

*Integrative approaches to investigate the
molecular basis of diseases and
adverse drug reactions:
from multivariate statistical analysis
to systems biology.*

Anna Bauer-Mehren

DOCTORAL THESIS

THESIS DIRECTORS

Drs. Ferran Sanz and Laura I. Furlong

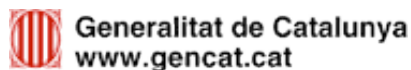


Department of Experimental and Health Services

Barcelona, 2010

The research in this PhD thesis was carried out at the Integrative Biomedical Informatics Laboratory (IBI) of the Research Programme on Biomedical Informatics (GRIB), which belongs to the IMIM and to the Department of Experimental and Health Sciences of UPF, and which is located at the Barcelona Biomedical Research Park (PRBB).

This research was funded in part by the EU-ADR and @neurist EU-funded projects, as well as by the Generalitat de Catalunya by means of a fellowship (Ajut FI_B) to the author of this thesis.



Für Papa, Mama und Frank

Acknowledgements

First of all I would like to thank my two thesis directors Ferran Sanz and Laura Furlong.

Ferran has always supported me in my scientific career. I would like to especially thank him for introducing me to the work in European projects and international conferences. I appreciate a lot that he always had confidence in my work and that he gave me a lot of responsibility from the very beginning. Ferran has always been a real mentor to me.

My very special thanks go to Laura, who brought me to the group and who was extremely supportive throughout the whole time. Laura was always very interested in my scientific career and taught me how to design and approach scientific questions, how to present the results at international meetings and conferences and finally how to publish the work in peer-reviewed journals. Next to being the best teacher possible to have as PhD student, Laura has become a really good friend. Thank you so much, Laura!

A fellowship by the Generalitat de Catalunya (Ajut FI_B) as well as funding by the EU-ADR and @neurist EU-funded projects is acknowledged.

I would like to thank my colleagues, who always supported me scientifically but also as friends. Special thanks go to Michael, colleague and friend, with whom working was not only very successful but also a lot of fun. I want to thank Laura L. for the great time here at the GRIB and especially while doing sports, the spinning, running etc. hours have been a great help on the way to finishing the PhD thesis. Moreover, I would like to thank Ingo (I will miss the coffee breaks in the morning), Sonja A. and Sonja H., Montse, Jana, Ferran, Ricard, Àngel, MarieCarmen, Kashif, Nils, Cesar, Emre, Carina, Martina, Chus and all the other members of the GRIB. Special thanks go to the members of the incredibly awesome “We play for fun” volleyball team. WPF rules!!! I also want to thank Rebecca, Angela, Bernd and Ben for the great time here. Barbecues, beach days and coffee breaks have always been so much fun with you. Also, I want to thank my friends back in Munich or England, Claudia, Caro, Mira, Markus, Fabian, Vicky, Ben and especially Doris for building our “PhD islands” that were a great help on the way to finishing the PhD thesis.

Very special thanks and kisses go to my whole family. Thank you all so much for being there. I want to thank Ricci, who always kept me up to date and never got tired of endless stories and moaning on the phone, Franzi for organizing amazing surprise visits and who always picks me up at the airport, Ala for giving perfect advices for the doctorate thesis and for the phone calls late at night and especially to my mother Renata who always supported my decision to go abroad for doing a PhD and who always has the best advices for life.

Finally, I would like to thank the most important person to me, Frank. Thank you for your help, support and patience in the last three years. Thank you for always believing in me and for encouraging me. Thank you for the amazing time we have spent together. I love you so much Franki!!!

Abstract

Despite some great success, many human diseases cannot be effectively treated, prevented or cured, yet. Moreover, prescribed drugs are often not very efficient and cause undesired side effects. Hence, there is a need to investigate the molecular basis of diseases and adverse drug reactions in more detail. For this purpose, relevant biomedical data needs to be gathered, integrated and analysed in a meaningful way.

In this regard, we have developed novel integrative analysis approaches based on both perspectives, classical multivariate statistics and systems biology. A novel multilevel statistical method has been developed for exploiting molecular and pharmacological information for a set of drugs in order to investigate undesired side effects. Systems biology approaches have been used to study the genetic basis of human diseases at a global scale. For this purpose, we have developed an integrated gene-disease association database and tools for user-friendly access and analysis. We showed that modularity applies for mendelian, complex and environmental diseases and identified disease-related core biological processes. We have constructed a workflow to investigate adverse drug reactions using our gene-disease association database. A detailed study of currently available pathway data has been performed to evaluate its applicability to build network models. Finally, a strategy to integrate information about sequence variations with biological pathways has been implemented to study the effect of the sequence variations onto biological processes.

In summary, the developed methods are of immense practical value for other biomedical researchers and can aid to improve the understanding of the molecular basis of diseases and adverse drug reactions.

Resumen

A pesar de que existen tratamientos eficaces para las enfermedades, no hay todavía una cura o un tratamiento efectivo para muchas de ellas. Asimismo los medicamentos pueden ser ineficaces o causar efectos secundarios indeseables. Por lo tanto, es necesario investigar en profundidad las bases moleculares de las enfermedades y de los efectos secundarios de los medicamentos. Para ello, es necesario identificar y analizar de forma integrada los datos biomédicos relevantes.

En este sentido, hemos desarrollado nuevos métodos de análisis e integración de datos biomédicos que van desde el análisis estadístico multivariante a la biología de sistemas. En primer lugar, hemos desarrollado un nuevo método estadístico multinivel para la explotación de la información molecular y farmacológica de un conjunto de drogas a fin de investigar efectos secundarios no deseados. Luego,

hemos usado métodos de biología de sistemas para estudiar las bases genéticas de enfermedades humanas a escala global. Para ello, hemos integrado en una base de datos asociaciones entre genes y enfermedades y hemos desarrollado herramientas para el fácil acceso y análisis de los datos. Mostramos que las enfermedades mendelianas, complejas y ambientales presentan modularidad e identificamos los procesos biológicos relacionados con dichas enfermedades. Hemos construido una herramienta para investigar las reacciones adversas a los medicamentos basada en nuestra base de datos de asociaciones entre genes y enfermedades. Realizamos un estudio detallado de los datos disponibles sobre los procesos biológicos para evaluar su aplicabilidad en la construcción de modelos dinámicos. Por último, desarrollamos una estrategia para integrar la información sobre las variaciones de secuencia de genes con los procesos biológicos para estudiar el efecto de dichas variaciones en los procesos biológicos.

En resumen, los métodos presentados en esta tesis constituyen una herramienta valiosa para otros investigadores y pueden ayudar a mejorar la comprensión de las bases moleculares de las enfermedades y de las reacciones adversas a los medicamentos.

Preface

Ever since, people have been suffering from diseases and thus scientists and health professionals have been trying to unravel and understand the underlying mechanisms in order to find effective remedies. In this context, drugs constitute key tools to treat the diseases. However, a significant number of diseases cannot be effectively treated yet. Additionally, many of the prescribed drugs are not efficient enough and cause undesirable adverse reactions. Thus, there is an urgent need to further investigate the molecular basis of diseases and adverse drug reactions. For this purpose relevant biomedical data has to be gathered, integrated and analysed in a meaningful way. However, biomedical data is typically very heterogeneous and scattered over various repositories and the scientific literature.

Hence, the main objective of this PhD thesis was the development and application of novel integrative data analysis approaches to investigate the molecular basis of diseases and adverse drug reactions in more detail. In this regard, the biomedical problems were studied from two different points of view: classical multivariate statistics and systems biology.

This PhD thesis addresses several studies on the mechanisms underlying diseases and adverse drug reactions and introduces several novel methods and tools, which are capable to deal with the magnitude, diversity and fragmentation of biomedical data. Each study started with an extensive evaluation of available and relevant biomedical data followed by the development of tools for its automatic retrieval, integration and subsequent analysis of the gained information.

In particular, we started the work with the more classical multivariate statistical approaches and developed a novel multilevel statistical method for its application in drug discovery projects (Selent *et al*, 2010). When studying the statistical approaches it became clear that statistical associations are not necessarily causal and other approaches have to be followed up to get a mechanistic understanding of the biological processes under study. Such mechanistic understanding requires the integration of all information available about the involved key players and how they interact in the cell. Hence, the focus shifted from statistical to systems biology approaches, which are trying to study a biological system at a more global scale by considering the interactions between the involved entities. These interactions are typically represented by means of biological networks. We systematically studied public pathway data regarding its accessibility and applicability for building models that simulate static and dynamic properties of biological networks (Bauer-Mehren *et al*, 2009b). Another work started with the extensive analysis of repositories of the state-of-the-art knowledge about the genetic origin of human diseases. We developed an integrated gene-disease association database and implemented tools for retrieval, integration and analysis (Bauer-Mehren *et al*, 2010b; Bundschus *et al*, 2010). A detailed study based on the newly integrated gene-disease associations revealed several interesting findings about human diseases (Bauer-Mehren *et al*, 2010a). We furthermore present a workflow for investigating the molecular

mechanisms underlying adverse drug reactions, which also uses the our gene-disease association database (see section 3.1.4). In another work, we introduce an approach to integrate biological pathways with information about the functional effects of sequence variations to build network models that allow assessing the effect of the variations on the dynamics of the biological processes. This approach can be of particular interest to investigate the mechanisms underlying diseases and adverse drug reactions being associated to sequence variations (Bauer-Mehren *et al*, 2009a).

In summary, the developed approaches and tools are very user-friendly and hence of immense value for other biomedical researchers. Their direct applicability in clinical practice and drug development projects has been demonstrated. It is therefore hoped that these methodologies and tools will eventually help to improve our understanding of the molecular basis of diseases and adverse drug reactions and will hence bring us closer to an improved clinical practice being more personalized and preventive.

This PhD thesis is divided into several chapters. In the first chapter typical biomedical problems are introduced regarding the understanding of the mechanisms underlying diseases and adverse drug reactions. Moreover, the need of data integration in biomedical research is discussed in detail. Current statistical and systems biology approaches are introduced. In chapter two, the objectives of this PhD thesis are explained and it is pointed out how they have been addressed. In the third chapter, the research carried out during this PhD thesis is presented. The final discussion critically evaluates the research carried out and puts the findings into the current state-of-the-art context. Finally, the list of publications resulting from the research carried out during this PhD thesis is provided.

Table of contents

| | |
|--|-----|
| ACKNOWLEDGEMENTS..... | VII |
| ABSTRACT | IX |
| PREFACE..... | XI |
| 1. INTRODUCTION | 1 |
| 1.1. TYPICAL PROBLEMS IN BIOMEDICAL RESEARCH | 3 |
| 1.1.1. HUMAN DISEASE MECHANISMS..... | 3 |
| 1.1.2. ADVERSE DRUG REACTION MECHANISMS..... | 7 |
| 1.1.3. SUMMARY | 9 |
| 1.2. NEED OF DATA INTEGRATION IN BIOMEDICAL RESEARCH | 11 |
| 1.2.1. BIOMEDICAL DATA SOURCES | 13 |
| 1.2.1.1. ENTREZ GENE | 15 |
| 1.2.1.2. UNIPROT | 15 |
| 1.2.1.3. GENE-DISEASE ASSOCIATION DATABASES | 15 |
| 1.2.1.4. PATHWAY DATABASES | 16 |
| 1.2.1.5. TEXT-MINING DERIVED INFORMATION | 18 |
| 1.2.2. DATA HETEROGENEITY AND STANDARDS..... | 19 |
| 1.2.3. COMPUTATIONAL DATA INTEGRATION APPROACHES..... | 20 |
| 1.2.4. ONTOLOGIES | 21 |
| 1.2.5. SUMMARY | 22 |
| 1.3. INTEGRATIVE BIOMEDICAL ANALYSIS APPROACHES..... | 23 |
| 1.3.1. STATISTICAL APPROACHES | 23 |
| 1.3.2. SYSTEMS BIOLOGY APPROACHES..... | 24 |
| 1.3.2.1. STATIC NETWORK ANALYSIS | 27 |
| 1.3.2.2. DYNAMIC NETWORK ANALYSIS..... | 30 |
| 1.3.3. SUMMARY | 31 |
| 2. OBJECTIVES..... | 33 |
| 2.1. OBJECTIVE 1: DEVELOPMENT AND APPLICATION OF STATISTICAL APPROACHES | 36 |
| 2.2. OBJECTIVE 2: DEVELOPMENT AND APPLICATION OF SYSTEMS BIOLOGY APPROACHES | 36 |
| 3. THESIS PUBLICATIONS | 39 |
| 3.1. A NOVEL MULTILEVEL STATISTICAL METHOD FOR THE STUDY OF THE RELATIONSHIPS BETWEEN MULTIRECEPTORIAL BINDING AFFINITY PROFILES AND IN VIVO ENDPOINTS | 41 |
| 3.2. PATHWAY DATABASES AND TOOLS FOR THEIR EXPLOITATION: BENEFITS, CURRENT LIMITATIONS AND CHALLENGES | 53 |

| | |
|--|-----|
| 3.3. NETWORK ANALYSIS OF AN INTEGRATED GENE-DISEASE ASSOCIATION DATABASE REVEALS FUNCTIONAL MODULES IN MENDELIAN, COMPLEX AND ENVIRONMENTAL DISEASES | 69 |
| 3.4. DEVELOPMENT OF WEBSERVICES AND WORKFLOWS FOR SUBSTANTIATION OF DRUG SAFETY SIGNALS | 111 |
| 3.5. DISGeNET: A CYTOSCAPE PLUGIN TO VISUALIZE, INTEGRATE, SEARCH AND ANALYZE GENE-DISEASE NETWORKS | 125 |
| 3.6. FROM SNPS TO PATHWAYS: INTEGRATION OF FUNCTIONAL EFFECT OF SEQUENCE VARIATIONS ON MODELS OF CELL SIGNALLING PATHWAYS .. | 161 |
| 4. DISCUSSION | 185 |
| 4.1. MULTIVARIATE STATISTICAL APPROACHES TO STUDY ADVERSE DRUG REACTIONS | 187 |
| 4.2. SYSTEMS BIOLOGY APPROACHES TO INVESTIGATE THE MOLECULAR BASIS OF DISEASES AND ADVERSE DRUG REACTIONS.. | 188 |
| 4.3. SUMMARY AND OUTLOOK | 194 |
| 5. CONCLUSIONS | 195 |
| 6. LIST OF PUBLICATIONS | 199 |
| ARTICLES | 201 |
| ORAL COMMUNICATIONS..... | 202 |
| POSTER COMMUNICATIONS | 203 |
| 7. REFERENCES | 205 |

1. INTRODUCTION

INTRODUCTION

1.1. Typical problems in biomedical research

Ever since, people have been suffering from diseases and have been trying to unravel and understand the underlying mechanisms in order to find effective remedies. Herbs that are for instance traditional in Chinese medicine, and later chemical substances or drugs have been used to prevent and cure the diseases. Despite the success of drug therapies, several diseases cannot be effectively treated yet. Additionally, many of the prescribed drugs and also herbs are not very efficient and cause undesirable side effects, which in severe cases can even lead to the withdrawal of the drug.

In 2001, the first draft of the human genome was published (Lander *et al*, 2001; Venter *et al*, 2001) and it was widely believed that this would revolutionize medicine. In the following years, the rapid maturation of low-cost high-throughput technologies for genotyping made possible genetic association studies which allowed to determine genetic variants associated to a large variety of common human diseases including heart disease (Helgadottir *et al*; McPherson *et al*, 2007), obesity (Herbert *et al*, 2006) and type I and II diabetes (Grant *et al*, 2006; Sladek *et al*). Nowadays, with the next-generation sequencing techniques, we are even able to sequence an individual's genome in few weeks at few costs.

However, most of the gained information has not been translated into clinical practice yet. While for some diseases the underlying mechanisms are known and have led to effective therapies for instance through drugs, for some others this is not the case. This can be explained as many common human diseases originate from complex interactions between genetic variations and environmental factors such as diet, age, sex and exposure to environmental toxins. Hence, for many diseases the underlying mechanisms are very complex and not completely understood. Consequently, we are still not able to fully treat common human diseases such as cardiovascular disorders, mental diseases or cancers. Moreover, there is a dramatically high number of drug candidates failing in clinical development due to lack of efficacy or because they cause severe adverse effects. In many cases such adverse effects only appear in some patients. Hence, even if the mechanism of the drug action seems to be understood, there are other factors such as genetic variants and environmental influences, which are responsible for individual drug response.

Human diseases and drug adverse reactions pose significant social and financial burdens to the public. Thus, the study of the underlying mechanisms is a major concern for biomedical research and therefore focus of this thesis. It is greatly hoped that success in this area will further support the vision of a personalized medicine being predictive and preventive.

1.1.1. Human disease mechanisms

For many years, scientists have been trying to understand the molecular and physiopathological mechanisms of diseases in order to design new preventive and therapeutic strategies. In the last decades the improvement of high-throughput technologies led to the generation of massive amounts of large-scale and high-dimensional molecular and physiological data. In this context, the combination of

INTRODUCTION

experimental and computational methods resulted in the discovery of disease-related genes, which were then studied to gain better understanding of disease mechanism (Botstein and Risch, 2003; Kann, 2010). A well-known example is Phenylketonuria, where first an association between the mutant PAH genotype and the disease was found, and then the function of the encoded PAH enzyme was studied with respect to the mechanism leading to the disease (Scriver and Waters, 1999). Another more recent example is the association of mutations of the BRCA1, respectively BRCA2 gene with an increased risk of breast and ovarian cancer, which had important implications for diagnosis and prognosis of these diseases (Futreal *et al*, 1994; Miki *et al*, 1994).

Despite some success, we are still far from understanding the molecular mechanisms underlying many human diseases. Although there are successful treatments for many human diseases we are still not able to cure or prevent common diseases, including asthma, cardiovascular diseases, mental disorders and cancer. This can mainly be explained by the complexity of the underlying mechanisms that is caused by several factors.

First, it has to be kept in mind that the associations found between genes and diseases are of statistical nature and do not necessarily mean that the found genes cause the disease (see figure 1A). Hence, the mere association between a genetic variant and a disease is usually not sufficient to explain disease development (Kann, 2010). It is therefore required to determine the exact role of the genetic variant in the disease mechanism. Second, DNA variations on their own do not lead to disease but affect molecular traits, such as the function of the proteins the genes are encoding, and these changes in turn affect the disease risk (see figure 1B). Finally, the biological context in which disease-related genes operate needs to be considered (see figure 1C). Many human diseases cannot be attributed to malfunction of single genes but arise due to complex interactions among multiple genetic variants (Hirschhorn and Daly, 2005). Ultimately, the influence of external variables such as environmental factors, infectious agents or drugs has to be considered when studying the occurrence and evolution of a disease. Hence, overall the mechanism underlying human diseases can be extremely complex. Even for mendelian diseases, such as Phenylketonuria, the underlying mechanisms are not fully understood. For instance, even if the mutant PAH genotype is known to be associated with the disease and also the role of the encoded PAH enzyme is well studied, there are many patients with the same mutation but greatly differing phenotype. Hence, phenotypic outcome cannot be predicted solely based on the genotype (Scriver *et al*, 1999). As a consequence, the interactions of genetic and environmental factors have to be studied in order to understand the molecular mechanism of disease development (Schadt, 2009). Such interactions can then be used to generate predictive models in order to study which molecular states drive disease development. This is furthermore of particular value for drug discovery, where the molecular states associated with the disease could be targeted in order to prevent or treat the disease. The identification and subsequent analysis of these molecular, disease-related states can have further implications, such as for the discovery of biomarkers that can be used for early detection and diagnosis of

diseases, or the prediction of clinical outcome and response to treatment (Chin and Gray, 2008). A well-known example for early detection and diagnosis is the aforementioned association of BRCA1/2 mutations with breast cancer. As a result, women with family breast cancer incidence can undergo genetic testing for BRCA1/2 mutations and hence can pursue stronger cancer surveillance and prevention regimens if tested positive or avoid unnecessary interventions if tested negative (Armstrong *et al*, 2000). Another example is where the genetic variation was found to be associated with individual drug response. For instance, it has become clinical practice to genotype patients with non-small cell lung cancer for mutations in the gene encoding the epidermal growth factor receptor (EGFR), in order to aid decision making about which drug to use for treatment (Sharma *et al*, 2007). Both examples can be seen as first attempts of personalized or patient-specific medication being one of the most promising and pushing fields in modern medicine.

In the last years we have made important improvements in understanding disease mechanisms and we have accumulated tremendous amounts of genomic data. Many findings, however, have not been translated into clinical practice, yet. Hence, to the best of our current knowledge, the mere associations between genomic information and diseases, although necessary, are not sufficient for understanding complex disease mechanisms. Besides, even the knowledge of a specific gene making substantial contribution to a disease phenotype can often not be directly translated to effective treatment due to complex downstream interactions in transcriptional, translational and posttranslational processes. Additionally, the influence of environmental factors on human health has to be considered when studying disease development and progression. An example is arsenic being a well-established human carcinogen. Accordingly, many studies support an association between arsenic exposure and increased incidence of solid tumours, such as lung, bladder, prostate, renal and skin tumours (Celik *et al*, 2008; Chiou *et al*, 1995; Radosavljević and Jakovljević, 2008; Smith *et al*, 1992; Tsuda *et al*, 1995; Yang *et al*, 2008). Studies conducted in developing countries show a general increase in the incidence of different types of cancers, which is hypothesised to be associated with exposure to environmental toxins among other factors, some of them of genetic origin (Park *et al*, 2008; Sankaranarayanan and Boffetta, 2010; Thun *et al*, 2010). Thus, there is a strong need to investigate the interactions among environmental carcinogens and genetic factors (Sankaranarayanan *et al*, 2010). Moreover, for many diseases, dysfunction of whole pathways or functional modules of interacting genes plays an important role in disease aetiology. For instance, (Lim *et al*, 2006) showed that a set of functionally related proteins is relevant to several forms of human ataxias, and (Jones *et al*, 2008) discovered a set of core signalling pathways, which are genetically altered in pancreatic cancers. Hence, in order to understand disease mechanisms, a network of the key players related to the disease and their interactions, for instance through biological pathways, has to be considered. These disease-related molecular networks can then be studied with respect to genetic and environmental perturbations and how they affect the disease risk (see figure 1C).

INTRODUCTION

At the end of the day, in order to achieve such understanding of disease mechanisms, the entire body of knowledge about genes, their association to diseases, their interactions, influence of environmental variables, etc. has to be taken into account. To achieve this, integrative data analysis approaches are required that cope with the heterogeneity and amount of biological data and are able to produce the complete picture. This picture can eventually be used to predict disease progression, treatment outcome and therefore aid development of safer and more efficient drugs.

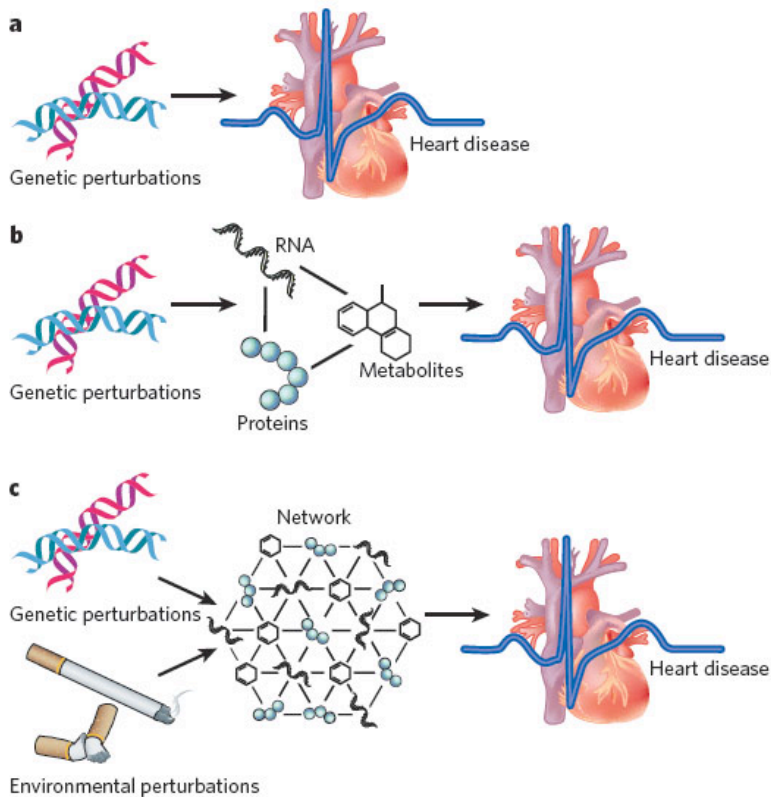


Figure 1: Causal gene-disease relationships, extracted from (Schadt, 2009).

