	*Nitrosomonas europaea*	2461	525	512	4843
			522	*Desulfotalea psychrophila*	3234	524	568	4788
			523	*Polynucleobacter sp. QLW-P1DMWA-1*	2077	521	588	4771
			524	*Actinobacillus pleuropneumoniae AP76 (serotype 7)*	2142	521	493	4866
			525	*Maricaulis maris*	3063	519	525	4836
			526	*Mannheimia succiniciproducens*	2380	515	516	4849
			527	*Bdellovibrio bacteriovorus*	3587	515	653	4712
			528	*Francisella tularensis subsp. novicida U112*	1719	515	512	4853
			529	*Actinobacillus pleuropneumoniae L20 (serotype 5b)*	2142	513	493	4874
			530	*Zymomonas mobilis*	1998	511	501	4868
			531	*Actinobacillus succinogenes*	2079	510	516	4854
			532	*Psychrobacter arcticum*	2120	510	594	4776
			533	*Thiobacillus denitrificans*	2827	506	608	4766
			534	*Pasteurella multocida*	2015	505	479	4896
			535	*Francisella tularensis subsp. holarctica LVS*	1754	504	483	4893
			536	*Halorhodospira halophila*	2407	503	527	4850
			537	*Magnetococcus sp. MC-1*	3716	495	690	4695
			538	*Desulfovibrio desulfuricans*	3775	494	586	4800
			539	*Actinobacillus pleuropneumoniae JL03 (serotype 3)*	2036	493	504	4883
			540	*Neisseria meningitidis FAM18 (serogroup C)*	1917	487	480	4913
			541	*Francisella tularensis subsp. tularensis WY96-3418*	1634	487	483	4910
			542	*Candidatus Desulfococcus oleovorans*	3265	487	610	4783
			543	*Francisella tularensis subsp. tularensis SCHU S4*	1603	485	485	4910
			544	*Desulfovibrio vulgaris Hildenborough*	3531	483	596	4801
			545	*Haemophilus influenzae Rd KW20 (serotype d)*	1657	483	452	4945
			546	*Neisseria meningitidis MC58 (serogroup B)*	2063	483	479	4918
			547	*Xylella fastidiosa 9a5c*	2832	482	505	4893
			548	*Neisseria meningitidis Z2491 (serogroup A)*	2049	481	489	4910
			549	*Desulfovibrio vulgaris DP4*	3091	481	595	4804
			550	*Syntrophus aciditrophicus*	3168	480	511	4889
			551	*Haemophilus influenzae 86-028NP (nontypeable)*	1792	478	450	4952
			552	*Xylella fastidiosa Temecula1*	2036	478	504	4898
			553	*Francisella tularensis subsp. mediasiatica FSC147*	1406	477	468	4935
			554	*Francisella tularensis subsp. holarctica FTNF002-00*	1580	476	460	4944
			555	*Thiomicrospira crunogena*	2196	474	534	4872
			556	*Francisella tularensis subsp. tularensis FSC 198*	1605	474	499	4907
			557	*Nitratiruptor sp. SB155-2*	1843	470	658	4752
			558	*Neisseria gonorrhoeae*	2002	467	466	4947
			559	*Coxiella burnetii Dugway 5J108-111*	2125	464	587	4829
			560	*Francisella tularensis subsp. holarctica OSU18*	1555	461	481	4938
			561	*Haemophilus somnus 2336*	1980	457	474	4949
			562	*Coxiella burnetii RSA 331*	1975	454	596	4830
			563	*Coxiella burnetii RSA 493*	2052	454	608	4818
			564	*Haemophilus somnus 129PT*	1798	452	462	4966
			565	*Haemophilus influenzae PittEE*	1619	447	445	4988
			566	*Sulfurovum sp. NBC37-1*	2438	445	569	4866
			567	*Haemophilus influenzae PittGG*	1667	427	420	5033
			568	*Candidatus Pelagibacter ubique*	1354	421	416	5043
			569	*Campylobacter fetus*	1719	418	446	5016
			570	*Candidatus Ruthia magnifica*	976	411	498	4971
			571	*Thiomicrospira denitrificans*	2096	410	492	4978

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Proteobacteria [continued...]	572	*Bartonella tribocorum*	2092	406	432	5042
			573	*Bartonella henselae*	1488	402	408	5070
			574	*Wolinella succinogenes*	2042	390	459	5031
			575	*Dichelobacter nodosus*	1280	385	448	5047
			576	*Bartonella bacilliformis*	1283	384	387	5109
			577	*Haemophilus ducreyi*	1717	383	360	5137
			578	*Bartonella quintana*	1142	380	381	5119
			579	*Candidatus Vesicomyosocius okutanii*	937	379	490	5011
			580	*Helicobacter hepaticus*	1875	378	482	5020
			581	*Campylobacter jejuni NCTC11168*	1634	369	481	5030
			582	*Campylobacter jejuni RM1221*	1838	366	480	5034
			583	*Baumannia cicadellinicola*	595	362	319	5199
			584	*Campylobacter jejuni 81116*	1626	362	485	5033
			585	*Campylobacter jejuni subsp. doylei 269.97*	1731	360	452	5068
			586	*Campylobacter jejuni 81-176*	1758	357	492	5031
			587	*Campylobacter curvus*	1931	356	431	5093
			588	*Wolbachia wMel*	1195	354	766	4790
			589	*Campylobacter hominis ATCC BAA-381*	1687	351	408	5121
			590	*Helicobacter pylori J99*	1489	347	418	5115
			591	*Candidatus Blochmannia pennsylvanicus*	610	345	299	5236
			592	*Helicobacter pylori 26695*	1576	344	458	5078
			593	*Helicobacter pylori HPAG1*	1544	343	438	5099
			594	*Buchnera aphidicola APS*	574	341	283	5256
			595	*Campylobacter concisus 13826*	1985	336	437	5107
			596	*Candidatus Blochmannia floridanus*	583	335	287	5258
			597	*Buchnera aphidicola Sg*	546	323	280	5277
			598	*Wolbachia wBm*	805	320	355	5205
			599	*Lawsonia intracellularis*	1337	320	411	5149
			600	*Anaplasma marginale*	949	319	359	5202
			601	*Rickettsia felis*	1512	318	414	5148
			602	*Ehrlichia canis*	925	316	358	5206
			603	*Buchnera aphidicola Bp*	507	316	277	5287
			604	*Rickettsia bellii RML369-C*	1429	315	407	5158
			605	*Ehrlichia ruminantium Welgevonden (France)*	958	314	351	5215
			606	*Rickettsia conorii*	1374	312	372	5196
			607	*Ehrlichia ruminantium Gardel*	950	311	349	5220
			608	*Ehrlichia chaffeensis*	1105	310	375	5195
			609	*Ehrlichia ruminantium Welgevonden (South Africa)*	888	310	352	5218
			610	*Anaplasma phagocytophilum*	1264	309	374	5197
			611	*Rickettsia rickettsii Sheila Smith*	1345	309	370	5201
			612	*Rickettsia massiliae*	980	309	357	5214
			613	*Wigglesworthia glossinidia*	617	307	336	5237
			614	*Rickettsia akari*	1259	306	375	5199
			615	*Rickettsia prowazekii*	835	306	351	5223
			616	*Rickettsia bellii OSU 85-389*	1475	304	397	5179
			617	*Neorickettsia sennetsu*	932	297	344	5239
			618	*Rickettsia canadensis*	1093	296	349	5235
			619	*Rickettsia typhi*	838	295	363	5222
			620	*Buchnera aphidicola Cc*	357	254	210	5416
			621	*Orientia tsutsugamushi Boryong*	1182	236	332	5312
			622	*Cupriavidus taiwanensis*	1550	218	410	5252
			623	*Helicobacter acinonychis*	1618	196	485	5199
			624	*Candidatus Carsonella ruddii*	182	75	136	5669

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
BACTERIA [continued...]	BACTERIA [continued...]	Spirochete	625	*Leptospira biflexa serovar Patoc Patoc 1 (Ames)*	3600	548	632	4700
			626	*Leptospira biflexa serovar Patoc Patoc 1 (Paris)*	3726	548	632	4700
			627	*Leptospira interrogans serovar lai*	4727	518	659	4703
			628	*Leptospira borgpetersenii L550*	2945	517	629	4734
			629	*Leptospira interrogans serovar Copenhageni*	3658	516	656	4708
			630	*Leptospira borgpetersenii JB197*	2880	508	643	4729
			631	*Treponema denticola*	2767	397	404	5079
			632	*Treponema pallidum subsp. pallidum Nichols*	1036	272	335	5273
			633	*Borrelia afzelii*	1214	245	301	5334
			634	*Borrelia burgdorferi*	1640	244	310	5326
			635	*Borrelia garinii*	932	244	313	5323
		Synergistetes	636	*Caldicellulosiruptor saccharolyticus*	2679	489	564	4827
			637	*Syntrophomonas wolfei*	2504	426	561	4893
		Tenericutes	638	*Acholeplasma laidlawii*	1380	335	367	5178
			639	*Mesoplasma florum*	682	295	344	5241
			640	*Mycoplasma penetrans*	1037	294	380	5206
			641	*Mycoplasma mobile*	633	268	309	5303
			642	*Mycoplasma capricolum*	812	266	374	5240
			643	*Mycoplasma pneumoniae*	689	264	324	5292
			644	*Mycoplasma mycoides*	1016	255	380	5245
			645	*Mycoplasma genitalium*	477	246	335	5299
			646	*Mycoplasma gallisepticum*	726	245	346	5289
			647	*Mycoplasma pulmonis*	782	241	336	5303
			648	*Mycoplasma agalactiae*	742	232	328	5320
			649	*Phytoplasma OY*	754	225	196	5459
			650	*Mycoplasma synoviae*	672	224	313	5343
			651	*Mycoplasma arthritidis*	631	219	314	5347
			652	*Phytoplasma AYWB*	693	218	202	5460
			653	*Ureaplasma parvum serovar 3 ATCC 700970*	614	215	305	5360
			654	*Mycoplasma hyopneumoniae J*	665	195	226	5459
			655	*Mycoplasma hyopneumoniae 7448*	663	195	226	5459
			656	*Mycoplasma hyopneumoniae 232*	691	192	212	5476
Archea	Archea	Crenarchaeota	657	*Sulfolobus solfataricus*	2977	508	628	4744
			658	*Sulfolobus acidocaldarius*	2223	477	669	4734
			659	*Sulfolobus tokodaii*	2825	472	675	4733
			660	*Metallosphaera sedula*	2256	450	677	4753
			661	*Caldivirga maquilingensis*	1963	447	653	4780
			662	*Pyrobaculum aerophilum*	2605	414	671	4795
			663	*Pyrobaculum islandicum*	1978	406	1204	4270
			664	*Pyrobaculum arsenaticum*	2299	405	658	4817
			665	*Pyrobaculum calidifontis*	2149	393	688	4799
			666	*Aeropyrum pernix*	1700	383	534	4963
			667	*Thermofilum pendens*	1876	372	1110	4398
			668	*Hyperthermus butylicus*	1602	363	458	5059
			669	*Ignicoccus hospitalis*	1434	362	497	5021
			670	*Staphylothermus marinus*	1570	356	499	5025

Table S1.1 [continued…]

Domain	KINGDOM	PHYLUM	NO.	ORGANISM NAME	Number of annotated proteins in genome	Number of *S. cerevisiae* proteins with orthologs in organism	Number of *S. cerevisiae* proteins with homologs in organism	Number of *S. cerevisiae* proteins absent in organism
Archea	Archea	Euryarchaeota	671	*Methanosarcina acetivorans*	4540	588	605	4687
			672	*Methanosarcina barkeri*	3624	575	670	4635
			673	*Methanosarcina mazei*	3370	555	556	4769
			674	*Methanospirillum hungatei*	3139	530	644	4706
			675	*Haloarcula marismortui*	4240	520	675	4685
			676	*Candidatus Methanoregula boonei*	2450	518	588	4774
			677	*Uncultured methanogenic archaeon RC-I*	3085	504	615	4761
			678	*Pyrococcus furiosus*	2125	498	600	4782
			679	*Picrophilus torridus*	1535	498	665	4717
			680	*Methanoculleus marisnigri*	2489	490	610	4780
			681	*Methanococcoides burtonii*	2273	481	523	4876
			682	*Natronomonas pharaonis*	2822	478	626	4776
			683	*Halobacterium sp. NRC-1*	2622	465	563	4852
			684	*Halobacterium salinarum R1*	2749	463	569	4848
			685	*Methanobacterium thermoautotrophicum*	1873	454	487	4939
			686	*Thermococcus kodakaraensis*	2306	453	539	4888
			687	*Methanobrevibacter smithii ATCC 35061*	1793	453	466	4961
			688	*Methanocorpusculum labreanum*	1739	451	492	4937
			689	*Haloquadratum walsbyi*	2646	447	589	4844
			690	*Methanosaeta thermophila*	1696	446	507	4927
			691	*Methanococcus vannielii*	1678	435	484	4961
			692	*Pyrococcus abyssi*	1898	433	538	4909
			693	*Methanococcus maripaludis S2*	1722	431	487	4962
			694	*Methanococcus jannaschii*	1786	426	425	5029
			695	*Thermoplasma volcanium*	1499	423	591	4866
			696	*Methanosphaera stadtmanae*	1534	421	473	4986
			697	*Methanococcus maripaludis C7*	1788	421	498	4961
			698	*Methanococcus maripaludis C5*	1822	420	486	4974
			699	*Thermoplasma acidophilum*	1482	417	586	4877
			700	*Methanococcus aeolicus*	1490	412	437	5031
			701	*Pyrococcus horikoshii*	1955	410	510	4960
			702	*Methanopyrus kandleri*	1687	366	433	5081
		Korarchaeota	703	*Candidatus Korarchaeum cryptofilum*	1602	401	638	4841
		Nanoarchaeum equitans	704	*Nanoarchaeum equitans*	536	201	310	5369

Table S1.2 Summary of the comparison between *S. cerevisiae sequences* and those of organisms from different groups for domains, kingdoms or phyla, classified by biological process, molecular function and cellular localization from the GOSLIM ontology.

(See the Tables S1.2 A-I. as following)

Table S1.2. A *S. cerevisiae* Orthology analysis with Biological Process(es) of GO

Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have orthologs in all organisms of the corresponding group. Columns numbered 1-32 represent each a biological process. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red colour scale indicates the similarity between the set of proteins associated to the biological process in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of orthologs to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

Supplementary Table S1.2. A. *S. cerevisiae* Orthology analysis with Biological Process(es) of GO [continued…].

Supplementary Table S1.2 (A) *S. cerevisiae* ORTHOLOGY ANALYSIS WITH BIOLOGICAL PROCESS(ES) OF GO.

BIOLOGICAL PROCESS TERMS		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
S. cerevisiae genome size (5880)		1706 (29%)	1133 (19%)	1054 (18%)	985 (17%)	589 (10%)	554 (9%)	521 (9%)	451 (8%)	415 (7%)	389 (7%)	272 (5%)	255 (4%)	251 (4%)	248 (4%)	228 (4%)	221 (4%)	218 (4%)	182 (3%)	169 (3%)	164 (3%)	162 (3%)	161 (3%)	150 (3%)	134 (2%)	132 (2%)	121 (2%)	99 (2%)	91 (2%)	86 (1%)	84 (1%)	65 (1%)	43 (1%)
ORTHOLOGS	E.1.2 (2310(39%))	393	455	420	523	184	248	282	199	184	146	152	140	169	120	108	98	107	85	98	67	87	106	55	81	57	45	38	52	48	49	29	12
	E.3.1 (2269(39%))	402	504	408	431	193	258	301	204	167	162	143	141	169	98	89	120	88	90	109	43	57	116	61	82	46	38	49	58	51	30	24	19
	E.3.3 (2244(38%))	421	485	394	422	185	258	298	195	168	157	144	131	164	99	91	117	91	89	105	43	58	115	57	81	44	39	48	59	51	32	23	19
	E.2.1 (2204(38%))	366	528	415	416	224	261	317	221	163	178	130	142	141	99	108	117	97	82	83	48	71	113	67	73	45	49	46	42	28	37	27	17
	E.4.2 (2174(37%))	387	497	413	390	202	262	304	214	162	169	129	144	139	101	108	122	101	81	85	43	69	108	62	81	41	43	41	42	36	37	27	20
	E.1.1 (2107(36%))	324	420	408	469	197	260	298	206	179	147	142	132	149	111	110	109	103	70	78	57	82	102	52	75	46	50	36	43	42	49	30	16
	E.4.3 (2007(34%))	334	455	366	368	151	232	279	194	157	151	126	114	153	81	95	109	99	84	90	43	66	104	56	73	39	43	44	48	36	38	25	16
	E.3.4 (1766(30%))	291	389	346	302	155	215	251	177	132	146	110	88	155	66	69	94	68	79	91	30	47	114	58	75	34	29	45	56	40	21	19	15
	E.2.2 (1728(29%))	286	407	339	330	170	200	269	177	136	131	118	101	117	85	85	99	90	68	70	41	62	90	56	65	39	37	34	28	20	31	22	12
	E.3.2 (1644(30%))	282	392	307	300	138	204	240	169	127	131	98	79	129	68	69	97	72	62	76	29	43	91	58	63	35	31	36	36	32	22	20	14
	E.2.3 (1592(27%))	266	390	326	332	161	197	244	167	126	135	91	91	88	84	83	86	82	70	48	41	60	70	53	63	35	38	41	25	13	30	24	15
	E.3 (1371(23%))	227	315	268	256	116	176	209	155	106	115	73	60	117	55	64	82	63	58	65	22	39	82	52	58	28	26	32	36	27	18	17	13
	E.4.5 (1300(22%))	236	353	261	245	129	168	217	157	104	120	78	50	59	69	83	79	80	36	18	29	48	24	49	54	32	33	8	10	7	26	16	11
	E.4.6 (1281(22%))	232	288	239	256	93	168	196	144	103	106	78	68	74	60	63	80	63	55	46	26	39	71	46	54	21	23	31	26	19	20	13	11
	E.4.7 (1257(21%))	219	316	249	236	111	182	230	157	106	121	79	47	66	64	73	83	70	28	19	20	46	39	52	48	29	28	8	22	10	18	15	14
	E.1.3 (1015(17%))	159	305	220	174	132	160	196	143	89	112	72	39	37	43	57	64	54	25	12	30	43	23	44	39	28	30	6	10	4	22	15	11
	E.4.4 (959(16%))	168	260	203	171	90	136	179	133	76	88	57	33	49	45	66	56	52	25	18	24	39	24	41	45	26	29	4	14	8	19	12	8
	E.4.1 (839(14%))	152	233	183	174	83	122	154	114	57	87	30	25	36	47	55	63	43	29	14	19	36	17	37	35	21	22	10	2	4	16	10	11
	E.1 (781(13%))	116	216	176	142	91	127	170	115	77	86	52	27	21	40	54	52	47	17	7	26	39	12	34	30	23	27	6	3	1	22	14	5
	B.1 (708(12%))	105	123	111	110	41	87	92	47	87	42	64	30	105	18	35	26	15	43	65	14	13	59	22	30	11	12	24	36	33	6	9	3
	B.12 (632(11%))	91	102	109	88	29	84	76	41	65	45	51	26	111	15	34	20	12	35	62	13	11	54	22	20	7	9	20	30	33	5	7	3
	E.4 (591(10%))	113	166	136	122	58	87	127	92	45	59	18	13	22	31	51	46	37	12	5	14	29	7	27	29	15	19	2	0	2	13	8	8
	E.2.4 (573(10%))	72	111	127	138	63	93	119	83	81	41	36	32	22	35	60	39	40	32	12	18	27	17	27	29	18	20	14	7	3	19	14	5
	E.2 (480(8%))	49	98	108	118	57	81	117	73	71	36	30	25	20	33	58	34	32	23	11	16	27	12	23	24	17	19	9	6	3	18	14	5
	B.8 (419(7%))	69	82	78	66	17	54	12	10	35	34	40	15	57	9	3	19	1	25	52	6	1	33	8	18	3	3	10	27	26	1	2	0
	A.3 (401(7%))	81	98	66	60	26	32	21	29	25	28	17	15	74	5	9	21	4	18	25	1	2	18	10	19	3	2	11	15	16	3	1	1
	B.6 (401(7%))	62	59	67	56	7	31	19	10	31	15	33	16	93	10	7	16	3	30	51	7	3	39	3	12	4	4	19	23	27	3	4	0
	B.10 (341(6%))	51	58	62	57	9	33	12	6	30	13	27	16	74	10	2	13	1	22	43	6	1	33	4	13	2	3	11	19	23	1	3	0
	B.15 (319(5%))	42	40	63	59	10	33	15	15	25	20	27	9	69	9	5	15	4	20	47	6	2	28	9	8	2	2	9	19	22	3	2	1
	B.9 (279(5%))	44	47	53	46	8	25	6	5	18	12	21	11	68	9	3	14	0	18	38	6	1	23	4	10	3	3	8	14	21	0	1	0
	B.11 (275(5%))	49	37	55	45	5	27	7	8	18	17	0	9	45	6	3	15	0	18	35	5	0	18	7	13	3	2	6	13	17	0	1	0
	E (263(5%))	23	53	74	67	25	45	92	50	33	22	12	7	2	23	45	25	22	8	2	12	23	0	18	14	10	15	2	0	0	12	8	2
	B.5 (263(5%))	35	36	52	45	4	30	8	5	25	13	0	10	60	8	2	12	0	21	38	7	0	31	4	12	1	1	10	17	15	1	2	0
	B.3 (236(4%))	37	50	46	40	7	24	8	8	19	15	22	10	42	7	3	13	1	13	28	5	1	17	7	7	1	2	7	8	13	0	1	0
	B.4 (233(4%))	39	41	64	37	12	35	7	7	19	22	0	11	32	9	4	16	1	22	11	0	0	21	5	12	1	1	11	14	6	0	1	0
	A.4 (201(3%))	50	76	32	29	17	16	5	16	10	23	0	3	25	3	2	19	0	4	1	0	1	0	4	7	1	0	3	0	0	0	0	0
	A.1 (164(3%))	36	67	22	26	15	8	0	12	6	15	0	4	25	4	2	18	0	6	3	0	0	4	3	6	1	0	4	0	3	0	1	0
	A.2 (148(3%))	30	43	21	27	10	6	2	11	3	12	0	6	19	5	2	17	0	2	11	0	0	3	3	2	1	0	0	4	7	0	0	0
	B.2 (146(3%))	25	17	39	31	2	8	3	2	4	2	0	6	24	6	2	12	1	13	25	5	0	10	0	4	1	1	4	8	14	1	1	0
	B.14 (138(2%))	28	26	47	29	6	17	1	5	8	8	0	5	22	6	2	15	0	10	7	1	0	3	3	4	0	1	2	1	6	0	1	0
	A (103(2%))	23	36	15	23	9	3	0	0	3	10	0	2	10	3	2	16	0	0	0	0	0	0	2	2	1	0	0	0	0	0	0	0
	B.7 (96(2%))	20	18	25	23	2	10	1	2	5	3	0	2	18	3	2	12	0	2	6	0	0	4	1	7	0	1	2	2	5	0	1	0
	B.13 (11(0.2 %))	3	0	2	3	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	A & B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	B.16 (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

1	NOT_ANNOTATED
2	RNA metabolic process
3	organelle organization
4	transport
5	transcription
6	response to stress
7	protein modification process
8	cell cycle
9	response to chemical stimulus
10	DNA metabolic process
11	carbohydrate metabolic process
12	lipid metabolic process
13	cellular amino acid and derivative metabolic process
14	membrane organization
15	signal transduction
16	protein catabolic process
17	cytoskeleton organization
18	generation of precursor metabolites and energy
19	heterocycle metabolic process
20	cell wall organization
21	anatomical structure morphogenesis
22	cofactor metabolic process
23	meiosis
24	cellular homeostasis
25	sporulation resulting in formation of a cellular spore
26	conjugation
27	cellular respiration
28	vitamin metabolic process
29	cellular aromatic compound metabolic process
30	cell budding
31	pseudohyphal growth
32	transposition

Supplementary Table I (A). Summary of the comparison between *S. cerevisiae* sequences and those of organisms from different groups for domains, kingdoms or phyla, classified by biological process from the GOSLIM ontology. Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have orthologs in all organisms of the corresponding group. Columns numbered 1-32 represent each a biological process. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red color scale indicates the similarity between the set of proteins associated to the biological process in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of orthologs to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

E	EUKARYOTA												A	ARCHEA	B	BACTERIA					
E.1	FUNGI	E.2	ANIMAL	E.3	PLANTS	E.4	PROTISTS						A.1	Crenarchaeota	B.1	Acidobacteria	B.6	Deinococcus-Thermus	B.11	Hyperthermophilic bacteria
E.1.1	Ascomycetes	E.2.1	Echinoderms	E.3.1	Dicotyledons	E.4.1	Alveolates	E.4.5	Entamoeba				A.2	Euryarchaeota	B.2	Actinobacteria	B.7	Firmicutes	B.12	Planctomyces
E.1.2	Basidiomycetes	E.2.2	Insects	E.3.2	Green algae	E.4.2	Cellular slime molds	E.4.6	Euglenozoa				A.3	Korarchaeota	B.3	Bacteroides	B.8	Fusobacteria	B.13	Proteobacteria
E.1.3	Microsporidians	E.2.3	Nematodes	E.3.3	Monocotyledons	E.4.3	Choanoflagellates	E.4.7					A.4	Nanoarchaeota	B.4	Chlamydia	B.9	Green nonsulfur bacteria	B.14	Spirochete
		E.2.4	Vertebrates	E.3.4	Red algae	E.4.4	Diplomonads		Paraba salids						B.5	Cyanobacteria	B.10	Green sulfur bacteria	B.15	Synergistetes
																			B.16	Tenericutes

Table S1.2. B *S. cerevisiae* Homology analysis with Biological Process(es) of GO *S. cerevisiae* Homology analysis with Biological Process(es) of GO.

Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have homologues in all organisms of the corresponding group. Columns numbered 1-32 represent each a biological process. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red colour scale indicates the similarity between the set of proteins associated to the biological process in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of homologues to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

(See the Table S1.2 B. on next page)

Supplementary Table S1.2. B. *S. cerevisiae* Homology analysis with Biological Process(es) of GO [continued…].

Supplementary Table 1.2 - (B) *S. cerevisiae* HOMOLOGY ANALYSIS WITH BIOLOGICAL PROCESS(ES) OF GO.

BIOLOGICAL PROCESS TERMS		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
S. cerevisiae genome size (5880)		1706 (29%)	1133 (19%)	1054 (18%)	985 (17%)	589 (10%)	554 (9%)	521 (9%)	451 (8%)	415 (7%)	389 (7%)	272 (5%)	255 (4%)	251 (4%)	248 (4%)	228 (4%)	221 (4%)	218 (4%)	182 (3%)	169 (3%)	164 (3%)	162 (3%)	161 (3%)	150 (3%)	134 (2%)	132 (2%)	121 (2%)	99 (2%)	91 (2%)	86 (1%)	84 (1%)	65 (1%)	43 (1%)
HOMOLOGS	E.4.3 (1162(20%))	273	222	240	244	145	111	125	86	84	88	50	74	49	68	51	50	38	33	21	26	36	28	26	21	21	24	15	16	13	26	16	6
	E.4.2 (1156(20%))	263	216	217	273	133	109	109	80	80	66	53	62	54	57	44	46	39	26	30	19	43	30	18	25	19	20	13	16	16	23	13	6
	E.3.1 (1147(20%))	304	210	220	240	136	101	108	86	74	85	45	55	39	57	42	45	31	29	20	16	33	24	24	23	14	17	17	13	14	19	14	8
	E.3.3 (1091(19%))	259	212	221	232	128	97	107	87	69	84	43	61	40	58	41	51	32	29	22	21	29	26	27	22	19	18	15	11	15	16	17	5
	E.2.1 (1075(18%))	293	196	201	233	128	94	85	70	66	66	43	52	52	52	37	44	41	34	27	18	31	23	20	21	17	17	14	17	18	20	14	9
	E.4.7 (1054(20%))	210	241	217	237	121	104	99	94	72	80	51	52	46	45	51	47	47	26	29	22	33	33	26	21	18	23	11	17	9	18	10	9
	E.3.4 (1051(18%))	236	226	189	216	126	100	111	79	67	77	57	68	52	48	38	47	30	31	31	12	21	29	17	20	11	20	16	13	24	8	13	6
	E.4.5 (945(16%))	187	215	169	199	121	108	111	75	66	66	46	55	46	44	40	45	33	21	27	26	40	33	19	27	18	22	9	17	9	24	15	9
	E.4.4 (752(13%))	146	168	157	161	67	79	97	67	43	73	26	42	25	32	26	33	37	12	13	10	11	23	24	18	9	13	5	12	7	2	2	5
	E.1.3 (749(13%))	127	208	161	147	109	81	100	78	47	74	27	54	18	30	32	36	36	7	11	14	15	15	27	28	15	15	2	7	5	10	5	6
	B.1 (740(13%))	164	114	98	138	36	95	67	47	63	58	49	45	64	15	24	31	16	33	30	14	8	39	23	26	8	9	16	22	20	2	0	5
	B.12 (712(12%))	155	121	107	124	46	82	90	54	62	63	43	39	49	16	21	26	26	32	33	12	10	46	22	30	9	12	13	22	19	6	1	6
	E.4.6 (695(12%))	162	143	135	146	56	70	73	48	44	58	29	47	30	33	27	27	25	17	16	8	12	18	13	14	7	14	6	11	6	9	6	4
	E.2.3 (693(12%))	143	136	142	166	83	75	60	52	40	56	22	38	38	41	33	29	25	13	17	13	28	13	11	11	7	14	8	9	12	20	11	5
	E.2.2 (638(11%))	129	117	115	154	73	60	63	42	38	34	21	36	41	33	30	26	29	15	22	14	24	16	14	11	7	13	7	10	16	16	11	5
	A.3 (638(11%))	128	143	91	106	52	84	97	49	64	56	40	21	61	18	33	16	17	17	35	16	19	31	19	17	15	12	7	16	20	8	6	4
	E.3.2 (636(11%))	189	101	107	128	57	48	69	39	44	42	29	31	28	29	17	20	13	15	15	8	7	12	10	10	7	9	9	6	12	6	3	6
	B.8 (491(8%))	126	76	76	57	23	47	37	33	32	49	19	40	66	15	10	12	9	19	25	8	5	31	16	22	6	7	9	18	17	5	0	2
	E.1.2 (469(8%))	117	101	101	78	83	48	43	42	36	36	14	21	18	23	18	20	19	8	10	10	17	7	10	8	8	12	5	4	5	11	10	6
	B.15 (343(6%))	72	61	52	47	18	39	72	30	33	27	24	17	39	13	16	9	7	16	18	8	10	16	15	12	6	5	6	9	17	2	1	3
	B.6 (326(6%))	64	63	46	43	21	37	58	35	28	29	23	15	40	8	12	6	9	21	16	6	7	19	20	16	4	4	9	11	16	2	0	1
	A.4 (310(5%))	72	102	46	37	26	46	27	27	21	32	8	12	28	2	7	10	2	6	3	1	3	4	12	6	5	3	5	0	1	0	1	3
	B.10 (260(4%))	63	55	40	35	15	25	13	12	18	21	18	11	35	3	5	6	5	18	15	4	2	12	7	9	3	4	9	8	10	1	0	0
	E.3 (253(4%))	69	44	52	48	26	25	30	19	16	24	8	15	12	11	7	7	5	7	6	4	4	8	3	4	3	4	3	3	5	2	1	1
	B.11 (251(4%))	61	55	43	27	12	21	16	18	23	19	17	16	32	4	2	9	7	13	14	2	3	12	8	8	3	2	3	8	10	4	0	0
	E.4.1 (194(3%))	46	40	49	56	23	19	24	18	7	11	3	11	3	11	8	8	5	3	1	3	11	1	3	4	1	6	2	0	0	3	1	1
	B.4 (190(3%))	46	33	32	27	14	26	32	20	23	17	10	10	15	7	8	3	7	10	6	5	5	11	11	9	1	3	5	3	2	0	1	0
	B.9 (156(3%))	48	21	19	28	4	14	5	9	10	10	7	7	17	0	1	2	1	10	12	1	1	8	4	9	1	1	4	1	9	1	0	0
	E.2.4 (125(2%))	26	27	15	32	23	12	3	7	11	8	0	6	11	5	5	3	2	2	3	2	3	2	5	3	2	3	2	1	2	2	2	1
	B.3 (104(2%))	26	25	16	9	7	15	5	10	6	12	4	7	6	0	1	3	1	4	4	2	0	8	4	5	1	1	4	4	3	0	0	0
	B.5 (89(2%))	27	20	15	9	1	7	3	7	3	10	2	4	11	0	1	1	1	7	4	0	0	3	3	1	1	1	4	0	2	0	0	0
	A.1 (82(1%))	23	17	11	3	5	17	1	7	3	11	3	4	12	0	1	2	0	0	4	1	0	6	5	1	1	0	0	4	3	0	1	0
	E.2 (81(1%))	13	17	13	25	15	7	3	4	4	5	0	5	7	4	1	1	2	2	1	2	2	2	3	3	1	3	2	1	1	1	1	1
	B.14 (64(1%))	18	19	14	7	4	5	6	8	0	8	4	0	4	0	1	3	3	2	2	0	0	2	2	1	0	0	0	1	0	0	0	0
	E.4 (43(0.7%))	11	9	10	10	6	5	9	4	2	1	0	3	0	3	1	1	1	0	0	0	3	0	1	1	0	0	0	0	0	0	0	0
	B.2 (34(0.6%))	10	7	5	6	2	6	10	5	4	5	0	1	0	1	1	0	1	0	0	1	1	1	3	1	0	0	0	1	0	0	0	0
	B.16 (33(0.6%))	7	9	8	8	0	3	0	1	4	2	0	0	2	0	0	1	0	3	0	0	0	0	1	0	1	1	3	0	0	0	0	0
	A.2 (30(0.5%))	5	9	6	3	2	5	1	3	0	5	0	5	3	0	0	1	0	0	0	0	0	1	2	0	1	0	0	0	1	0	1	0
	A (18(0.3%))	4	4	6	3	2	5	0	3	0	5	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	B.7 (8(0.1%))	1	2	1	3	0	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	B.13 (1(0.01 %))	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E.1 (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E.1.1 (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	A & B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

1	NOT_ANNOTATED
2	RNA metabolic process
3	organelle organization
4	transport
5	transcription
6	response to stress
7	protein modification process
8	cell cycle
9	response to chemical stimulus
10	DNA metabolic process
11	carbohydrate metabolic process
12	lipid metabolic process
13	cellular amino acid and derivative metabolic process
14	membrane organization
15	signal transduction
16	protein catabolic process
17	cytoskeleton organization
18	generation of precursor metabolites and energy
19	heterocycle metabolic process
20	cell wall organization
21	anatomical structure morphogenesis
22	cofactor metabolic process
23	meiosis
24	cellular homeostasis
25	sporulation resulting in formation of a cellular spore
26	conjugation
27	cellular respiration
28	vitamin metabolic process
29	cellular aromatic compound metabolic process
30	cell budding
31	pseudohyphal growth
32	transposition
31	pseudohyphal growth
32	transposition

Supplementary Table I (B). Summary of the comparison between *S. cerevisiae* sequences and those of organisms from different groups for domains, kingdoms or phyla, classified by biological process from the GOSLIM ontology. Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have homologues in all organisms of the corresponding group. Columns numbered 1-32 represent each a biological process. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red color scale indicates the similarity between the set of proteins associated to the biological process in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of homologues to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

E	EUKARYOTA								A	ARCHEA	B	BACTERIA					
E.1	FUNGI	E.2	ANIMAL	E.3	PLANTS	E.4	PROTISTS		A.1	Crenarchaeota	B.1	Acidobacteria	B.6	Deinococcus-Thermus	B.11	Hyperthermophilic bacteria
E.1.1	Ascomycetes	E.2.1	Echinoderms	E.3.1	Dicotyledons	E.4.1	Alveolates	E.4.5	Entamoeba	A.2	Euryarchaeota	B.2	Actinobacteria	B.7	Firmicutes	B.12	Planctomyces
E.1.2	Basidiomycetes	E.2.2	Insects	E.3.2	Green algae	E.4.2	Cellular slime molds	E.4.6	Euglenozoa	A.3	Korarchaeota	B.3	Bacteroides	B.8	Fusobacteria	B.13	Proteobacteria
E.1.3	Microsporidians	E.2.3	Nematodes	E.3.3	Monocotyledons	E.4.3	Choanoflagellates	E.4.7		A.4	Nanoarchaeota	B.4	Chlamydia	B.9	Green nonsulfur bacteria	B.14	Spirochete
		E.2.4	Vertebrates	E.3.4	Red algae	E.4.4	Diplomonads		Paraba salids			B.5	Cyanobacteria	B.10	Green sulfur bacteria	B.15	Synergistetes
																B.16	Tenericutes

Table S1.2. C ***S. cerevisiae*** **Absent genes analysis with Biological Process(es) of GO**

Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that are simultaneously absent in all organisms of the corresponding group. Columns numbered 1-32 represent each a biological process. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red colour scale indicates the similarity between the set of proteins associated to the biological process in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity (low absent), while red entries indicate low similarity (highly absence) between the two sets of proteins. The numbers in each entry indicate the number of *S. cerevisiae* proteins that are absent in all organisms from the group represented in the row.

(See the Table S1.2 C. on next page)

Supplementary Table S1.2. C - *S. cerevisiae* Absent genes analysis with Biological Process(es) of GO. [continued…]

Supplemetary Table - S1.2 (C) *S. cerevisiae* ABSENT GENES ANALYSIS WITH BIOLOGICAL PROCESS(ES) OF GO.

BIOLOGICAL PROCESS TERMS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
S. cerevisiae genome size (5880)	1706 (29%)	1133 (19%)	1054 (18%)	985 (17%)	589 (10%)	554 (9%)	521 (9%)	451 (8%)	415 (7%)	389 (7%)	272 (5%)	255 (4%)	251 (4%)	248 (4%)	228 (4%)	221 (4%)	218 (4%)	182 (3%)	169 (3%)	164 (3%)	162 (3%)	161 (3%)	150 (3%)	134 (2%)	132 (2%)	121 (2%)	99 (2%)	91 (2%)	86 (1%)	84 (1%)	65 (1%)	43 (1%)
E.(0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.1 (10(0.2%))	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.1.1 (10(0.2%))	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.3 (142(2%))	71	10	26	17	10	10	6	13	11	7	7	2	1	4	6	3	7	6	5	5	3	0	2	0	3	2	2	0	1	1	2	0
E.3.1 (282(5%))	149	21	37	27	14	22	8	16	21	11	10	3	1	6	11	5	8	8	5	15	7	0	4	1	3	7	4	0	1	1	5	0
E.1.2 (1525(26%))	764	205	211	120	142	103	57	104	88	97	27	20	5	39	49	34	40	23	10	58	21	2	46	14	44	36	16	4	4	9	13	10
E.4 (1978(34%))	889	259	327	226	209	142	82	130	128	115	50	36	19	73	71	43	69	43	31	84	39	8	56	22	52	49	27	11	18	19	19	14
E.2 (2028(34%))	894	238	299	228	181	139	78	121	142	101	69	38	40	71	67	35	58	49	45	86	31	13	51	28	55	46	30	27	33	15	19	11
E.2.4 (2185(37%))	948	252	326	254	198	154	90	130	161	110	83	42	50	76	68	43	65	58	50	87	34	20	54	32	57	47	33	31	37	18	19	13
E.2.2 (2371(40%))	1005	285	364	276	211	162	93	139	161	124	78	62	60	84	75	43	74	58	50	94	41	24	57	33	58	50	35	34	36	21	22	13
E.3.2 (2519(43%))	1017	354	439	314	282	203	103	171	174	144	79	75	39	97	109	61	105	67	35	97	75	19	68	31	66	62	34	17	17	41	35	14
E.3.3 (2527(43%))	1027	347	422	325	270	197	116	168	171	142	85	63	47	90	96	53	95	64	41	100	70	20	66	30	64	64	35	21	20	36	25	19
E.2.3 (2540(43%))	1043	313	397	319	233	183	115	158	174	126	89	69	58	87	80	58	77	66	54	97	43	28	64	37	66	52	39	30	37	20	24	14
E.3.4 (2554(43%))	960	359	435	403	254	199	139	174	179	142	89	83	39	115	103	73	102	64	41	95	75	15	69	32	75	61	34	19	20	48	26	20
E.2.1 (2582(44%))	1047	320	421	330	231	197	119	159	179	139	99	61	58	96	83	60	80	66	59	98	55	25	63	39	65	55	38	32	40	27	24	17
E.4.2 (2603(44%))	1071	342	424	328	257	185	122	170	174	152	93	66	60	92	84	58	83	75	57	104	53	23	72	30	69	62	44	33	34	28	28	18
A & B (2642(45%))	916	477	541	446	341	220	202	225	163	167	73	81	20	142	109	106	119	64	24	88	89	10	66	34	65	66	39	4	10	47	38	23
E.4.1 (2740(47%))	1058	370	464	382	295	205	150	172	182	147	111	71	75	108	100	72	92	75	55	102	62	29	66	39	71	60	41	28	31	31	28	19
E.4.3 (2755(47%))	1111	375	442	380	283	217	136	178	176	150	101	72	51	99	85	67	83	67	58	96	58	32	71	41	68	58	39	30	37	22	26	21
B (2881(49%))	1005	564	573	475	379	239	214	238	176	182	75	84	22	150	113	112	124	66	25	90	92	10	72	36	70	68	40	4	10	50	41	27
E.4.6 (3157(54%))	1151	495	558	423	374	256	187	214	215	181	125	83	78	117	123	83	107	86	80	120	92	37	78	44	80	72	51	38	47	47	40	23
B.13 (3256(55 %))	1107	638	654	549	419	277	262	279	202	207	89	97	24	171	131	133	147	72	27	96	97	13	84	47	75	77	41	4	10	54	45	29
E.4.7 (3616(62%))	1289	497	587	521	355	277	208	209	241	186	145	158	138	141	110	93	108	125	121	123	81	92	75	65	81	70	79	54	67	48	40	20
A (3672(62%))	1225	625	725	610	422	308	302	298	224	219	126	156	48	186	144	137	160	99	45	121	113	46	87	55	87	89	58	22	15	64	48	27
E.4.5 (3678(63%))	1297	483	620	546	338	283	212	224	242	203	154	152	148	139	108	101	108	126	125	110	71	105	85	56	77	68	81	64	70	35	34	23
B.7 (3692(63%))	1228	719	732	611	463	317	305	307	212	239	112	128	37	193	150	158	161	84	38	114	118	25	96	48	92	84	53	11	13	62	54	32
A.2 (3807(65%))	1258	649	755	643	443	320	313	313	236	226	136	158	52	193	146	142	167	104	48	126	119	49	90	58	92	96	63	24	16	70	50	28
B.2 (3889(66%))	1305	755	760	636	480	331	312	331	232	253	123	135	44	199	156	142	176	96	37	118	120	32	99	63	94	87	54	10	13	65	51	33
B.5 (3940(67%))	1302	731	743	670	474	333	319	320	237	246	133	159	48	188	147	158	166	98	50	121	111	32	100	61	100	93	55	17	21	64	54	34
B.3 (4076(69%))	1345	768	786	694	484	344	340	338	245	254	127	150	53	204	160	160	178	98	49	127	127	39	103	69	102	98	59	21	22	68	55	34
E.1.3 (4097(70%))	1420	531	656	658	342	311	226	229	272	197	173	162	196	174	139	121	129	150	145	120	99	122	79	66	83	76	90	74	77	52	45	26
B.9 (4178(71%))	1369	772	774	727	488	358	349	338	257	258	146	161	61	204	157	162	175	96	48	129	126	46	104	77	104	96	53	18	16	69	54	34
E.4.4 (4195(71%))	1399	625	686	659	431	344	257	258	294	227	191	181	178	172	137	132	133	148	138	131	110	114	89	74	93	81	89	66	71	64	52	30
A.1 (4209(72%))	1348	733	824	699	467	372	349	338	270	259	151	186	69	203	165	166	181	112	69	138	123	74	107	72	101	97	62	37	27	65	52	34
B.14 (4363(74%))	1412	795	825	774	499	369	368	358	265	276	157	154	62	210	174	174	190	106	58	137	128	46	108	81	106	101	58	22	22	73	57	35
B.10 (4393(75%))	1406	785	810	757	500	385	394	362	281	264	166	180	75	213	177	167	190	103	70	139	134	61	113	76	110	106	58	36	32	71	58	35
B.6 (4408(75%))	1433	802	819	788	497	375	359	363	267	278	153	185	71	213	173	168	184	107	61	134	129	53	106	65	106	102	59	25	20	73	57	35
B.1 (4413(75%))	1437	807	828	731	506	370	362	356	258	283	159	180	82	214	169	164	187	106	74	136	136	63	105	77	108	100	58	33	33	76	56	35
B.12 (4517(77%))	1460	821	821	767	508	386	355	365	281	275	172	190	91	216	173	176	180	115	74	139	136	61	106	83	111	100	65	39	34	73	57	34
B.15 (4677(80%))	1461	856	861	810	522	417	383	375	304	296	176	202	94	217	183	184	199	122	70	144	137	81	113	85	111	107	70	39	27	76	56	37
B.11 (4685(80%))	1465	840	857	810	520	434	388	391	308	293	175	197	79	228	207	179	207	121	75	145	149	79	119	91	112	111	72	32	28	77	60	38
B.4 (4712(80%))	1483	836	844	816	521	401	381	365	294	278	187	194	139	223	175	176	195	120	89	143	136	75	111	90	110	104	65	36	43	76	58	38
B.16 (4727(80%))	1467	848	842	815	521	411	399	365	295	292	181	198	137	214	176	189	192	129	97	136	126	88	108	83	111	100	79	40	46	70	55	37
A.3 (4822(82%))	1497	803	891	813	505	436	403	372	319	298	216	219	116	224	186	184	197	147	109	147	136	112	121	97	109	107	80	60	50	73	58	38
B.8 (4951(84%))	1512	886	883	896	543	451	472	407	341	300	213	200	138	223	216	190	208	138	91	154	151	97	126	93	118	111	79	46	43	78	63	41
A.4 (5350(91%))	1584	866	969	913	540	491	489	407	377	328	264	240	198	242	219	192	216	172	165	163	153	157	134	120	121	118	90	91	85	84	64	40

	ABSENT GENES

	NOT_ANNOTATED
2	RNA metabolic process
3	organelle organization
4	transport
5	transcription
6	response to stress
7	protein modification process
8	cell cycle
9	response to chemical stimulus
10	DNA metabolic process
11	carbohydrate metabolic process
12	lipid metabolic process
13	cellular amino acid and derivative metabolic process
14	membrane organization
15	signal transduction
16	protein catabolic process
17	cytoskeleton organization
18	generation of precursor metabolites and energy
19	heterocycle metabolic process
20	cell wall organization
21	anatomical structure morphogenesis
22	cofactor metabolic process
23	meiosis
24	cellular homeostasis
25	sporulation resulting in formation of a cellular spore
26	conjugation
27	cellular respiration
28	vitamin metabolic process
29	cellular aromatic compound metabolic process
30	cell budding
31	pseudohyphal growth
32	transposition
31	pseudohyphal growth
32	transposition

Supplementary Table I (C). Summary of the comparison between *S. cerevisiae* sequences and those of organisms from different groups for domains, kingdoms or phyla, classified by biological process from the GOSLIM ontology. Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that are simultaneously absent in all organisms of the corresponding group. Columns numbered 1-32 represent each a biological process. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red color scale indicates the similarity between the set of proteins associated to the biological process in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity (low absent), while red entries indicate low similarity (highly absence) between the two sets of proteins. The numbers in each entry indicate the number of *S. cerevisiae* proteins that are absent in all organisms from the group represented in the row.

E	EUKARYOTA								
E.1	FUNGI	E.2	ANIMAL	E.3	PLANTS	E.4	PROTISTS		
E.1.1	Ascomycetes	E.2.1	Echinoderms	E.3.1	Dicotyledons	E.4.1	Alveolates	E.4.5	Entamoeba
E.1.2	Basidiomycetes	E.2.2	Insects	E.3.2	Green algae	E.4.2	Cellular slime molds	E.4.6	Euglenozoa
E.1.3	Microsporidians	E.2.3	Nematodes	E.3.3	Monocotyledons	E.4.3	Choanoflagellates	E.4.7	
		E.2.4	Vertebrates	E.3.4	Red algae	E.4.4	Diplomonads		Paraba salids

A	ARCHEA
A.1	Crenarchaeota
A.2	Euryarchaeota
A.3	Korarchaeota
A.4	Nanoarchaeota

B	BACTERIA				
B.1	Acidobacteria	B.6	Deinococcus-Thermus	B.11	Hyperthermophilic bacteria
B.2	Actinobacteria	B.7	Firmicutes	B.12	Planctomyces
B.3	Bacteroides	B.8	Fusobacteria	B.13	Proteobacteria
B.4	Chlamydia	B.9	Green nonsulfur bacteria	B.14	Spirochete
B.5	Cyanobacteria	B.10	Green sulfur bacteria	B.15	Synergistetes
				B.16	Tenericutes

Table S1.2. D ***S. cerevisiae*** **Orthology analysis with Molecular Function(s) of GO**

Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have orthologs in all organisms of the corresponding group. Columns numbered 1-20 represent each a molecular function. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red colour scale indicates the similarity between the set of proteins associated to the molecular function in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of orthologs to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

(See the Table S1.2 D. on next page)

Supplementary Table S1.2. D - *S. cerevisiae* Orthology analysis with Molecular Function(s) of GO. [continued…]

Supplementary Table - S1.2 (D) *S. cerevisiae* ORTHOLOGY ANALYSIS WITH MOLECULAR FUNCTION(S) OF GO.

MOLECULAR FUNCTION TERMS		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
S. cerevisiae genome size (5880)		2782 (47%)	764 (13%)	629 (11%)	346 (6%)	328 (6%)	319 (5%)	292 (5%)	269 (5%)	208 (4%)	166 (3%)	108 (2%)	83 (1%)	83 (1%)	73 (1%)	70 (1%)	56 (1%)	52 (1%)	47 (1%)	44 (1%)	16 (0.3%)
ORTHOLOGS	E.1.2 (2310(39%))	693	464	381	167	106	75	104	174	84	89	69	59	55	42	40	43	37	33	25	12
	E.3.1 (2269(39%))	599	475	393	169	90	72	125	186	75	117	71	58	63	45	21	49	40	32	24	11
	E.3.3 (2244(38%))	609	472	383	163	89	66	118	179	75	114	70	58	60	46	23	50	40	34	23	11
	E.2.1 (2204(38%))	558	501	363	159	109	87	120	178	78	110	79	60	42	48	28	45	38	39	23	12
	E.4.2 (2174(37%))	563	486	376	159	94	75	111	158	88	113	74	60	48	45	25	45	36	34	23	12
	E.1.1 (2107(36%))	611	441	375	119	109	84	85	143	92	80	66	59	50	32	33	32	32	36	27	11
	E.4.3 (2007(34%))	479	451	347	161	100	59	111	167	72	112	74	59	43	38	20	47	34	31	24	12
	E.3.4 (1766(30%))	364	412	318	166	77	49	96	151	60	110	59	58	46	38	8	45	36	29	20	6
	E.2.2 (1728(29%))	400	415	301	139	97	68	84	119	63	103	61	58	33	35	20	39	32	32	21	12
	E.3.2 (1644(30%))	386	393	295	151	69	36	91	126	49	102	60	58	44	39	12	43	32	26	20	11
	E.2.3 (1592(27%))	352	389	278	136	91	59	89	111	58	82	55	56	25	38	21	28	32	29	20	12
	E.3 (1371(23%))	252	349	246	147	64	30	75	100	42	97	48	54	34	30	8	40	31	23	19	5
	E.4.5 (1300(22%))	263	362	222	140	71	40	88	40	50	66	46	58	21	30	17	34	27	24	21	11
	E.4.6 (1281(22%))	267	345	227	124	56	13	67	98	36	78	49	53	24	27	5	34	26	23	18	12
	E.4.7 (1257(21%))	253	358	236	128	59	24	65	55	46	63	52	59	21	30	11	31	22	28	23	7
	E.1.3 (1015(17%))	163	307	189	108	75	54	60	42	26	66	36	55	13	31	6	24	20	18	19	11
	E.4.4 (959(16%))	160	288	190	114	49	14	59	32	33	55	30	54	12	22	3	29	21	28	18	7
	E.4.1 (839(14%))	159	260	139	103	51	19	54	20	29	53	31	51	6	20	4	20	20	19	17	11
	E.1 (781(13%))	119	266	159	70	59	35	38	28	23	38	30	52	9	20	5	17	16	17	18	10
	B.1 (708(12%))	131	180	167	25	22	9	22	92	9	51	21	34	36	8	1	25	14	0	15	0
	B.12 (632(11%))	82	168	155	23	21	9	17	93	11	57	15	39	35	8	1	19	14	0	11	0
	E.4 (591(10%))	93	215	106	80	37	9	38	7	16	40	23	48	1	10	1	12	14	16	16	5
	E.2.4 (573(10%))	92	187	98	35	46	30	19	39	22	30	22	31	11	2	4	14	8	17	20	5
	E.2 (480(8%))	71	168	94	24	39	27	13	30	16	27	21	31	4	1	3	12	7	16	19	5
	B.8 (419(7%))	52	143	75	21	16	3	11	44	6	44	13	30	27	5	0	16	13	2	0	0
	A.3 (401(7%))	47	120	51	60	17	3	24	50	2	34	19	10	19	8	0	4	15	8	0	0
	B.6 (401(7%))	50	112	79	22	5	0	11	56	6	47	14	18	34	4	0	6	9	0	1	0
	B.10 (341(6%))	32	105	61	16	5	2	9	44	6	45	11	22	25	5	1	11	13	0	0	0
	B.15 (319(5%))	27	98	73	21	13	1	6	31	8	34	12	1	27	5	2	8	9	0	4	0
	B.9 (279(5%))	26	94	47	18	9	1	5	26	6	38	10	21	22	5	1	9	11	0	2	0
	B.11 (275(5%))	31	91	44	23	9	0	6	31	6	31	11	1	15	3	1	8	11	0	1	0
	E (263(5%))	33	123	63	10	13	5	5	0	4	16	11	31	0	0	0	8	6	13	14	0
	B.5 (263(5%))	17	80	51	10	6	0	4	41	3	38	10	3	22	3	0	8	9	0	0	0
	B.3 (236(4%))	16	91	36	13	8	1	8	34	6	36	11	23	13	4	1	5	13	0	1	0
	B.4 (233(4%))	23	78	32	20	13	3	6	30	5	33	12	4	12	3	0	5	9	0	0	0
	A.4 (201(3%))	21	64	16	52	13	2	19	5	2	24	17	5	3	8	1	1	12	0	0	0
	A.1 (164(3%))	16	52	14	34	10	2	13	5	0	22	16	5	5	6	0	2	10	0	0	0
	A.2 (148(3%))	14	46	16	32	9	2	9	3	0	17	15	4	7	5	0	2	10	0	0	0
	B.2 (146(3%))	10	61	26	11	2	0	3	9	3	23	10	1	10	2	0	1	9	0	0	0
	B.14 (138(2%))	13	54	15	16	5	0	2	7	5	22	11	1	5	4	0	2	9	0	0	0
	A (103(2%))	9	36	6	29	8	2	8	0	0	11	14	3	1	5	0	1	8	0	0	0
	B.7 (96(2%))	7	46	10	4	1	0	3	7	3	16	9	1	0	2	0	1	10	0	0	0
	B.13 (11(0.2 %))	0	5	0	1	0	0	1	1	0	0	0	0	1	0	0	0	3	0	0	0
	A & B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	B.16 (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

1	NOT_ANNOTATED
2	hydrolase activity
3	transferase activity
4	structural molecule activity
5	DNA binding
6	transcription regulator activity
7	RNA binding
8	oxidoreductase activity
9	enzyme regulator activity
10	ligase activity
11	peptidase activity
12	helicase activity
13	lyase activity
14	nucleotidyltransferase activity
15	lipid binding
16	isomerase activity
17	translation regulator activity
18	phosphoprotein phosphatase activity
19	signal transducer activity
20	motor activity

Supplementary Table I (D). Summary of the comparison between *S. cerevisiae* sequences and those of organisms from different groups for domains, kingdoms or phyla, classified by molecular function from the GOSLIM ontology. Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have orthologs in all organisms of the corresponding group. Columns numbered 1-20 represent each a molecular function. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red color scale indicates the similarity between the set of proteins associated to the molecular function in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of orthologs to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

E	EUKARYOTA										A	ARCHEA	B	BACTERIA					
E.1	FUNGI	E.2	ANIMAL	E.3	PLANTS	E.4	PROTISTS				A.1	Crenarchaeota	B.1	Acidobacteria	B.6	Deinococcus-Thermus	B.11	Hyperthermophilic bacteria
E.1.1	Ascomycetes	E.2.1	Echinoderms	E.3.1	Dicotyledons	E.4.1	Alveolates	E.4.5	Entamoeba		A.2	Euryarchaeota	B.2	Actinobacteria	B.7	Firmicutes	B.12	Planctomyces
E.1.2	Basidiomycetes	E.2.2	Insects	E.3.2	Green algae	E.4.2	Cellular slime molds	E.4.6	Euglenozoa		A.3	Korarchaeota	B.3	Bacteroides	B.8	Fusobacteria	B.13	Proteobacteria
E.1.3	Microsporidians	E.2.3	Nematodes	E.3.3	Monocotyledons	E.4.3	Choanoflagellates	E.4.7			A.4	Nanoarchaeota	B.4	Chlamydia	B.9	Green nonsulfur bacteria	B.14	Spirochete
		E.2.4	Vertebrates	E.3.4	Red algae	E.4.4	Diplomonads		Paraba salids				B.5	Cyanobacteria	B.10	Green sulfur bacteria	B.15	Synergistetes
																	B.16	Tenericutes

Table S1.2. E ***S. cerevisiae*** **Homology analysis with Molecular Function(s) of GO**

Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have homologues in all organisms of the corresponding group. Columns numbered 1-20 represent each a molecular function. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red colour scale indicates the similarity between the set of proteins associated to the molecular function in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of homologues to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

(See the Table S1.2 E. on next page)

Supplementary Table S1.2. E - *S. cerevisiae* Homology analysis with Molecular Function(s) of GO. [continued…]

Supplementary Table - S1.2 (E) *S. cerevisiae* HOMOLOGY ANALYSIS WITH MOLECULAR FUNCTION(S) OF GO.

MOLECULAR FUNCTION TERMS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
S. cerevisiae genome size (5880)	2782 (47%)	764 (13%)	629 (11%)	346 (6%)	328 (6%)	319 (5%)	292 (5%)	269 (5%)	208 (4%)	166 (3%)	108 (2%)	83 (1%)	83 (1%)	73 (1%)	70 (1%)	56 (1%)	52 (1%)	47 (1%)	44 (1%)	16 (0.3%)
E.4.3 (1162(20%))	515	151	130	39	76	78	53	66	48	28	26	21	22	15	33	7	6	11	7	2
E.4.2 (1156(20%))	522	141	108	47	74	77	57	66	52	34	27	18	18	15	22	5	7	11	5	0
E.3.1 (1147(20%))	547	140	113	45	72	79	51	51	44	29	23	22	11	13	24	5	4	13	2	0
E.3.3 (1091(19%))	496	137	112	49	73	79	54	52	41	33	21	22	14	10	20	3	4	10	3	0
E.2.1 (1075(18%))	533	93	109	47	76	77	47	42	48	30	13	3	20	14	24	6	4	3	5	1
E.4.7 (1054(20%))	393	171	95	54	53	58	76	62	64	44	31	23	13	15	22	11	13	10	6	1
E.3.4 (1051(18%))	436	135	143	38	69	78	57	64	34	30	18	20	22	16	15	8	6	11	3	1
E.4.5 (945(16%))	357	141	112	33	61	68	58	50	50	38	29	8	14	12	28	9	12	14	4	1
E.4.4 (752(13%))	266	142	103	41	27	22	48	34	40	35	24	22	12	18	9	9	15	8	6	2
E.1.3 (749(13%))	240	126	86	48	63	63	66	30	40	26	25	19	5	11	15	8	8	5	6	1
B.1 (740(13%))	226	157	111	38	17	13	29	91	17	28	26	31	29	7	2	18	7	12	4	0
B.12 (712(12%))	236	134	116	33	22	12	29	74	17	27	15	28	29	10	1	17	9	12	8	0
E.4.6 (695(12%))	281	102	93	31	19	19	44	39	23	26	16	21	10	13	18	10	9	9	5	0
E.2.3 (693(12%))	298	88	73	29	45	56	45	22	33	18	19	19	11	6	20	3	5	2	3	0
E.2.2 (638(11%))	297	60	71	24	41	48	34	35	26	18	16	2	14	6	13	4	1	1	1	1
A.3 (638(11%))	152	154	151	51	24	12	31	51	7	19	16	35	22	17	3	8	9	7	11	0
E.3.2 (636(11%))	301	90	74	26	29	31	35	27	16	17	17	19	6	2	9	2	2	4	4	0
B.8 (491(8%))	112	140	75	44	19	11	13	47	10	26	13	32	19	7	2	16	6	9	3	1
E.1.2 (469(8%))	207	46	36	20	64	58	25	17	23	16	9	3	1	6	6	1	3	1	1	1
B.15 (343(6%))	80	63	86	31	10	6	7	41	6	19	8	8	12	4	0	9	6	1	6	0
B.6 (326(6%))	72	58	91	21	15	3	6	43	5	13	4	9	13	4	1	13	4	1	7	0
A.4 (310(5%))	41	119	46	56	9	7	23	15	5	16	13	35	4	13	2	4	11	0	2	0
B.10 (260(4%))	66	47	41	21	9	4	19	34	4	17	2	11	12	4	0	9	2	0	1	0
E.3 (253(4%))	99	43	29	16	14	12	15	11	7	9	3	17	3	0	5	1	0	3	0	0
B.11 (251(4%))	58	59	34	30	10	3	9	31	7	19	9	6	3	4	1	9	6	7	1	0
E.4.1 (194(3%))	94	30	16	7	7	11	13	5	5	8	6	4	0	0	5	4	2	2	2	0
B.4 (190(3%))	52	29	51	22	9	4	5	17	1	9	3	3	4	2	0	8	2	1	4	0
B.9 (156(3%))	48	37	15	18	5	1	0	14	4	6	6	4	10	2	0	9	2	0	1	0
E.2.4 (125(2%))	60	11	6	1	22	22	8	9	1	2	5	2	3	0	2	0	0	0	0	0
B.3 (104(2%))	25	24	13	17	7	1	3	19	1	4	1	2	3	1	0	2	0	0	0	0
B.5 (89(2%))	16	20	13	11	7	1	9	10	1	6	3	4	1	0	0	4	2	0	0	0
A.1 (82(1%))	13	29	20	11	4	2	9	2	4	3	4	4	0	3	2	1	1	0	0	0
E.2 (81(1%))	40	5	4	1	17	15	3	4	0	2	2	2	2	0	2	0	0	0	0	0
B.14 (64(1%))	20	9	7	10	4	1	0	6	1	6	0	2	1	1	0	2	2	0	0	0
E.4 (43(0.7%))	25	5	3	1	2	3	2	0	1	3	1	1	0	0	1	1	0	1	0	0
B.2 (34(0.6%))	14	6	10	4	3	0	1	0	1	1	0	1	0	1	0	0	0	0	0	0
B.16 (33(0.6%))	3	13	0	9	1	0	1	0	1	3	0	0	1	0	0	1	2	0	0	0
A.2 (30(0.5%))	5	7	5	6	1	1	1	1	3	3	1	0	0	0	1	0	1	0	0	0
A (18(0.3%))	4	5	1	4	1	1	1	0	2	2	0	0	0	0	1	0	0	0	0	0
B.7 (8(0.1%))	2	4	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
B.13 (1(0.01 %))	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.1 (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.1.1 (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A & B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

HOMOLOGS

1	NOT_ANNOTATED
2	hydrolase activity
3	transferase activity
4	structural molecule activity
5	DNA binding
6	transcription regulator activity
7	RNA binding
8	oxidoreductase activity
9	enzyme regulator activity
10	ligase activity
11	peptidase activity
12	helicase activity
13	lyase activity
14	nucleotidyltransferase activity
15	lipid binding
16	isomerase activity
17	translation regulator activity
18	phosphoprotein phosphatase activity
19	signal transducer activity
20	motor activity

Supplementary Table I (E). Summary of the comparison between *S. cerevisiae* sequences and those of organisms from different groups for domains, kingdoms or phyla, classified by molecular function from the GOSLIM ontology. Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have homologues in all organisms of the corresponding group. Columns numbered 1-20 represent each a molecular function. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red color scale indicates the similarity between the set of proteins associated to the molecular function in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of homologues to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

E	EUKARYOTA								
E.1	FUNGI	E.2	ANIMAL	E.3	PLANTS	E.4	PROTISTS		
E.1.1	Ascomycetes	E.2.1	Echinoderms	E.3.1	Dicotyledons	E.4.1	Alveolates	E.4.5	Entamoeba
E.1.2	Basidiomycetes	E.2.2	Insects	E.3.2	Green algae	E.4.2	Cellular slime molds	E.4.6	Euglenozoa
E.1.3	Microsporidians	E.2.3	Nematodes	E.3.3	Monocotyledons	E.4.3	Choanoflagellates	E.4.7	
		E.2.4	Vertebrates	E.3.4	Red algae	E.4.4	Diplomonads		Paraba salids

A	ARCHEA
A.1	Crenarchaeota
A.2	Euryarchaeota
A.3	Korarchaeota
A.4	Nanoarchaeota

B	BACTERIA				
B.1	Acidobacteria	B.6	Deinococcus-Thermus	B.11	Hyperthermophilic bacteria
B.2	Actinobacteria	B.7	Firmicutes	B.12	Planctomyces
B.3	Bacteroides	B.8	Fusobacteria	B.13	Proteobacteria
B.4	Chlamydia	B.9	Green nonsulfur bacteria	B.14	Spirochete
B.5	Cyanobacteria	B.10	Green sulfur bacteria	B.15	Synergistetes
				B.16	Tenericutes

Table S1.2. F ***S. cerevisiae*** **Absent genes analysis with Molecular Function(s) of GO**

Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that are absent in all organisms of the corresponding group. Columns numbered 1-20 represent each a molecular function. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red colour scale indicates the similarity between the set of proteins associated to the molecular function in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity (low absent), while red entries indicate low similarity (highly absence) between the two sets of proteins. The numbers in each entry indicate how many *S. cerevisiae* proteins are commonly absent from the genome of all organisms in the group represented in the row.

(See the Table S1.2 F. on next page)

Supplementary Table S1.2. F - *S. cerevisiae* Homology analysis with Molecular Function(s) of GO. [continued…]

Supplementary Table - S1.2 (F) *S. cerevisiae* ABSENT GENES ANALYSIS WITH MOLECULAR FUNCTION(S) OF GO.