- (A) Classical approach where DNA variation correlates directly with disease
- (B) DNA variation leads to modified molecular traits that in turn affect disease risk
- (C) More realistic view where DNA variation and environmental perturbations affect molecular states of networks that in turn affect disease risk

1.1.2. Adverse drug reaction mechanisms

It is widely accepted that any substance that has a therapeutic effect can also produce an undesirable adverse reaction (Edwards and Aronson, 2000). Not surprisingly, such unwanted effects or adverse drug reactions (ADRs) caused by drugs are a major clinical problem. Depending on severity, which ranges from mild to life threatening, they can cause immediate withdrawals of drugs and hence are significant financial burden to health care and pharmaceutical industry (Wilke *et al*, 2007).

The reasons for ADRs are diverse and include inappropriate use or administration of the drug, nonselective action of the drug and genetic predisposition. In this context, genetic variants have gained more and more attention since the 1950/60s. Genetic variants can determine susceptibility to ADRs by affecting both, pharmacokinetics (absorption, distribution, metabolism and excretion of the drug), and pharmacodynamics (mechanisms of the drug such as receptor binding or signalling) (Pirmohamed and Park, 2001; Wilke *et al*, 2007). In addition, drug-drug interactions and environmental factors can play an important role.

Typically, ADRs are divided into type A, being predictable from the pharmacology of the drug, and type B, being bizarre and unpredictable (Pirmohamed *et al*, 2001). Figure 2 shows a schematic representation of possible mechanisms of drug adverse events, which are explained in more detail the next section.

In general, mechanisms of ADRs can be grouped into the following, sometimes overlapping, categories related to: (i) on-target pharmacology (ii) off-target pharmacology, (iii) immunological reactions, (iv) biological activation to toxic metabolites, and (v) idiosyncratic toxicities, which are specific to an individual and usually difficult to predict or explain.

A drug or its metabolites can react directly with an intended (target) or an unintended (off-target) receptor, which may cause adverse reactions related to on-target or off-target pharmacology.

On-target effects are highly predictable as they are related to the mechanism of the drug and typically arise due to improper activation or inhibition of the intended target, for instance due to expression of the target in an another tissue not considered during drug development (Taniguchi *et al*, 2007).

Off-target effects are common since most drugs are not selective and interact with more than one target. For instance, a growing body of evidence confirms that the therapeutic effects of antipsychotic drugs are complex and cannot be ascribed to a single receptor. Hence, the traditional concept of a single receptor must be expanded to a whole set of biomolecules that are putatively involved in the pharmacological effect of the drugs (Roth *et al*, 2004). Consequently, activation or inhibition of unexpected receptors can lead to unwanted ADRs.

Furthermore, adverse reactions are possible due to impaired drug metabolism during detoxication and excretion processes. In this respect, genetic variations are a key determinant of drug metabolism phenotypes and individual drug response

INTRODUCTION

(Liebler and Guengerich, 2005). An interesting example highlighting the influence of genetic variation on drug response is CYP2D6. CYP2D6, which is responsible for the metabolism of many drugs is also prone to genetic variations causing loss of function, decreased activity of the enzyme, altered substrate specificity or increased activity (Pirmohamed *et al*, 2001). These effects in turn lead to decreased metabolism and elimination, accumulation of the drug or re-routing of the metabolism. For instance, the anticoagulant drug Warfarin, shows major risk for haemorrhage and could be made safer by considering the patients genotype for CYP2D6 and VKORC1, another gene related to Warfarin resistance (Gurwitz and Motulsky, 2007). Furthermore, mechanisms of genetically induced drug effects can be polygenic (as in the case for Warfarin) and also depend on environmental factors such as diet, exposure to xenobiotics, smoking or alcohol (Gurwitz *et al*, 2007). Hence, complexity is an issue when studying the mechanisms underlying ADRs.

Another possible mechanism of ADRs is related to the binding of drug metabolites with nucleophiles in the cell, such as DNA, proteins or small molecules. These drug-nucleophile aggregates can trigger regulatory processes that lead to inflammation, apoptosis and necrosis. Reactions involving DNA can furthermore lead to impaired DNA repair or DNA mutations, which in turn may cause carcinogenesis.

Immunological reactions are very common, usually unpredictable and lead to hypersensitivity or autoimmune responses of the body. They range from mild skin rashes to immune-mediated organ failure. Moreover, there is evidence that idiosyncratic drug reactions are mainly immune-related (Uetrecht, 2007).

In the last years, idiosyncratic drug reactions have gained attention, as they are least understood and least predictable and pursuant to their definition difficult to reproduce in human populations (Liebler *et al*, 2005). Nowadays, idiosyncratic drug reactions are mostly responsible for drug withdrawal and therefore major factor contributing to cost and uncertainty in drug development. As already mentioned, many idiosyncratic drug reactions are immune-related or appear due to reactive metabolites. Moreover, the identification of genetic variants that are strongly associated with idiosyncratic drug reactions is promising to predict the risk of such ADRs.

In summary, in order to predict and prevent ADRs, a deep comprehension of the molecular mechanisms of disease, drug action as well as the influence of genetic variants and environmental factors on the human body is required.

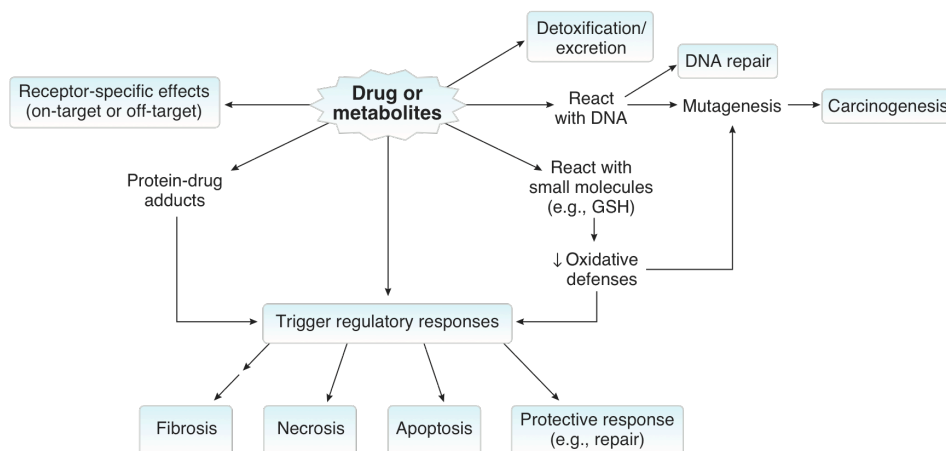


Figure 2: Mechanism of adverse drug reaction, taken from (Taniguchi et al, 2007)

A drug or its metabolite interacts with a specific receptor to mediate on-target or off-target adverse effects (upper left side). In addition, metabolites can be detoxified and excreted, here adverse effects can occur, for instance due to genetic variants in metabolizing enzymes leading to impaired drug metabolism. Moreover, metabolites can react with nucleophiles including DNA, proteins and small molecules. The formation of unrepaired or mispaired DNA adducts is often mutagenic and may lead to carcinogenesis (upper right side). The impairment of oxidative defence can lead to inflammation and cell death (apoptosis or necrosis, lower part in the middle). The formation of drug-protein adducts can trigger immune responses from protective to apoptosis and necrosis can result. Chronic inflammation and repair can also lead to tissue fibrosis (lower left side).

1.1.3. Summary

Despite some break-through discoveries in the last century, including the sequencing of the whole human genome, we are still far from having a complete understanding of the molecular processes involved in diseases. We are therefore not able to prevent and cure common diseases such as asthma, Alzheimer or cancer. Ever since, people have been trying to find remedies, and as a consequence drugs are prescribed to prevent or treat the diseases. However, in some patients the use of drugs leads to the development of undesired side effects. These adverse drug reactions remain a major clinical problem and have led to the withdrawal of a variety of drugs.

Human diseases as well as adverse drug reactions impose major burdens, including financial costs, on both patients and society, and are therefore major problems to be addressed by biomedical research. The pharmaceutical industry and public institutes spend significant amounts of money and effort on the development of new technologies, analysis methods and computational approaches to study the molecular mechanisms underlying complex human traits. However, still most

INTRODUCTION

available drugs are not as effective and safe as they should be. For instance, the efficacy rate of drugs within major disease areas such as asthma, cancer, psychiatric illnesses and cardiovascular diseases are in the range of 25–60 % (Jørgensen and Winther, 2009). A more profound understanding of human disease mechanisms would clearly lead to a better understanding of drug action and would therefore benefit current drug discovery in order to develop more efficient and safer drugs. Such understanding could moreover pave the way for a personalized medicine being predictive and preventive and will therefore benefit our society in a variety of ways.

1.2. Need of data integration in biomedical research

Within the recent years, the development and improvement of high-throughput technologies for DNA sequencing and analysis of transcriptomes, proteomes and metabolomes have led to an explosion of “omics” data. In contrast to the classical hypothesis-driven research, where data is gathered in a focused manner to answer a specific question (Searls, 2005), this accumulation of large-scale data allowed the evolvement of data-driven research, which aims at gathering extensive data and making it available for sampling and interpretation. At the same time, technical innovations in computer science yield to construction of advanced repositories for storage and novel computational methods for analysis of large-scale data.

Nowadays, biomedical data is typically stored in online databases that are automatically as well as manually populated. In most of the cases, experts of the field curate the information manually from the literature or directly from experiments and annotate the data with additional information. This process not only ensures high quality of the data but also facilitates data integration, for instance by means of incorporation of cross-links to further databases by the curators. However, such curation process is very work-intensive and time-consuming, and requires considerable expertise. Moreover, much of the information is still not available in databases but only as free text in published articles. Clearly, this unstructured form of data hinders automatic retrieval tremendously. In this regard, text-mining has evolved as a useful tool not only to extract the information that is locked in the literature but also to make the task of curation easier and less time-consuming (Zweigenbaum *et al*, 2007). In this manner, the field of biocuration has evolved very recently to make biomedical information accessible to both humans and computers (Howe *et al*, 2008).

Biomedical data is typically fragmented, distributed over various databases and partly locked in the literature. This poses difficulties to the individual scientist to acquire, validate and analyze the data. In some cases, information even seems to be hidden since it is stored in databases or part of articles that are not accessible to or simply not known by the researcher. Conversely, large-scale data properties can only be revealed considering the whole information available (Cokol *et al*, 2005).

INTRODUCTION

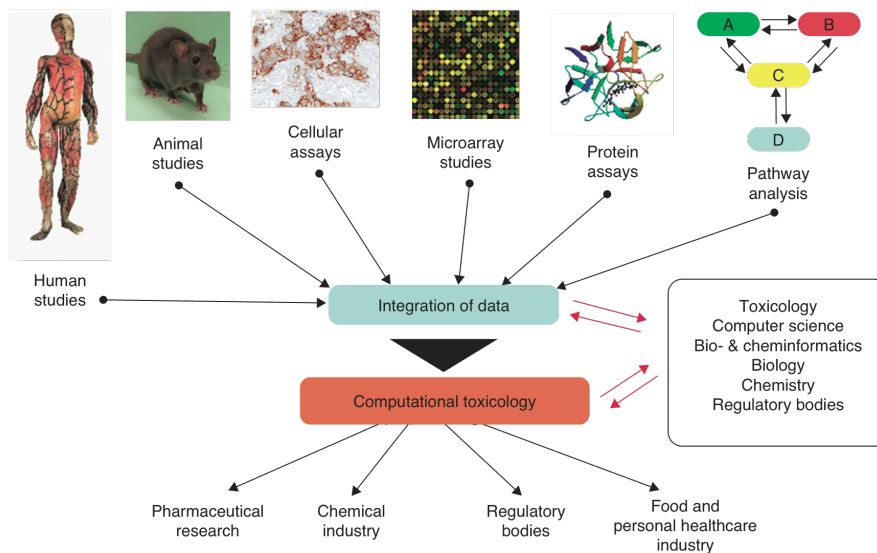


Figure 3: Data integration in biomedical research, taken from (Nigsch *et al*, 2009)

Different kinds of data need to be integrated in biomedical research in order to shed light on the molecular mechanisms underlying complex human traits. Then, computational approaches are required to interpret the data and propose hypotheses. Finally, the mere data is converted into knowledge and aids further decision-making for instance regarding drug development or drug safety regulation.

Another peculiarity of biomedical data is that properties differ immensely depending on the type of technique to produce the data and on the type of biomedical entity the data describes. An illustrative example for this is given by studies on drug-toxicity, where diverse experiments are carried out, accumulating data of completely different types (see figure 3). Among them are categorical values derived from clinical surveys on the patient's well being, for example. Further data sources might be microarray experiments producing continuous values of gene expression levels measured for all genes before and after drug use. Here, the genes are the biomedical entities while the expression values represent their attributes. Eventually, the study might include information about the interactions between the biomedical entities involved, for example through biological pathways. Though difficult, all the different kind of data could be combined by means of computational approaches in order to give valuable clues on the mechanisms underlying drug toxicity (see figure 3).

All in all, data integration and analysis approaches are required to collect and combine biomedical data in a meaningful way and to subsequently produce novel

knowledge from the integrated data. Today, it is widely recognized that the challenge of combining biomedical data is not due to the sheer data quantity but its diversity (Searls, 2005). In this regard, bioinformatics and computational approaches have evolved to cope with the unprecedented wealth of diverse, complex and distributed information (Goble and Stevens, 2008).

This section depicts some of the major current sources of biomedical information, including online-databases and literature accessed through text-mining (section 1.2.1). Furthermore, it describes some data standards that have been developed to ease automatic data exchange (section 1.2.2). In the last two sections (sections 1.2.3 and 1.2.4), some approaches for data integration are discussed.

1.2.1. Biomedical data sources

The conjunction of the extraordinary amount of information generated by high-throughput methods and the improved capabilities of computer hardware resulted in a remarkable increase of data repositories. In 2001, the Nucleic Acid Research journal database supplement listed only 96 databases covering different aspects of molecular and cell biology while this year, in 2010, there are 1230 recorded (Cochrane and Galperin, 2010).

Moreover, there is a lot of information available in the literature and the amount of publications is dramatically increasing, it has been demonstrated that the number of publications in PubMed/MEDLINE grows exponentially (Hunter and Cohen, 2006). Hence it is difficult for individual scientists to keep up with the relevant publications in their own discipline. In this regard, text-mining methods are required to extract relevant information automatically from text.

In this section, some of the major sources, including databases and literature, are described that are important for biomedical research, especially for investigating the molecular mechanism underlying human traits and that were used throughout this thesis (see table 1). While some of the presented sources cover more general information about biomedical entities, some others focus on a specific matter.

Table 1: Data sources for biomedical research

| Data source | Link | Short description |
|--------------------|---|---|
| Entrez Gene | http://www.ncbi.nlm.nih.gov/entrez | General information about genes |
| UniProt | http://www.uniprot.org/ | General information about proteins |
| OMIM | http://www.ncbi.nlm.nih.gov/omim | General information about genetic diseases and their associated genes (links directly to Entrez Gene) |
| PharmGKB | http://www.pharmgkb.org/ | Information about pharmacogenes and associations to diseases or drug adverse events |
| CTD | http://ctd.mdibl.org/ | Information about the effects of chemicals on human health, also about the underlying genes |
| Reactome | http://www.reactome.org/ | Information about human cell signalling pathways |
| KEGG | http://www.genome.jp/kegg/ | Conjunction of databases containing, gene, drug, pathway information, among others |
| NCI PID | http://pid.nci.nih.gov/ | Information about human cell signalling pathways |
| Pathway Commons | http://www.pathwaycommons.org/pc/ | Combination of Reactome, NCI PID and Cancer Cell Map |
| WikiPathways | http://www.wikipathways.org/ | Open community pathway database |
| dbSNP | http://www.ncbi.nlm.nih.gov/projects/SNP/ | Broad collection of simple genetic polymorphisms |
| MEDLINE | http://www.ncbi.nlm.nih.gov/pubmed/ | Online database of biomedical journal citations and abstracts (accessible through PubMed) |
| UMLS | http://www.nlm.nih.gov/research/umls/ | Compendium of a variety of biomedical terminology systems including different disease vocabularies |

1.2.1.1. Entrez Gene

NCBI Entrez Gene is a database for gene-specific information including sequence, chromosomal localization, gene products, associated markers, and phenotypes. Moreover, it provides links to citations, related sequences, variation, maps, expression, homologs, protein domain content and external databases (Maglott *et al*, 2006). Entrez Gene identifiers (GeneIDs) are species-specific and hence unique gene identifiers. Most biomedical data sources provide cross-links to Entrez Gene such as GO, KEGG, Reactome or UniProt, which are discussed in the following. Thus, Entrez Gene can be considered as primary source for gene information.

1.2.1.2. UniProt

The Universal Protein Resource (UniProt) is a centralized resource for protein sequences and functional information. It unites Swiss-Prot, TrEMBL and PIR protein database activities and creates three layers of protein sequence databases: the UniProt Archive (UniParc), the UniProt Knowledgebase (UniProt) and the UniProt Reference (UniRef) databases. This section only discusses the UniProt Knowledgebase, a comprehensive database storing information on protein sequence, structure and function (Apweiler *et al*, 2004; The UniProt, 2010). It consists of two sections, UniProt/Swiss-Prot and UniProt/TrEMBL, where the first contains fully manually curated entries, and the latter automatically annotated information. The manual curation process of Swiss-Prot involves extensive cross-referencing to other biomedical databases, functional and feature annotation as well as annotation to evidence found in literature. Hence, UniProt/SwissProt serves as a reliable source of protein information and is a good starting point for studies involving proteins.

UniProt/SwissProt provides curated information on the functional and phenotypic effects of natural variations, including SNPs, as well as on mutations of protein sequences. For several of these natural variants and mutants, it furthermore provides associations to disease phenotypes. Thus, it provides a comprehensive framework to extract information about the association of sequence variations and human diseases.

1.2.1.3. Gene-disease association databases

Aim of this thesis is to investigate the molecular mechanisms underlying complex human traits including diseases and adverse drug reactions. In this regard, it is of importance to connect genotypic with phenotypic information. Hence, in this section specific sources of gene-disease associations are briefly introduced.

In the 60s, Dr. McKusick started collecting information about genes and their association to diseases first as a book and later as online database. His Online Mendelian Inheritance in Man (OMIM) database has become a highly popular source in medical genetics (Hamosh *et al*, 2005). OMIM traditionally focused on monogenic diseases and later started to include complex diseases as well. Next to providing a summary of the clinical features of the disorder, it links to numerous

INTRODUCTION

databases, such as DNA and protein sequence, PubMed and mutation databases, among others.

In the last years, further databases with different focus have been built. Among them PharmGKB has evolved, a database specialized on the knowledge about genes that are involved in modulating drug response (pharmacogenes) (Klein *et al*, 2001). In general, genes are classified as pharmacogenes because they are involved in the pharmacokinetics of a drug (how is the drug absorbed, distributed, metabolized and eliminated), or the pharmacodynamics of a drug (how does the drug act on its target and what are the downstream effects) (Altman, 2007).

Another important source is the Comparative Toxicogenomics Database (CTD), which contains manually curated information about gene-disease relationships with focus on understanding the effects of environmental chemicals on human health (Mattingly *et al*, 2006). It is, similar to the other databases, highly cross-linked with other biomedical databases.

Moreover, as mentioned above, UniProt/SwissProt not only contains curated information about protein sequence, structure and function but also provides information on the functional effect of sequence variants and their association to disease.

1.2.1.4. Pathway databases

A biological pathway can circumscribe several types of biological processes including regulatory, metabolic and signalling processes or protein-protein interactions. Throughout this thesis, we will use the term pathway for regulatory, metabolic or signalling processes but not referring to protein-protein interactions. Currently, there exist a vast variety of databases containing information about pathways or protein-protein interactions. The Pathguide resource serves as a good overview of these databases (Bader *et al*, 2006). More than 200 pathway repositories are listed, from which over 60 are specialized on reactions in human.

In this section some of the major state-of-the-art pathway databases are discussed, namely Reactome, KEGG, WikiPathways, the Nature Pathway Interaction Database (PID) and Pathway Commons. They present pathways in a graphical format comparable to the representation in text books, as well as in standard formats allowing exchange between different software platforms and further processing by network analysis, visualization and modelling tools

Reactome is currently one of the most complete and best curated pathway databases (Joshi-Tope *et al*, 2005). It covers reactions for any type of biological process, including metabolic, regulatory and cell signalling pathways, and organizes them in a hierarchal manner. Expert biologists curate all pathways and reactions from biomedical literature or experiments (Joshi-Tope *et al*, 2005).

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is not only a database for pathways but consists of 19 highly interconnected databases, containing genomic, chemical and phenotypic information (Kanehisa *et al*, 2008; Kanehisa and Goto, 2000). KEGG categorizes its pathways into metabolic processes, genetic information processing, environmental information processing including signalling

pathways, cellular processes, information on human diseases and drug development. However, the best-organized and most complete information can be found for metabolic pathways. KEGG is not organism specific but covers a wide range of organisms including human.

A recently developed resource for pathway information that strongly differs from other pathway repositories is WikiPathways. WikiPathways is an open source project based, like Wikipedia, on the MediaWiki software (Adriaens *et al*, 2008). It serves as an open and collaborative platform for creation, edition and curation of biological pathways in different species. WikiPathways aims to achieve a public commitment to pathway storage and curation by keeping pathway creation and curation processes simple. Any user with an account on WikiPathways can create new pathways, and edit already existing ones. WikiPathways does not use standard formats like BioPAX (<http://biopax.org>) or SBML (Hucka *et al*, 2003) but offers a much simpler representation called GenMAPP Pathway Markup Language (GPML). Hence, interoperability with other pathway databases is impeded, and substantial efforts towards combining WikiPathways with the other pathway repositories are required.

The Pathway Interaction Database (PID) contains data on cell signalling for human (Schaefer, 2006). It combines three different sources, the NCI-curated pathways that are obtained from peer-reviewed literature, as well as pathways imported from Reactome and BioCarta. Similar to Reactome, PID structures pathways hierarchically into pathways and its sub-pathways.

Pathway Commons is a compilation of the public pathway databases Reactome, PID and Cancer Cell Map as well as protein-protein interaction databases such as HPRD (Mishra *et al*, 2006), HumanCyc, IntAct (Kerrien *et al*, 2007) and MINT (Zanzoni *et al*, 2002). Herein, the pathway hierarchies of Reactome and PID are conserved. Hence, it serves as an access point for a collection of public databases and provides technology for integrating pathway information. Pathway creation, extension and curation remain duty of the source pathway databases.

Most of the here presented pathway databases allow querying and browsing their data through a web interface but also provide programmatic access through webservices or APIs. They contain manually curated information and provide cross-references to other biomedical databases such as Entrez Gene or UniProt and annotate entities and processes with GO terms. Furthermore they mostly support current pathway exchange formats such as BioPAX and SBML, which is not the case for many other pathway databases. However, there is overlap in the information available; for specific pathways some databases offer more accurate and complete information than others. (Adriaens *et al*, 2008) describe a workflow developed for gathering and curating all information on a pathway in order to obtain a broad and correct representation. Nonetheless, the described process heavily relies on manual intervention and is very work-intensive and time-consuming. Consequently, there is a need for automation of both the pathway retrieval process and the integration of different data sources.

INTRODUCTION

In the last years, several tools for pathway visualization and analysis have been developed and the aforementioned standard formats ease import of pathways into these tools (Suderman and Hallett, 2007). One of the most known and widely used open-source software is Cytoscape. It offers, next to pathway visualization, a vast variety of analysis tools ranging from advanced network analysis to webservices (Shannon *et al*, 2003), for instance to retrieve pathways from the here described pathway repositories.

The databases presented above allow the access to a wide range of data on biological pathways. However, the data is fragmented and the representation of the biochemical reactions differs between databases, as well as coverage and accuracy of annotations. In addition, often data is not provided in interchangeable formats hampering its automatic integration. Nevertheless, integration is needed to obtain a complete view of the biological process of interest. This is of particular interest for studying the molecular mechanisms underlying human diseases and adverse drug reactions, aim of this thesis. Currently, an automatic integration of data from the diverse repositories is difficult due the aforementioned limitations concerning missing possibility to automatically access the data, lack of annotation and misuse of standard formats. As a consequence, development of improved tools for data integration but also more accurate data curation and annotation are needed.

1.2.1.5. Text-mining derived information

Due to the vast increase of published literature in health and life sciences, it is not possible (not even for expert curators of the aforementioned databases) to keep track of the relevant knowledge that is regularly published (Cokol *et al*, 2005). In this regard, text-mining has evolved as a useful tool to automatically extract information about biomedical entities and their relationships reported in the literature (Ananiadou *et al*, 2010; Jensen *et al*, 2006). Clearly, this task is very difficult since the extraction of information from text typically requires a human understanding of the text, which also takes into account background knowledge to make inferences (Ananiadou *et al*, 2010). Goal of text-mining is to convert unstructured into structured information to aid knowledge discovery and hypothesis making.

In biomedical research, we not only study the biomedical entities themselves, such as genes or proteins, but also how they interact. In this regard, automatic information extraction from text has to address the identification of the biomedical entities and their relationships in free text.

Text-mining has already been successfully used to extract biomedical entities from literature. This process contains two steps, first the synonyms are located in the text (named entity recognition) (Ananiadou *et al*, 2004) and second they are mapped to unique and standard identifiers of curated databases (normalization) (Cohen *et al*, 2008), such as UniProt or Entrez Gene. Some examples are OSIRIS, a system for the identification of DNA variation terms in text and their mapping to dbSNP identifiers (Furlong *et al*, 2008), ProMiner for the recognition of gene and protein mentions (Hanisch *et al*, 2005) and Peregrine, which is based on

dictionaries and can thus be used for the recognition of any biomedical entity, such as small molecules and drugs, genes or proteins (Schuemie *et al*, 2007).

The difficult task of relationship extraction has gained more attention recently, since we have moved from classical molecular biology focusing on single entities to Systems biology studying their interactions (Ananiadou *et al*, 2010). Some examples for the extraction of different kind of associations between biomedical entities are systems for extraction of protein-protein interactions (Donaldson *et al*, 2003; Fundel *et al*, 2007) or gene-disease associations (Bundschuh *et al*, 2008). Text-mining has also been applied to link pathways to literature evidence and even to construct biological reactions and pathways from literature (Oda *et al*, 2008).

As for any other computational and experimental method, data quality and reliability are an issue. Typically, results are compared to gold standards that are prepared by domain experts for the training and testing of text mining applications. Then, the quality of the system in comparison to the gold standard is defined by measurements such as recall and precision or the F-Score, a combination of recall and precision (Hersh, 2005). Moreover, recently the BioCreative challenge has established as a community-wide effort for the evaluation of text-mining and information extraction systems applied to the biological domain (Krallinger *et al*, 2008). This challenge poses tasks to the community for systems to be developed that are not only useful to general researchers but also for specific end users such as database curators.

In summary, text-mining has important influence on biomedical research as it allows the automatic extraction of biomedical entities and their relationships from free text. Moreover, it is of major importance for database curation since the vast amount of information published in the biomedical domain impedes to keep track manually.

1.2.2. Data heterogeneity and standards

Nowadays, main challenges of data integration are not related to data quantity but data heterogeneity and the fact that data is spread among diverse often overlapping and redundant repositories. Some examples for data heterogeneity, hindering data integration, are differing data formats or the use of different vocabularies (Searls, 2005). For instance, the protein WS-1 has 21 different accession numbers and 10 distinct names (Goble *et al*, 2008). Multiple identifiers can be explained by rediscovery, assignment of new function, alternative transcription and post-translational modifications, among others. However, for proper data integration, shared and common identities and names are essential. Same identities and unambiguous naming of biomedical entities does not only allow individual researchers to easily combine information about the same entity from various sources but also facilitates automatic integration of information by means of computational tools. For instance, UniProt and KEGG both use their own identifiers to represent proteins. Common practice to solve the problem of diverse identifiers is the use of cross-links to connect between the same entities in the different databases, such as the case for UniProt and KEGG. Such cross-linking, which can

INTRODUCTION

be seen as a tool for data integration itself, is based on accurate identifier mapping and usually requires manual work by means of database curation. However, the advantages and disadvantages of a single global unique identity schema, which would in principle allow an unobstructed automatic data integration, are still discussed (Goble *et al*, 2008).

Moreover, there is a need for shared semantics or standards for data schemas and values. This is especially important as different conceptualisations and representations hinder the process of data integration immensely and urge manual intervention and interpretation by humans, yet again achieved by expensive data curation.

In the systems biology community, several standards for representation and exchange of pathway data have evolved. For instance, Systems Biology Markup language (SBML) (Hucka *et al*, 2003) is widely accepted standard to represent mathematical models of biochemical pathways. A database storing such models in SBML format is BioModels, a resource of quantitative models of biomedical interest, which furthermore curates the models by annotating entities with terms from controlled vocabularies and by adding cross-links to other data resources. (Le Novère *et al*, 2006).

Another important standard for the systems biology community is Biological Pathway Exchange (BioPAX) (<http://biopax.org>), a unified framework for pathway representation, which is supported by the major pathway databases such as Reactome and KEGG. Although some of the formats have established as standards, conversion between formats and misuse of formats are still issues to be solved. In this regard, some attempts have been made to define the minimal information required to represent a certain data type, such as MIAME for microarray data (Brazma *et al*, 2001) and MIRIAM for biochemical models (Le Novère *et al*, 2005).

Nevertheless, some limitations to data integration still exist, for instance due to incorrect use of standard identifiers and data formats or the lack of annotation of required attributes of biomedical entities.

1.2.3. Computational data integration approaches

Several computational approaches have been developed to cope with the vast amount of data that is spread over isolated and overlapping repositories. In principle, these approaches have to tackle two problems: first, to automatically access the data and second to merge it correctly.

Common approaches allowing programmatic access to data and tools are web services or the use of APIs. Although almost all biomedical databases provide such programmatic access to their data, real data integration is not provided and the automated process of data collection by means of bioinformatic scripts and tools is only facilitated. In this context, workflow approaches, in which data is gathered, integrated and even analysed in several, consecutive steps, have become popular. Software has been developed that allows manually creating such workflows without the need of programming skills. One example is Taverna, a tool to integrate

resources that are shared as web services into a scientific workflow to perform *in silico* experiments (Oinn *et al*, 2004).

Probably the simplest way of connecting biomedical entities is the aforementioned use of cross-links. Here, same entities with different identifiers are directly linked through hyperlinks on the web or by providing identifier-mapping tables. However, yet again, real data integration is only possible by means of bioinformatics tools making use of the available mapping.

Another approach, allowing true data integration, is data warehousing. Here, the data is first collected, systematically combined and stored in a central repository. Then, the data is regularly updated which causes high maintenance costs due to data and format changes in the original data sources (Goble *et al*, 2008). Prominent approaches in the biomedical domain include BioMART (Haider *et al*, 2009) and BioWarehouse (Lee *et al*, 2006), among others.

One of the most promising and more recent solutions for an automated information retrieval and data management system is the use of semantic web technologies. They provide the aforementioned warehouse capabilities on the fly. Hence, they do not move all data in a central repository but leave it at the original repositories and simply link it through unique URLs. One example is the Bio2RDF project, which aims at building a mashup combining major databases relevant for the biomedical and bioinformatics domain by making use of semantic web technologies (Belleau *et al*, 2008). However, the applicability of such approaches is still limited but remains promising (Antezana *et al*, 2009; Dumontier and Villanueva-Rosales, 2009).

1.2.4. Ontologies

Ultimate goal of data integration in biomedical research is not only to gather all data regarding a specific question but also to automate its analysis and to provide testable hypotheses (Slater *et al*, 2008).

However, in order to uncover the meaning of the gathered data, or in other words to convert mere data into knowledge, information has to be interpreted and put into context.

In this regard, ontologies have emerged providing vocabularies to describe biomedical data and it is hoped that such described data aids the automated analyses of the data eventually (Bodenreider and Stevens, 2006). An ontology represents explicit formal specifications of the concepts and their relationships in a domain (Gruber, 1993). Ontologies go well beyond controlled vocabularies by not only providing the same vocabulary for the entities but also by organizing them within classifications and hierarchies (Bodenreider *et al*, 2006). The basic idea is that given such well-defined structure automatic inference of knowledge about the entities and their relations is automatically possible.

With the advent of semantic web technologies RDF (Resource Description Framework) and OWL (Web Ontology Language) have evolved that are particularly well suited for representing bio-ontologies (Bodenreider *et al*, 2006). Some examples of bio-ontologies are introduced in the following.

INTRODUCTION

Probably the most used and known ontology in the biomedical domain is The Gene Ontology (GO) (Ashburner *et al*, 2000), which captures three major aspects about a gene product that are represented by the three branches Molecular Function, Biological Process and Cellular Component (Bodenreider *et al*, 2006).

Another example is the aforementioned BioPAX, a standard for pathway representation, which uses OWL to represent the pathways and is widely accepted by the pathway databases explained in section 1.2.1.4.

Even longer practice have ontology-like clinical and medical terminologies such as ICD (International Classification of Disease) (Nahler, 2009), SNOMED (Systematized Nomenclature of Medicine) (Cote and Robboy, 1980) and MeSH (Medical Subject Headings) (Sewell, 1964). In this regard, UMLS (Unified Medical Language System) was built to integrate these different disease vocabularies into one system (Bodenreider, 2004).

In summary, the use of ontologies in biomedicine has become a trend recently. Ontologies aim at describing entities and their relationships in a structured way to allow automatic analysis of the described data.

1.2.5. Summary

In the last years, the generation and analysis of genomic, transcriptomic, proteomic and other genome-scale data have become routine approaches of biomedical research. The parallel development and improvement of computational approaches allowed the building of various online databases storing the large-scale data produced by high-throughput technologies (see section 1.2.1). In parallel, great effort has been made to manually curate the information such that the available data serve as valid sources to study the molecular processes underlying human diseases and adverse drug reactions. Moreover, these improvements triggered the implementation of data standards and ontologies with the aim to allow automated data access, integration and subsequent analysis (see sections 1.2.2, 1.2.3 and 1.2.4).

Nevertheless, there are still major issues to be solved regarding annotation, lack of standards or their misuse, as well as lack of computational approaches for data integration and analysis. Also, data fragmentation still poses obstacles to our understanding of the molecular processes in the human body. This is especially visible considering the fact that individual researchers are often restricted to so called knowledge pockets (Cokol *et al*, 2005), which constitute only small fractions of the complete knowledge being available and are furthermore distributed over distinct databases or literature.

To some extent, text-mining approaches have helped to alleviate this problem of data fragmentation and are promising tools to aid database maintenance and curation. However, there is still an urgent need for improved integrative approaches to analyse the data in a meaningful way in order to create novel hypotheses and to eventually aid to understand the molecular mechanisms underlying complex human traits.

1.3. Integrative biomedical analysis approaches

While the previous section has discussed the need for data integration approaches in biomedical research and has introduced some methods for data integration, this section concentrates on approaches focusing on data analysis and therefore can be referred to as integrative biomedical data analysis approaches. In this regard, two major types of methodologies are introduced, statistical approaches that mainly aim at establishing associations between biomedical entities or phenomena and systems biology approaches, which study in more detail how biomedical entities interact and how perturbations affect the whole system.

1.3.1. Statistical approaches

As already discussed, one major challenge in biomedical research is to make sense out of the vast amount of available data. Also, it has been discussed that much of this data is redundant, cross-linked and therefore highly correlated. Hence, data reduction and data visualization are crucial to allow high-quality data analysis. In this regard, mathematical and statistical methods have been established for a long time, such as principal component analysis (PCA) or clustering (Howe *et al*, 2007). PCA, for instance, transforms a number of correlated variables into a smaller number of uncorrelated principal components such that the first principal component reflects the greatest variance of the data, the second principal component the second greatest variance, and so on. Thus, by creating new views of the data, it detects trends, groupings and outliers of the data. PCA has been used, for instance, to cluster gene expression data (Yeung and Ruzzo, 2001) or to cluster drugs according to their receptor binding profiles (Lange *et al*, 2007).

Next to methods that focus on analysing the structure underlying the data, biomedical research requires methods studying relationships between variables. In this regard, regression analyses are among the most flexible and most used approaches. These methods allow studying how a set of predictor (independent) variables influences some outcome (dependent) variables. In addition, a regression model can predict the outcome for new predictor values.

One such approach is logistic regression. It has wide application in health science and clinical studies (Harre *et al*, 1988) due to its ability to model dichotomous outcome (i.e. yes/no or healthy/disease) (Bagley *et al*, 2001). For instance, a logistic regression model was used for the early diagnosis of acute myocardial infarction (Kennedy *et al*, 1996).

Another related approach for multivariate data analysis is projections to latent structures by means of partial least squares (PLS) (Wold, 1982). PLS combines data projection, similar to the aforementioned PCA, and regression analysis and is particularly well suited when the number of variables describing some objects is much higher than the number of objects. The most simple and most used form of PLS is the PLS regression (PLSR). However PLS can be extended in various directions, for instance to perform discriminative analysis (PLS-DA), among others (Wold *et al*, 2001a). PLS is applied in various fields including multivariate calibration and quantitative structure-activity relationships (QSAR) (Wold *et al*,

INTRODUCTION

2001b). QSAR models are typically used to predict the activity of a chemical with respect to a specific biological target or to predict its likelihood to produce a specific toxic effect. For instance, it has been used to detect relationships between structural properties of a series of antipsychotic drug candidates and their binding affinities to G-protein-coupled receptors (GPCRs) (Dezi *et al*, 2007). Moreover, PLS has found various applications in the biomedical domain including the analysis of microarray data to predict clinical outcome (Pérez-Enciso and Tenenhaus, 2003). Examples are classification of different types of human tumours (Tan *et al*, 2004) and prediction of transcription factor activities (Boulesteix and Strimmer, 2005).

In summary, there is a wide range of methods being widely used to detect relationships between objects or between objects and their properties by means of data reduction, projection and regression analysis. Nevertheless, it has to be kept in mind that such found relationships represent statistical associations rather than causation. Therefore, a combination of these approaches with further analysis of the underlying mechanism is required in order to explain the found relationships. For instance, to mechanistically explain a found association between a drug and a specific side effect. However, these statistical methods represent meaningful tools to uncover relationships between different types of entities and to build predictive models, being able to cope with typical issues of biomedical data such as high dimensionality, high correlation between variables and missing data.

1.3.2. Systems biology approaches

Systems biology is an analytical approach to investigate the relationships among the components of a biological system in order to understand its emergent properties (Arrell *et al*, 2010).

In order to achieve such system level understanding, the biological system or network of study is systematically perturbed, the responses of the involved genes, proteins, and pathways are monitored, and ultimately mathematical or computational models are built that describe the underlying structure of the system and its response to individual perturbations (Ideker *et al*, 2001). This process is performed in an iterative manner, where experimental observations are matched against model predictions which in turn allows the formation of new models, new predictions, and new experiments to test them (Ideker *et al*, 2001). Figure 4 shows an overview of systems biology. The required combination of experimental and computational methods, as well as the iterative nature of the approach is visible. Moreover, the important role of network biology, which will be discussed in more detail in this section, is shown.

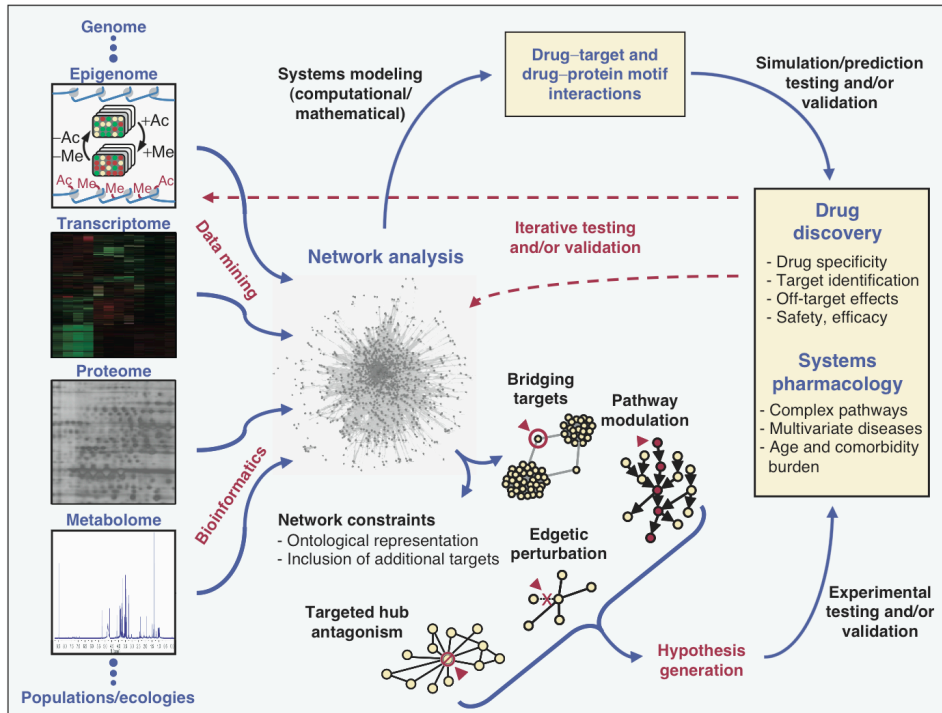


Figure 4: Systems Biology overview, taken from (Arrell and Terzic, 2010)

Systems Biology aims at understanding biological processes at a systems level. For this purpose different kinds of data are integrated. In this context, experimental and computational approaches are combined in an iterative manner. A network of interacting biomedical entities is perturbed, the responses of the involved genes, proteins and pathways are measured. A mathematical or computational model is built, which in turn provides hypotheses that can then be tested experimentally. This process is iterated until a model is built, which is able to predict the outcome of the experiments and hence can be used in drug discovery or systems pharmacology projects, for example.

INTRODUCTION

Systems biology approaches aim at understanding how a biological system is maintained, or in other words, how growth, development and adaptation are achieved. Moreover, it tries to explain and predict the influence of environmental or genetic perturbation on the system. This process is studied at different levels, ranging from genes, proteins, their interactions through biological pathways, organelles, cells, tissues up to the whole organism (see figure 5). While cell- and organ-scale models have a long history such as for glucose metabolism and homeostasis (Bergman *et al*, 1979; Kansal, 2004), the advent of high-throughput technologies led to the construction of models at lower levels, such as models of signalling pathways. For instance, the epidermal growth factor receptor (EGFR) signalling cascade is one of the best-studied and most important signalling pathways in mammals and is important for regulation of cell growth, proliferation and differentiation. Therefore, in the last years several models have been built, each focusing on different aspects of EGFR signalling (Birtwistle *et al*, 2007; Borisov *et al*, 2009; Hornberg *et al*, 2005; Kholodenko *et al*, 1999; Li *et al*, 2009a; Schoeberl *et al*, 2002).

However, such low-level pathway models are quite disconnected from systematic disease biology, and hence it has been argued that they have only limited use in drug discovery (Butcher *et al*, 2004). For instance, signalling cascades are not isolated units within the cell, but form part of a mesh of interconnected networks through which the signal elicited by an environmental cue can traverse (Yaffe, 2008). Ultimately, each cell is exposed to a variety of signalling cues, and the specificity of the response will be determined by the signalling mechanisms that are activated by the cue (Alberts *et al*, 2007). In this regard, recent research has highlighted the importance of so-called cross-talks between pathways, such as connections between signalling through the purinergic receptors and the Ca²⁺ sensing (Chaumont *et al*, 2008) or the link between extracellular glyocalyx structure and nitric oxide signalling pathway (Tarbell and Ebong, 2008). Hence, in order to understand biological processes at all levels, it has to be considered that pathways do not function as isolated units but instead they form a 'network of networks' that is responsible for the behaviour of the cell (Barabási and Oltvai, 2004). In this regard, network analysis has emerged as a representational formalism to study the interactions between biomedical entities (Butts, 2009). In this section, static and dynamic network analysis approaches are presented. Static network analysis can be used to derive emergent biological features from the topology of the networks, such as feedback loops. And dynamic network analysis approaches study the evolution of a biological process, over time under certain conditions, such as upon external perturbations of genetic or environmental origin.

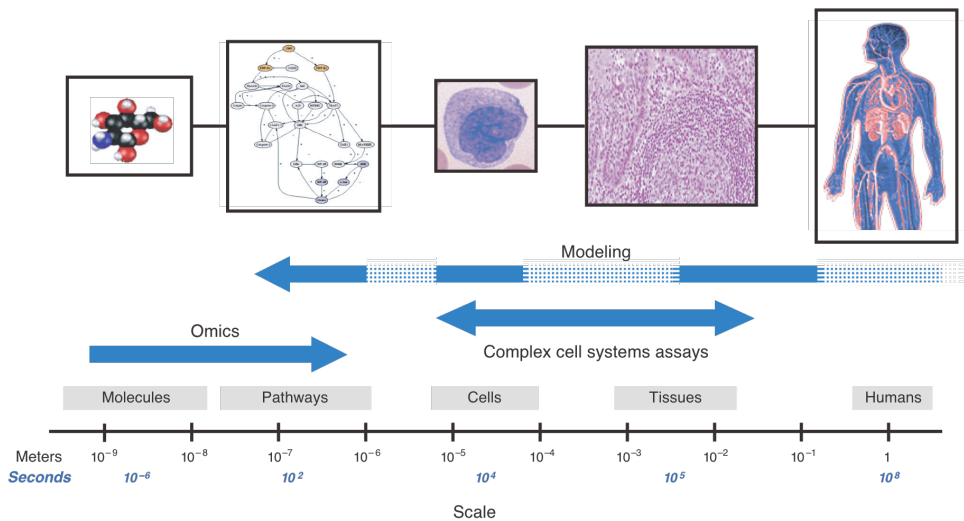


Figure 5: Granularity of systems biology, taken from (Butcher *et al*, 2004)

Systems Biology approaches try to understand biological processes at all levels. From the identification of molecules and their properties (“omics”), to the building of interaction networks or pathways, up to modelling the processes at the cell, tissue, organ or whole human body level. In this regard, experimental and computational approaches are combined.

1.3.2.1. Static network analysis

The development of high-throughput technologies has not only helped to identify genes or proteins but also to determine how these molecules interact with each other. In this context, various types of networks are possible including protein–protein interaction, metabolic, signalling and transcription-regulatory networks. The study of these molecular interactions, also called network analysis, is therefore of major importance to achieve the goal of fully understanding biological processes.

Typically, a network consists of nodes (for instance the genes and/or proteins) and edges, representing their interactions or associations. Recent examples are protein-protein interactions (Przulj *et al*, 2004), drug-target interactions (Yildirim *et al*, 2007), as well as genotype-phenotype relationships (Goh *et al*, 2007; Lee *et al*, 2008; Li and Agarwal, 2009b). Such network formalism allows the representation of integrated data sources as a single framework and the subsequent analysis of its emergent properties, such as robustness.