MOLECULAR FUNCTION TERMS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
S. cerevisiae genome size (5880)	2782 (47%)	764 (13%)	629 (11%)	346 (6%)	328 (6%)	319 (5%)	292 (5%)	269 (5%)	208 (4%)	166 (3%)	108 (2%)	83 (1%)	83 (1%)	73 (1%)	70 (1%)	56 (1%)	52 (1%)	47 (1%)	44 (1%)	16 (0.3%)
E (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.1 (10(0.2%))	8	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.1.1 (10(0.2%))	8	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.3 (142(2%))	106	4	2	8	11	9	1	1	4	1	0	1	0	0	1	0	0	0	3	0
E.3.1 (282(5%))	212	11	4	19	15	14	3	1	6	1	1	1	0	0	1	0	1	0	6	0
E.1.2 (1525(26%))	1167	50	37	63	75	83	27	3	43	10	2	5	1	5	7	0	7	0	9	3
E.4 (1978(34%))	1422	63	78	96	126	138	26	18	58	12	2	3	4	8	13	0	7	1	15	2
E.2 (2028(34%))	1426	97	93	88	118	129	25	30	59	13	6	4	14	4	13	4	7	4	15	3
E.2.4 (2185(37%))	1507	115	110	96	125	139	26	40	68	15	6	4	18	5	13	5	8	5	15	3
E.2.2 (2371(40%))	1598	129	140	109	135	140	33	49	68	15	10	7	22	11	17	6	10	5	16	3
E.3.2 (2519(43%))	1671	137	119	112	187	197	39	35	103	18	12	4	8	13	27	2	10	3	17	4
E.3.3 (2527(43%))	1679	152	133	120	164	174	40	38	92	19	17	3	9	16	27	3	8	3	18	5
E.2.3 (2540(43%))	1683	152	149	119	146	156	35	50	80	22	15	5	21	15	17	7	9	8	18	4
E.3.4 (2554(43%))	1643	191	145	101	154	161	52	53	96	24	29	5	13	16	39	3	8	7	15	9
E.2.1 (2582(44%))	1691	157	156	126	141	155	45	49	81	26	16	20	21	10	18	5	10	5	16	3
E.4.2 (2603(44%))	1728	148	154	126	159	169	45	48	76	22	11	8	18	12	23	7	10	2	17	4
A & B (2642(45%))	1673	132	118	116	206	220	75	20	127	58	21	2	2	18	39	3	11	7	17	10
E.4.1 (2740(47%))	1789	153	177	120	170	197	43	50	94	24	12	5	14	16	21	6	10	9	15	3
E.4.3 (2755(47%))	1807	167	164	132	154	184	49	54	91	28	11	5	21	19	19	2	12	6	17	2
B (2881(49%))	1737	152	139	186	220	240	104	20	139	58	23	3	3	34	45	3	17	7	17	10
E.4.6 (3157(54%))	1938	226	216	148	225	261	63	62	113	36	28	5	28	23	32	8	10	9	21	4
B.13 (3256(55 %))	1925	202	176	214	236	256	127	27	149	67	40	5	3	39	50	4	21	12	18	13
E.4.7 (3616(62%))	2153	254	303	153	215	238	71	161	106	60	31	5	50	27	38	14	18	9	17	8
A (3672(62%))	2179	266	242	178	251	284	108	70	160	76	46	9	11	30	55	8	19	24	18	14
E.4.5 (3678(63%))	2178	277	309	161	196	212	67	181	112	64	38	22	48	30	30	14	14	10	21	4
B.7 (3692(63%))	2119	272	226	236	262	284	139	47	167	71	58	7	6	46	60	5	24	22	23	11
A.2 (3807(65%))	2227	286	256	194	261	278	112	79	164	76	51	13	12	32	59	8	20	25	18	14
B.2 (3889(66%))	2209	300	258	237	276	286	171	58	176	69	46	7	10	49	59	5	29	31	23	16
B.5 (3940(67%))	2257	316	267	245	271	288	144	52	171	75	62	8	12	48	60	12	24	23	18	15
B.3 (4076(69%))	2309	344	296	252	275	286	155	59	174	77	63	9	9	53	63	9	28	29	22	15
E.1.3 (4097(70%))	2379	328	354	176	188	202	86	197	142	74	47	9	65	30	49	24	24	24	19	4
B.9 (4178(71%))	2363	348	302	254	277	292	163	70	176	78	59	14	14	51	65	14	29	30	20	16
E.4.4 (4195(71%))	2367	345	349	177	251	284	106	203	138	78	54	9	59	32	58	20	19	12	22	8
A.1 (4209(72%))	2382	354	321	193	271	284	145	98	184	89	55	14	19	35	63	29	23	29	24	14
B.14 (4363(74%))	2464	414	316	280	276	291	158	74	182	82	70	26	15	51	63	12	30	42	28	16
B.10 (4393(75%))	2440	404	365	254	279	295	152	86	176	80	77	13	20	53	65	17	27	33	28	16
B.6 (4408(75%))	2441	406	337	257	281	289	174	88	182	86	71	27	16	58	65	11	31	33	25	16
B.1 (4413(75%))	2426	424	360	268	287	297	161	86	182	87	61	18	18	57	67	13	31	35	25	16
B.12 (4517(77%))	2466	459	367	276	283	298	166	102	180	82	78	16	19	54	68	20	29	36	25	16
B.15 (4677(80%))	2529	477	360	281	286	304	188	133	186	89	75	38	24	56	68	18	32	43	28	16
B.11 (4685(80%))	2519	433	444	289	284	302	180	125	186	95	74	17	26	61	65	28	31	34	37	15
B.4 (4712(80%))	2546	471	404	284	286	303	186	135	186	89	67	22	45	60	67	17	31	40	28	16
B.16 (4727(80%))	2562	438	397	251	291	307	183	127	189	105	80	29	43	58	64	22	34	39	28	15
A.3 (4822(82%))	2583	487	426	221	285	304	157	158	199	113	73	38	42	47	67	44	28	32	33	16
B.8 (4951(84%))	2618	478	478	257	291	305	188	178	192	96	82	21	37	60	68	24	33	36	41	15
A.4 (5350(91%))	2720	578	566	224	304	310	170	249	201	126	78	43	78	51	67	51	29	47	42	16

ABSENT GENES

1	NOT_ANNOTATED
2	hydrolase activity
3	transferase activity
4	structural molecule activity
5	DNA binding
6	transcription regulator activity
7	RNA binding
8	oxidoreductase activity
9	enzyme regulator activity
10	ligase activity
11	peptidase activity
12	helicase activity
13	lyase activity
14	nucleotidyltransferase activity
15	lipid binding
16	isomerase activity
17	translation regulator activity
18	phosphoprotein phosphatase activity
19	signal transducer activity
20	motor activity

Supplementary Table I (F). Summary of the comparison between *S. cerevisiae* sequences and those of organisms from different groups for domains, kingdoms or phyla, classified by molecular function from the GOSLIM ontology. Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that are absent in all organisms of the corresponding group. Columns numbered 1-20 represent each a molecular function. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red color scale indicates the similarity between the set of proteins associated to the molecular function in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity (low absent), while red entries indicate low similarity (highly absence) between the two sets of proteins. The numbers in each entry indicate how many *S. cerevisiae* proteins are commonly absent from the genome of all organisms in the group represented in the row.

E	EUKARYOTA								
E.1	FUNGI	E.2	ANIMAL	E.3	PLANTS	E.4	PROTISTS		
E.1.1	Ascomycetes	E.2.1	Echinoderms	E.3.1	Dicotyledons	E.4.1	Alveolates	E.4.5	Entamoeba
E.1.2	Basidiomycetes	E.2.2	Insects	E.3.2	Green algae	E.4.2	Cellular slime molds	E.4.6	Euglenozoa
E.1.3	Microsporidians	E.2.3	Nematodes	E.3.3	Monocotyledons	E.4.3	Choanoflagellates	E.4.7	Paraba salids
		E.2.4	Vertebrates	E.3.4	Red algae	E.4.4	Diplomonads		

A	ARCHEA
A.1	Crenarchaeota
A.2	Euryarchaeota
A.3	Korarchaeota
A.4	Nanoarchaeota

B	BACTERIA				
B.1	Acidobacteria	B.6	Deinococcus-Thermus	B.11	Hyperthermophilic bacteria
B.2	Actinobacteria	B.7	Firmicutes	B.12	Planctomyces
B.3	Bacteroides	B.8	Fusobacteria	B.13	Proteobacteria
B.4	Chlamydia	B.9	Green nonsulfur bacteria	B.14	Spirochete
B.5	Cyanobacteria	B.10	Green sulfur bacteria	B.15	Synergistetes
				B.16	Tenericutes

Table S1.2. G ***S. cerevisiae*** **Orthology analysis with Cellular component(s) of GO**

Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have orthologs in all organisms of the corresponding group. Columns numbered 1-21 represent each a cellular localization. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red colour scale indicates the similarity between the set of proteins associated to the cellular component in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of orthologs to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

(See the Table S1.2 G. on next page)

Supplementary Table S1.2. G - *S. cerevisiae* Orthology analysis with Cellular Component(s) of GO. [continued…]

Supplementary Table - S1.2 (G) *S. cerevisiae* ORTHOLOGY ANALYSIS WITH CELLULAR COMPONENT(S) OF GO.

	CELLULAR COMPONENT TERMS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	S. cerevisiae genome size (5880)	3108 (53%)	1864 (32%)	1098 (19%)	1049 (18%)	948 (16%)	363 (6%)	353 (6%)	321 (6%)	318 (5%)	303 (5%)	301 (5%)	207 (4%)	204 (4%)	198 (3%)	191 (3%)	170 (3%)	114 (2%)	104 (2%)	98 (2%)	71 (1%)	20 (0.3%)
ORTHOLOGS	E.1.2 (2310(39%))	1339	673	490	224	408	135	192	105	110	141	110	103	92	114	91	90	56	56	21	28	2
	E.3.1 (2269(39%))	1342	715	505	198	355	131	208	110	122	131	122	75	70	108	83	66	32	48	9	24	4
	E.3.3 (2244(38%))	1311	692	486	229	340	127	205	107	117	124	125	77	72	107	81	59	35	47	13	24	6
	E.2.1 (2204(38%))	1322	736	470	156	363	134	190	121	122	135	130	85	82	97	80	67	44	47	10	26	3
	E.4.2 (2174(37%))	1300	700	482	162	357	135	195	114	120	129	121	78	78	101	76	66	37	45	7	26	1
	E.1.1 (2107(36%))	1218	646	456	186	372	145	137	104	113	132	107	84	89	101	94	83	54	51	22	28	5
	E.4.3 (2007(34%))	1216	627	448	150	296	117	187	103	99	127	125	60	86	86	77	66	46	47	8	26	1
	E.3.4 (1766(30%))	1090	553	432	118	251	97	201	96	83	121	99	46	53	85	57	42	18	36	5	26	1
	E.2.2 (1728(29%))	1075	552	378	106	257	103	162	101	96	100	97	57	79	77	60	58	42	36	5	25	3
	E.3.2 (1644(30%))	1004	522	373	130	227	78	181	89	82	102	107	50	67	79	55	45	23	39	5	22	0
	E.2.3 (1592(27%))	960	505	363	104	253	91	171	94	97	95	94	61	75	79	69	58	38	41	10	23	2
	E.3 (1371(23%))	858	427	318	90	187	58	175	75	66	84	88	39	45	71	48	38	15	33	5	21	0
	E.4.5 (1300(22%))	776	458	211	84	173	63	172	81	74	36	98	49	68	73	49	53	32	32	5	21	1
	E.4.6 (1281(22%))	798	386	284	79	194	74	156	72	68	69	88	50	54	72	45	43	21	32	5	23	1
	E.4.7 (1257(21%))	743	436	212	93	161	61	148	77	67	38	90	42	86	72	48	41	21	27	5	26	2
	E.1.3 (1015(17%))	575	391	164	67	112	50	122	76	46	29	72	25	51	56	43	42	23	23	7	20	1
	E.4.4 (959(16%))	590	326	158	67	110	41	137	66	48	29	84	33	48	52	31	41	17	19	5	24	0
	E.4.1 (839(14%))	503	289	148	56	102	29	121	61	41	35	62	28	41	44	35	34	14	23	5	19	0
	E.1 (781(13%))	430	292	126	59	92	41	77	57	33	27	55	22	45	46	36	40	18	15	7	18	0
	B.1 (708(12%))	403	167	220	94	74	26	48	19	18	31	33	22	15	38	5	20	8	2	7	5	1
	B.12 (632(11%))	373	148	208	82	68	19	45	16	16	30	24	18	13	31	4	19	6	1	5	5	1
	E.4 (591(10%))	352	192	98	43	72	18	92	47	29	24	54	18	32	37	26	30	9	17	5	16	0
	E.2.4 (573(10%))	334	183	124	51	94	27	41	31	28	39	27	22	33	42	22	34	11	10	2	17	1
	E.2 (480(8%))	278	162	98	42	74	22	30	27	26	30	25	16	28	33	22	32	10	10	0	14	1
	B.8 (419(7%))	265	91	154	36	52	16	40	9	12	24	21	14	0	29	3	1	1	1	5	0	0
	A.3 (401(7%))	266	83	100	32	45	12	77	24	10	19	19	7	3	19	7	3	1	4	3	2	0
	B.6 (401(7%))	259	58	159	41	47	12	38	2	10	22	14	11	2	26	2	6	2	1	6	0	0
	B.10 (341(6%))	227	62	133	29	41	16	34	1	11	23	15	8	1	29	1	1	1	0	5	0	0
	B.15 (319(5%))	206	64	124	33	42	8	31	5	5	20	4	10	3	27	2	6	2	1	5	1	0
	B.9 (279(5%))	183	48	103	27	36	11	32	3	9	17	11	9	0	19	1	1	1	1	5	0	0
	B.11 (275(5%))	181	49	113	29	35	8	39	2	6	17	4	11	0	21	2	0	1	1	5	0	0
	E (263(5%))	136	68	42	32	38	9	12	15	11	18	19	7	19	20	15	23	6	8	0	11	0
	B.5 (263(5%))	179	39	111	23	36	8	19	2	5	21	2	8	0	21	1	0	1	1	5	0	0
	B.3 (236(4%))	163	46	87	11	29	8	32	3	7	14	15	8	1	21	1	1	1	0	5	0	0
	B.4 (233(4%))	156	50	115	12	31	7	32	5	7	19	4	9	0	19	3	1	1	1	5	0	0
	B.4 (201(3%))	137	49	43	7	18	5	66	13	2	9	15	3	0	4	0	0	1	0	0	0	0
	A.1 (164(3%))	119	38	31	9	17	5	45	13	2	8	9	3	0	5	0	0	1	0	0	0	0
	A.2 (148(3%))	111	30	24	9	16	5	41	10	2	7	7	4	0	6	0	0	1	0	0	0	0
	B.2 (146(3%))	103	16	61	14	25	7	19	0	4	12	1	5	0	16	1	1	1	1	5	0	0
	B.14 (138(2%))	93	24	72	7	20	7	25	2	3	11	1	4	0	16	0	0	1	0	5	0	0
	A (103(2%))	78	24	16	4	12	4	38	9	2	5	6	2	0	4	0	0	1	0	0	0	0
	B.7 (96(2%))	63	14	40	8	22	6	12	0	4	11	1	3	0	8	1	0	1	1	2	0	0
	B.13 (11(0.2 %))	7	0	7	0	4	0	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
	A & B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	B.16 (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

1	cytoplasm
2	nucleus
3	mitochondrion
4	NOT_ANNOTATED
5	membrane
6	endoplasmic reticulum
7	ribosome
8	chromosome
9	endomembrane system
10	mitochondrial envelope
11	nucleolus
12	vacuole
13	cytoskeleton
14	membrane fraction
15	Golgi apparatus
16	cellular bud
17	cell cortex
18	cytoplasmic membrane-bounded vesicle
19	cell wall
20	microtubule organizing center
21	extracellular region

Supplementary Table I (G). Summary of the comparison between *S. cerevisiae* sequences and those of organisms from different group for domains, kingdoms or phyla, classified by cellular localization from the GOSLIM ontology. Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have orthologs in all organisms of the corresponding group. Columns numbered 1-21 represent each a cellular localization. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red color scale indicates the similarity between the set of proteins associated to the cellular component in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of orthologs to *S. cerevisiae* proteins that are common to all organisms from the group represented in the row.

E	EUKARYOTA								
E.1	FUNGI	E.2	ANIMAL	E.3	PLANTS	E.4	PROTISTS		
E.1.1	Ascomycetes	E.2.1	Echinoderms	E.3.1	Dicotyledons	E.4.1	Alveolates	E.4.5	Entamoeba
E.1.2	Basidiomycetes	E.2.2	Insects	E.3.2	Green algae	E.4.2	Cellular slime molds	E.4.6	Euglenozoa
E.1.3	Microsporidians	E.2.3	Nematodes	E.3.3	Monocotyledons	E.4.3	Choanoflagellates	E.4.7	Paraba salids
		E.2.4	Vertebrates	E.3.4	Red algae	E.4.4	Diplomonads		

A	ARCHEA
A.1	Crenarchaeota
A.2	Euryarchaeota
A.3	Korarchaeota
A.4	Nanoarchaeota

B	BACTERIA				
B.1	Acidobacteria	B.6	Deinococcus-Thermus	B.11	Hyperthermophilic bacteria
B.2	Actinobacteria	B.7	Firmicutes	B.12	Planctomyces
B.3	Bacteroides	B.8	Fusobacteria	B.13	Proteobacteria
B.4	Chlamydia	B.9	Green nonsulfur bacteria	B.14	Spirochete
B.5	Cyanobacteria	B.10	Green sulfur bacteria	B.15	Synergistetes
				B.16	Tenericutes

Table S1.2. H *S. cerevisiae* Homology analysis with Cellular component(s) of GO

Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have homologues in all organisms of the corresponding group. Columns numbered 1-21 represent each a cellular localization. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red colour scale indicates the similarity between the set of proteins associated to the cellular component in S. cerevisiae and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of *S. cerevisiae* proteins that have homologues in all organisms from the group represented in the row.

(See the Table S1.2 H. on next page)

Supplementary Table S1.2. H - *S. cerevisiae* Homology analysis with Cellular Component(s) of GO. [continued…]

Supplementary Table - S1.2 (H) *S. cerevisiae* HOMOLOGY ANALYSIS WITH CELLULAR COMPONENT(S) OF GO.

CELLULAR COMPONENT TERMS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
S. cerevisiae genome size (5880)	3108 (53%)	1864 (32%)	1098 (19%)	1049 (18%)	948 (16%)	363 (6%)	353 (6%)	321 (6%)	318 (5%)	303 (5%)	301 (5%)	207 (4%)	204 (4%)	198 (3%)	191 (3%)	170 (3%)	114 (2%)	104 (2%)	98 (2%)	71 (1%)	20 (0.3%)
E.4.3 (1162(20%))	582	411	193	159	207	70	41	74	69	54	40	56	37	46	48	37	31	20	10	11	2
E.4.2 (1156(20%))	571	393	199	171	202	82	46	56	76	61	41	51	36	26	45	38	32	21	6	9	3
E.3.1 (1147(20%))	548	374	188	207	200	79	45	65	60	57	48	55	31	26	40	27	28	19	7	9	6
E.3.3 (1091(19%))	541	371	183	160	198	82	45	64	66	65	37	49	31	22	38	22	26	16	6	10	2
E.2.1 (1075(18%))	503	351	182	184	179	72	41	57	67	48	38	45	40	31	39	33	30	24	4	12	2
E.4.7 (1054(20%))	558	391	187	121	191	63	52	62	72	54	46	44	37	31	44	34	22	31	5	11	0
E.3.4 (1051(18%))	535	390	200	135	187	67	42	55	64	58	48	42	26	27	47	27	20	25	0	6	0
E.4.5 (945(16%))	508	359	164	102	169	69	30	54	62	46	45	36	31	19	42	36	26	27	4	8	1
E.4.4 (752(13%))	409	283	144	73	135	44	41	46	55	28	42	32	26	31	29	15	16	19	0	12	0
E.1.3 (749(13%))	348	308	126	71	126	54	47	53	47	24	48	40	23	27	23	25	16	13	2	9	0
B.1 (740(13%))	409	207	180	105	118	49	44	27	28	43	20	33	7	25	13	8	3	7	2	2	2
B.12 (712(12%))	370	225	174	86	99	35	36	33	22	32	32	22	12	36	12	9	7	6	0	4	1
E.4.6 (695(12%))	375	237	140	86	127	47	32	37	38	35	32	33	22	23	30	16	14	14	1	7	0
E.2.3 (693(12%))	334	246	115	96	133	48	23	40	39	33	23	37	20	21	20	24	21	16	2	3	0
E.2.2 (638(11%))	314	211	97	92	124	51	17	31	38	28	20	36	25	19	20	21	20	13	2	5	0
A.3 (638(11%))	355	203	136	79	79	25	64	27	17	28	40	25	18	28	10	21	7	6	0	6	0
E.3.2 (636(11%))	283	195	103	139	100	40	25	31	32	30	20	23	12	24	32	8	9	12	0	4	0
B.8 (491(8%))	284	137	129	63	63	31	52	24	13	26	11	17	3	13	8	3	0	2	1	2	0
E.1.2 (469(8%))	202	182	81	70	83	27	17	34	31	24	18	8	17	7	15	15	17	10	7	6	1
B.15 (343(6%))	208	94	74	36	34	15	39	17	6	10	12	9	8	8	5	12	4	2	1	3	0
B.6 (326(6%))	194	91	81	33	31	14	26	17	6	6	9	10	7	10	3	12	3	2	0	3	0
A.4 (310(5%))	175	102	75	14	35	15	71	25	9	12	26	7	3	7	1	5	0	0	1	2	0
B.10 (260(4%))	149	78	72	29	40	16	21	12	9	16	16	6	2	9	4	1	1	2	0	0	0
E.3 (253(4%))	125	87	45	36	40	14	14	13	11	11	7	9	8	6	12	6	6	6	0	3	0
B.11 (251(4%))	156	58	69	26	30	10	35	12	5	13	9	5	3	12	1	2	0	1	0	1	0
E.4.1 (194(3%))	98	69	41	22	40	12	6	15	8	19	11	6	6	5	4	8	4	3	1	1	0
B.4 (190(3%))	111	46	54	16	17	10	25	8	3	8	6	3	6	11	1	5	3	0	0	3	0
B.9 (156(3%))	103	30	51	13	28	10	20	7	1	10	3	6	1	5	0	2	0	0	0	0	0
E.2.4 (125(2%))	54	47	18	24	20	4	1	9	2	4	1	5	3	2	1	3	4	0	3	0	0
B.3 (104(2%))	60	32	30	8	9	7	20	7	2	1	4	1	0	1	0	0	0	0	0	0	0
B.5 (89(2%))	55	23	27	3	11	6	13	8	2	7	5	0	0	0	0	0	0	0	0	0	0
A.1 (82(1%))	43	28	23	8	8	4	12	6	3	3	6	2	0	0	0	0	0	0	0	0	0
E.2 (81(1%))	32	30	15	16	17	4	1	8	2	4	1	2	3	1	0	2	4	0	1	0	0
B.14 (64(1%))	37	17	19	3	4	2	12	5	0	1	6	0	0	1	0	0	0	0	0	0	0
E.4 (43(0.7%))	20	13	6	8	8	4	1	3	2	1	3	2	1	0	2	3	1	2	0	0	0
B.2 (34(0.6%))	19	10	4	4	2	2	4	3	0	0	3	1	1	2	0	1	0	0	0	1	0
B.16 (33(0.6%))	21	3	14	1	5	0	10	1	0	2	2	0	0	1	0	0	0	0	0	0	0
A.2 (30(0.5%))	16	5	10	2	5	3	6	3	2	0	1	1	0	0	0	0	0	0	0	0	0
A (18(0.3%))	10	5	7	0	4	1	4	3	1	0	0	1	0	0	0	0	0	0	0	0	0
B.7 (8(0.1%))	6	1	3	0	3	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
B.13 (1(0.01 %))	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
E (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.1 (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.1.1 (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A & B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

HOMOLOGS

1	cytoplasm
2	nucleus
3	mitochondrion
4	NOT_ANNOTATED
5	membrane
6	endoplasmic reticulum
7	ribosome
8	chromosome
9	endomembrane system
10	mitochondrial envelope
11	nucleolus
12	vacuole
13	cytoskeleton
14	membrane fraction
15	Golgi apparatus
16	cellular bud
17	cell cortex
18	cytoplasmic membrane-bounded vesicle
19	cell wall
20	microtubule organizing center
21	extracellular region

Supplementary Table I (H). Summary of the comparison between *S. cerevisiae* sequences and those of organisms from a group for domains, kingdoms or phyla, classified by cellular localization from the GOSLIM ontology. Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that have homologues in all organisms of the corresponding group. Columns numbered 1-21 represent each a cellular localization. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red color scale indicates the similarity between the set of proteins associated to the cellular component in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of *S. cerevisiae* proteins that have homologues in all organisms from the group represented in the row.

E	EUKARYOTA								
E.1	FUNGI	E.2	ANIMAL	E.3	PLANTS	E.4	PROTISTS		
E.1.1	Ascomycetes	E.2.1	Echinoderms	E.3.1	Dicotyledons	E.4.1	Alveolates	E.4.5	Entamoeba
E.1.2	Basidiomycetes	E.2.2	Insects	E.3.2	Green algae	E.4.2	Cellular slime molds	E.4.6	Euglenozoa
E.1.3	Microsporidians	E.2.3	Nematodes	E.3.3	Monocotyledons	E.4.3	Choanoflagellates	E.4.7	
		E.2.4	Vertebrates	E.3.4	Red algae	E.4.4	Diplomonads		Paraba salids

A	ARCHEA	B	BACTERIA				
A.1	Crenarchaeota	B.1	Acidobacteria	B.6	Deinococcus-Thermus	B.11	Hyperthermophilic bacteria
A.2	Euryarchaeota	B.2	Actinobacteria	B.7	Firmicutes	B.12	Planctomyces
A.3	Korarchaeota	B.3	Bacteroides	B.8	Fusobacteria	B.13	Proteobacteria
A.4	Nanoarchaeota	B.4	Chlamydia	B.9	Green nonsulfur bacteria	B.14	Spirochete
		B.5	Cyanobacteria	B.10	Green sulfur bacteria	B.15	Synergistetes
						B.16	Tenericutes

Table S1.2. I ***S. cerevisiae*** **Absent genes analysis with Cellular component(s) of GO**

Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that absent in all organisms of the corresponding group. Columns numbered 1-21 represent each a cellular localization. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red colour scale indicates the similarity between the set of proteins associated to the cellular component in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity (low absent), while red entries indicate low similarity (highly absence) between the two sets of proteins. The numbers in each entry indicate the number of *S. cerevisiae* proteins that are absent in all organisms from the group represented in the row.

(See the Table S1.2 I. on next page)

Supplementary Table S1.2. I - *S. cerevisiae* Absent genes analysis with Cellular Component(s) of GO. [continued…]

Supplementary Table - S1.2 (I) *S. cerevisiae* ABSENT GENES ANALYSIS WITH CELLULAR COMPONENT(S) OF GO.

CELLULAR COMPONENT TERMS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
S. cerevisiae genome size (5880)	3108 (53%)	1864 (32%)	1098 (19%)	1049 (18%)	948 (16%)	363 (6%)	353 (6%)	321 (6%)	318 (5%)	303 (5%)	301 (5%)	207 (4%)	204 (4%)	198 (3%)	191 (3%)	170 (3%)	114 (2%)	104 (2%)	98 (2%)	71 (1%)	20 (0.3%)
E (0(0%))	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E.1 (10(0.2%))	3	1	0	6	0	1	2	0	0	0	0	0	0	1	0	0	0	0	1	0	0
E.1.1 (10(0.2%))	3	1	0	6	0	1	2	0	0	0	0	0	0	1	0	0	0	0	1	0	0
E.3 (142(2%))	44	28	14	57	28	12	4	6	7	9	2	4	7	3	2	4	3	1	9	2	0
E.3.1 (282(5%))	74	52	24	115	37	20	11	8	12	9	8	7	10	5	2	9	4	1	32	3	3
E.1.2 (1525(26%))	488	415	170	512	181	79	47	107	71	47	39	37	53	34	31	36	17	16	61	27	10
E.4 (1978(34%))	681	545	274	586	303	107	70	123	97	85	44	50	72	57	46	58	35	23	73	31	13
E.2 (2028(34%))	698	518	283	623	303	110	67	110	90	86	36	55	68	57	53	52	30	23	77	29	12
E.2.4 (2185(37%))	779	555	315	655	329	118	75	117	102	92	39	64	72	62	56	56	31	25	78	31	13
E.2.2 (2371(40%))	868	625	357	676	353	133	86	131	104	107	56	70	80	67	63	66	35	26	82	35	15
E.3.2 (2519(43%))	963	719	366	654	393	152	87	152	140	117	59	80	103	64	66	90	58	38	82	36	14
E.3.3 (2527(43%))	962	712	387	661	404	153	86	150	134	124	62	81	101	69	72	89	53	41	79	37	12
E.2.3 (2540(43%))	964	681	393	693	393	152	96	143	121	123	65	75	82	72	68	64	35	31	85	37	16
E.3.4 (2554(43%))	1027	714	371	644	437	167	71	154	146	112	62	100	107	73	79	79	65	37	60	34	13
E.2.1 (2582(44%))	989	688	404	700	400	156	105	143	128	120	56	77	82	70	72	70	40	33	84	33	15
E.4.2 (2603(44%))	977	702	390	724	396	148	96	156	123	118	63	81	93	72	73	75	48	39	85	36	16
A & B (2642(45%))	1084	888	369	562	464	175	91	193	181	141	96	75	128	65	95	91	75	59	59	45	11
E.4.1 (2740(47%))	1051	764	418	701	454	175	92	156	144	126	63	87	92	71	88	80	50	43	83	36	16
E.4.3 (2755(47%))	1052	755	424	748	449	181	109	149	154	126	60	93	85	68	70	72	41	39	80	35	17
B (2881(49%))	1230	988	376	577	480	180	168	207	190	144	123	81	131	71	98	96	77	63	59	45	11
E.4.6 (3157(54%))	1238	931	490	782	491	182	121	179	160	155	80	102	111	79	89	99	66	42	91	38	18
B.13 (3256(55 %))	1417	1105	431	636	546	205	196	227	213	163	135	103	148	88	115	110	85	69	66	53	12
E.4.7 (3616(62%))	1542	973	566	845	605	242	135	184	182	213	92	124	113	95	102	97	71	48	88	36	18
A (3672(62%))	1579	1147	566	763	651	249	139	230	231	213	123	121	155	106	150	126	97	77	84	38	17
E.4.5 (3678(63%))	1571	978	691	886	612	234	137	190	183	222	83	128	108	107	102	85	58	46	89	43	18
B.7 (3692(63%))	1632	1232	522	705	622	240	212	250	237	198	156	118	155	110	146	123	90	77	74	55	16
A.2 (3807(65%))	1645	1187	589	786	677	260	151	236	243	218	124	123	173	111	154	133	99	83	85	60	18
B.2 (3889(66%))	1733	1284	552	751	659	250	218	257	251	203	153	129	159	111	155	127	96	79	76	60	15
B.5 (3940(67%))	1730	1253	562	784	680	257	221	253	244	200	156	135	159	106	152	125	96	88	87	58	17
B.3 (4076(69%))	1826	1315	593	785	702	268	227	259	262	215	163	137	175	117	160	130	99	91	81	63	17
E.1.3 (4097(70%))	1891	1076	766	911	704	258	167	192	224	250	103	142	130	115	125	103	75	68	89	42	19
B.9 (4178(71%))	1857	1324	599	833	726	281	228	259	267	214	162	153	180	119	160	136	104	87	83	63	17
E.4.4 (4195(71%))	1841	1182	766	913	704	279	159	214	215	248	100	143	134	118	131	118	84	66	93	36	20
A.1 (4209(72%))	1876	1330	683	823	735	291	161	256	265	238	156	151	176	134	163	134	102	87	93	59	18
B.14 (4363(74%))	1937	1385	619	896	750	279	229	269	273	221	176	161	179	136	171	139	102	94	91	61	19
B.10 (4393(75%))	1975	1363	637	877	752	288	229	263	267	224	165	152	185	136	166	146	106	92	90	65	19
B.6 (4408(75%))	1974	1388	649	896	760	302	235	269	276	231	171	156	182	133	169	142	104	90	87	64	19
B.1 (4413(75%))	2002	1401	656	890	760	287	244	275	271	229	171	152	182	135	173	142	103	95	89	64	17
B.12 (4517(77%))	2071	1402	674	881	777	308	255	272	280	241	168	167	179	131	175	142	101	97	93	62	18
B.15 (4677(80%))	2114	1480	701	904	804	310	238	276	288	248	189	171	185	145	178	144	107	97	88	64	18
B.11 (4685(80%))	2133	1480	716	895	806	314	236	276	279	251	178	169	195	140	171	163	110	95	87	69	18
B.4 (4712(80%))	2157	1455	708	923	803	306	239	280	281	244	174	173	183	142	179	138	102	100	92	65	20
B.16 (4727(80%))	2138	1483	721	919	807	306	235	276	280	246	184	168	180	145	174	137	98	97	84	64	18
A.3 (4822(82%))	2191	1489	820	938	818	325	196	270	290	256	166	175	183	151	174	146	106	94	95	63	20
B.8 (4951(84%))	2265	1547	773	951	827	316	244	288	292	253	192	176	201	155	180	166	113	101	92	69	20
A.4 (5350(91%))	2502	1624	938	1028	889	342	199	283	306	282	183	197	201	187	190	165	113	104	97	69	20

ABSENT GENES

1	cytoplasm
2	nucleus
3	mitochondrion
4	NOT_ANNOTATED
5	membrane
6	endoplasmic reticulum
7	ribosome
8	chromosome
9	endomembrane system
10	mitochondrial envelope
11	nucleolus
12	vacuole
13	cytoskeleton
14	membrane fraction
15	Golgi apparatus
16	cellular bud
17	cell cortex
18	cytoplasmic membrane-bounded vesicle
19	cell wall
20	microtubule organizing center
21	extracellular region

Supplementary Table I (I). Summary of the comparison between *S. cerevisiae* sequences and those of organisms from different group for domains, kingdoms or phyla, classified by cellular localization from the GOSLIM ontology. Each row represents a group from a given domain of life. The group is identified in column 2. The numbers shown in parenthesis on that column indicate the number of *S. cerevisiae* proteins that absent in all organisms of the corresponding group. Columns numbered 1-21 represent each a cellular localization. The number shown in the second row for each of these columns indicates the number of *S. cerevisiae* proteins associated to each GOSLIM category. In each column, a green to red color scale indicates the similarity between the set of proteins associated to the cellular component in *S. cerevisiae* and that of proteins commonly associated to all organisms from the group. Green entries indicate high similarity (low absent), while red entries indicate low similarity (highly absence) between the two sets of proteins. The numbers in each entry indicate the number of *S. cerevisiae* proteins that are absent in all organisms from the group represented in the row.

E	EUKARYOTA									
E.1	FUNGI	E.2	ANIMAL	E.3	PLANTS	E.4	PROTISTS			
E.1.1	Ascomycetes	E.2.1	Echinoderms	E.3.1	Dicotyledons	E.4.1	Alveolates	E.4.5	Entamoeba	
E.1.2	Basidiomycetes	E.2.2	Insects	E.3.2	Green algae	E.4.2	Cellular slime molds	E.4.6	Euglenozoa	
E.1.3	Microsporidians	E.2.3	Nematodes	E.3.3	Monocotyledons	E.4.3	Choanoflagellates	E.4.7		
		E.2.4	Vertebrates	E.3.4	Red algae	E.4.4	Diplomonads		Paraba salids	

A	ARCHEA
A.1	Crenarchaeota
A.2	Euryarchaeota
A.3	Korarchaeota
A.4	Nanoarchaeota

B	BACTERIA				
B.1	Acidobacteria	B.6	Deinococcus-Thermus	B.11	Hyperthermophilic bacteria
B.2	Actinobacteria	B.7	Firmicutes	B.12	Planctomyces
B.3	Bacteroides	B.8	Fusobacteria	B.13	Proteobacteria
B.4	Chlamydia	B.9	Green nonsulfur bacteria	B.14	Spirochete
B.5	Cyanobacteria	B.10	Green sulfur bacteria	B.15	Synergistetes
				B.16	Tenericutes

Table S1. 3 Summary of the comparison between *S. cerevisiae* sequences and those of organisms from different groups for domains, kingdoms or phyla, classified with the different KEGG pathways.

(See the Tables S1.3 A-C. as following)

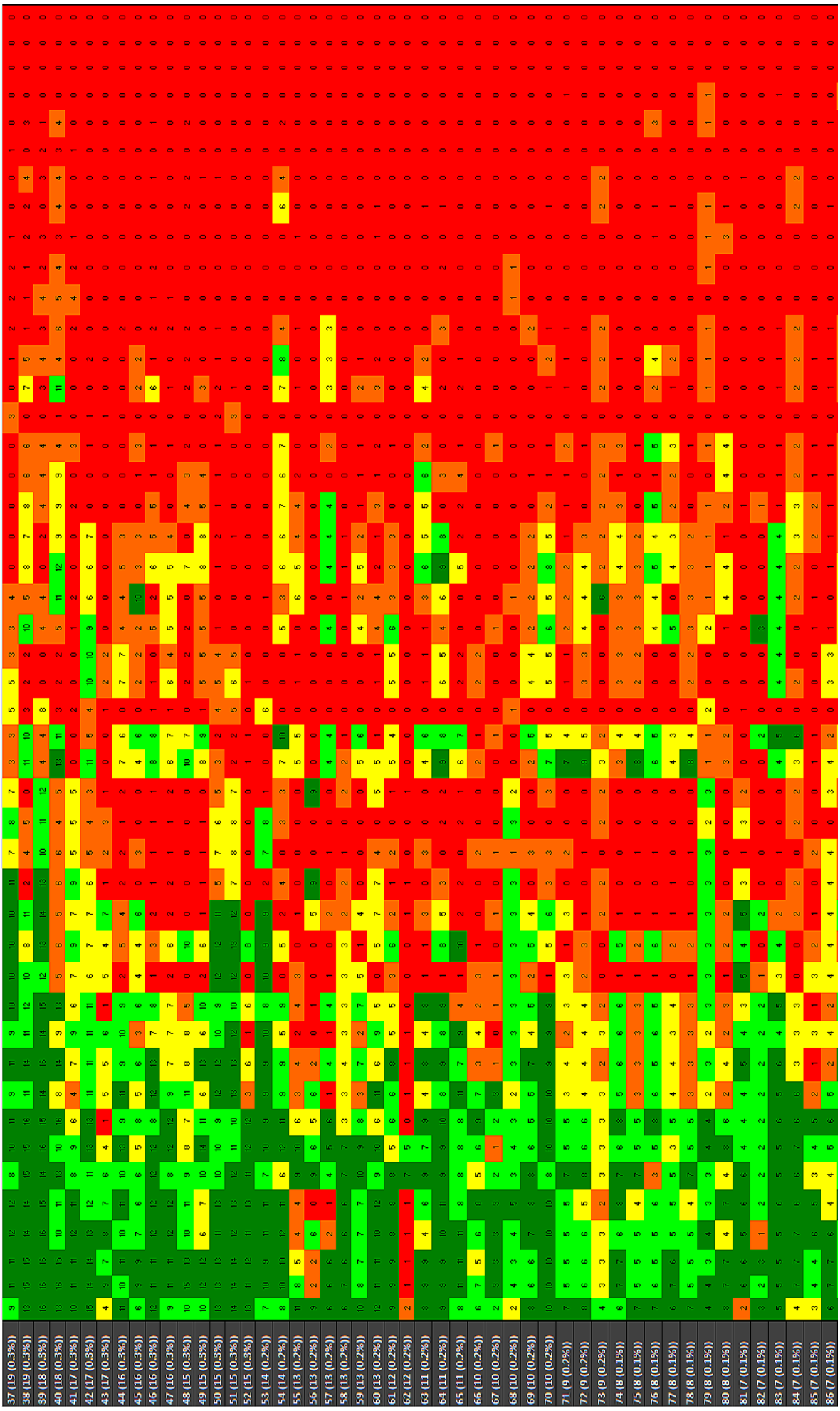
Table S1.3. A *S. cerevisiae* Orthology analysis with Pathways of KEGG

Each row represents a pathway, while each column represents a group from a given domain of life. The numbers shown in parenthesis on the first row of each column indicate the number of *S. cerevisiae* proteins that have orthologs in all organisms of the corresponding group. Rows numbered 1-106 represent each a pathway term and their included proteins of *S. cerevisiae*. In each row, a green to red color scale indicates the similarity between the set of proteins associated to the KEGG pathway in *S. cerevisiae* and that of proteins commonly associated to all organisms from the groups. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of orthologs to *S. cerevisiae* proteins that are common to all organisms from the groups represented in the columns.

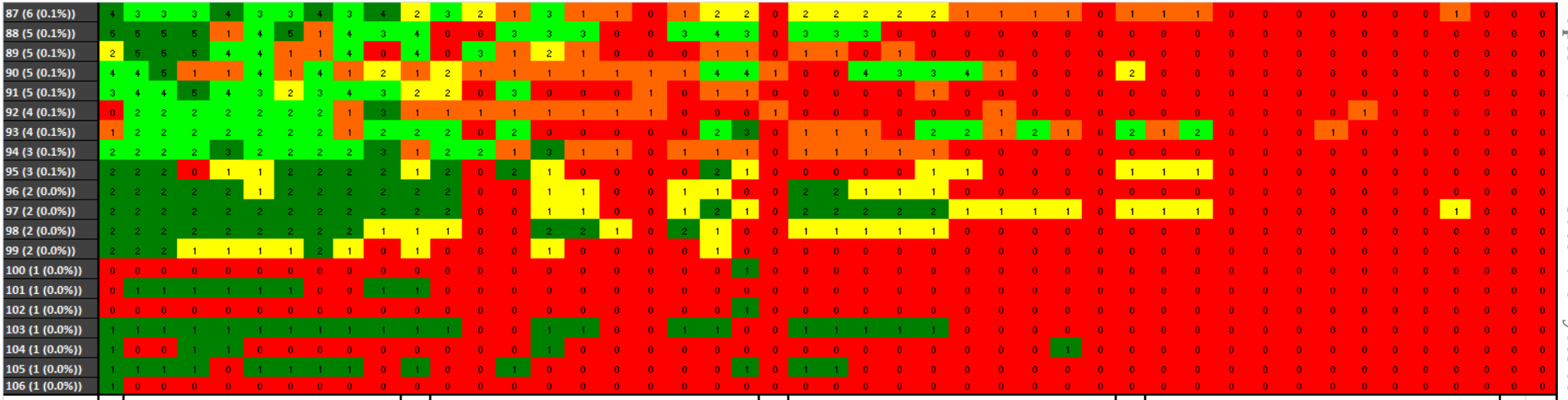
(See the Table S1.3 A. on next page)

Supplementary Table S1.3 A – *S. cerevisiae* Orthology analysis with Pathways of KEGG.

Supplementary Table S1.3 A [continued….]



Supplementary Table S1.3 A [continued….]



1	NOT_ANNOTATED	28	Galactose metabolism	55	Lysine biosynthesis	82	Biotin metabolism
2	Metabolic pathways	29	Glycerophospholipid metabolism	56	O-Mannosyl glycan biosynthesis	83	Limonene and pinene degradation
3	Ribosome	30	Glycine, serine and threonine metabolism	57	Riboflavin metabolism	84	Methane metabolism
4	Cell cycle - yeast	31	SNARE interactions in vesicular transport	58	Sphingolipid metabolism	85	Nicotinate and nicotinamide metabolism
5	Purine metabolism	32	Butanoate metabolism	59	Sulfur metabolism	86	Pentose and glucuronate interconversions
6	Oxidative phosphorylation	33	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	60	Terpenoid backbone biosynthesis	87	Benzoate degradation via CoA ligation
7	Pyrimidine metabolism	34	Phenylalanine, tyrosine and tryptophan biosynthesis	61	Glycerolipid metabolism	88	Arachidonic acid metabolism
8	MAPK signaling pathway - yeast	35	Tyrosine metabolism	62	High-mannose type N-glycan biosynthesis	89	Ether lipid metabolism
9	Glycolysis / Gluconeogenesis	36	Glutathione metabolism	63	Pantothenate and CoA biosynthesis	90	Thiamine metabolism
10	Ubiquitin mediated proteolysis	37	Homologous recombination	64	Propanoate metabolism	91	Ubiquinone and other terpenoid-quinone biosynthesis
11	Starch and sucrose metabolism	38	Selenoamino acid metabolism	65	Reductive carboxylate cycle (CO2 fixation)	92	Androgen and estrogen metabolism
12	Aminoacyl-tRNA biosynthesis	39	Mismatch repair	66	Biosynthesis of unsaturated fatty acids	93	Fatty acid biosynthesis
13	Proteasome	40	Valine, leucine and isoleucine biosynthesis	67	gamma-Hexachlorocyclohexane degradation	94	Ethylbenzene degradation
14	Nucleotide excision repair	41	Base excision repair	68	Non-homologous end-joining	95	Lipoic acid metabolism
15	Citrate cycle (TCA cycle)	42	Fatty acid metabolism	69	Phenylalanine metabolism	96	alpha-Linolenic acid metabolism
16	Endocytosis	43	Regulation of autophagy	70	Valine, leucine and isoleucine degradation	97	Benzoate degradation via hydroxylation
17	Pyruvate metabolism	44	Lysine degradation	71	1- and 2-Methylnaphthalene degradation	98	Synthesis and degradation of ketone bodies
18	DNA replication	45	Nitrogen metabolism	72	3-Chloroacrylic acid degradation	99	Taurine and hypotaurine metabolism
19	Alanine, aspartate and glutamate metabolism	46	Porphyrin and chlorophyll metabolism	73	Cyanoamino acid metabolism	100	1,4-Dichlorobenzene degradation
20	Arginine and proline metabolism	47	Tryptophan metabolism	74	beta-Alanine metabolism	101	Fatty acid elongation in mitochondria
21	Cysteine and methionine metabolism	48	Glyoxylate and dicarboxylate metabolism	75	Drug metabolism - cytochrome P450	102	Fluorobenzoate degradation
22	RNA polymerase	49	Histidine metabolism	76	Drug metabolism - other enzymes	103	Geraniol degradation
23	N-Glycan biosynthesis	50	Inositol phosphate metabolism	77	Folate biosynthesis	104	Other glycan degradation
24	Pentose phosphate pathway	51	Phosphatidylinositol signaling system	78	Metabolism of xenobiotics by cytochrome P450	105	Peptidoglycan biosynthesis
25	Amino sugar and nucleotide sugar metabolism	52	Steroid biosynthesis	79	Protein export	106	Styrene degradation
26	Fructose and mannose metabolism	53	Lysosome	80	Vitamin B6 metabolism		
27	Basal transcription factors	54	One carbon pool by folate	81	Aminophosphonate metabolism		

Table S1.3. B ***S. cerevisiae*** **Homology analysis with Pathways of KEGG**

Each row represents a pathway, while each column represents a group from a given domain of life. The numbers shown in parenthesis on the first row of each column indicate the number of *S. cerevisiae* proteins that have homologues in all organisms of the corresponding group. Rows numbered 1-106 represent each a pathway term and their included proteins. In each row, a green to red color scale indicates the similarity between the set of proteins associated to the KEGG pathway in *S. cerevisiae* and that of proteins commonly associated to all organisms from the groups. Green entries indicate high similarity, while red entries indicate low similarity between the two sets of proteins. The numbers in each entry indicate the number of homologues to *S. cerevisiae* proteins that are common to all organisms from the groups represented in the columns.

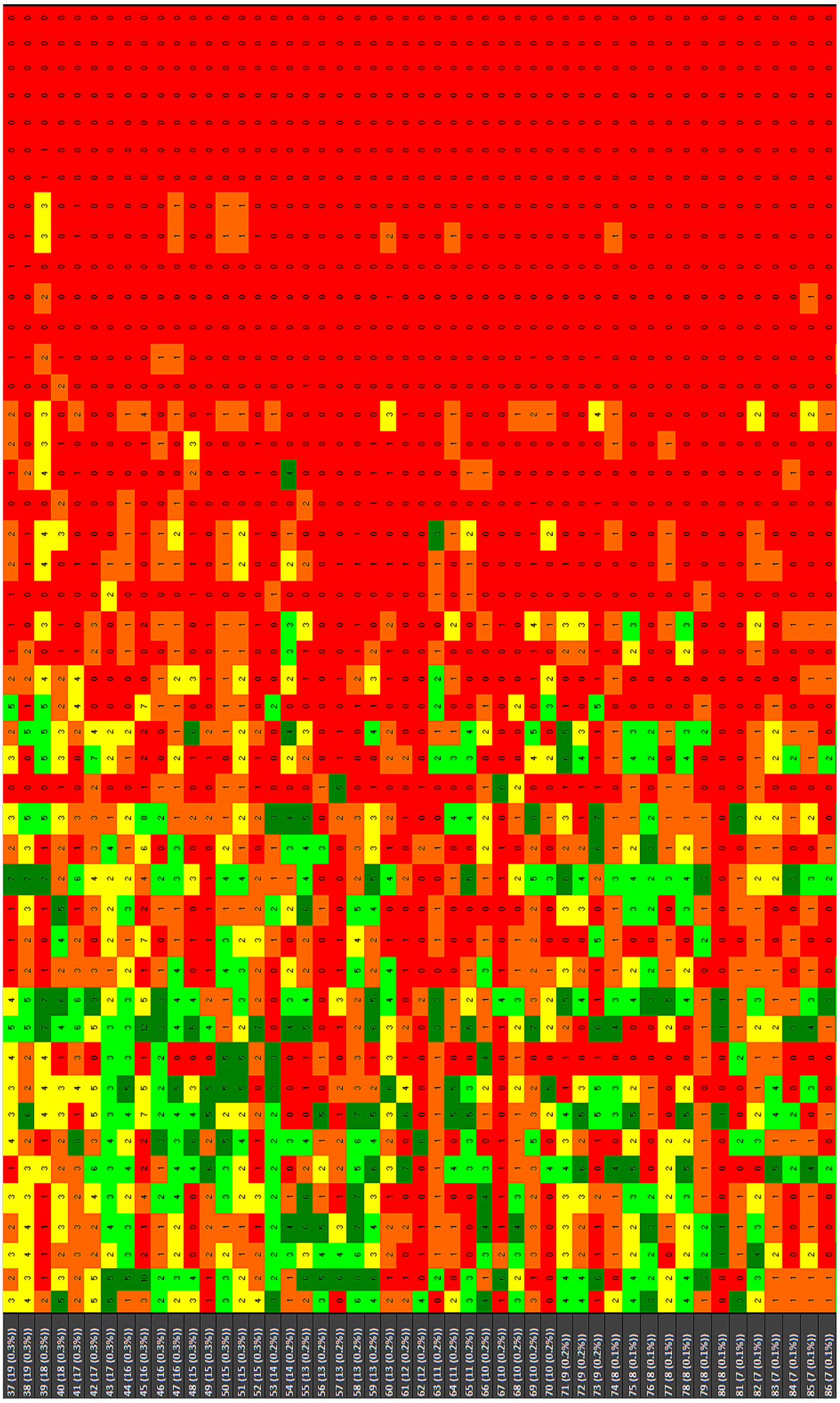
(See the Table S1.3 B. on next page)

Supplementary Table S1.3 B – *S. cerevisiae* Homology analysis with Pathways of KEGG.

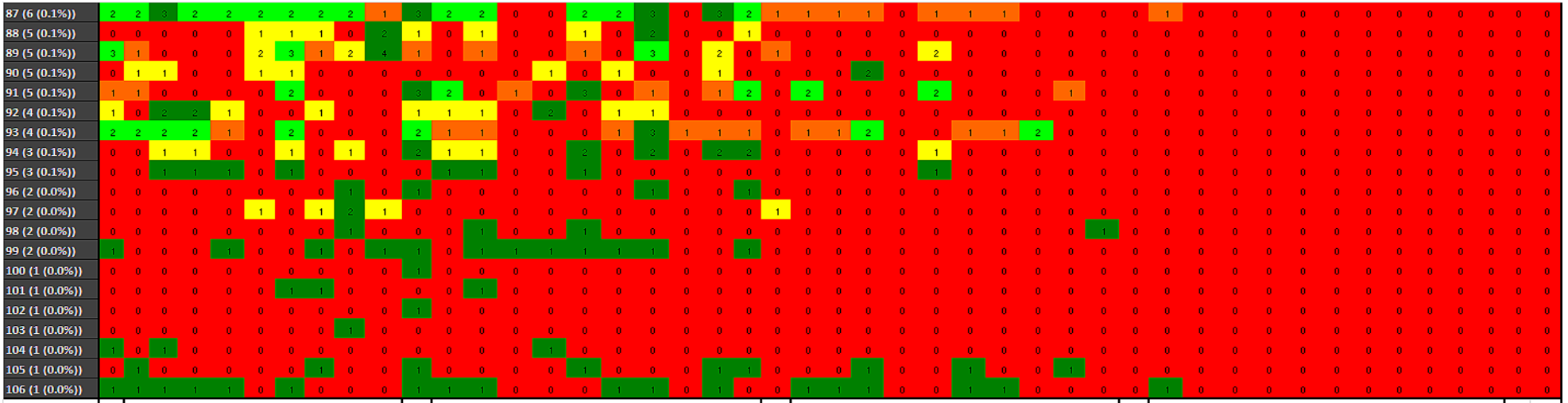
Supplementary Table 1.3 (B). *S.cerevisiae* HOMOLOGY ANALYSIS WITH PATHWAYS OF KEGG

Frequency of *S. cerevisiae* genes found orthologs in phylogenetic groups / Frequency of *S. cerevisiae* genes distributed in KEGG pathways	PROTIST::Choanoflagellates (1162 (20%))	PROTIST::Cellular slime molds (1156 (20%))	PLANTS::Dicotyledons (1147 (20%))	PLANTS::Monocotyledons (1091 (19%))	ANIMAL::Echinoderms (1075 (18%))	PROTIST::Paraba salids (1054 (20%))	PLANTS::Red algae (1051 (18%))	PROTIST::Entamoeba (945 (16%))	PROTIST::Diplomonads (752 (13%))	FUNGI::Microsporidians (749 (13%))	BACTERIA::Acidobacteria (740 (13%))	BACTERIA::Planctomyces (712 (12%))	PROTIST::Euglenozoa (695 (12%))	ANIMAL::Nematodes (693 (12%))	ANIMAL::Insects (638 (11%))	ARCHEA::Korarchaeota (638 (11%))	PLANTS::Green algae (636 (11%))	BACTERIA::Fusobacteria (491 (8%))	FUNGI::Basidiomycetes (469 (8%))	BACTERIA::Synergistetes (343 (6%))	BACTERIA::Deinococcus-Thermus (326 (6%))	ARCHEA::Nanoarchaeota (310 (5%))	BACTERIA::Green sulfur bacteria (260 (4%))	PLANTS (253 (4%))	BACTERIA::Hyperthermophilic bacteria (251 (4%))	PROTIST::Alveolates (194 (3%))	BACTERIA::Chlamydia (190 (3%))	BACTERIA::Green nonsulfur bacteria (156 (3%))	ANIMAL::Vertebrates (125 (2%))	BACTERIA::Bacteroides (104 (2%))	BACTERIA::Cyanobacteria (89 (2%))	ARCHEA::Crenarchaeota (82 (1%))	ANIMAL (81 (1%))	BACTERIA::Spirochete (64 (1%))	PROTIST (43 (0.7%))	BACTERIA::Actinobacteria (34 (0.6%))	BACTERIA::Tenericutes (33 (0.6%))	ARCHEA::Euryarchaeota (30 (0.5%))	ARCHEA (18 (0.3%))	BACTERIA::Firmicutes (8 (0.1%))	BACTERIA::Proteobacteria (1 (0.01 %))	EUKARYOTES (0 (0%))	FUNGI (0 (0%))	FUNGI::Ascomycetes (0 (0%))	PROKARYOTES (0 (0%))	BACTERIA (0 (0%))
1 (4571 (77.7%))	935	922	938	878	875	800	805	723	562	573	482	480	534	563	512	385	501	291	400	199	183	169	154	198	142	165	109	89	108	62	48	38	74	40	40	24	22	12	9	6	0	0	0	0	0	0
2 (610 (10.4%))	103	107	91	94	87	107	131	103	78	59	155	131	72	56	67	128	59	111	24	71	70	42	55	24	54	8	34	38	7	13	21	23	3	6	0	3	0	4	0	0	0	0	0	0	0	0
3 (142 (2.4%))	8	10	10	12	12	22	8	13	22	26	24	21	15	7	5	40	8	31	1	25	15	50	13	6	24	2	16	14	0	13	9	9	0	7	0	3	5	5	3	0	0	0	0	0	0	0
4 (115 (2.0%))	27	28	28	27	21	29	24	24	24	23	15	17	17	18	11	12	14	10	16	9	15	8	5	6	4	7	7	0	4	2	0	0	2	4	2	2	0	0	0	0	0	0	0	0	0	0
5 (89 (1.5%))	16	14	11	7	7	21	16	16	13	10	8	7	15	3	7	24	6	9	6	2	3	12	8	1	5	2	2	1	1	3	0	4	0	1	0	1	0	0	0	0	0	0	0	0	0	0
6 (76 (1.3%))	9	8	8	6	12	8	12	5	8	10	18	15	6	4	3	9	4	10	4	5	8	7	8	1	2	2	4	8	0	0	6	1	0	1	0	0	2	0	0	0	0	0	0	0	0	0
7 (69 (1.2%))	14	11	11	8	9	14	13	11	11	8	9	8	14	4	7	20	9	10	3	5	8	12	8	2	4	0	1	2	1	1	1	3	1	2	0	1	0	0	0	0	0	0	0	0	0	0
8 (55 (0.9%))	7	7	7	5	10	9	7	6	2	5	4	2	5	6	3	13	5	3	2	5	4	2	1	2	1	2	2	1	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
9 (48 (0.8%))	7	7	6	6	11	13	9	13	9	3	11	16	5	3	7	13	4	7	1	13	12	1	6	4	8	1	5	7	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
10 (44 (0.7%))	10	7	6	6	5	11	9	10	10	9	8	9	8	5	3	3	5	7	3	3	2	0	5	3	2	2	0	0	0	1	1	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0
11 (40 (0.7%))	8	2	11	10	4	6	9	0	1	5	10	4	0	6	0	1	5	9	0	2	4	0	2	5	4	0	3	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12 (39 (0.7%))	1	1	2	1	3	10	2	8	9	5	9	4	10	0	0	10	3	9	1	10	8	14	12	1	12	0	7	3	0	3	4	2	0	3	0	0	3	3	2	1	0	0	0	0	0	0
13 (35 (0.6%))	1	2	2	2	3	4	2	2	7	5	1	0	2	2	1	5	1	0	2	0	0	7	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0
14 (34 (0.6%))	6	4	7	8	6	10	11	7	3	9	6	7	5	6	5	12	3	5	1	6	4	7	3	2	3	0	4	3	0	3	3	3	0	2	0	2	0	2	2	1	1	0	0	0	0	0
15 (33 (0.6%))	6	5	2	0	2	6	4	5	4	0	9	10	1	0	0	9	0	4	2	1	7	1	3	0	1	1	2	2	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16 (33 (0.6%))	8	7	7	9	4	8	11	9	5	3	0	0	6	3	1	0	3	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17 (33 (0.6%))	3	5	0	3	2	9	5	11	10	2	8	8	0	1	2	8	2	11	0	5	6	1	2	0	2	1	1	2	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18 (30 (0.5%))	3	1	3	2	3	5	5	6	4	4	8	7	4	1	3	8	2	8	0	6	6	6	4	0	5	0	3	3	0	4	3	3	0	2	0	2	0	2	2	1	1	0	0	0	0	0
19 (29 (0.5%))	5	11	4	3	7	7	2	11	6	2	17	9	3	9	3	10	7	14	1	5	4	7	3	0	5	1	2	4	0	0	4	7	0	0	0	0	0	1	0	0	0	0	0	0	0	0
20 (29 (0.5%))	6	9	8	7	6	8	6	6	4	1	11	8	4	5	3	8	3	11	1	5	4	3	4	2	4	0	2	3	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
21 (29 (0.5%))	7	4	7	7	4	2	6	7	2	0	12	10	1	2	5	11	5	9	0	3	6	0	3	2	6	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22 (29 (0.5%))	6	5	4	2	3	5	5	5	6	4	1	1	6	1	1	10	1	2	3	0	2	8	1	0	2	0	1	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
23 (28 (0.5%))	6	6	5	4	5	6	10	6	2	2	4	4	4	4	5	2	3	1	2	3	0	5	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24 (28 (0.5%))	1	5	3	3	1	13	3	6	5	4	12	7	6	0	0	6	3	4	1	4	4	0	3	2	5	1	4	2	0	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0
25 (27 (0.5%))	1	0	2	1	2	4	3	5	1	5	7	8	1	2	2	8	0	3	0	5	4	2	3	0	6	0	4	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26 (27 (0.5%))	4	8	6	5	2	4	8	6	5	5	9	14	3	0	5	12	6	2	0	6	4	2	2	2	8	1	3	3	1	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0
27 (23 (0.4%))	8	8	8	9	8	9	11	7	2	13	2	0	2	10	8	2	5	0	4	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28 (23 (0.4%))	3	0	11	10	0	4	11	0	2	3	9	6	3	3	0	4	5	8	0	3	3	0	2	5	3	1	5	2	0	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0
29 (23 (0.4%))	5	4	4	5	2	5	6	6	5	6	7	5	6	2	3	4	3	12	2	5	5	1	3	2	2	3	8	2	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
30 (23 (0.4%))	5	1	5	5	3	5	0	5	5	2	2	4	1	1	3	7	2	6	0	2	2	1	2	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31 (23 (0.4%))	10	7	8	10	8	13	10	9	4	4	0	0	6	5	4	0	4	1	4	0	0	0	0	4	0	2	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
32 (22 (0.4%))	5	6	4	4	3	8	7	6	5	2	8	10	2	2	7	11	3	2	0	5	3	1	1	1	6	0	0	4	1	0	1	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0
33 (22 (0.4%))	6	8	4	6	4	1	1	3	5	5	4	1	1	3	6	1	1	1	2	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
34 (22 (0.4%))	3	2	3	2	3	4	6	3	2	1	11	3	2	3	3	6	4	7	1	5	6	3	4	1	6	0	1	2	0	0	1	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0
35 (20 (0.3%))	7	4	5	5	5	8	6	8	3	1	7	7	4	2	6	9	3	9	1	7	8	1	0	2	6	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36 (19 (0.3%))	3	0	1	0	0	1	0	1	2	2	9	5	0	0	0	7	0	2	0	2	3	1	3	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Supplementary Table S1.3 B [continued…]



Supplementary Table S1.3 B [continued….]



	1	NOT_ANNOTATED	28	Galactose metabolism	55	Lysine biosynthesis	82	Biotin metabolism		
	2	Metabolic pathways	29	Glycerophospholipid metabolism	56	O-Mannosyl glycan biosynthesis	83	Limonene and pinene degradation		
	3	Ribosome	30	Glycine, serine and threonine metabolism	57	Riboflavin metabolism	84	Methane metabolism		
	4	Cell cycle - yeast	31	SNARE interactions in vesicular transport	58	Sphingolipid metabolism	85	Nicotinate and nicotinamide metabolism		
	5	Purine metabolism	32	Butanoate metabolism	59	Sulfur metabolism	86	Pentose and glucuronate interconversions		
	6	Oxidative phosphorylation	33	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	60	Terpenoid backbone biosynthesis	87	Benzoate degradation via CoA ligation		
	7	Pyrimidine metabolism	34	Phenylalanine, tyrosine and tryptophan biosynthesis	61	Glycerolipid metabolism	88	Arachidonic acid metabolism		
	8	MAPK signaling pathway - yeast	35	Tyrosine metabolism	62	High-mannose type N-glycan biosynthesis	89	Ether lipid metabolism		
	9	Glycolysis / Gluconeogenesis	36	Glutathione metabolism	63	Pantothenate and CoA biosynthesis	90	Thiamine metabolism		
	10	Ubiquitin mediated proteolysis	37	Homologous recombination	64	Propanoate metabolism	91	Ubiquinone and other terpenoid-quinone biosynthesis		
	11	Starch and sucrose metabolism	38	Selenoamino acid metabolism	65	Reductive carboxylate cycle (CO2 fixation)	92	Androgen and estrogen metabolism		
	12	Aminoacyl-tRNA biosynthesis	39	Mismatch repair	66	Biosynthesis of unsaturated fatty acids	93	Fatty acid biosynthesis		
	13	Proteasome	40	Valine, leucine and isoleucine biosynthesis	67	gamma-Hexachlorocyclohexane degradation	94	Ethylbenzene degradation		
	14	Nucleotide excision repair	41	Base excision repair	68	Non-homologous end-joining	95	Lipoic acid metabolism		
	15	Citrate cycle (TCA cycle)	42	Fatty acid metabolism	69	Phenylalanine metabolism	96	alpha-Linolenic acid metabolism		
	16	Endocytosis	43	Regulation of autophagy	70	Valine, leucine and isoleucine degradation	97	Benzoate degradation via hydroxylation		
	17	Pyruvate metabolism	44	Lysine degradation	71	1- and 2-Methylnaphthalene degradation	98	Synthesis and degradation of ketone bodies		
	18	DNA replication	45	Nitrogen metabolism	72	3-Chloroacrylic acid degradation	99	Taurine and hypotaurine metabolism		
	19	Alanine, aspartate and glutamate metabolism	46	Porphyrin and chlorophyll metabolism	73	Cyanoamino acid metabolism	100	1,4-Dichlorobenzene degradation		
	20	Arginine and proline metabolism	47	Tryptophan metabolism	74	beta-Alanine metabolism	101	Fatty acid elongation in mitochondria		
	21	Cysteine and methionine metabolism	48	Glyoxylate and dicarboxylate metabolism	75	Drug metabolism - cytochrome P450	102	Fluorobenzoate degradation		
	22	RNA polymerase	49	Histidine metabolism	76	Drug metabolism - other enzymes	103	Geraniol degradation		
	23	N-Glycan biosynthesis	50	Inositol phosphate metabolism	77	Folate biosynthesis	104	Other glycan degradation		
	24	Pentose phosphate pathway	51	Phosphatidylinositol signaling system	78	Metabolism of xenobiotics by cytochrome P450	105	Peptidoglycan biosynthesis		
	25	Amino sugar and nucleotide sugar metabolism	52	Steroid biosynthesis	79	Protein export	106	Styrene degradation		
	26	Fructose and mannose metabolism	53	Lysosome	80	Vitamin B6 metabolism				
	27	Basal transcription factors	54	One carbon pool by folate	81	Aminophosphonate metabolism				

Table S1.3. C ***S. cerevisiae*** **Absent genes analysis with Pathways of KEGG**

Each row represents a pathway, while each column represents a group from a given domain of life. The numbers shown in parenthesis on the first row of each column indicate the number of *S. cerevisiae* proteins that are absent in all organisms of the corresponding group. Rows numbered 1-106 represent each a pathway term. In each row, a green to red color scale indicates the similarity between the set of proteins associated to the KEGG pathway in *S. cerevisiae* and that of proteins commonly associated to all organisms from the kingdom or phylum. Green entries indicate high similarity (low absent), while red entries indicate low similarity (highly absence) of proteins between the two sets of proteins. The numbers in each entry indicate the number of *S. cerevisiae* proteins that are absent in all organisms from the groups represented in the columns.