In this respect, it has been shown that network topology properties of biological networks differ from random networks. For instance, biological networks tend to have few hub nodes, which are connected to many neighbours and many other

INTRODUCTION

nodes with only few neighbours. On the contrary, in random networks all nodes have roughly the same number of neighbours (Barabási *et al*, 2004). Moreover, it was found that biological networks show certain local interaction patterns, so called network motifs, such as feedback or feed-forward loops (Han, 2008). Network motifs and their functions have been extensively studied for transcription regulation networks in *Escherichia coli* suggesting that biological networks have a degree of structural simplicity which could ultimately help to understand the behaviour of large and complex networks in terms of elementary circuit patterns (Alon, 2007; Shen-Orr *et al*, 2002). Furthermore, biological networks tend to be modular by forming local neighbourhoods of shared properties. For instance, the analysis of protein-protein interaction networks revealed clusters or modules of highly interacting proteins that were found to have functional relevance (Przulj *et al*, 2004; Sharan *et al*, 2007). This concept of modularity has also been discussed in the context of human genetic diseases using malformation syndromes as examples, where the relevance of the identified modules was studied with respect to disease development (Oti and Brunner, 2007; Suthram *et al*, 2010; Zaghoul and Katsanis, 2010). These modules might be responsible for properties such as robustness to environmental perturbations and evolutionary conservation (Hartwell *et al*, 1999). Hence, it is crucial to identify these modules, as they can play a key role in the aetiology of diseases. In addition, they can be used for different purposes such as the identification and prioritization of candidate disease genes (Cerami *et al*, 2010; Oti *et al*, 2006), the development of new treatments or drug repurposing. Within this context, (Suthram *et al*, 2010) identified modules which are affected in a variety of distinct diseases and showed that they contain several proteins being targets of drugs already known to effectively treat several diseases. Network topology analysis can be used to identify such functionally relevant modules (Aittokallio and Schwikowski, 2006).

An example where different types of networks are used for drug discovery and safety studies is given in figure 6. Here, drug information and biological data, such as protein-protein interactions, are integrated into a global drug network. Then, the network properties can give information about historical drug development trends and suggest new drug targets. Another network type combining disease and drug information can be used, for instance, to identify new indications for drugs, to detect new drug targets or to unravel unwanted adverse drug reactions.

In summary, topological properties of biological networks differ strongly from random networks and hence can be used to explain a variety of biological behaviours. Network analysis approaches have gained a lot of importance especially in systems biology and are more and more used to analyse the complex mechanisms underlying disease and adverse drug reactions.

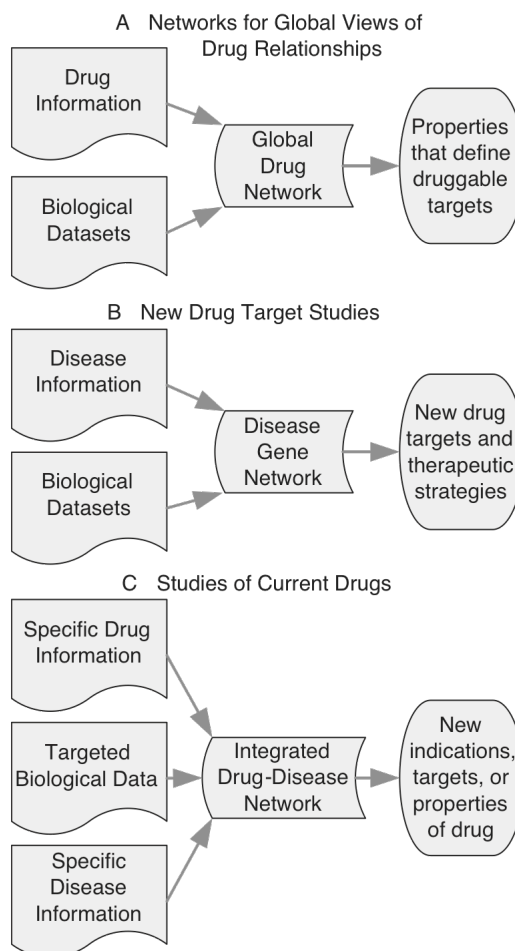


Figure 6: Different types of biological networks and their application in drug discovery and safety studies, taken from (Berger and Iyengar, 2009)

- (A) Drug information and biological data, such as protein-protein interactions, are integrated into a global drug network. The network properties can then give information about historical drug development trends and suggest new drug targets.
- (B) Information about a specific disease is used to identify potential new drug targets and therapeutic strategies.
- (C) Disease and drug information is integrated to identify new indications for drugs, unknown targets of drugs, and other interesting properties of the drugs, for instance to avoid unwanted adverse drug reactions.

1.3.2.2. Dynamic network analysis

Besides the analysis of the static properties of biological networks, mathematical and computational approaches have been proposed to simulate their dynamical behaviour. These approaches study the evolution of a biological process under certain conditions or upon perturbations over time. Such perturbations can be of genetic origin or due to external factors including drugs. Dynamic modelling approaches range from qualitative to quantitative formalisms.

Qualitative approaches are suitable to induce dynamical properties of complex systems, when few data is accessible and especially if kinetic data is missing (Chaouiya *et al*, 2008). Examples are Boolean or Petri net models. In contrast, quantitative models aim at representing the system in a detailed way and require accurate kinetic data. Here, typical models are based on differential or stochastic equations. The applicability of the modelling approach strongly depends on the kind of biological process to be modelled (metabolic, regulatory or signalling pathway), the question to be answered, and the amount and detail of data available. For instance, Boolean models become suitable if quantitative data is missing, and differential equation models can be used if details on kinetic parameters are measurable and available. Both approaches are explained in more detail in the following.

In a Boolean model, pathways are represented as interaction, directed graphs in which the nodes represent the molecules (e.g. proteins) and the edges are signed arcs denoting the direct influence of one species upon another, being either activating (+) or inhibiting (-). This first representation requires little a priori knowledge about the network under study and allows detection of important features such as feedback loops. On top of this interaction graph, a logical model is built in which each component can be either ON ("1") or OFF ("0"). Here, Boolean rules or functions define how different causal effects converging at a certain species are combined (e.g. by using AND, OR or NOT operators) (Kauffman, 1969). Examples of applications are T-cell receptor signalling (Saez-Rodriguez *et al*, 2007) or EGFR signalling (Samaga *et al*, 2009). A widely used software for Boolean models is CellNetAnalyzer (Klamt *et al*, 2007). The clear advantage of Boolean models is that they can interpret and predict behaviour based on qualitative topological data, which is readily available in the literature (Hendriks, 2010).

However, in order to get a deeper understanding of the underlying dynamics, quantitative aspects such as kinetic rates have to be considered. In this regard, differential equations have been used as mathematical formalism to create kinetic models, which typically represent protein-protein interactions and enzymatic events with a series of mass-action kinetic reactions. There are plenty examples of such quantitative models, including the aforementioned EGFR signalling (Birtwistle *et al*, 2007; Borisov *et al*, 2009; Hornberg *et al*, 2005; Kholodenko *et al*, 1999; Li *et al*, 2009a; Schoeberl *et al*, 2002), Akt signalling (Chen *et al*, 2009; Hatakeyama *et al*, 2003) or apoptosis (Bentele *et al*, 2004; Fussenegger *et al*, 2000). A widely used software for quantitative modelling is COPASI (Hoops *et al*, 2006), among others (Alves *et al*, 2006). Moreover, markup languages such as SBML (Hucka *et al*, 2003) and CellML (Lloyd *et al*) were created to allow model exchange and storage.

Finally, BioModels, a database storing and curating quantitative models of biochemical and cellular systems was built (Le Novère *et al*, 2006).

So far, the here presented examples model the behaviour of a single specific biological process. Recently, some attempts have been made to make use of these models for clinical practice. For instance, to predict individual patient response to treatment (Hendriks *et al*, 2006) or to identify pathway nodes for therapeutic intervention (Schoeberl *et al*, 2009).

1.3.3. Summary

All in all, the complexity and multilevel nature of biological functions and processes pose an extraordinary challenge to our understanding of biological processes. Nevertheless, the improvements of high-throughput technologies, computational approaches as well as the development of data standards and online-databases, allowed the identification of biomedical entities and how they interact in the cell.

Statistical methods capable of dealing with the multivariate nature of biomedical data are suitable to study the underlying data structure, for instance in order to detect clusters of biomedical entities with similar properties. Moreover, they allow the prediction of outcome dependent on a set of descriptive biomedical properties such as it is the case for predicting clinical outcome based on patient data or predicting binding affinities based on drug properties. However, the relationships that are established by such methods are statistical associations not to be confounded with causation. In this regard, systems biology methods have gained attention and application in the biomedical domain. They aim at gaining a systems level understanding by studying interconnected biomedical entities at different levels. In particular, network analysis has widely been used to uncover topological and dynamic properties to gain better understanding of the studied biological system. For instance, it has been found that biological networks, different to random networks, contain network motifs and modules of particular biological function. Also, a variety of predictive mathematical and computational models have been built that simulate the dynamics of important cell signalling and regulatory pathways. Hence, systems biology and in particular network approaches serve as suitable frameworks to investigate how genetic and environmental perturbations lead to complex traits such as disease or adverse drug reaction in human. In this context, they will have large impact on drug discovery and development and will provide the foundation for a prospective medicine that overcomes the current limitations of disease complexity and drug discovery (Auffray *et al*, 2009). However, in order to build such predictive models, heterogeneous data from various sources needs to be united and thus improved integrative data analysis approaches are required.

INTRODUCTION

2. OBJECTIVES

OBJECTIVES

The previous section has extensively discussed the most important achievements and difficulties in studying the molecular mechanisms underlying human diseases and adverse drug reactions. Moreover, the strong need for integrative analysis approaches to study these mechanisms was discussed and some approaches were presented. Against this background, this section discloses the main objectives of this PhD thesis and points out in which particular publications they have been addressed.

The main objectives of this PhD project can be summarized as follows:

- 1 . Development and application of statistical approaches to study associations in multivariate biomedical data.
In particular, development of new multivariate statistical approaches integrating different levels of drug-related data with the purpose of aiding drug development projects.
- 2 . Development and application of systems biology approaches to investigate the molecular basis of diseases and adverse drug reactions.
 - a In particular, exploitation of available repositories of biomedical data to evaluate their suitability for automatic data extraction and use in integrative biomedical research.
 - b Subsequent development of new integrative bioinformatics tools for the merging, visualization and analysis of the gathered biomedical data in order to tackle typical problems in biomedical research, especially regarding the mechanisms leading to diseases and drug adverse reactions.

2.1. Objective 1: Development and application of statistical approaches

As described in the introduction, there is a large number of drugs failing in clinical development due to a lack of efficacy or because they are causing severe side effects. It has been recognized that many of these failures relate to the fact that the classic concept of a single pharmacological receptor to be targeted by the drug is usually not valid. On the contrary, most drugs are not as selective as thought and target many receptors that are then either involved in the pharmacological effect of the drug or responsible for undesired side effects or sometimes both.

Statistical methods have been previously introduced that can be used to study the relationship of drug properties with some biological endpoints. It has also been shown that a deep and detailed understanding of those relationships is required to comprehend the molecular basis of diseases and therapies. Hence, the first objective of this thesis was to develop new multivariate statistical approaches that advance in this way. In this regard, we developed a novel multilevel statistical method, which is presented in section 3.1. The method is based on the sequential building of linked multivariate statistical models, each one introducing a new level of drug description. On the one hand, this allows overcoming the one-target assumption that typically does not apply to most drugs and on the other hand takes into account important information about the drugs at different levels, which are integrated by this method. A variety of studies about multivariate data analysis in drug design preceded this publication, presented as oral or poster communications at international conferences (see publications 15, 20, 21, 22 in section 6). Moreover, in collaboration with the Erasmus Medical Center Rotterdam, we developed a logistic regression model for the prediction of 60 days mortality of patients after aneurismal subarachnoid haemorrhage (Risselada *et al*, 2010).

2.2. Objective 2: Development and application of systems biology approaches

The second objective addresses the development and application of systems biology approaches to study the molecular mechanisms underlying human diseases and adverse drug reactions. This objective can be split into two parts. First, suitable biomedical data has to be identified, collected and subsequently exploited by integrative analysis approaches. And second, these approaches have to be developed making use of the evaluated biomedical data. Throughout this thesis, both objectives (2a and 2b) were tackled in this order.

In section 1.2 problems relevant to data integration were presented. Biomedical data is typically heterogeneous, fragmented and distributed over various online databases and literature. Despite some major improvements in defining data standards needed for correct and automatic data integration and exchange, there are still major problems related to a lack of common identifiers or standards, misuse of data formats and issues regarding curation and annotation of biomedical data. An

example for a deep analysis of biomedical repositories (objective 2a) was the extensive study of publicly available biological pathways with the aim to use the data to realize systems biology approaches (see section 3.2). In particular, in this work we present a vision of how to employ publicly available pathway data to generate network models in an automatic manner that can subsequently be used to answer practical biological problems.

Objective 2b addresses the major goal of this thesis, the development of new integrative bioinformatics tools for integration, visualization and analysis of biomedical data in order to solve typical problems in biomedical research, especially regarding disease mechanisms and drug adverse reactions. This objective was addressed in several publications.

First, a new comprehensive database on human gene-disease associations including information about mendelian, complex and environmental diseases was developed (see sections 3.3 and 3.5). For this purpose, first an extensive analysis of available repositories of gene-disease associations was performed (objective 2a). This is of particular interest as data about the genetic origin of human diseases is dispersed over various databases and literature. Hence integration is required in order to get a comprehensive view of all the genes associated to human diseases. The integrated gene-disease association database combines information about the genetic origin of mendelian, complex and environmental diseases, where the latter refers to diseases caused by environmental chemicals including drugs. Hence, this new integrated gene-disease association database serves as a suitable framework for studying human diseases and adverse drug reactions (objective 2b). It was furthermore used as a gold standard to be compared with associations extracted from the information extraction system *Text2SemRel* (Bundschuh *et al.*, 2010).

Systems biology approaches, in particular network analysis tools, have been introduced since they serve as suitable frameworks to investigate how genetic and environmental perturbations lead to complex traits, such as diseases or adverse drug reactions in human. Hence, in this work we used network analysis tools to detect functional modules related to human diseases (see section 3.3).

In another work, we present a strategy for the study of the mechanisms of adverse drug reactions, which makes use of our new integrated gene-disease association database. In brief, a Taverna workflow is presented to substantiate signals consisting of drug-adverse event pairs by checking if there are genes associated to the adverse event that are also targets of the drug. This work is presented in section 3.1.4 and oral presentations (see publications 9 and 10 in section 6).

Moreover, we developed DisGeNET, a Cytoscape plugin for user-friendly access to our new integrated gene-disease association database. DisGeNET represents a user-friendly tool allowing integration, analysis, interpretation and visualization of human gene-disease association networks, which will intensely aid investigating the molecular basis of human diseases and adverse drug reactions (see section 3.5).

OBJECTIVES

As mentioned above, a statistical association between a genetic variant and a disease or an adverse drug event is usually not sufficient to explain the underlying processes. In order to study the effect of genetic variations, it is first crucial to determine how they influence the function of the protein encoded by the affected gene. Furthermore, a more detailed study is required to investigate how this modification reflects in further downstream processes, such as in signalling pathways. Consequently, in another project (3.6), we integrated data about sequence variations and their effect on protein function with biological networks for visualization, analysis and modelling purposes. This approach aids building of predictive models that simulate the effect of sequence variations on the dynamics of cell signalling processes and is therefore useful to shed light on the molecular basis of human diseases and adverse drug reactions. The presented method moves towards the analysis of disease-related states important to understand how genetic and environmental perturbations affect processes in the human body.

Moreover, our work related to objective 2 was presented in various oral and poster communications at international conferences (see publications 8, 11, 12, 15, 16 and 18 in section 6).

3. THESIS PUBLICATIONS

THESIS PUBLICATIONS

3.1. A Novel Multilevel Statistical Method for the Study of the Relationships between Multireceptorial Binding Affinity Profiles and In Vivo Endpoints

Selent J*, Bauer-Mehren A*, López L, Loza MI, Sanz F, Pastor M

Molecular Pharmacology. 2010;77(2):149-58

* Both authors contributed equally to this work

THESIS PUBLICATIONS

Selent J, Bauer-Mehren A, López L, Loza MI, Sanz F, Pastor M. [A novel multilevel statistical method for the study of the relationships between multireceptorial binding affinity profiles and in vivo endpoints](#). Mol Pharmacol. 2010; 77(2): 149-58.

3.2. Pathway databases and tools for their exploitation: benefits, current limitations and challenges

Bauer-Mehren A, Furlong LI, Sanz F

Mol Syst Biol. 2009;**5**:290

THESIS PUBLICATIONS

Bauer-Mehren A, Furlong LI, Sanz F. [Pathway databases and tools for their exploitation: benefits, current limitations and challenges](#). Mol Syst Biol. 2009; 5:290.

3.3. Network analysis of an integrated gene-disease association database reveals functional modules in mendelian, complex and environmental diseases

Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI

submitted. 2010

THESIS PUBLICATIONS

Network analysis of an integrated gene-disease association database reveals functional modules in mendelian, complex and environmental diseases

Anna Bauer-Mehren¹, Markus Bundschuh², Michael Rautschka¹, Miguel A. Mayer¹, Ferran Sanz¹, Laura I. Furlong^{1,*}

¹ Research Programme on Biomedical Informatics (GRIB) IMIM, DCEX, Universitat Pompeu Fabra, C/Dr. Aiguader 88, 08003 Barcelona, Spain

² Institute for Computer Science, Ludwig-Maximilians-University Munich, Oettingenstr. 67, 80538 Munich, Germany

* To whom correspondence should be addressed.

Running Title: Integrated view of human gene-disease associations

Abstract

Background

For many years, scientists have been trying to understand the mechanisms underlying diseases in order to design new preventive and therapeutic strategies. Most human diseases arise due to interactions between multiple genetic variants and environmental factors. The fragmentation of information, so called knowledge pockets, poses obstacles to our understanding of the molecular processes underlying human diseases.

Methodology/Principal Findings

We developed a comprehensive database of gene-disease associations by integrating data from diverse sources including text-mining derived associations. We furthermore built gene-disease association networks and used network topology analysis to study their emergent properties. The global network analysis confirms the need of integrating gene-disease associations to bridge the aforementioned knowledge pockets. The analysis also shows that human diseases have many gene associations in common indicating a highly shared genetic origin. We furthermore extracted disease-related modules by means of clustering and demonstrate that most diseases are associated to a core set of biological processes. More strikingly, similar findings are obtained when studying groups of diseases. This suggests that the diseases in these groups, which can be very similar but also very unrelated, might arise due to dysfunction of the same biological processes in the cell. Our analysis also shows that only few diseases are solely caused by defects in direct interactions between proteins. We present in several case studies how the detection of disease-related modules and their adjacent functional analysis can be used to shed light on

THESIS PUBLICATIONS

disease development or the influences of environmental variables on human health.

Conclusions/Significance

For the first time, we include mendelian, complex and environmental diseases in an integrated gene-disease association database. We furthermore show that the concept of modularity applies for all of them and provide a functional analysis of disease-related modules. Such analysis can provide important new biological insights, which might not be discovered when considering each of the gene-disease association repositories independently. Hence, we present a suitable framework for the study of mechanisms leading to diseases and adverse drug reactions. Additionally, we make all data publicly available through DisGeNET, a plugin for Cytoscape to access and analyse our data with the aim to aid future studies of disease development and drug discovery projects.

Introduction

For many years, scientists have been trying to understand the molecular and physiopathological mechanisms of diseases in order to design new preventive and therapeutic strategies. In the near past, the combination of experimental and computational methods led to the discovery of disease-related genes (Botstein and Risch, 2003; Kann, 2010). A well-known example is Phenylketonuria, where the function of the gene encoding the PAH enzyme was studied with respect to the mechanism of the disease (Scriver and Waters, 1999). However, we are still far from fully understanding disease causation, especially regarding complex diseases such as cancer (Kann, 2010). Even for mendelian diseases, such as Phenylketonuria, this is not fully achieved because phenotypic outcome cannot be predicted solely based on the genotype (Scriver *et al*, 1999). It has become evident, that many human diseases cannot be attributed to malfunction of single genes but arise due to complex interactions among multiple genetic variants (Hirschhorn and Daly, 2005). Moreover, influence of external variables such as environmental factors, infectious agents or drugs have to be considered when studying the occurrence and evolution of a disease. Furthermore, for many complex diseases alterations in several genes can make subtle contributions to the susceptibility of a particular individual. At the end of the day, how a disease is caused and thus how it can be treated can only be studied on the basis of the entire body of knowledge including all genes that are associated to the disease and their interactions through biological pathways. However, with the unprecedented wealth of information available, it is extremely difficult to obtain a complete picture of the genetic basis of diseases. As a consequence, in order to obtain such a complete picture, data integration from different sources is required. This is of special interest considering the fact that individual researchers are often restricted to so called knowledge pockets (Cokol *et al*, 2005) that are much smaller than all the knowledge available but spread over literature or distinct databases. This fragmentation of information poses obstacles to our understanding of the molecular processes underlying human disease.

In the 60s, Dr. McKusick started collecting information about genes and their association to diseases first as a book and later as a database. His Online Mendelian

Inheritance in Man (OMIM) database has become a highly popular source in medical genetics (Hamosh *et al*, 2005). OMIM traditionally focused on monogenic diseases and later started to include complex diseases. In the last years, further databases have been built, among them PharmGKB, a database specialized on the knowledge about genes that are involved in modulating drug response (pharmacogenes) (Klein *et al*, 2001) or CTD, which is focused on the effect of environmental chemicals on human disease (Mattingly *et al*, 2006). However, each of the databases focuses on different aspects of phenotype-genotype relationships. Moreover, due to the vast increase of published literature in health and life sciences, no one (not even expert curators of such databases) can keep track of the relevant knowledge that is regularly published (Cokol *et al*, 2005). Here, text-mining has evolved as a useful tool to automatically extract information about the relationships between biomedical entities reported in the literature (Ananiadou *et al*, 2010). Thus, to obtain a comprehensive picture of the state of the art knowledge about the genes influencing human diseases, integration of information from different databases and literature is needed.

In the last years network analysis has emerged as a representational formalism to study the interactions between biomedical entities (Butts, 2009). Recent analysis examples are protein-protein (Przulj *et al*, 2004a), drug-target interactions (Yildirim *et al*, 2007), as well as genotype-phenotype relationships (Goh *et al*, 2007; Lee *et al*, 2008; Li and Agarwal, 2009). Such network formalism allows the representation of integrated data sources as a single framework and the subsequent analysis of its emergent properties. In this respect, it has been shown that network topology properties of biological networks differ from random networks. For instance, biological networks tend to have few hub nodes, which are connected to many neighbours and many other nodes with a much smaller degree, whereas in random networks most nodes have roughly the same number of neighbours (Barabási and Oltvai, 2004).

In this context, analysis of protein-protein interaction networks revealed clusters or modules of highly interacting proteins that were found to have functional relevance (Przulj *et al*, 2004b; Sharan *et al*, 2007). This concept of modularity has also been discussed in the context of human genetic diseases using malformation syndromes as examples (Oti and Brunner, 2007; Suthram *et al*, 2010; Zaghoul and Katsanis, 2010). Many human diseases show overlap in their phenotype and hence diseases have been clustered according to phenotypic similarity (Freudenberg and Propping, 2002; Hidalgo *et al*, 2009). Many phenotypically similar diseases are caused by functionally related genes, such as Stickler, Marshall and OSMED syndromes (Ahmad *et al*, 1991; Melkonimi *et al*, 2000; Snead and Yates, 1999). Also, it was found that several forms of human ataxias are related to the same set of interacting proteins (Lim *et al*, 2006). Thus, many diseases are caused by dysfunction of interacting proteins or biological pathways (D'Andrea and Grompe, 2003; Jones *et al*, 2008; Lim *et al*, 2006). It is crucial to identify these modules, as they play a key role in the aetiology of the diseases. In addition, they can be used for different purposes such as the identification and prioritization of candidate disease genes (Cerami *et al*, 2010; Oti *et al*, 2006), the development of new

treatments or drug repurposing. For instance, Suthram *et al.* identified modules, which are affected in a variety of distinct diseases and showed that they contain several proteins being targets of drugs that are already known to effectively treat several diseases (Suthram *et al.*, 2010). Network topology analysis can be used to identify such functionally relevant modules (Aittokallio and Schwikowski, 2006).

Similarly, Goh *et al.* used a global analysis of a human gene-disease network based on OMIM to show that gene products related to the same disease have a higher likelihood to physically interact (Goh *et al.*, 2007). In this article, we pick up the concept of modularity of human genetic diseases with the aim of assessing it at a global scale for the whole spectrum of human diseases including mendelian, complex and environmental diseases. For this purpose, we first create a comprehensive database of human gene-disease associations including data from several databases and text-mining derived associations in order to bridge the aforementioned knowledge pockets. The resulting database comprises the whole spectrum of human diseases with genetic origin, including mendelian, complex and environmental diseases, and represents, to the best of our knowledge, the most complete view on human gene-disease associations that is currently publicly available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download>. Moreover, we represent our database as graphs and study global properties by means of network analysis. Our results indicate that for most human diseases, and even for sets of related diseases, functional modules exist. Such modules are in general comprised of more than one biological process. We show in several case studies how our network representation of human genetic diseases and the adjacent detection of functionally related gene modules can be used not only to shed light on the molecular basis of human diseases but also to gain a better understanding of the influence of environmental factors, including drugs, on human health. Our results confirm the need of integrating human gene-disease associations from various sources. Moreover, they support the concept of modularity of human genetic diseases, which we studied for the first time at a global level for the whole spectrum of human diseases. Finally, we make all data publicly available through DisGeNET, a plugin for Cytoscape to access and analyse our data (Bauer-Mehren *et al.*, 2010).

Results

Global network analysis of a comprehensive database on gene-disease associations

A comprehensive database on gene-disease associations was developed by integrating information from four repositories: Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.*, 2005), UniProt/SwissProt (UNIPROT) (Apweiler *et al.*, 2004), Pharmacogenomics Knowledge Base (PHARMGKB) (Altman, 2007), and Comparative Toxicogenomics Database (CTD) (Mattingly *et al.*, 2006). In addition, associations from a literature-derived human gene-disease network (LHGDN) (Bundschuh *et al.*, 2008) were included to increase the coverage of our database (see Materials and methods). In this regard, two aspects had to be considered: the different data sources represent gene-disease associations in

different ways and they use different gene and disease vocabularies. To ensure an accurate integration of gene-disease association data, we developed a gene-disease association ontology (see Figure 1 in Supplementary Material). Moreover, we performed a mapping of disease and gene vocabularies (see Materials and methods). Diseases were classified into 26 disease classes according to the MeSH hierarchy allowing the analysis of groups of related diseases based on standard disease classification. Using this disease classification, many diseases are assigned to more than one disease class as several systems or organs are affected.

This integration resulted in an increase in coverage of (i) diseases, (ii) genes and (iii) their associations compared to the original data sources (see Figure 1). Moreover, the overlap among databases was surprisingly small, highlighting the need of integrating different data sources to obtain a comprehensive source of current knowledge about gene-disease associations (see Figure 2 in the Supplementary Material).

We used bipartite graphs as network formalism to represent the gene-disease association database. We created four different bipartite graphs called OMIM (only including data from OMIM), CURATED (including data from expert curated databases), LHGDN (text mining data only) and ALL (including all available gene-disease associations) (Figure 2). The more data sources were considered the denser the networks became indicating that many more diseases share genetic origin than reflected in a single source. In OMIM, most diseases are associated to one or few genes. Contrasting, in the other networks most diseases are part of a large connected component. This largest connected component increases noticeably when integrating more data, while the number of diseases associated to only one gene decreases concomitantly (see curly brackets in Figure 2). This indicates that most diseases are associated to more than one gene, even for mendelian diseases. These findings are in agreement with several studies showing that in many monogenic disorders, the observed phenotype is the result of the combined effect of a primary gene, modifier genes and other factors (Dipple and McCabe, 2000; Scriver *et al*, 1999). Thus, even for mendelian diseases, complexity of the gene-disease associations is an issue.

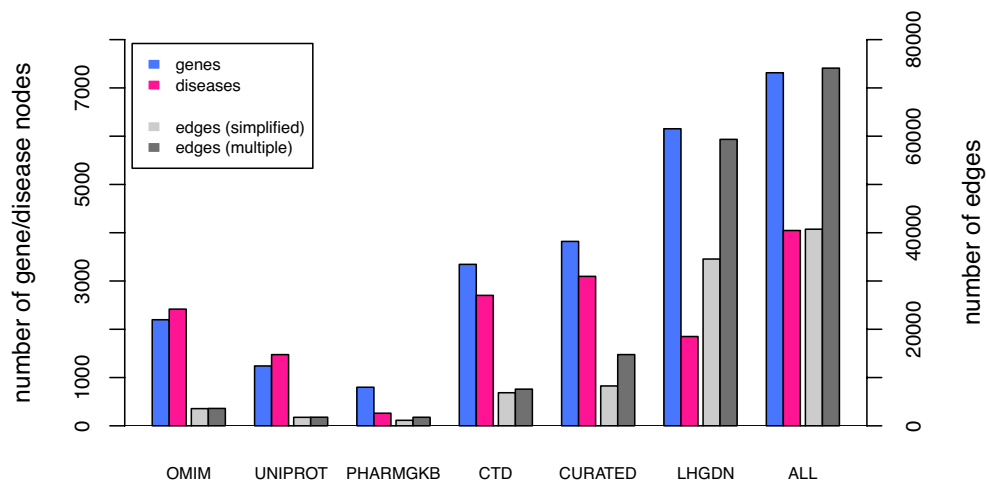


Figure 1: Number of distinct gene/disease nodes and edges per data source

The number of diseases refers to the actual number of disease nodes in the networks after mapping of disease vocabularies. The number of edges (simplified) refers to the number of distinct gene-disease associations. The number of edges (multiple) represents all edges, considering one edge for each source or evidence reporting the gene-disease association.

Topological analysis of the networks can uncover important properties of gene-disease associations. For example, the node degree follows distinct frequency distributions according to the type of network (Newman, 2003). Our analysis shows that the degree distributions of diseases and genes are different from degree distribution of random networks, but none of them follows a power law distribution (Figure 3 in Supplementary Material). Nevertheless, there are two main trends visible. Both, the number of hubs (nodes that are highly connected with other nodes) and the average degree size increase dramatically through the integration process. For the gene nodes, the average degree ranges from 1.6 in OMIM to 5.6 in ALL and for the disease nodes, from 1.5 in OMIM to 10.1 in ALL (see Figure 3 in Supplementary Material). The degree of a disease node represents the number of associated genes and hence can be used as a measure for the locus heterogeneity of the disease. There is a dramatic increase in the maximum locus heterogeneity observed in each data set; there are 30 genes annotated to Diabetes Mellitus Type II in OMIM, 350 genes associated to Prostatic Neoplasms in CURATED, 1133 genes associated to Neoplasms in LHGDN and 1274 genes associated to Breast Neoplasms in ALL. With respect to the genes, the increase in the node degree is less dramatic but still visible (see Figure 3 in Supplementary Material). Two network projections were obtained from the bipartite gene-disease networks to generate disease centric and gene centric representations of the data. Both projections allow studying diseases with genetic origin in a more detailed way (see next section). We can also consider the degree distributions of the disease and gene

projection networks. In contrast to the bipartite graph degree distribution, the degree of a gene (disease) node indicates the number of gene (disease) neighbours in the gene (disease) projection network. Interestingly, the degree distributions of the projected networks are much broader than the degree distributions of the bipartite graph (data not shown). The right tails of the distributions get much more populated the more data sources are included (more hubs in CURATED than in OMIM and many more hubs in LHGDN than in CURATED). Moreover, in the disease projection the average number of diseases connected to another disease is 2.2 in OMIM, 8.5 in CURATED and 103.6 in ALL. All these findings suggest a much higher level of interrelation of human diseases than observed by solely considering a single data source (e.g. OMIM).

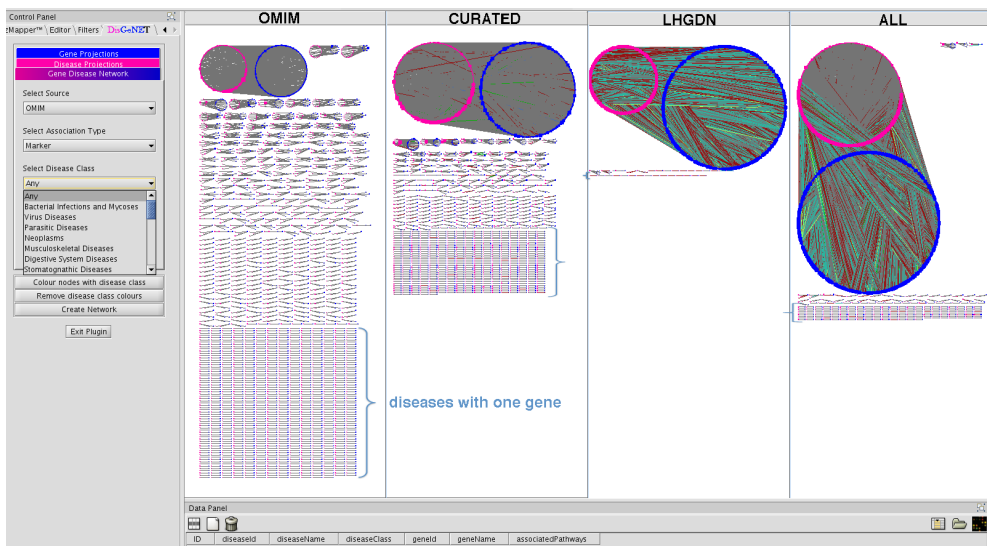


Figure 2: Cytoscape screenshot of the four gene-disease networks

Cytoscape screenshot depicting the four gene-disease networks. The Cytoscape layout “Group attributes layout” was used to group gene (blue) and disease (magenta) nodes. The colour of the edges corresponds to the type of gene-disease association according to our gene-disease association ontology. Grey edges represent *Marker* association, red denotes *GeneticVariation*, blue corresponds to *Therapeutic* class, green to *RegulatoryModification*. The curly brackets frame diseases with only one gene associated, the number of diseases with single gene annotation decreases visibly when incorporating more data (from OMIM to ALL).

Functional analysis at the level of individual diseases

Integration of different sources resulted in an increase in the locus heterogeneity for many diseases. This is particularly evident when incorporating text-mining derived associations. Several studies based on OMIM database indicated that, for diseases with high locus heterogeneity, the associated genes are involved in the same biological process (Freudenberg *et al*, 2002; Goh *et al*, 2007; van Driel *et al*, 2006). Goh *et al*. introduced the homogeneity measure, defined as the maximum fraction of genes sharing the same biological annotation (Goh *et al*, 2007). Thus, we used the homogeneity measure to test if the aforementioned concept still applies to our integrated data set. We calculated GO biological process (GO-BP) homogeneity for each disease as defined in equation 2. We obtained similar results for OMIM data as other authors (Goh *et al*, 2007), however for the larger networks (CURATED, LHGDN, ALL) the homogeneity values decreased. In addition, we also calculated homogeneity using annotations to biological pathways. For single diseases, we obtained similar results as for GO-BP with average values ranging between 56 % and 77 % (77 % in OMIM, 67 % in CURATED, 56 % in LHGDN and 59 % in ALL). Hence, for the integrated data sets, more than one biological process is associated to a single disease.

In order to further explore why there is a decrease in homogeneity when integrating more data into the network, we studied the dependency of homogeneity values on the number of associated genes. Interestingly, homogeneity values varied with the number of genes associated to a given disease (see Figure 3). For all data sources, even for OMIM, the homogeneity decreases with increasing number of associated gene products ($r \gg -0.25$) and was significantly higher (p-value < 0.05) than for random controls (see Figure 3). For instance, in CURATED, diseases with two to five annotated gene products have on average 75 % of the gene products annotated to the same pathway, while this value decreases to 38 % if 50 to 100 gene products are annotated to the disease. For diseases with two to five annotated gene products, approximately 70 % of them participate in the same pathway for all data sources. On the other hand, for diseases with more than 10 gene products annotated, it is more likely that more than one pathway is involved. Moreover, it is striking that although the text-mining derived network is very dense with an average of 18.7 genes per disease, the pathway homogeneity still differs significantly from random, with values comparable to the ones observed in the CURATED set. Similar results were obtained for GO-BP term homogeneity analysis (Figure 4 in Supplementary Material).

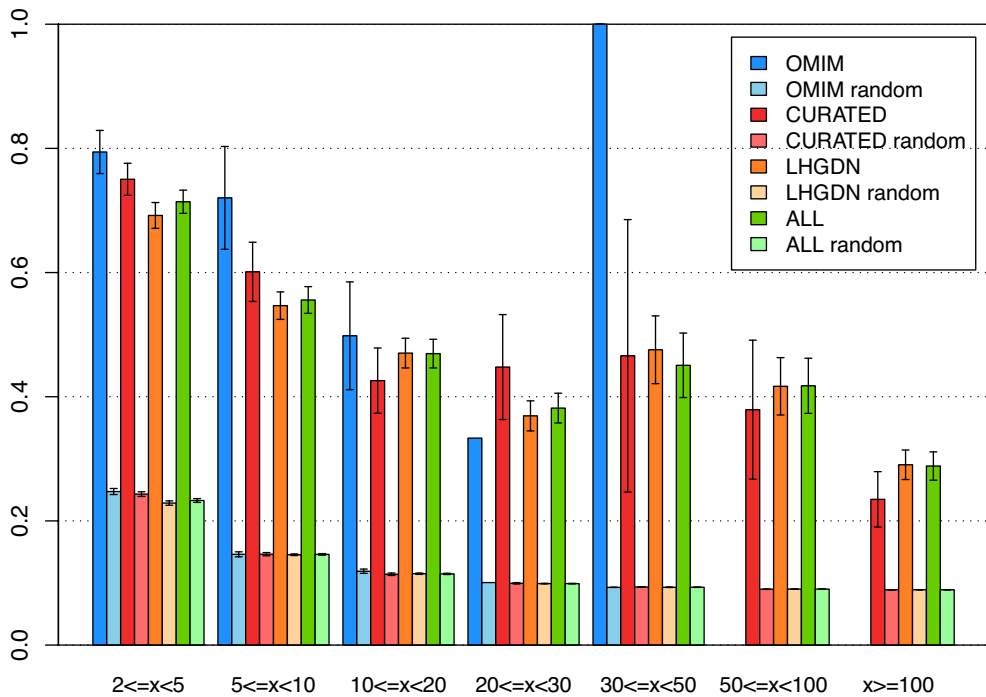


Figure 3: Pathway homogeneity for diseases

Mean pathway homogeneity values of single diseases and random controls are plotted for all four networks binned by the number of associated gene products per disease. Pathway homogeneity values range from 0 to 1, where 1 means that all gene products associated to the disease are annotated to the same pathway. Confidence intervals of 95% were added to allow comparison of real to random values. For OMIM, there are only two diseases with more than 30 gene products annotated, both with a pathway homogeneity of 1.

Functional analysis of gene and disease clusters

In the disease projection network, most diseases are highly connected with average degrees of 2.2 in OMIM up to 103.6 in ALL, and depending on the data source, the largest connected component gets very dense. Hence, we applied a graph-clustering algorithm to identify highly connected units, so called clusters or modules, in the projected networks (see Materials and methods). Diseases in the resulting disease clusters have a common genetic aetiology. The clustering revealed some disease clusters being homogeneous in terms of disease class and other clusters containing diseases from different disease classes (disease clusters can be visualized within Cytoscape using the session file available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download>). We then investigated if the genes associated to the disease clusters are more likely to participate in the same biological processes, by analysing the GO-BP and pathway homogeneity (see Supplementary Material). On average, pathway homogeneity is

THESIS PUBLICATIONS

68 % for OMIM and 59 % for CURATED, suggesting that in these datasets 60-70 % of the gene products belonging to a disease cluster participate in the same pathway. For the more populated networks (LHGDN and ALL) the average pathway homogeneity values of disease clusters decreases to approximately 50 %. Compared to single diseases, homogeneity values are slightly smaller, but still significantly different from random. No correlation was observed between the homogeneity values and the number of diseases per cluster or the disease class (data not shown). However, similarly to single diseases, we observe a decrease of the pathway homogeneity with increasing number of gene products annotated to the diseases in a cluster ($r \gg -0.26$) (Figure 5 in Supplementary Material). In addition, the analysis identified disease clusters with different degrees of pathway/GO-BP homogeneity; some disease clusters are characterized by one main pathway or biological process, while other disease clusters might involve several biological processes. Interestingly, in CURATED, a high proportion of the disease clusters (71 %) have more than one pathway annotated (similar values were obtained for GO-BP).

The gene projection networks are also very dense and a graph-clustering algorithm was applied in order to identify groups of phenotypically related genes (see Materials and methods). These clusters represent genes that share disease associations. By exploring their functional annotation we can gain insight into biological processes that are common to the genes of each cluster and hence common to their associated diseases. Several studies presented evidence arguing for a modular nature of human diseases, especially for congenital malformations and related syndromes (Lim *et al.*, 2006). Thus, we wanted to assess at a global scale including the whole spectrum of diseases, to what extent the groups of phenotypically related genes represent functional modules in the cell.

Gene products can be functionally related to each other in different ways. For example, they can be related by means of direct, physical protein-protein interactions or by more indirect associations, as observed between enzymes in the context of a metabolic pathway.

First, we assessed to which degree the proteins encoded by the genes in the clusters physically interact in the cell. For this purpose, we used a recently published human interaction network (HIN) based on protein-protein and signalling interactions (Cerami *et al.*, 2010). We introduce the HINscore, which represents the number of connected components in a subgraph of HIN containing all nodes of the gene cluster (see equation 3). The score is defined as the fraction of disease-clustered gene products that are actually involved in physical interactions. For CURATED and OMIM, clusters including less than 50 nodes show HINscores significantly higher than for random clusters, while for the other networks the difference is significant for clusters of less than 15 nodes (see Figure 4A). In larger clusters, it is more likely to find the same number of connected components as in randomly selected gene clusters. Figure 4B illustrates some selected clusters from CURATED with high HINscores. The upper part shows the clusters from the gene projection network, highlighting in red the edges corresponding to physical interactions between the gene products. The lower part shows their corresponding

subgraphs in HIN highlighting the physically interacting modules of genes related to the disease. For instance, cluster B.1 contains genes mainly associated to mitochondrial complex I deficiency (genes in the lower right part of the gene cluster), and Leigh and Alexander diseases (genes in the upper left part of the gene cluster). The latter are neurometabolic disorders that result from defects in the mitochondrial respiratory chain. Genes associated to these diseases encode proteins that form a physically interacting module as illustrated by the HIN subgraph (bottom). Other examples of clusters with high HIN score are related to peroxisomal disorders (e.g. Zellweger syndrome), different types of anemia (Diamond-Blackfan anemia or Heinz body anemia) or Walker-Warburg and Fukuyama syndromes. Thus, the HINscore can be used to identify phenotypically derived gene clusters in which physical, direct interactions between the gene products might play an important role.

Second, we evaluated the indirect relationships between disease gene products. For this purpose we calculated the degree of homogeneity of the phenotypically related gene clusters with respect to biological processes and pathways. The average GO-BP homogeneity for clusters smaller than 50 nodes is significantly (p-value < 0.05) higher than for randomly selected clusters (except for ALL). Moreover, similar to disease clusters, we observe that homogeneity decreases with increasing size of the cluster for all data sets ($r \gg -0.20$). For very large clusters (clusters larger than 50 nodes), the results are not significantly different to random controls; however, such clusters under represented in our dataset. 74 % of the clusters in CURATED have a GO-BP homogeneity value smaller or equal to 75 %, hence for most of the clusters, there are at least two GO-BP terms annotated (see Figure 6 in the Supplementary Material). Similar results were obtained when assessing pathway homogeneity (see Figure 7 in the Supplementary Material).

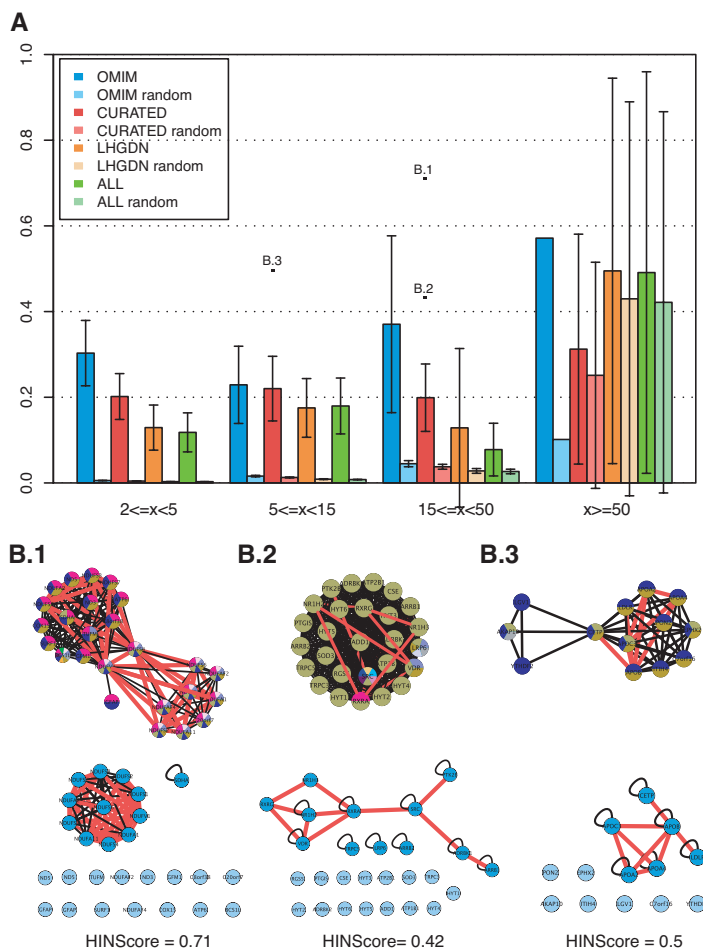


Figure 4: HINscores for phenotypically derived gene clusters

A: Mean HINscores plotted for different cluster sizes for all networks and random controls.

B: Selected gene clusters denoted as B.1, B.2, B.3 and their corresponding HIN subgraphs from the CURATED dataset. Upper part shows the gene clusters and lower part the HIN subgraphs. In the phenotypically derived gene clusters, red edges represent physical interactions among the gene products. In the HIN subgraphs, red edges denote phenotypic relationship among the corresponding genes. Nodes in light blue belong to the phenotypically derived gene clusters that are not present in HIN. B.1 is associated to mitochondrial respiratory chain deficiencies, Leigh and Alexander Disease. B.2 corresponds to Hypertension and Cardiovascular Diseases. B.3 represents different types of Hyperlipoproteinemia.

Most of the gene clusters (72 %) in CURATED are of size smaller than 15. Interestingly, for such cluster sizes, HINscores and homogeneity values differ significantly from random for all four networks. In general, clusters with high HINscore, pathway or GO-BP homogeneity are homogeneous in terms of associated diseases, meaning that the genes are annotated to similar diseases. For example, gene products from the cluster depicted in Figure 4 (B.1), which corresponds to mitochondrial respiratory chain deficiencies and Alexander and Leigh Disease, physically interact and hence the HINscore is very high. In addition, they are all annotated to the same pathway resulting in a pathway homogeneity value of 1.

In contrast, clusters with very low homogeneity values (< 0.25) are heterogeneous in terms of disease annotation, and are underrepresented in the dataset. Such clusters contain genes with very high allelic heterogeneity. In CURATED, for instance, genes that have more than 20 associated diseases, belong to heterogeneous clusters with low GO-BP (mean = 0.17) and pathway homogeneity values (mean = 0.28).

The majority of clusters show medium range HINscore, pathway and GO-BP values, suggesting that more than one biological process is relevant to a disease or a group of related diseases. All in all, the results show that the concept of modularity applies for most diseases, even for clusters of related diseases, as homogeneity and HINscore values differ significantly from random.