(See the Table S1.3 C. on next page)

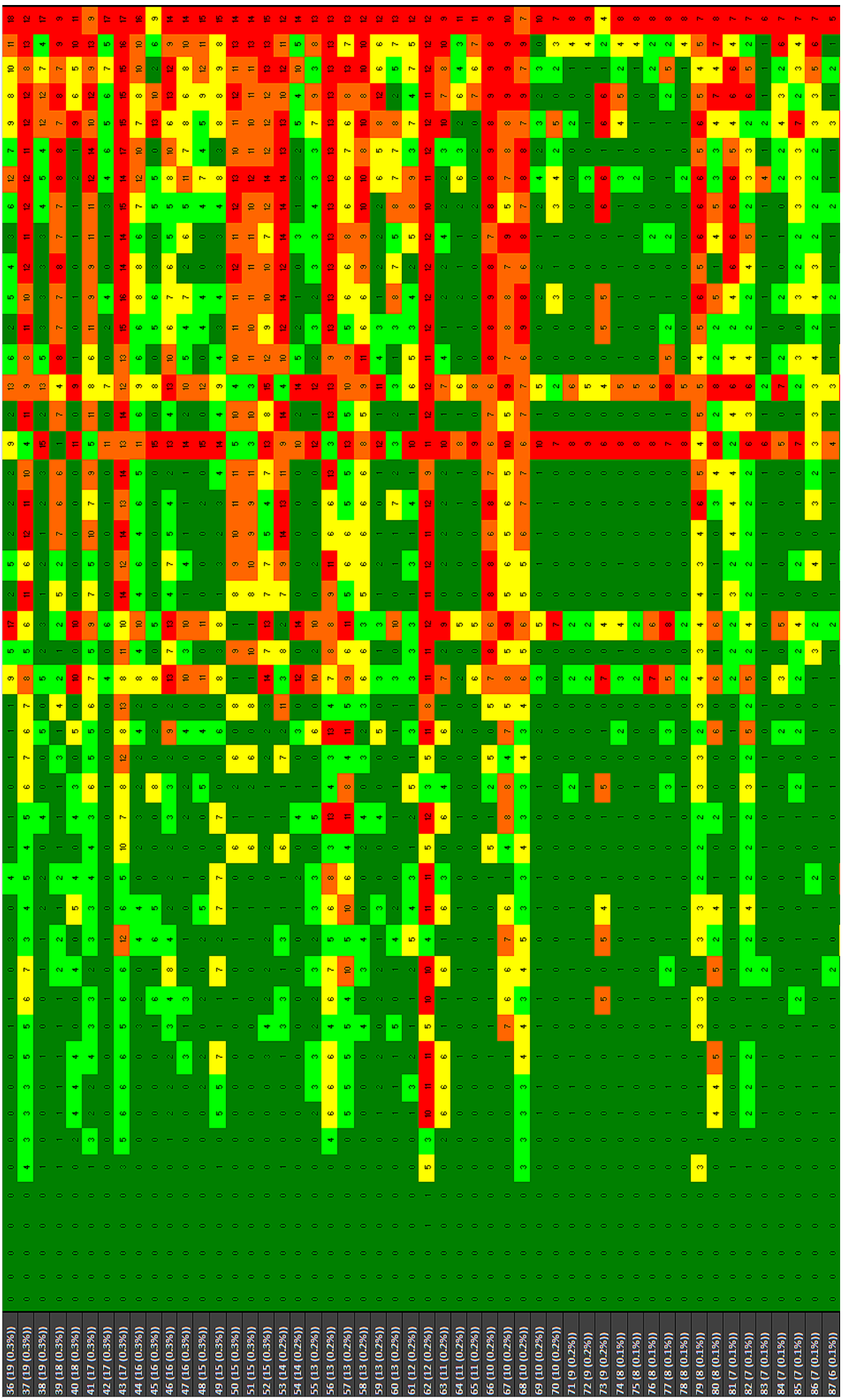
Supplementary Table S1.3 C – *S. cerevisiae* Absent genes analysis with Pathways of KEGG.

Supplementary Table 1.3 (C). *S.cerevisiae* NON-HOMOLOGY (ABSENT GENES) ANALYSIS WITH PATHWAYS OF KEGG

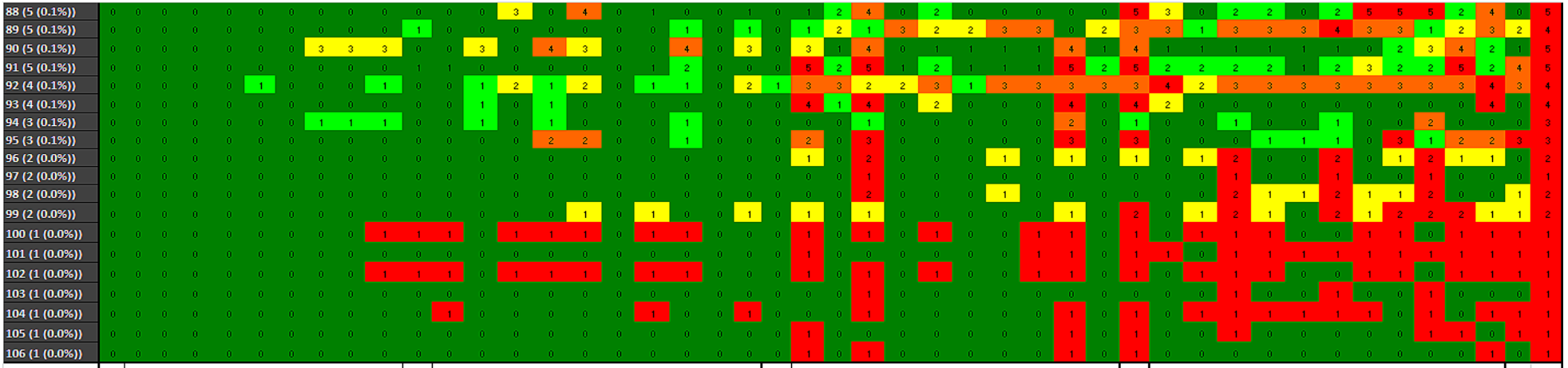
Frequency of *S. cerevisiae* genes found orthologs in phylogenetic groups / Frequency of *S. cerevisiae* genes distributed in KEGG pathways	EUKARYOTES (0 (0%))	FUNGI (10 (0.2%))	FUNGI::Ascomycetes (10 (0.2%))	PLANTS (142 (2%))	PLANTS::Dicotyledons (282 (5%))	FUNGI::Basidiomycetes (1525 (26%))	PROTIST (1978 (34%))	ANIMAL (2028 (34%))	ANIMAL::Vertebrates (2185 (37%))	ANIMAL::Insects (2371 (40%))	PLANTS::Green algae (2519 (43%))	PLANTS::Monocotyledons (2527 (43%))	ANIMAL::Nematodes (2540 (43%))	PLANTS::Red algae (2554 (43%))	ANIMAL::Echinoderms (2582 (44%))	PROTIST::Cellular slime molds (2603 (44%))	PROKARYOTES (2642 (45%))	PROTIST::Alveolates (2740 (47%))	PROTIST::Choanoflagellates (2755 (47%))	BACTERIA (2881 (49%))	PROTIST::Euglenozoa (3157 (54%))	BACTERIA::Proteobacteria (3256 (55 %))	PROTIST::Paraba salids (3616 (62%))
1 (4571 (77.7%))	0	8	8	132	265	1457	1860	1867	1995	2162	2298	2293	2288	2294	2309	2355	2349	2438	2466	2484	2749	2787	3057
2 (610 (10.4%))	0	0	0	5	8	24	48	84	99	110	95	107	130	108	137	125	83	144	141	101	204	121	335
3 (142 (2.4%))	0	2	2	0	2	2	2	2	2	2	2	2	2	1	6	6	17	3	8	83	4	93	8
4 (115 (2.0%))	0	0	0	3	3	17	22	19	22	21	33	32	30	37	30	27	51	37	34	55	46	64	42
5 (89 (1.5%))	0	0	0	0	0	5	7	9	9	12	12	20	16	19	14	11	12	21	17	27	27	29	40
6 (76 (1.3%))	0	0	0	3	3	3	18	14	19	17	24	25	21	20	20	24	28	26	21	30	39	34	55
7 (69 (1.2%))	0	0	0	0	0	5	6	4	5	6	9	12	9	10	12	12	11	12	12	26	18	28	27
8 (55 (0.9%))	0	0	0	2	4	10	21	19	19	20	26	26	23	23	20	24	25	26	22	27	34	31	29
9 (48 (0.8%))	0	0	0	0	0	0	0	2	2	2	0	2	4	5	5	5	0	1	10	0	2	3	14
10 (44 (0.7%))	0	0	0	1	1	4	5	4	5	4	4	4	5	5	5	5	24	7	5	24	7	31	7
11 (40 (0.7%))	0	0	0	0	1	3	7	12	16	12	9	8	13	15	16	15	4	15	15	4	29	8	22
12 (39 (0.7%))	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	3	0	5
13 (35 (0.6%))	0	0	0	0	0	1	2	1	2	1	3	3	2	3	2	2	12	3	4	14	4	20	4
14 (34 (0.6%))	0	0	0	0	0	1	3	1	1	2	4	3	2	4	3	7	14	4	8	16	10	17	10
15 (33 (0.6%))	0	0	0	0	0	0	1	2	3	2	4	4	3	3	6	2	2	4	4	2	4	2	24
16 (33 (0.6%))	0	0	0	0	0	0	1	0	0	0	5	1	1	5	2	1	18	5	3	20	4	25	6
17 (33 (0.6%))	0	0	0	0	0	0	0	2	3	2	1	3	2	3	3	1	0	2	2	0	4	0	14
18 (30 (0.5%))	0	0	0	0	0	1	3	1	1	1	3	3	2	4	2	5	4	4	4	8	4	9	6
19 (29 (0.5%))	0	0	0	0	0	0	0	0	0	0	1	6	2	6	5	1	0	1	7	0	4	0	13
20 (29 (0.5%))	0	0	0	0	0	0	0	1	1	2	2	1	4	2	2	4	0	4	3	0	5	0	14
21 (29 (0.5%))	0	0	0	0	0	1	0	2	5	5	4	3	5	2	3	4	0	7	1	0	7	0	16
22 (29 (0.5%))	0	0	0	0	0	4	3	3	3	5	6	9	7	6	6	4	7	7	8	19	9	21	9
23 (28 (0.5%))	0	0	0	1	2	3	5	3	3	3	6	6	4	6	7	6	12	8	12	12	11	13	12
24 (28 (0.5%))	0	0	0	0	0	0	0	1	1	2	0	2	1	4	2	1	0	3	1	0	2	0	4
25 (27 (0.5%))	0	0	0	0	0	0	1	3	3	3	7	6	3	6	7	11	1	4	8	1	7	4	11
26 (27 (0.5%))	0	0	0	0	0	0	1	2	2	2	0	1	4	4	2	6	0	4	6	0	2	4	12
27 (23 (0.4%))	0	0	0	0	0	0	1	0	1	1	2	5	2	4	2	3	12	11	12	16	18	18	11
28 (23 (0.4%))	0	0	0	0	0	0	0	1	1	1	1	1	1	2	1	5	0	2	4	0	8	3	4
29 (23 (0.4%))	0	0	0	0	0	0	0	2	2	3	5	3	3	3	5	4	0	2	10	0	3	1	8
30 (23 (0.4%))	0	0	0	0	0	1	0	3	5	7	1	1	6	5	4	8	0	7	2	0	7	0	10
31 (23 (0.4%))	0	0	0	0	0	1	5	5	6	9	9	8	9	10	12	11	16	12	9	16	12	20	10
32 (22 (0.4%))	0	0	0	0	0	0	1	1	1	1	0	0	3	0	1	2	0	1	4	0	2	0	8
33 (22 (0.4%))	0	0	0	0	1	4	3	3	3	4	8	8	10	14	7	3	14	9	12	15	13	16	21
34 (22 (0.4%))	0	0	0	0	0	0	4	8	10	11	3	4	9	2	10	12	1	7	4	1	14	1	14
35 (20 (0.3%))	0	0	0	0	0	1	0	1	1	2	1	2	2	2	2	2	0	1	2	0	3	0	6

Frequency of *S. cerevisiae* genes found orthologs in phylogenetic groups / Frequency of *S. cerevisiae* genes distributed in KEGG pathways	ARCHEA (3672 (62%))	PROTIST::Entamoeba (3678 (63%))	BACTERIA::Firmicutes (3692 (63%))	ARCHEA::Euryarchaeota (3807 (65%))	BACTERIA::Actinobacteria (3889 (66%))	BACTERIA::Cyanobacteria (3940 (67%))	BACTERIA::Bacteroides (4076 (69%))	FUNGI::Microsporidians (4097 (70%))	BACTERIA::Green nonsulfur bacteria (4178 (71%))	PROTIST::Diplomonads (4195 (71%))	ARCHEA::Crenarchaeota (4209 (72%))	BACTERIA::Spirochete (4363 (74%))	BACTERIA::Green sulfur bacteria (4393 (75%))	BACTERIA::Deinococcus-Thermus (4408 (75%))	BACTERIA::Acidobacteria (4413 (75%))	BACTERIA::Planctomyces (4517 (77%))	BACTERIA::Synergistetes (4677 (80%))	BACTERIA::Hyperthermophilic bacteria (4685 (80%))	BACTERIA::Chlamydia (4712 (80%))	BACTERIA::Tenericutes (4727 (80%))	ARCHEA::Korarchaeota (4822 (82%))	BACTERIA::Fusobacteria (4951 (84%))	ARCHEA::Nanoarchaeota (5350 (91%))
1 (4571 (77.7%))	3215	3087	3148	3307	3325	3325	3461	3372	3541	3457	3623	3686	3674	3715	3687	3748	3887	3890	3874	3884	4004	###	4297
2 (610 (10.4%))	171	359	148	187	170	190	181	450	203	424	235	215	248	238	246	271	302	283	347	356	380	368	553
3 (142 (2.4%))	19	7	98	27	99	101	100	24	104	16	24	102	102	104	112	118	106	102	106	100	44	106	44
4 (115 (2.0%))	66	47	71	68	79	75	77	50	76	59	76	85	84	82	83	85	92	98	88	94	91	104	101
5 (89 (1.5%))	20	44	38	21	41	38	39	49	42	44	29	43	43	46	47	49	52	49	57	56	45	50	69
6 (76 (1.3%))	39	57	36	42	43	44	45	59	43	60	44	46	44	46	44	52	59	62	50	54	59	57	66
7 (69 (1.2%))	14	34	33	15	36	33	35	27	37	31	19	35	37	37	42	41	43	40	41	39	31	43	47
8 (55 (0.9%))	27	28	35	31	34	34	35	32	34	36	32	37	40	37	36	38	40	51	38	38	42	52	53
9 (48 (0.8%))	2	14	1	2	3	3	0	26	5	26	6	0	4	4	4	8	12	7	11	7	20	16	47
10 (44 (0.7%))	33	7	31	33	30	32	33	10	33	16	38	37	32	33	33	35	40	40	39	44	40	37	44
11 (40 (0.7%))	13	21	6	15	11	19	9	27	13	31	18	18	23	18	24	25	23	22	26	24	39	28	40
12 (39 (0.7%))	0	4	0	0	0	0	0	6	0	4	1	0	0	0	4	4	1	0	0	1	4	4	2
13 (35 (0.6%))	14	4	28	14	17	27	29	5	28	7	15	29	29	29	28	29	29	29	28	29	15	29	15
14 (34 (0.6%))	16	12	22	17	24	22	23	8	23	18	19	23	23	26	26	25	25	27	26	25	18	28	22
15 (33 (0.6%))	4	24	4	5	4	5	4	31	4	27	4	4	5	4	4	7	12	10	5	20	17	23	32
16 (33 (0.6%))	23	5	26	26	24	26	28	18	27	13	28	28	28	28	28	28	28	28	26	26	32	28	32
17 (33 (0.6%))	0	18	0	0	0	2	1	22	1	17	0	2	5	0	0	8	7	9	12	9	15	11	32
18 (30 (0.5%))	4	7	13	5	14	13	15	5	16	9	9	17	13	17	22	22	20	16	21	22	10	22	12
19 (29 (0.5%))	0	13	0	1	1	1	1	26	1	16	0	6	9	6	1	6	9	3	15	10	6	7	22
20 (29 (0.5%))	2	19	1	3	2	2	3	28	3	19	4	3	6	5	6	3	8	5	14	12	10	12	26
21 (29 (0.5%))	2	16	2	4	2	2	3	29	4	23	5	2	8	7	5	6	6	6	22	18	10	11	29
22 (29 (0.5%))	9	10	21	10	23	22	23	9	23	10	10	23	23	23	23	23	23	23	23	23	11	23	13
23 (28 (0.5%))	16	13	15	18	16	17	19	26	23	25	18	19	21	20	22	23	21	22	24	24	25	27	23
24 (28 (0.5%))	5	12	1	6	1	0	0	15	2	11	10	1	2	1	3	6	11	3	4	6	21	13	28
25 (27 (0.5%))	9	6	1	10	3	14	8	12	14	20	14	8	15	12	13	14	11	11	16	16	18	18	25
26 (27 (0.5%))	6	9	3	6	5	6	3	14	9	17	10	4	8	8	10	8	8	8	11	10	14	20	25
27 (23 (0.4%))	17	12	22	20	21	20	20	4	22	20	19	23	22	22	21	22	22	22	23	22	20	23	21
28 (23 (0.4%))	4	2	0	4	2	8	2	14	5	17	6	2	9	6	9	7	6	5	8	7	19	9	23
29 (23 (0.4%))	8	9	4	8	8	8	9	9	9	14	13	6	9	14	12	13	13	12	9	9	18	8	22
30 (23 (0.4%))	0	15	0	0	0	1	0	19	2	18	4	0	2	2	3	2	5	3	11	6	4	10	22
31 (23 (0.4%))	23	11	22	23	22	22	23	19	23	19	23	23	23	23	23	23	23	23	23	22	23	22	23
32 (22 (0.4%))	0	12	0	0	0	1	0	13	0	13	1	1	3	2	1	4	5	2	5	3	4	14	21
33 (22 (0.4%))	19	14	18	19	17	20	18	15	19	17	21	20	20	20	18	21	21	21	21	21	21	21	21
34 (22 (0.4%))	2	17	2	4	2	3	4	18	3	16	2	2	4	4	5	6	5	6	5	10	10	8	17
35 (20 (0.3%))	2	8	1	3	2	3	3	18	3	12	3	1	4	4	3	4	7	4	6	4	6	6	19

Supplementary Table S1.3 C [continued….]



Supplementary Table S1.3 C [continued….]



1	NOT_ANNOTATED	30	Glycine, serine and threonine metabolism	59	Sulfur metabolism	88	Arachidonic acid metabolism
2	Metabolic pathways	31	SNARE interactions in vesicular transport	60	Terpenoid backbone biosynthesis	89	Ether lipid metabolism
3	Ribosome	32	Butanoate metabolism	61	Glycerolipid metabolism	90	Thiamine metabolism
4	Cell cycle - yeast	33	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	62	High-mannose type N-glycan biosynthesis	91	Ubiquinone and other terpenoid-quinone biosynthesis
5	Purine metabolism	34	Phenylalanine, tyrosine and tryptophan biosynthesis	63	Pantothenate and CoA biosynthesis	92	Androgen and estrogen metabolism
6	Oxidative phosphorylation	35	Tyrosine metabolism	64	Propanoate metabolism	93	Fatty acid biosynthesis
7	Pyrimidine metabolism	36	Glutathione metabolism	65	Reductive carboxylate cycle (CO2 fixation)	94	Ethylbenzene degradation
8	MAPK signaling pathway - yeast	37	Homologous recombination	66	Biosynthesis of unsaturated fatty acids	95	Lipoic acid metabolism
9	Glycolysis / Gluconeogenesis	38	Selenoamino acid metabolism	67	gamma-Hexachlorocyclohexane degradation	96	alpha-Linolenic acid metabolism
10	Ubiquitin mediated proteolysis	39	Mismatch repair	68	Non-homologous end-joining	97	Benzoate degradation via hydroxylation
11	Starch and sucrose metabolism	40	Valine, leucine and isoleucine biosynthesis	69	Phenylalanine metabolism	98	Synthesis and degradation of ketone bodies
12	Aminoacyl-tRNA biosynthesis	41	Base excision repair	70	Valine, leucine and isoleucine degradation	99	Taurine and hypotaurine metabolism
13	Proteasome	42	Fatty acid metabolism	71	1- and 2-Methylnaphthalene degradation	100	1,4-Dichlorobenzene degradation
14	Nucleotide excision repair	43	Regulation of autophagy	72	3-Chloroacrylic acid degradation	101	Fatty acid elongation in mitochondria
15	Citrate cycle (TCA cycle)	44	Lysine degradation	73	Cyanoamino acid metabolism	102	Fluorobenzoate degradation
16	Endocytosis	45	Nitrogen metabolism	74	beta-Alanine metabolism	103	Geraniol degradation
17	Pyruvate metabolism	46	Porphyrin and chlorophyll metabolism	75	Drug metabolism - cytochrome P450	104	Other glycan degradation
18	DNA replication	47	Tryptophan metabolism	76	Drug metabolism - other enzymes	105	Peptidoglycan biosynthesis
19	Alanine, aspartate and glutamate metabolism	48	Glyoxylate and dicarboxylate metabolism	77	Folate biosynthesis	106	Styrene degradation
20	Arginine and proline metabolism	49	Histidine metabolism	78	Metabolism of xenobiotics by cytochrome P450		
21	Cysteine and methionine metabolism	50	Inositol phosphate metabolism	79	Protein export		
22	RNA polymerase	51	Phosphatidylinositol signaling system	80	Vitamin B6 metabolism		
23	N-Glycan biosynthesis	52	Steroid biosynthesis	81	Aminophosphonate metabolism		
24	Pentose phosphate pathway	53	Lysosome	82	Biotin metabolism		
25	Amino sugar and nucleotide sugar metabolism	54	One carbon pool by folate	83	Limonene and pinene degradation		
26	Fructose and mannose metabolism	55	Lysine biosynthesis	84	Methane metabolism		
27	Basal transcription factors	56	O-Mannosyl glycan biosynthesis	85	Nicotinate and nicotinamide metabolism		
28	Galactose metabolism	57	Riboflavin metabolism	86	Pentose and glucuronate interconversions		
29	Glycerophospholipid metabolism	58	Sphingolipid metabolism	87	Benzoate degradation via CoA ligation		

Table S1.4 A comparison of dynamic and adaptive responses of different organisms with *S. cerevisiae*

We find the organisms that are more distant to *S. cerevisiae* in **Figure 2.2-Figure 2.5**, (Supplementary **Figure S1.2-Figure S1.5**) with respect to some biological process also have phenotypic behavior that is more different from the yeast than those that are predicted to be closer with respect to that process.

Pathway/ Biological Process	Reference behavior in *S. cerevisiae*	Closer Organism	Behavior of process in closer organism	More distant organism	Behavior of process in more distant organism
Cell wall organization	Ideal temperature for Pheromone stimulated growth is between 24º C and 37° C. [**100**, **101**]	------	------	***Candida albicans***	Pheromone causes smoothing, the initial step in the mating process, only in a/a cells expressing the opaque phenotype. Does not grow at 37º C.
	Glycosidase-I does not play role in outer chain formation or mannosylation and elongation of oligosaccharide residues in cell wall organization. [**102**, **103**]	------	------	***Candida glabrata***	Glycosidase-I or Glycosidase-II play important role in mobility of β-N-acetylhexosaminidase to initiate outer-N-chain elongation in cell wall organization.
Meiosis	*S. cerevisiae* has two silent mating cassettes (HML & HMR) and an active MAT locus. The morphogenesis is regulated by MAPK signaling pathway. [**100**, **104-106**]	***Candida glabrata***	*C. glabrata* has two silent mating cassettes and an active MAT locus and undergoes mating-type interconversions via a Ho-type endonuclease, regulated by a MAP kinase cascade.	***Candida albicans***	*C. albicans* has one MAT locus and a mating-type pleiotropic switching event is required for mating to occur.

Table S1.4 [continued…]

Pathway/ Biological Process	Reference behavior in *S. cerevisiae*	Closer Organism	Behavior of process in closer organism	More distant organism	Behavior of process in more distant organism
Meiosis [continued..]	Ideal temperature for Pheromone stimulated growth is between 24º C and 37° C. [**100**, **101**]	------	------	***Candida albicans***	Pheromone causes smoothing, the initial step in the mating process, only in a/a cells expressing the opaque phenotype. Does not grow at 37º C.
	Only diploid cells of *S. cerevisiae* shows bipolar budding pattern. [**107**, **108**]	***Kluyveromyces lactis***	Both haploid and diploid cells of *K. lactis* shows bipolar budding pattern.	***Candida albicans***	Axial budding pattern.
Pseudohyphal growth	Switching from axial to bipolar mode is required for germ tube emission. [**108**]	***Candida albicans***	Switching from axial to bipolar mode is required for germ tube emission because the mechanism requires a polar budding pattern.	***Yarrowia lipolytica***	Switch from axial to bipolar mode is not necessary because *Y. lipolytica* shows budding pattern in haploid and diploid form.
Thiamine metabolism	Thiamine biosynthetic pathway component plays dual role, cooperating with repair mechanisms in mitochondria. [**109-111**]	***Arabidopsis thaliana***	Thiamine biosynthetic pathway component probably plays dual role, cooperating with repair mechanisms.	***Pseudomonas fluorescens***	Thiamine biosynthetic pathway component does not play any role in repair mechanisms.

Table S1.4 [continued…]

Pathway/ Biological Process	Reference behavior in *S. cerevisiae*	Closer Organism	Behavior of process in closer organism	More distant organism	Behavior of process in more distant organism
Non-homologous end-joining (NHEJ) recombination	Haploid strain performs NHEJ efficiently and diploid strain performs NHEJ inefficiently. [**112, 113]**	***Candida glabrata***	------	***Kluyveromyces lactis***	NHEJ transcription is not regulated by cell type (both haploid and diploid show same efficiency in repair**).**
	Illegitimate recombination (IR) by NHEJ pathway occurs at rate of 1-5 transformants/µg **[113]**	***Candida glabrata***	------	***Kluyveromyces lactis***	Illegitimate recombination (IR) by NHEJ pathway occurs 1000 fold faster than in *S. cerevisiae*.
	IR mechanism is based on microhomology and the target site is near to the consensus sequence for TOP1 binding (Topoisomerase-1). [**110, 114, 115]**	***Candida glabrata***	IR mechanism is based on microhomology and the target site is near to the consensus sequence for TOP1 binding.	***Kluyveromyces lactis***	IR mechanism is not based on microhomology and the target site is not specific to the consensus sequence for TOP1 binding. **[14, 59,70]**
	Mitotic and ORF IR are equally compromised by HR (homologous recombination) and NHEJ. **[113]**	***Candida glabrata***	Mitotic and ORF IR are equally compromised by HR and NHEJ.	***Kluyveromyces lactis***	Mitotic (cell cycle error) IR occurs 6 folds more frequently than in ORFs (transcription error).

Table S1.4 [continued…]

Pathway/ Biological Process	Reference behavior in *S. cerevisiae*	Closer Organism	Behavior of process in closer organism	More distant organism	Behavior of process in more distant organism
Isoleucine biosynthesis	1-butanol production flux comes mostly from an intermediate of the isoleucine biosynthesis pathway produced through the use of threonine. [**116-119**]	***Escherichia coli***	1-butanol production flux comes mostly from an intermediate of the isoleucine biosynthesis pathway produced through the use of threonine.	***Laptospira interrogans***, ***Methanococcus jannaschii***, ***Geobacter sulfurreducens***	1-butanol production is only observed when threonine-mediated isoleucine biosynthesis is shut down and pyruvate mediated isoleucine biosynthesis is overexpressed. [5, 29, 56, 68]
One carbon pool by folate	One-carbon pool in cytoplasm required for synthesis of purines, thymidylate and regeneration of methionine. [**120-122**]	***Rattus norvegicus***	One-carbon pool in cytoplasm required for synthesis of purines, thymidylate and regeneration of methionine.	***Arabidopsis thaliana***	One-carbon pool in cytoplasm is only requiring for regeneration of methylation. It is not required for purine synthesis.

2.7.4. Supplementary File containing the ScCOGs

A text formatted file[2], in which, each *S. cerevisiae* proteins of the ScCOGs (see first column NCBI Reference ID), and its orthologs (rows) from the corresponding 704 organisms (columns) are provided. If any corresponding organism does not find ortholog of the reference *S. cerevisiae* protein, it is provided with dashed line.

2.7.5. Supplementary File containing the ScCHGs

A text formatted file[3], in which, each *S. cerevisiae* proteins of the ScCHGs (see first column NCBI Reference ID), and its homologous (rows) from the corresponding 704 organisms (columns) are provided. If any corresponding organism does not find homologous of the reference *S. cerevisiae* protein, it is provided with dashed line.

2.8. Acknowledgments

We acknowledge the reviewer's work and thank him and the editor for helping us focus this work and improve its quality.

2.9. Author Contributions

Conceived and designed the experiments: HK, RA.

Performed the experiments: HK, RA.

Analyzed the data: HK RA EV AS.

Contributed reagents/materials/analysis tools: HK, RA.

Wrote the paper: HA, RA, EV, AS.

[2] The text file is available in the CD with file name "ScCOGs.txt".
[3] The text file is available in the CD with file name "ScCHGs.txt".

Chapter 3. A human centric comparison of eukaryotic proteomes: Implications for the study of human biology

3.1. Abstract

In Chapter 2 we proposed a methodology to assist in the choice of appropriate model organisms for the study of specific biological processes and networks in large classes of organisms. We applied that methodology to analyze the adequacy of *S. cerevisiae* as a model to study various biological phenomena at the molecular level.

In this chapter we apply the same methodology to study the human proteome and compared it to that of other eukaryotes with fully sequenced genomes in order to reveal the unique differences between humans and other organisms at the protein level. A more detailed analysis of the comparison between the proteomes of primates reveals both, proteins that are unique to humans and which primates appear to be more closely related to human with respect to the sets of proteins associated to various biological phenomena.

We find that the proteome of gorilla is functionally closer to that of human than the chimp and monkey proteomes. In addition, at the sequence level, a significant fraction of the proteins of gorilla are more similar to those of human than the corresponding orthologs from chimp proteins. Our analysis also identifies which animals could be good model organism to study of the physiology of different human tissues, such as brain, bone and muscle. We also identify lower eukaryotes that could be good models to study different aspects of human biology. For example, *C. elegans* is likely to be suitable for studying EGFR mediated MAPK pathways regulatory processes.

3.2. Introduction

"*Cogito ergo sum.*" [**123**] With this short sentence Descartes was in a way trying to answer the question of what makes human different from other living beings. In recent years Descartes views have been somewhat disputed by Damasio, which placed more emphasis on the synthesis of reason and emotion [**124**]. In both cases the brain and its activity is placed at the center of what makes us different from other organisms. However, recent studies in animal metacognition make it apparent that some animals are capable of abstract reasoning, tool building and using, and making social and emotional ties that are similar to those made by humans [**125-129**]. Therefore, the question of what makes humans human is still very much under debate.

An answer to the question of what makes us unique is more likely to be found in the combination of physiological and developmental processes that allow human being to exist and survive. Understanding the molecular mechanisms that lead to such uniqueness will require identifying the differences between humans and other organisms at the molecular level. Such identification can only be done by systematically comparing the molecular components of human to those of other species. Given the availability of full genome sequences for more than 1000 different species, comparative genomics allows us to perform these comparisons. In addition, such comparisons will also provide information about the emergence and evolution of the differences one will find [**19**, **130-134**].

Many of the important differences found between the different genomes exist at the regulatory level, in parts of the genomic sequences that do not code for expressed genes [**135-140**]. However, another part of the differences is also bound to be found in the varying set of proteins encoded in each genome and in the functional interactions that occur between the proteins coded in those genes [**141-145**] .

Hence, comparing the complete proteome of humans to that of other organisms with fully sequenced genomes is bound to identify the protein(function)s that are specific to human. In addition, because functional annotations also associate proteins to specific biological processes and circuits, one can also get a measure of the differences between other organisms and humans with respect to those processes and circuits [**146**]. Such a human-centric proteome comparison is likely to identify some of the molecular components and processes that make us what we are.

In addition, it will also identify those organisms that are likely to be more similar to us with respect to important biological processes and circuits, by pinpointing the species with a set of proteins involved in those processes and circuits that is the most similar to the corresponding human set. This has immediate implications for the understanding of human biology and biomedicine. If one knows the set of human proteins involved in specific biological processes or pathologies, one can then identify the organism with the set of proteins that is more similar to that of humans. This organism or set of organisms are then likely to be reasonable models to study that pathology [**147**].

A study such as the one described in the previous paragraph requires performing several tasks. First, one should consider individual protein function. Then, one should consider the processes, pathways, and circuits in which each protein is involved, thus defining the sets of proteins for each process. Afterwards, and because not all organisms have the same quality and exhaustiveness in the functional annotation of their proteomes, one should also have a way to transfer functional annotation from the better annotated organisms to the others. Finally, one should be able to integrate all this information and study how these functions are conserved between the different organisms.

In this paper we performed such a human centric proteome comparison between 55 eukaryotic organisms with fully sequenced genomes. To do so we downloaded different levels of functional annotation for the human proteome, and developed methods to integrate the different levels and transfer that annotation, when needed, to other organisms.

Performing such transfer at a multigenomic scale can only be effectively done through sequence comparisons. Using BLAST, we systematically compared the proteome of humans to that of other eukaryotes with fully sequenced genomes. Through these comparisons we were able to identify the proteins that are unique to humans, and establish ranks of similarity between the sets of proteins involved in different biological processes and circuits in humans and in the other eukaryotes. We identify what is unique of the human proteome with respect to different eukaryotic organisms and clades and pinpoint the most likely eukaryotes to be good model organisms in which to study specific biological processes and phenomena that have biological and biomedical relevance.

3.3. Results

As stated above, the goal of this work is to identify what unique in the human proteome and in its sets of proteins, associated to different biological processes and phenomena. To do so we start by systematically comparing the complete human proteome to that of 54 other eukaryotes with fully sequenced genomes (**Figure 3.1**). This is done by BLASTing the human proteome with each of the other proteomes and separating the results into four main groups: clusters of Absent (***A***) proteins that are unique to each of the organisms with respect to human; clusters of general homologues (***S***), which include proteins that have at least some low sequence similarity to a given human protein; clusters of exclusive Homologues (***H***), which include proteins that only have low sequence similarity to a specific human protein; clusters of general orthologs (***Og***), which include proteins that have strong similarity to a specific human protein in at least a fraction of its sequence; clusters of Domain orthologs (***D***), which have strong similarity to a specific human protein only in a fraction of its sequence; and clusters of Orthologs (***O***), which have strong similarity to a specific human protein over the entire length of both proteins. In addition, clusters of Functional Orthologs (***FO***) were further identified and analyzed within the ***O***-clusters. ***FO***-clusters are subset of the ***O***-cluster that contains only one human protein and at most one protein from each of the other eukaryotes. This protein is deemed to be the most likely functional ortholog of the human protein in the relevant eukaryote, based on similarity of the sequences. Other proteins that pass the filtering procedure to identify likely functional orthologs (see methods) are considered to be paralogs and analyzed in a protein duplication study (see below). The procedure to identify the different types of clusters is summarized in **Figure 3.1**.

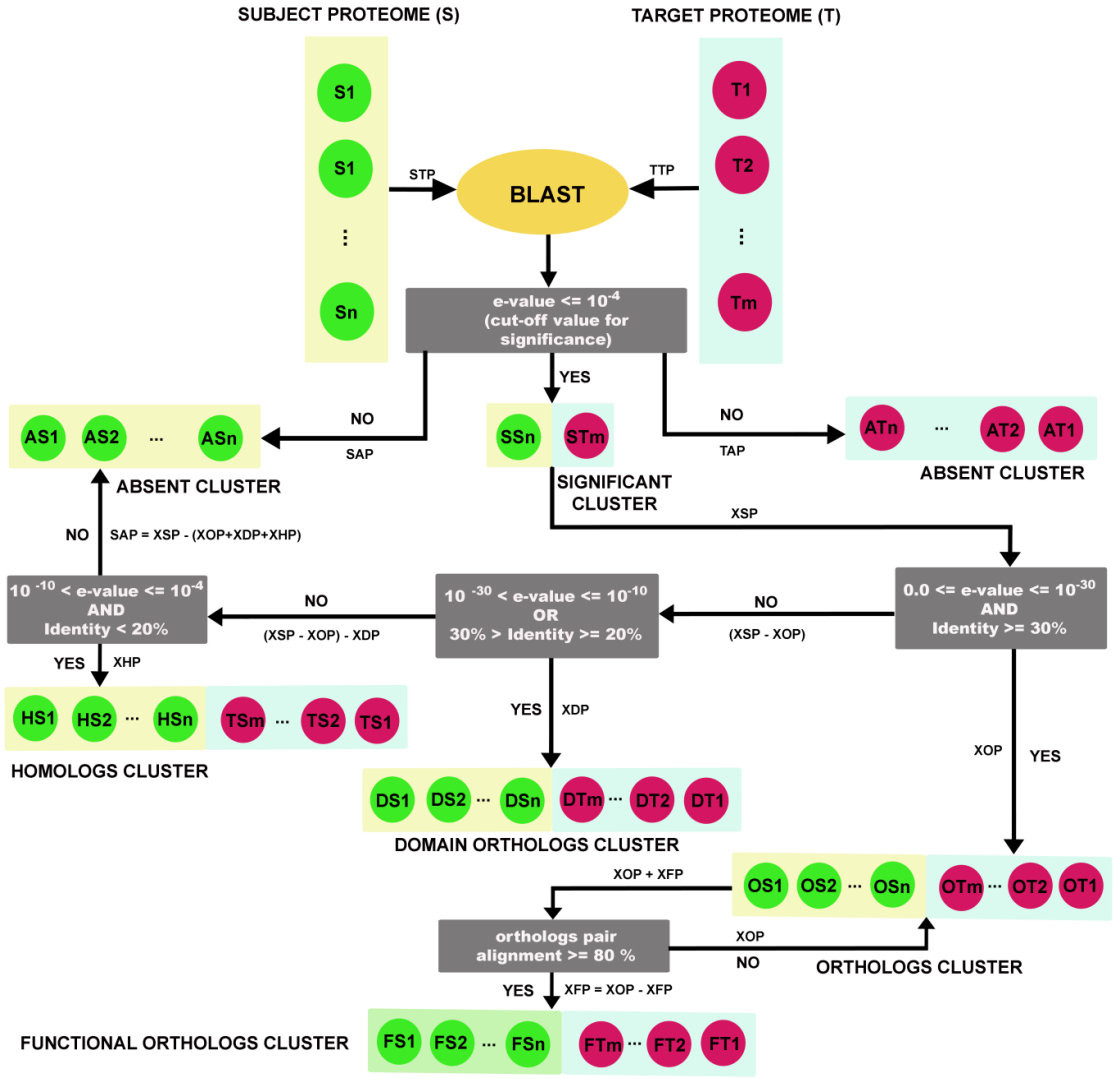


Figure 3.1 **Method for the human centric proteome comparisons** - The human proteome is blasted against the proteome of each organism from Table 3.1. Protein from one proteome that do not match any protein in the other proteome with $\boldsymbol{e-value \geq 10^{-4}}$ are declared as being absent in the later proteome (***A*-clusters**). If a match with $\boldsymbol{e-value < 10^{-4}}$ the pair is added to an ***S*-cluster**. ***S***-clusters are further categorized into ***H*-clusters** (protein pairs that match with $\boldsymbol{10^{-4} < e-value \leq 10^{-10}}$), ***D*-clusters** (protein pairs that match with $\boldsymbol{10^{-10} < e-value \leq 10^{-30}}$ **OR** $\boldsymbol{20\% \leq sequence\ identity < 30\%}$), ***O*-clusters** (protein pairs that match with $\boldsymbol{e-value \leq 10^{-30}}$, **AND** $\boldsymbol{30\% \geq sequence\ identity}$) ***FO*-clusters** (the human-target protein pair with highest sequence similarity in a given ***O*-cluster**, that align $\boldsymbol{\geq 80\%}$ to the total length human protein). See results and methods for details. For each abbreviation terms of the figure, see **Box 3.1** in the next page.

Box 3.1

SAP –Absent Proteins in Subject proteome (S)

SSn **–** 'n' number of significant proteins in Subject.

DSn – 'n' number of Domain ortholog proteins Subject.

HSn – 'n' number of Homologous proteins in Subject.

OSn – 'n' number of Ortholog proteins in Subject.

FSn – 'n' number of Functions ortholog proteins in Subject.

ASn – 'n' number of Absent proteins in Subject.

XSP – Pairs of Significant Proteins, (**X = *SSn* ←→ *STm***).

XOP – Pairs of Ortholog Proteins (**X = *OSn* ←→*OTm***).

XHP – Pairs of Homologous Proteins (**X = *HSn* ←→*HTm***)

TAP **–**Absent Proteins in Target proteome (T)

TSm **–** 'm' number of significant proteins in Target.

DTm – 'm' number of Domain ortholog proteins in Target

HTm – 'm' number of Homologous proteins in Target

OTm – 'm' number of Ortholog proteins in Target

FTm – 'm' number of Functional ortholog proteins in Target

ATm – 'm' number of Absent proteins in Target

XFP – Pairs of Significant Proteins, (**X = *FSn* ←→ *FTm***)

XDP – Pairs of Domain ortholog Proteins (**X = *DSn* ←→*DTm***)

3.3.1. Large Scale Proteome Comparisons

As one would expect, the proteome of human is most similar to that of other vertebrates. The vertebrate with a proteome having the least similarity to that of human is the African clawed frog. 28% of all human proteins are absent in this animal. In contrast, only 7% of the frog's proteins are absent in human. On the opposite end of the scale, and to our surprise, gorilla has the proteome that is the most similar to that of human. 97% of the human proteins have at least one homologue in gorilla. In addition, only 5% of the gorilla proteins are absent in the human proteome.

The non-vertebrate animal with a proteome that is closest to that of human is *Drosophila melanogaster,* while the multicellular animal with the least similar proteome is *Brugia malayi* (filaria). Only 64% of the human proteins have at least one homologue in *B. malayi*, while 40% of this organism's proteins are absent in human.

In decreasing order of proteome similarity with respect to humans, the eukaryotic clades are: Animals, Fungi, Plants and Protista. *Giardia lamblia* is the eukaryotic organism with a proteome that is the most dissimilar to that of human, with only 15% of its proteins being present in human. Interestingly, only 1250 human proteins form ***S***- clusters that include sequences from all 55 eukaryotes (see **Figure 3.2. B**). In addition, 180 proteins are unique to human and absent in all other eukaryotes. Details can be found in **Figure 3.2-Figure 3.5,** and in Supplementary **Figure S2.1.**

(See the Table 3.1 and Figure 3.2 on following pages)

Table 3. 1 Human centric comparison of full proteomes for 54 eukaryotes with fully sequenced genomes. 20125 human proteins were considered. Columns: ***HSP*** – number of human proteins making ***S***-clusters with the proteome of the target organism. ***HTP*** – total number of human proteins. ***TTP*** – total number of proteins in target genome. ***HSP*** – number of human proteins making ***S***-clusters with the proteome of the target organism. ***TSP*** – number of target proteins making ***S***-clusters with the human proteome. ***HFP*** – number of human proteins making ***FO***-clusters with the proteome of the target organism. ***TFP*** – number of target proteins making ***FO***-clusters with the human proteome. ***HOP –*** number of human proteins making ***Og***-clusters with the proteome of the target organism. ***TOP –*** number of target proteins making Og-clusters with the human proteome. ***HDP*** – number of human proteins making ***D***-clusters with the proteome of the target organism. ***TDP*** – number of target proteins making ***D***-clusters with the human proteome. ***HHP*** – number of human proteins making ***H***-clusters with the proteome of the target organism. ***THP*** – number of target proteins making ***H***-clusters with the human proteome. ***HAP*** – number of human proteins making ***A***-clusters with the proteome of the target organism. ***TAP*** – number of target proteins making ***A***-clusters with the human proteome.

Table 3.1

OrgName	Phylum	Domain	HTP	HSP	HFP	HOP	HDP	HHP	HAP	TTP	TSP	TFP	TOP	TDP	THP	TAP
Homo sapiens (human)	Vertebrates	ANIMAL	20125*	20125	20003	20005	120	0	0	25680*	25679	25361	25361	318	0	1
Gorilla gorilla	Vertebrates	ANIMAL	20125	19510	18843	18980	440	90	615	27473	26106	25284	25397	564	145	1367
Mus musculus (mouse)	Vertebrates	ANIMAL	20125	19274	18269	18397	719	158	851	29311	24841	22917	22989	1346	506	4470
Pan troglodytes (chimpanzee)	Vertebrates	ANIMAL	20125	19270	18162	18355	732	183	855	25175	20809	18929	19091	1370	348	4366
Rattus norvegicus (rat)	Vertebrates	ANIMAL	20125	19132	18001	18146	796	190	993	26157	23597	21380	21464	1506	627	2560
Macaca mulatta (rhesus monkey)	Vertebrates	ANIMAL	20125	19088	17858	18027	839	222	1037	23998	21206	18405	18556	2130	520	2792
Bos taurus (cow)	Vertebrates	ANIMAL	20125	18214	17484	17572	641	1	1911	22368	18365	17913	17941	424	0	4003
Canis familiaris (dog)	Vertebrates	ANIMAL	20125	18205	17519	17609	596	0	1920	19835	18813	18235	18273	540	0	1022
Danio rerio (zebrafish)	Vertebrates	ANIMAL	20125	17946	15289	15391	1945	610	2179	27234	25039	21264	21313	2857	869	2195
Monodelphis domestica (opossum)	Vertebrates	ANIMAL	20125	17625	16506	16578	1046	1	2500	19157	16943	16059	16084	858	1	2214
Xenopus tropicalis (western clawed frog)	Vertebrates	ANIMAL	20125	15778	12455	12559	2541	678	4347	8741	8623	8137	8161	414	48	118
Drosophila melanogaster (fruit fly)	Insects	ANIMAL	20125	14526	9427	9507	3638	1381	5599	13827	9434	6482	6494	2312	628	4393
Xenopus laevis (African clawed frog)	Vertebrates	ANIMAL	20125	14473	12367	12438	2033	2	5652	10894	10174	9811	9824	350	0	720
Drosophila pseudoobscura pseudoobscura	Insects	ANIMAL	20125	14468	9327	9405	3650	1413	5657	16071	9819	6592	6608	2488	723	6252
Aedes aegypti (yellow fever mosquito)	Insects	ANIMAL	20125	14214	9313	9395	3470	1349	5911	15428	10769	7208	7225	2795	749	4659
Strongylocentrotus purpuratus (purple sea urchin)	Echinoderms	ANIMAL	20125	13955	10922	10997	2956	2	6170	28881	19243	12924	12982	6261	0	9638
Caenorhabditis elegans (nematode)	Nematodes	ANIMAL	20125	13425	7519	7586	4306	1533	6700	20185	9443	5101	5110	3274	1059	10742
Brugia malayi (filaria)	Nematodes	ANIMAL	20125	12928	7267	7340	4233	1355	7197	11371	6802	4010	4025	2126	651	4569
Neurospora crassa	Ascomycetes	FUNGI	20125	8911	3315	3381	3866	1664	11214	9824	4113	1932	1957	1529	627	5711
Aspergillus nidulans	Ascomycetes	FUNGI	20125	8848	3431	3500	3940	1408	11277	9541	4691	2134	2165	1848	678	4850
Laccaria bicolor	Basidiomycetes	FUNGI	20125	8727	3418	3475	3724	1528	11398	18215	4774	2164	2173	1828	773	13441
Aspergillus oryzae	Ascomycetes	FUNGI	20125	8722	3430	3499	3709	1514	11403	12074	5282	2390	2414	2112	756	6792
Cryptococcus neoformans B-3501A	Basidiomycetes	FUNGI	20125	8653	3320	3376	3819	1458	11472	6500	3767	1918	1932	1339	496	2733
Debaryomyces hansenii	Ascomycetes	FUNGI	20125	8164	2934	2987	3645	1532	11961	6324	3453	1676	1687	1313	453	2871
Schizosaccharomyces pombe (fission yeast)	Ascomycetes	FUNGI	20125	8153	3167	3210	3526	1417	11972	5003	3392	1886	1887	1125	380	1611
Pichia stipitis	Ascomycetes	FUNGI	20125	8078	2913	2959	3600	1519	12047	5816	3414	1673	1682	1270	462	2402
Candida glabrata	Ascomycetes	FUNGI	20125	7743	2730	2773	3439	1531	12382	5191	3117	1577	1584	1126	407	2074
Fusarium graminearum	Ascomycetes	FUNGI	20125	7015	3470	3514	3495	6	13110	11656	4094	2251	2262	1832	0	7562
Cryptococcus neoformans JEC21	Basidiomycetes	FUNGI	20125	6607	3271	3310	3296	1	13518	6273	3015	1900	1906	1107	2	3258
Ustilago maydis	Basidiomycetes	FUNGI	20125	6598	3542	3665	2932	1	13527	6538	2970	1806	1876	1094	0	3568
Magnaporthe grisea	Ascomycetes	FUNGI	20125	6235	2951	2997	3233	5	13890	14010	3802	1967	1979	1822	1	10208
Kluyveromyces lactis	Ascomycetes	FUNGI	20125	5958	2807	2841	3110	7	14167	5335	2553	1551	1554	997	2	2782
Saccharomyces cerevisiae (budding yeast)	Ascomycetes	FUNGI	20125	5918	2812	2850	3061	7	14207	5881	2829	1736	1743	1082	4	3052
Candida albicans	Ascomycetes	FUNGI	20125	4043	1728	1752	2287	4	16082	14629	2093	1228	1231	860	2	12536
Encephalitozoon cuniculi	Microsporidians	FUNGI	20125	3215	1235	1259	1955	1	16910	1996	879	487	488	391	0	1117
Arabidopsis thaliana (thale cress)	Dicotyledons	PLANTS	20125	9561	4144	4196	3746	1619	10564	27361	12364	5684	5696	4746	1922	14997
Oryza sativa japonica (Japanese rice)	Monocotyledons	PLANTS	20125	7375	3967	4006	3368	1	12750	26937	9184	4853	4860	4322	2	17753
Ostreococcus lucimarinus	Green algae	PLANTS	20125	7029	2794	2837	2655	1537	13096	7603	4010	1843	1844	1517	649	3593
Cyanidioschyzon merolae	Red algae	PLANTS	20125	7028	2527	2586	3053	1389	13097	5013	2932	1394	1418	1106	408	2081
Chlamydomonas reinhardtii	Green algae	PLANTS	20125	6275	3458	3500	2772	3	13850	14416	4480	2545	2555	1925	0	9936
Monosiga brevicollis	Choanoflagellates	PROTISTS	20125	8829	4997	5049	3776	4	11296	9171	4926	3218	3224	1700	2	4245
Dictyostelium discoideum (cellular slime mold)	Cellular slime molds	PROTISTS	20125	7685	4084	4120	3561	4	12440	13437	5067	3036	3043	2022	2	8370
Trypanosoma cruzi	Euglenozoa	PROTISTS	20125	6781	2417	2468	2767	1546	13344	19607	7070	2843	2867	3000	1203	12537
Entamoeba histolytica	Entamoeba	PROTISTS	20125	6570	2056	2091	3116	1363	13555	8162	4094	1704	1705	1710	679	4068
Trypanosoma brucei	Euglenozoa	PROTISTS	20125	6544	2284	2333	2704	1507	13581	8712	3666	1552	1567	1480	619	5046
Tetrahymena thermophila	Alveolates	PROTISTS	20125	6307	3011	3037	3264	6	13818	24770	6631	3242	3246	3375	10	18139
Plasmodium yoelii	Alveolates	PROTISTS	20125	6227	1815	1858	2827	1542	13898	7353	2291	919	925	921	445	5062
Trichomonas vaginalis	Parabasalids	PROTISTS	20125	5058	2161	2182	2869	7	15067	59679	7717	3733	3733	3965	19	51962
Leishmania major	Euglenozoa	PROTISTS	20125	4683	2290	2333	2347	3	15442	8265	3015	1650	1664	1349	2	5250
Plasmodium falciparum 3D7	Alveolates	PROTISTS	20125	4327	1881	1906	2416	5	15798	5263	1791	956	961	824	6	3472
Cryptosporidium parvum	Alveolates	PROTISTS	20125	4313	1910	1941	2371	1	15812	3805	1615	874	885	729	1	2190
Cryptosporidium hominis	Alveolates	PROTISTS	20125	4075	1750	1779	2294	2	16050	3885	1403	760	762	641	0	2482
Theileria annulata	Alveolates	PROTISTS	20125	3879	1702	1728	2148	3	16246	3792	1473	811	817	653	3	2319
Theileria parva	Alveolates	PROTISTS	20125	3586	1712	1733	1852	1	16539	4061	1453	816	819	633	1	2608
Giardia lamblia	Diplomonads	PROTISTS	20125	2971	1339	1357	1613	1	17154	6502	1594	659	662	931	1	4908

***The complete downloaded human proteome had 25680 proteins. This number reduces to 25125 non/redundant proteins acording to the November 2011 genome version deposited in NCBI.**

Figure 3.2

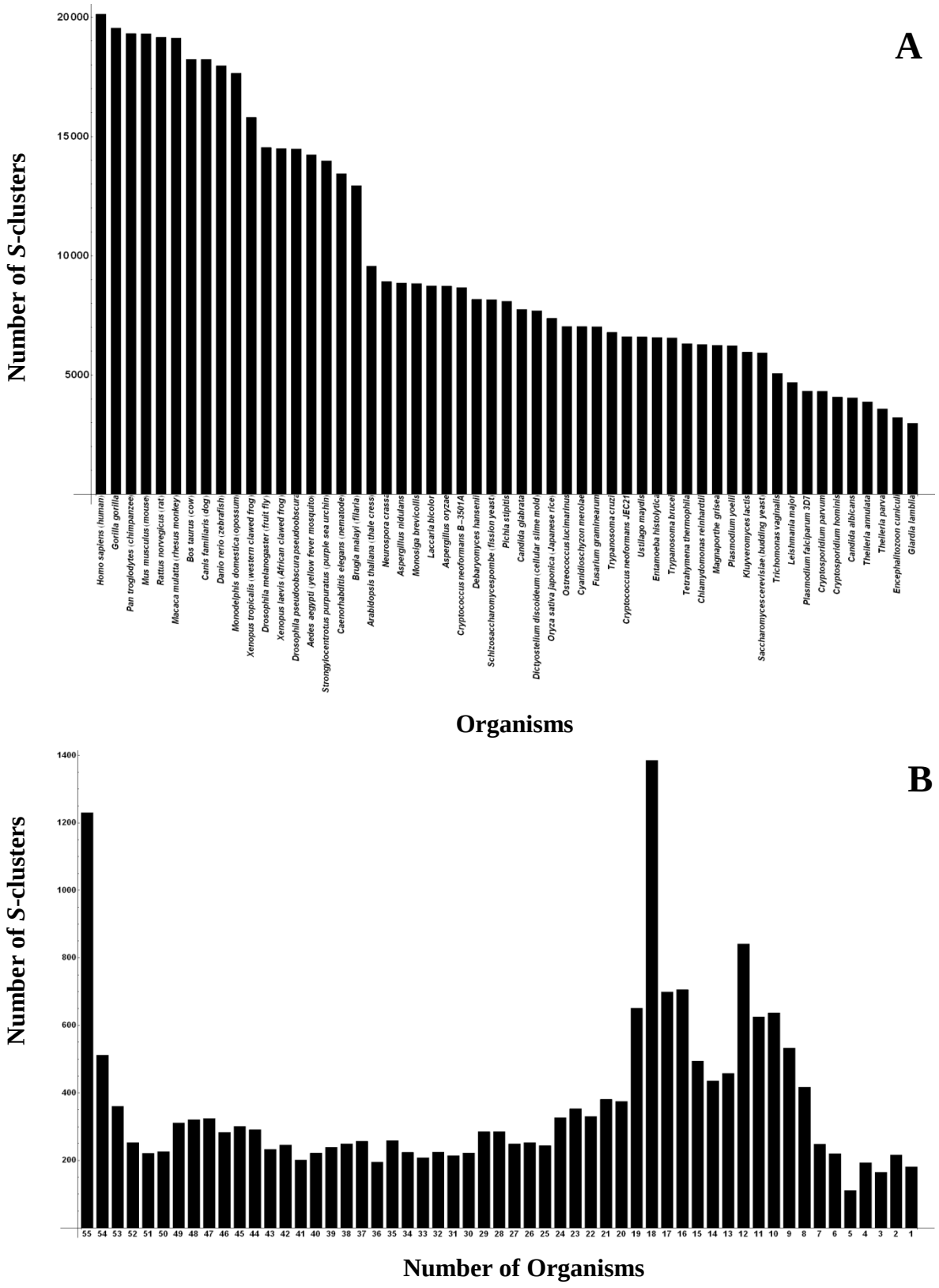


Figure 3.2 Coarse analysis of protein conservation in eukaryotes with fully sequenced genomes. **A** – Number of human centric ***S***-clusters of protein found in each eukaryotic proteome. **B** – Histogram showing how many ***S***-clusters (y-axis) contain sequences from a given number of organisms (x-axis).

Figure 3.2 [continued…]

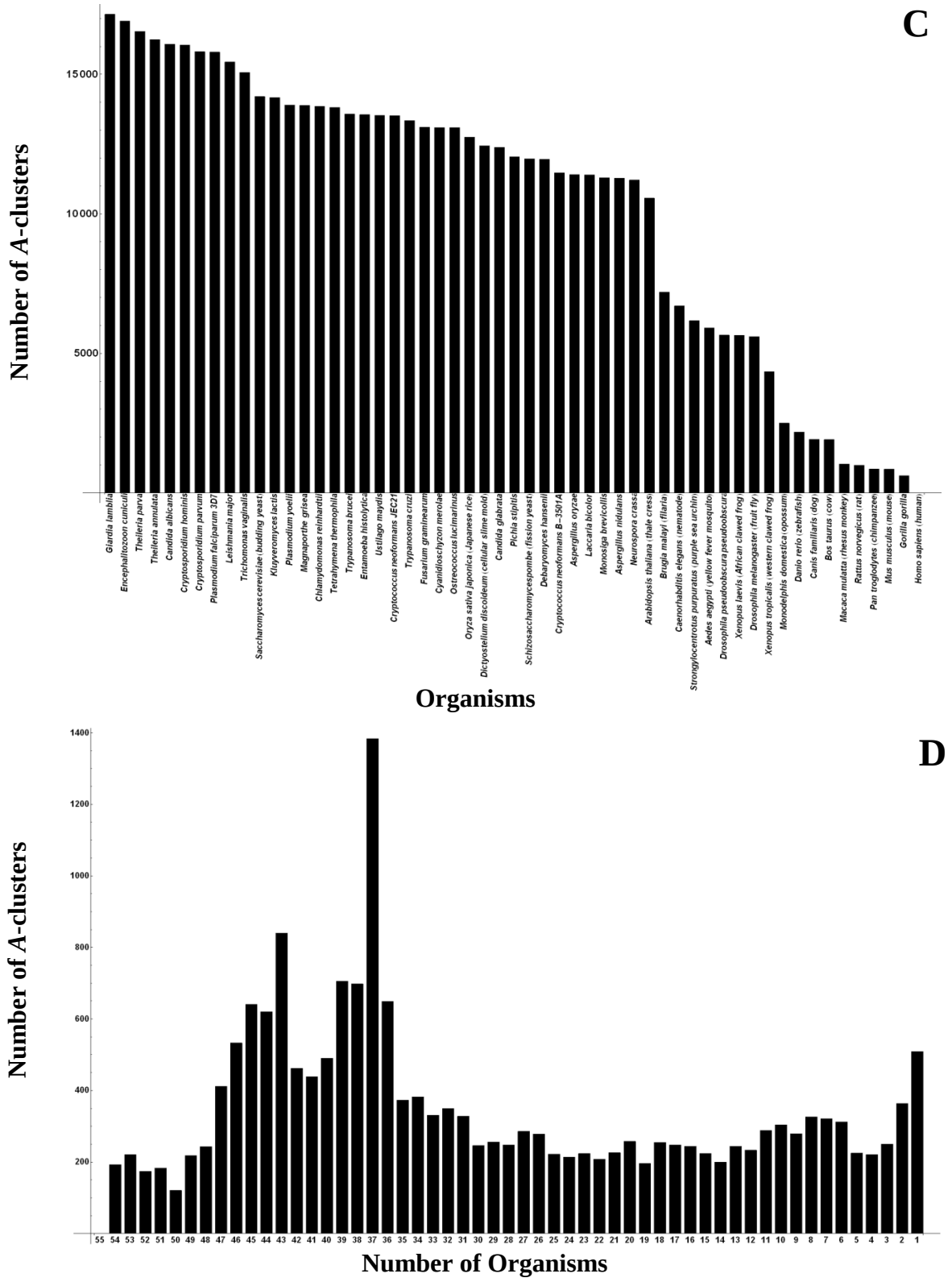


Figure 3.2 Coarse analysis of protein conservation in eukaryotes with fully sequenced genomes. C – Number of human centric ***A***-clusters of protein found in each eukaryotic proteome. **D** – Histogram showing how many ***A***-clusters (y-axis) contain sequences from a given number of organisms (x-axis).

3.3.2. Large Scale Comparison of Clusters of Homologues

To study conservation of proteins at high granularity, we analyze the set of protein pairs in the ***S***- and ***H***-cluster sets. The protein pairs from the first set have a reciprocal BLAST hit with $\boldsymbol{e-value \leq 10^{-4}}$ in the pairwise comparison between the human and the target proteomes. This includes all protein pairs that have some level of sequence conservation, whether that level is high or low. The protein pairs from the second set have a reciprocal BLAST hit with $\boldsymbol{10^{-10} \leq e-value \leq 10^{-4}}$ and thus are only distantly related. Because this criterion is not very stringent, one finds proteins that only have at most the same general function within a given ***H***-cluster. For example, oxidases with different substrate and product specificities would most likely be found in the same ***H***-cluster. Therefore a comparison of ***H***-clusters between proteomes will only provide information about functional conservation at a high level of granularity. A plot of the number of ***S***-clusters vs. the number of organisms included in those clusters is shown in **Figure 3.2**. Approximately 1200 human proteins form ***S***-clusters with all eukaryotes. Another set of approximately 1400 proteins forms ***S***-clusters with all 18 eukaryotes, amongst them, *Monodelphis domestica* (opossum) form the highest number of significant clusters with human proteins (1368), while an additional 6 proteins form ***S***-clusters with all Fungi and Protists.

To analyze the human proteins that only have distant relatives in the other eukaryotes we plot the number of ***H***-clusters *vs.* the number of organisms included in those clusters (**Figure 3.3**). No human protein forms ***H***-clusters with more than 21 eukaryotes and only 2 human proteins form ***H***-clusters simultaneously with the 21 eukaryotes. Approximately 3400 human proteins simultaneously form ***H***-clusters with at least one eukaryote, amongst these, 23% proteins with Protists, 18% proteins with Plants, 14% proteins with Fungi and 44% proteins with Animal forms the ***H***-clusters. The two organisms with the highest number of proteins forming ***H***-clusters with the human proteome are *Neurospora crassa* (627 proteins forming clusters with 1664 human proteins) and *Arabidopsis thaliana* (1922 proteins forming clusters with 1619 human proteins).

Figure 3.3

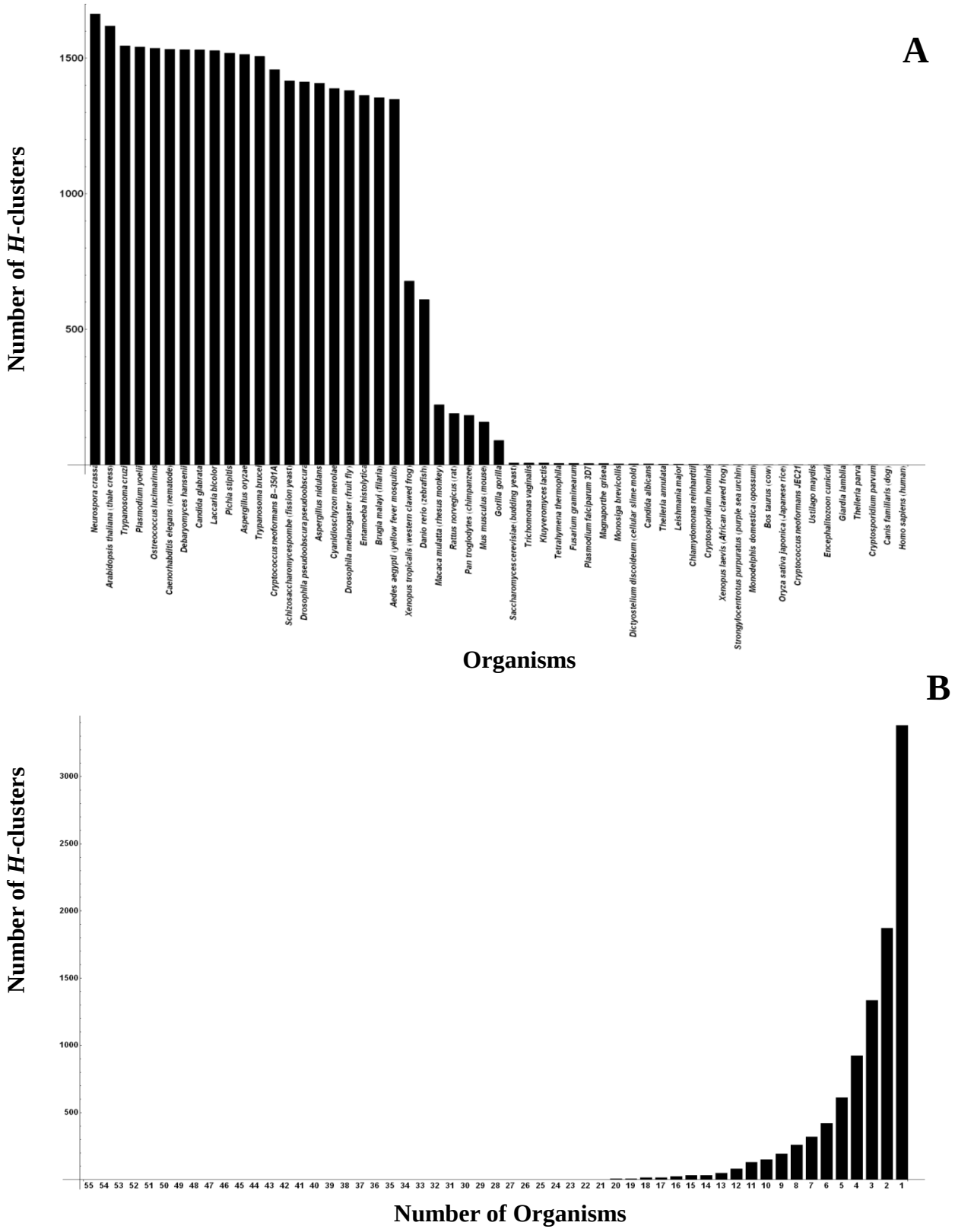


Figure 3.3 Coarse analysis of protein conservation in eukaryotes with fully sequenced genomes. A – Number of human centric ***H***-clusters of protein found in each eukaryotic proteome. **B** – Histogram showing how many ***H***-clusters (y-axis) contain sequences from a given number of organisms (x-axis).