Case studies

Our database represents a suitable framework to study human diseases with genetic origin, including those in which environmental factors play an important role. As environmental factors we consider toxins to which we are exposed in our daily life but also therapeutic drugs. Thus, the database can serve to explore all the genes known to be associated to a disease, to study relationships among diseases at the genetic level or to identify biological processes associated to certain diseases.

Here we provide some exemplary case studies to illustrate the kind of outcomes that can be obtained using our integrated database and to inspire future studies. The results presented in the previous section indicate that the gene clusters resemble functional units, in terms of shared biological processes, and these processes can be studied more deeply to shed light on the mechanisms related to the diseases. Hence, we determined the specific biological processes relevant for each gene cluster by calculating GO term and pathway enrichment (see Materials and methods). In total, we obtained significant (p -value < 0.05) GO and pathway enrichment for 94 % of the clusters in CURATED. The first two clusters are very large (740 and 144 genes). Hence, we applied a second round of clustering on these two clusters to obtain smaller modules to ease their analysis. Details on the enrichment results are available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#ClusterAnalysis>. Moreover, we provide a Cytoscape session file (DisGeNET.cys) including the examples presented here, available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download>.

THESIS PUBLICATIONS

The following examples illustrate the (i) prediction of disease candidate genes, (ii) study of the interactions between environmental factors and diseases at the genetic level, (iii) identification of shared mechanisms of distinct diseases and (iv) analysis of mechanisms of adverse drug reactions.

Example 1 - Gene clusters and pathway analysis to predict new disease candidate genes

One of the clusters is composed of 20 genes, most of them associated to melanoma and developmental diseases affecting pigmentation, eye and ear functions, such as Tietz and Waardenburg syndromes. Most genes of the cluster are associated to melanoma with the exception of MITF. GO enrichment analysis resulted in terms like “melanin biosynthetic process from tyrosine” (GO:0006583), “eye pigment biosynthetic process” (GO:0006726) and “melanocyte differentiation” (GO:0030318), among others. These are all processes relevant to skin, hair and eye pigmentation, hearing function in the cochlea, and skin carcinogenesis. Figure 5B shows the Melanogenesis pathway (KEGG hsa:04916), which is the most significantly enriched pathway for this cluster. The proteins encoded by genes TYR and ASIP (ASP in KEGG) and the transcription factor encoded by MITF regulating expression of the TYR gene, are highlighted in red in the pathway. Since MITF appears not only in the same phenotypically derived cluster but also in the same pathway as the genes associated to Melanoma, it could be proposed that MITF is a candidate disease gene for Melanoma. In fact, we could confirm this finding by checking the disease neighbourhood of MITF in the gene-disease network (ALL) that also includes text-mining derived information (see Figure 5C). The information extracted by text-mining indicates that MITF has been reported as a gene involved in melanocyte development and characterized as melanoma oncogene (Carreira *et al*, 2006; Garraway and Sellers, 2006). In conclusion, clustering analysis of the gene projection network followed by functional enrichment analysis can be used to propose new candidate disease genes.

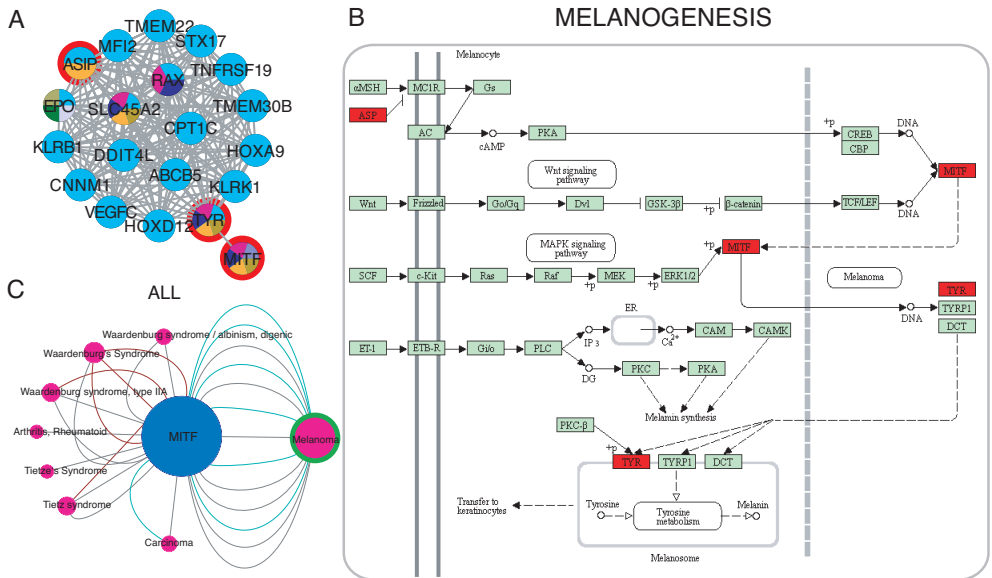


Figure 5: Candidate disease gene prediction

A: Phenotypically derived gene cluster associated to Melanoma. MITF is the only gene in the cluster not associated to Melanoma. B: The Melanogenesis pathway (KEGG: hsa:04916) with genes MITF, TYR and ASP (ASIP in A) coloured in red. C: Neighbourhood of MITF gene in network ALL.

Example 2 - Interaction between environmental exposure with arsenic compounds and cancer at the genetic level

Another cluster contains 67 genes mostly associated to Arsenic Poisoning, skin and nervous system diseases, and different types of neoplasms. Arsenic is a well established human carcinogen, and many studies support an association between arsenic exposure and increased incidence of solid tumours, such as lung, bladder, prostate, renal and skin tumours (Celik *et al*, 2008; Chiou *et al*, 1995; Radosavljević and Jakovljević, 2008; Smith *et al*, 1992; Tsuda *et al*, 1995; Yang *et al*, 2008). Moreover, studies conducted in developing countries show a general increase in the incidence of different types of cancers, which is hypothesised to be associated with exposure to environmental toxins, among other factors, some of them of genetic origin (Park *et al*, 2008; Sankaranarayanan and Boffetta, 2010; Thun *et al*, 2010). Thus, there is a need to investigate the interactions among environmental carcinogens and genetic factors (Sankaranarayanan *et al*, 2010). Although more studies are needed to determine a linkage between arsenic exposure and Breast Cancer incidence (Navarro Silvera and Rohan, 2007), this cluster indicates a possible association at the genetic level. Some of the genes that are associated to Arsenic Poisoning are also known to be associated to Breast Cancer, such as TNF, CCL20, CXCL2, CXCL3 and IL1B. Apoptosis-inducing factors IL1B and TNF are down regulated by arsenic compounds (C, 2006), as indicated by the supporting

evidence of one of the associations in the dataset. This observation combined with the DNA damaging effect of arsenic (Bau *et al*, 2002) may provide a mechanistic hypothesis for the tumorigenic effects of arsenic.

All in all, cluster analysis of the gene projection network uncovered an interesting relationship between environmental exposure to arsenic compounds and cancer. This relationship deserves further investigation at the epidemiological and molecular levels.

Example 3 – Identification of shared mechanisms of different diseases

Another cluster containing 79 genes is an example of a heterogeneous cluster in which genes are associated to different diseases. Figure 6 shows the three main disease groups, Atopic Dermatitis (an autoimmune skin disease), Diabetes Mellitus Type I (an early onset, insulin-dependent, autoimmune disease), and Inflammatory Bowel Diseases (including Crohn Disease and Ulcerative Colitis). All these diseases are related as they share many gene associations. Interestingly, according to MeSH, one of the diseases (Crohn Disease) is not classified as Immune Systems Disease but only as Digestive Systems Disease (genes coloured in pink). However, it is well established that Crohn Disease is an autoimmune disease (Duerr, 2003; Scaldaferrri and Fiocchi, 2007).

GO term and pathway enrichment analysis showed that for this heterogeneous cluster, there are common biological processes associated to the distinct diseases. For instance, although there are 59 pathways annotated to this cluster, the pathway homogeneity is 41 % indicating that almost half of the gene products appear in the same pathway. The most significantly enriched processes are related to immune (GO:0006955) and inflammatory response (GO:0006954), while the most significantly enriched pathway is the Jak-STAT signalling pathway (KEGG hsa:04630). Figure 6 shows the Jak-STAT signalling pathway, which contains genes associated to all three diseases of this cluster. Interestingly, the connections of the different diseases can be seen on different levels of this signalling pathway, from receptor-ligand interactions towards downstream signalling and transcriptional regulation.

This example shows the value of clustering and subsequent GO and pathway enrichment analysis to identify mechanisms that are common to distinct diseases.

Example 4 - Knowledge about genetic basis of diseases can shed light on mechanisms underlying drug toxicity

Rhabdomyolysis can result from a traumatic injury, but also appears as a consequence of other diseases or due to intoxication with recreational and prescription drugs. We use Rhabdomyolysis as an example to illustrate how to use our database to understand mechanisms underlying drug toxicity. Using the gene-disease network from ALL, we found several Myopathies, CPT Deficiencies and other diseases, such as Acute Renal Failure and Malignant Hyperthermia, in the neighbourhood of Rhabdomyolysis. One of the genes associated to Rhabdomyolysis is CPT2, which encodes the mitochondrial carnitine palmitoyltransferase II. Inherited deficiencies in this enzyme lead to CPT2 deficiency, an autosomal

recessive disorder characterized by recurrent Myoglobinuria, episodes of muscle pain, stiffness, and Rhabdomyolysis. On the basis of this knowledge it is possible to create a hypothesis on the mechanisms by which certain drugs such as Perhexiline can lead to Rhabdomyolysis. Perhexiline, which is prescribed for severe Angina Pectoris (De Luca *et al*, 2008) inhibits CPT1, shifting myocardial substrate utilization from long chain fatty acids to carbohydrates. Perhexiline can also target, but to a lesser extent, CPT2 (Kennedy *et al*, 2000), which would explain the toxic effects of the drug in skeletal muscles due to the association of CPT2 with Rhabdomyolysis. This example shows the power of using our gene-disease data in combination with drug-target data for the analysis of drug toxicities.

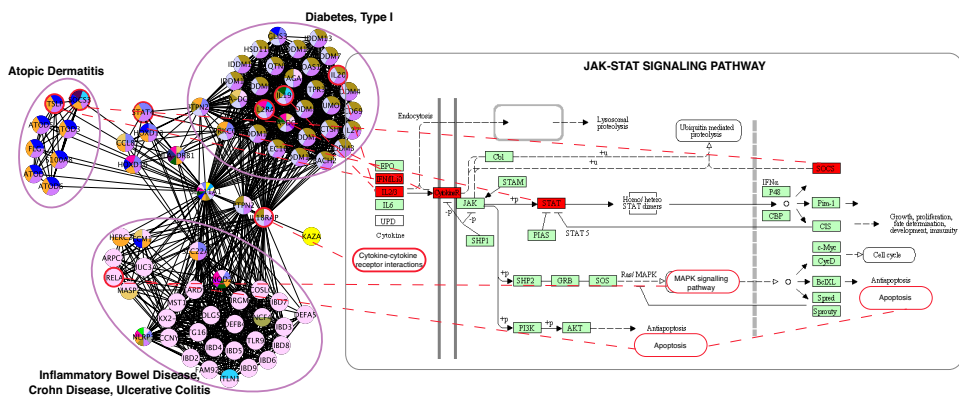


Figure 6: Identification of shared mechanisms of distinct diseases

A cluster containing genes associated to distinct diseases is shown on the left part of the figure. There are three main disease groups, Atopic Dermatitis (an autoimmune skin disease), Diabetes Mellitus Type I (an early onset, insulin-dependent, autoimmune disease), and Inflammatory Bowel Diseases (including Crohn Disease and Ulcerative Colitis). Diseases are coloured according to their disease class (see Figure 8 in Supplementary Material). The most significantly enriched Jak-STAT signalling pathway is displayed with some nodes from the cluster coloured in red (right part).

Discussion

We presented a global analysis of a comprehensive database of gene-disease associations revealing a modular nature of the whole spectrum of human genetic diseases. We compiled our database by integrating data from various expert-curated sources and text-mining derived associations. The overlap among databases is surprisingly small, which highlights the need of integrating different data sources to obtain a more complete picture of the current knowledge of gene-disease associations. This is of special interest considering the vast amount of literature published in the biomedical field that precludes individual researchers and even expert curators from keeping track of all accessible knowledge. Moreover, distinct databases set their focus differently and hence cover different aspects of genotype-

phenotype relationships.

We furthermore represented gene-disease associations as networks to obtain an overall overview of the genetic origin of human diseases and to subsequently study global properties of the data by means of network analysis. The topological analysis revealed several trends, which provide interesting new insights concerning the understanding of human diseases. First, the more data sources we consider, the denser the networks become indicating that the aforementioned knowledge pockets exist and that an integration of diverse repositories is required to bridge the gap between them. Second, the degree distributions of diseases and genes show a highly populated right tail differing from the typical degree distribution of random networks. This supports the idea that many diseases cannot be attributed to a single gene but to several genes that might interact in complex ways leading to the disease. Interestingly, the degree distributions of our networks, though different from random, do not follow power-law. Although many early studies on topology of biological networks proposed power-law behaviour, recent re-evaluation of these studies indicate that this is not always the case and other models need to be considered (Lima-Mendez and Helden, 2009).

The small number of associated genes in OMIM (on average 1.5 genes per disease) can be partly explained as OMIM was formerly built as a database focusing on mendelian, monogenic diseases and just recently started including complex diseases. However, the increase in the number of associated genes per diseases (2.7 in CURATED, 18.7 in LHGDN, 10.1 in ALL) is in agreement with discussions about the complexity of human genetic diseases, even in case of monogenic traits (Dipple *et al*, 2000; Scriver *et al*, 1999). These studies suggest that the distinction between mendelian and multiple gene disorders is rather artificial and that the influence of a variety of genes, including the so-called modifier genes, and environmental factors, cannot be neglected. Moreover, when studying the relationships among diseases in the disease projection networks, we observe that the average number of diseases connected to other diseases increases, suggesting a higher level of interrelation between human diseases than observed by solely considering a single data source (e.g. OMIM). This result points out that the genetic origin of mendelian, complex and environmental diseases is much more common than expected.

For single diseases, the functional analysis showed that homogeneity values (GO-BP and pathway) are significantly higher than for random controls. This shows that genes related to the same disease are more likely to be involved in the same biological processes than randomly selected disease genes. These findings support the concept of modularity of human diseases not only for mendelian diseases, as already shown by other authors (Freudenberg *et al*, 2002; Goh *et al*, 2007; Lim *et al*, 2006; Oti *et al*, 2007; van Driel *et al*, 2006), but also for complex and environmental diseases which are all integrated in our database. Interestingly, similar findings were obtained for disease clusters. This implies that there are biological processes common to groups of different diseases, which might represent shared underlying mechanisms of the groups of different diseases.

Intriguingly, we found an indirect correlation between the number of disease

related gene products and the homogeneity values. This shows that the number of biological processes that are related to a disease or a disease cluster increases with increasing number of gene products. A more detailed analysis of disease clusters revealed that the majority has medium pathway and GO-BP homogeneity values and there are only few extremely homogeneous or heterogeneous disease clusters. Overall, the results indicate that with the exception of a small set of diseases, for most of the diseases there are at least two biological processes or pathways associated. Similar results were recently reported by (Li *et al*, 2009) showing on average 12 pathways associated to a disease. Moreover, there is evidence that for many human diseases more than just one pathway, a set of so-called core pathways, are playing an important role such as the case for Pancreatic Cancer (Jones *et al*, 2008) or Glioblastoma (Cerami *et al*, 2010). Hence, for several diseases or groups of diseases, for example those with more genes associated and thus lower homogeneity values, cross-talks of pathways could play an important role. For instance, the cross-talk between Integrin and TGF- β pathways has been found to be related to several human pathologies including systemic sclerosis, idiopathic pulmonary fibrosis, chronic obstructive pulmonary disease and cancer (Margadant and Sonnenberg, 2010).

Regarding the phenotypically derived gene clusters, an indirect correlation between homogeneity values and cluster sizes was also observed. However, for very large clusters the difference to randomly generated clusters is not statistically significant. For the majority of gene clusters, pathway and GO-BP homogeneity values lie in a medium range indicating that more than just one biological process is associated to each set of phenotypically related genes.

We furthermore introduced another measurement, the HINscore, to identify gene clusters in which the proteins encoded by the genes physically interact and hence, the associated diseases might be related to dysfunction of whole protein machineries. Clusters with high HINscore are homogeneous in terms of disease association and also show high pathway and GO-BP homogeneity values. However, the percentage of clusters with high homogeneity or HINscore values is small and for most clusters these values are within medium range suggesting that only few diseases are solely caused by defects in protein complexes. Clusters with very low homogeneity values (< 0.25) are heterogeneous in terms of disease annotation, and are underrepresented in the dataset. Interestingly, such clusters contain genes with very high allelic heterogeneity. In CURATED, for instance, genes having more than 20 associated diseases belong to heterogeneous clusters with low GO-BP and pathway homogeneity values. It could be argued that such genes encode multifunctional proteins, and mutations in these proteins affecting different functions can then lead to different disease phenotypes. This set of genes might be classified as pleiotropic genes (Chavali *et al*, 2010) or represent genes that “moonlight” between different functions (Huberts and van der Klei, 2010). It would be interesting to further investigate the role of these proteins with respect to disease development.

There are some limitations in our analysis, which we would like to mention. Although we provide a comprehensive database of gene-disease associations, it is

THESIS PUBLICATIONS

not complete due to natural limitations in the curation process of the original databases, and might contain inaccurate associations derived from text-mining. Moreover, incomplete annotation of genes to GO terms and biological pathways is another issue. Only about half of the disease genes are annotated to pathways or appear in HIN. Also, the pathway databases suffer from annotation issues such as incomplete coverage of cross-talks, and the integration of pathways from different databases is still not fully achieved (Bauer-Mehren *et al*, 2009b).

Even taking into account the aforementioned limitations, to the best of our knowledge this is the first analysis of human genetic diseases including mendelian, complex and environmental diseases at a global scale. We observe good quality of text-mining derived associations, as values for LHGDN are comparable to the networks derived from expert-curated databases. We also demonstrate how the integration of text-mining derived gene-disease associations can close knowledge gaps found in the curated databases such as shown in Example 1. For instance, the association between MITF, a transcription factor regulating the expression of TYR gene, and Melanoma (Carreira *et al*, 2006; Garraway *et al*, 2006) was not found in any of the curated databases but was present in the text-mining derived network.

Finally, our results point out that for most diseases not a single but several biological processes might be affected. We believe that this has important implications for disease treatment and drug development. If a disease is associated to several pathways, a therapy considering the diversity of biological processes could be of advantage. And if a set of diseases is related to the same pathways, a treatment already successful for one of the diseases could also be applied to the other diseases (Berger and Iyengar, 2009).

In addition to the topological analysis of the networks, we identified the core biological processes related to each phenotypically derived gene cluster. In several exemplary use cases we demonstrated the value of such analysis to unveil biological processes related to diseases in order to gain a better understanding of the mechanism underlying them. We explored a variety of scenarios to (i) determine candidate disease genes, (ii) discover associations between environmental factors and diseases at the genetic level, (iii) identify shared mechanisms of different diseases and (iv) exploit the data to shed light on drug adverse reaction mechanism. Nevertheless, many other applications are feasible such as the identification of potential new drug targets and therapies, drug repurposing (Berger *et al*, 2009), prediction of disease comorbidity (Park *et al*, 2009), and prediction of candidate disease genes (Kann, 2010). All these approaches strongly depend on data quality and coverage. Hence, the use of the here presented unified gene-disease association database can provide important new biological insights, which might not be discovered when considering each of the single data repositories independently.

In summary, we provide a comprehensive database of gene-disease associations covering mendelian, complex and environmental diseases, as well as a detailed analysis on the modularity of the whole spectrum of human diseases at a global scale. Moreover, we make all data publicly available through DisGeNET, a plugin for Cytoscape (Bauer-Mehren *et al*, 2010), with the aim of easing future studies on human diseases.

Materials and methods

Data sources

In this study we combine five repositories of gene-disease associations to generate a comprehensive view of human diseases with genetic origin.

OMIM: Online Mendelian Inheritance in Man (OMIM) focuses on inherited or heritable diseases. Gene-disease associations were obtained by parsing the mim2gene file for associations of type “phenotype” (data was downloaded from <ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene> on June, 6th 2009). All associations were labelled “phenotype” as provided in the mim2gene file and classified as *Marker* in our gene-disease association ontology (see Figure 1, Supplementary Material). In total, we obtained for 2198 distinct genes and 2473 distinct disease terms 3432 gene-disease associations. After mapping of disease vocabularies, the OMIM network contained 2417 distinct diseases.

UNIPROT: UniProt/SwissProt is a database containing curated information about protein sequence, structure and function. Moreover, it provides information on the functional effect of sequence variants and their association to disease. We extracted this information from UniProt/SwissProt release 57.0 (March 2009) as described in (Bauer-Mehren *et al.*, 2009a). All protein identifiers were converted to Entrez Gene identifiers in order to allow integration with the other data sources. All gene-disease associations were classified as *GeneticVariation*. UniProt provided 1746 distinct gene-disease associations for 1240 distinct genes and 1475 distinct diseases.

PHARMGKB: The Pharmacogenomics Knowledge Base (PharmGKB) is specialized on the knowledge about pharmacogenes, genes that are involved in modulating drug response. Genes are classified as pharmacogenes because they are (i) involved in the pharmacokinetics of a drug (how the drug is absorbed, distributed, metabolized and eliminated) or (ii) the pharmacodynamics of a drug (how the drug acts on its target and its mechanisms of action) (Altman, 2007). Hence, it covers less broadly human gene-disease associations but was found to be complementary to the other sources, as it contains some gene-disease associations not present in the other repositories. We downloaded the genes.zip, diseases.zip and relationships.zip from http://www.pharmgkb.org/resources/downloads_and_web_services.jsp on June 6th 2009 and parsed the files to extract gene-disease associations. We furthermore made use of the perl webservices to obtain all available annotations and supporting information. We included 1772 associations for 79 distinct genes and 261 distinct diseases. PharmGKB associations were classified as *Marker* if the original label was “Related” and as *RegulatoryModification* if the original label was “Positively Related” or “Negatively Related”.

CTD: The Comparative Toxicogenomics Database (CTD) contains manually curated information about gene-disease relationships with focus on understanding the effects of environmental chemicals on human health. We downloaded the CTD_gene_disease_relations.tsv file from <http://ctd.mdibl.org/downloads/> on June 2nd 2009 and parsed it for gene-disease associations of type “marker” or “therapeutic” (see <http://ctd.mdibl.org/help/glossary.jsp> for description of the

original labels). CTD includes associations from OMIM but with some differences (i) for some associations extra information such as cross-links to PubMed are available and (ii) some associations are missing in either of the two databases. Hence, we kept all available gene-disease associations from both sources. All CTD gene-disease associations were classified as *Marker* if the original label was “marker” and as *Therapeutic* if the original label was “therapeutic”. All cross-links to PubMed were kept. In total CTD data provided 6469 associations for 2702 distinct diseases and 3345 distinct genes.

LHGDN: The literature-derived human gene-disease network (LHGDN) is a text mining derived database with focus on extracting and classifying gene-disease associations with respect to several biomolecular conditions. It uses a machine learning based algorithm to extract semantic gene-disease relations from a textual source of interest. The semantic gene-disease relations were extracted with F-measures of 78 (see (Bundschuh *et al*, 2008) for further details). More specifically, the textual source utilized here originates from Entrez Gene’s GeneRIF (Gene Reference Into Function) database (Mitchell *et al*, 2003). This database represents a rapidly growing knowledge repository and consists of high-quality phrases created or reviewed by MeSH indexers. Hereby, the phrases refer to a particular gene in the Entrez Gene database and describe its function in a concise phrase. Using this textual repository for text mining has recently gained increasing attention, due to the high quality of the provided textual data in the GeneRIF database (Bundschuh *et al*, 2008; Lu *et al*, 2007; Rubinstein and Simon, 2005). LHGDN was created based on a GeneRIF version from March 31st, 2009, consisting of 414241 phrases. These phrases were further restricted to the organism *Homo sapiens*, which resulted in a total of 178004 phrases. We extracted all data from LHGDN and classified the original associations using our ontology. In total, LHGDN provided 59342 distinct gene-disease associations for 1850 diseases and 6154 distinct genes. The LHGDN is also available in the Linked Life Data Cloud (<http://linkedlifedata.com/sources>).

Generation of gene-disease networks

Gene-disease associations were collected from several sources. The source databases use two different disease vocabularies (MIM and MeSH). EntrezGene identifiers are used for genes (except for UniProt/SwissProt which uses UniProt identifiers). Moreover, the kind of association differs among the databases and ranges from the generic term *related* to more specific terms such as *altered expression*. In order to merge all gene-disease associations and to present them in one comprehensive gene-disease network, we (i) mapped UniProt identifiers to EntrezGene identifiers if necessary, (ii) mapped MIM to MeSH vocabulary if possible (see Mapping of disease vocabularies) and (iii) integrated associations through our gene-disease association ontology. We furthermore constructed four different gene-disease networks, OMIM (only containing OMIM data), CURATED (containing gene-disease associations of OMIM, UNIPROT, PHARMGKB or CTD), LHGDN (only containing text mining data) and ALL (containing all gene-disease associations). Our comprehensive database is available as sqlite database as well as through DisGeNET, a plugin for Cytoscape for visualization and analysis of

the gene-disease association networks (Bauer-Mehren *et al*, 2010). Moreover, we provide a Cytoscape session including all examples discussed in this article; all data is available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download>.

All gene-disease networks are represented as bipartite graphs. A bipartite graph has two types of vertices and the edges run only between vertices of un-like types (Newman, 2003). The bipartite graphs are multigraphs in which two vertices can be connected by more than one edge. In our networks, the multiple edges represent the multiple data sources reporting the gene-disease association. Bipartite graphs have two degree distributions, one for each of the two types of vertices. We generated two projections, one for the diseases and one for the genes using the igraph library in R (Gabor and Tamas, 2006). The projected graphs contain only vertices of the same kind (monopartite) and two nodes are connected if they share a neighbour in the original bipartite graph. Before calculating node degree distributions and projecting the networks, we simplified the graphs and removed multiple edges. Hence, nodes that are connected by multiple edges are only connected by one edge in the simplified graph. This simplification is needed in order to correctly run the projection as implemented in the igraph library. Moreover, the node degree in the simplified graphs represents the number of first neighbours.

Gene-disease association ontology

For a correct integration of gene-disease association data, we developed a gene-disease association ontology (see Figure 1 in Supplementary Material). We classified all association types as found in the original source databases into *Association* if there is a relationship between the gene/protein and the disease, and into *NoAssociation* if there is no association between a gene/protein and a certain disease (in other words, if there is evidence for the independence between a gene/protein and a disease). The different association types from the original databases were mapped to the ontology for a seamless integration. In this study, we only considered gene-disease associations of type *Association*. The ontology is available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download>.

Mapping of disease vocabularies and disease classification

We used the MeSH hierarchy for disease classification. The repositories of gene-disease associations use two different disease vocabularies, MIM terms for OMIM diseases (used by OMIM, UniProt, CTD) and MeSH terms (used by CTD, PharmGKB, LHGDN). We used the UMLS metathesaurus to map from MIM to MeSH vocabularies. This step was performed to merge disease terms representing the same disorder, thus reducing redundancy. We were able to map 497 MIM terms directly to MeSH using UMLS and we additionally mapped 23 MIM terms by using a string mapping approach. Briefly, we searched the UMLS metathesaurus for MeSH terms for which there is at least one synonym exactly matching one of the synonyms describing the MIM term of interest. The resulting 63 matched terms were manually checked and reduced to 23 terms. For disease classification, we considered all 23 upper level concepts of the MeSH tree branch C (Diseases), plus two concepts (“Psychological Phenomena and Processes” and “Mental Disorders”)

of the F branch (Psychiatry and Psychology). Moreover, we added one disease class “Unclassified” for all disease terms for which a classification was not possible. We categorized all diseases into one or more of the 26 possible disease classes. For MeSH disease terms we directly used its position in the MeSH hierarchy, for MIM disease terms that were not mapped to MeSH, we used the disease classification of (Goh *et al*, 2007). Then, we mapped their disease classification to the MeSH hierarchy and extended the mapping using a disease classification available at CTD (CTD_disease_hierarchy.tsv downloaded August, 8th 2009). In total, we were able to classify 3980 (98.39 %) diseases. The disease classification allows filtering and searching of the network restricted to disease class, all implemented within DisGeNET (Bauer-Mehren *et al*, 2010).

Graph clustering

We used a graph-clustering algorithm with edge weights to identify functional units in the disease and gene projection networks. We used a local installation of the MCL graph cluster algorithm (van Dongen, 2000), which had successfully been applied to protein family detection (Enright *et al*, 2002). We calculated edge weights as follows:

$$w_{e(v_1, v_2)} = \frac{\bigcap a_{v_1}, a_{v_2}}{\min(a_{v_1}, a_{v_2})}, w_{e(v_1, v_2)} \in]0, 1], \quad (1)$$

where $e(v_1, v_2)$ is the edge connecting vertices v_1 and v_2 and a_v is the number of annotations to vertex v (genes to disease nodes or diseases to gene nodes).

Edge weights range from zero to one (excluding zero), where one means that the two vertices share all annotations of the node with less annotation. The most critical parameter of the MCL is the inflation value, which has large impact on the number of clusters, cluster sizes and cluster densities. We run the MCL cluster algorithm on the gene and disease projection networks with different inflation values. For OMIM we chose an inflation value of 1.8 following the suggestion by (Brohee and van Helden, 2006). The CURATED network is much denser and thus we used a higher inflation value of 3.6 to ensure better granularity. For LGDH and ALL, an inflation value of 5.0, respectively 3.6 was chosen.

GO-BP and pathway homogeneity

We calculated GO and pathway homogeneity as first defined by (Goh *et al*, 2007) for (i) each disease separately, (ii) for the disease clusters resulting from graph clustering with edge weights, and (iii) the gene clusters. Homogeneity is defined as the maximum fraction of genes sharing the same biological annotation:

$$H_i = \max_j \left[\frac{n_i^j}{n_i} \right], \quad (2)$$

where n_i is the total number of genes in the disease, disease cluster or gene cluster (i) with annotations, and n_j^i is the number of genes sharing the same biological annotation (j).

GO annotation was downloaded October, 6th 2009 from (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz) and then restricted to the branch biological process. We used all annotations and did not restrict to evidence type. The pathway annotation was downloaded from KEGG (<ftp://ftp.genome.jp/pub/kegg/genes/organisms/hsa>) and Reactome (<http://www.reactome.org/download/index.html>) on November, 11th 2009.

To calculate random controls for pathway (see Figure 3) and GO-BP homogeneity for single diseases and disease clusters, we randomly sampled genes from the set of disease genes of the studied network with annotation to pathway or GO-BP. We then took the annotation of the corresponding gene products and calculated pathway and GO-BP homogeneity values. Random controls for gene clusters were obtained by randomly assigning genes to clusters while total number of clusters and original cluster sizes were maintained. Random sampling was repeated 10^4 times to reach statistical significance and averages were compared to real values. A two-sided Kolmogorov-Smirnov test was performed to calculate p-values for comparison of real and random homogeneity values. Moreover, binning of cluster sizes was performed to show dependence of cluster sizes and homogeneity values, for the bin-wise comparison of mean values, 95 % confidence intervals were calculated. For this purpose, we calculated the Pearson correlation coefficient between homogeneity values and number of associated gene products.

HINscore calculation

Cerami et al. recently published a human interaction network (HIN) based on protein-protein interaction data from HPRD and pathway data from Reactome, NCI/Pathway Interaction database and the MSKCC Cancer Cell map (Cerami *et al*, 2010). We used this network to evaluate if disease genes belonging to clusters were more likely to be connected in HIN than expected from randomly generated clusters. For this purpose, we calculated the HINscore for each gene cluster defined as:

$$HINscore_{cluster_i} = 1 - \frac{cc_{sg_{cluster_i}} - 1}{n - 1}, \quad (3)$$

where cc is the number of connected components of subgraph sg built using all nodes in $cluster_i$ connected by edges appearing in the human interaction network (HIN).

We compared the HINscores calculated for each gene cluster of the four networks with random controls. For the random controls, we randomly selected the same number of genes per cluster from the samples set consisting of all genes in the network of study being present in HIN. We repeated the randomization process 10^4 times to achieve statistical significance and took the average as random reference.

THESIS PUBLICATIONS

We display mean HINscore for different cluster sizes and 95 % confidence intervals to show statistically significant difference to random controls.

GO and pathway enrichment

For the functional enrichment analysis, we used the R package GOstat (Falcon and Gentleman, 2007) and calculated for each gene cluster in CURATED the enrichment of terms in each GO category (biological process, molecular function, cellular component), as well as enriched pathways (KEGG). As reference background we used the list of disease genes that have at least one term annotated. We applied conditional hypergeometric test using a p-value cut-off of 0.05 and restricted the result to terms for which there were at least two genes annotated to in the reference background. The annotation of gene ids to GO terms was taken from the annotation package “org.Hs.eg.db” based on data provided by Entrez Gene (<ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/>) with a date stamp of September 2009. Annotation to pathways was taken from “KEGG.db” with mappings to pathways from KEGG Genome of March 2009. We only calculated GO term and pathway enrichment for clusters containing more than 3 genes.

Acknowledgements

We thank Robert Castelo for useful discussions. This work was generated in the framework of the EU-ADR project co-financed by the European Commission through the contract no. ICT-215847 and the eTOX project from the European Community's Seventh Framework Program (FP7/2007-2013) for the Innovative Medicine Initiative under grant agreement no. 115002. The Research Unit on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB) and member of the COMBIOMED network. We thank the Departament d'Innovació, Universitat i Empresa (Generalitat de Catalunya) for a grant to author ABM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary Material

Supplementary information is available at the journal web site and at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html>. This includes the DisGeNET database and plugin, as well as the ontology, and a Cytoscape session including all the use cases of this article.

References

- Ahmad, N.N., Ala-Kokko, L., Knowlton, R.G., Jimenez, S.A., Weaver, E.J., Maguire, J.I., Tasman, W. & Prockop, D.J. Stop codon in the procollagen II gene (COL2A1) in a family with the Stickler syndrome (arthroophthalmopathy). *Proc. Natl. Acad. Sci.* **88**, 6624-6627 (1991).
- Aittokallio, T. & Schwikowski, B. Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.* **7**, 243-255 (2006).
- Altman, R.B. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat. Genet.* **39**, 426 (2007).

- Ananiadou, S., Pyysalo, S., Tsujii, J.i. & Kell, D.B. Event extraction for systems biology by text mining the literature. *Trends Biotechnol.* **28**, 381-390 (2010).
- Apweiler, R. *et al* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115-119 (2004).
- Barabási, A.-L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101-113 (2004).
- Bau, D.-T., Wang, T.-S., Chung, C.-H., Wang, A.S.S. & Jan, K.-Y. Oxidative DNA adducts and DNA-protein cross-links are the major DNA lesions induced by arsenite. *Environ. Health Perspect.* **110 Suppl 5**, 753-756 (2002).
- Bauer-Mehren, A., Furlong, L., Rautschka, M. & Sanz, F. From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. *BMC Bioinformatics* **10**, S6-S6 (2009a).
- Bauer-Mehren, A., Furlong, L.I. & Sanz, F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol.* **5**, 290 (2009b).
- Bauer-Mehren, A., Rautschka, M., Sanz, F. & Furlong, L.I. DisGeNET - a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *submitted* (2010).
- Berger, S.I. & Iyengar, R. Network analyses in systems pharmacology. *Bioinformatics* **25**, 2466-2472 (2009).
- Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, 228-237 (2003).
- Brohee, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**, 488-488 (2006).
- Bundschuh, M., Dejori, M., Stetter, M., Tresp, V. & Kriegel, H.-P. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics* **9**, 207-207 (2008).
- Butts, C.T. Revisiting the Foundations of Network Analysis. *Science* **325**, 414-416 (2009).
- C, F. Gene expression profiles in peripheral lymphocytes by arsenic exposure and skin lesion status in a Bangladeshi population. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1367-1375 (2006).
- Carreira, S., Goodall, J., Denat, L., Rodriguez, M., Nuciforo, P., Hoek, K.S., Testori, A., Larue, L. & Goding, C.R. Mitf regulation of Dial controls melanoma proliferation and invasiveness. *Genes Dev.* **20**, 3426-3439 (2006).
- Celik, I. *et al* Arsenic in drinking water and lung cancer: a systematic review. *Environ. Res.* **108**, 48-55 (2008).
- Cerami, E., Demir, E., Schultz, N., Taylor, B.S. & Sander, C. Automated Network Analysis Identifies Core Pathways in Glioblastoma. *PLoS ONE* **5** (2010).
- Chavali, S., Barrenas, F., Kanduri, K. & Benson, M. Network properties of human disease genes with pleiotropic effects. *BMC Systems Biology* **4**, 78 (2010).
- Chiou, H.Y., Hsueh, Y.M., Liaw, K.F., Horng, S.F., Chiang, M.H., Pu, Y.S., Lin, J.S., Huang, C.H. & Chen, C.J. Incidence of internal cancers and ingested inorganic arsenic: a seven-year follow-up study in Taiwan. *Cancer Res.* **55**,

THESIS PUBLICATIONS

- 1296-1300 (1995).
- Cokol, M., Iossifov, I., Weinreb, C. & Rzhetsky, A. Emergent behavior of growing knowledge about molecular interactions. *Nat. Biotechnol.* **23**, 1243-1247 (2005).
- D'Andrea, A.D. & Grompe, M. The Fanconi anaemia/BRCA pathway. *Nat. Rev. Cancer* **3**, 23-34 (2003).
- De Luca, L. *et al* Overview of emerging pharmacologic agents for acute heart failure syndromes. *Eur. J. Heart. Fail.* **10**, 201-213 (2008).
- Dipple, K.M. & McCabe, E.R. Modifier genes convert "simple" Mendelian disorders to complex traits. *Mol. Genet. Metab.* **71**, 43-50 (2000).
- Duerr, R.H. Update on the genetics of inflammatory bowel disease. *J. Clin. Gastroenterol.* **37**, 358-367 (2003).
- Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* **30**, 1575-1584 (2002).
- Falcon, S. & Gentleman, R. Using GStats to test gene lists for GO term association. *Bioinformatics* **23**, 257-258 (2007).
- Freudenberg, J. & Propping, P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18**, S110-115 (2002).
- Gabor, C. & Tamas, N. The igraph Software Package for Complex Network Research. *InterJournal Complex Systems*, 1695 (2006).
- Garraway, L.A. & Sellers, W.R. From integrated genomics to tumor lineage dependency. *Cancer Res.* **66**, 2506-2508 (2006).
- Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. & Barabási, A.-L. The human disease network. *Proc. Natl. Acad. Sci.* **104**, 8685-8690 (2007).
- Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-517 (2005).
- Hidalgo, C.A., Blumm, N., Barabási, A.-L. & Christakis, N.A. A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Comp. Biol.* **5** (2009).
- Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95-108 (2005).
- Huberts, D.H.E.W. & van der Klei, I.J. Moonlighting proteins: an intriguing mode of multitasking. *Biochim. Biophys. Acta* **1803**, 520-525 (2010).
- Jones, S. *et al* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science* **321**, 1801-1806 (2008).
- Kann, M.G. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief. Bioinform.* **11**, 96-110 (2010).
- Kennedy, J.A., Kiosoglous, A.J., Murphy, G.A., Pelle, M.A. & Horowitz, J.D. Effect of perhexiline and oxfenicine on myocardial function and metabolism during low-flow ischemia/reperfusion in the isolated rat heart. *J. Cardiovasc. Pharmacol.* **36**, 794-801 (2000).
- Klein, T.E. *et al* Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J.* **1**, 167-170 (2001).

- Lee, D.S., Park, J., Kay, K.A., Christakis, N.A., Oltvai, Z.N. & Barabási, A.L. The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci.* **105**, 9880-9885 (2008).
- Li, Y. & Agarwal, P. A Pathway-Based View of Human Diseases and Disease Relationships. *PLoS ONE* **4**, e4346 (2009).
- Lim, J. *et al* A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration. *Cell* **125**, 801-814 (2006).
- Lima-Mendez, G. & Helden, J.v. The powerful law of the power law and other myths in network biology. *Mol. Biosyst.* **5**, 1482-1493 (2009).
- Lu, Z., Cohen, K.B. & Hunter, L. GeneRIF QUALITY ASSURANCE AS SUMMARY REVISION. *Pac. Symp. Biocomput.*, 269-280 (2007).
- Margadant, C. & Sonnenberg, A. Integrin-TGF-beta crosstalk in fibrosis, cancer and wound healing. *EMBO Reports* **11**, 97-105 (2010).
- Mattingly, C.J., Rosenstein, M.C., Davis, A.P., Colby, G.T., Forrest, J.N. & Boyer, J.L. The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.* **92**, 587-595 (2006).
- Melkonieni, M. *et al* Autosomal Recessive Disorder Otospondylomegapiphysal Dysplasia Is Associated with Loss-of-Function Mutations in the COL11A2 Gene. *Am. J. Hum. Genet.* **66**, 368-377 (2000).
- Mitchell, J.A., Aronson, A.R., Mork, J.G., Folk, L.C., Humphrey, S.M. & Ward, J.M. Gene Indexing: Characterization and Analysis of NLM's GeneRIFs. *AMIA Annu. Symp. Pro.* **2003**, 460-464 (2003).
- Navarro Silvera, S.A. & Rohan, T.E. Trace elements and cancer risk: a review of the epidemiologic evidence. *Cancer Causes Control* **18**, 7-27 (2007).
- Newman, M.E.J. The structure and function of complex networks. *SIAM Review* **45**, 167-256 (2003).
- Oti, M. & Brunner, H.G. The modular nature of genetic diseases. *Clin. Genet.* **71**, 1-11 (2007).
- Oti, M., Snel, B., Huynen, M.A. & Brunner, H.G. Predicting disease genes using protein-protein interactions. *J. Med. Genet.* **43**, 691-698 (2006).
- Park, J., Lee, D.-S., Christakis, N.A. & Barabasi, A.-L. The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.* **5** (2009).
- Park, S., Bae, J., Nam, B.-H. & Yoo, K.-Y. Aetiology of cancer in Asia. *Asian Pac. J. Cancer Prev.* **9**, 371-380 (2008).
- Przulj, N., Corneil, D.G. & Jurisica, I. Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508-3515 (2004a).
- Przulj, N., Wigle, D.A. & Jurisica, I. Functional topology in a network of protein interactions. *Bioinformatics* **20**, 340-348 (2004b).
- Radosavljević, V. & Jakovljević, B. Arsenic and bladder cancer: observations and suggestions. *J. Environ. Health* **71**, 40-42 (2008).
- Rubinstein, R. & Simon, I. MILANO - custom annotation of microarray results using automatic literature searches. *BMC Bioinformatics* **6**, 12 (2005).
- Sankaranarayanan, R. & Boffetta, P. Research on cancer prevention, detection and management in low- and medium-income countries. *Ann. Oncol.* (2010).
- Scaldaferri, F. & Fiocchi, C. Inflammatory bowel disease: progress and current

THESIS PUBLICATIONS

- concepts of etiopathogenesis. *J. Dig. Dis.* **8**, 171-178 (2007).
- Scriver, C.R. & Waters, P.J. Monogenic traits are not simple: lessons from phenylketonuria. *Trends Genet.* **15**, 267-272 (1999).
- Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3** (2007).
- Smith, A.H., Hopenhayn-Rich, C., Bates, M.N., Goeden, H.M., Hertz-Picciotto, I., Duggan, H.M., Wood, R., Kosnett, M.J. & Smith, M.T. Cancer risks from arsenic in drinking water. *Environ. Health Perspect.* **97**, 259-267 (1992).
- Snead, M.P. & Yates, J.R.W. Clinical and molecular genetics of Stickler syndrome. *J. Med. Genet.* **36**, 353-359 (1999).
- Suthram, S., Dudley, J.T., Chiang, A.P., Chen, R., Hastie, T.J. & Butte, A.J. Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. *PLoS Comp. Biol.* **6** (2010).
- Thun, M.J., DeLancey, J.O., Center, M.M., Jemal, A. & Elizabeth, M.W. The global burden of cancer: priorities for prevention. *Carcinogenesis* **31**, 100-110 (2010).
- Tsuda, T., Babazono, A., Yamamoto, E., Kurumatani, N., Mino, Y., Ogawa, T., Kishi, Y. & Aoyama, H. Ingested arsenic and internal cancer: a historical cohort study followed for 33 years. *Am. J. Epidemiol.* **141**, 198-209 (1995).
- van Dongen, S., Centers for Mathematics and Computer Science, University of Utrecht, (2000).
- van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G. & Leunissen, J.A.M. A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* **14**, 535-542 (2006).
- Yang, C.-Y., Chang, C.-C. & Chiu, H.-F. Does arsenic exposure increase the risk for prostate cancer? *J. Toxicol. Environ. Health A* **71**, 1559-1563 (2008).
- Yildirim, M.A., Goh, K.-I., Cusick, M.E., Barabasi, A.-L. & Vidal, M. Drug-target network. *Nat. Biotechnol.* **25**, 1119-1126 (2007).
- Zaghloul, N.A. & Katsanis, N. Functional modules, mutational load and human genetic disease. *Trends Genet.* **26**, 168-176 (2010).

Supplementary material

Network analysis of an integrated gene-disease association database reveals functional modules in mendelian, complex and environmental diseases

Anna Bauer-Mehren¹, Markus Bundschuh², Michael Rautschka¹, Miguel A. Mayer¹, Ferran Sanz¹, Laura I. Furlong^{1,*}

¹ Research Programme on Biomedical Informatics (GRIB) IMIM, DCEX, Universitat Pompeu Fabra, C/Dr. Aiguader 88, 08003 Barcelona, Spain

² Institute for Computer Science, Ludwig-Maximilians-University Munich, Oettingenstr. 67, 80538 Munich, Germany

* To whom correspondence should be addressed.

1. Data integration

A comprehensive database on gene-disease associations was developed by integrating information from four repositories: Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al*, 2005), UniProt/SwissProt (UNIPROT) (The UniProt, 2010), Pharmacogenomics Knowledge Base (PHARMGKB) (Altman, 2007), and Comparative Toxicogenomics Database (CTD) (Mattingly *et al*, 2006). In addition, associations from a literature-derived human gene-disease network (LHGDN) (Bundschus *et al*, 2008) were included to increase the coverage of the database. For a correct integration of gene-disease association data, we developed a gene-disease association ontology (see Figure 1). The ontology is available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download>.

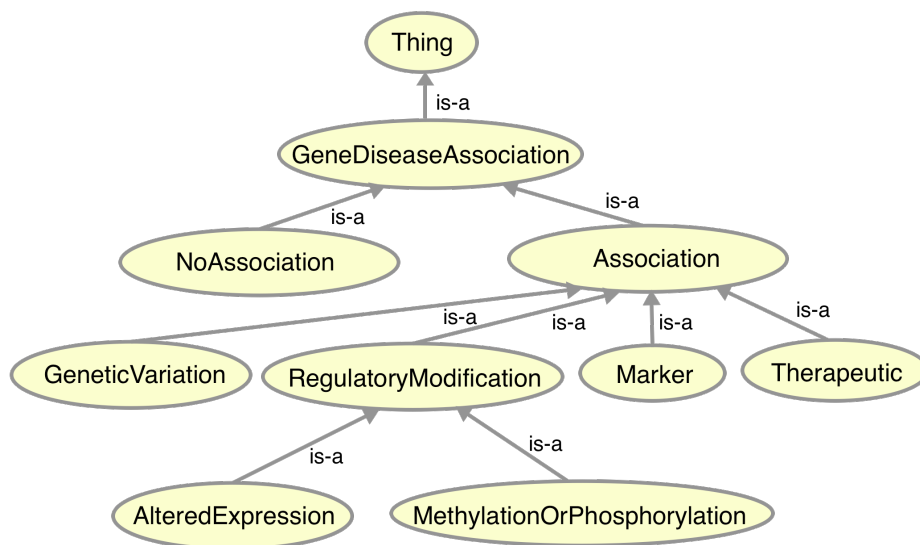


Figure 1: Gene-disease association ontology
 Gene-disease association ontology developed to allow correct integration of information from diverse repositories.