3.3.3. Large Scale Comparison for Clusters of Domain Orthologs

To study conservation of protein function at intermediate granularity, we analyze both the ***Og***- and the ***D***-clusters. ***Og***-clusters include all protein pairs that have a reciprocal BLAST hit with $\boldsymbol{e-value \leq 10^{-10}}$. These protein pairs have a closer sequence relationship than that for the ***H***-cluster pairs. ***D***-clusters include all protein pairs that are matched with $\boldsymbol{10^{-30} \leq e-value \leq 10^{-10}}$ and an identity between 20% and 30%. These exclude all protein pairs that are more likely to be true functional orthologs.

Because the criteria defined for both clusters are more stringent than those defined for the ***S***- and ***H***-clusters, we expect to find proteins that have functions that are somewhat more similar than those found in ***S***- and ***H***-cluster. Therefore a comparison of ***Og***- and ***D***-clusters between proteomes will provide information about functional conservation at an intermediate level of granularity.

A plot of the number of ***Og***-clusters vs. the number of organisms included in those clusters is shown in **Figure 3.4**. Approximately 1200 human proteins simultaneously form ***Og***-clusters with all eukaryotes. Another set of approximately 1400 proteins forms the highest number of of ***Og***-clusters simultaneously with 18 eukaryotes. Approximately 3400 human proteins simultaneously form ***H***-clusters with atleast one eukaryote. 214 human proteins found unique to human that do not form ***Og***-clusters with any of the eukaryotes.

To analyze the human proteins that only have relatives in the other eukaryotes at the ***D***-cluster level, we plot the number of ***D***-clusters vs. the number of organisms included in those clusters (**Figure 3.5**). On average, 13% human proteins form ***D***-clusters with any given eukaryotic organism. *Caenorhabditis elegans* (Nematode) is the organism with the highest fraction of ***D***-clusters, at 21% of the human proteome, whereas gorilla is the organism with the lowest fraction of ***D***-clusters, at only 2% of the human proteome. Only 3 human proteins form ***D***-clusters simultaneously with 49 eukaryotes. 2600 human proteins form ***D***-clusters with at least one eukaryote. Amongst these, 72% form ***D***-clusters with one animal, 4% of the proteins form ***D***-clusters with a Plant, 7% of the proteins form ***D***-clusters a Fungus, and 16% of the proteins form ***D***-clusters a Protist. 24 human proteins do not form ***D***-clusters.

Figure 3.4

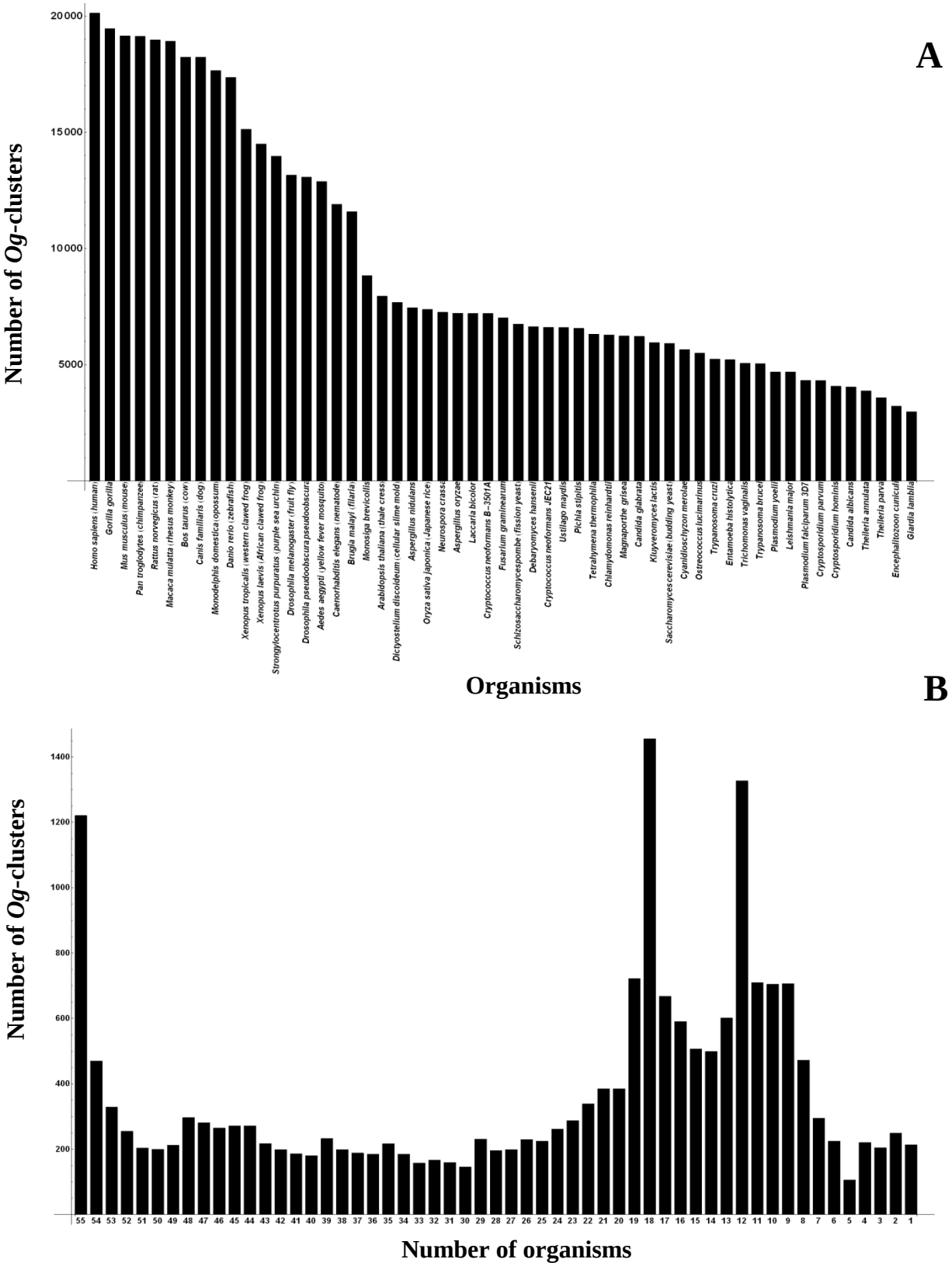


Figure 3.4 **Analysis of protein conservation in eukaryotes with fully sequenced genomes. A –** Number of human centric ***Og***-clusters of protein found in each eukaryotic proteome. **B** – Histogram showing how many ***Og***-clusters (y-axis) contain sequences from a given number of organisms (x-axis).

Figure 3.5

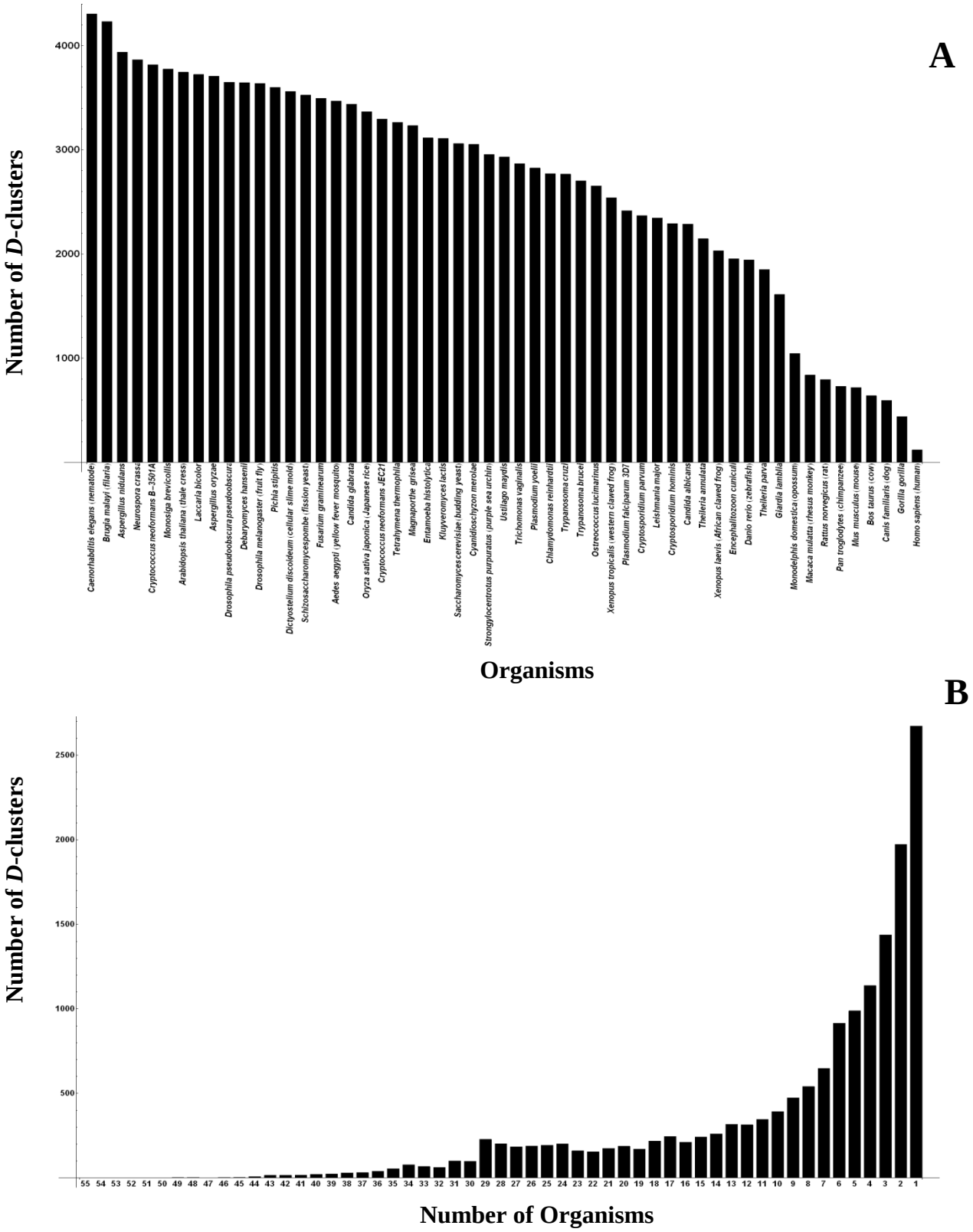


Figure 3.5 Analysis of protein conservation in eukaryotes with fully sequenced genomes. A – Number of human centric ***D***-clusters of protein found in each eukaryotic proteome. **B** – Histogram showing how many ***D***-clusters (y-axis) contain sequences from a given number of organisms (x-axis).

3.3.4. Large Scale Comparison for Clusters of Orthologs

To study conservation of protein function at low granularity, we analyze the ***O***-clusters. ***O***-clusters include all protein pairs that have a reciprocal BLAST hit with $\boldsymbol{e-value \leq 10^{-30}}$ and an identity equal to or larger than 30%. True functional orthologs are included in these clusters.

Approximately 500 human proteins form ***O***-clusters with all eukaryotes, while 380 human proteins form no ***O***-clusters with any other eukaryotes (**Figure 3.6**). Approximately 9% of all human proteins form ***O***-clusters with all animals. Somewhat surprisingly, both gorilla and mouse proteins form a larger number of ***O***-clusters with human proteins than chimp proteins. On the opposite side of the scale, *Giardia Lambia* and *Encephalitozoon cuniculli* form the smallest number of pairs in the ***O***-clusters with the human proteome.

3.3.5. Large Scale Comparison for Clusters of Functional Orthologs

Finally, we study the ***FO***-clusters. These clusters include only those protein pairs that are more likely to be true functional orthologs between humans and the organism of interest (see methods). The same number of human protein form ***O***- and ***FO***-clusters (**Figure 3.6**). The difference is that while ***O***-clusters can include more than one hit between a human protein and the proteome of the eukaryote being analyzed (or vice versa), ***FO***-clusters will include only the best of these hits. ***FO***-clusters can be compared to ***O***-clusters in order to get a picture of functional duplication and evolution between human and the other eukaryotes. We do so in more detail below.

Figure 3.6

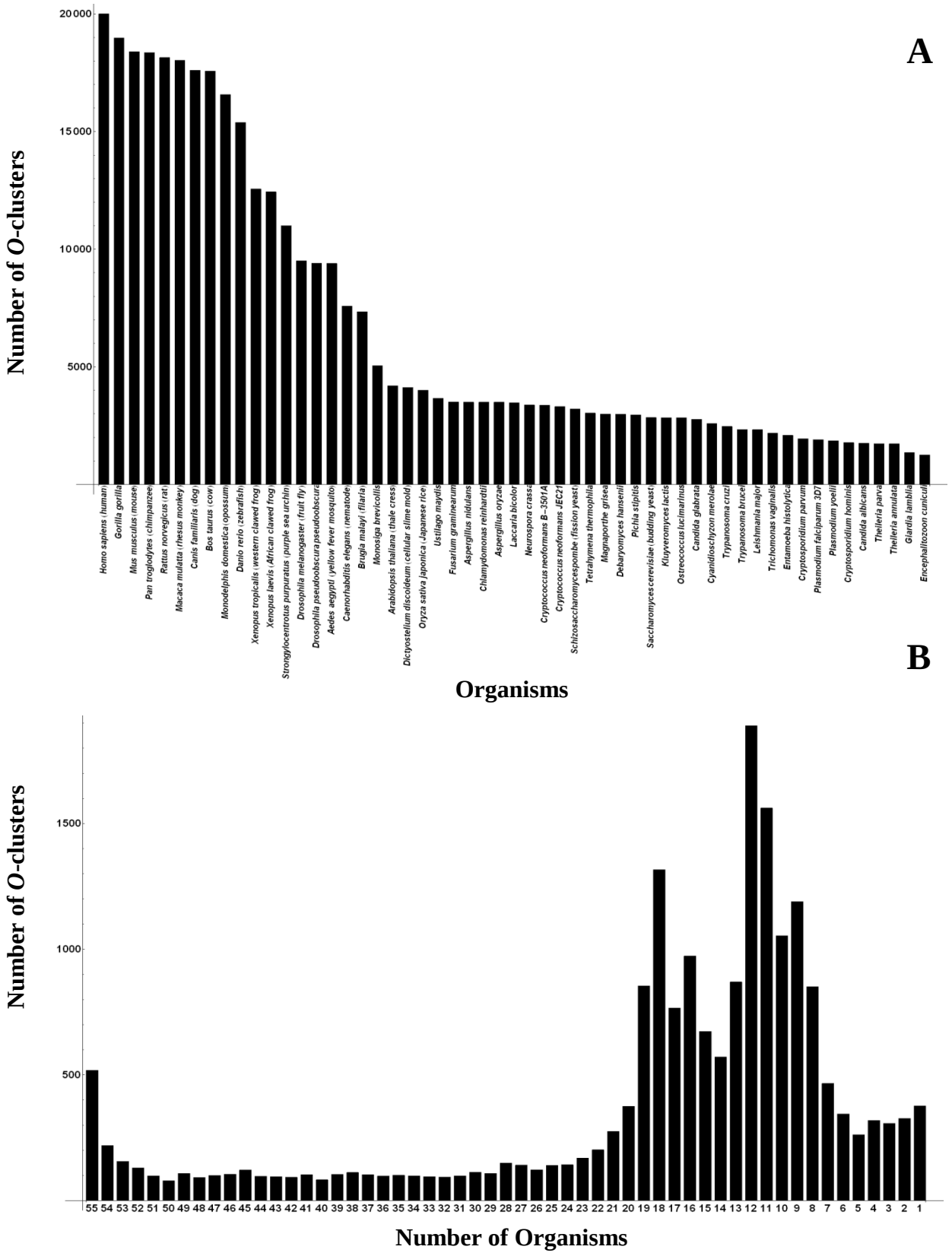


Figure 3.6 **Analysis of protein conservation in eukaryotes with fully sequenced genomes. A** – Number of human centric ***O***-clusters of protein found in each eukaryotic proteome. **B** – Histogram showing how many ***O***-clusters (y-axis) contain sequences from a given number of organisms (x-axis).

3.3.6. Large scale comparative analyses of functional conservation

Because proteins in the ***FO***-clusters are those that are more likely to have the same function in different organisms, these clusters permit studying functional conservation between organisms at a higher confidence level. This enables identifying the best alternative organism(s) on which to study a given process that cannot be studied, for whatever reason, in human. Such identification relies on the assumption that the dynamics and regulation of a given process will be the most similar between organisms whose set of proteins involved in the process is the most similar [147, 148].

To perform this type of analysis we first downloaded the functional annotation for the human proteome with respect to: (I) GO (Gene Ontology) categories i.e., biological processes, molecular functions and localizations [149], (II) catalytic proteins (Enzymes), (III) substrate proteins, that are modified by enzymes in signaling pathways [150], (IV) receptors [151-153], (V) ligands [152], (VI) proteins that are involved in various biological circuits [154], and (VII) human proteins that specifically express in tissues or organs [155].

Then, to compare the differences between the set of proteins involved in a given process or circuit between humans and other eukaryotes we proceeded in the following way. First, we identified the set of proteins involved in the process, both in humans and in the other eukaryotes. Second, each protein function was coded as an element in a vector of functions. Third, the vector of the protein functions was compared between human and the relevant eukaryotic organism of interest, by calculating the Normalized Hamming Distance (***NHD***) between the human and eukaryotic vectors (see methods for details). The smaller this distance is, the more similar the relevant sets of proteins being compared are. Finally, organisms were ordered by increasing order of average ***NHD***, considering all functional categories. Results are summarized in **Figure 3.7**.

Interestingly, gorilla is, on average, the organism with proteins sets associated to specific biological processes that are more similar to those of human. Mice, rats, and chimps also have protein sets that are quite similar to those of human, as have other animals. On the opposite end, *G. lamblia* is the organism in which the smallest fraction of human proteins forms ***FO***-clusters.

The functional category with the highest degree of protein conservation between human and each of the other eukaryotes is that of catalytic proteins, followed by the substrate proteins functional category. Conservation of protein functions in this set is very high among animals, and decreases between animal and more distant phyla, suggesting that the metabolism of animals is, in general terms, quite similar. In fact, the fraction of human proteins from the catalytic protein set that is absent from all other eukaryotes is minimal (<15 % of the total human catalytic proteins; [results not shown]).

In contrast, functional categories receptors, immunologic proteins, and ligands have the lowest degree of protein conservation between human and the other eukaryotes. This argues for the specificity of these proteins and for their importance in making humans different from other organisms.

We now focus our analysis of functional conservation on specific sets of human proteins that are involved in a few important biological processes and categories. These processes and categories were chosen for their involvement in the following important biological phenomena:

(a) Tissue specific roles,

b) Regulating interactions with the environment (ligands and receptors),

c) Association with specific phenotypic responses in health and disease.

To perform that analysis at a high level of confidence we focus on the ***FO***-cluster, because the pairs of proteins in these clusters are the likeliest to be functional orthologs.

3.3.7. Conservation of the human tissue-specific proteome

In this analysis we can a) identify organisms that might be appropriate models to study tissue specific diseases, and b) provide a protein centric view of how tissue-specific functionality evolved in eukaryotes.

For example, proteins that are specific to the olfactory tract have some degree of conservation in a large fraction of eukaryotic organisms (**Figure 3.7** and Supplementary **Figure S2.2**). The proteins annotated as being specific to these tissues are few. Such conservation is further indication that the mechanisms for nutrient detection and environmental recognition evolved from an original rudimentary set of proteins, as suggested in [130].

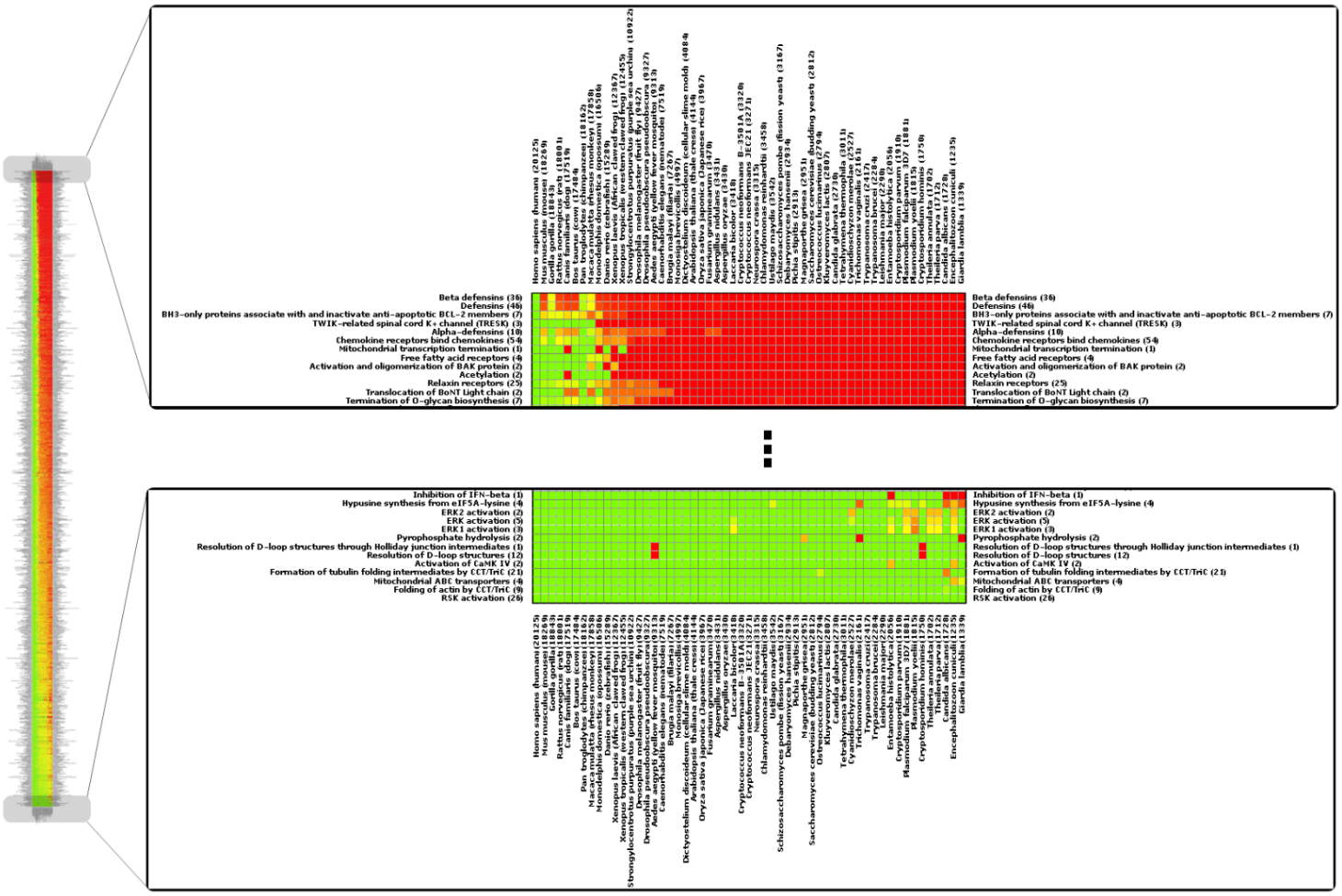


Figure 3.7 Summary of protein conservation for the human proteome associated to specific tissues. Each column summarizes the results for an eukaryotic organism. Each row summarize summarizes the results for a broad functional category of the proteome. The greener the square, the more similar the protein set associated with the functional category of the row in the organism of the column is to the correspondent human protein set. The redder the square, the less similar the protein set associated with the functional category of the row in the organism of the column is to the correspondent human protein set. The complete results can be analysed in Supplementary **Figure S2.2**.

Human "tissues" with large specifically associated proteomes that have the highest degree of conservation in eukaryotes are the lacrimal gland (452 proteins) and tears (459 proteins). The proteins from these "tissues" have functional orthologs in all animals and some lower eukaryotes. Interestingly, functional orthologs for the human proteins CACNA1D, KCNH4, KCNN3, and PRR4 are present in all animals but the African clawed frog.

The first three proteins regulate calcium and potassium channels and their activities, while the last appears to play a role in protecting the human eye [156]. The pattern of conservation for PRR4 suggests that the eye-protection systems may have evolved in frogs in ways that are different from other animals. Consistent with that view is the fact that the

secretory glands frogs have behind their eyes produce venomous liquids that protect them from predation and are absent in vertebrates [157, 158].

The set of proteins associated with most types of reproductive tissues of humans is highly conserved in mammals. For example, PAEP progestagen-associated endometrial protein (gene id 5047) is conserved in all mammals. This is a glycoprotein that contributes for making the uterine environment suitable for reproduction and is used for predicting pregnancy following an IVF (*in-vitro* fertilization) cycle [159]. In contrast, protein sets from the urethra, the human scalp, and the seminal vesicle are only highly conserved in primates.

Overall, Gorilla and monkey are the animals that share the largest fraction of common ***FO***-clusters for each tissue specific protein set. This was somewhat surprising, as we were expecting that role to fall on chimps. On the other hand, as expected, *Giardia lamblia* is the organism that contains less functional orthologs that are specifically annotated in human tissues.

There are some interesting differences between primates with respect to conservation of some individual proteins. For example GHRL (ghrelin/obestatin prepropeptide, gene-ID: 51738) is annotated as being specifically expressed in vena cava. This proteins is absent in chimps and present in the other primates. It regulates growth hormone release and is involved in inhibiting thirst and anxiety [160]. It is also a good marker for studying type-2 diabetes [150], ischemic stroke [161], cardiovascular functions [162] and Rett syndrome [160]. Its absence in chimp could be telling us that the other primates would be better models to study some aspects of those diseases.

There are also proteins that are specific to human alone. For example, the STATH statherin protein (gene-ID: 6779) from the enamel pellicleis absent at the ***FO***-cluster level from all non-human eukaryotes. This protein appears to be crucial in the maintenance of tooth enamel integrity and health. It is involved in lubrication, maintenance of mineral homeostasis, and early phases of microbial colonization [148]. Our result suggest that studies involving this protein should be done in humans because it is absent from other eukaryotes.

3.3.8. Conservation of the human ligand/receptor-specific proteome

A set of ligand and receptor proteins was previously identified and annotated [156]. We further added information of textual annotations to this set of proteins by manually searching

for the terms “receptor”, “receptor associated”, “ligand” and “ligand associated” in the text of annotated human protein entries in NCBI. By analyzing the conservation of these human proteins in other eukaryotes we gain perspective about how different human cells are from those of other eukaryotes with respect to signal sensing and response and about how these processes may have evolved in eukaryotes. This is so because this set of proteins can be taken as a proxy of the mechanism that human cells use to sense and respond to environmental cues.

By and large, proteins involved in MAPK and TOR signaling, in AKT1 apoptotic pathways, in heat shock response, and in myosin mechanic-sensing are conserved in the eukaryotic domain (**Figure 3.8** and Supplementary **Figure S2.3** and **Figure S2.4**). The DOCK family of dedicator proteins, involved in cytokinesis is present in all animals but not in the African frog. Other proteins that are specific to mammals are TLR4-like receptors, some interferons and interleukins, and other immune system related proteins.

The most conserved receptor in all eukaryotes is TNF receptor associated protein 1 (gene-ID: 10131). In contrast tumor necrosis factors have a variety of conservation patterns. TNFRSF17, which plays roles in cell survival and proliferation, B-cell maturation and inflammation [163], is conserved in human, gorilla, mouse and rat, whereas TNFRSF10A and TNFRSF10B are conserved in gorilla, cow, monkey, mouse and chimpanzee. TNFRSF10C and TNFRSF10D, which mediate stress induced apoptosis [164], are present in gorilla, cow, monkey, and chimpanzee. This result suggests an intricate evolutionary pattern for TNF families in vertebrates that could be associated with specialized functions for each TNF. This view is consistent with previously reported results about the diversification of function in TNF families [165]. Some experimental results also support a functional specialization, even within the same TNF family (see for example [166]).

Overall, gorilla is the organism with a larger number of receptor proteins that have ***FO*** in the set of human receptors, followed by mouse, cow, rat, monkey, dog and chimpanzee. Opossum is the mammal with the smaller number of ***FO***-cluster with respect to human receptors. All non-human primates have a fairly similar high number of ***FO*** to human ligand proteins. Taken together, these results seem to indicate that ligand proteins are functionally more conserved than receptor proteins in mammals. Why this is so is unclear, but it indicates that signals (ligands) are more conserved than the mechanisms and pathways through which those signals are transduced. The entry points to the later are the receptors and

Figure 3.8

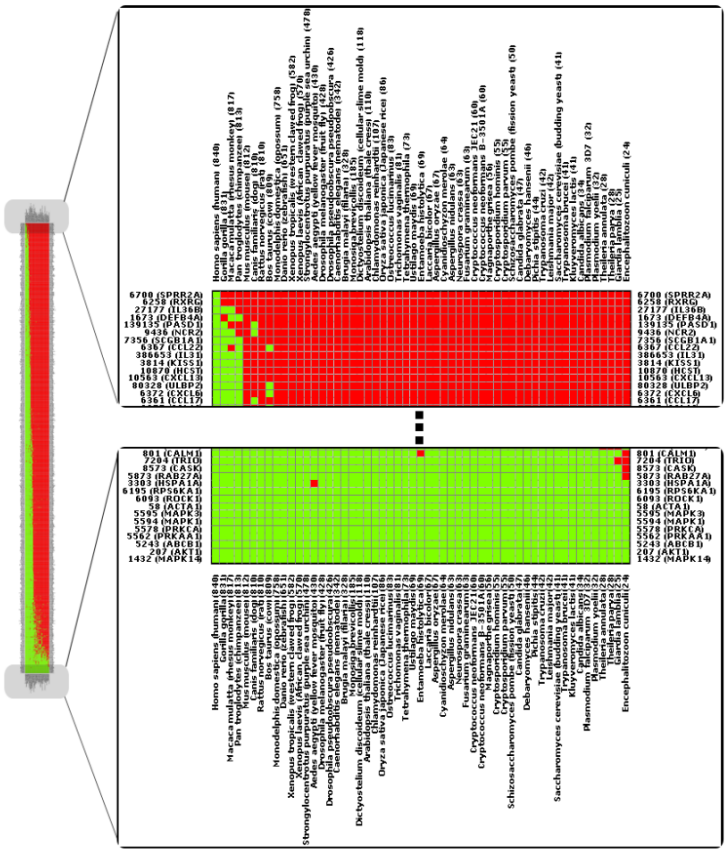


Figure 3.8. A Summary of protein conservation for the human proteome associated to specific ligand functions. Each column summarizes the results for an eukaryotic organism. Each row summarizes the results for a ligand category of the proteome. The greener the square, the more similar the protein set associated with the ligand function category of the row in the organism of the column is to the correspondent human protein set. The redder the square, the less similar the protein set associated with the ligand function category of the row in the organism of the column is to the correspondent human protein set. The complete results can be analysed in Supplementary **Figure S2.3**.

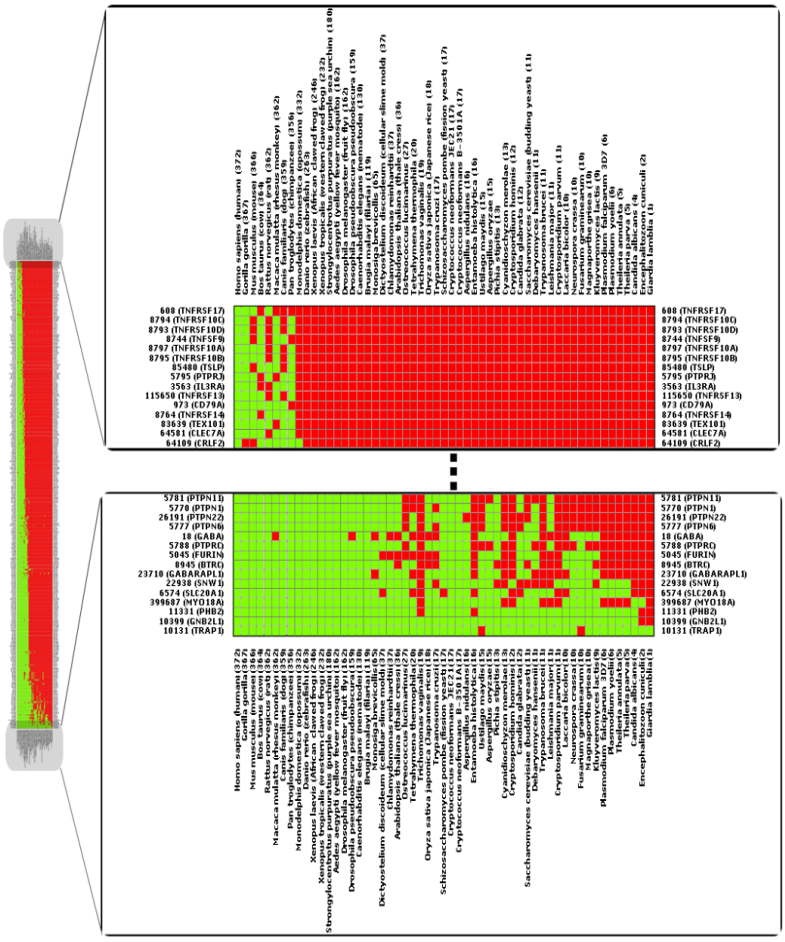


Figure 3.8 B. Summary of protein conservation for the human proteome associated to specific receptor functions. Each column summarizes the results for an eukaryotic organism. Each row summarizes the results for a receptor category of the proteome. The greener the square, the more similar the protein set associated with the ligand function category of the row in the organism of the column is to the correspondent human protein set. The redder the square, the less similar the protein set associated with the receptor function category of the row in the organism of the column is to the correspondent human protein set. The complete results can be analysed in Supplementary **Figure S2.4**.

such increased variation can provide additional fine tuning to a variety of cellular responses. The full comparative study is shown in Supplementary **Figure S2.3** and **Figure S2.4**.

There are some primate specific receptors/ligands. IL-36ß is conserved only in human, gorilla and monkey and absent in chimp, while DEFB4A is conserved in human, monkey and chimp and absent in gorilla. The former protein is involved in in regulation of dendritic and T-cell activity, while the later protein is a *"Defensing"* that is also associated with bone innate immunity [167]. These results suggest that chimp would not be as good a model as the other primates to study regulation of human dendritic and T-cell activity, while gorilla should be disfavored as a model to study some aspects of bone innate immunity. KISS1, a protein that specifically suppresses metastasis in melanomas and breast cancer [168-171] is also specific to primates, probably indicating a recently evolved mechanism for cancer control in this lineage.

There are also human specific receptor/ligands. An example is SPRR2A, a small proline-rich protein 2A that functionally interacts with IL6 and regulates biliary epithelial cell modifications in response to stress [172]. This result suggests that the results of some experiments regarding the effect of IL6 done in mice should be extrapolated to humans with great care.

3.3.9. Conservation of human metabolism-specific proteome

We also analyzed the set of human proteins annotated to specific metabolic pathways [154]. With this comparison we can a) identify organisms that might be appropriate models to study pathway specific diseases, or b) provide a protein centric view of how pathway-specific functionality evolved in eukaryotes.

The proteins involved in many human pathways are highly conserved throughout all eukaryotes. These pathways include "RSK activation", "folding of actin by CCT/Tric", "mitochondrial ABC transporter", "folding of intermediated by CCT/Tric", and "activation of CaMK IV" (**Figure 3.9** and Supplementary **Figure S2.5**). Interestingly, among the highly conserved pathways, functional orthologs of human proteins involved in "resolution of D-loop structure" and "holiday junction intermediates" are absent from yellow fever mosquitos and *Cryptosporidium hominis*. In addition, these proteins are also absent from ***Og***- and ***D***-clusters in the mosquitos. One of the absent proteins is LIG1 ligase I (Gene ID-3978), which is associated to "DNA replication" and "base excision repair functions" in humans. These results suggest that the proteins involved in such functions in mosquitos may have significantly diverged from those of other eukaryotes. If this so, mosquitos could be an interesting model to study these biological processes in order to identify alternative mechanisms for their regulation and execution. In addition, four proteins from "translation synthesis by HREV 1" are absent as functional orthologs in monkey, although they are present in the other higher animals. This pinpoints other animals in which it is likely that some line specific evolutionary events led to different ways of resolving recombination-related DNA repair problems.

Some human pathways appear to be specific for animals, while others are specific for primates. Many of the later are related to immunological responses. This conservation agrees with the tissue-specific analysis made above. For example, proteins from the Defensins related pathways are conserved only in primates. These proteins are engaged in host defense against a broad spectrum of bacterial, fungal and viral pathogens [173-175].

Another surprising result from our analysis is that, overall, there is a larger fraction of human metabolic proteins conserved at the ***FO*** level in mouse, followed by gorilla, rat, dog, cow, chimpanzee and monkey. Full details of the comparison are given for all the eukaryotes in Supplementary **Figure S2.5**.

Figure 3.9

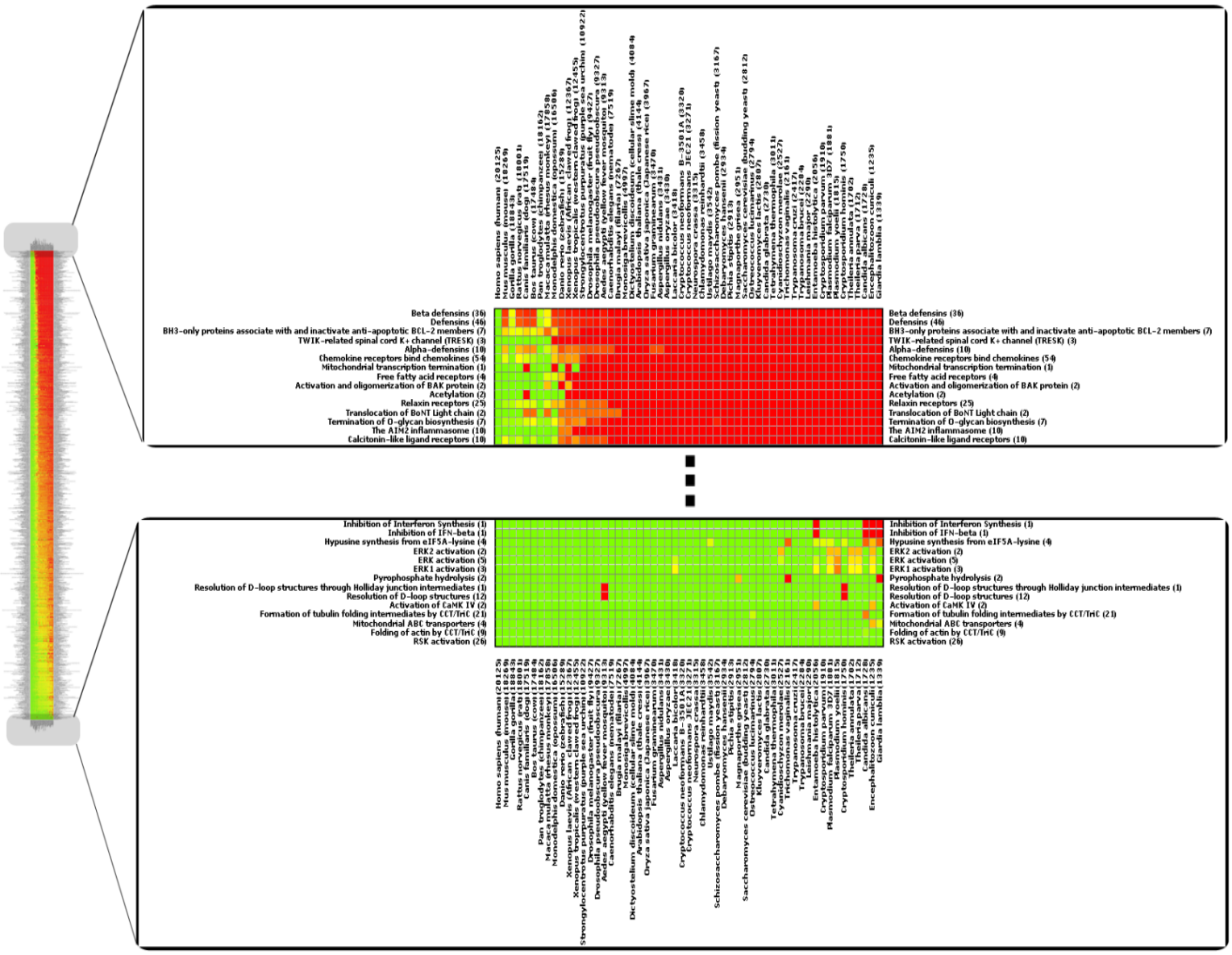


Figure 3.9 Summary of protein conservation for the human proteome associated to specific metabolic functions. Each column summarizes the results for an eukaryotic organism. Each row summarize summarizes the results for metabolic functional category of the proteome. The greener the square, the more similar the protein set associated with the functional category of the row in the organism of the column is to the correspondent human protein set. The redder the square, the less similar the protein set associated with the functional category of the row in the organism of the column is to the correspondent human protein set. The complete results can be analysed in Supplementary **Figure S2.5**.

3.3.10. Conservation of the amino acid biosynthesis-specific proteome between humans and *Saccharomyces cerevisiae*

It is well known that humans depend on diet to supply the essential amino acids. This is so because we do not synthesize them. In contrast, *S. cerevisiae*, an organism that is used as a model to study many processes, can synthesize all twenty amino acids. Therefore, we wanted to compare the generalized amino acid metabolism between the two organisms.

To do so we manually identified the proteins that are involved in amino acid biosynthesis, based on the information from SGD [88, 89] and separated them into two classes: (a) enzymes that are involved in the core catalytic processes, and (b) interacting proteins to the enzymes that are involved in regulation of the catalysis. Then, we searched for each of these yeast proteins in the different levels of protein clusters generated in our study. The human proteins were assumed to have the same function as the yeast protein in the relevant ***FO***-cluster. The reconstruction of the amino acid biosynthesis network in both organisms is shown in **Figure 3.10** and analyzed in **Table 3.2**. Supplementary **Figure S2.6** shows results for the biosynthesis of the individual amino acids.

Overall, 268 *S. cerevisiae* proteins are involved in amino acid biosynthesis. 114 are enzymes and 154 are interaction proteins. 100 out of the 268 proteins are absent in human. Out of these 48 are enzymes and 54 are interaction proteins. 113 yeast proteins are found in the ***O***-clusters (51 enzymes and 62 interaction proteins), while 53 additional yeast proteins are found in the ***DO***-clusters (15 enzymes and 38 interaction proteins). No yeast protein was found in the ***H***-clusters. Our results are in total agreement with what is known about amino acid biosynthesis in humans. For example, 40% of the enzymes for the biosynthesis of methionine are absent in humans.

Figure 3.10

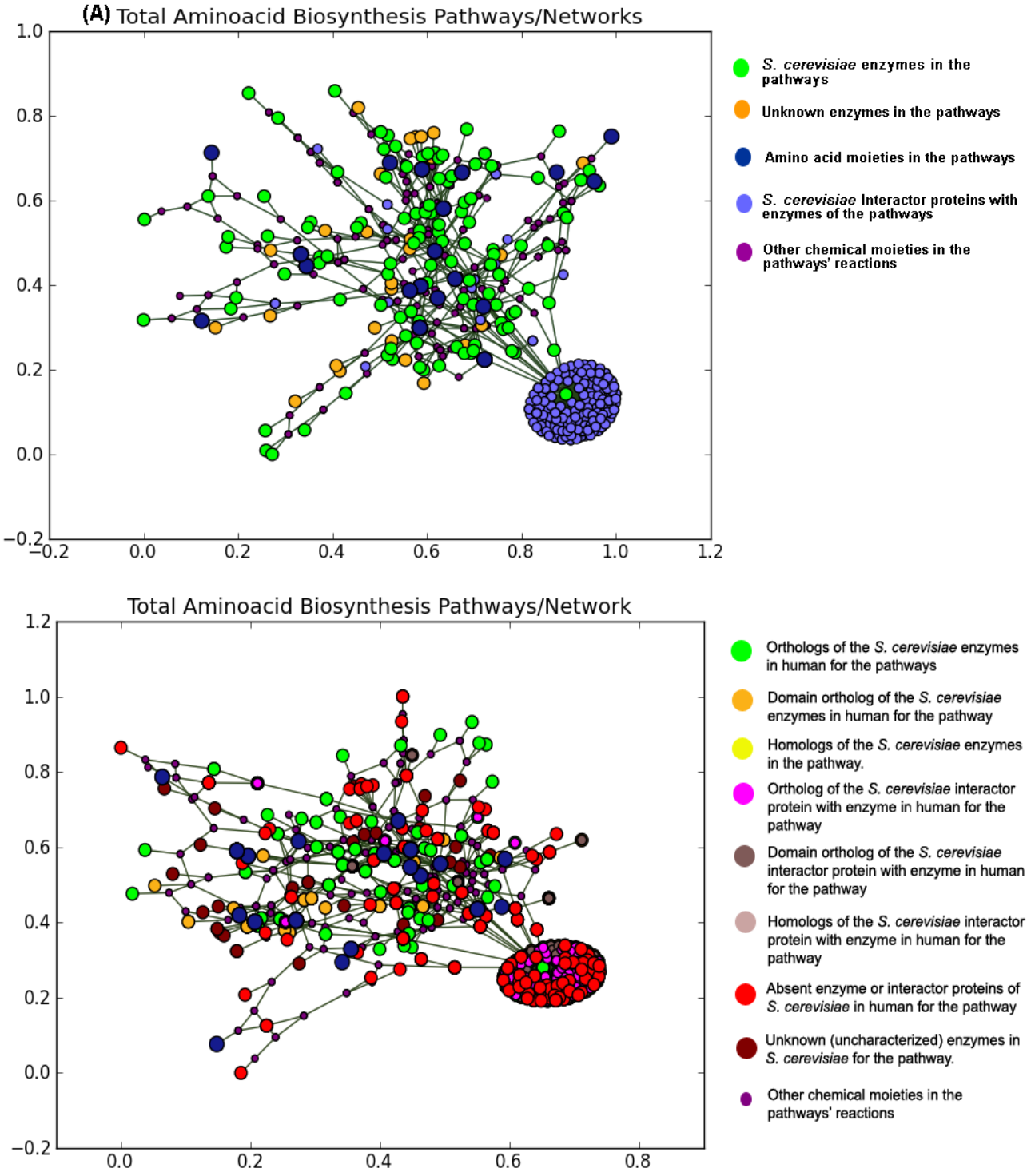


Figure 3.10 **Summary of comparative study of protein conservation between *S. cerevisiae* and human proteomes associated to all 20 amino acid biogenesis pathways.**

Table 3.2 **Frequency of enzymes and their regulatory interactor proteins associated with amino acid biogenesis pathways in *S. cerevisiae* and that found as orthologs and absent in human.**

Comparison of all amino acid biogenesis pathways involved enzymes and interacting proteins between Yeast and Human						
Amino acid biogenesis pathway	*S. cerevisiae*		Human			
	Enzymes	Interactor	Orthologs		Absent	
			Enzymes	Interactor	Enzymes	Interactor
Alanine	3	1	3	0	0	1
Arginine	18	5	15	3	3	2
Asparagine	4	2	4	1	0	1
Aspartate	9	2	4	1	5	1
Cysteine	6	148	6	91	0	57
Glutamate	4	4	3	2	1	2
Glutamine	6	4	5	2	1	2
Glycine	4	1	3	1	1	0
Histidine	7	2	0	1	7	1
Isoleucine	9	4	7	1	2	3
Leucine	10	4	6	2	4	2
Lysine	11	4	4	3	7	1
Methionine	25	153	15	97	10	56
Phenylalanine	9	5	3	3	6	2
Proline	6	3	6	2	0	1
Serine	14	3	13	2	1	1
Threonine	6	5	1	5	5	0
Tryptophan	8	8	0	5	8	3
Tyrosine	11	6	4	4	7	2
Valine	7	5	4	3	3	2
Total	**177**	**369**	**106**	**229**	**71**	**140**

3.3.11. Conservation of developmental proteins

Our study also permits analyzing the role of the protein complement of man in making us different from other organisms, as opposed to the role of differences at the genome sequence and gene expression levels. For such an analysis we focus on the ***FO***-clusters for protein that are annotated as participating in the development of some of the tissues and organs that do have large phenotypical differences between us and other animals: cancer associated proteins, immune system, bone, muscle and brain.

Some proteins from the energy metabolism (UQCR10, UQCR11), cell adhesion (TIMM8A) and proliferation (RXRG, FOXP1), and circadian rhythms (STRA13) appear to be unique to humans, ***FO***-wise. Thus, these proteins make promising targets to study and

identify protein-dependent differences between human and other animals, rather than regulatory dependent differences.

For example, RXRG is a retinoid X nuclear receptor (RXR) family member, mediating anti-proliferative effect of retinoic acid (RA) [176, 177]. RXR proteins appear to have evolved in vertebrate through 2 rounds of duplications [178]. Although other RXR family members have ***FO***-clusters in primates, the human RXRG only has ***H***-clusters, showing similarity with respect to the other family member only in the first 17 amino acids of the protein. Thus, our results suggest that extrapolating the specific role of RXRG in human brain development from the roles of other RXR family members [179] should be done with care. Another example, FOX1P, has variable roles and it either promotes or suppresses tumor progress in different cancers [180]. This protein has recently been shown to mediate regulation of miRNA processing in response to cytokines [180]. Our results suggest that this role might have uniquely evolved in the human lineage.

TMSB4X, USMG5, PLN and SLN (muscle), STATH and CEMP1 (teeth/bone), HMHB1, CD24 and CD52 (immune system) are proteins that are also specific to humans (supplementary figures S6-S8). This fact suggests that these proteins may have unique contributions to the developmental events that differentiate humans from other animals. Such an interpretation is consistent with results found in [181], where the authors report major differences in the receptors of killer T-cells between human and other primates.

Some proteins, such as Phospholamban (PLN), are common to all primates. This protein mediates the ß-adrenergic effect and has role in heart failure associated with dilated cardiomyopathy etiology [182], suggesting that primates could be a good model to study this type of disease. The complete list of these proteins can be seen in Supplementary **Figure S2.7, Figure S2.8** and **Figure S2.9**.

3.3.12. Comparative analysis of functional duplication

As stated above, the eukaryotic proteins forming ***FO***-clusters are those that are deemed more likely to be functional orthologs to a corresponding human protein. Because of this, it is important to understand how such orthologs are conserved and/or have been duplicated during evolution. To perform such an analysis we further divided the ***FO***-clusters into four protein groups. The first group is that of proteins that have a single copy in each organism (One to one group ***[O-O]***). The second group is that of protein that have a single

copy in humans and more than one copy in the other eukaryote (One to many group ***[O-M]***). The third group is that of protein that have more than one copy in humans and a single copy in the other eukaryote (Many to one group ***[M-O]***). The fourth group is that of protein that have more than one copy in both humans and in the other eukaryote (Many to many group ***[M-M]***).

37% of all human proteins form ***O-O***-clusters with other eukaryotes. Based on these clusters, the organism that has the most similar pattern of protein conservation with respect to humans is the chimp (**Figure 3.11. A**). This analysis also reveals that only 25% of all human proteins have no paralogs in the human proteome. 22% of the human proteins form ***O-O***-clusters with the proteome of gorilla, 26% of the human proteins form ***O-O***-clusters -with the proteome of *M. mullatta*, and 37% of the human proteins form ***O-O***-clusters -with the proteome of chimp.

Approximately 10% of the human proteome only forms ***O-M***-clusters. Analyzing these clusters shows that gorilla has the highest number of duplicated proteins with respect to unique proteins of human (**Figure 3.11. B**).

Approximately 63% of the human proteome form ***M-M***-clusters with at least one eukaryote. Gorilla is again the organism with the highest similarity to human, when these clusters are analyzed, with chimp coming in at a close second (**Figure 3.11. D**). 62% of the gorilla proteome is involved in ***M-M***-clusters as opposed to 57% in chimp.

We find that the functional categories of proteins that have patterns of orthology and duplication that are more specific to human are Receptors and Immunological proteins. 28% of the human receptors and 52% of the human immunological proteins form no ***FO***-clusters with other organisms. Transcription factors, proteins involved in brain development, muscle development, and ligand proteins also have patterns of orthology and duplication that is very specific to human.

Figure 3.11

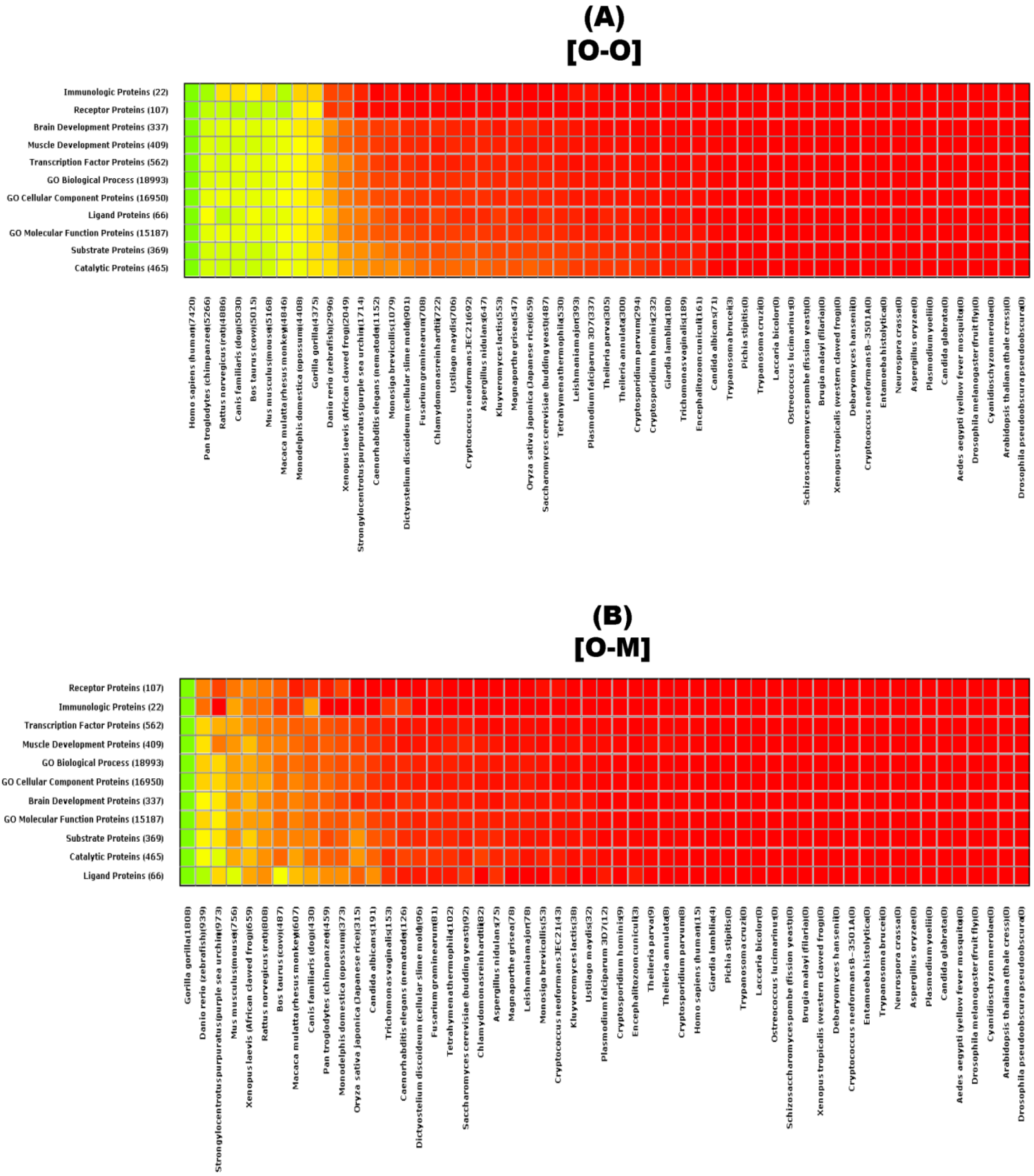

(A)
[O-O]
Immunologic Proteins (22)
Receptor Proteins (107)
Brain Development Proteins (337)
Muscle Development Proteins (409)
Transcription Factor Proteins (562)
GO Biological Process (18993)
GO Cellular Component Proteins (16950)
Ligand Proteins (66)
GO Molecular Function Proteins (15187)
Substrate Proteins (369)
Catalytic Proteins (465)
Homo sapiens (human)(7420)
Pan troglodytes (chimpanzee)(5266)
Rattus norvegicus (rat)(4886)
Canis familiaris (dog)(5030)
Bos taurus (cow)(5015)
Mus musculus (mouse)(5168)
Macaca mulatta (rhesus monkey)(4846)
Monodelphis domestica (opossum)(4408)
Gorilla gorilla(4375)
Danio rerio (zebrafish)(2996)
Xenopus laevis (African clawed frog)(2049)
Strongylocentrotus purpuratus (purple sea urchin)(1714)
Caenorhabditis elegans (nematode)(1152)
Monosiga brevicollis(1079)
Dictyostelium discoideum (cellular slime mold)(901)
Fusarium graminearum(708)
Chlamydomonas reinhardtii(722)
Ustilago maydis(706)
Cryptococcus neoformans JEC21(692)
Aspergillus nidulans(647)
Kluyveromyces lactis(553)
Magnaporthe grisea(547)
Oryza sativa Japonica (Japanese rice)(659)
Saccharomyces cerevisiae (budding yeast)(487)
Tetrahymena thermophila(530)
Leishmania major(393)
Plasmodium falciparum 3D7(337)
Theileria parva(305)
Theileria annulata(300)
Cryptosporidium parvum(294)
Cryptosporidium hominis(232)
Giardia lamblia(180)
Trichomonas vaginalis(189)
Encephalitozoon cuniculi(161)
Candida albicans(71)
Trypanosoma brucei(3)
Pichia stipitis(0)
Trypanosoma cruzi(0)
Laccaria bicolor(0)
Ostreococcus lucimarinus(0)
Schizosaccharomyces pombe (fission yeast)(0)
Brugia malayi (filaria)(0)
Xenopus tropicalis (western clawed frog)(0)
Debaryomyces hansenii(0)
Cryptococcus neoformans B-3501A(0)
Entamoeba histolytica(0)
Neurospora crassa(0)
Aspergillus oryzae(0)
Plasmodium yoelii(0)
Candida glabrata(0)
Aedes aegypti (yellow fever mosquito)(0)
Drosophila melanogaster (fruit fly)(0)
Cyanidioschyzon merolae(0)
Arabidopsis thaliana (thale cress)(0)
Drosophila pseudoobscura pseudoobscura(0)
(B)
[O-M]
Receptor Proteins (107)
Immunologic Proteins (22)
Transcription Factor Proteins (562)
Muscle Development Proteins (409)
GO Biological Process (18993)
GO Cellular Component Proteins (16950)
Brain Development Proteins (337)
GO Molecular Function Proteins (15187)
Substrate Proteins (369)
Catalytic Proteins (465)
Ligand Proteins (66)
Gorilla gorilla(1808)
Danio rerio (zebrafish)(939)
Strongylocentrotus purpuratus (purple sea urchin)(973)
Mus musculus (mouse)(756)
Xenopus laevis (African clawed frog)(659)
Rattus norvegicus (rat)(808)
Bos taurus (cow)(487)
Macaca mulatta (rhesus monkey)(607)
Canis familiaris (dog)(430)
Pan troglodytes (chimpanzee)(459)
Monodelphis domestica (opossum)(373)
Oryza sativa Japonica (Japanese rice)(315)
Candida albicans(191)
Trichomonas vaginalis(153)
Caenorhabditis elegans (nematode)(126)
Dictyostelium discoideum (cellular slime mold)(96)
Fusarium graminearum(81)
Tetrahymena thermophila(102)
Saccharomyces cerevisiae (budding yeast)(92)
Chlamydomonas reinhardtii(82)
Aspergillus nidulans(75)
Magnaporthe grisea(78)
Leishmania major(78)
Monosiga brevicollis(53)
Cryptococcus neoformans JEC21(43)
Kluyveromyces lactis(38)
Ustilago maydis(32)
Plasmodium falciparum 3D7(12)
Cryptosporidium hominis(9)
Encephalitozoon cuniculi(3)
Theileria parva(9)
Theileria annulata(8)
Cryptosporidium parvum(8)
Homo sapiens (human)(15)
Giardia lamblia(4)
Pichia stipitis(0)
Trypanosoma cruzi(0)
Laccaria bicolor(0)
Ostreococcus lucimarinus(0)
Schizosaccharomyces pombe (fission yeast)(0)
Brugia malayi (filaria)(0)
Xenopus tropicalis (western clawed frog)(0)
Debaryomyces hansenii(0)
Cryptococcus neoformans B-3501A(0)
Entamoeba histolytica(0)
Trypanosoma brucei(0)
Neurospora crassa(0)
Aspergillus oryzae(0)
Plasmodium yoelii(0)
Candida glabrata(0)
Aedes aegypti (yellow fever mosquito)(0)
Drosophila melanogaster (fruit fly)(0)
Cyanidioschyzon merolae(0)
Arabidopsis thaliana (thale cress)(0)
Drosophila pseudoobscura pseudoobscura(0)

Figure 3.11 [continued…]

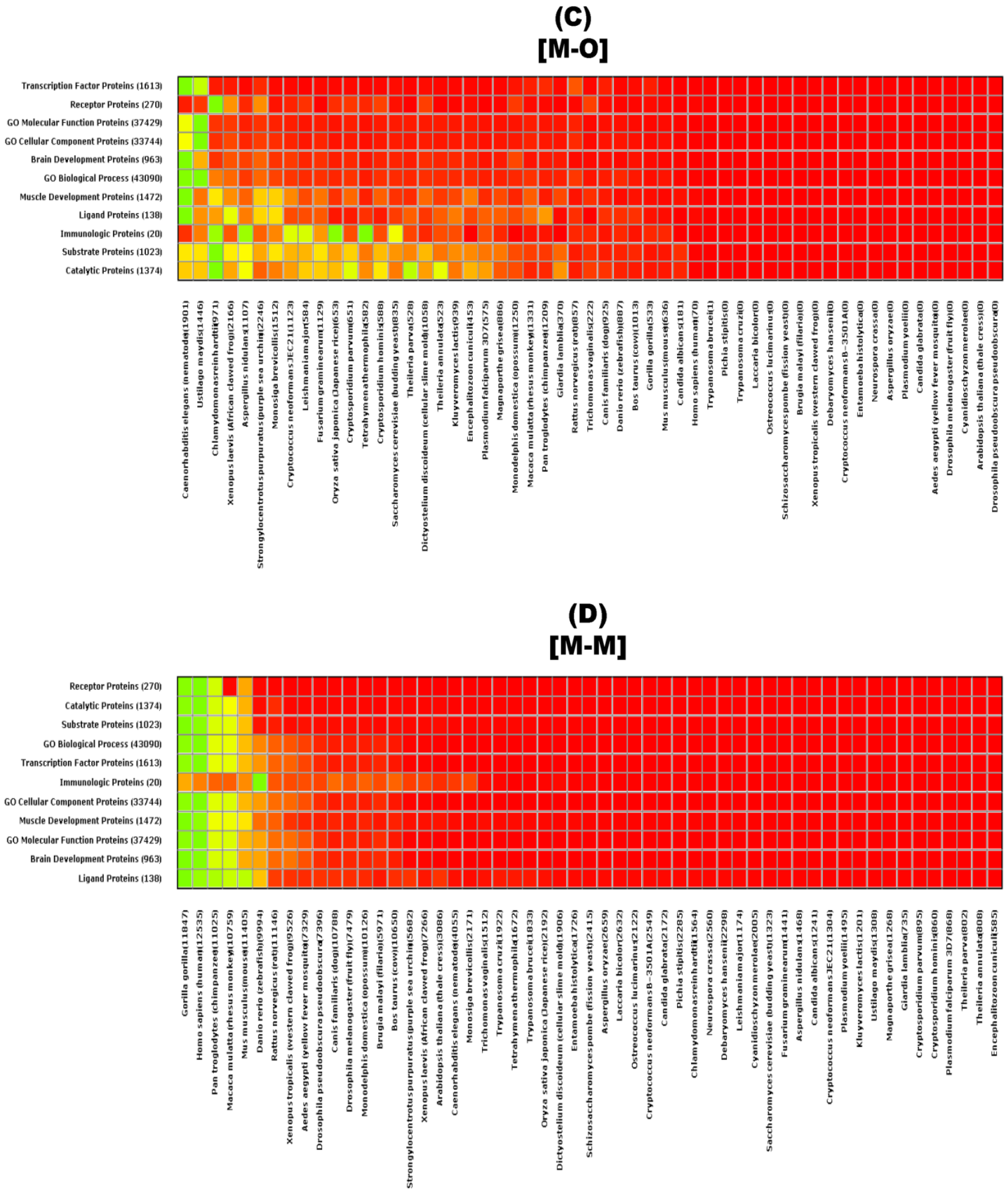


Figure 3.11 Summary of protein conservation of unique and duplicated proteins in *FO*-clusters. Each column summarizes the results for an eukaryotic organism. Each row summarize summarizes the results for a broad functional category of the proteome. The greener the square, the more similar the protein set associated with the functional category of the row in the organism of the column is to the correspondent human protein set. The redder the square, the less similar the protein set associated with the functional category of the row in the organism of the column is to the correspondent human protein set. **A** – ***O*-*O***-clusters. **B** – ***O*-*M***-clusters. **C** – ***M*-*O***-clusters. **D** – ***M*-*M***-clusters.

Macaca mulatta stands out as the primate where less duplication of receptor proteins occurred, when compared to the human proteome. That primate has the highest number of receptor ***O*-*O***-clusters and the lowest number of receptor ***M*-*O***-, ***O*-*M***-, and ***M*-*M***-clusters. A similar statement can be made about gorilla in the immunological protein category.

3.3.13. Conservation study of HIV-Tat regulated human proteins with the FO clusters of eukaryotes

As a final example of the possibilities for this type of analysis, we focus on the proteins that are either regulated by or regulate the HIV-Tat protein in human (***TAT***-Set). This set of proteins was downloaded from NCBI [183] and combined with the experimental ***TAT***-human interactome data previously published [184]. ***TAT*** is a protein that binds to various host proteins, indirectly and directly causing a diversity of post translational modifications in those proteins. These modifications lead to strong increases in the level of transcription of HIV dsDNA, which facilitates spreading of the infection. In addition, Tat appears to be exported by HIV infected cells and have a role in the HIV disease process [185, 186].

The proteins involved in the HIV-Tat regulated processes were mapped to the ***FO***-clusters of all the eukaryotes (**Figure 3.12**). The conservation of proteins in the different categories is quite extensive in most vertebrates. In other eukaryotes, a smaller number of these proteins is conserved. A striking result regards the 16 subunits of human tubulin. When ***Tat*** binds tubulin, it leads to disruption of microtubule polymerization. All subunits are fully conserved in all eukaryotes but *Candida albicans*. *C. albicans* is a common intestinal fungus that invades mucosal tissues and becomes an opportunistic parasite in immune compromised hosts [187-189]. A protein that is absent in human, DUR31, is required for establishment of *C. albicans* microcolonies after mucosal invasion [188]. The absence of human tubulin orthologs together with the absence of DUR31 in humans makes us think that, by studying the process of mucosal invasion by *C. elegans* one could gain a better understanding of what happens upon disruption of tubulin polymerization by Tat in Humans. In addition, it is also known that microtubules in *C. albicans* and *S. cerevisiae* hyphae use similar tubulin subunits.

These are different from the tubulin that *S. cerevisiae* uses in normal growth situations [190, 191]. Hence, studying the transition between normal growth and hyphal growth in *S. cerevisiae* could provide useful information regarding what happens in human cells after tubulin disruption by Tat binding.

Another interesting finding is that the importins that mediate nuclear import of Tat to the nucleus are absent from the genomes of *T. vaginalis* and *E. histolytica*. *T. vaginalis* is believed to disrupt the urogenital monolayer and activate local immune T-cell load in order to increase viral replication [192]. This could provide mechanistic explanation to the observation that patients infected with STDs are more likely to become infected by HIV upon contact with an HIV-carrying partner [193]. The parasite could be using HIV's importin proteins. Furthermore, 173 and 114 various human proteins respectively up-regulate and down-regulate under the effect of HIV-Tat. Pathways associated with the genomic deregulation may be an important area from the drug discovery and diagnostic point of view during AIDS development in human. For these, all vertebrate appear to be good model organisms for human.

187 human proteins were experimentally determined to bind to Tat from Figure 3.13. Out of these, 183 interact with HIV in the nucleus. Surprisingly, gorilla has the highest number of proteins (142, of which 111 belong to the nuclear interaction subset) that form ***FO***-clusters with those in the ***TAT***-set. In contrast only 113 chimp proteins (52 nuclear) and 98 *M. mulatta* proteins (33 nuclear) are present in the ***FO***-clusters for the 187 human proteins. A more detailed analysis of the conservation of the ***TAT***-Set in primates is provided in Supplementary **Figure S2.10**. This analysis suggests that that gorilla is more adequate as a model to study the role of Tat in HIV infection than chimps.

Figure 3.12

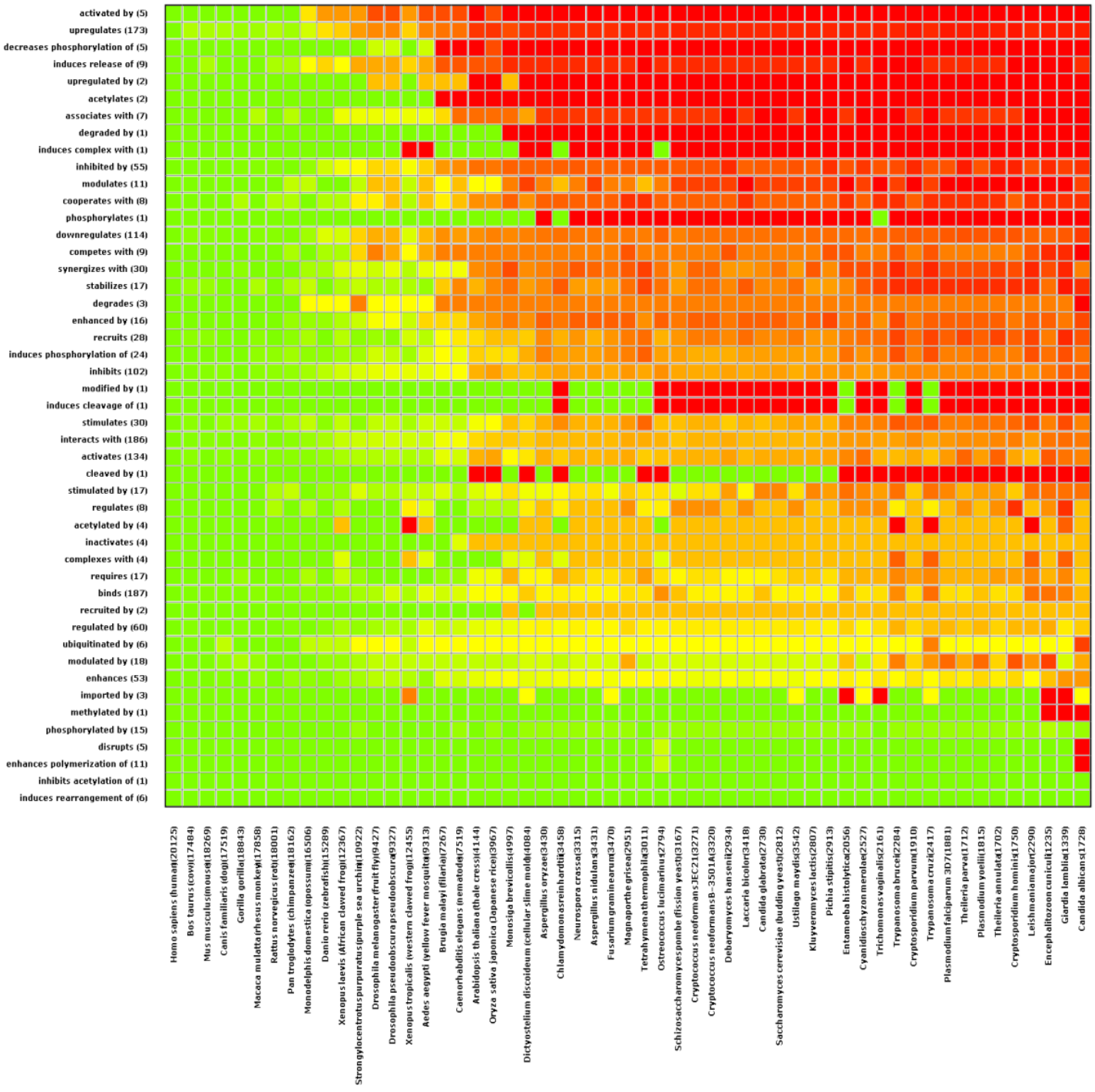


Figure 3.12 Summary of protein conservation in *FO*-clusters for proteins associated to HIV-Tat proteins. Each column summarizes the results for an eukaryotic organism. Each row summarize summarizes the results for a broad functional category of the proteome. The greener the square, the more similar the protein set associated with the functional category of the row in the organism of the column is to the correspondent human protein set. The redder he square, the less similar the protein set associated with the functional category of the row in the organism of the column is to the correspondent human protein set.

3.4. Discussion

In this article we systematically compare the human proteome to that of other eukaryotes in order to identify the proteins that are unique to human. We also analyze how similar the sets of proteins that participate in different biological phenomena are between human and each of the other eukaryotes. With these comparisons we hope to partially contribute to answer two questions. The first question is what makes *H. sapiens* unique among the eukaryotes. The second is what eukaryotes are likely to be the best model organisms to study different biological aspects of human biology.

There are technical challenges involved in answering these questions. One challenge is that of identifying proteins that are either unique in the human proteome or unique in the proteome of the eukaryote of interest. The only effective way to do so is by comparing the sequences of each protein from one of the proteomes to each of the sequences of the other. Another challenge is that of identifying the most likely functional ortholog pairs when comparing two proteomes. The final challenge is that of comparing the proteomes from a functional perspective.

We address the first challenge at the level of sequence similarity. If no sufficient similarity is found either between a human protein and any protein in the other eukaryote or a eukaryotic protein and any human protein, the protein is said to be unique.

We also address the second challenge at the level of sequence similarity. One can also identify the proteins in one proteome that are conserved in the other, based on the sequence similarity. Similar proteins between proteomes can be organized into clusters, classified in different categories. If the similarity between sequences is low, the proteins within a cluster are termed homologues. If the similarity is high over the entire sequence the proteins within a cluster are termed orthologs. The protein in a given eukaryote that is the most similar to a given human protein is termed its functional ortholog. Using sequence to infer such functional orthology was shown to be more accurate than using structure or a number of other protein features [194]. Various methods to identify orthologs are available [147, 195-198]. Of these we choose the one described in methods. A benchmark of this method done by comparing the human and the baker's yeast proteomes with themselves shows that this method identifies the real ortholog 100% of the times (data not shown).

The third challenge is addressed by taking advantage of the annotation of full proteomes with respect to different functional categories (tissue-specific, GO categories, enzymes, receptors, ligands, reactome, pathways, circuits). By integrating these different categories and comparing functional orthologs between human and any of the other organisms, we identify those organisms that have the set of protein is more similar to human in any given functional category.

The curated human proteome has 20125 proteins. 37% of all human proteomes are absent in at least one of the analysed eukaryotes. The set of human proteins that is more highly conserved in all other eukaryotes is that of catalytic proteins. In contrast, the sets of proteins that have the highest fraction of unique proteins in human are immunological proteins and receptors.

The organisms with functional protein sets that are the most similar to those of human are chimp and gorilla. The proteome of gorilla is 7% larger and that of chimp 2% smaller. When comparing humans with these closely related primates we find that 3% of the human proteins are absent in gorilla and 4% are absent in chimp. Many of the absent proteins are receptors or ligands involved in the immune system, although there are also some proteins from other categories that differ. In addition, we find that, for many functional categories and subgroups, the gorilla proteome is more similar to that of human than the chimp proteome. This is consistent with previously published results chimpanzee [199]. However, the pattern of gene (protein) duplication has diverged more between human and gorilla than between human and gorilla. This can be seen by the fact that 71% of the ***FO***-clusters between chimp and human are ***O-O***-clusters while only 59% fall in this category for the comparison between human and gorilla. In summary, ***O-O***-clusters between human and chimp are more numerous than between human and gorilla. In contrast, the gorilla proteome forms the highest number of ***O-M-*** and ***M-M-***clusters with the human proteome.

Our analysis also identifies lower eukaryotes that could be good models to study different aspects of human biology. For example, *C. elegans* is likely to be suitable for studying EGFR mediated MAPK pathways regulatory processes. In fact this organism was used to investigate the role of one specific domain of CDC25 in cancer prevention and development [200]. Fungi are also identified as being likely to be a suitable model to study role of bioactive peptides or neuropeptides in regulation and fine-tuning of metabolism, as was already done [201]. A final example can be found again in fungi, which are again likely

to be suitable models to study the role of small peptides in regulation of host-pathogens interactions [202, 203]. Another aspect of this analysis which we chose not to focus on is that of identifying proteins in eukaryotic parasites that could mediate the effect that the parasites have in humans. These proteins would be those that are more similar to their functional orthologs in human in the functional categories that are involved in the disease phenotype [204]. The methodology presented here could be used to study other organism in similar fashion [147]. In addition our results could have strong implication regarding the use of primates or other eukaryotes to study disease/related issues that cannot be studied in humans due to ethical, scientific or legal issues. It is expected that the more similar the protein set that is involved in a given biological process in a specific organism is to that of human, the more likely it is that the results of studying this process in that organism can be extrapolated to human. Hence, to study that process one should consider this issue in conjunction with technical considerations before choosing the model organism for the study.

3.5. Methodology

3.5.1. Proteome sequences

The complete proteomes of *Homo sapiens* (25679 proteins), *Pan troglodytes* (24732 proteins), and *Macaca mulatta* (23272 proteins) were downloaded from NCBI. The complete proteome from *Gorilla gorilla* (27335 proteins) was downloaded from Ensemble (version 66.31). The complete proteomes of 51 eukaryotic organisms with fully sequenced genomes (**Table 3.** 1) was downloaded from the KEGG database (December 2009), cross-referenced and complemented with the corresponding proteomes found in the NCBI database.

3.5.2. Homology analysis

BLAST+ version 2.2.26 was downloaded from NCBI and used locally. All protein sequences were formatted using FormatDB. A pipeline that identifies various levels of homology (orthologs, domain orthologs, homologues) and classifies two proteomes relative to each other based on these relationships was developed and implemented using Python 2.7, Numpy (Numerical Python), Scipy and Matplotlib.

3.5.3. Proteome Comparison and Classification

The comparison between the complete proteomes of any two organisms was done as described previously [147]. In short, first the proteomes are BLASTed against each other. Afterwards, orthologs and homologues pairs are separated based on e-value, identity of residues between the two aligned sequences and fraction of the total protein sequences that align to each other. **Figure 3.1** details the different groups of proteins that are generated from this analysis.

Each protein of the total proteome from a reference organism $\boldsymbol{R}$ is blasted against the entire proteome of a target organism $\boldsymbol{T}$. The result set $\boldsymbol{S}$ includes all protein pairs that generate a hit with $\boldsymbol{e-value \leq 10^{-4}}$, with one $\boldsymbol{S}$-cluster per human protein. All proteins from $\boldsymbol{R}$ that do not generate positive hits in the proteome of $\boldsymbol{T}$ are grouped in absent protein clusters ($\boldsymbol{A}$-cluster). All proteins from $\boldsymbol{T}$ that do not generate positive hits in $\boldsymbol{R}$ are also grouped in $\boldsymbol{A}$-clusters.

Each pair of sequences in $\boldsymbol{S}$ is further analyzed in order to identify homologues, domain orthologs, and orthologs. General homologues are defined as all pairs of proteins that

are matched with $\boldsymbol{e-value \leq 10^{-4}}$ and identity smaller than or equal to 20%. Because it is important to identify those protein pairs that are distantly related, for the analysis of functional evolution, we also separate the set of exclusive homologue pairs. These are defined as all pairs of proteins that are matched with $\boldsymbol{10^{-10} \leq e-value \leq 10^{-4}}$ and identity smaller than or equal to 20%. These are used to build the protein set $\boldsymbol{H}$. Exclusive domain orthologs are defined as all pairs that are matched with $\boldsymbol{10^{-30} \leq e-value \leq 10^{-10}}$ and an identity between 20% and 30%. These are used to build the protein set $\boldsymbol{D}$. Orthologs are defined as all pairs that are matched with $\boldsymbol{e-value < 10^{-30}}$ and an identity larger than 30%. These are used to build the protein set $\boldsymbol{O}$. We also consider the set of general orthologs, $\boldsymbol{O_g}$, defined as the union set between $\boldsymbol{D}$ and $\boldsymbol{O}$. This set is important for the analysis of functional evolution.

Figure 3.13

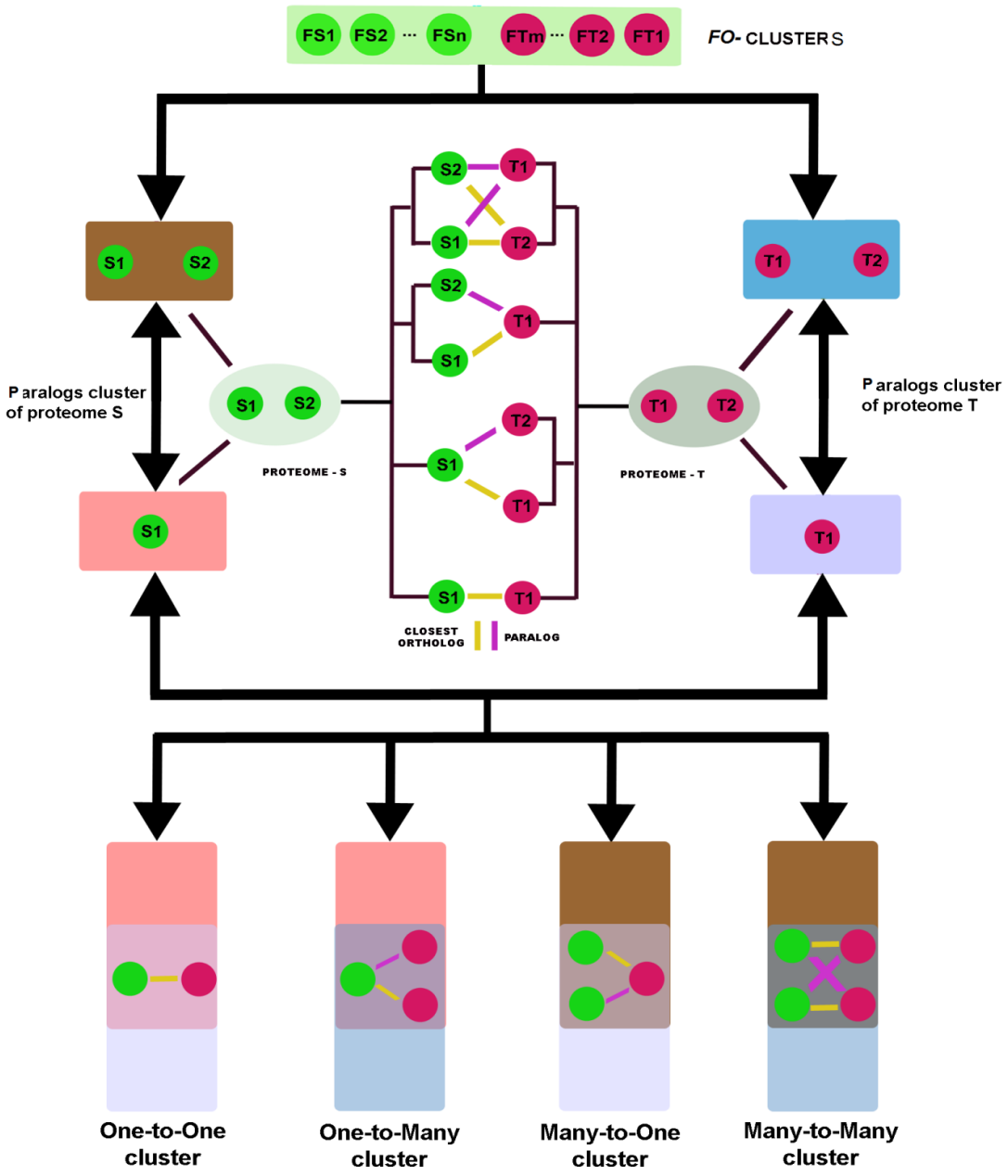


Figure 3.13 Classification of human centric *FO*-clusters. *FO*-clusters can be of several types. One-to-one (***[O-O]***) clusters include one human protein and one protein from the target eukaryote. One-to-many (***[O-M]***) clusters include one human protein and more than one protein (set of paralogous proteins) from the target eukaryote. Many-to-one (***[M-O]***) clusters include more than one human protein (set of paralogous proteins) and one protein from the target eukaryote. Many- to-many (***[M-M]***) clusters include more than one human protein and more than one protein from the target eukaryote.

3.5.4. Functional orthology and duplication analysis

Cluster of Functional Orthologs (***FO***) are defined as a subset of ***O***. ***FO*** includes all protein pairs of the ***O*** protein set that share an alignment for more than 80% of the proteins' lengths. There can be four types of ***FO***-clusters (**Figure 3.13**). A one-to-one ***FO***-cluster (***F[O-O]***) is composed of one protein from ***S*** and one protein from ***T***. A one-to-many ***FO***-cluster (***F[O-M]***) is composed of one protein from ***S*** and more than one protein from ***T***. A many-to-one ***FO***-cluster (***F[M-O]***) is composed of more than one protein from ***S*** and only one protein from ***T***. A many-to-many ***FO***-cluster (***F[M-M]***) is composed of more than one protein from ***S*** and more than one protein from ***T***.

Given that ***FO***-clusters are composed of proteins that are very close, sequence-wise, such clusters can be analyzed to infer information about duplication of proteins and protein function. Whenever the ***FO***-cluster has more than one protein from any of the organisms (***F[O-M]***, ***F[M-O]***, and ***F[M-M]*** clusters) we use a function score **F,** defined in [147]**,** to predict which pair of proteins within the cluster is more likely to include true functional orthologs. This score is given by:

$$F = (F1 + F2) - F3 \qquad \textbf{Eq. 1}$$

Factor ***F1*** is calculated as follows.

$$F1 = 1 - (S - I)/S \qquad \textbf{Eq. 2}$$

In Eq. 2, **S** represents the similarity residues of amino acids found over the alignment, and **I** represent the identical residues found over the alignment. Both these values are outputs of BLAST. ***F1*** is always between zero and 1.

Factor ***F2*** is calculated as follows.

$$F2 = (AL - G1 - G2)/PL \qquad \textbf{Eq. 3}$$

In Eq. 3 ***G1*** is the number of gaps within the aligned region of the query sequence, **G2** is the number of gaps within the aligned region of the target sequence, **AL** represents the length of the alignment, and ***PL*** is the total length of the query sequence. **F2** is always between zero and 1.

Factor ***F3*** is calculated as follows.

$$\boldsymbol{F3 = (G1/L1) + (G2/L2)} \qquad \textbf{Eq. 4}$$

In Eq. 4 ***L1*** is the length of the query sequence, and ***L2*** is the full length of the target sequence. ***F3*** is always between 0 and 2.Hence, ***F*** is always between 0 and 2.

The pair of proteins with the highest ***F***-score in each ***FO***-cluster is considered to be the one including the real functional orthologs.

We have benchmarked this assumption by BLASTing the human genome against itself and the baker's yeast genome against itself. In every single case, the highest ***F****-score* is that of a protein with itself. Unlike the e-value, the ***F***-score provides a measure that is symmetric between two proteomes. The highest ***F***-score pair for a ***FO***-cluster between two organisms is always the same, whether the target genome is ***S*** or ***T***.

3.5.5. Functional re-annotation of the human proteome

To attribute function to human and baker's yeast proteins, we downloaded the GOSLIM classification for human and baker's yeast from the GO database [88, 89], including categories for biological process, molecular functions and cellular localization. This information was used to re-annotate function in the remaining eukaryotic proteomes under comparison. If not annotated, the protein from a specific eukaryote with the highest ***F****-score* with respect to a given human protein was attributed the same GO classification as that of the human.

3.5.6. Calculating the difference between corresponding sets of proteins in different organisms

We wanted to compare how different the set of proteins involved in a given biological function is between different organisms. To do so we calculate the Hamming Distance (HD) between the vector $\overrightarrow{V1}$ of protein functions associated to a specific process, localization or pathway in humans and the vector $\overrightarrow{V2}$ of corresponding protein functions in another organism. ***HD*** is given by $HD = \sum_{i=1}^{n}(1 - \delta_i)$, where δ_i is the Kronecker delta. δ_i is 1 if the elements in position i of both vectors are homologue and 0 otherwise. ***HD*** is normalized (***NHD***) by dividing it by the maximum ***HD*** between corresponding vectors of all organisms.

The smaller ***NHD*** is, the more similar the two vectors are and the more similar is the set of functions executing a specific process in both organisms. Consequently, the smaller ***NHD*** is, the more likely it is that the process of interest works in a similar way in the organisms being compared.

All calculations were performed using Mathematica [91].

3.6. Supporting Materials

3.6.1. Supporting Figures

Figure S2.1

Domain clade level frequency representation of proteins that linked in functional orthologs (*HFP-TFP*), orthologs (*HOP-TOP*), domain orthologs (*HDP-TDP*), homologous (*HHP-THP*), significant (*HSP-TSP*) and absent (*HAP-TAP*) clusters of human and all the eukaryotes.

(See the all figures of the clusters in next pages)

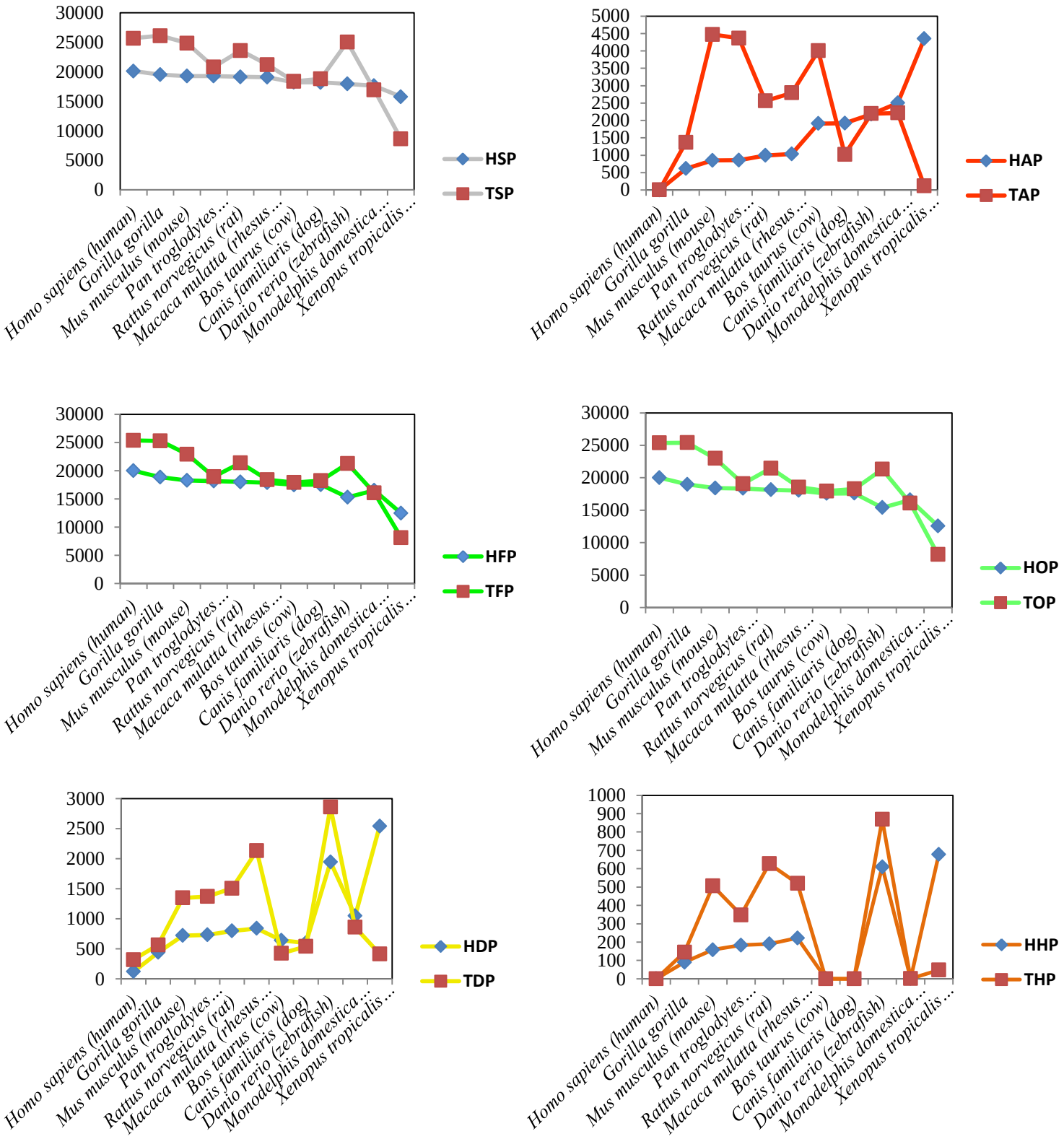


Figure S2.1. A Frequency of proteins found from **Human and Vertebrates** in the clusters of Functional orthologs (*HFP-TFP*), Orthologs (*HOP-TOP*), Domain orthologs (*HDP-TDP*), Homologous (*HAP-TAP*), Significance (*HSP-TSP*) and Absent (*HAP-TAP*).

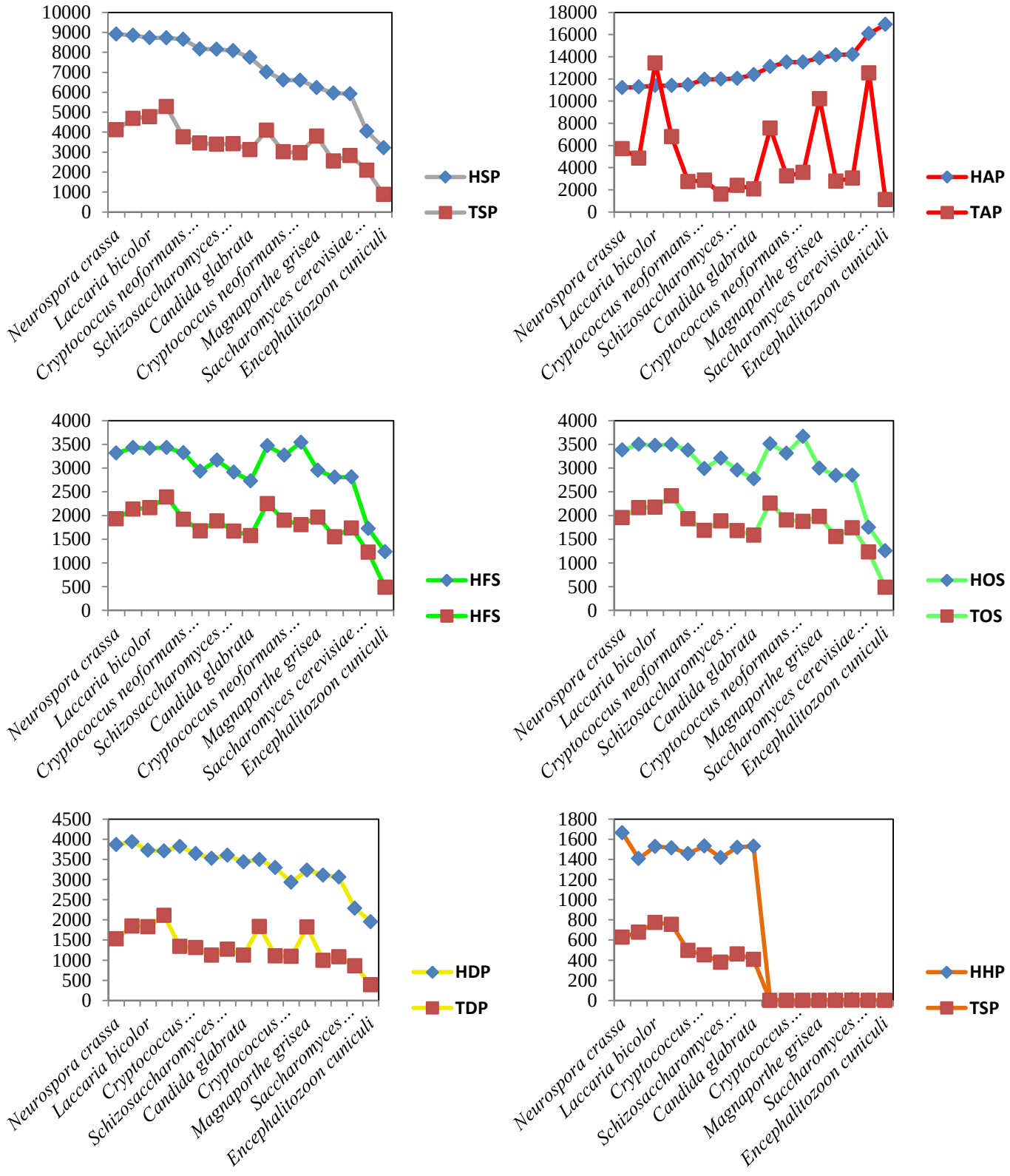


Figure S2.1. B Frequency of proteins found from **Human and Fungi** domain in the clusters of Functional orthologs (*HFP-TFP*), Orthologs (*HOP-TOP*), Domain orthologs (*HDP-TDP*), Homologous (*HAP-TAP*), Significance (*HSP-TSP*) and Absent (*HAP-TAP*).

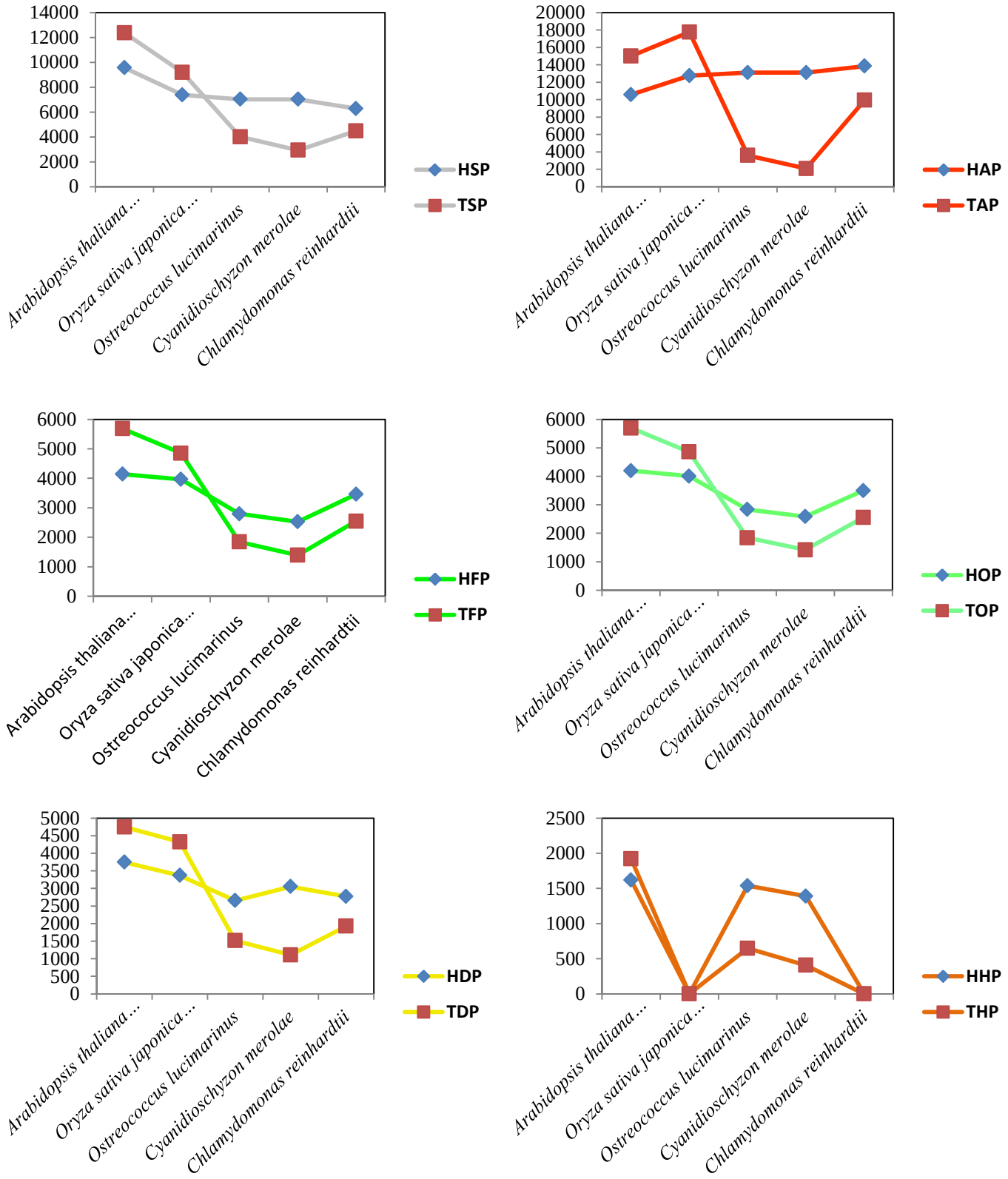


Figure S2.1. C Frequency of proteins found from **Human and Plant** domain in the clusters of Functional orthologs (*HFP-TFP*), Orthologs (*HOP-TOP*), Domain orthologs (*HDP-TDP*), Homologous (*HAP-TAP*), Significance (*HSP-TSP*) and Absent (*HAP-TAP*).

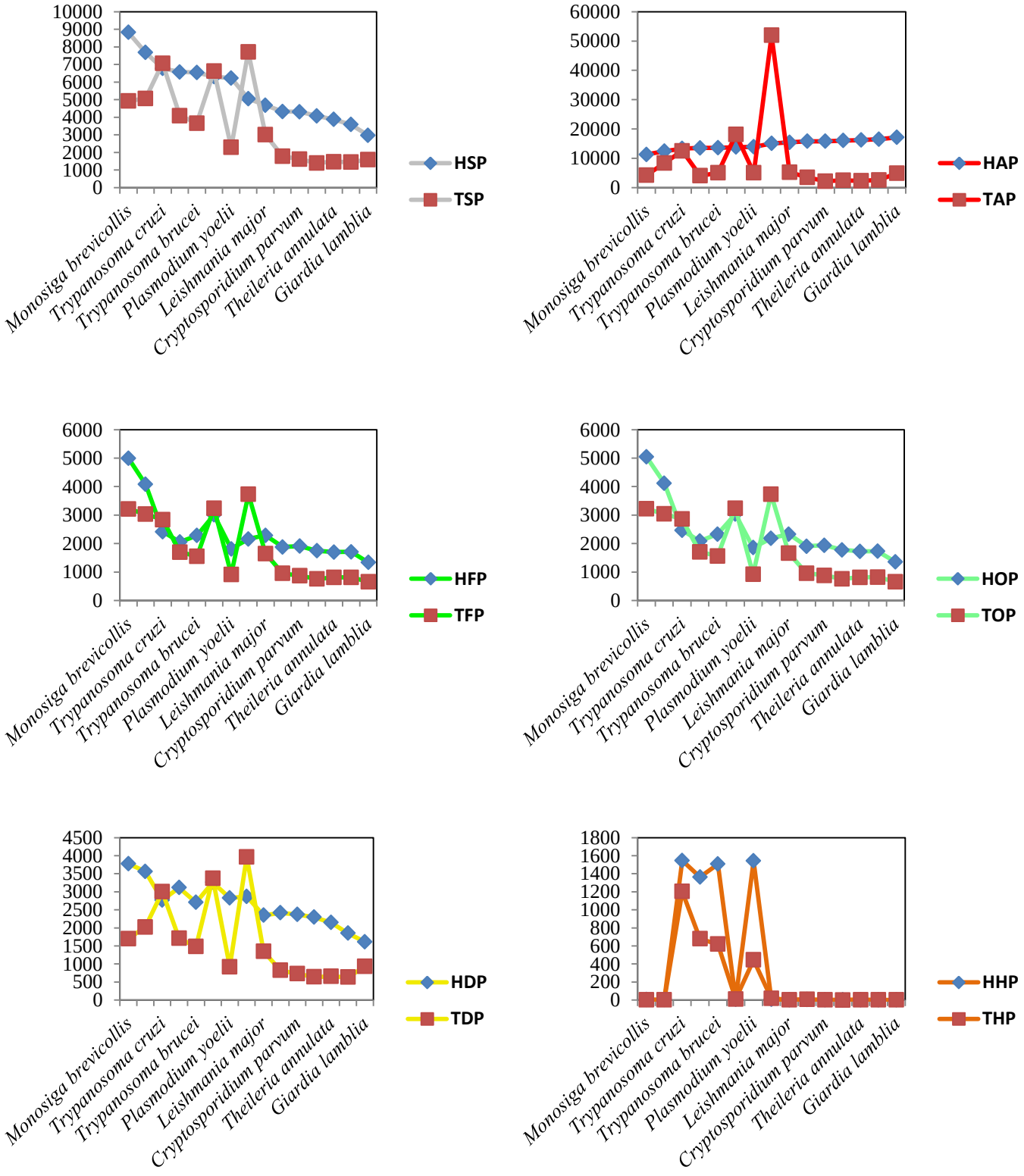


Figure S2.1. D Frequency of proteins found from Human and Protist domain in the clusters of Functional orthologs (*HFP-TFP*), Orthologs (*HOP-TOP*), Domain orthologs (*HDP-TDP*), Homologous (*HAP-TAP*), Significance (*HSP-TSP*) and Absent (*HAP-TAP*).

Figure S2.2*

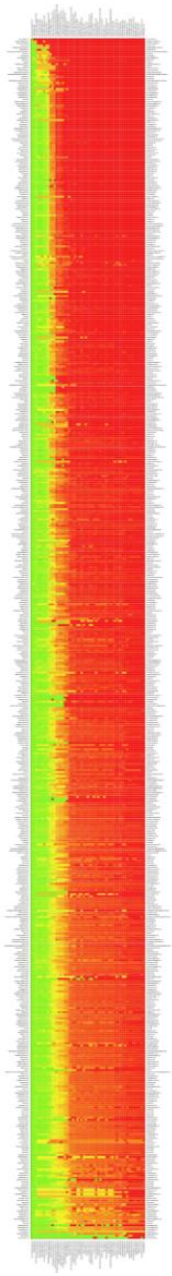


Figure S2.2 Complete heat map for comparison of proteome associated to human tissues and that found conservation at functional orthologs levels in the [FO] clusters with 56 eukaryotes. Each row corresponds to specific term of tissue in which the human protein(s) are expressed and each column corresponds to one of the eukaryotes under analysis. The numbers between parentheses indicate the number of proteins associated to that category in human tissue (rows) and in the other eukaryotes (columns). The protein of the tissues is sorted by increasing average network distance between human and each of the other eukaryotes. The organisms are sorted by increased average distance of the networks for all tissue categories between the organisms represented in the column and human. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to human. Green color indicates similar sets of proteins with respect to human. *Enlarged figure is available as Figure S2.2 in the CD that is provided with this thesis.

Figure S2.3*

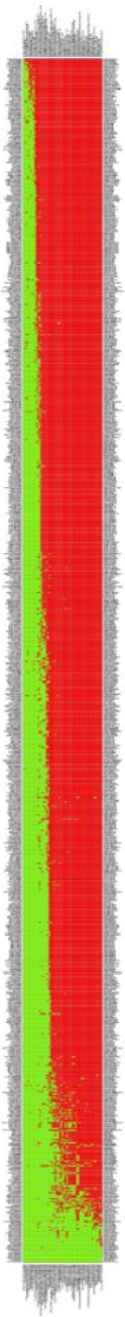


Figure S2.3 Complete heat map for comparison of proteome associated to ligand activity in human and that found conservation at functional orthologs levels in the [FO] clusters with 56 eukaryotes. Each row corresponds to a ligand protein (NCBI Gene ID) of human and each column corresponds to one of the eukaryotes under analysis. The text in parentheses represent gene symbol of the ligand proteins of human (rows) and in the other frequency of total functional orthologs found for the total human ligands in eukaryotes (columns). The ligand proteins are sorted by increasing average network distance between human and each of the other eukaryotes. The organisms are sorted by increased average distance of the networks for all the ligands between the organisms represented in the column and human. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to human. Green color indicates similar sets of proteins with respect to human. *Enlarged figure is available as Figure S2.3 in the CD that is provided with this thesis.

Figure S2.4*

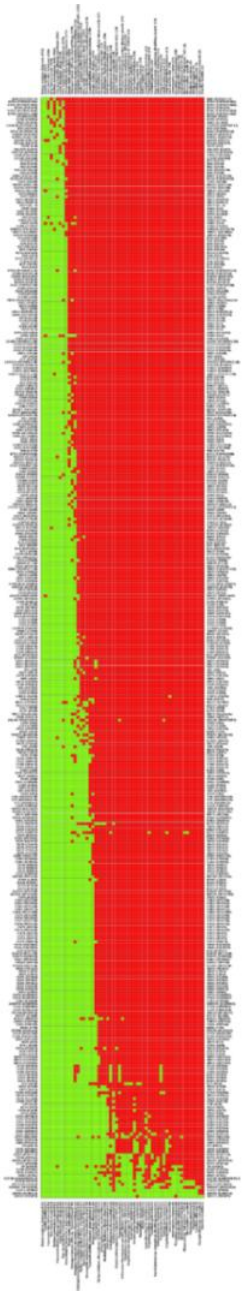


Figure S2.4 Complete heat map for comparison of proteome associated to receptor activity in human and that found conservation at functional orthologs levels in the [FO] clusters with 56 eukaryotes. Each row corresponds to a receptor protein (NCBI Gene ID) of human and each column corresponds to one of the eukaryotes under analysis. The text in parentheses represent gene symbol of the receptor proteins of human (rows) and in the other frequency of total functional orthologs found for the total human receptors in eukaryotes (columns). The receptor proteins are sorted by increasing average network distance between human and each of the other eukaryotes. The organisms are sorted by increased average distance of the networks for all the receptors between the organisms represented in the column and human. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to human. Green color indicates similar sets of proteins with respect to human. *Enlarged figure is available as Figure S2.4 in the CD that is provided with this thesis.

Figure S2.5*

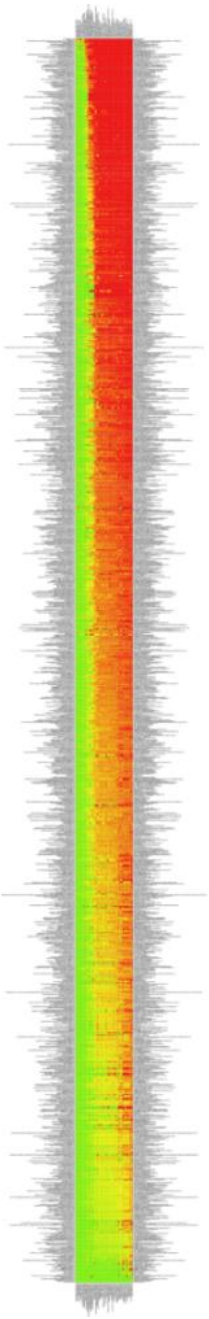
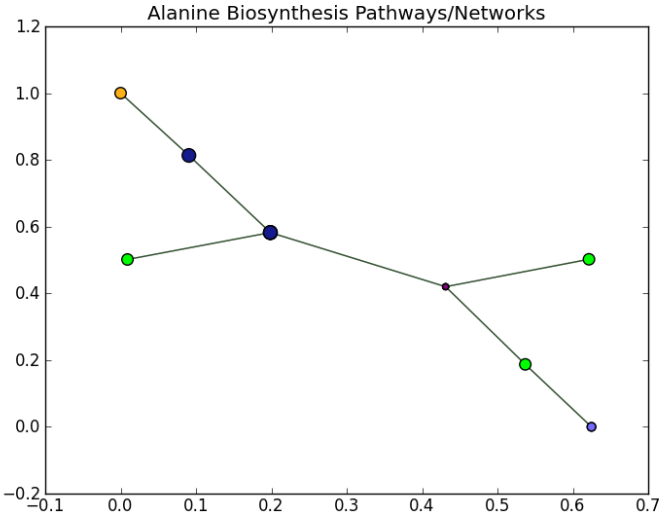
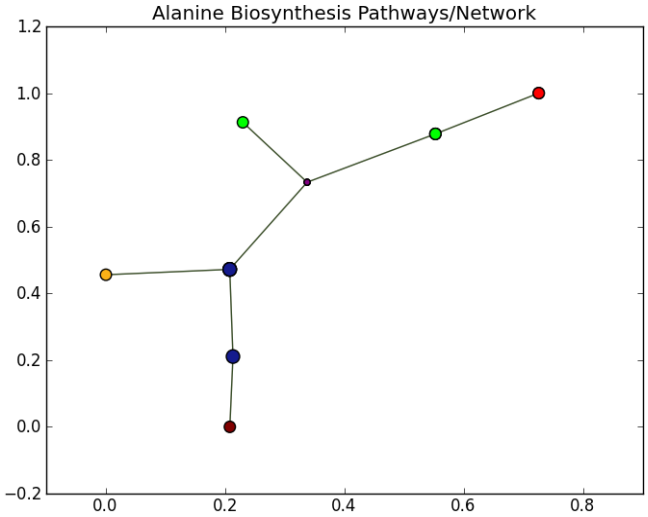


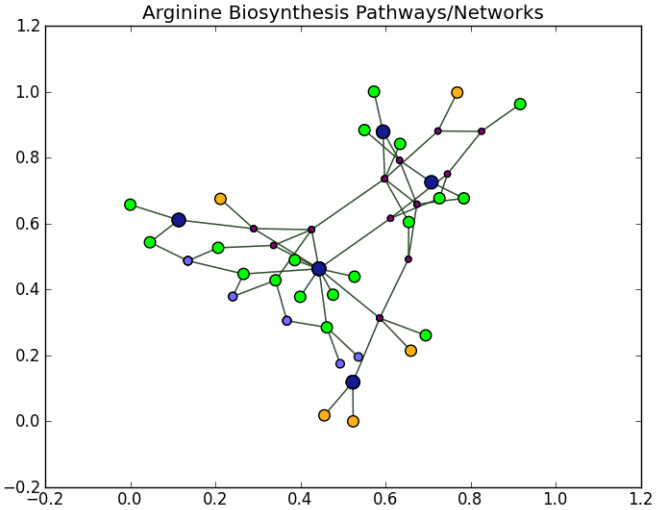
Figure S2.5 Complete heat map for comparison of proteome associated to specific reactions in metabolic or signaling pathways of human and that found conservation at functional orthologs levels in the [FO] clusters with 56 eukaryotes. Each row corresponds to specific term of the reaction pathways in human and each column corresponds to one of the eukaryotes under analysis. The numbers in parentheses represent human proteins that are associated with each of the reaction pathways (rows) and in the other frequency of total functional orthologs found in eukaryotes (columns). The reacting proteins are sorted by increasing average network distance between human and each of the other eukaryotes. The organisms are sorted by increased average distance of the networks for all the reactions between the organisms represented in the column and human. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to human. Green color indicates similar sets of proteins with respect to human. *Enlarged figure is available as Figure S2.5 in the CD that is provided with this thesis.

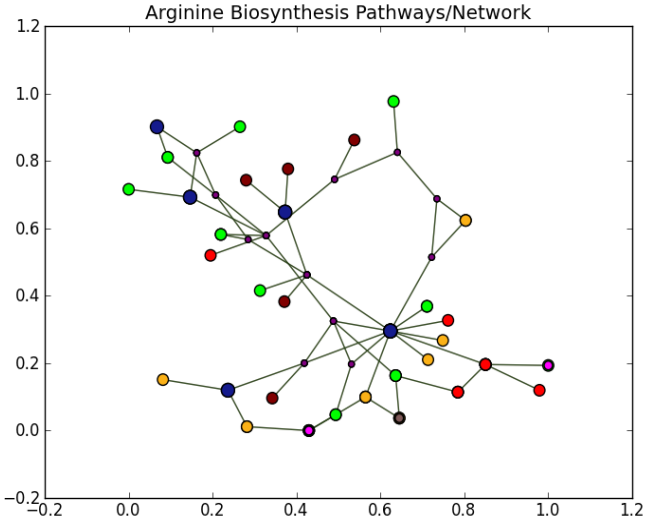
Figure S2.6

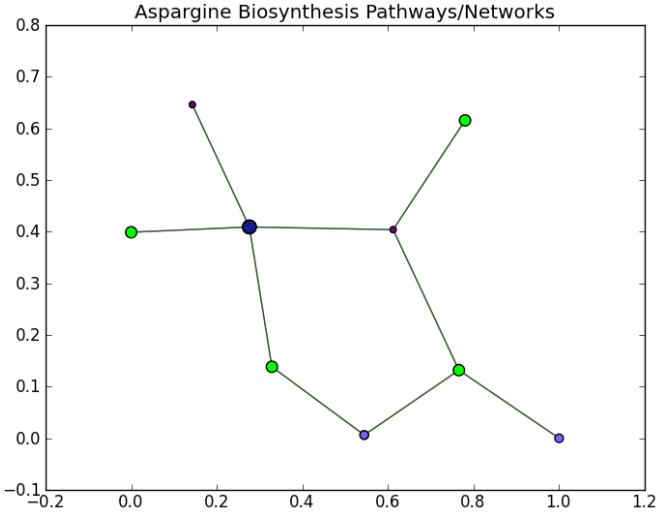
***Saccharomyces cerevisiae* pathways** **Human pathways**











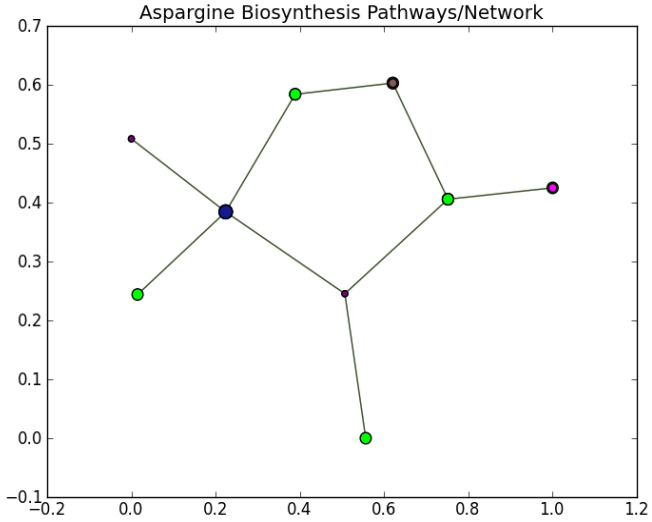
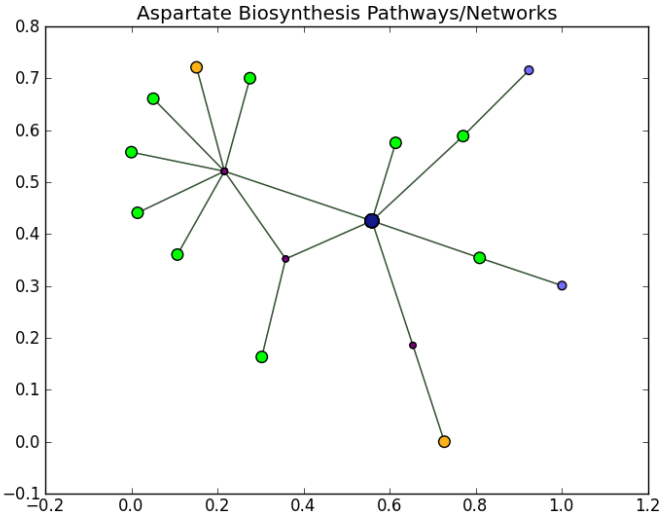
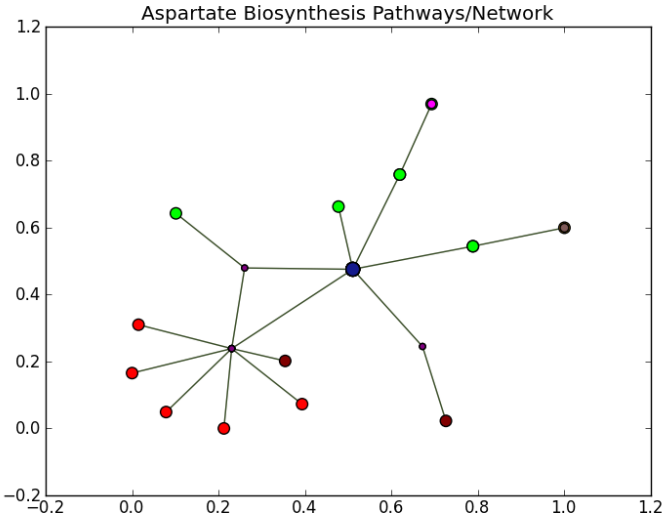
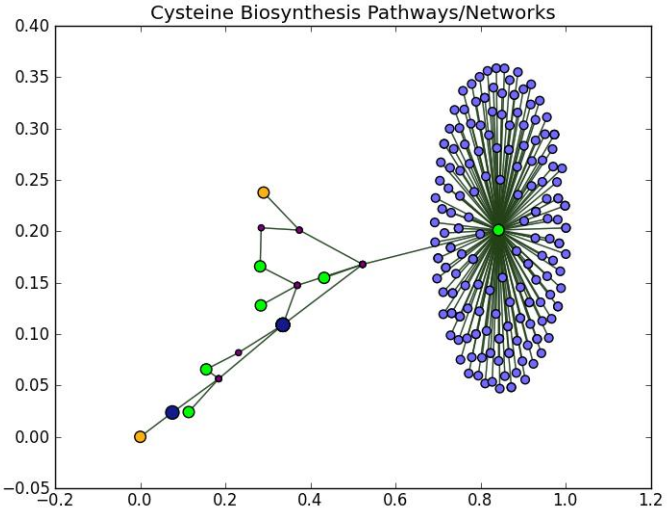


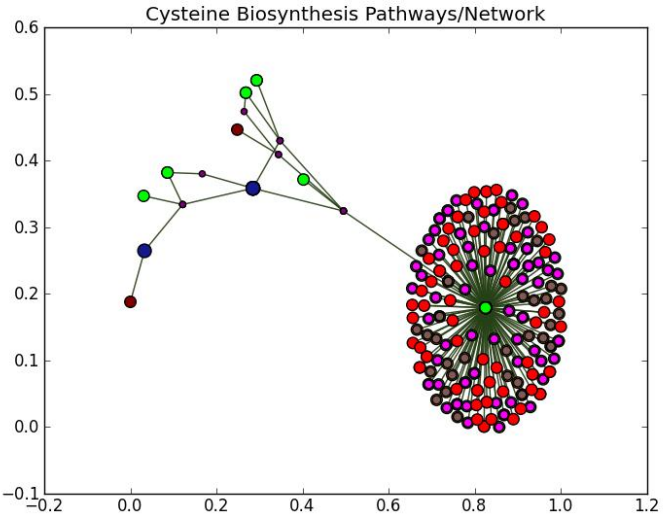
Figure S2.6 [continued…]

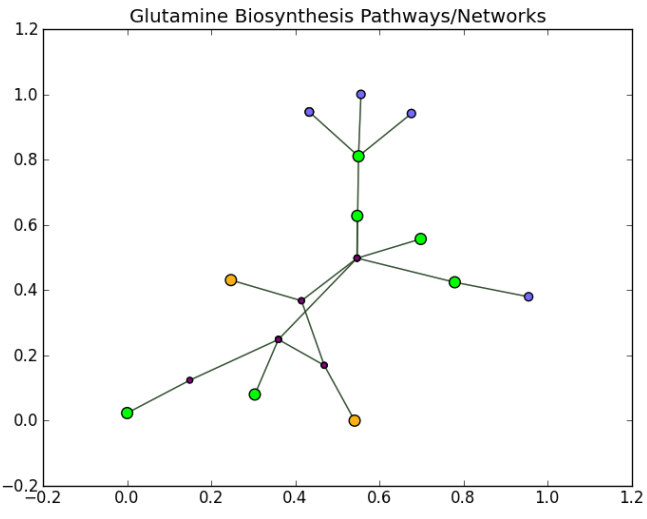
***Saccharomyces cerevisiae* pathways** | **Human pathways**











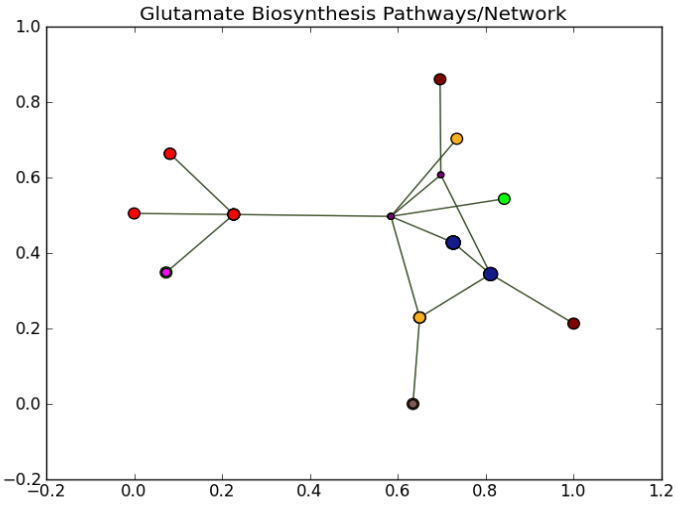
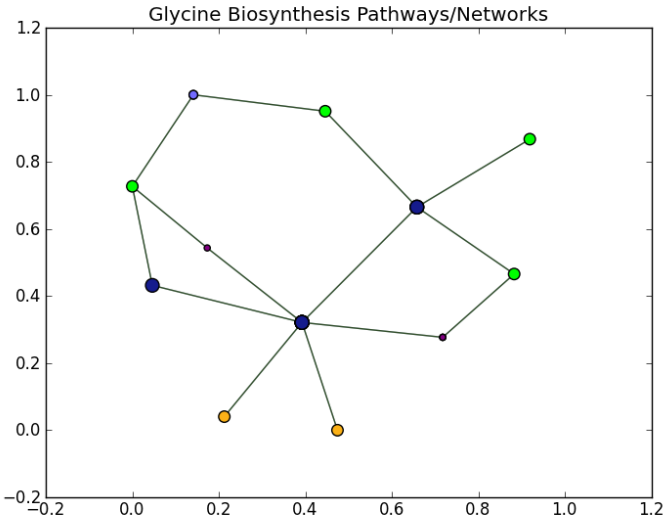
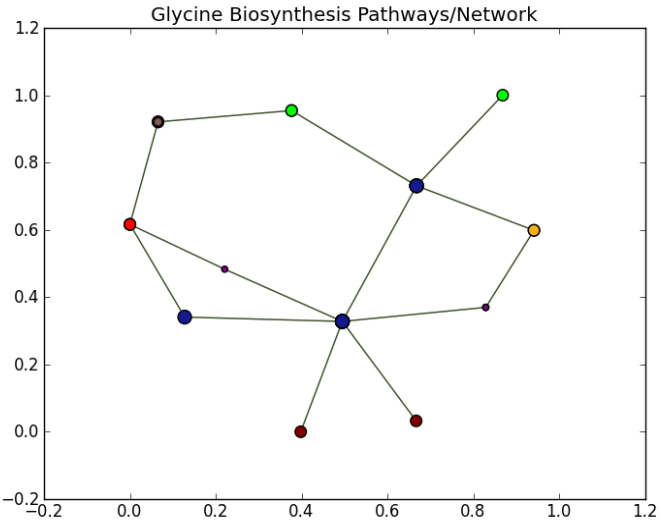
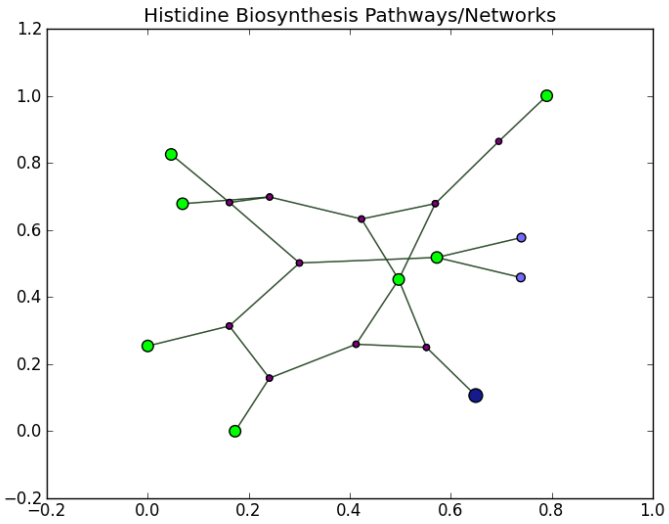


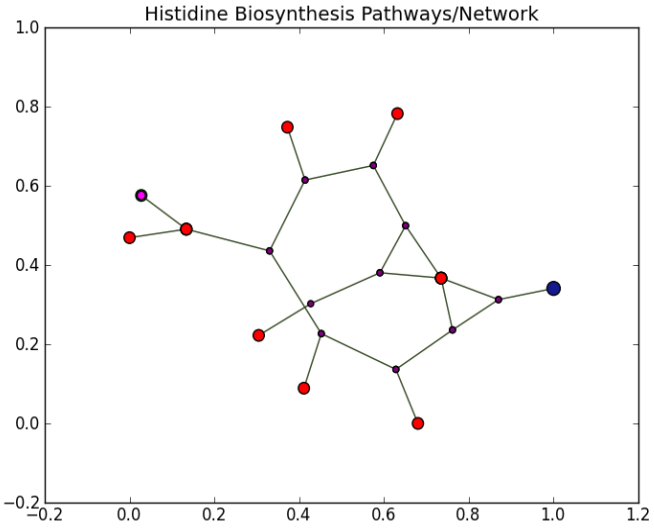
Figure S2.6 [continued…]

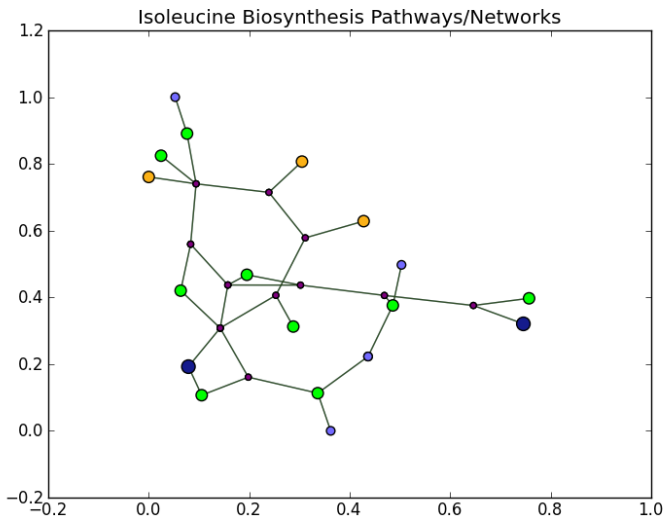
***Saccharomyces cerevisiae* pathways** | **Human pathways**











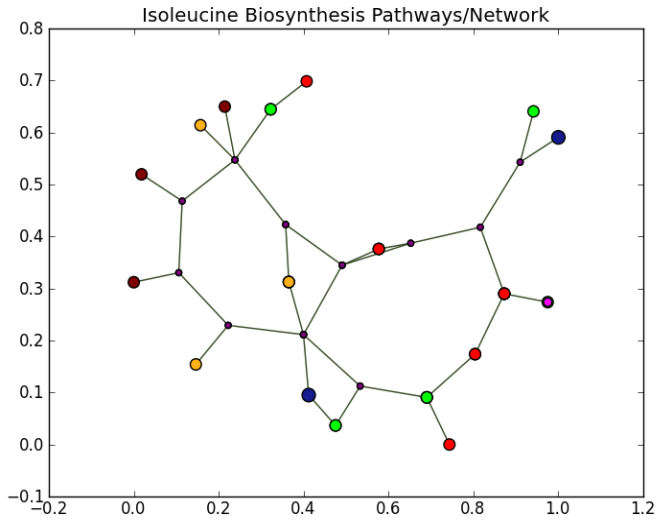
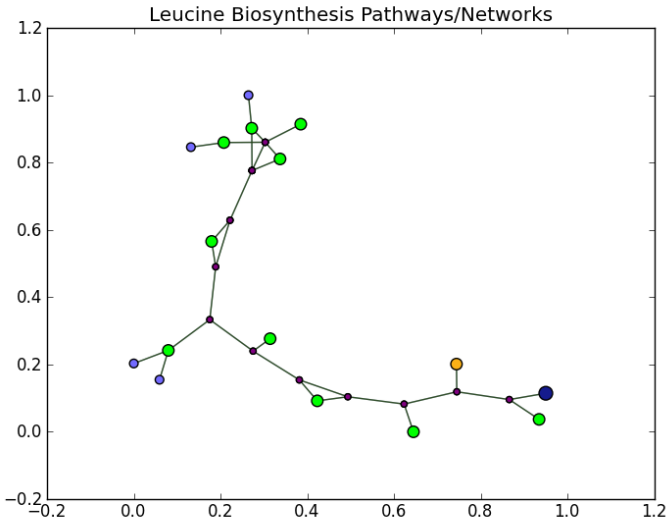
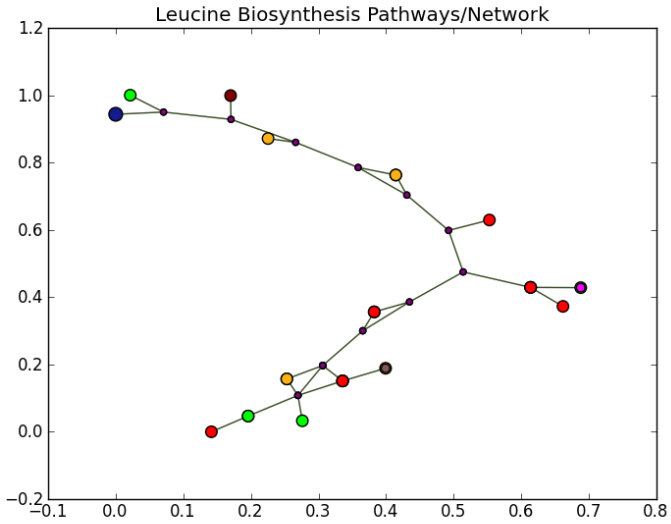
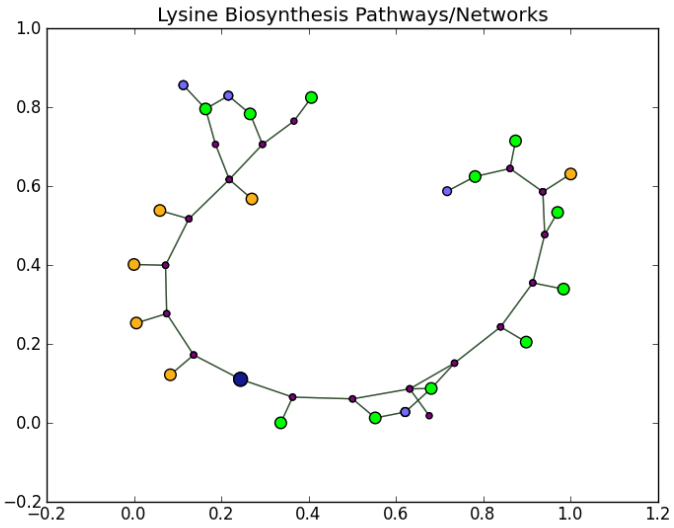


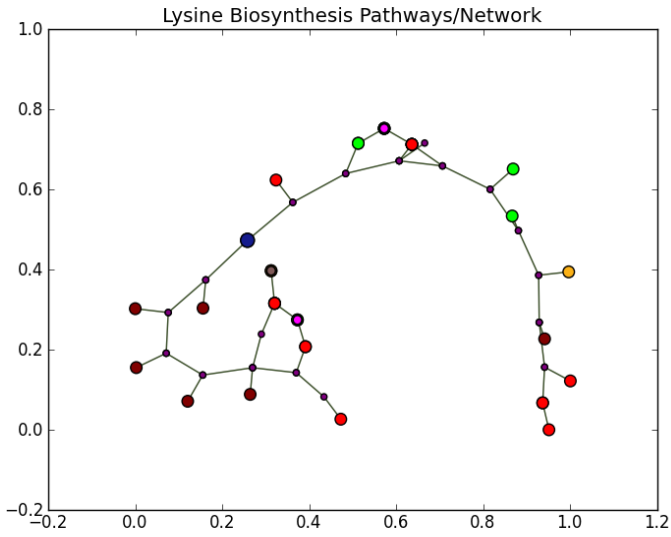
Figure S2.6 [continued…]

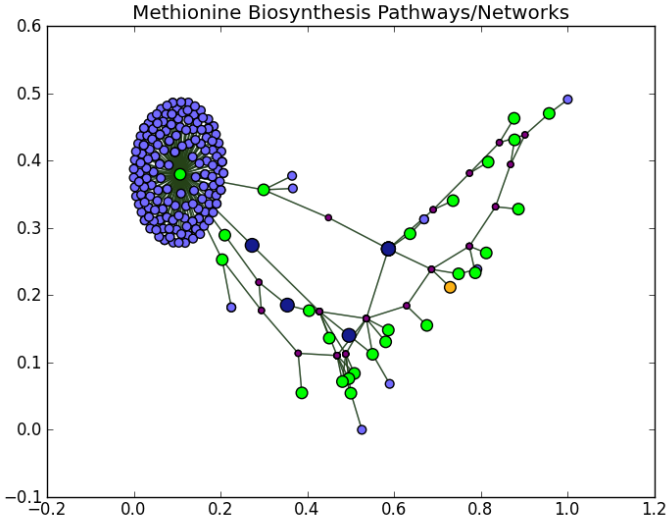
***Saccharomyces cerevisiae* pathways** | **Human pathways**











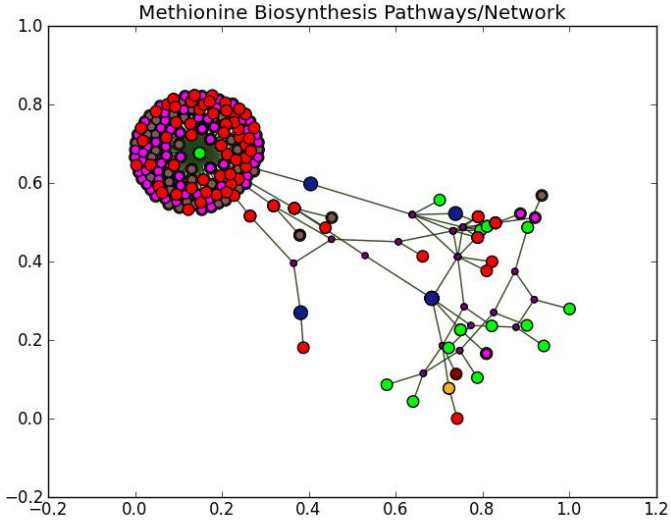
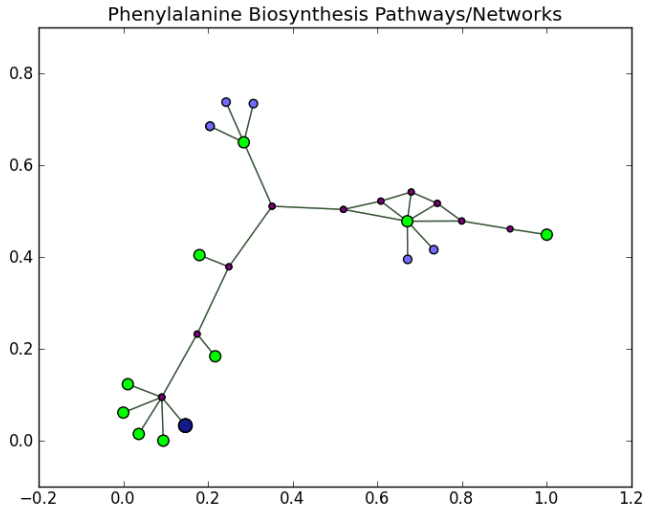
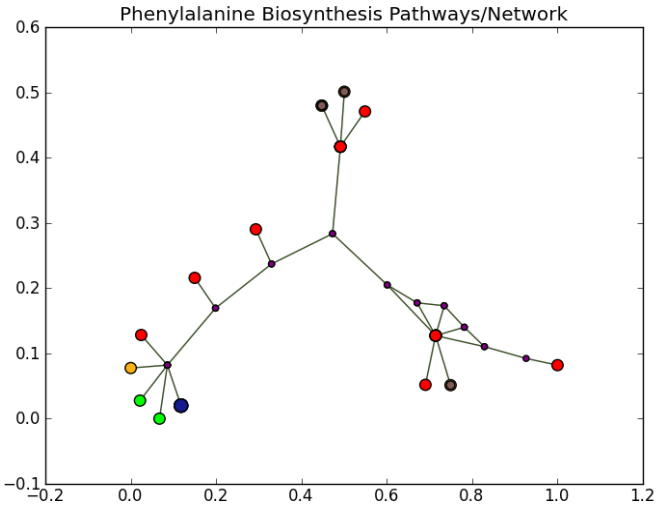
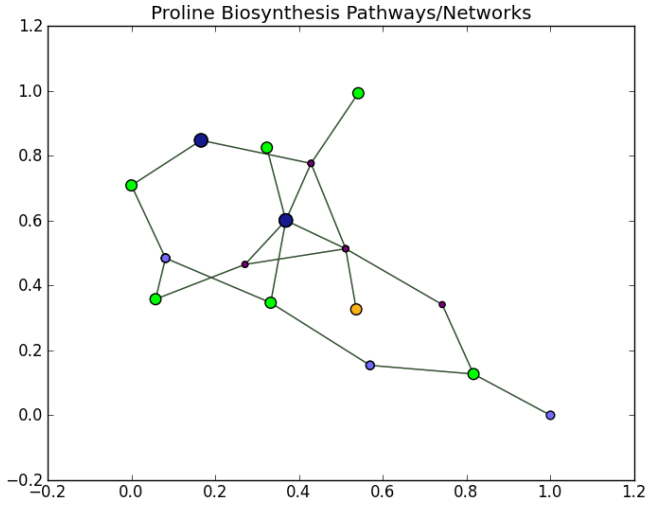


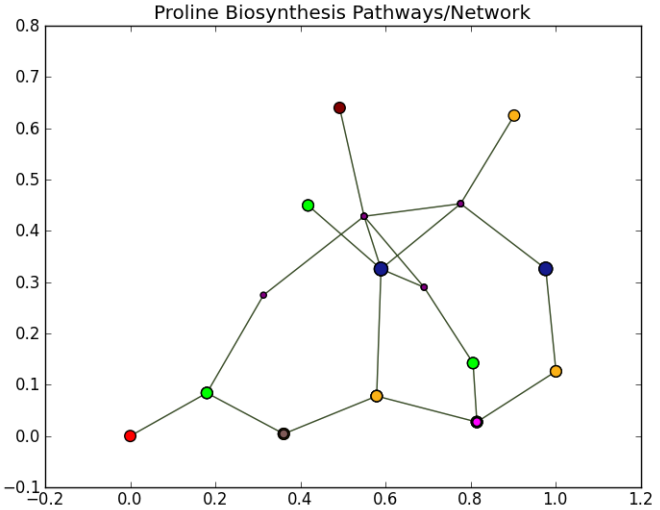
Figure S2.6 [continued…]

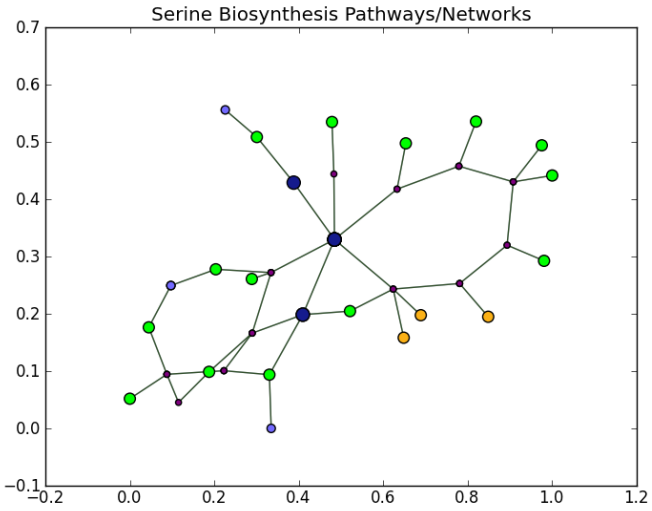
***Saccharomyces cerevisiae* pathways** | **Human pathways**











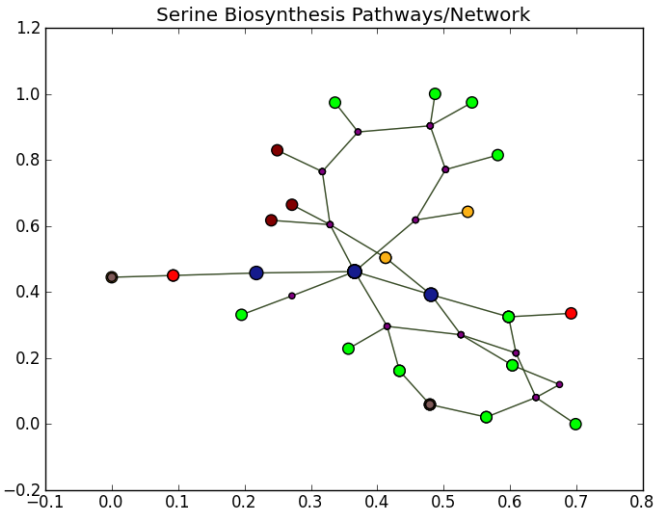
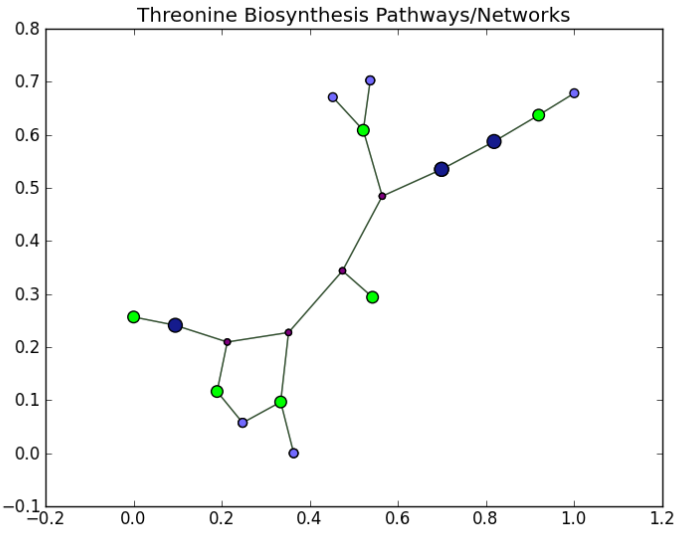
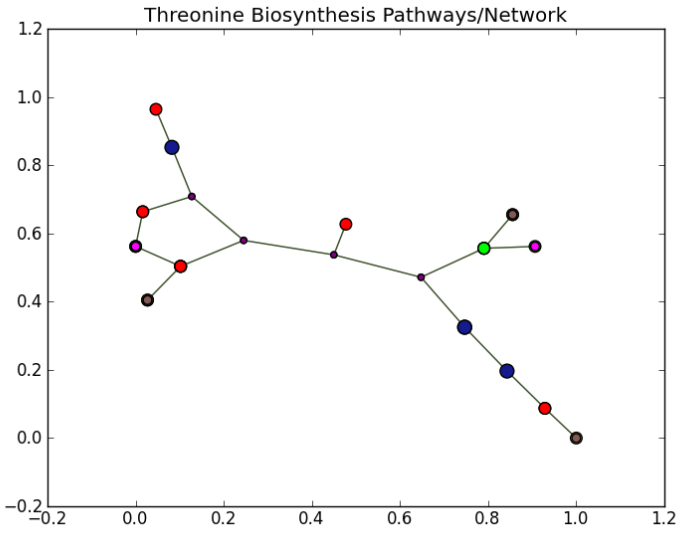
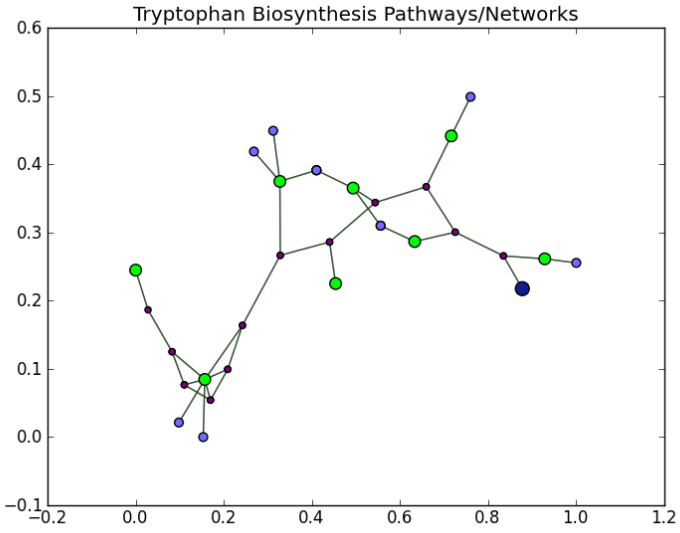


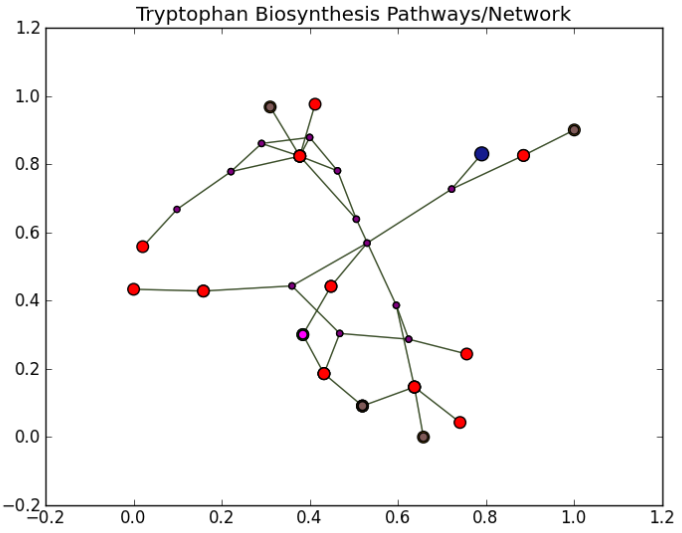
Figure S2.6 [continued…]

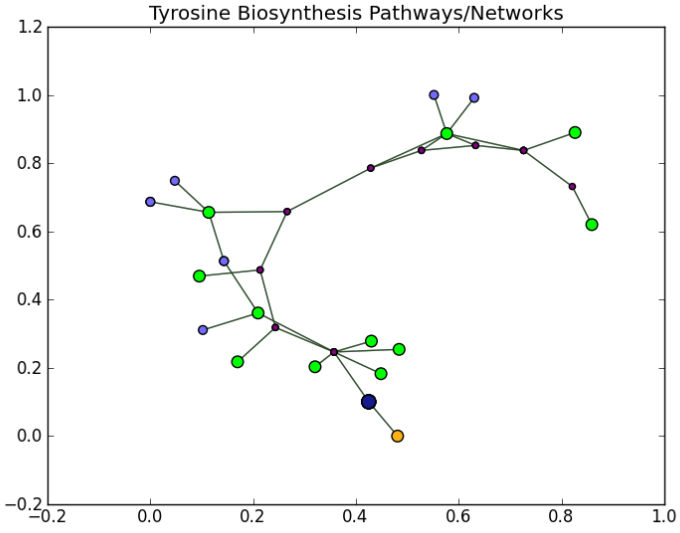
***Saccharomyces cerevisiae* pathways** | **Human pathways**











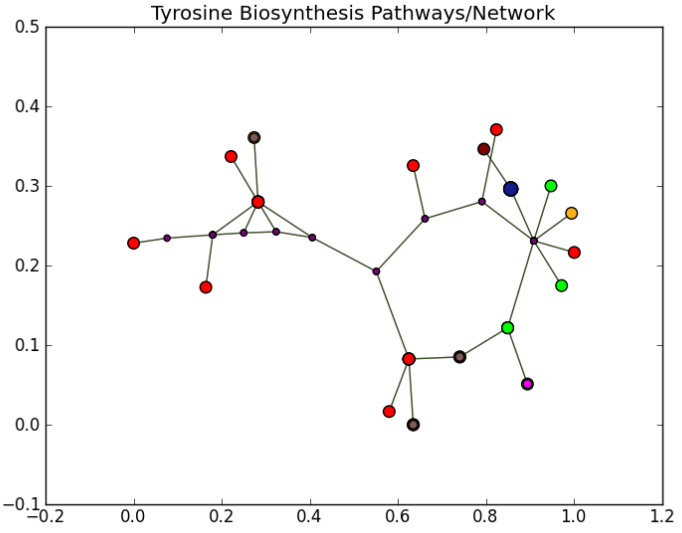


Figure S2.6 [continued…]

***Saccharomyces cerevisiae* pathways** **Human pathways**

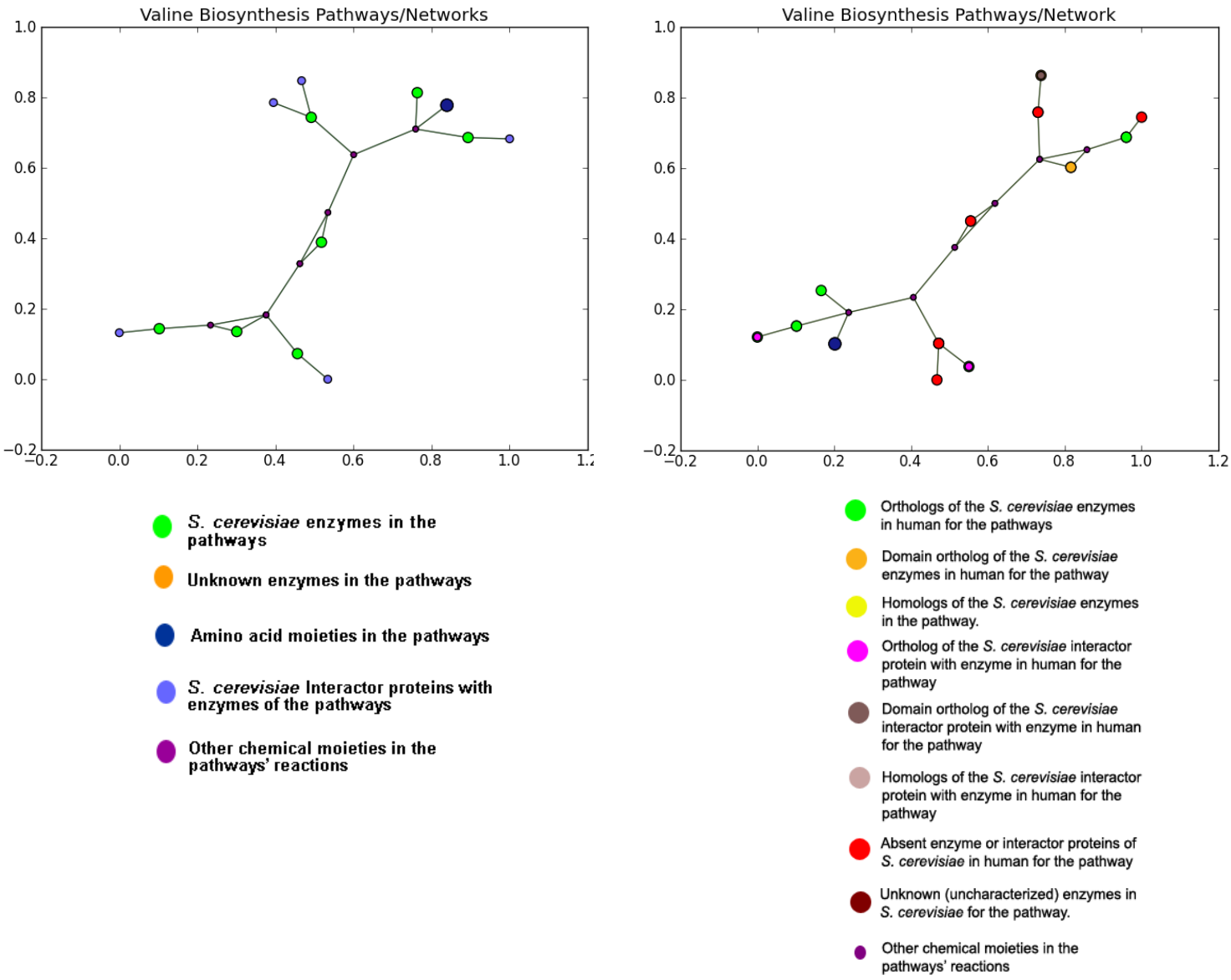


Figure S2.6 Comparative study of regulatory catalytic processes in each of the 20 amino acid biogenesis pathways between *S. cerevisiae* and human. Each node represents total molecules that are involved in module of an amino acid biogenesis pathway, manually separated from KEGG database. Color of the node represents type of the molecule i.e., enzyme, interactor protein, undefined protein, intermediate metabolite and amino acid moiety. The colored node also represent whether the same proteins of *S. cerevisiae* found as orthologs, domain orthologs, homologs or absent in human.

Figure S2.7*

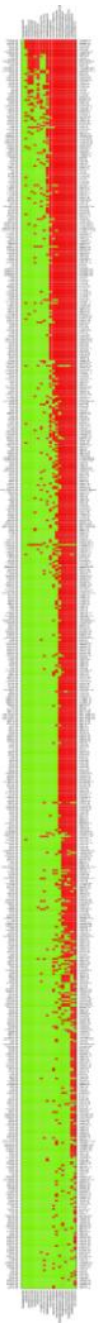


Figure S2.7 Complete heat map for comparison of proteome associated to bone development of human and that found conservation at functional orthologs levels in the [FO] clusters with 18 animals. Each row corresponds to proteins (NCBI Gene ID) that are associated with human bone development and each column corresponds to one of the animals under analysis. The numbers in parentheses represent human protein frequency (rows) and in the other frequency of total functional orthologs found in animals (columns). The bone development proteins are sorted by increasing average network distance between human and each of the other animals. The organisms are sorted by increased average distance of the networks for all the bone proteins between the organisms represented in the column and human. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to human. Green color indicates similar sets of proteins with respect to human. *Enlarged figure is available as Figure S2.7 in the CD that is provided with this thesis.

Figure S2.8*

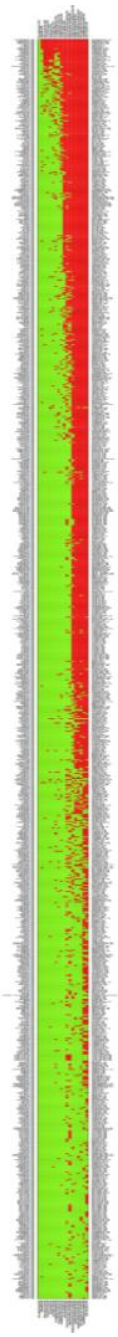


Figure S2.8 Complete heat map for comparison of proteome associated to muscle development of human and that found conservation at functional orthologs levels in the [FO] clusters with 18 animals. Each row corresponds to proteins (NCBI Gene ID) that are associated with human muscle development and each column corresponds to one of the animals under analysis. The numbers in parentheses represent human protein frequency (rows) and in the other frequency of total functional orthologs found in animals (columns). The muscle development proteins are sorted by increasing average network distance between human and each of the other animals. The organisms are sorted by increased average distance of the networks for all the bone proteins between the organisms represented in the column and human. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to human. Green color indicates similar sets of proteins with respect to human. *Enlarged figure is available as Figure S2.8 in the CD that is provided with this thesis.

Figure S2.9*

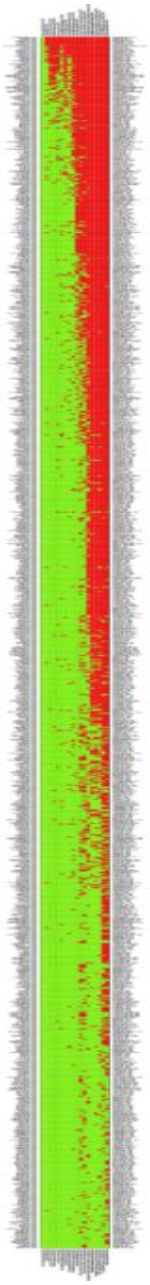


Figure S2.9 Complete heat map for comparison of proteome associated to brain development of human and that found conservation at functional orthologs levels in the [FO] clusters with 18 animals. Each row corresponds to proteins (NCBI Gene ID) that are associated with human brain development and each column corresponds to one of the animals under analysis. The numbers in parentheses represent human protein frequency (rows) and in the other frequency of total functional orthologs found in animals (columns). The brain development proteins are sorted by increasing average network distance between human and each of the other animals. The organisms are sorted by increased average distance of the networks for all the bone proteins between the organisms represented in the column and human. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to human. Green color indicates similar sets of proteins with respect to human. *Enlarged figure is available as Figure S2.9 in the CD that is provided with this thesis.

Figure S2.10*

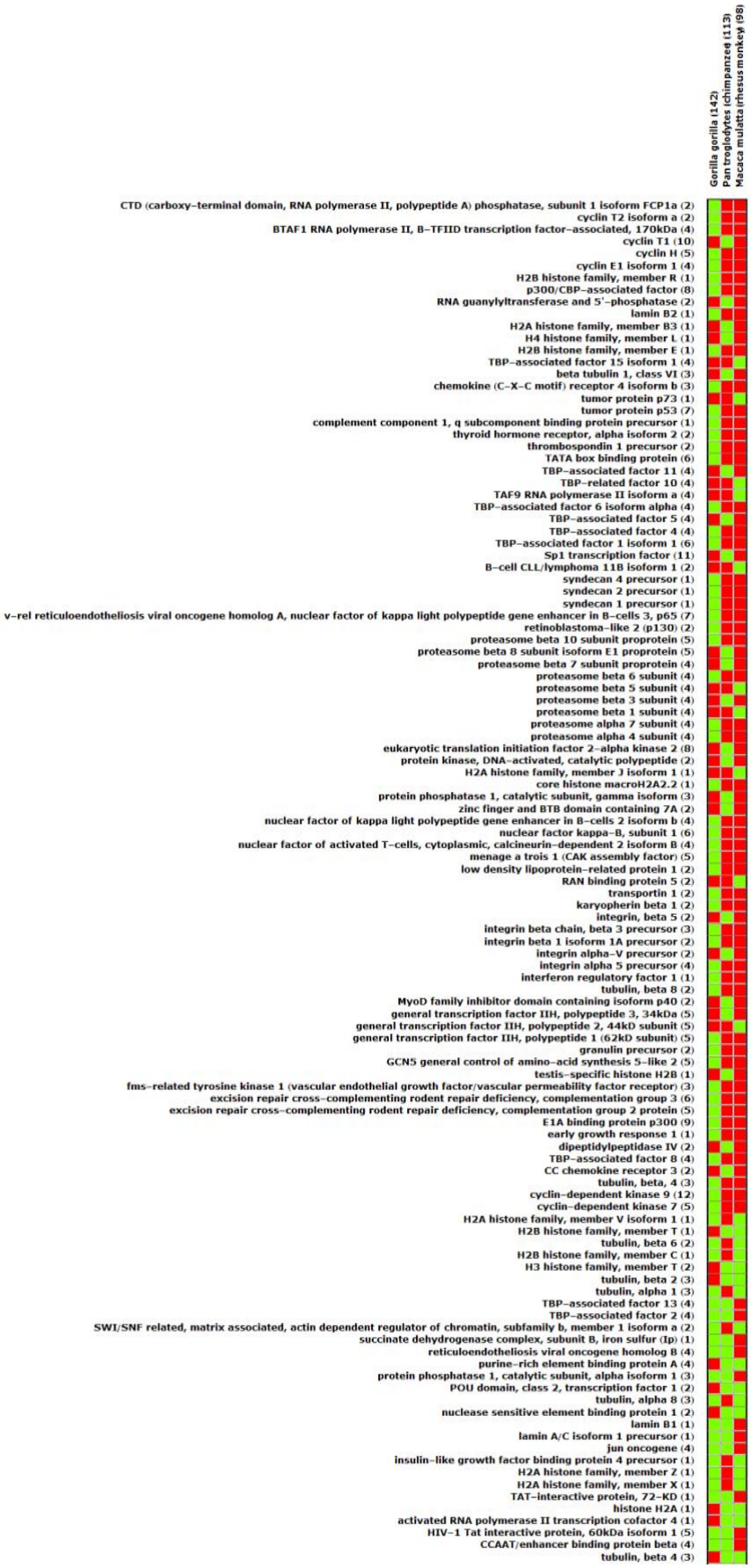


Figure S2.10 Complete heat map for comparison of human proteome associated to HIV-Tat binding activity and that found conservation at functional orthologs level in the [FO] clusters with 3 primates.

Each row corresponds to human proteins (NCBI protein name) that physically bind to HIV-Tat protein and each column corresponds to one of the three non-human primates under analysis. The numbers in parentheses represent frequency of regulatory functions those are associated with the HIV-Tat interactions (rows) and in the other frequency of total functional orthologs found in the primates (columns). The interacting proteins are sorted by increasing average network distance between human and each of the other primates based on sequence proximities found between protein pairs of the ***FO***-clusters. The proximities were calculated by *F-scores* (see methods). The organisms are sorted by increased average distance of the networks for all the interacting proteins between the organisms represented in the column and human. Distance is calculated as described in methods. Red color indicates relatively absent sets of the sequence proximities between proteins with respect to human. Green color indicates similar sets of the closest sequenced proteins with respect to human. ***Enlarged figure is available as Figure S2.10 in the CD that is provided with this thesis.**

Chapter 4. Homol-MetReS: An integrated framework tool to study evolutionary molecular systems biology

4.1. Abstract

Model organisms facilitate the study of other species that may be hard to analyze directly. However, the ability to characterize a given process in specific organisms does not ensure that those organisms will be appropriate models for the way the process works in the organism one is interested in. Recently, a method that compares the set of proteins involved in a given process in different organisms was proposed as a way to identify specific organisms that are likely to be appropriate models of the way a process works in larger classes of living beings.

Here, we report the development of this method into a web application, Homol-MetReS, that will allow the users to efficiently apply the method to compare molecular circuits between any numbers of organisms. To facilitate such comparisons, the tool permits functional (re)annotation of proteomes, to properly identify both, the individual proteins that are involved in the process(es) of interest, and their function. It also permits direct comparison of the sets of proteins involved in the process(es) in different organisms.

In order to illustrate the automation of the tool, we apply it to compare the whole proteome of *Saccharomyces cerevisiae* to that of 57 other eukaryotes and compare these results with those presented in chapter 2. We thus better identify the processes and organisms for which the yeast is likely to be a good model. In addition we apply the tool to analyze and compare the full proteomes of different malaria parasites, identifying the differences in the enzyme complement of those parasites. Those differences are then related to the disparities in the virulence and drug sensitivity of the various parasites.

4.2. Introduction

Understanding at the molecular level how an organism responds to an environmental stimulus or executes a specific biological process requires a laborious research process. A critical step of this process is identifying the proteins and genes that mediate the response. Once this is done, their individual functions must be established. In addition, the physical and functional interactions between the different proteins that coordinate the process of interest must also be understood. Integrating all this information facilitates reconstructing the molecular network that regulates the process. In principle, that reconstruction can be used to predict how the network will respond in different situations and to alternative stimulus, for example through the use of mathematical models. If the predictions are confirmed and a sufficient level of understanding is achieved, the organism may become a model to study the way in which that process might work in other living beings.

Model organisms facilitate the study of other species that may be hard to analyze directly because of one or more of the following reasons [147]:

a) They are technically hard to experiment upon,

b) Ethical issues hinder experimentation, or

c) The developmental processes and time scale of the organism's response is too slow.

However, characterizing a given process in specific organisms does not provide any guidance about which of these organisms to choose as a model for the same process in an organism that may be hard to study.

Recently, a method was proposed to choose appropriate model organisms to study a specific biological process [147] (as defined in Chapters 2 and 3). This method consists in first identifying the proteins that participate in the processes of interest in the different organism. Then, a comparison of the network of protein functions between the different organisms is performed. Finally, we select the organism with a set of proteins as similar as possible to that of our organism of interest and with the possibility of being experimented upon. The method was applied to study the adequateness of *Sacharomyces cerevisiae* as a model organism to study different processes in approximately seven hundred organisms.

To efficiently apply that method to other cases, users need a tool that simultaneously allows them to

a) Properly identify both, the individual proteins that are involved in the network(s) of interest and their function, and

b) Compare the networks of interest between different organisms.

Proper functional identification of genes on a full genome scale and for all organisms with fully sequenced genomes is only possible by using the functional information that is available for proteins in other genomes. Such functional information is transferred to the proteins in the new genomes through the use of sequence homology. In short, orthology (and homology) between an uncharacterized protein and another with known function is used to transfer, either partially or in full, the functional annotation of the latter protein to the former [205, 206]. This procedure relies on the ortholog conjecture [207]. Using additional information, such as synteny [208], or metagenomics context can improve the accuracy of this information transfer [209]. The information transfer process is often automated and its accuracy critically depends, among other things, on the correctness of the functional annotation that is available. Manual partial re-annotation of functional information by researchers often improves that accuracy. Even though many high quality tools and workflows are available for proteome/genome (re)annotation [210-218], using them typically require programming skills that experimental scientists often lack the time to develop. Proper comparison of networks on a large scale is, as far as we know, not a functionality that is available on widely used genome analysis tools and needs to be done almost manually, for example using PathBlast [219], KEGG [220], or MetaCYC [81]. In addition, no tool that we are aware of permits simultaneous large scale re-annotation, functional integration and network comparisons.

In this work we aimed at developing and testing a prototype web application that would enable researcher to apply the method described in [147], by providing the functionality discussed in the previous paragraph. This web application, Homol-MetReS, is available at **http://homolmetres.udl.cat**. It was designed to provide a user-friendly pipeline of methods to re-annotate or transfer functional annotation between proteomes, and to evaluate how similar the networks of proteins involved in a specific biological process are between different organisms. The application implements functionality for

a) Comparing full proteomes using sequence homology,

b) Functionally (re)annotating the relevant proteins of said proteomes,

c) Generating heat maps that easily rank the similarity between organisms in a list with respect to the set of proteins that may be associated to the relevant functions, reactions or biological processes and pathways/networks.

We verify the accuracy of the prototype by reanalyzing two previously published case studies and deriving new information from the comparisons in each of them.

4.3. Results

4.3.1. Homol-MetReS

Homol-MetReS is a web application that enables rationally identifying appropriate model organisms in which to study the functioning of a specific function, biological process or circuit and from which to extrapolate the results to other organisms where that function or process is hard to study. The method to identify appropriate model organisms requires that at least some of the alternatives being considered have a reasonably accurate functional annotation of the proteins that participate in the process or circuit of interest. As explained in more detail in [147], the method compares the set of proteins that participate in the relevant process(es) in the group of organisms being considered and identifies the organism in which the network is the most similar to that of one's organism of interest. Doing so requires that Homol-MetReS provides distinct functionality to the user in a tightly integrated manner. In order to facilitate this task, Homol-MetReS provides three modular, yet integrated, central functionalities to the user.

First, users can (re)annotate the function of each of the proteins in the proteome of an organism of interest with respect to many different classifications of biological function: GO [221], Pathways [220], EC Number [222], Protein name, Interactions [223], Receptor function [224], Ligand function [224], and Substrate [222]. In addition, the application provides a category for personal functional classification definitions. Homol-MetReS includes manual and automated functional annotation modes (**Figure 4.1**).

Second, users can compare the sequences of the individual proteins from their proteome of interest to those of the full proteome from more than 1200 other organisms that have fully sequenced and annotated genomes. This functionality can be integrated with that for (re)annotation. Functional information from one organism can be transferred to another by the user, based on sequence homology. The application also permits identifying candidate proteins for missing functions, proteins that are absent in specific organisms and gene duplication events. The process for doing this is illustrated in .

Third, users can graphically compare the network of proteins that participate in a given set of functional categories between alternative organisms, in order to identify which organism have networks that are more similar to all others within each category **Figure 4.3**. This functionality can be integrated with the other two and permits the navigation between graphical representations of the analysis for different functional categories. It also visually identifies the organisms that could be appropriate models to study a process of interest in an organism in which that process is hard to study. Data can also be downloaded in zipped format.

4.3.2. Using Homol-MetReS

Figure 4.1, **Figure 4.2** and **Figure 4.3** (see above) represent a summary of what users can expect in Homol-MetReS. To facilitate both, security and experiment management, one needs to register to use the application for the first time. This allows the server to create a set of directories where the results of all experiments developed by that user will be stored. Once registered, the user can login into the application (**Figure 4.1 A**) and create different projects. Each project corresponds to the analysis and comparison of an organism of interest to other organisms from the database. Once a specific organism is chosen, the user can select any set of proteins that are of interest in the full proteome of that organism (**Figure 4.1 B**). The application provides a search facility to identify the different proteins of interest. If a (set of) protein(s) is either not included in the database or not properly annotated, the user has the option of adding or annotating that (set of) protein(s) to be used in subsequent proteome comparisons. At this stage, users can set about to perform the comparison of the proteins from their organism with the proteome of other organisms that they also must select (**Figure 4.1 C**).

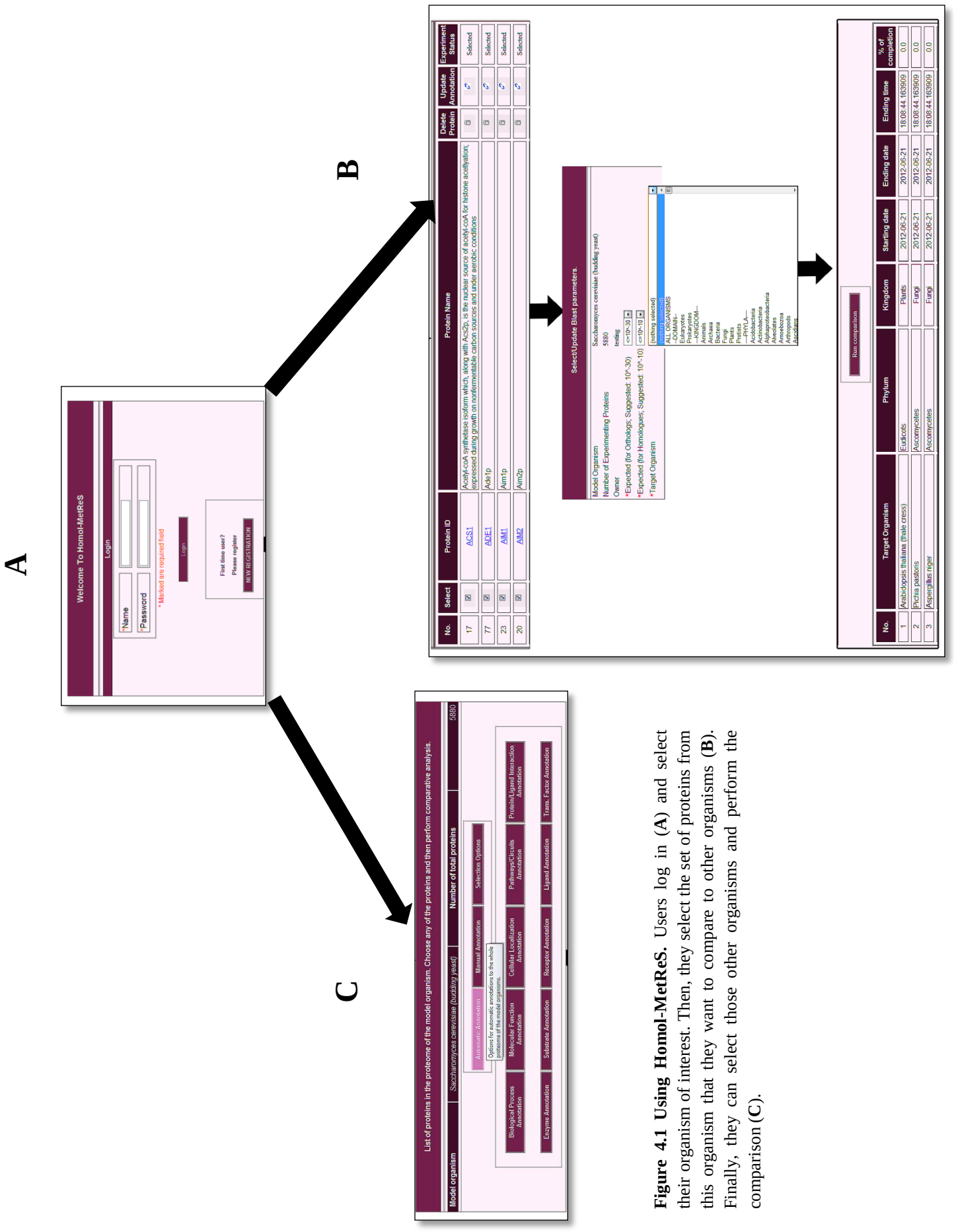


Figure 4.1 Using Homol-MetReS. Users log in (**A**) and select their organism of interest. Then, they select the set of proteins from this organism that they want to compare to other organisms (**B**). Finally, they can select those other organisms and perform the comparison (**C**).

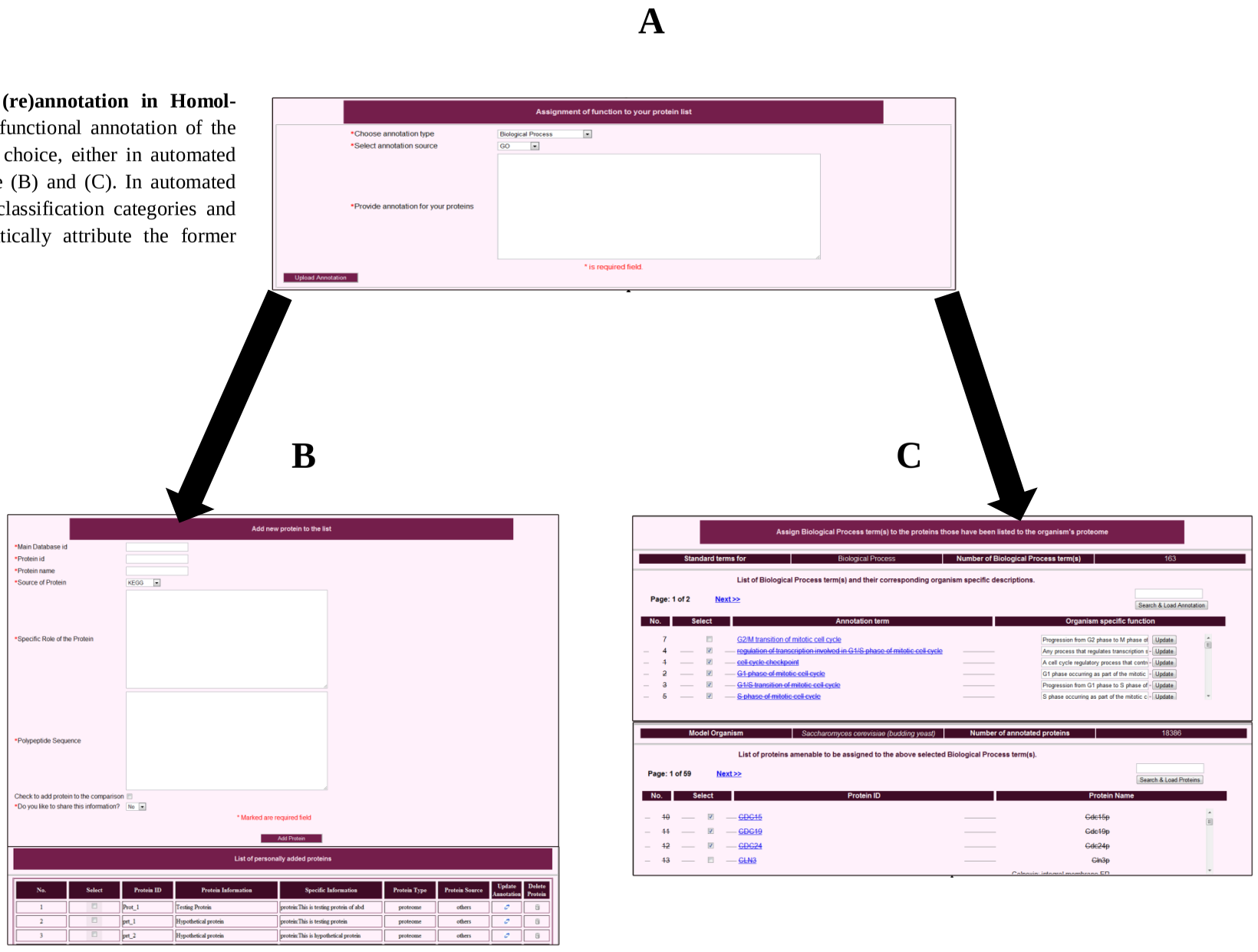


Figure 4.2 Functional (re)annotation in Homol-MetReS. Users can perform functional annotation of the proteins in their organism of choice, either in automated mode (A) or in manual mode (B) and (C). In automated mode, users select different classification categories and various proteins and automatically attribute the former function to the later proteins.

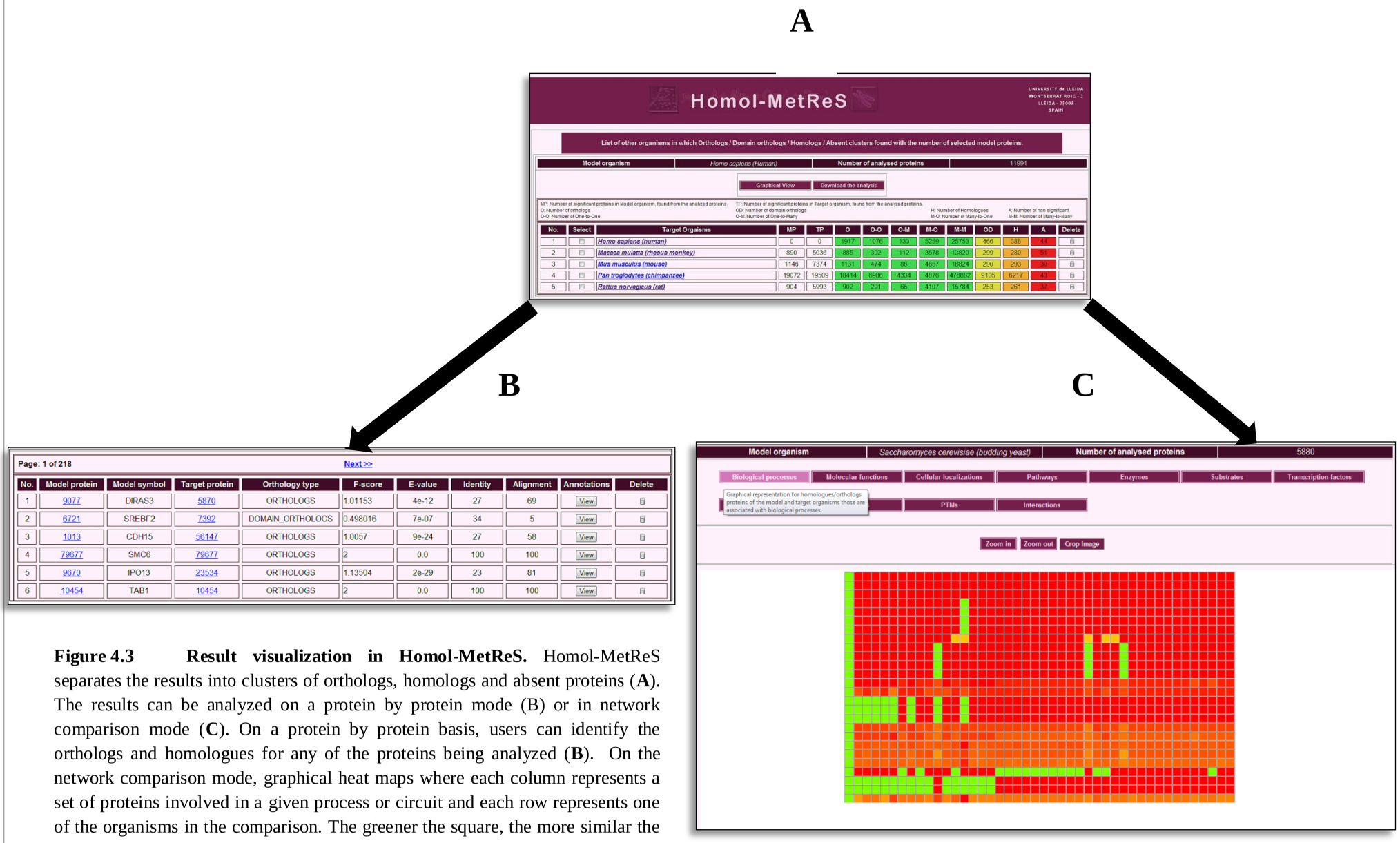


Figure 4.3 Result visualization in Homol-MetReS. Homol-MetReS separates the results into clusters of orthologs, homologs and absent proteins (**A**). The results can be analyzed on a protein by protein mode (B) or in network comparison mode (**C**). On a protein by protein basis, users can identify the orthologs and homologues for any of the proteins being analyzed (**B**). On the network comparison mode, graphical heat maps where each column represents a set of proteins involved in a given process or circuit and each row represents one of the organisms in the comparison. The greener the square, the more similar the set of proteins in that organism is to the corresponding set of proteins from the original organisms of interest.

To perform the comparison, the user should select cut-off values for what are to be considered orthologs and homologs. Users should also select the threshold value below which no sequence similarity is considered as significant between organisms. Although such values are case specific, a default value of 10^{-30} for the e-value and 30% for the identity appears to work well in most cases for the ortholog selection. For homologue selection, reasonable default values are 10^{-10} for the e-value and 20% for the identity. Once the comparison is set and running, the user can log off. Homol-MetReS will send a message to the registered e-mail once the comparison is fully done.

Alternatively, before performing the comparison, users can provide functional annotation for the proteins in their organism of interest (**Figure 4.2. A**, **B**). They can do so in automated (**Figure 4.2. A**) or in manual mode (**Figure 4.2 B**). In automated mode, the user will be able to associate the protein(s) of interest to different subcategories within the categories mentioned above, simply by clicking on boxes and saving the result. In manual mode, users provide a tab or coma separated text where each protein is identified with its NCBI entry, followed by the functional annotation. This annotation can be done with respect to the following categories: GO ontology, Pathways and circuits, Transcription Factors, Receptor and/or Ligand functions, Enzymes and/or Substrates, Interactions, and Post Translational Modifications. Homol-MetReS automatically verifies if the entry is already present and avoids storing repeated entries.

Once the comparison is done (**Figure 4.3 A**) the user can analyze the results for each protein individually, looking for gene duplications/deletions between organisms (**Figure 4.3 B**). This provides useful information about the relative evolution of the organisms being compared. In addition, users can transfer functional annotation between organisms. Finally, the user can visualize the results in different forms. Pie charts display information about the percentage of proteins that is found for each functional category in the genomes of the comparison. In addition, heat maps are used to represent the differences between the sets of proteins that participate in specific process/circuit/other biological functional category in different organisms (**Figure 4.3 C**). A green colored square indicates high similarity, while a red colored square indicates low similarity. At this stage the user can decide which alternative organism is the best for performing experiments.

Homol-MetReS is highly modular and efficient when it comes to comparing different organisms. Comparisons are centered on modular categories of biological function (GO,

Enzymes, Interactions, Pathways, etc.). When a comparison is done in the functional context of one of those categories, for example Pathways, Homol-MetReS permits transferring the sequence relationships that are found to a different comparison between the same organisms, now in the context of another category, for example GO Biological Function. This saves significant time because BLAST does not have to be run again for comparison of proteins that simultaneously belong to different modular categories. All the results are stored in the user's folder for 7 days. These results can be downloaded in ZIP format. After 7 days, the application considers the results to be obsolete and automatically deletes them.

4.3.3. Case studies in Homol-MetReS: *Saccharomyces cerevisiae*

In order to benchmark Homol-MetReS we have repeated a subset of the analysis reported in [147] and compared the results obtained then with the results obtained automatically through the use of the Homol-MetReS application. In the previous analysis we had analyzed *Saccharomyces cerevisiae* as a model organism for different biological processes and pathways in bacteria, archaea and eukaryotes. Here we update only the eukaryotic part of the comparison. As *S. cerevisiae* is widely used as a model organism to study many different molecular aspects of eukaryotic cell behavior, this analysis is important to establish the appropriateness of that use. The analysis using Homol-MetReS found 5880 proteins in the most recent version of the *S. cerevisiae* proteome. The assignment of biological function for the whole proteome was done for five functional categories: Enzyme assignments, Biological process assignments, Pathway assignments, Molecular function assignments, and Localization assignments. The functional annotation for the proteome was downloaded from SGD (Saccharomyces Genome Database), and introduced into the *S. cerevisiae* database in Homol-MetReS using automatic annotation (see supplementary figures for details). The frequency of the subcategories in the whole proteome is shown in **Figure 4.4**. In the case study we annotated terms for 388 enzymes, 110 biological processes, 53 molecular functions, and 19 localizations to the full proteome of the *S. cerevisiae*. Within each of these functional categories, Homol-MetReS provides graphical pie chart representation of each of the functional categories for percentage frequency occurrence of each protein that are attributed to the corresponding terms.

Supplementary **Figure S3-S5** of [147] (**Figure S1.3-Figure S1.5** of the **Chapter 2**) analyze how similar the networks for circuits involved in different biological processes, molecular functions and pathways are between *S. cerevisiae* and 704 other organisms. In the

current work we performed a similar comparison of the proteome of *S. cerevisiae* to that of the 57 eukaryotes that were analyzed in [147]. The organisms that have biological circuits predicted to be more similar to those of *S. cerevisiae* are humans, rhesus monkeys, and chimpanzees. This can be seen in Supplementary **Figure S3.3-Figure S3.9**. As it was described in [147] *S. cerevisiae* is likely to be a reasonable model to study processes like "Cytoplasmic translocation" (168 proteins), "Translational elongation" (38 proteins), "RNA modification" (84 proteins), and iron sulfur cluster assembly (5 proteins) in primates. In contrast, *S. cerevisiae* is less likely to be a good model to study the following processes in primates: "vitamin metabolic process" (47 proteins), "amino acid transport" (44 proteins), "mitochondrial translation" (115 proteins), "cell wall organization" (219 proteins), among others.

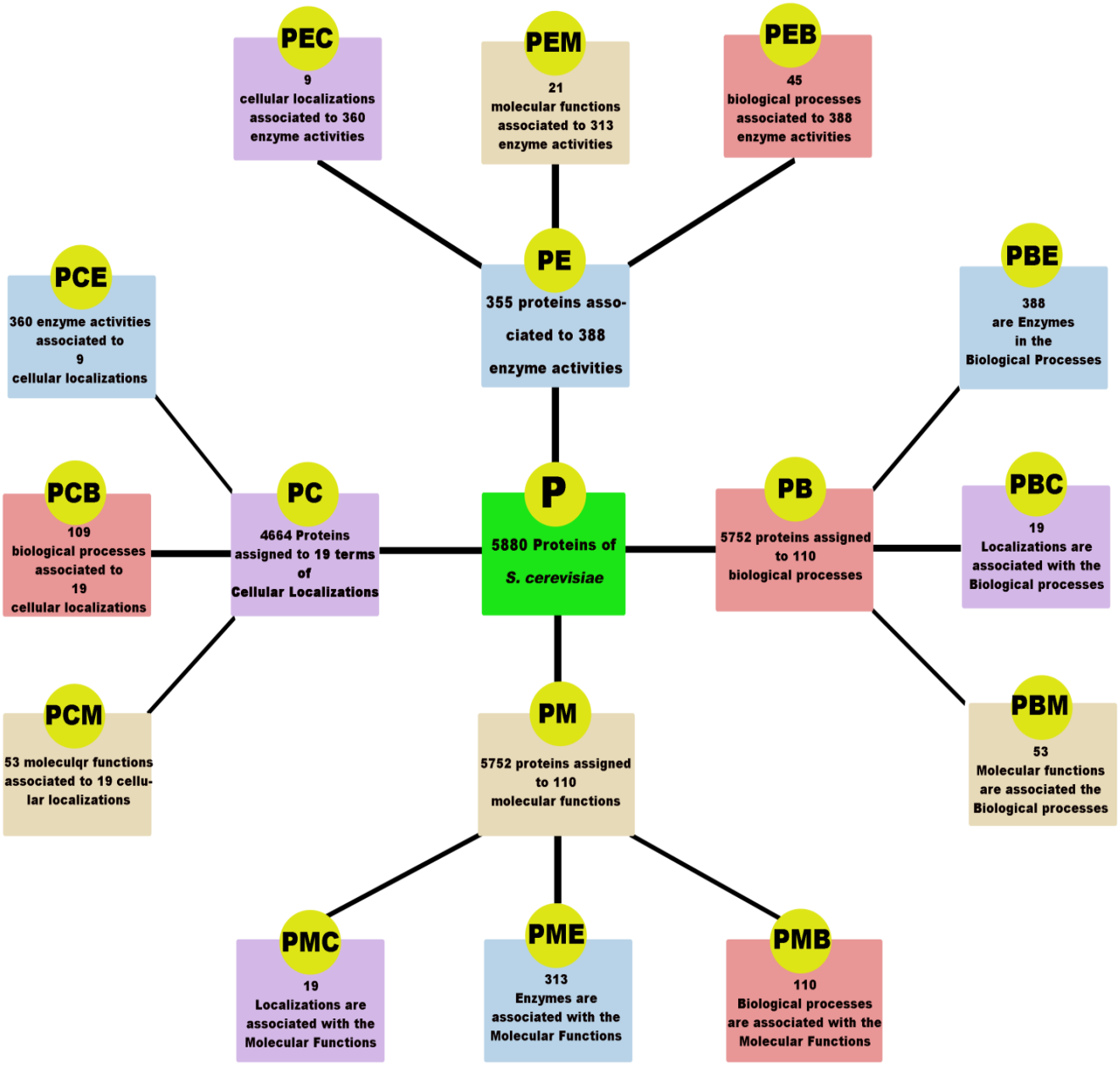


Figure 4.4 Integration of functional annotation for the 5880 proteins in the *Saccharomyces cerevisiae* proteome. Homol-MetReS permits integrating the functional annotation of these proteins between different classification schemes. For example, out of the total 5880 proteins in the yeast proteome (P), 355 proteins are enzymes (PE), spawning 388 different enzyme activities. These enzymes are further distributed into 21 different molecular functions terms (PEM), 9 cellular localization terms (PEC) and 45 biological process terms (PEB). The same way molecular function attributed proteins (PM) integrated with the localizations (PMC) and biological processes (PMB). Such an integrated analysing can start at the level of any functional category.

In general the results of these comparisons are similar to those in Chapter 2 [147]. Differences are due to two factors. First, Homol-MetReS uses the more detailed full GO classification, as opposed to the previous analysis, which relied on GOSLIM. The later, simpler, classification has 32 biological processes, 21 molecular functions and 21 cellular component terms, while the full GO has 20912 biological processes, 9812 molecular function and 2931 cellular component terms. Second, more proteins in the proteome of *S. cerevisiae* have functional annotation in our current database than when the analysis reported in **Chapter 2** [147] was done. In that paper 71%, 53% and 82% of all *S. cerevisiae* proteins had functional annotation respectively for biological process, molecular functions and localizations. Now, the corresponding numbers are 98%, 89% and 79%. It must be remarked that in the current analysis we used only eukaryotes (57 organisms), whereas in the chapter 2, 700 organisms were used. In addition, the current study also compares the enzyme complement of the different organisms directly. As in chapter 2, we ranked the organisms based on the proportion of homologues, domain orthologs, orthologs and duplications that they have in the various functional categories with respect to *S. cerevisiae*, . The organisms that are more similar to *S. cerevisiae* for each of those categories are summarized in **Figure 4.5**.

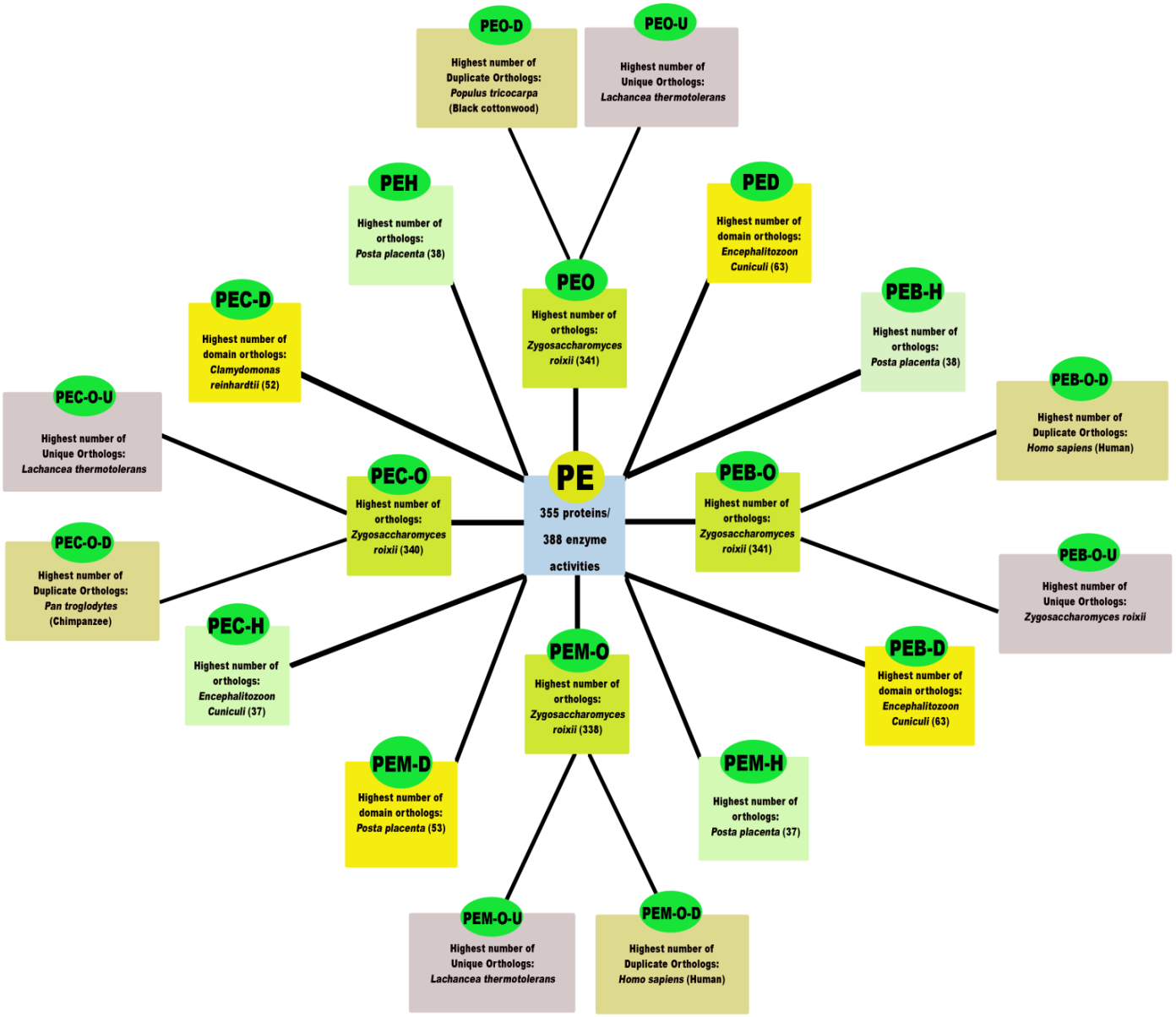


Figure 4.5 Comparative functional analysis of the integrated enzyme component of the yeast proteome with other eukaryotes at different homology levels. Homol-MetReS permits an integrated visualization of the homology and duplication patterns of yeast proteins in other organisms, accounting for the functional information associated to the proteins. We show the results of such an analysis for the enzyme component (**PE**) of the proteome from *S. cerevisiae*, annotated in the Figure 4.4. Orthologs (***O***); Domain ortholog (***D***); Homologues (***H***); proteins with orthologs that are unique (***O-U***); proteins with orthologs that are duplicated (***O-D***). In addition to this, Homol-MetReS automatically associates this information to proteins' localization (**PEC**), molecular function (**PEM**) and biological process (**PEB**) information terms.

During the process of analysis Homol-MetReS permits correlating different functional aspects of the proteins. For example, the annotated enzyme complement of *S. cerevisiae* contains 388 enzyme activities (EC numbers), and it appears to be 30% similar to that of primates (see Supplementary **Figure S3.6** for more details). Many of the enzymes that are not conserved between the yeast and primates are involved in *S. cerevisiae* processes that are also not well conserved in primates with respect to the yeast (Figure 4.5 for summary, Supplementary **Figure S3.3-Figure S3.9** for complete comparison with respect to different functional classifications). One of those processes is "Vitamin Metabolic process" (Supplementary **Figure S3.3**), with 40 proteins annotated as participating in *S. cerevisiae*. Out of these 47 proteins, 17 are enzymes. These enzymes are absent from the enzyme complement of primates (Supplementary **Figure S3.6**). This example illustrates how the Homol-MetReS platform can be used to perform and facilitate such comparisons, enabling the correlation between different functional classification categories because it tightly integrates those classifications together.

4.3.4. Case studies in Homol-MetReS: Malaria Parasites

Parasites from the *Plasmodium* genus are responsible for malaria, a disease with an enormous human and economic worldwide impact. Several species cause the disease, with different etiologies. *Plasmodium falciparum* appears to be the most lethal, while *Plasmodium vivax* is less lethal but more recurrent [225]. Here, we select some of these parasites as a case study to further illustrate the usefulness of Homol-MetReS. The application is used to compare the proteomes of the different *Plasmodium* species and strains with fully sequenced genomes. The results of this comparison can help understand the varying effects of some anti-malaria drugs on different parasites. In addition, we also compare their genomes with those of human and chimp, suggesting explanations for some differences in the etiology of the disease between the two primates.

We anchor the study in *Plasmodium vivax*, which is the malaria parasite more frequently associated with recurring malaria. The latest version of *P. vivax*'s genome has 5393 proteins in its full proteome. 222 enzyme terms are automatically identified using the text mining facilities available in Homol-MetReS. These are used to identify the proteins with an associated E.C. number in the annotation of their sequences, which are stored in FASTA format. These 222 distinct enzymes are associated with 314 proteins in the *P. vivax* proteome.

The difference in numbers is a consequence of enzymes that are composed by more than one subunit and of enzyme activities that can be performed by more than one protein.

After the enzymes annotation stage is over, we compare the full proteome of *P. vivax* to those of *Plasmodium berghei*, *Plasmodium chabaudi*, *Plasmodium falciparum 3D7*, *Plasmodium falciparum Dd2*, *Plasmodium falciparum HB3*, *Homo sapiens* (human), and *Pan troglodytes* (chimpanzee). Some of the results are summarized in **Figure 4.6**. The comparative analysis for the complete enzyme list is available in **Figure S3.10** (the **Figure S3.10** is provided in the CD of the thesis). Most of the identified enzyme activities are common to all *Plasmodium* parasites in the comparison. The ones that are specific to some of the species or strains could be used as markers for understanding physiology of host-pathogen relationships. A few examples of how this can be done will now be discussed.

"*Adenylosuccinate synthetase*" (EC: 6.3.4.4) is the first enzyme in the purine salvage pathway of *Plasmodium*. This pathway is involved in the salvaging of the host's purines for the synthesis of DNA by the parasite [226], and it is an important target pathway for drugs that kill the parasite [227]. That enzyme is absent in *Plasmodium falciparum Dd2*, a clone from Indochina/Laos. It is well established that the *P. falciparum Dd2* strain is more resistance to chloroquine than the HB3 strain [228, 229]. There is also evidence that the Dd2 clone has a high propensity to acquire resistance against drugs that target the purine salvage pathway, whereas *P. falciparum* HB3 does not [230, 231]. Given that these drugs are high affinity analogues of the transition state of an enzyme in the pathway, the absence of enzyme orthologs in the pathway implies that there is either an alternative salvage pathway or an alternative enzyme that replaces the one working in other *P. falciparum* strains.

(See Figure 4.6 on next page)

Figure 4.6*

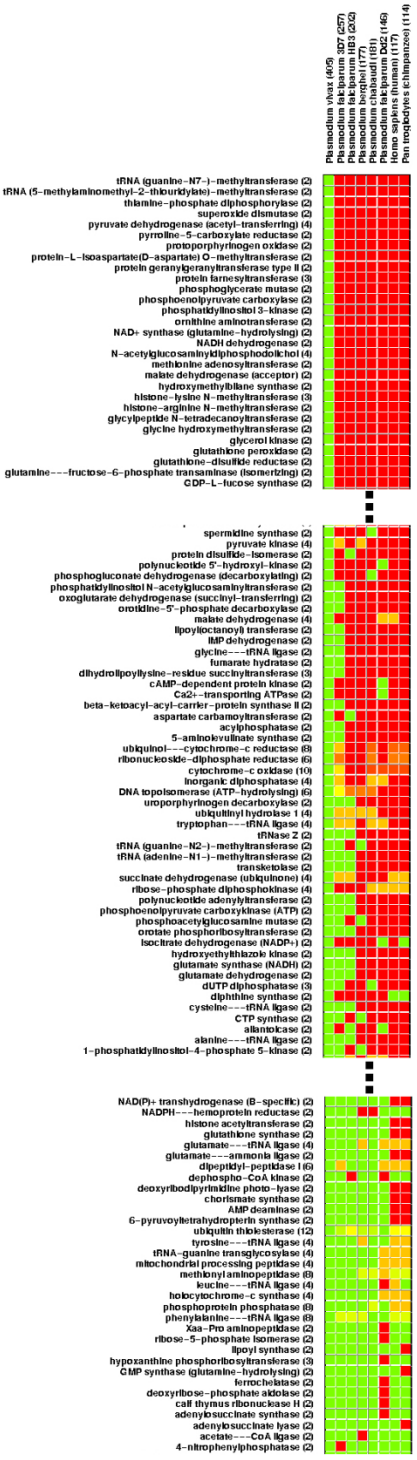


Figure 4.6 Comparing the enzyme complement of different organisms from the *Plasmodium* genus. Humans and chimps are also included in the comparison. This figure shows a detail of the enzymes that are not common to each of the organisms. Enzymes were annotated using the automated annotation mode of Homol-MetReS. *P. vivax*, has 314 annotated enzymes and was used as the central organism in the comparison. Orthologs for that enzyme set were searched in the other organisms. Red colour indicates that the sequence ortholog is absent in an organism, while green colour suggests presence of the enzyme. The complete analysis is shown in Supplementary **Figure S3.10**. ***Enlarged figure is available as Figure S3.10 in the CD that is provided with this thesis.**

"*Porphobilinogen synthase*" (EC: 4.2.1.24) is an enzyme that is involved in heme biosynthesis. It is present in all Plasmodium species but *Plasmodium falciparum* HB3, where it is absent. The *Plasmodium* enzyme localizes to various compartments of the parasite (apicoplast, mitochondria and cytosol). It has a low catalytic efficiency when compared with the corresponding enzyme of the host [232]. In fact, it was reported that the parasite can import the host enzyme and use it for heme biosynthesis during the intraerythrocytic stage of infection [233, 234]. Thus, it is conceivable that the HB3 strain has completely lost the gene coding for this enzyme and that this strain only uses the protein when imported from the host. We also note that the low catalytic efficiency of the parasitic enzyme is probably due to the fact that it has more than one enzyme activity [232]. This leads us to speculate that there may be some other enzyme with multiple activities that could replace the native "*Porphobilinogen deaminase*" when the parasite is not infecting a host.

Our analysis also reveals interesting results with respect to enzymes that are usually involved in energy metabolism. First, it was found that both subunits of "*Pyruvate dehydrogenase*" (EC: 1.2.1.51), which is localized in apicoplast and involved in lipoylation of proteins [235, 236], are absent only in the genome of *P. chabaudi*. This enzyme appears to be important only in the late stages of the development of the disease in liver [237, 238]. *P. chabaudi*, together with *P. yoelii* are mouse specific malaria parasite. Previous experiments show that mice that are deficient in enzymes related to pyruvate metabolism are more resistant to infection by *P. chabaudi* [239]. Taken together, these observations suggest that *P. chaubadi* might be using the host's enzymes to perform the function that its cognate PDH should perform.

Second, "*Aspartate transaminase*" (EC: 2.6.1.1) is absent from the genome of *P. chabaudi*. How this correlates to any phenotype that is *P. chabaudi*-specific is unknown. However, given that fumarate is generated as a side product of the purine salvage pathway and that tricarboxilic acid cycle related enzymes appear to function in a biosynthetic capacity in the malaria parasites [240], it could well be that some less specific enzyme replaces this activity in *P. chabaudi*.

4.4. Discussion

In this work we present a web application, Homol-MetReS, whose purpose is two-fold. On one hand it aims at facilitating whole proteome functional (re)annotation. On the other, it aims at using this annotation, combined with sequence comparison, to predict how similar the network of proteins involved in a given biological process is between different organisms. Together, these two features facilitate identifying an appropriate model organism to study a given process, if for some reason that process cannot be appropriately studied in one's organism of interest. This identification is done through the comparison of the set of proteins involved in the same process in the different organisms.

Such comparisons are possible because of the accumulation of fully sequenced and annotated genomes since 1995. The methods implemented in the application have been previously developed and tested manually using *S. cerevisiae* as a case study [147]. In this earlier work *S. cerevisiae* was thoroughly analyzed and compared to 700 other organisms in order to identify the biological processes and pathways in each of those organisms for which the yeast might be a good study model [147]. In this paper we update the analysis from that study considering only eukaryotes, as a benchmark to ensure that Homol-MetReS is working appropriately. We find that results are similar, yet more accurate and specific, as we now use more detailed functional classifications. In addition, the application saves approximately 80% of the time it would take to perform the same study in a similar way as in [147].This is partially due to the fact that earlier sequence comparisons remain stored and need not be performed again.

We have also performed a comparative analysis of the proteomes of various malaria parasites among themselves. They were also compared to the human and chimp proteomes. Interestingly, some differences are found in the enzyme complement of the different parasites and, in some cases, those differences can be correlated with differences between the different strains of malaria parasites in their infectious behavior or resistance to treatments.

Because Homol-MetReS does not focus on the genome, but rather on the proteome, it is useful for understanding the comparative functional evolution between the proteins of different organisms. The comparison method used by the application permits differentiating between paralogs and orthologs. This differentiation is crucial for appropriate functional comparison. However, our network comparison is robust to mistakes in that differentiation,

because as long as one of the paralogs is a real functional ortholog, the network being compared will be similar.

Homol-MetReS has partial functional overlap with other tools [81, 195, 219, 241-255]. A list of some of the most widely cited is shown in **Table 4.1**. A comparative summary of their functionality is shown in **Table 4.2**. Homol-MetReS is unique in allowing users to define new functional categories and re-annotate preexisting ones. This is a plus for those users that need to reduce semantic gaps in between existing and required functional definitions. Because such personalized functional annotation could also hinder use by others, Homol-MetReS separates these personal functional definitions and keeps them user-specific. This avoids clashing definitions between different users.

Table 4.1 Summary of functional comparison of other web applications and Homol-MetReS.

Tool	Method for sequence comparison	Classification Schemes	Organims in database
COG	Pairwise sequence analysis using Blastp	Protein names, EC numbers, GO Biological processes	66, Eukaryotes + Prokaryotes
OrthoMCL	Pairwise sequence analysis using Blastp	Protein names	150 Eukaryotes
Homologene	Pairwise sequence analysis using Blastp	Protein names	Eukaryotes
InParanoid	Pairwise sequence analysis using Blastp	Protein names, Interactions	100 Eukaryotes
PHOG	Pairwise and Global sequence analysis	Protein names	25 Eukaryotes
TreeFam	TreeFam infers homology analysis by mean of gene trees	Protein names	28 Eukaryotes
Homol-MetReS	Pairwise sequence analysis using Blastp	Protein names, EC numbers, Full GO ontology, KEGG Pathways, Interactions, Substrates, Receptors, Ligands, Transporters, Post translational modifications	1257, Eukaryotes + Prokaryotes
DODO	Using Domain information and rpsBlast	Protein names, Full GO ontology	----
MicrobesOnline	Pairwise sequence analysis using Blastp	Protein names, EC numbers, Full GO ontology, KEGG Pathways	3705, Eukaryotes + Prokaryotes
EggNOG	Non supervised orthologs grouping	Proteins names, based on COG and KOG	630 Eukaryotes + Prokaryotes
Roundup	Reciprocal Smallest Distance (RSD) algorithm	Protein names, Full GO ontology	1501 Eukaryotes + Prokaryotes
OMA	Co-occurence of orthologous genes in different genomes	Protein names, Chromosome locus, Full GO classification, KEGG Pathways, COG	352 Eukaryotes + Prokaryotes
YETI	Pairwise sequence analysis using Blastp	Protein names, GO Biological process, Molecular functions, Localizations, Interactions, Pathways	----
PathBlast	Protein-protein interaction network based pair wise analysis	Protein names, KEGG Pathways, Interactions	----
BioCyc, MetaCyc, Pathway Tools	Manually curated Pairwise sequence analysis	Protein names, EC numbers, Full GO ontology, KEGG Pathways, Interactions, Substrates, Receptors, Ligands, Transporters, Post translational modifications	>1000, Eukaryotes + Prokaryotes
PROCOM	Pairwise sequence analysis using Blastp	----	32 Eukaryotes
NetAligner	Pairwise sequence analysis using Blastp	Protein name, Interactions and Pathways	Eukaryotes and E. coli
PROMPT	Pairwise sequence analysis using Blastp	Protein names, EC numbers, Full GO ontology	----
Negative Proteome Database	Pairwise sequence analysis using Blastp	Protein name	Eukaryotes

Table 4.2 Summary of functional comparison of other web applications and Homol-MetReS.

Tool	Multiple full proteome comparisons	Discriminates various levels of homology	Detects missing proteins	Functional (re)annotation by user	Integrates various classifications	Graphical representation of analysis	Comparison of functional networks	Allow users to manage their own definition of proteins and functions
COG	•	•						
OrthoMCL		•	•	•				
Homologene				•	•			
InParanoid		•		•	•			
PHOG		•		•				
TreeFam								
Homol-MetReS	•	•	•	•	•	•	•[a]	•
DODO		•		•				
MicrobesOnline	•	•		•	•			
EggNOG		•		•		•		
Roundup			•	•	•			
OMA					•			
YETI		•		•	•	•		
PathBlast			•	•	•	•	•	
BioCyc, MetaCyc, Pathway Tools			•	•	•	•	•	
PROCOM	•	•						
NetAligner			•		•	•	•	
PROMPT	•				•			
Negative Proteome Database	•		•					

[a] As opposed to the other tools, Homol-MetReS permits simultaneous comparison of multiple networks in multiples organisms.

With the exception of Pathway Tools, BioCyc and MetaCYC [81, 256], none of the other tools permits integrating functional information using all the different classification schemes that are available in Homol-MetReS. In addition, the only tool that permits comparing pathways or circuits using homology search is PathBlast [219]. However, doing this in PathBlast requires manually identifying the proteins involved in a specific process and comparing those proteins to only one other organism. In Homol-MetReS this type of comparison can be done in large scale and automatically. An important feature of Homol-MetReS is that it identifies proteins that are absent between any pair of organisms chosen for comparison. This feature is shared with PathBlast [219], the BioCYC suite of applications [81, 256], Netaligner [245] and OrthoMCL [248].

Some tools integrate the determination of homologues or orthologs clusters with the functional annotation of the proteins in the clusters. Homol-MetReS is one of them. However, in Homol-MetReS, this integration covers a wider range of functional classifications than any other we are aware of. In addition, Homol-MetReS permits comparing as many organisms simultaneously as the user decides.

In short, Homol-MetReS provides a wider range of functionality than other comparable tools. It facilitates full proteome annotation and comparison, enabling the identification of appropriate model organisms from which the results of studying specific biological phenomena can be more securely extrapolated to other organisms of interest.

4.5. Materials & Methods

4.5.1. Homol-MetReS implementation

Homol-MetReS runs remotely through a web service on a Linux cluster, under TOMCAT. It is a modular application that is built using Zope 3.0 [257], Python [258], MySQL [259], and Mathematica [260]. It can be accessed using any or the major web browsers, from any of the major operating systems. The application uses Mathematica™ to compare the similitude of networks and build the graphical representation of those comparisons.

4.5.2. Internal database

Underlying and supporting Homol-MetReS, there is a database that was designed and developed specifically for this application (**Figure 4.7**). This database contains all the information for each of the more than 1200 fully sequenced and annotated proteomes available in the tool. In brief, a series of different tables store the information of a list of available organisms, the proteome of each organism, and the functional information about the different categories in GO classification, pathways and circuits, EC numbers, transcription factors, post-translational modifications and interactions. When an individual organism is analyzed, a new database that is organism-specific is automatically created. All information about the proteome of this organism is then stored into this new database to make analysis more efficient. Each protein entry in the proteome table of an organism is connected to the adequate function in those tables, whenever that information is available. Users can insert additional annotation information for proteins that will then be included in the organism-specific database. A summary of this information is shown in **Figure 4.7**.

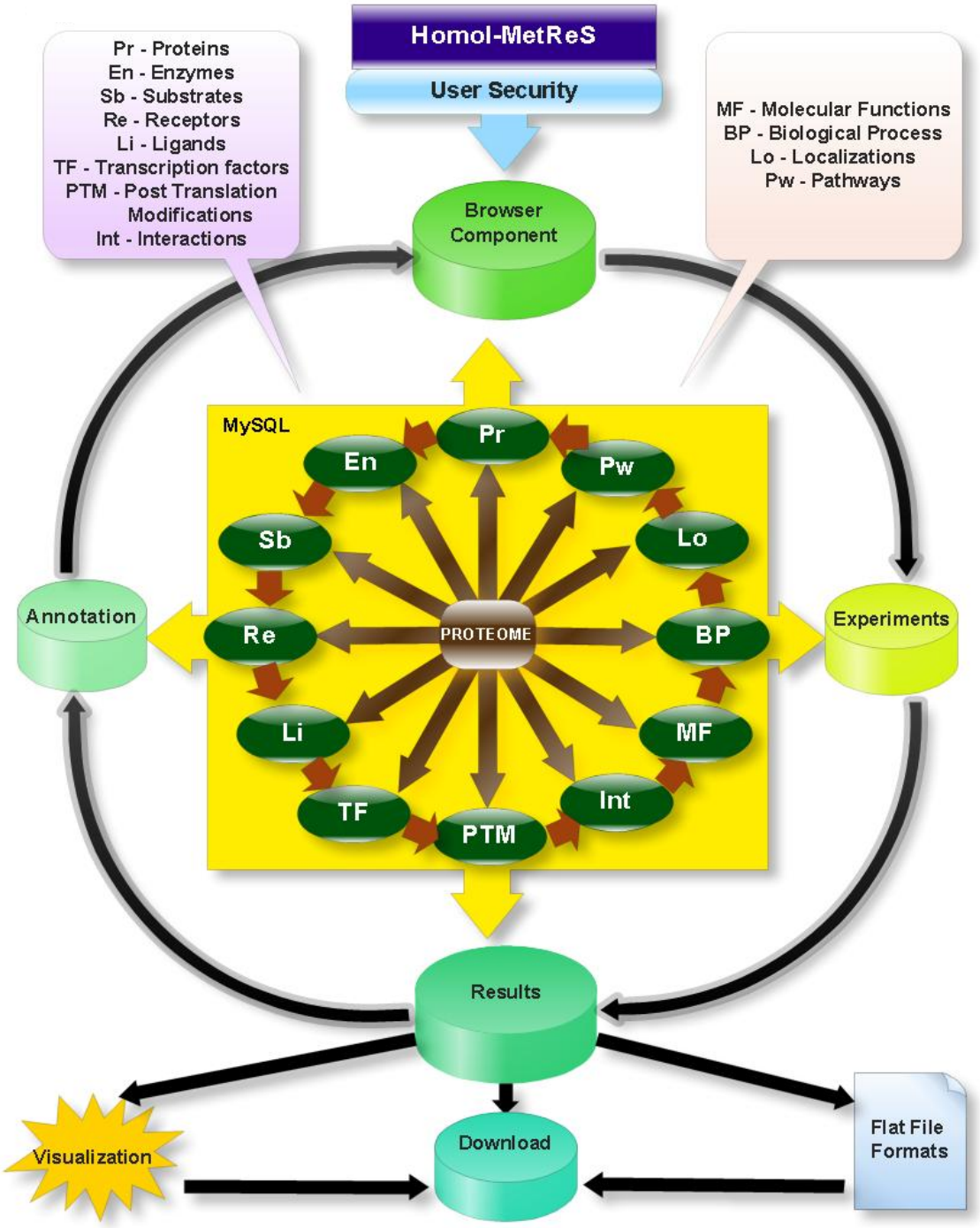


Figure 4.7 Summary of database structure and connectivity between functional modules in Homol-MetReS. Each project starts with an organism. The functional information about the different aspects of the proteome of the organism is stored in the central database. This information includes sequences, GO ontology, EC numbers, transcription factors, receptors, ligands, substrates, interactions, pathways and post-translational modifications. This information can be retrieved and updated through the annotation modules, to facilitate functional (re)annotation of proteins in an integrated manner with any of the annotated information. Once appropriate functional annotation is ready, the comparison modules can be used to perform proteome scale sequence comparisons. The results module accesses the results of the comparison. These can be analysed either protein-by-protein, in individual lists, or in bulk, through the analysis of orthologs homologues and absent genes. The results are accessible to the Visualization Module, which can generate heat maps that compare networks of proteins classified as having similar functionality in different organisms.

The enzyme information terms come from BRENDA [261]. The receptor and ligand information terms are obtained from IUPHAR-DB [224]. The GO ontology information terms about biological process, molecular functions and cellular localization terms are downloaded from the GO database [221, 262]. The pathways information terms are obtained from KEGG [263]. The post translational modification information terms are derived from the Human Protein Reference Database [155]. Because there is no standard general classification for transcription factors, users are not provided with such a standard table. However, they can define such a classification themselves.

Homol-MetReS currently contains 1207 organisms with fully sequenced genomes, together with the functional annotation for each of the genes from any of that organisms. Each organisms are classified based on domain (2 domains: Prokaryotes- 1082, Eukaryotes-129), kingdom (6 kingdoms: Bacteria-999, Animals-46, Archaea-79, Fungi-43, Plants-11, Protists-29), phylum (54 phyla) and class (447 classes). The master databases for standard functional terms include Pathways (KEGG), GO (gene ontology terms: biological processes – 20912, molecular function – 9812, cellular components - 2931), Enzymes (Brenda enzyme terms - 4253), Receptor (IUPHAR receptor terms - 558), Ligand (IUPHAR ligand terms - 2756), and Chemical compounds (KEGG compound terms - 14774).

4.5.3. Proteome Comparison

Comparison of individual protein sequences is done using BLAST [264, 265], which is downloaded from NCBI and incorporated into Homol-MetReS. The comparison of proteomes is done using a pipeline that implements the methodology described in [147] and classifies proteins into clusters of orthologs, homologs, and absent proteins. In brief, this pipeline uses the following process to identify the different types of protein clusters. First, the selected proteins of the organism of interest are blasted against the entire proteome of the target organisms selected by the user. Then, for each individual protein in a proteome, only proteins that have appropriate user-defined e-value, identity and coverage in another proteome are flagged and appended to the cluster of orthologs of the original protein. If more than one protein is identified as being a possible ortholog, a metric described below is used to identify which of them is more likely to be the “true ortholog” of the query protein. This metric, **F,** is used to further classify cluster of orthologs into four different types:

One-to-One Clusters: only one protein in the target organism matches the protein of the organism of interest according to the orthology criteria defined by the user.

One-to-Many Clusters: more than one protein in the target organism matches the protein of the organism of interest, according to the orthology criteria defined by the user.

Many-to-One Clusters: only one protein in the target organism matches more than one protein of the organism of interest, according to the orthology criteria defined by the user.

Many-to-Many Clusters: multiple query proteins in the organism of interest match multiple proteins in a target proteome, according to the orthology criteria defined by the user.

If proteins meet only some of those user-defined criteria, they are appended to the cluster of homologs for the original protein. If homology is further identified in only a portion of the homologous proteins, these are further classified as domain homologues. Proteins of one proteome that have no homologues in another proteome are classified as clusters of absent proteins. At the end of the comparison, clusters of orthologs, domain orthologs, homologs, and absent proteins are provided. This implements the methods described in [147].

4.5.4. Metric for prediction of orthologs

The set of all proteins from a target organism that are appended to the cluster of orthologs of a specific protein from the organism of interest are ranked using a score function ***F***, as defined in [147]. The protein with the highest score function is predicted to be the ortholog of the query protein in the target organism, while the remaining proteins are flagged as in-paralogs of that ortholog and used for gene duplication analysis. ***F*** is calculated as follows:

$$\boldsymbol{F} = (\boldsymbol{F1} + \boldsymbol{F2}) - \boldsymbol{F3} \qquad \textbf{Eq. 1}$$

Factor ***F1*** is calculated as follows.

$$\boldsymbol{F1} = \mathbf{1} - (\boldsymbol{S} - \boldsymbol{I})/\boldsymbol{S} \qquad \textbf{Eq. 2}$$

In Eq. 2, ***S*** represents the similarity score of the alignment (combined score of identical and similar amino acid residues in the alignment), and ***I*** represent the identity score of alignment. Both these values are outputs of BLAST. ***F1*** is always between zero and 1.

Factor ***F2*** is calculated as follows.

$$\boldsymbol{F2 = (AL - G1 - G2)/PL} \qquad \textbf{Eq. 3}$$

In Eq. 3 ***AL*** represents the length of the alignment and ***PL*** is the total length of the query sequence, ***G1*** and ***G2*** represent number of gaps within the aligned region of the query and target sequence, respectively. F2 is always between zero and 1.

Factor ***F3*** is calculated as follows.

$$\boldsymbol{F3 = (G1/L1) + (G2/L2)} \qquad \textbf{Eq. 4}$$

In Eq. 4, ***G1*** is the number of gaps within the aligned region of the query sequence, ***L1*** is the length of the query sequence, ***G2*** is the number of gaps within the aligned region of the target sequence, and ***L2*** is the full length of the target sequence.

Theoretically, $-\infty \leq \boldsymbol{F} \leq 2$. However, in practice its value is always found to be between 0 and 2, if reasonable cut-off values for e-value and identity are chosen.

4.5.5. Network Comparison

One of the purposes of Homol-MetReS is to identify alternative organism on which to study a given process that cannot be studied, for some reason, in the organism one is interested in. The application identifies such alternatives by comparing how similar the set of proteins that execute a given process or participate in a specific circuit is between our organism of interest and the organisms to which it is being compared. Doing this accurately requires that at least some of the organisms one is comparing have appropriate functional annotation associating proteins to the relevant biological process or circuit. A vector V_{HDPi} containing all proteins types associated to the process is created for each proteome P_i in the comparison. Each entry in the vector is one of the protein types. Next, each P_i is individually searched for orthologs to each of the protein types. When a protein type has an ortholog in P_i, the entry corresponding to that protein type is set as 1; otherwise it is set as 0. Subsequently, the Hamming Distance (***HD***) between the vector of the organism of interest and that of the alternative organisms is calculated using the formula $\boldsymbol{HD = \sum_{i=1}^{n} \delta_{i,i}}$. Here, $\delta_{i,i}$is a kronecker delta. It takes the value one if the elements in position *i* of both vectors is the same and zero otherwise. The smaller ***HD***, the more similar the two vectors are and the more similar the set of proteins executing a specific process or functions from the biological component in both organisms, referenced to the organism of interest. Consequently, the more likely it is that the model organism is a good model to study the relevant function or process and generalize the results for the other

organism. Graphical representation of the network similarity data is done using Mathematica [260].

4.5.6. Management of Homol-MetReS jobs and user-specific information

On top of the application structure, each user is provided with their own environment, where they can create independent workspaces that are organism specific. When a user starts a session, the information regarding that organism's workspace is transferred from the MySQL database and temporarily stored into a ZODB (Zope Object Database), which enables faster access times and better memory management. This ZODB is specifically created for each individual session and user. The user has total control over the information generated and stored in this ZODB during each experiment. The whole architecture and procedure is summarized in **Figure 4.8**.

.

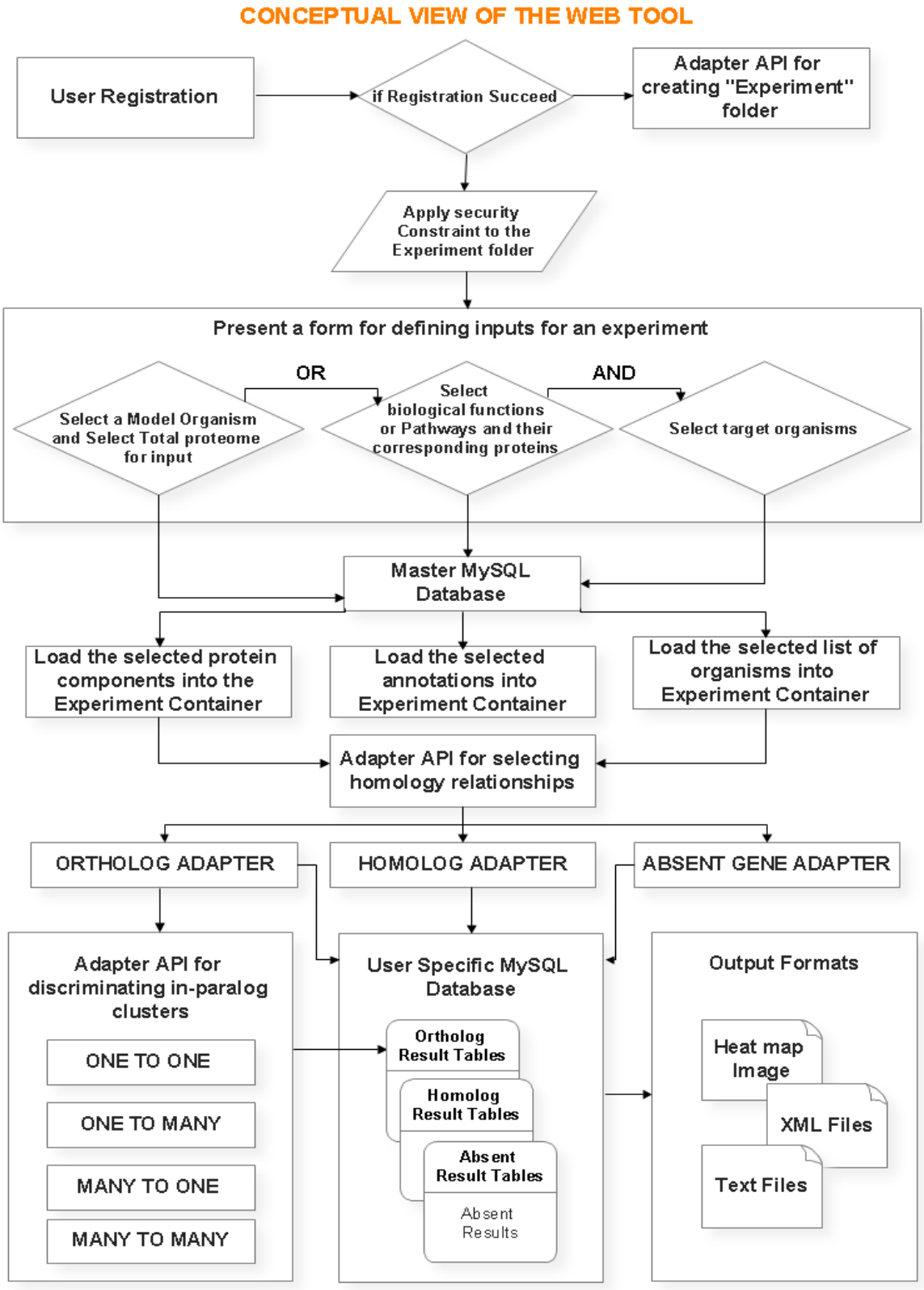


Figure 4.8 **Flow chart for Homol-MetReS functioning.** Users must register, before logging in and creating their organism centric *in silico* proteome comparison. The application retrieves the data for a central MySQL database and performs the sequence comparison between the organisms of choice and any other organism(s) in the database. Results for the comparison are sorted in different clusters of proteins. These can be visualized as xml files, heat maps, or text files. The detail of the implementation is provided in **Appendix 2**.

4.6. Supporting Materials

4.6.1. Appendix 2 - Details for implementation of Homol-MetReS

Overall conceptual implementation of Homol-MetReS

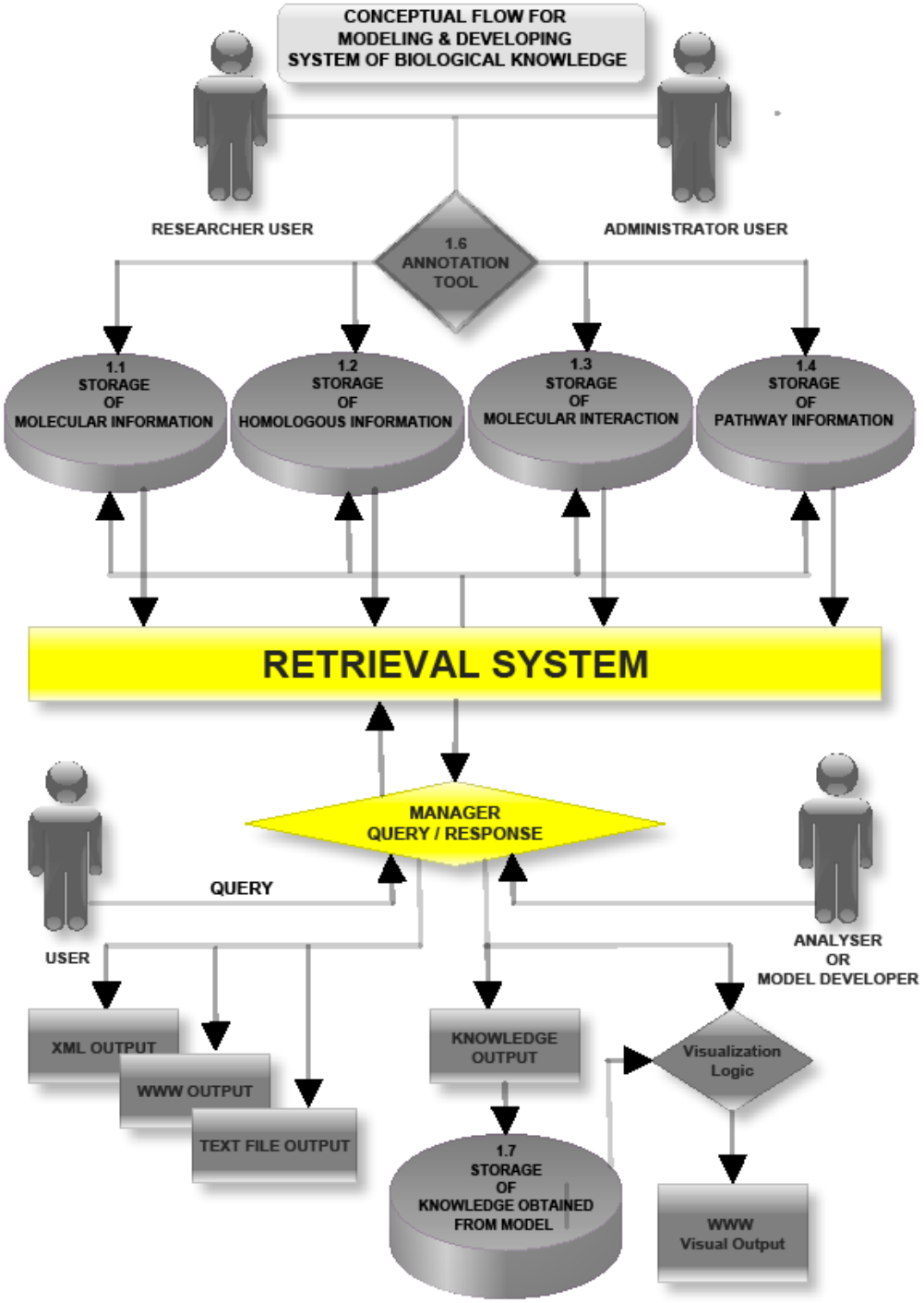


Figure S3.1 Architecture of web application conceptual model for Homol-MetReS.

Homol-MetReS is a common platform that includes various applications. It can be used for annotation, searches, sequence comparisons, information management and visualization of proteome organization from different organisms in an integrated manner. The overall implementation is shown in the **Figure S3.1**.

Homol-MetReS uses a central algorithm to integrate different applications at variable computation/communication granularity. It is designed with a user-centric, organism-centric perspective, permitting the comparison of the proteome organization between two or more organisms, considering different levels of biological function and taking one of the organisms as the reference. The platform is user-centric because user specific tasks are independent from one user to others. In addition, the analysis must be anchored to a specific organism, which also makes the platform organism-centric. The proteome of each organism is stored in a central database and tightly coupled with functional annotation components, such as enzyme, substrate, pathways, etc. When a user compares a model organism with another organism(s), the two proteomes can be coupled via sequence comparison.

To deal with such a multilevel complexity in Homol-MetReS requires appropriate software engineering approaches to address interoperability, maintenance, and software composition challenges. At the same time the architecture of the platform and its underlying database should optimize performance and scalability to variable levels of parallelism. Considering these factors, Common Component Architecture (CCA) was adopted as a technology for building the Homol-MetReS applications as a collection of reusable interfaces that were implemented to components and provided plug-and-play environment for high-performance computing that encapsulate the required fundamental algorithms, solvers, and methods.

Common Component Architecture in Homol-MetReS

Typically in a web application, one will have the concept of a model layer, where the data model is described and implemented. The model layer is separate from the View (presentation layer) and the Template layers. In Homol-MetReS a model layer is typically defined as inheriting from proteome analysis functional classes that inherits from three top levels of components of Grok. These are (a) Application, (b) Container and (c) Model.

The application schemas in Homol-MetReS are part of a general Object Relation Mapper (ORM) that is used with the Python objects generated during the utilization of the

server. Consider three sources of Python objects which contain and deal with data: a Model object which is stored in a database, a Form object which is submitted from an HTTP Request and a call to an Adapter object which pulls data from a Container object that receives relational information from the backend database and returns it to the application. The user will then receive the results via HTTP Response. All these types of objects contain data, so it is helpful to be able to use the same system for formally describing the data in any object of the Homol-MetReS.

Architecture for database model in Homol-MetReS

Homol-MetReS employs client-server architecture to communicate between user-specific database and the central server's database. The design of the databases considers both how and what type of data is to be acquired, presented, edited, and entered. At the server-side, information is stored in relational model, MySQL. In addition, the session-specific information updates are performed under the access control of a ZODB (Zope Object Database), an object oriented data-access that is coupled to the relational database by ORM. Homol-MetReS applications work as interfaces by which the object oriented architecture and the relational database can communicate in order to provide flexible access to data and to prevent major changes in the application when a schema change occurs during the evolution of the Homol-MetReS system.

In Homol-MetReS there are three types of functional components (I) Proteome, (II) Annotations and (III) Results. The proteome component (Container component) contains the complete set of proteins of any selected organism. Each selected model organism leads to the creation of an independent database. There are six classified proteome subcomponents, implemented in form of tables in the database model: (i) Molecular entities, (ii) Functions, (iii) Processes, (iv) Localizations, (v) Interactions and (vi) Pathways. Each of the components is then further separated and organized into categories and subcategories. For example, Functions are separated into Enzymes, Receptors, Ligands, Substrates, Transcription Factors and Post Translation Modification categories. All these components and categories are tightly connection in a biological hierarchy (see **Figure 1.1** of **Chapter 1**). **Figure S3.2** shows how all the components are connected within three top level working bins.

Functionally, a protein is polymorphic molecule and different aspects of its functional description are emphasized by different people. Keeping this in mind, the Annotation

component was implemented to automatically coordinate different annotation terms that are attributed to common set of proteins by different users. If a given user annotates a given subcomponent for a set of proteins, Homol-MetReS automatically checks within other functionally categorized proteome subcomponents and integrates both subcomponents.

Results can be of two types in Homol-MetReS: (i) annotation results and (ii) sequence comparison results. Annotation results retrieve information from all the Annotation subcomponents of the list of selected proteins. Sequence comparison results access proteins from other organisms and classify these proteins based on sequence similarity. Results are sent to applications that perform the analysis and visualization tasks.

Clustering implementation in Homol-MetReS

An intermediate list is used to create sequence similarity-based protein clusters during the analysis. This list contains protein sets from corresponding functional modules and it is loaded whenever users need to access it, either for annotation or for sequence comparison.

Integration implementation in Homol-MetReS

One of the major concepts behind component architecture is to divide different types of functionality into different components in order to keep the amount of functionality provided by a single component and integrate them as on demands. Such coupling-decoupling way of component interaction needs to work within a well-defined framework and is achieved using adapters that are designed for identifying the functional context of the protein annotation. **Figure S3.2** summarizes this integration. Two classes of adapters were developed and built into Homol-MetReS: (i) Application functionality adapters and (ii) Web publishing adapters. The Web publishing adapters adapts any of the data model to the visualization model. For example if an application is showing graphics of orthology of Enzymes, in next page the same data model adapts to visualization adapters to show pie chart of the enzymes.

Utilities in Homol-MetReS

The architecture of Homol-MetReS includes “utilities”. Like adapters, such utilities provide specific sets of functionality to the platform. The difference between adapters and utilities is that utilities do not operate on other data components. They simply provide a specific service, such as database connectivity, indexing, searching, mail delivery, browser session etc.

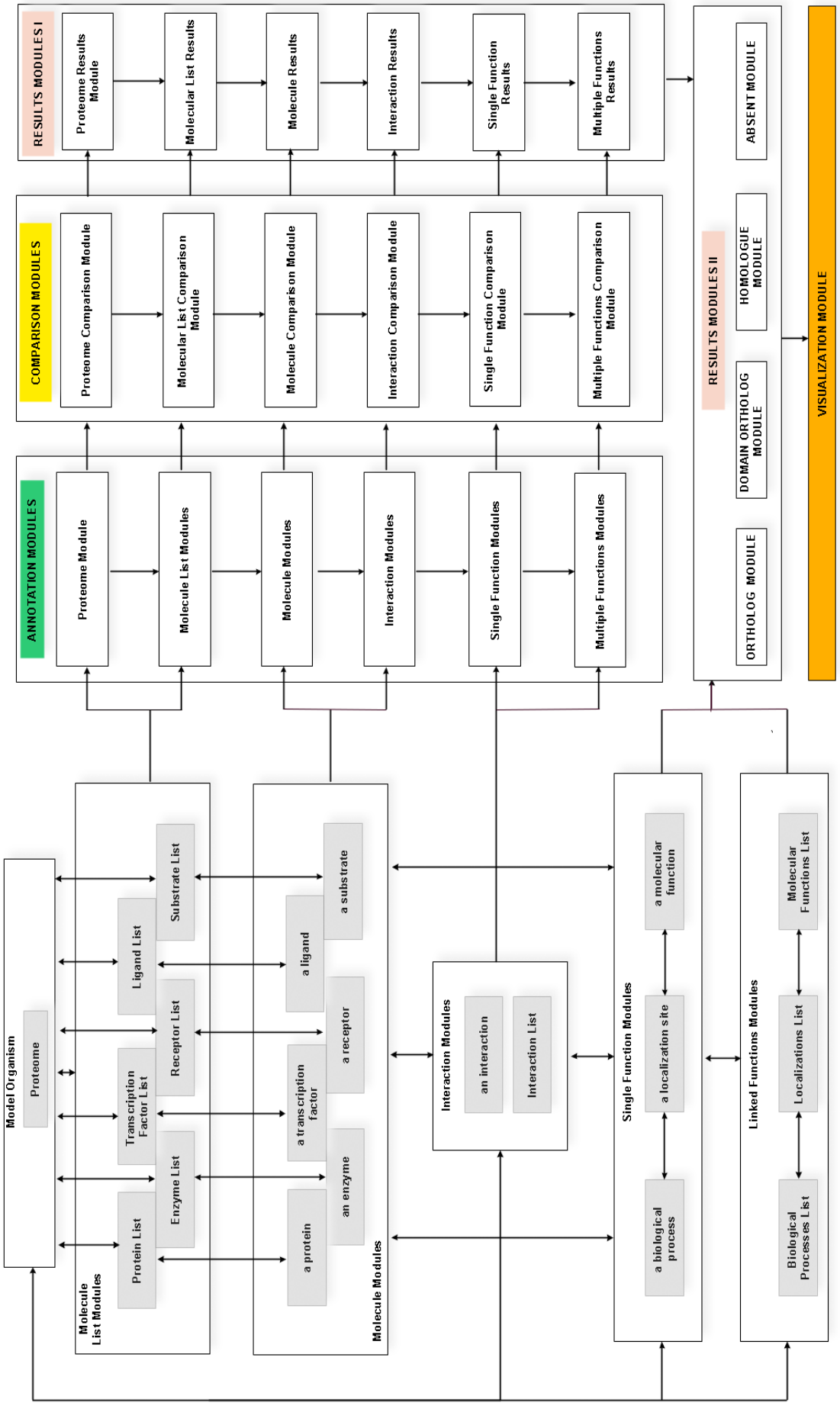


Figure S3.2 Data model integration in Homol-MetReS.

4.6.2. Supplementary Figures

Figure S3.3*

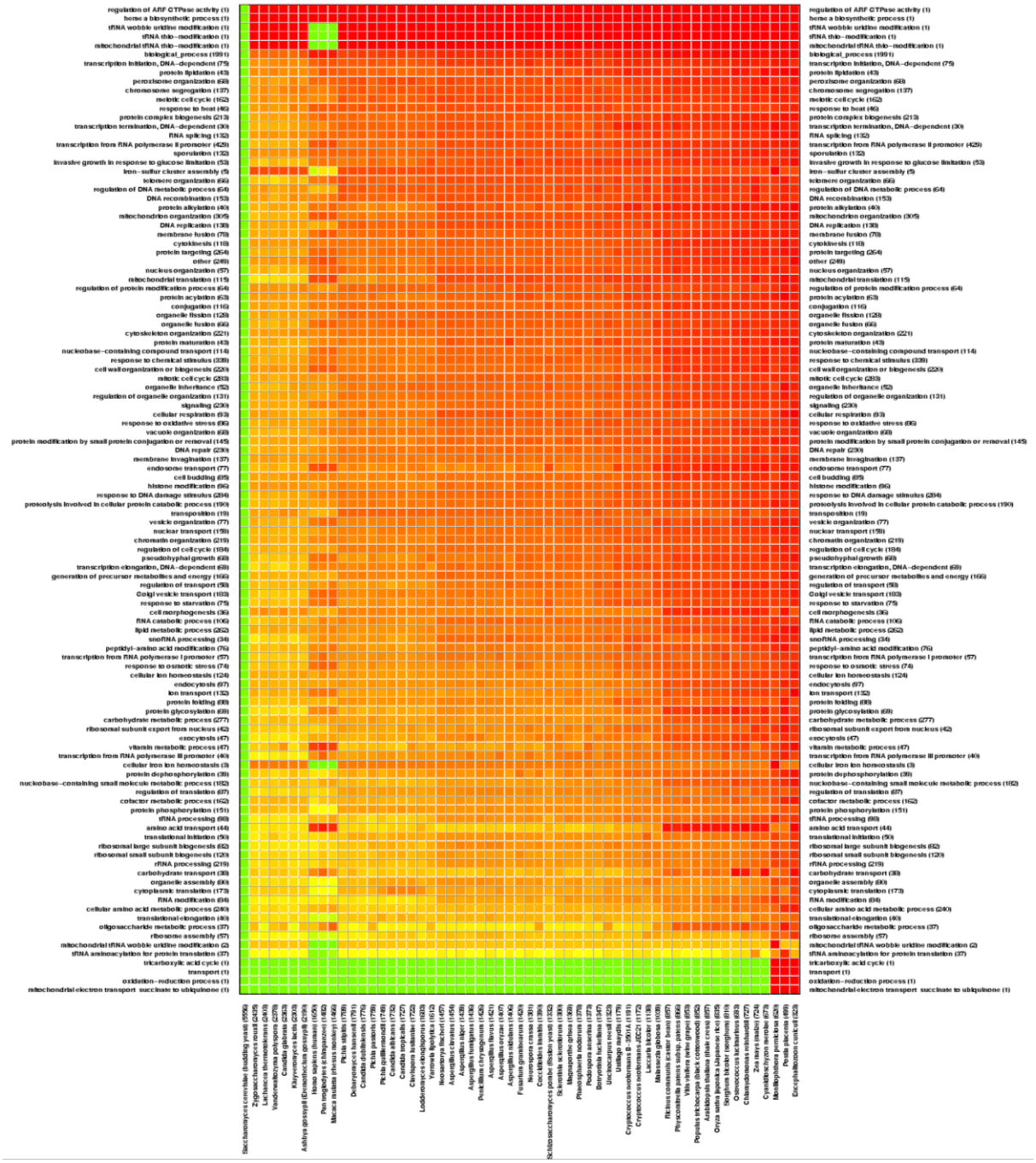


Figure S3.3 Complete heat map for the comparison between the sets of proteins involved in the different biological processes in S. cerevisiae and in 56 other eukaryotes. Each column corresponds to a specific biological process and each row corresponds to one of the eukaryotes under analysis. The rows are sorted by increasing network distance for all GO categories between the organisms represented in the row and *S. cerevisiae*. Distance is calculated as described in methods. Red colour indicates dissimilar sets of proteins with respect to *S. cerevisiae*. Green colour indicates similar sets of proteins with respect to *S. cerevisiae*. ***Enlarged figure is available as Figure S3.3 in the CD that is provided with this thesis.**

Figure S3.4*

Figure S3.4 Complete heat map for the comparison between the sets of proteins involved in the different molecular functions in *S. cerevisiae* and in 56 other eukaryotes. Each column corresponds to a specific molecular function and each row corresponds to one of the eukaryotes under analysis. The rows are sorted by increasing network distance for all GO categories between the organisms represented in the row and *S. cerevisiae*. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to *S. cerevisiae*. Green color indicates similar sets of proteins with respect to *S. cerevisiae*. ***Enlarged figure is available as Figure S3.4 in the CD that is provided with this thesis.**

Figure S3.5*

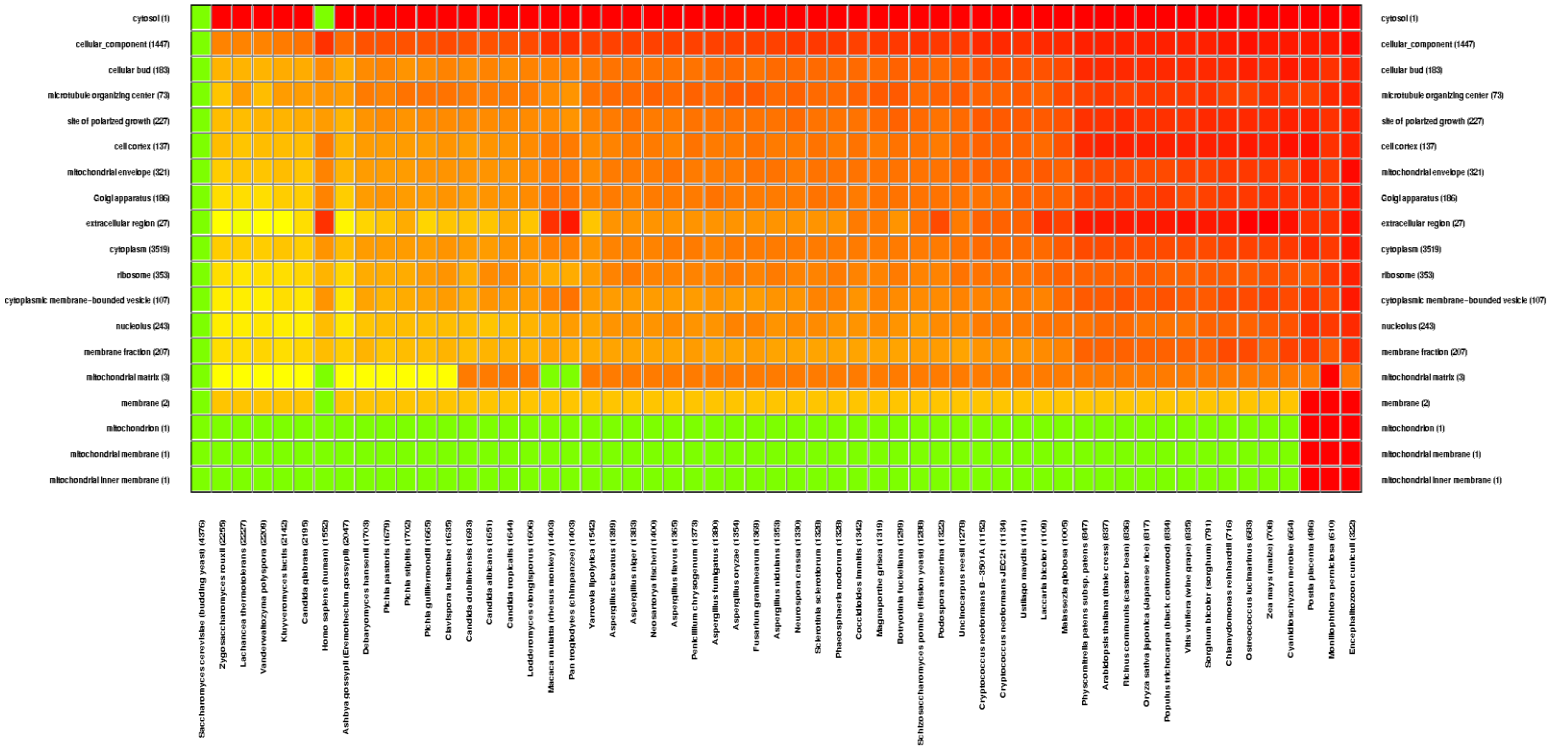


Figure S3.5 Complete heat map for the comparison between the sets of proteins involved in the different localizations in *S. cerevisiae* and in 56 other eukaryotes. Each column corresponds to a specific localization and each row corresponds to one of the eukaryotes under analysis. The rows are sorted by increasing network distance for all GO categories between the organisms represented in the row and *S. cerevisiae*. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to *S. cerevisiae*. Green color indicates similar sets of proteins with respect to *S. cerevisiae*. ***Enlarged figure is available as Figure S3.5 in the CD that is provided with this thesis.**

Figure S3.6*

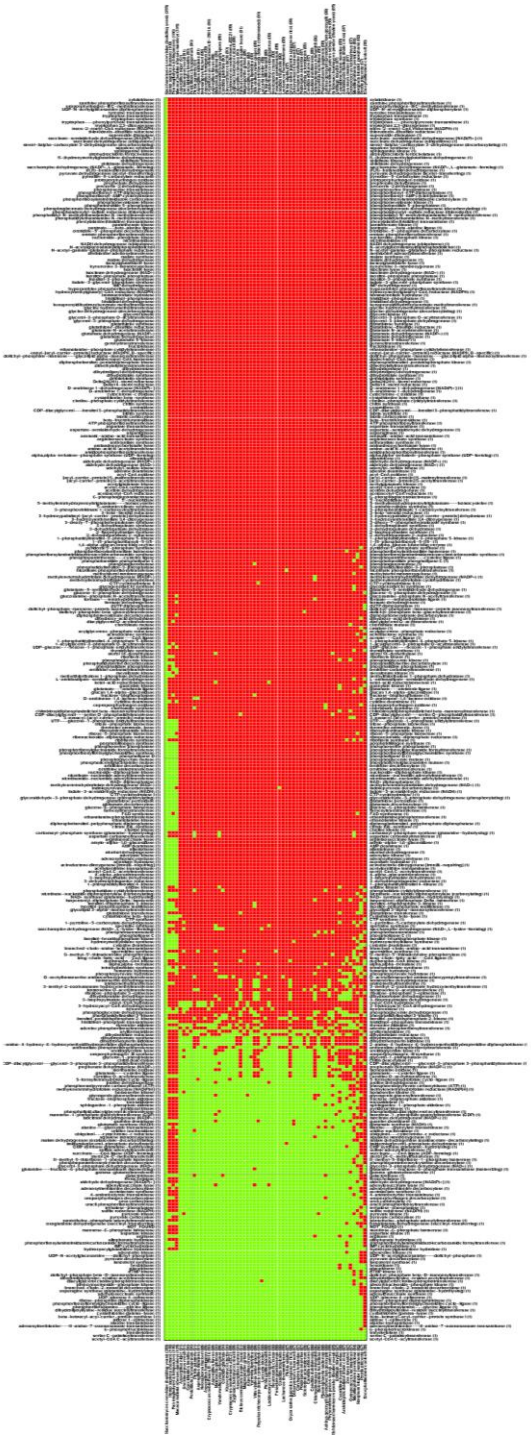


Figure S3.6 Complete heat map for the comparison between the sets of enzymes in *S. cerevisiae* and in 56 other eukaryotes. Each column corresponds to a specific enzyme activity as defined in the EC classification and each row corresponds to one of the eukaryotes under analysis. The rows are sorted by increasing network distance for all the enzyme categories between the organisms represented in the row and *S. cerevisiae*. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to *S. cerevisiae*. Green color indicates similar sets of proteins with respect to *S. cerevisiae*. ***Enlarged figure is available as Figure S3.6 in the CD that is provided with this thesis.**

Figure S3.7*

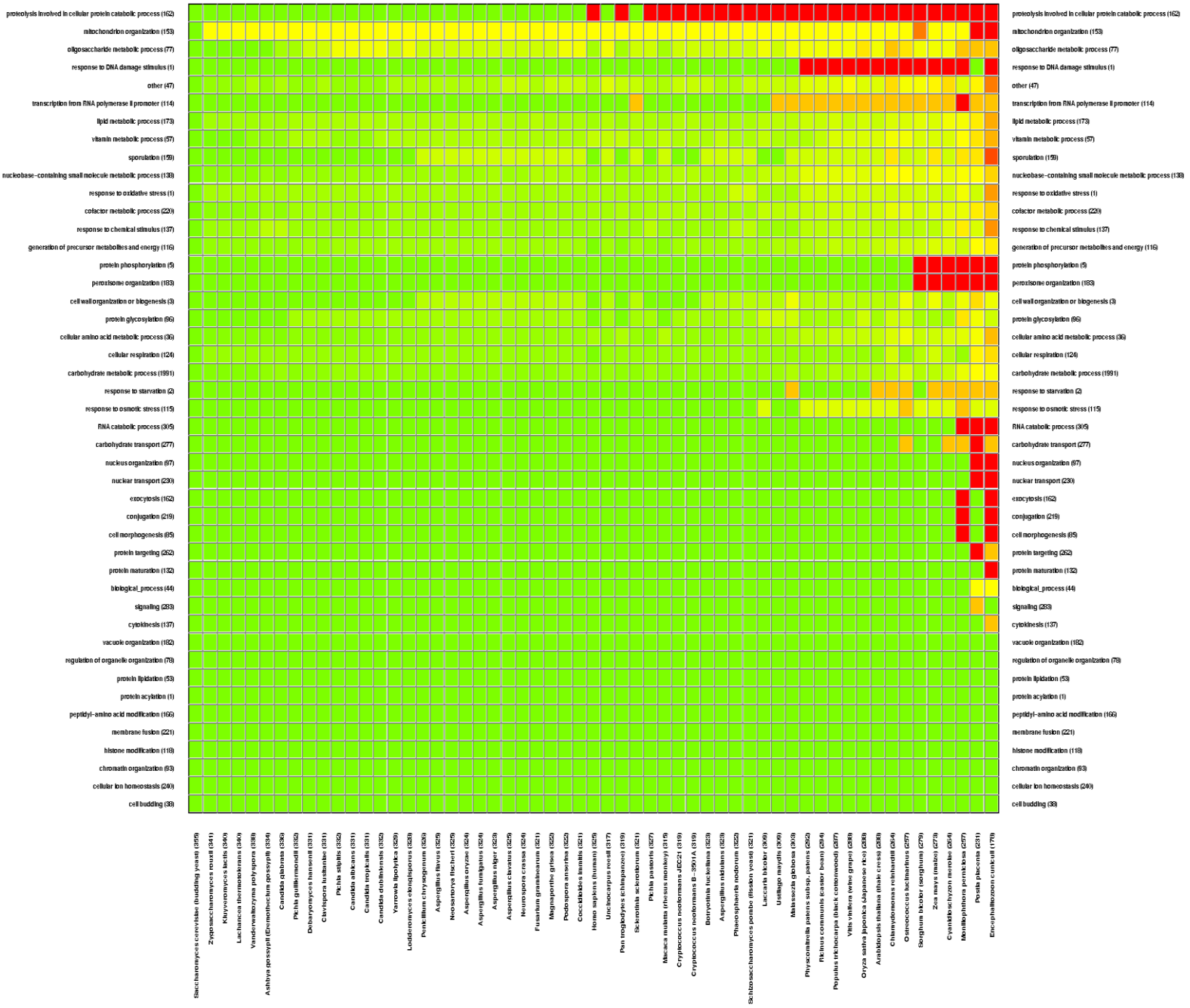


Figure S3.7 Complete heat map for the comparison between the sets of enzymes involved in biological processes in *S. cerevisiae* and in 56 other eukaryotes. Each column corresponds to a specific GO biological process in which the sets of enzymes are involved and each row corresponds to one of the eukaryotes under analysis. The rows are sorted by increasing network distance for all GO categories between the organisms represented in the row and *S. cerevisiae*. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to *S. cerevisiae*. Green color indicates similar sets of proteins with respect to *S. cerevisiae*. ***Enlarged figure is available as Figure S3.7 in the CD that is provided with this thesis.**

Figure S3.8*

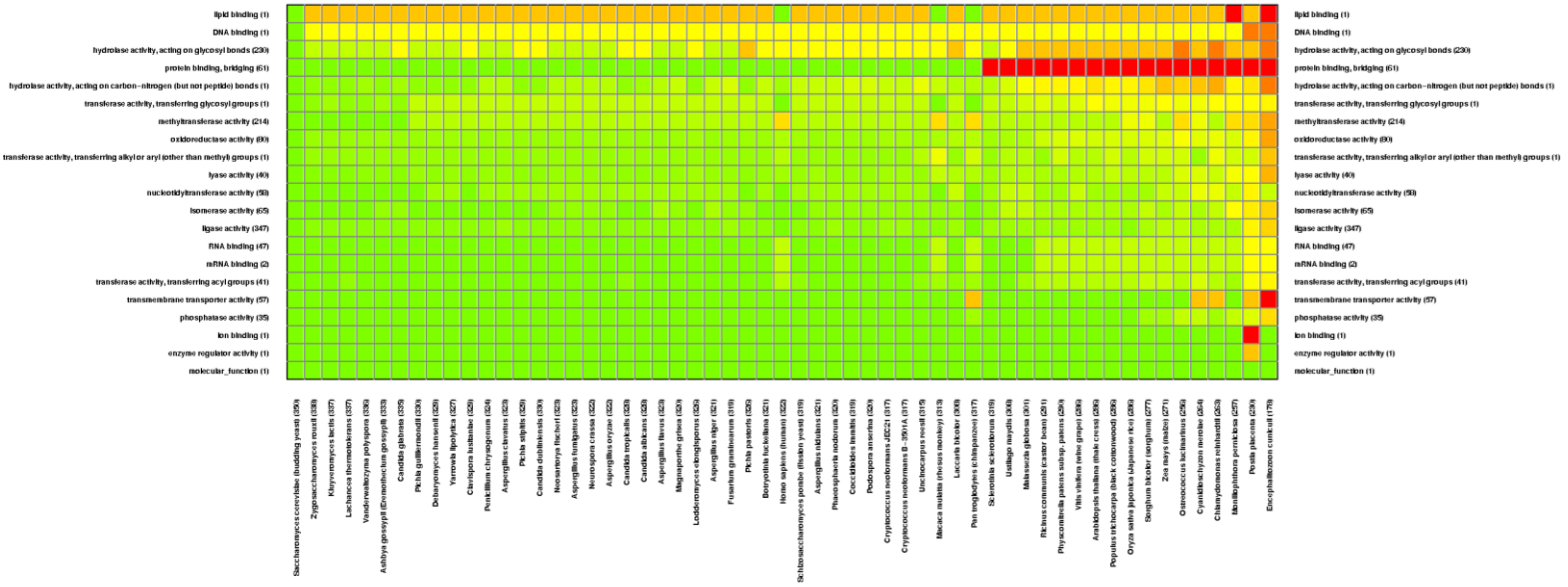


Figure S3.8 Complete heat map for the comparison between the sets of enzymes involved in molecular functions in *S. cerevisiae* and in 56 other eukaryotes. Each column corresponds to a specific GO molecular function term in which the sets of enzymes involved and each row corresponds to one of the eukaryotes under analysis. The rows are sorted by increasing network distance for all GO categories between the organisms represented in the row and *S. cerevisiae*. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to *S. cerevisiae*. Green color indicates similar sets of proteins with respect to *S. cerevisiae*. ***Enlarged figure is available as Figure S3.8 in the CD that is provided with this thesis.**

Figure S3.9*

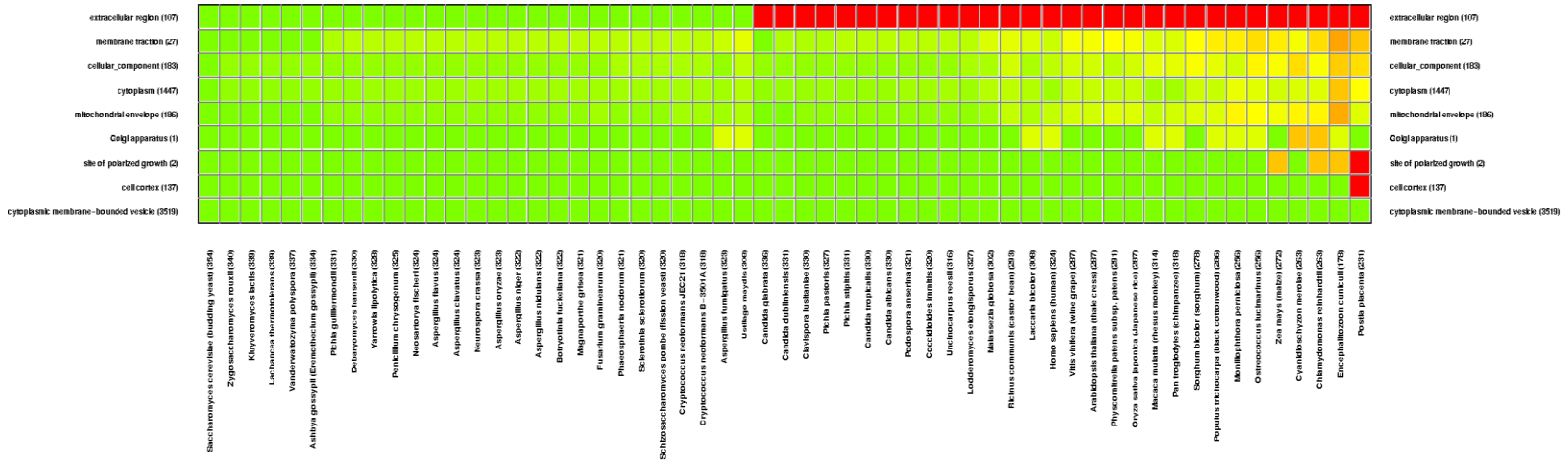


Figure S3.9 Complete heat map for the comparison between the sets of enzymes associated with a cellular localization in *S. cerevisiae* and in 56 other eukaryotes. Each column corresponds to a specific GO localization term and each row corresponds to one of the eukaryotes under analysis. The rows are sorted by increasing network distance for all GO categories between the organisms represented in the row and *S. cerevisiae*. Distance is calculated as described in methods. Red color indicates dissimilar sets of proteins with respect to *S. cerevisiae*. Green color indicates similar sets of proteins with respect to *S. cerevisiae*. ***Enlarged figure is available as Figure S3.9 in the CD that is provided with this thesis.**

4.6.3. Supplementary Tables

Table S3.1 Current Statistics of Homol-MetReS Database

Number of Organisms		**1207**
Kingdom of Organisms	**Bacteria**	**999**
	Archaea	**79**
	Fungi	**43**
	Plants	**11**
	Animals	**46**
Number of Protein Sequences		**5337216**
Number of Annotation Categories		**10**
Number of standard Enzyme terms		**4253**
Number of standard Receptor terms		**558**
Number of standard Ligand terms		**2753**
Number of standard Biological Process terms		**20912**
Number of standard Molecular Function terms		**9813**
Number of standard Cellular Component terms		**2931**
Number of standard Chemical Compounds terms		**14733**

Chapter 5. Final Discussion

5.1. Overview

Systems Biology ultimately aims at understanding how the molecular components of organisms work in an integrated manner, reacting to the environment and keeping the organism alive and healthy. Currently, the question is what we wish to accomplish in modern Systems Biology. Do we want to understand less about more, using a systems biology approach to understand global networks at the expense of mechanistic detail, or do we go on understanding more about less, using reductionist approaches aimed at understanding the mechanistic details of molecular machineries at the expense of comprehensive analysis. Each approach clearly has its strengths and limitations, depending on what biological question needs to be answered. However, the real issue is how we can use both together.

High-throughput approaches, such as whole genome gene expression measurements, proteomics (quantification and identifications of protein & their modifications), and metabolomics (quantification of metabolites) provide only a part of the cellular picture. Comparing the dynamic changes between different experiments and/or environmental conditions allows the generation of molecular or genetic networks of interdependence. This information can provide usefull insights into the dynamics of the genetic and proteomic programs of the cell. However, the information can provide few mechanistic details of how that dynamics is regulated. Such details can only be obtained through reductionist approaches.

Thus, to fully understand the workings of a biological system in detail, both approaches are needed, because they provide complementary data. The key issue is how to provide a flexible solution that enable biologists to combine them, taking into account the development of new experimental technologies and the large amounts of data that are already publicly available to reconstruct the proteome of both, new and well known organisms.

This thesis contribute to the development of such solutions by developing a methodological pipeline that integrates high-throughput and mechanistic datasets and uses the integration for detailed comparisons between the proteomes of different organisms. The Homol-MetReS platform implements the pipeline and makes it available to other researchers.

In the remainder of this discussion we will focus on providing an integrated discussion of the work presented in the previous chapters, highlighting how the work presented here can contribute to the progress of systems biology. We will conclude by

proposing several lines of research that could be taken to further develop the research presented here.

5.2. General discussion and future perspectives

Because proteome reconstruction is a central issue of the thesis, integration of the various types of functional information about proteins is necessary. An appropriate integration enables easy updates of the tool as new information becomes available. Homol-MetReS transfers information between well annotated and new proteomes by using sequence similarity between proteins using BLAST results and combining the e-value, (%) identity or similarity scores, and gaps in the alignment to build a composite score. This score helps identifying the most likely functional ortholog of any given protein in a new organism. The accumulation of fully sequenced and annotated genomes facilitates the use of such sequence comparison to reconstruct maps of metabolic, signal transduction, and gene circuits in new genomes.

For example in Chapter 2, *Saccharomyces cerevisiae* was compared as model organism with 704 other organisms from almost all clades of life. Amongst these, *Yarrowia lipolytica* (Ascomycetes) is another yeast, with a poorly annotated genome. This complete proteome was compared to that of *S. cerevisiae.* That comparison led to the reconstruction of 102 different metabolic pathways in *Y. lipolitica*. The results (see **Figure S1.2**) reveal that 3459 genes of *S. cerevisiae* were found to be mapped with orthologs in the *Y. lipolitica* proteome. Such type of mapping of pathways/process provide further insights about comparison of physiologies of two organisms, helping in identifying the function of proteins that were previously not characterized at the functional level.

When no homology is found between a gene in newly or uncharacterized sequenced genome and previously characterized genes with know function, sequence based annotation is not possible. However, if structural information is available, structural homology comparisons may facilitate attributing general or specific functions to individual genes, for example using classification such as SCOP or CATH. Integrating such structural information in Homo-MetReS is one of the possible ways in which its functionality could be extended in the future.

When no structural or sequence homology exists between a gene/protein and other of known function, similar patterns of co-occurrence of this protein with others of known

function in large numbers of organisms could also provide some functional information. The logic behind phylogenetically conserved group prediction of function is as follows. If a set of homologous proteins with unknown function is present (absent) in the same genomes when compared to proteins of known function, then it is possible that evolution acted simultaneously on both sets of proteins because somehow they share a function. This could allow the researcher to predict that some genes are involved in the same processes, although their individual function(s) may remain uncertain. Nevertheless, it should be emphasized that functionally related proteins do not necessarily coevolve, and functional modules need not behave as evolutionary modules.

Homol-MetReS also provides an alternative to building phylogenetic trees. By considering the sets of proteins that have orthologs, homologs or are absent between organisms one can cluster that organisms in the following way. First, construct a meta-proteome that contains ***O***-clusters for all proteins of interest to the researcher from all organisms of interest. Then, build a matrix where each row represents one cluster and each collumn represents an organism. Finally, cluster the organisms with respect to their row similarity. Examples of this are discussed in the results of **Chapter 2** and **Chapter 3**. Such trees are more likely to provide information about how close the organisms are with respect to the specific processes in which the proteins of interest are involved than about the global information history of the organism set.

An example is shown in **Figure 5.1** for 18 animals. The proteins chosen to build these trees are involved in the development of brain (**Figure 5.1. A**), bone (**Figure 5.1. B**), and muscle (**Figure 5.1. C**). One can see that such an analysis clearly separates vertebrates from non-vertebrates with respect to brain development. Furthermore, and because this tree is similar to the corresponding phylogenetic tree found in NCBI, one can assume that the pattern of conservation for proteins involved in brain development is reasonably close to of the global evolution for this set of organisms. The same can not be said about the trees built using proteins involved in muscle and bone development. For example, with respect to bone development, dog forms an ingroup with human and gorilla, with monkey and chimpanzee as outgroup. This suggests that the protein networks involved in human bone and muscle development are more similar to that of dogs and gorilla and somewhat less similar to that of chimp and monkey.

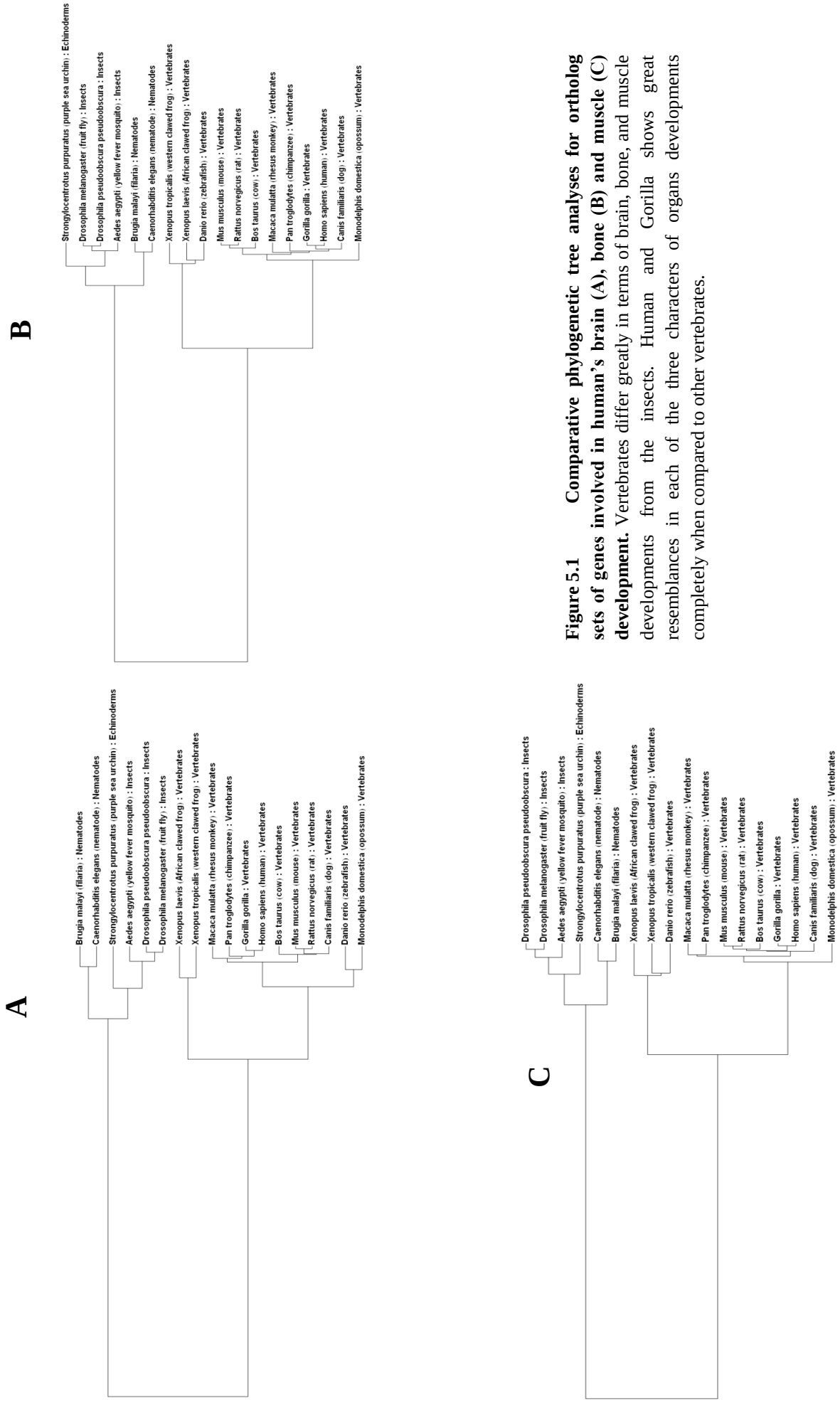


Figure 5.1 Comparative phylogenetic tree analyses for ortholog sets of genes involved in human's brain (A), bone (B) and muscle (C) development. Vertebrates differ greatly in terms of brain, bone, and muscle developments from the insects. Human and Gorilla shows great resemblances in each of the three characters of organs developments completely when compared to other vertebrates.

Homol-MetReS also allows users to perform a comparative analysis of protein duplication/deletion between organisms. This is an important aspect of genome analysis, as many genomes underwent several round of whole genome duplication during their evolution. What duplicate proteins were retained through evolution and which were lost can provide important insights into how the environment modulates such evolution. For example, we observe that the percentage of human receptors that have been duplicated and remain active throughout evolution is larger than that of other types of proteins. Given that such receptors sense how the enviroment changes, such duplications could be associated to adaptation to a wider range of environmental conditions.

Homol-MetReS can be used to compare how the network of proteins responsible for a given process, pathway, or circuit have evolved in groups of different organisms. This can be done by comparing the network of protein from a reference organism to that of other organisms and identifying which proteins have been retained and lost in the different organisms. This also allows users to identify protein fusion and domain shuffling events.

Figure 5.2 shows a schematic example of such a comparison for a hypothetical pathway in three organisms.

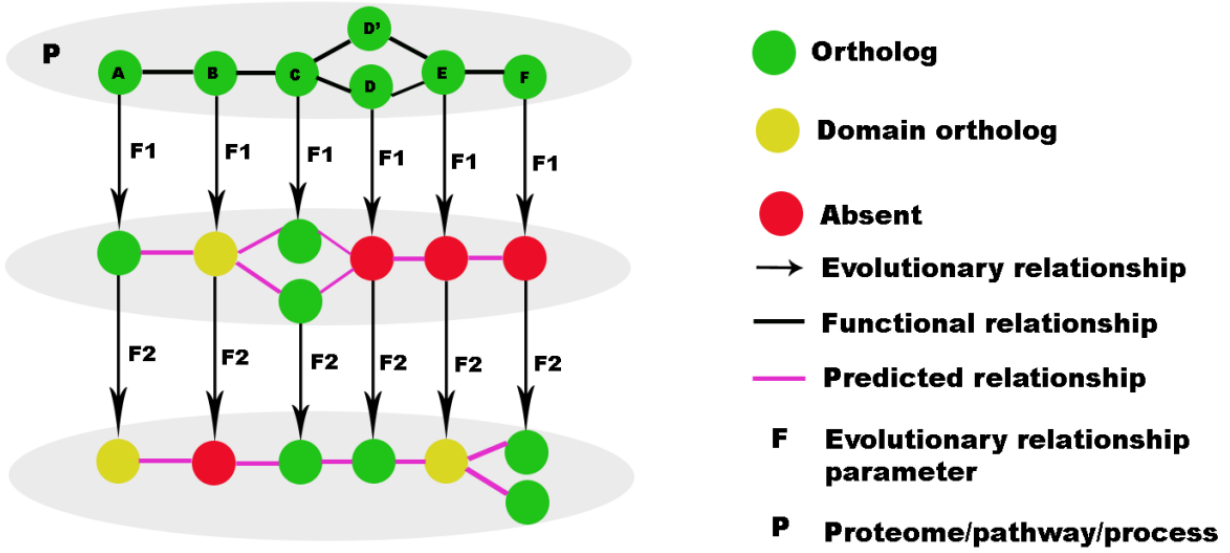


Figure 5.2 **Molecular network/pathways alignment at evolution levels. P1**, **P2**, and **P3** represent three alternative organisms. The pathway has six steps (**A-F**), with step four being executed by two alternative proteins. The events represented here can be interpreted in the following way. The first step is conserved in **P2** and partially conserved in **P3**. The second step is partially conserved in **P2** and absent in **P3**. The third step was duplicated in P2 and is conserved in **P3**. The last three steps of the pathway are absent in **P2** and partially conserved and/or duplicated in **P3**.

With all its functionality, Homol-MetReS is adequate for identifying appropriate model organisms to study various biological phenomena. This application was further developed and illustrated in **Chapters 2** and **3**. In **Chapter 3** we further illustrate how this analysis can identify those proteins that are unique to each organism by identifying the unique proteins of human.

Homol-MetReS provides functional annotation that integrates information of standard terms from GO, BRENDA, KEGG, NCBI, and HPRD. However, one aspect where Homol-MetReS is lacking is that of transcription factor (TF) classification. To our knowledge, there is no global classification that considers all possible types of TF. Once such a classification is made available, it could be promtly integrated into the central database that underlies the application and tightly coupled with the other functional classifications already being considered. This would improve the functionality of the application with respect to the reconstruction of gene circuits. Nevertheless, the functionality that Homol-MetReS provides allows for users to define their own functional classification of TF (or other types of proteins) and rely on them for (manual or automated) protein functional annotation and circuit reconstruction. This functionality also permits users to modify preexisting functional classifications to better suit their research.

Another limitation of Homol-MetReS is that it does not provide quantitative information regarding physical-chemical properties of proteins, metabolites, or mRNAs, levels of protein abundance, or correlation between changes in gene expression and protein activity. This is important information from the systemic point of view. However, reliable data about these aspects at the genome scale are still limited. Subsequent iterations of Homol-MetReS will consider including such data if and when it becomes widely available. In fact, users can include some of that information in the manual (re)annotation that they can perform in the application.

Two alternative ways in which Homol-MetReS can be improved were already suggested. On one hand, it could be improved with a general transcription factor classification. On the other it could include quantitative information about different aspects of the molecular components of cells. Another aspect in which Homol-MetReS could be imporved is by including a text mining tool [268] to permit automated extraction of functional and quantitative information about proteins and circuits from the scientific literature.

A final suggestion that would significantly improve Homol-MetReS is to include functionality that permits semi-automated creation of mathematical models to study the dynamic behavior of reconstructed protein circuits. Such models provide a systematic way for integrating genetic and biochemical information. Including this functionality is a big challenge, as it would require also including information about parameter values and regulation of biological processes that is scarce. One way around this limitation is by allowing users to manually modify models and include that information. Simulation and analysis of those models provides a deeper understanding of the organization and complexity of biological systems with respect to different aspects of biology. For example, they can be used to study principles of organization and operation in the adaptive response of organisms and identify the constraints that shape those principles. When Homol-MetReS includes such functionality, it should come coupled with the capacity to export models in SBML or CelML formats that would allow users to take those models and analyze them in other software applications. It might also be useful to develop a Molecular Systems Markup Language (MSML) to more easily integrate functional annotation of proteins/genes in the context of model building.

5.3. Possible pitfalls and how to avoid them

There are several pitfalls that could hinders Homol-MetReS' future development. First, the amount of new data to be included in the central database could become unmanageable. To avoid that, data should be stored in an organized structure in order to allow efficient data mining. In addition, the quality of genome wide experimental data and functional annotation should be improved if Homol-MetReS-like tools are to become even more useful. Second, as new types of large scale data becomes available, the pipeline on which Homol-MetReS relies should be modified and updated in such a way to maintain its cohesiveness and permit an accurate and flexible semi-automated construction of mathematical models. Third, to maintain and develop its functionality in such a way, that it is useful to both, researchers interested in proteome wide analysis and researchers interested in the analysis of smaller subsets of that proteome. The conciliation of both currents is not trivial, remaining one of the most challenging aspects in the development of methods for analysis of genetic, biochemical, signal transduction processes from systemic perspective.

5.4. Final Remarks

The work presented in this thesis contributes to the functional understanding of molecular circuits and their organization. It seems to us that a time should come when a complete hierarchy of integration molecular networks will be identified. If this is so, we believe that the work presented in this thesis constitutes a very encouraging head start towards the goal of such classification. A skeleton for such a hierarchy is shown as an example in **Figure 5.3**.

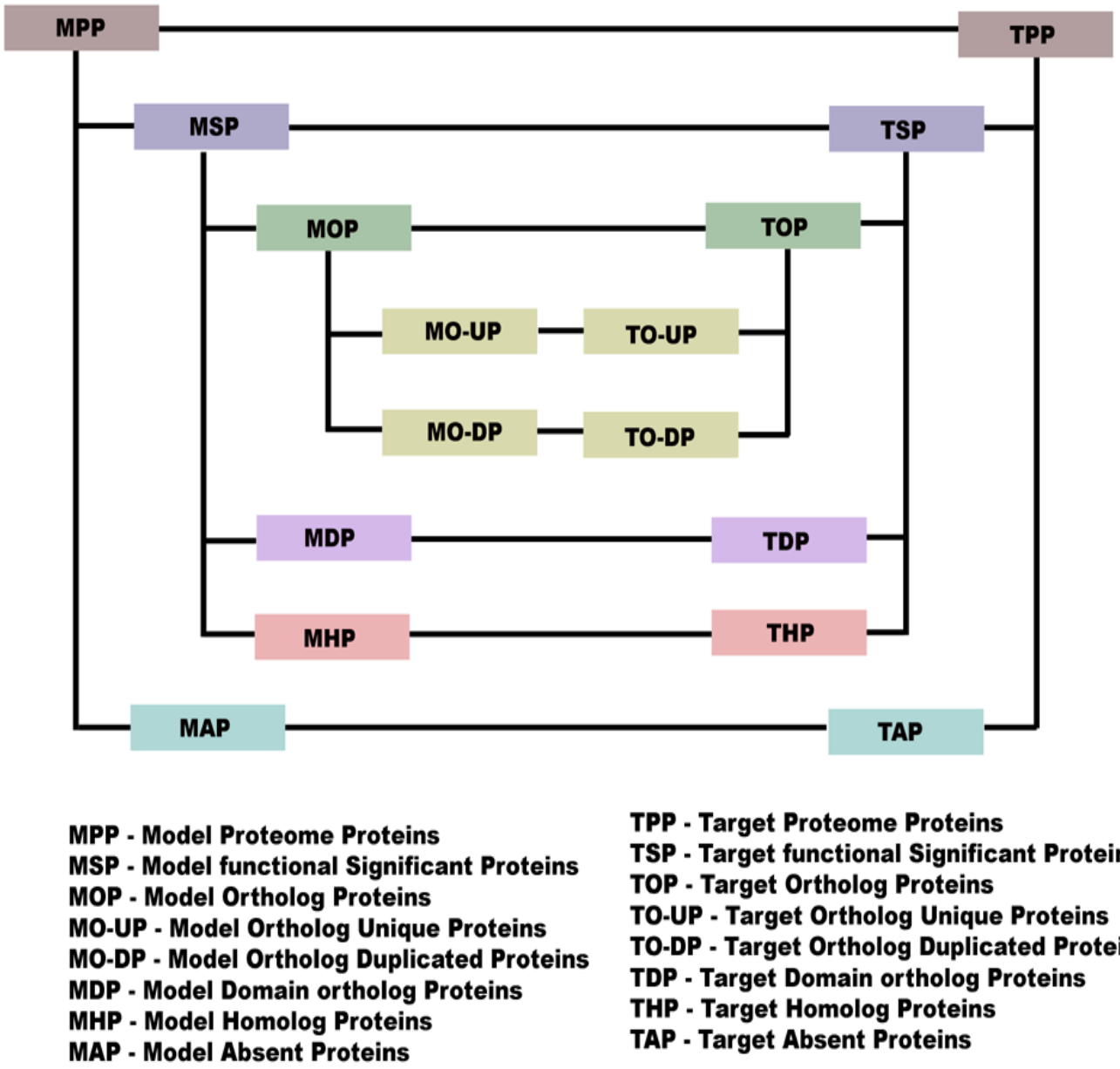


Figure 5.3 **Integration of relational scheme used to link protein information of cells and evolution.**

Chapter 6. Conclusions

1. We developed a methodology for functional comparison of full proteomes among organisms with fully sequenced genomes.

2. Our methodology classifies any two complete proteomes into distinct clusters of Significant, Orthologs, Domain orthologs, Homologs and Absent proteins.

3. We propose a score function that appears to be appropriate to distinguish between orthologs and paralogs of protein function.

4. Our method proposes ways to identify appropriate model organisms to study the dynamics of different biological processes and pathways in specific organisms, as long as the proteins that participate in the processes are known.

5. The method we propose here could be especially relevant to assist in the choice of appropriate model organisms for both, the study of human specific biological processes and the characterization of a specific biological phenomenon in a large class of organisms.

6. We provide Homol-MetReS to the community, an application where the method is implemented.

7. Homol-MetReS can accurately identify duplications or deletion of protein coding genes and reconstruct a form of functional phylogeny over the set of proteins involved in specific processes in selected organisms.

8. Applying our methods shows that *Saccharomyces cerevisiae* is good general model to study: DNA replication, metabolic pathways, purine metabolism, and amino acyl t-RNA synthesis.

9. Human is one of the top six non-fungal organisms whose proteome is the most similar to that of *S. cerevisiae*.

10. Through our methodology we identified the proteins that are unique to humans, and establish ranks of similarity between the sets of proteins involved in different processes, functions, localizations, biochemical circuits and tissue components in humans and in the other eukaryotes.

11. We pinpointed the most likely eukaryotes to be good model organisms in which to study specific biological processes and phenomena that have biological and biomedical relevance in human.

12. We provide a first complete functional characterization of the gorilla proteome.

13. We find that the proteome of gorilla is functionally more similar to that of human than the chimp proteome, at the ***FO*** level. A more detailed analysis reveals that the ***O-O*** clusters of ***FO*** between human and chimp are more numerous than between human and gorilla. In contrast, the gorilla proteome forms the highest number of ***O-M*** and ***M-M*** clusters with the human proteome.

Bibliography

1. Ideker, T., T. Galitski, and L. Hood, *A new approach to decoding life: systems biology.* Annu Rev Genomics Hum Genet, 2001. **2**: p. 343-72.
2. Ishii, N., et al., *Multiple high-throughput analyses monitor the response of E. coli to perturbations.* Science, 2007. **316**(5824): p. 593-7.
3. Villoslada, P., L. Steinman, and S.E. Baranzini, *Systems biology and its application to the understanding of neurological diseases.* Ann Neurol, 2009. **65**(2): p. 124-39.
4. Albulescu, L.O., et al., *A quantitative, high-throughput reverse genetic screen reveals novel connections between Pre-mRNA splicing and 5' and 3' end transcript determinants.* PLoS Genet, 2012. **8**(3): p. e1002530.
5. Brush, G.S., et al., *Yeast IME2 functions early in meiosis upstream of cell cycle-regulated SBF and MBF targets.* PLoS One, 2012. **7**(2): p. e31575.
6. Gifford, M.L., et al., *Cell-specific nitrogen responses mediate developmental plasticity.* Proc Natl Acad Sci U S A, 2008. **105**(2): p. 803-8.
7. Calvano, S.E., et al., *A network-based analysis of systemic inflammation in humans.* Nature, 2005. **437**(7061): p. 1032-7.
8. Chen, S., et al., *Structure of N-terminal domain of ZAP indicates how a zinc-finger protein recognizes complex RNA.* Nature structural & molecular biology, 2012. **19**(4): p. 430-5.
9. Bender, R.A., *A NAC for regulating metabolism: the nitrogen assimilation control protein (NAC) from Klebsiella pneumoniae.* J Bacteriol, 2010. **192**(19): p. 4801-11.
10. Del Vecchio, D., A.J. Ninfa, and E.D. Sontag, *Modular cell biology: retroactivity and insulation.* Mol Syst Biol, 2008. **4**: p. 161.
11. Reitzer, L., *Nitrogen assimilation and global regulation in Escherichia coli.* Annu Rev Microbiol, 2003. **57**: p. 155-76.
12. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761 Suppl): p. C47-52.
13. Noble, D., *Claude Bernard, the first systems biologist, and the future of physiology.* Exp Physiol, 2008. **93**(1): p. 16-26.
14. Al-Nuaimi, Y., et al., *A prototypic mathematical model of the human hair cycle.* Journal of theoretical biology, 2012.
15. Hughes, M.W., et al., *In search of the Golden Fleece: unraveling principles of morphogenesis by studying the integrative biology of skin appendages.* Integr Biol (Camb), 2011. **3**(4): p. 388-407.

16. Savageau, M.A., R. Metter, and W.W. Brockman, *Statistical significance of partial base-pairing potential: implications for recombination of SV40 DNA in eukaryotic cells.* Nucleic Acids Res, 1983. **11**(18): p. 6559-70.
17. Savageau, M.A., *Regulation of differentiated cell-specific functions.* Proc Natl Acad Sci U S A, 1983. **80**(5): p. 1411-5.
18. Salvado, B., et al., *Two component systems: physiological effect of a third component.* PLoS One, 2012. **7**(2): p. e31095.
19. Salvado, B., et al., *Methods for and results from the study of design principles in molecular systems.* Math Biosci, 2011. **231**(1): p. 3-18.
20. Ideker, T., et al., *Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.* Science, 2001. **292**(5518): p. 929-34.
21. Li, M., et al., *Chemotaxis kinase CheA is activated by three neighbouring chemoreceptor dimers as effectively as by receptor clusters.* Mol Microbiol, 2011. **79**(3): p. 677-85.
22. Amin, D.N. and G.L. Hazelbauer, *Chemoreceptors in signalling complexes: shifted conformation and asymmetric coupling.* Mol Microbiol, 2010. **78**(5): p. 1313-23.
23. Hazelbauer, G.L. and W.C. Lai, *Bacterial chemoreceptors: providing enhanced features to two-component signaling.* Curr Opin Microbiol, 2010. **13**(2): p. 124-32.
24. Amin, D.N. and G.L. Hazelbauer, *The chemoreceptor dimer is the unit of conformational coupling and transmembrane signaling.* J Bacteriol, 2010. **192**(5): p. 1193-200.
25. Atkinson, M.R. and A.J. Ninfa, *Role of the GlnK signal transduction protein in the regulation of nitrogen assimilation in Escherichia coli.* Mol Microbiol, 1998. **29**(2): p. 431-47.
26. Liu, E.T., *Systems biology, integrative biology, predictive biology.* Cell, 2005. **121**(4): p. 505-6.
27. Ng, P., et al., *Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation.* Nat Methods, 2005. **2**(2): p. 105-11.
28. Wei, C.L., et al., *Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state.* Stem Cells, 2005. **23**(2): p. 166-85.
29. Constante, M., R. Grunberg, and M. Isalan, *A biobrick library for cloning custom eukaryotic plasmids.* PLoS One, 2011. **6**(8): p. e23685.

30. Quackenbush, J., *Extracting biology from high-dimensional biological data.* J Exp Biol, 2007. **210**(Pt 9): p. 1507-17.
31. Junker, B.H., C. Klukas, and F. Schreiber, *VANTED: a system for advanced data analysis and visualization in the context of biological networks.* BMC Bioinformatics, 2006. **7**: p. 109.
32. Shah, S.P., et al., *Atlas - a data warehouse for integrative bioinformatics.* BMC Bioinformatics, 2005. **6**: p. 34.
33. Ruebenacker, O., et al., *Integrating BioPAX pathway knowledge with SBML models.* IET Syst Biol, 2009. **3**(5): p. 317-28.
34. Shen, S.Y., F. Bergmann, and H.M. Sauro, *SBML2TikZ: supporting the SBML render extension in LaTeX.* Bioinformatics, 2010. **26**(21): p. 2794-5.
35. Matsuoka, Y., et al., *Payao: a community platform for SBML pathway model curation.* Bioinformatics, 2010. **26**(10): p. 1381-3.
36. Swainston, N. and P. Mendes, *libAnnotationSBML: a library for exploiting SBML annotations.* Bioinformatics, 2009. **25**(17): p. 2292-3.
37. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
38. Celniker, S.E., et al., *Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence.* Genome biology, 2002. **3**(12): p. RESEARCH0079.
39. Leone, G., L. Fianchi, and M.T. Voso, *Therapy-related myeloid neoplasms.* Curr Opin Oncol, 2011. **23**(6): p. 672-80.
40. Bench, A.J., *The role of molecular genetic analysis within the diagnostic haemato-oncology laboratory.* Int J Lab Hematol, 2012. **34**(1): p. 21-34.
41. Cronin, M. and J.S. Ross, *Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology.* Biomark Med, 2011. **5**(3): p. 293-305.
42. Oricchio, E., et al., *Mouse models of cancer as biological filters for complex genomic data.* Dis Model Mech, 2010. **3**(11-12): p. 701-4.
43. Maciejewski, J.P., R.V. Tiu, and C. O'Keefe, *Application of array-based whole genome scanning technologies as a cytogenetic tool in haematological malignancies.* Br J Haematol, 2009. **146**(5): p. 479-88.
44. de Leval, L. and P. Gaulard, *Pathobiology and molecular profiling of peripheral T-cell lymphomas.* Hematology Am Soc Hematol Educ Program, 2008: p. 272-9.

45. Wilkins, M.R., et al., *From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis.* Biotechnology (N Y), 1996. **14**(1): p. 61-5.
46. Saqi, M.A. and D.L. Wild, *Expectations from structural genomics revisited: an analysis of structural genomics targets.* Am J Pharmacogenomics, 2005. **5**(5): p. 339-42.
47. Gevaert, K. and J. Vandekerckhove, *Protein identification methods in proteomics.* Electrophoresis, 2000. **21**(6): p. 1145-54.
48. Steen, H. and A. Pandey, *Proteomics goes quantitative: measuring protein abundance.* Trends Biotechnol, 2002. **20**(9): p. 361-4.
49. Jabbour, R.E., et al., *A protein processing filter method for bacterial identification by mass spectrometry-based proteomics.* Journal of proteome research, 2011. **10**(2): p. 907-12.
50. Basle, E., N. Joubert, and M. Pucheault, *Protein chemical modification on endogenous amino acids.* Chemistry & biology, 2010. **17**(3): p. 213-27.
51. Png, E., et al., *A new method of high-speed cellular protein separation and insight into subcellular compartmentalization of proteins.* Anal Bioanal Chem, 2011. **400**(3): p. 767-75.
52. Claydon, A.J. and R.J. Beynon, *Protein turnover methods in single-celled organisms: dynamic SILAC.* Methods Mol Biol, 2011. **759**: p. 179-95.
53. Altmann, K., M. Durr, and B. Westermann, *Saccharomyces cerevisiae as a model organism to study mitochondrial biology: general considerations and basic procedures.* Methods Mol Biol, 2007. **372**: p. 81-90.
54. Beck, H., D. Dobritzsch, and J. Piskur, *Saccharomyces kluyveri as a model organism to study pyrimidine degradation.* FEMS Yeast Res, 2008. **8**(8): p. 1209-13.
55. Cogburn, L.A., et al., *Functional genomics of the chicken--a model organism.* Poult Sci, 2007. **86**(10): p. 2059-94.
56. Jin, T., et al., *How human leukocytes track down and destroy pathogens: lessons learned from the model organism Dictyostelium discoideum.* Immunol Res, 2009. **43**(1-3): p. 118-27.
57. Jonsson, K.I., *Tardigrades as a potential model organism in space research.* Astrobiology, 2007. **7**(5): p. 757-66.

58. Meyer, M. and J. Vilardell, *The quest for a message: budding yeast, a model organism to study the control of pre-mRNA splicing.* Brief Funct Genomic Proteomic, 2009. **8**(1): p. 60-7.
59. Veldman, M.B. and S. Lin, *Zebrafish as a developmental model organism for pediatric research.* Pediatr Res, 2008. **64**(5): p. 470-6.
60. Siva, N., *1000 Genomes project.* Nat Biotechnol, 2008. **26**(3): p. 256.
61. Ellegren, H., *Comparative genomics and the study of evolution by natural selection.* Mol Ecol, 2008. **17**(21): p. 4586-96.
62. Tettelin, H., et al., *Comparative genomics: the bacterial pan-genome.* Curr Opin Microbiol, 2008. **11**(5): p. 472-7.
63. Alves, R., E. Vilaprinyo, and A. Sorribas, *Integrating bioinformatics and computational biology: Perspectives and possibilities for in silico network reconstruction in molecular systems biology.* Current Bioinformatics, 2008. **3**(2): p. 98-129.
64. Murakami, C. and M. Kaeberlein, *Quantifying yeast chronological life span by outgrowth of aged cells.* J Vis Exp, 2009(27).
65. Biddick, R. and E.T. Young, *The disorderly study of ordered recruitment.* Yeast, 2009. **26**(4): p. 205-20.
66. Hohmann, S., M. Krantz, and B. Nordlander, *Yeast osmoregulation.* Methods Enzymol, 2007. **428**: p. 29-45.
67. Nasheuer, H.P., et al., *Initiation of eukaryotic DNA replication: regulation and mechanisms.* Prog Nucleic Acid Res Mol Biol, 2002. **72**: p. 41-94.
68. Brocard-Masson, C. and B. Dumas, *The fascinating world of steroids: S. cerevisiae as a model organism for the study of hydrocortisone biosynthesis.* Biotechnol Genet Eng Rev, 2006. **22**: p. 213-52.
69. Lopez-Mirabal, H.R. and J.R. Winther, *Redox characteristics of the eukaryotic cytosol.* Biochim Biophys Acta, 2008. **1783**(4): p. 629-40.
70. Owsianowski, E., D. Walter, and B. Fahrenkrog, *Negative regulation of apoptosis in yeast.* Biochim Biophys Acta, 2008. **1783**(7): p. 1303-10.
71. Miller-Fleming, L., F. Giorgini, and T.F. Outeiro, *Yeast as a model for studying human neurodegenerative disorders.* Biotechnol J, 2008. **3**(3): p. 325-38.
72. Foury, F., *Human genetic diseases: a cross-talk between man and yeast.* Gene, 1997. **195**(1): p. 1-10.

73. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol, 2007. **25**(11): p. 1251-5.
74. Caspi, R.R., *Tregitopes switch on Tregs.* Blood, 2008. **112**(8): p. 3003-4.
75. Dwight, S.S., et al., *Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO).* Nucleic Acids Res, 2002. **30**(1): p. 69-72.
76. Jensen, L.J., et al., *Prediction of human protein function according to Gene Ontology categories.* Bioinformatics, 2003. **19**(5): p. 635-42.
77. Buza, T.J., et al., *Gene Ontology annotation quality analysis in model eukaryotes.* Nucleic Acids Res, 2008. **36**(2): p. e12.
78. Hu, J.C., et al., *What we can learn about Escherichia coli through application of Gene Ontology.* Trends Microbiol, 2009. **17**(7): p. 269-78.
79. Scheer, M., et al., *JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W510-5.
80. Arakawa, K., et al., *KEGG-based pathway visualization tool for complex omics data.* In Silico Biol, 2005. **5**(4): p. 419-23.
81. Caspi, R., et al., *The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res, 2008. **36**(Database issue): p. D623-31.
82. Novak, B. and J.J. Tyson, *Design principles of biochemical oscillators.* Nat Rev Mol Cell Biol, 2008. **9**(12): p. 981-91.
83. Tyson, J.J. and B. Novak, *Temporal organization of the cell cycle.* Curr Biol, 2008. **18**(17): p. R759-R768.
84. Lee, I., et al., *Predicting genetic modifier loci using functional gene networks.* Genome Res, 2010.
85. McGary, K.L., et al., *Systematic discovery of nonobvious human disease models through orthologous phenotypes.* Proc Natl Acad Sci U S A, 2010. **107**(14): p. 6544-9.
86. Pena-Castillo, L., et al., *A critical assessment of Mus musculus gene function prediction using integrated genomic evidence.* Genome Biol, 2008. **9 Suppl 1**: p. S2.
87. Abascal, F. and A. Valencia, *Clustering of proximal sequence space for the identification of protein families.* Bioinformatics, 2002. **18**(7): p. 908-21.

88. Hirschman, J.E., et al., *Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome.* Nucleic Acids Res, 2006. **34**(Database issue): p. D442-5.
89. Hong, E.L., et al., *Gene Ontology annotations at SGD: new data sources and annotation methods.* Nucleic Acids Res, 2008. **36**(Database issue): p. D577-81.
90. Okuda, S., et al., *KEGG Atlas mapping for global analysis of metabolic pathways.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W423-6.
91. Wolfram, S., *The Mathematica Book*. 4th ed1999: Cambridge University Press.
92. Dacks, J.B. and M.C. Field, *Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode.* J Cell Sci, 2007. **120**(Pt 17): p. 2977-85.
93. Keeling, P., *Five questions about microsporidia.* PLoS Pathog, 2009. **5**(9): p. e1000489.
94. Barbrook, A.C., et al., *Organization and expression of organellar genomes.* Philos Trans R Soc Lond B Biol Sci, 2010. **365**(1541): p. 785-97.
95. Martin, W., et al., *Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus.* Proc Natl Acad Sci U S A, 2002. **99**(19): p. 12246-51.
96. Raven, J.A. and J.F. Allen, *Genomics and chloroplast evolution: what did cyanobacteria do for plants?* Genome Biol, 2003. **4**(3): p. 209.
97. Keeling, P.J., *The endosymbiotic origin, diversification and fate of plastids.* Philos Trans R Soc Lond B Biol Sci, 2010. **365**(1541): p. 729-48.
98. Martin, W., *Evolutionary origins of metabolic compartmentalization in eukaryotes.* Philos Trans R Soc Lond B Biol Sci, 2010. **365**(1541): p. 847-55.
99. Poole, R.L., *The TAIR database.* Methods Mol Biol, 2007. **406**: p. 179-212.
100. Lockhart, S.R., et al., *Alpha-pheromone-induced "shmooing" and gene regulation require white-opaque switching during Candida albicans mating.* Eukaryot Cell, 2003. **2**(5): p. 847-55.
101. Xue, C.B., et al., *A covalently constrained congener of the Saccharomyces cerevisiae tridecapeptide mating pheromone is an agonist.* J Biol Chem, 1989. **264**(32): p. 19161-8.
102. Mora-Montes, H.M., et al., *Endoplasmic reticulum alpha-glycosidases of Candida albicans are required for N glycosylation, cell wall integrity, and normal host-fungus interaction.* Eukaryot Cell, 2007. **6**(12): p. 2184-93.

103. Tsai, P.K., et al., *Isolation of glucose-containing high-mannose glycoprotein core oligosaccharides.* Proc Natl Acad Sci U S A, 1984. **81**(20): p. 6340-3.
104. Herskowitz, I., *MAP kinase pathways in yeast: for mating and more.* Cell, 1995. **80**(2): p. 187-97.
105. Srikantha, T., S.A. Lachke, and D.R. Soll, *Three mating type-like loci in Candida glabrata.* Eukaryot Cell, 2003. **2**(2): p. 328-40.
106. Strathern, J.N., et al., *Homothallic switching of yeast mating type cassettes is initiated by a double-stranded cut in the MAT locus.* Cell, 1982. **31**(1): p. 183-92.
107. Gimeno, C.J., et al., *Unipolar cell divisions in the yeast S. cerevisiae lead to filamentous growth: regulation by starvation and RAS.* Cell, 1992. **68**(6): p. 1077-90.
108. Herrero, A.B., et al., *Candida albicans and Yarrowia lipolytica as alternative models for analysing budding patterns and germ tube formation in dimorphic fungi.* Microbiology, 1999. **145 (Pt 10)**: p. 2727-37.
109. Machado, C.R., et al., *Thi1, a thiamine biosynthetic gene in Arabidopsis thaliana, complements bacterial defects in DNA repair.* Plant Mol Biol, 1996. **31**(3): p. 585-93.
110. Schiestl, R.H., M. Dominska, and T.D. Petes, *Transformation of Saccharomyces cerevisiae with nonhomologous DNA: illegitimate integration of transforming DNA into yeast chromosomes and in vivo ligation of transforming DNA to mitochondrial DNA sequences.* Mol Cell Biol, 1993. **13**(5): p. 2697-705.
111. Jin, H., et al., *Characterization of the SOS response of Pseudomonas fluorescens strain DC206 using whole-genome transcript analysis.* FEMS Microbiol Lett, 2007. **269**(2): p. 256-64.
112. Astrom, S.U., S.M. Okamura, and J. Rine, *Yeast cell-type regulation of DNA repair.* Nature, 1999. **397**(6717): p. 310.
113. Kegel, A., et al., *Genome wide distribution of illegitimate recombination events in Kluyveromyces lactis.* Nucleic Acids Res, 2006. **34**(5): p. 1633-45.
114. Cormack, B.P. and S. Falkow, *Efficient homologous and illegitimate recombination in the opportunistic yeast pathogen Candida glabrata.* Genetics, 1999. **151**(3): p. 979-87.
115. Zhu, J. and R.H. Schiestl, *Topoisomerase I involvement in illegitimate recombination in Saccharomyces cerevisiae.* Mol Cell Biol, 1996. **16**(4): p. 1805-12.
116. Xu, H., et al., *Isoleucine biosynthesis in Leptospira interrogans serotype lai strain 56601 proceeds via a threonine-independent pathway.* J Bacteriol, 2004. **186**(16): p. 5400-9.

117. Atsumi, S., T. Hanai, and J.C. Liao, *Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels.* Nature, 2008. **451**(7174): p. 86-9.
118. Howell, D.M., H. Xu, and R.H. White, *(R)-citramalate synthase in methanogenic archaea.* J Bacteriol, 1999. **181**(1): p. 331-3.
119. Risso, C., et al., *Elucidation of an alternate isoleucine biosynthesis pathway in Geobacter sulfurreducens.* J Bacteriol, 2008. **190**(7): p. 2266-74.
120. Christensen, K.E. and R.E. MacKenzie, *Mitochondrial one-carbon metabolism is adapted to the specific needs of yeast, plants and mammals.* Bioessays, 2006. **28**(6): p. 595-605.
121. Pasternack, L.B., D.A. Laude, Jr., and D.R. Appling, *Whole-cell detection by 13C NMR of metabolic flux through the C1-tetrahydrofolate synthase/serine hydroxymethyltransferase enzyme system and effect of antifolate exposure in Saccharomyces cerevisiae.* Biochemistry, 1994. **33**(23): p. 7166-73.
122. Prabhu, V., et al., *13C nuclear magnetic resonance detection of interactions of serine hydroxymethyltransferase with C1-tetrahydrofolate synthase and glycine decarboxylase complex activities in Arabidopsis.* Plant Physiol, 1996. **112**(1): p. 207-16.
123. Descartes, R., *A discourse of a method for the well guiding of reason, and the discovery of truth in the sciences (1649)*, 1649 (2011), EEBO Editions, ProQuest.
124. Damasio, A., *Descartes' Error: Emotion, Reason, and the Human Brain*2005: Penguin.
125. Smith, J.D., *The study of animal metacognition.* Trends Cogn Sci, 2009. **13**(9): p. 389-96.
126. Beran, M.J. and J.D. Smith, *Information seeking by rhesus monkeys (Macaca mulatta) and capuchin monkeys (Cebus apella).* Cognition, 2011. **120**(1): p. 90-105.
127. Redford, J.S., *Evidence of metacognitive control by humans and monkeys in a perceptual categorization task.* J Exp Psychol Learn Mem Cogn, 2010. **36**(1): p. 248-54.
128. Rutz, C., et al., *The ecological significance of tool use in New Caledonian crows.* Science, 2010. **329**(5998): p. 1523-6.
129. Emery, N.J. and N.S. Clayton, *The mentality of crows: convergent evolution of intelligence in corvids and apes.* Science, 2004. **306**(5703): p. 1903-7.
130. Ye, K. and Z. Gu, *Recent advances in understanding the role of nutrition in human genome evolution.* Adv Nutr, 2011. **2**(6): p. 486-96.

131. Nakashima, H., K. Nishikawa, and T. Ooi, *Differences in dinucleotide frequencies of human, yeast, and Escherichia coli genes.* DNA Res, 1997. **4**(3): p. 185-92.
132. Noonan, J.P., *Neanderthal genomics and the evolution of modern humans.* Genome Res, 2010. **20**(5): p. 547-53.
133. Lieschke, G.J. and P.D. Currie, *Animal models of human disease: zebrafish swim into view.* Nat Rev Genet, 2007. **8**(5): p. 353-67.
134. Hughes, J.F., et al., *Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes.* Nature, 2012. **483**(7387): p. 82-6.
135. Somel, M., et al., *MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates.* PLoS Biol, 2011. **9**(12): p. e1001214.
136. Molaro, A., et al., *Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates.* Cell, 2011. **146**(6): p. 1029-41.
137. Crisci, J.L., et al., *On characterizing adaptive events unique to modern humans.* Genome Biol Evol, 2011. **3**: p. 791-8.
138. McLean, C.Y., et al., *Human-specific loss of regulatory DNA and the evolution of human-specific traits.* Nature, 2011. **471**(7337): p. 216-9.
139. Pai, A.A., et al., *A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues.* PLoS Genet, 2011. **7**(2): p. e1001316.
140. Paar, V., et al., *Large tandem, higher order repeats and regularly dispersed repeat units contribute substantially to divergence between human and chimpanzee Y chromosomes.* J Mol Evol, 2011. **72**(1): p. 34-55.
141. de Lichtenberg, U., et al., *Dynamic complex formation during the yeast cell cycle.* Science, 2005. **307**(5710): p. 724-7.
142. Pache, R.A. and P. Aloy, *A novel framework for the comparative analysis of biological networks.* PLoS One, 2012. **7**(2): p. e31220.
143. Ideker, T. and N.J. Krogan, *Differential network biology.* Mol Syst Biol, 2012. **8**: p. 565.
144. Bandyopadhyay, S., et al., *Rewiring of genetic networks in response to DNA damage.* Science, 2010. **330**(6009): p. 1385-9.
145. Beltrao, P. and L. Serrano, *Specificity and evolvability in eukaryotic protein interaction networks.* PLoS Comput Biol, 2007. **3**(2): p. e25.
146. Srivas, R., et al., *Assembling global maps of cellular function through integrative analysis of physical and genetic networks.* Nat Protoc, 2011. **6**(9): p. 1308-23.

147. Karathia, H., et al., *Saccharomyces cerevisiae as a model organism: a comparative study.* PLoS One, 2011. **6**(2): p. e16015.
148. Helmerhorst, E.J., et al., *Mass spectrometric identification of key proteolytic cleavage sites in statherin affecting mineral homeostasis and bacterial binding domains.* Journal of proteome research, 2010. **9**(10): p. 5413-21.
149. Carbon, S., et al., *AmiGO: online access to ontology and annotation data.* Bioinformatics, 2009. **25**(2): p. 288-9.
150. Oz, O., et al., *Effect of sitagliptin monotherapy on serum total ghrelin levels in people with type 2 diabetes.* Diabetes Res Clin Pract, 2011. **94**(2): p. 212-6.
151. Harmar, A.J., et al., *IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels.* Nucleic Acids Res, 2009. **37**(Database issue): p. D680-5.
152. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2005. **33**(Database issue): p. D54-8.
153. Kandasamy, K., et al., *NetPath: a public resource of curated signal transduction pathways.* Genome Biol, 2010. **11**(1): p. R3.
154. Matthews, L., et al., *Reactome knowledgebase of human biological pathways and processes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D619-22.
155. Mishra, G.R., et al., *Human protein reference database--2006 update.* Nucleic Acids Res, 2006. **34**(Database issue): p. D411-4.
156. Fung, K.Y., et al., *Characterization of the in vivo forms of lacrimal-specific proline-rich proteins in human tear fluid.* Proteomics, 2004. **4**(12): p. 3953-9.
157. Daly, J.W., et al., *Individual and geographic variation of skin alkaloids in three species of Madagascan poison frogs (Mantella).* J Chem Ecol, 2008. **34**(2): p. 252-79.
158. Saporito, R.A., et al., *Alkaloids in the mite Scheloribates laevigatus: further alkaloids common to oribatid mites and poison frogs.* J Chem Ecol, 2011. **37**(2): p. 213-8.
159. Bentin-Ley, U., et al., *Glycodelin in endometrial flushing fluid and endometrial biopsies from infertile and fertile women.* Eur J Obstet Gynecol Reprod Biol, 2011. **156**(1): p. 60-6.
160. Hara, M., et al., *Ghrelin levels are reduced in Rett syndrome patients with eating difficulties.* Int J Dev Neurosci, 2011. **29**(8): p. 899-902.
161. Kantorova, E., et al., *Leptin, adiponectin and ghrelin, new potential mediators of ischemic stroke.* Neuro Endocrinol Lett, 2011. **32**(5): p. 716-21.

162. Ledderose, C., S. Kreth, and A. Beiras-Fernandez, *Ghrelin, a novel peptide hormone in the regulation of energy balance and cardiovascular function.* Recent Pat Endocr Metab Immune Drug Discov, 2011. **5**(1): p. 1-6.
163. Alexaki, V.I., et al., *B-cell maturation antigen (BCMA) activation exerts specific proinflammatory effects in normal human keratinocytes and is preferentially expressed in inflammatory skin pathologies.* Endocrinology, 2012. **153**(2): p. 739-49.
164. Park, K.J., et al., *Death receptors 4 and 5 activate Nox1 NADPH oxidase through riboflavin kinase to induce reactive oxygen species-mediated apoptotic cell death.* The Journal of biological chemistry, 2012. **287**(5): p. 3313-25.
165. Wiens, G.D. and G.W. Glenney, *Origin and evolution of TNF and TNF receptor superfamilies.* Dev Comp Immunol, 2011. **35**(12): p. 1324-35.
166. Copeland, S., et al., *Acute inflammatory response to endotoxin in mice and humans.* Clin Diagn Lab Immunol, 2005. **12**(1): p. 60-7.
167. Stockmann, P., et al., *Increased human defensine levels hint at an inflammatory etiology of bisphosphonate-associated osteonecrosis of the jaw: an immunohistological study.* J Transl Med, 2011. **9**: p. 135.
168. Nash, K.T., et al., *Requirement of KISS1 secretion for multiple organ metastasis suppression and maintenance of tumor dormancy.* J Natl Cancer Inst, 2007. **99**(4): p. 309-21.
169. Cho, S.G., et al., *KiSS1 suppresses TNFalpha-induced breast cancer cell invasion via an inhibition of RhoA-mediated NF-kappaB activation.* J Cell Biochem, 2009. **107**(6): p. 1139-49.
170. Ohtaki, T., et al., *Metastasis suppressor gene KiSS-1 encodes peptide ligand of a G-protein-coupled receptor.* Nature, 2001. **411**(6837): p. 613-7.
171. Nash, K.T. and D.R. Welch, *The KISS1 metastasis suppressor: mechanistic insights and clinical utility.* Front Biosci, 2006. **11**: p. 647-59.
172. Demetris, A.J., et al., *Small proline-rich proteins (SPRR) function as SH3 domain ligands, increase resistance to injury and are associated with epithelial-mesenchymal transition (EMT) in cholangiocytes.* J Hepatol, 2008. **48**(2): p. 276-88.
173. Lisitsyn, N.A., et al., *Enteric alpha defensins in norm and pathology.* Ann Clin Microbiol Antimicrob, 2012. **11**: p. 1.
174. Semple, F. and J.R. Dorin, *beta-Defensins: Multifunctional Modulators of Infection, Inflammation and More?* J Innate Immun, 2012.

175. Arnett, E. and S. Seveau, *The multifaceted activities of mammalian defensins.* Curr Pharm Des, 2011. **17**(38): p. 4254-69.
176. Huang, J.K., et al., *Retinoid X receptor gamma signaling accelerates CNS remyelination.* Nat Neurosci, 2011. **14**(1): p. 45-53.
177. Papi, A., et al., *RXRgamma and PPARgamma ligands in combination to inhibit proliferation and invasiveness in colon cancer cells.* Cancer Lett, 2010. **297**(1): p. 65-74.
178. Philip, S., et al., *Adaptive evolution of the Retinoid X receptor in vertebrates.* Genomics, 2012. **99**(2): p. 81-9.
179. Dawson, M.I. and Z. Xia, *The retinoid X receptors and their ligands.* Biochim Biophys Acta, 2012. **1821**(1): p. 21-56.
180. Yang, M.H., et al., *Connective tissue growth factor modulates oral squamous cell carcinoma invasion by activating a miR-504/FOXP1 signalling.* Oncogene, 2012. **31**(19): p. 2401-11.
181. Abi-Rached, L., et al., *Human-specific evolution and adaptation led to major qualitative differences in the variable receptors of human and chimpanzee natural killer cells.* PLoS Genet, 2010. **6**(11): p. e1001192.
182. Medeiros, A., et al., *Mutations in the human phospholamban gene in patients with heart failure.* Am Heart J, 2011. **162**(6): p. 1088-1095 e1.
183. Fu, W., et al., *Human immunodeficiency virus type 1, human protein interaction database at NCBI.* Nucleic Acids Res, 2009. **37**(Database issue): p. D417-22.
184. Gautier, V.W., et al., *In vitro nuclear interactome of the HIV-1 Tat protein.* Retrovirology, 2009. **6**: p. 47.
185. Barboric, M. and B.M. Peterlin, *A new paradigm in eukaryotic biology: HIV Tat and the control of transcriptional elongation.* PLoS biology, 2005. **3**(2): p. e76.
186. Hetzer, C., et al., *Decoding Tat: the biology of HIV Tat posttranslational modifications.* Microbes and infection / Institut Pasteur, 2005. **7**(13): p. 1364-9.
187. Correa, S.G., et al., *High dissemination and hepatotoxicity in rats infected with Candida albicans after stress exposure: potential sensitization to liver damage.* Int Immunol, 2004. **16**(12): p. 1761-8.
188. Mayer, F.L., et al., *The novel Candida albicans transporter Dur31 Is a multi-stage pathogenicity factor.* PLoS pathogens, 2012. **8**(3): p. e1002592.
189. Odds, F.C., *Pathogenesis of Candida infections.* Journal of the American Academy of Dermatology, 1994. **31**(3 Pt 2): p. S2-5.

190. Finley, K.R. and J. Berman, *Microtubules in Candida albicans hyphae drive nuclear dynamics and connect cell cycle progression to morphogenesis.* Eukaryotic cell, 2005. **4**(10): p. 1697-711.

191. Yin, H., et al., *Stu1p is physically associated with beta-tubulin and is required for structural integrity of the mitotic spindle.* Mol Biol Cell, 2002. **13**(6): p. 1881-92.

192. Guenthner, P.C., W.E. Secor, and C.S. Dezzutti, *Trichomonas vaginalis-induced epithelial monolayer disruption and human immunodeficiency virus type 1 (HIV-1) replication: implications for the sexual transmission of HIV-1.* Infection and immunity, 2005. **73**(7): p. 4155-60.

193. Wasserheit, J.N., *Epidemiological synergy. Interrelationships between human immunodeficiency virus infection and other sexually transmitted diseases.* Sex Transm Dis, 1992. **19**(2): p. 61-77.

194. Jensen, L.J., et al., *Prediction of human protein function from post-translational modifications and localization features.* J Mol Biol, 2002. **319**(5): p. 1257-65.

195. Chen, T.W., et al., *DODO: an efficient orthologous genes assignment tool based on domain architectures. Domain based ortholog detection.* BMC Bioinformatics, 2010. **11 Suppl 7**: p. S6.

196. Lee, Y., et al., *Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA).* Genome Res, 2002. **12**(3): p. 493-502.

197. Shi, G., M.C. Peng, and T. Jiang, *MultiMSOAR 2.0: an accurate tool to identify ortholog groups among multiple genomes.* PLoS One, 2011. **6**(6): p. e20892.

198. Shi, G., L. Zhang, and T. Jiang, *MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement.* BMC Bioinformatics, 2010. **11**: p. 10.

199. Scally, A., et al., *Insights into hominid evolution from the gorilla genome sequence.* Nature, 2012. **483**(7388): p. 169-75.

200. Modzelewska, K., et al., *An activating mutation in sos-1 identifies its Dbl domain as a critical inhibitor of the epidermal growth factor receptor pathway during Caenorhabditis elegans vulval development.* Mol Cell Biol, 2007. **27**(10): p. 3695-707.

201. Boonen, K., J.W. Creemers, and L. Schoofs, *Bioactive peptides, networks and systems biology.* Bioessays, 2009. **31**(3): p. 300-14.

202. Amaya, M., A. Baranova, and M.L. van Hoek, *Protein prenylation: a new mode of host-pathogen interaction.* Biochem Biophys Res Commun, 2011. **416**(1-2): p. 1-6.

203. Stolf, B.S., et al., *Protein disulfide isomerase and host-pathogen interaction.* ScientificWorldJournal, 2011. **11**: p. 1749-61.

204. De Jesus, J.B., et al., *A further proteomic study on the effect of iron in the human pathogen Trichomonas vaginalis.* Proteomics, 2007. **7**(12): p. 1961-72.

205. Kirichenko, E.B., Y.V. Orlova, and D.V. Kurilov, *Component composition of the essential oil of the Artemisia lerchiana Web. from the southeastern Russia.* Dokl Biochem Biophys, 2008. **422**: p. 292-5.

206. Mikriakov, V.R., M.A. Stepanova, and D.V. Mikriakov, *[Dependence of the infestation intensity of zope Abramis ballerus with Dactylogyrus chranilowi Bychowsky, 1931 on the level of antimicrobial effect of the host's blood serum]*. Parazitologiia, 2011. **45**(1): p. 50-3.

207. Le Naour, F., et al., *Tetraspanins connect several types of Ig proteins: IgM is a novel component of the tetraspanin web on B-lymphoid cells.* Cancer Immunol Immunother, 2004. **53**(3): p. 148-52.

208. Berkowicz, D.A., G.O. Barnett, and H.C. Chueh, *Component architecture for web based EMR applications.* Proc AMIA Symp, 1998: p. 116-20.

209. Myers, D.L., K.S. Culp, and R.S. Miller, *Use of a Web-based process model to implement security and data protection as an integral component of clinical information management.* Proc AMIA Symp, 1999: p. 897-900.

210. Zhang, H., et al., *Quantifying aggregation dynamics during Myxococcus xanthus development.* J Bacteriol, 2011. **193**(19): p. 5164-70.

211. Xie, C., et al., *Statistical image analysis reveals features affecting fates of Myxococcus xanthus developmental aggregates.* Proc Natl Acad Sci U S A, 2011. **108**(14): p. 5915-20.

212. Tiwari, A., et al., *Bistable responses in bacterial genetic networks: designs and dynamical consequences.* Math Biosci, 2011. **231**(1): p. 76-89.

213. Narula, J. and O.A. Igoshin, *Thermodynamic models of combinatorial gene regulation by distant enhancers.* IET Syst Biol, 2010. **4**(6): p. 393-408.

214. Chaudhury, S. and O.A. Igoshin, *Dynamic disorder in quasi-equilibrium enzymatic systems.* PLoS One, 2010. **5**(8): p. e12364.

215. Tiwari, A., et al., *The interplay of multiple feedback loops with post-translational kinetics results in bistability of mycobacterial stress response.* Phys Biol, 2010. **7**(3): p. 036005.

216. Narula, J., et al., *Modeling reveals bistability and low-pass filtering in the network module determining blood stem cell fate.* PLoS Comput Biol, 2010. **6**(5): p. e1000771.

217. Eswaramoorthy, P., et al., *Single-cell measurement of the levels and distributions of the phosphorelay components in a population of sporulating Bacillus subtilis cells.* Microbiology, 2010. **156**(Pt 8): p. 2294-304.

218. Ray, J.C. and O.A. Igoshin, *Adaptable functionality of transcriptional feedback in bacterial two-component systems.* PLoS Comput Biol, 2010. **6**(2): p. e1000676.

219. Kelley, B.P., et al., *PathBLAST: a tool for alignment of protein interaction networks.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W83-8.

220. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.

221. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

222. *Enzyme nomenclature. Report on the recommendations (1964) of the International Union of Biochemistry on Nomenclature and Classification of Enzymes.* Science, 1965. **150**(3697): p. 719-21.

223. Stark, C., et al., *The BioGRID Interaction Database: 2011 update.* Nucleic Acids Res, 2011. **39**(Database issue): p. D698-704.

224. Sharman, J.L., et al., *IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data.* Nucleic Acids Res, 2011. **39**(Database issue): p. D534-8.

225. *Wikipedia. Malaria. Available from:* ***http://en.wikipedia.org/wiki/Malaria***.

226. Alves, R. and M.A. Savageau, *Extending the method of mathematically controlled comparison to include numerical comparisons.* Bioinformatics, 2000. **16**(9): p. 786-98.

227. Xia, X.Q., M. McClelland, and Y. Wang, *TabSQL: a MySQL tool to facilitate mapping user data to public databases.* BMC Bioinformatics, 2010. **11**: p. 342.

228. Ursos, L.M., S.M. Dzekunov, and P.D. Roepe, *The effects of chloroquine and verapamil on digestive vacuolar pH of P. falciparum either sensitive or resistant to chloroquine.* Molecular and biochemical parasitology, 2000. **110**(1): p. 125-34.

229. Dzekunov, S.M., L.M. Ursos, and P.D. Roepe, *Digestive vacuolar pH of intact intraerythrocytic P. falciparum either sensitive or resistant to chloroquine.* Molecular and biochemical parasitology, 2000. **110**(1): p. 107-24.

230. Smith, R.P., et al., *EST Express: PHP/MySQL based automated annotation of ESTs from expression libraries.* BMC Bioinformatics, 2008. **9**: p. 186.

231. McDonald, A.G., et al., *ExplorEnz: a MySQL database of the IUBMB enzyme nomenclature.* BMC Biochem, 2007. **8**: p. 14.

232. Nagaraj, V.A., et al., *Unique properties of Plasmodium falciparum porphobilinogen deaminase.* The Journal of biological chemistry, 2008. **283**(1): p. 437-44.

233. Sato, S., et al., *Enzymes for heme biosynthesis are found in both the mitochondrion and plastid of the malaria parasite Plasmodium falciparum.* Protist, 2004. **155**(1): p. 117-25.

234. Padmanaban, G., V.A. Nagaraj, and P.N. Rangarajan, *An alternative model for heme biosynthesis in the malarial parasite.* Trends in biochemical sciences, 2007. **32**(10): p. 443-9.

235. Gunther, S., et al., *Apicoplast lipoic acid protein ligase B is not essential for Plasmodium falciparum.* PLoS pathogens, 2007. **3**(12): p. e189.

236. Morgans, D. and P.R. Carroll, *A direct acting adrenergic component of the venom of the Sydney funnel-web spider, Atrax robustus.* Toxicon, 1976. **14**(3): p. 185-9.

237. Pei, Y., et al., *Plasmodium pyruvate dehydrogenase activity is only essential for the parasite's progression from liver infection to blood infection.* Molecular microbiology, 2010.

238. Ray, J.C., J.J. Tabor, and O.A. Igoshin, *Non-transcriptional regulatory processes shape transcriptional network dynamics.* Nat Rev Microbiol, 2011. **9**(11): p. 817-28.

239. Lundberg, J.G., et al., *A major food web component in the orinoco river channel: evidence from planktivorous electric fishes.* Science, 1987. **237**(4810): p. 81-3.

240. Bulusu, V., V. Jayaraman, and H. Balaram, *Metabolic fate of fumarate, a side product of the purine salvage pathway in the intraerythrocytic stages of Plasmodium falciparum.* The Journal of biological chemistry, 2011. **286**(11): p. 9236-45.

241. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.

242. Schneider, A., C. Dessimoz, and G.H. Gonnet, *OMA Browser--exploring orthologous relations across 352 complete genomes.* Bioinformatics, 2007. **23**(16): p. 2180-2.

243. Schmidt, T. and D. Frishman, *PROMPT: a protein mapping and comparison tool.* BMC Bioinformatics, 2006. **7**: p. 331.

244. Reiter, L.T., et al., *Accentuate the negative: proteome comparisons using the negative proteome database.* Fly (Austin), 2007. **1**(3): p. 164-71.

245. Pache, R.A., A. Ceol, and P. Aloy, *NetAligner--a network alignment server to compare complexes, pathways and whole interactomes.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W157-61.

246. Ostlund, G., et al., *InParanoid 7: new algorithms and tools for eukaryotic orthology analysis.* Nucleic Acids Res, 2010. **38**(Database issue): p. D196-203.

247. Orton, R.J., W.I. Sellers, and D.L. Gerloff, *YETI: Yeast Exploration Tool Integrator.* Bioinformatics, 2004. **20**(2): p. 284-5.

248. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.

249. Li, J.B., et al., *Procom: a web-based tool to compare multiple eukaryotic proteomes.* Bioinformatics, 2005. **21**(8): p. 1693-4.

250. Li, H., et al., *TreeFam: a curated database of phylogenetic trees of animal gene families.* Nucleic Acids Res, 2006. **34**(Database issue): p. D572-80.

251. Jensen, L.J., et al., *eggNOG: automated construction and annotation of orthologous groups of genes.* Nucleic Acids Res, 2008. **36**(Database issue): p. D250-4.

252. Homologene, N., *Available from:* ***http://www.ncbi.nlm.nih.gov/homologene****.* 2011.

253. Deluca, T.F., et al., *Roundup: a multi-genome repository of orthologs and evolutionary distances.* Bioinformatics, 2006. **22**(16): p. 2044-6.

254. Dehal, P.S., et al., *MicrobesOnline: an integrated portal for comparative and functional genomics.* Nucleic Acids Res, 2010. **38**(Database issue): p. D396-400.

255. Datta, R.S., et al., *Berkeley PHOG: PhyloFacts orthology group prediction web server.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W84-9.

256. Karp, P.D., et al., *Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.* Brief Bioinform, 2010. **11**(1): p. 40-79.

257. Gordon, R.S., *Zope 3 developer's handbook.* Library Journal, 2005. **130**(8): p. 114-114.

258. Igoshin, O.A., R. Alves, and M.A. Savageau, *Hysteretic and graded responses in bacterial two-component signal transduction.* Mol Microbiol, 2008. **68**(5): p. 1196-215.

259. Gordon, R.S., *MySQL: The complete reference.* Library Journal, 2004. **129**(20): p. 152-152.

260. Sun, W., et al., *The crystal structure of Aquifex aeolicus prephenate dehydrogenase reveals the mode of tyrosine inhibition.* The Journal of biological chemistry, 2009. **284**(19): p. 13223-32.

261. Scheer, M., et al., *BRENDA, the enzyme information system in 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D670-6.

262. Barrell, D., et al., *The GOA database in 2009--an integrated Gene Ontology Annotation resource.* Nucleic Acids Res, 2009. **37**(Database issue): p. D396-403.

263. Chaudhury, S. and O.A. Igoshin, *Dynamic disorder-driven substrate inhibition and bistability in a simple enzymatic reaction.* J Phys Chem B, 2009. **113**(40): p. 13421-8.

264. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

265. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

Index