The integration of data from various sources allowed an increase in coverage of the resulting database. Moreover, the overlap among the different databases is surprisingly small. Hence data integration from diverse sources is needed to get a comprehensive picture of current gene-disease associations. Figure 2 shows the overlap among databases regarding diseases, genes and their associations.

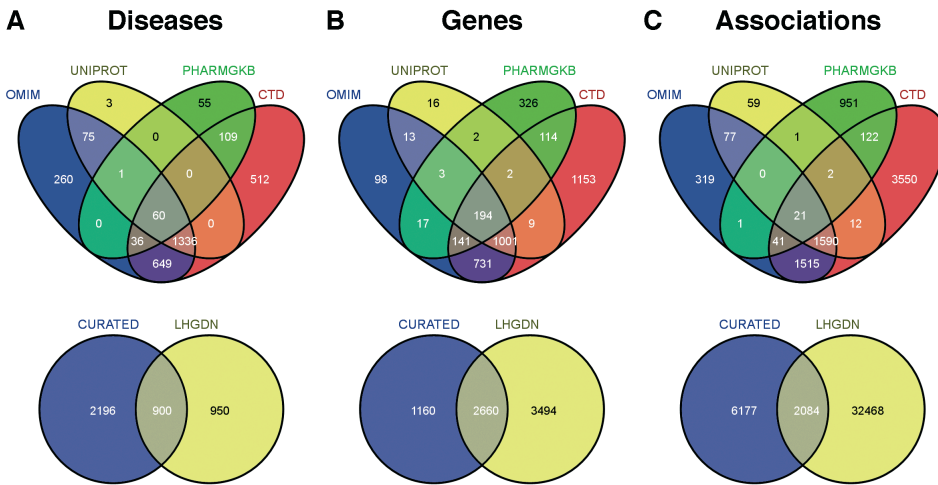


Figure 2: Venn diagrams of data overlap among databases

The upper panel shows the overlaps among the individual expert curated databases. The lower panel displays the overlap of CURATED and the text-mining derived network (LHGDN).

2. Network properties - node degree distributions

Studying the degree distribution of networks allows us to distinguish between different types of networks. For example, random networks show a typical peak corresponding to the average degree in the degree distribution.

In a bipartite graph there exist two degree distributions, one for each vertex type (disease and genes) (Newman, 2003). The first important observation is that the degree distributions for diseases and genes are different from degree distribution of random networks, but none of them follow a power law distribution.

For the diseases, the average node degree increases from 1.5 in OMIM to 10.1 in ALL. The node degree can be used as a measure of the locus heterogeneity of a given disease. There is a dramatic increase in the maximum locus heterogeneity observed in each data set, from 30 genes annotated to Diabetes Mellitus Type II in OMIM, 350 genes associated to Prostatic Neoplasms in CURATED, 1133 genes associated to Neoplasms in LHGDN and 1274 genes associated to Breast Neoplasms in ALL (see Figure 3). Interestingly, when considering the 10 top-ranking diseases in terms of locus heterogeneity, three diseases in OMIM belong to the “Neoplasm” disease class, 7 in CURATED and 10 in ALL. This may be due to the fact that cancer is one of the most studied diseases and hence more knowledge is available on the relationship of genes and different cancer types.

With respect to the genes, the increase in the node degree is less dramatic but still visible (from an average degree of 1.6 in OMIM to 5.6 in ALL). The degree of the gene in the bipartite graph can be used as a measure of the allelic heterogeneity (the number of diseases associated to a gene). In OMIM, collagen type II alpha 1

THESIS PUBLICATIONS

(COL2A1) has most disease annotations and there is another collagen, collagen type I alpha 1 in the list of the 10 top-ranked genes. In CURATED, collagen type II is in the top-ranked 30 genes but not for LHGDN or ALL. Moreover, the 10 top-ranked genes of OMIM and CURATED include some cancer related genes such as PTEN and TP53, which is also one of the genes with most disease associations in LHGDN and ALL. The list of the 10 top-ranked genes in CURATED includes cancer related genes (TNF, KRAS) but also many genes related to inflammation such as PTGS2 and IL6. In LHGDN and ALL the 10 top-ranked lists are very similar and contain mainly cancer related genes (TNF, TP53, TGFB1 and genes involved in immune system responses (IL6, IL10, IL1B).

We can also consider the degree distributions of the disease and gene projection networks. In contrast to the bipartite graph degree distribution, the degree of a gene (disease) node indicates the number of gene (disease) neighbours in the gene (disease) projection network. Interestingly, the degree distributions of the projected networks are much broader than the degree distributions of the bipartite graph (data not shown). The right tail of the distributions get much more populated the more data sources are included (more hubs in CURATED than in OMIM and again many more hubs in the LHGDN than in CURATED). Moreover, in the disease projection the average number of diseases connected to any disease is 2.2 in OMIM, 8.5 in CURATED and 103.6 in ALL, suggesting a higher degree of relatedness of human diseases than expected by solely considering a single data source (e.g. OMIM).

In summary, the degree distributions for diseases and genes are different from degree distribution of typical random networks, but none of them follows a power law. Moreover, there is a large dispersion of the right tail that is more evident the more data is incorporated into the networks. There is an increase in the average degree of the nodes, in the number of hubs and also in the degree of the hubs as a consequence of including more information in the network.

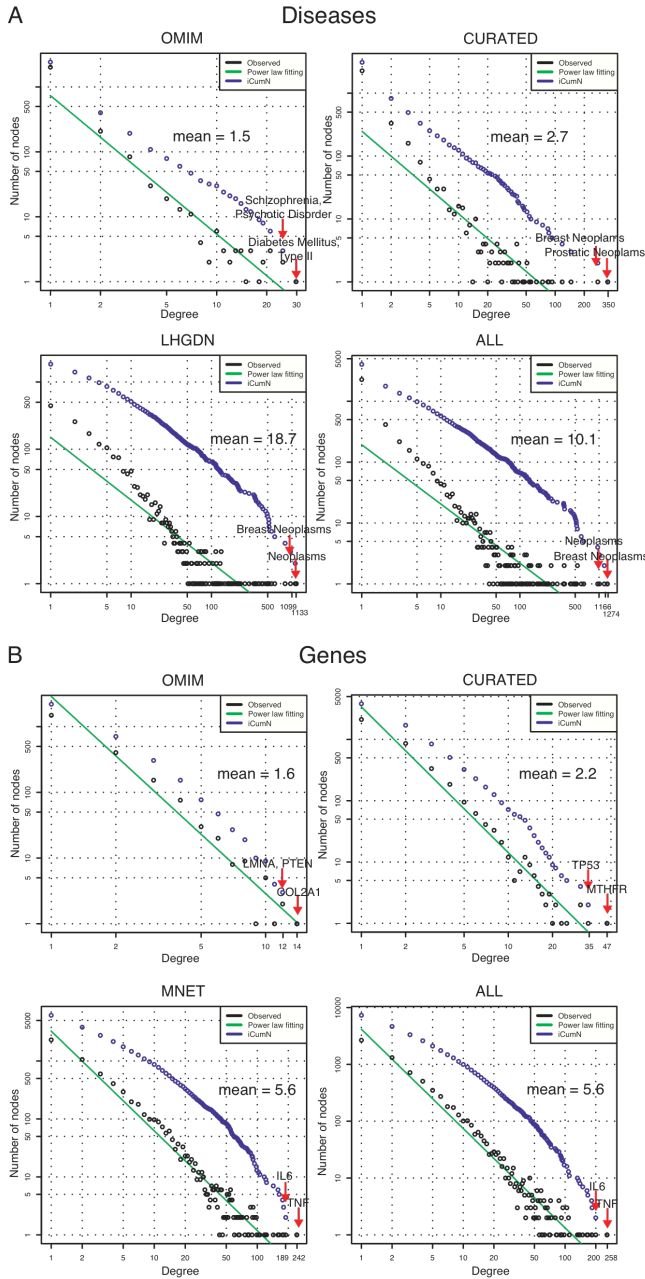


Figure 3: Degree distributions of the bipartite networks

The node degree distributions of the bipartite networks are plotted showing (A) the number of associated genes per disease and (B) the number of associated diseases per gene. Red arrows highlight the two disease- or gene-nodes with highest degree. Moreover, average degree values are plotted.

3. Functional analysis

3.1. GO homogeneity – individual diseases

It has been shown that, for OMIM diseases, the associated genes are involved in the same biological and cellular processes (Goh *et al*, 2007; Hartwell *et al*, 1999; Ravasz *et al*, 2002). In order to test if this concept still applies for our data set, we calculated GO term and pathway homogeneity for each disease. Mean GO-BP homogeneity values are plotted for different numbers of associated gene products and compared to random controls. First, for all data sources, the homogeneity values decrease with increasing number of associated gene products. Second, for all networks homogeneity values are significantly different (p -value < 0.05) from random control. For instance, in CURATED, diseases with two to five annotated gene products have on average 68 % of the gene products annotated to the biological process, while this value decreases to 38 % if there are between 50 and 100 gene products annotated to the disease. Moreover, it is striking that although the text-mining derived network is very dense with an average of 18.7 genes associated per disease, the GO-BP homogeneity still differs significantly from random. Similar results were obtained for pathway homogeneity analysis (Figure 3 in main text).

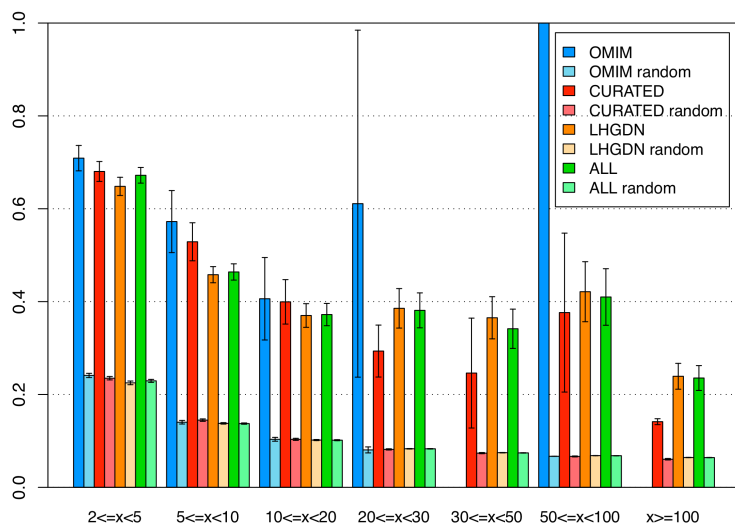


Figure 4: GO-BP homogeneity for individual diseases

Mean pathway homogeneity values for different number of associated gene products are plotted and compared to random controls (IC 95 %). For OMIM, there are only two diseases with more than 50 gene products annotated, both with a GO-BP homogeneity value of 1.

3.2. GO-BP and pathway homogeneity – disease clusters

We calculated GO-BP and pathway homogeneity as before for individual diseases for our disease clusters derived from graph clustering of the disease projection networks. Overall, we obtained similar results as for individual diseases. Figure 5 shows the average pathway homogeneity values of disease clusters plotted for different sizes of associated gene products. Similarly to individual diseases, pathway homogeneity decreases with increasing size of associated gene products. On average, pathway homogeneity for OMIM is 68 % and 59 % for CURATED suggesting that 60 - 70 % of the gene products belonging to a disease cluster participate in the same pathway. For the larger networks (LHGDN and ALL) the average pathway homogeneity values of disease clusters slightly decreases to approximately 50 %. All values are significantly different from random (p-value < 0.05). For complexity reason, GO-BP homogeneity values are not shown but are similar to pathway homogeneity.

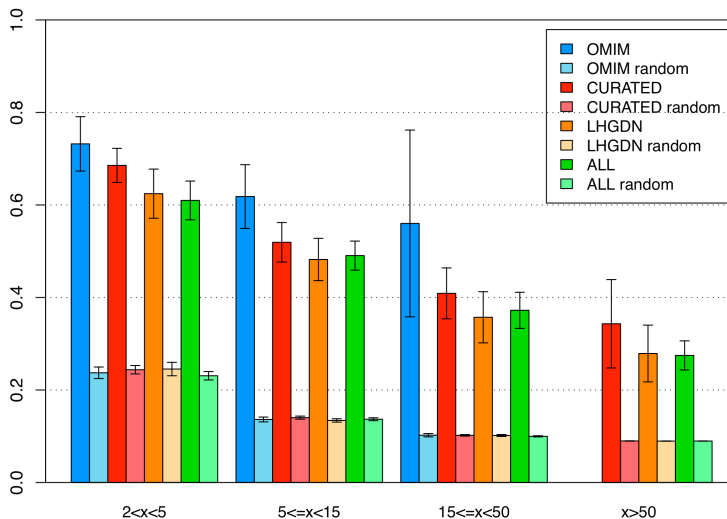


Figure 5: Pathway homogeneity – disease clusters

Mean pathway homogeneity values for different number of associated gene products are plotted and compared to random controls (IC 95 %).

3.3. GO-BP and pathway homogeneity – gene clusters

We calculated GO-BP and pathway homogeneity for gene clusters. Figure 6 and 7 show average GO-BP, respectively pathway homogeneity values for different cluster sizes. Here, the cluster size refers to the number of associated gene products of the cluster with annotation to GO-BP, respectively pathway. Up to cluster size 50, the average GO-BP homogeneity is significantly (p-value < 0.05) higher with respect to randomly selected clusters (except for ALL). On average, 72 % of the

clusters have a GO-BP homogeneity value larger than 0.5 or higher, hence more than half of the genes are annotated to the same GO-BP term.

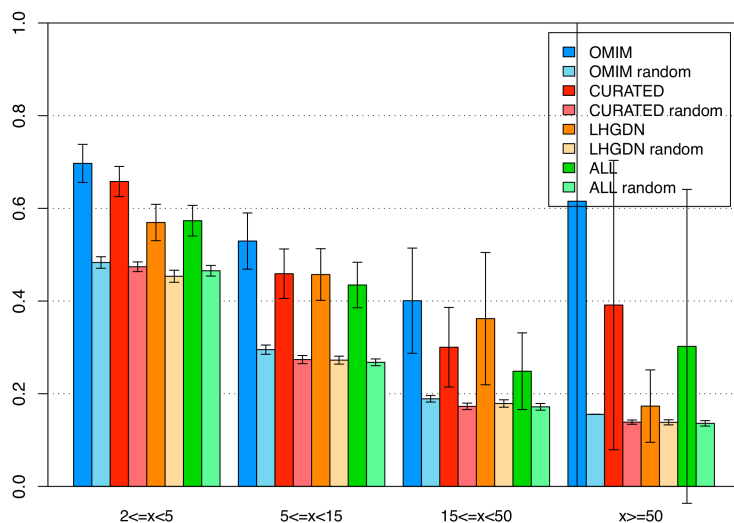


Figure 6: GO-BP homogeneity for gene clusters

Mean GO-BP homogeneity values for different number of associated gene products are plotted and compared to random controls (IC 95 %).

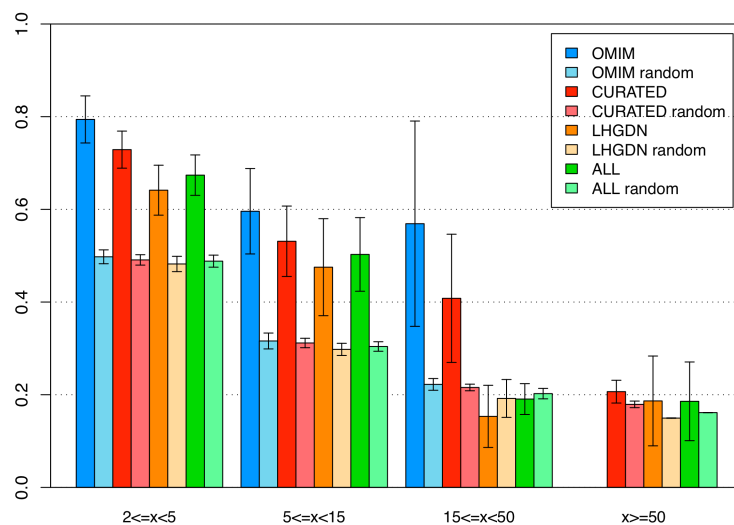


Figure 7: Pathway homogeneity for gene clusters

Mean pathway homogeneity values for different number of associated gene products are plotted and compared to random controls (IC 95 %).

4. Disease classification

Diseases were classified into 26 disease classes according to the MeSH hierarchy allowing the analysis of groups of related diseases based on standard disease classification. Many diseases are assigned to more than one disease class as several systems or organs are affected. Figure 8 shows the disease colour mapping used in DisGeNET (Bauer-Mehren *et al*, 2010). Disease and gene nodes can be coloured according to their disease class and can have multiple colours if they are assigned to more than one disease class.



























| DiseaseColourMapping \ | |
|---|---|
|  | C01 Bacterial Infections and Mycoses |
|  | C02 Virus Diseases |
|  | C03 Parasitic Diseases |
|  | C04 Neoplasms |
|  | C05 Musculoskeletal Diseases |
|  | C06 Digestive System Diseases |
|  | C07 Stomatognathic Diseases |
|  | C08 Respiratory Tract Diseases |
|  | C09 Otorhinolaryngologic Diseases |
|  | C10 Nervous system Diseases |
|  | C11 Eye Diseases |
|  | C12 Male Urogenital Diseases |
|  | C13 Female Urogenital Diseases and Pregnancy Complications |
|  | C14 Cardiovascular Diseases |
|  | C15 Hemic and Lymphatic Diseases |
|  | C16 Congenital, Hereditary, and Neonatal Diseases and Abnormalities |
|  | C17 Skin and Connective Tissue Diseases |
|  | C18 Nutritional and Metabolic Diseases |
|  | C19 Endocrine System Diseases |
|  | C20 Immune System Diseases |
|  | C21 Disorders of Environmental Origin |
|  | C22 Animal Diseases |
|  | C23 Pathological Conditions, Signs and Symptoms |
|  | F01 Behavior and Behavior Mechanisms |
|  | F03 Mental Disorders |
|  | NA Not Available |

Figure 8: Disease classes colouring

5. Gene annotations

We used annotation of genes to GO-BP, pathways and HIN. Table 1 shows the number of disease genes per network that actually have annotation to GO-BP, pathways or were part of HIN.

Table 1: GO and pathway annotation

| With annotation to | OMIM (2198) | CURATED (3820) | LHGDN (6154) | ALL (7314) |
|--------------------------------|----------------|-------------------|-----------------|---------------|
| GO-BP | 2117 | 3417 | 5704 | 6460 |
| Pathway (KEGG and Reactome) | 1249 | 2007 | 3271 | 3620 |
| HIN | 1628 | 2685 | 4670 | 5175 |

6. References

- Altman, R.B. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat Genet* **39**, 426 (2007).
- Bauer-Mehren, A., Rautschka, M., Sanz, F. & Furlong, L.I. DisGeNET - a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *submitted to Bioinformatics* (2010).
- Bundschuh, M., Dejori, M., Stetter, M., Tresp, V. & Kriegel, H.-P. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics* **9**, 207 (2008).
- Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. & Barabási, A.-L. The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-8690 (2007).
- Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-517 (2005).
- Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47-52-C47-52 (1999).
- Mattingly, C.J., Rosenstein, M.C., Davis, A.P., Colby, G.T., Forrest, J.N. & Boyer, J.L. The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci* **92**, 587-595 (2006).
- Newman, M.E.J. The structure and function of complex networks. *SIAM Review* **45**, 167-256 (2003).
- Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabási, A.L. Hierarchical organization of modularity in metabolic networks. *Science (New York, N.Y.)* **297**, 1551-1555 (2002).
- The UniProt, C. The Universal Protein Resource (UniProt) in 2010. *Nucl. Acids Res.* **38**, D1-D4 (2010).

3.4. Development of webservice and workflows for substantiation of drug safety signals

in preparation. 2010

THESIS PUBLICATIONS

Development of webservice and workflows for substantiation of drug safety signals

Introduction

Drug safety issues can arise during pre-clinical screening, clinical trials and, more importantly, after the drug is marketed and tested for the first time on the population (Giacomini *et al*, 2007a). Although relatively rare once a drug is marketed, drug safety issues constitute a major cause of morbidity and mortality worldwide. Every year about 2 million patients in the US are affected by a serious adverse drug reaction (ADR) resulting in approximately 100000 fatalities, ranking ADRs between the fourth and sixth cause of death in the US, not far behind cancer and heart diseases (Lazarou *et al*, 1998). Similar figures were estimated from other western countries (van der Hooft *et al*, 2006). Serious ADRs resulting from the treatment with thalidomide prompted modern drug legislation more than 40 years ago (Härmark and Grootheest, 2008). Over the past 10 years, 19 broadly used marketed drugs were withdrawn after presenting unexpected side effects (Giacomini *et al*, 2007a). The current and future challenges of drug development and drug utilization, and a number of recent high-impact drug safety issues (e.g. rofecoxib (Vioxx)) highlight the need of an improvement of safety monitoring systems (Olsson, 1998).

Moreover, unravelling the molecular mechanisms by which the ADR is elicited is of great relevance as well, since it has important implications in both public health and drug development. Understanding the molecular mechanisms of ADRs can be achieved by placing the drug adverse reaction in the context of current biomedical knowledge, which might explain it. Due to the huge amounts of data generated by the “omics” experiments, and the ever increasing volume of data and knowledge stored in databases and knowledge bases for studying ADRs, the application of bioinformatic analysis tools is essential in order to study and analyse ADRs.

Although the factors that determine the susceptibility to ADR are not completely well understood, accumulating evidence over the years indicate an important role of genetic factors (Chiang and Butte, 2009; Gurwitz and Motulsky, 2007; Wilke *et al*, 2007). Most of the ADRs are mechanistically related to drug metabolism phenomena, leading for instance to an unusual drug accumulation in the body (Gurwitz *et al*, 2007; Wilke *et al*, 2007). In addition, they can also be associated to inter individual genetic variants, most notably single nucleotide polymorphisms (SNPs), in genes encoding drug metabolising enzymes and drug target genes (Gurwitz *et al*, 2007; Wilke *et al*, 2007). One of the first ADRs explained by a genetic determinant was the inherited deficiency of the enzyme glucose-6-phosphate dehydrogenase causing severe anaemia in patients treated with the antimalarial drug primaquine (Giacomini *et al*, 2007b). Alternatively, an ADR can be caused by the interaction of the drug with a target different from the originally intended target (also known as anti-targets) (Ekins, 2004). A well known

THESIS PUBLICATIONS

example of an anti-target ADR is provided by aspirin, whose anti-inflammatory effect, exerted by inhibition of prostaglandin production by COX-2, comes at the expense of irritation of the stomach mucosa by its unintended inhibition of COX-1 (Kawai, 1998). Furthermore, in addition to mechanisms related to off-target pharmacology, it is becoming evident that ADRs may often be caused by the combined action of multiple genes (Gurwitz *et al*, 2007). The anticoagulant warfarin, which shows a varying degree of anticoagulant effects, is often associated with haemorrhages, and leads the list of drugs with serious ADR in the US and Europe. A 50 % of the variable effects of warfarin are explained by polymorphisms in the genes CYP2C9 and VKORC1, while the associated genes accounting to the remaining variability in the response to warfarin in the population are unknown (Chiang *et al*, 2009).

Other cases of ADRs may arise as a consequence of drug-drug interactions, or the interaction of the action of the drug with environmental factors (Chiang and Yu, 2003; Gurwitz *et al*, 2007). Indeed, the interplay between genotype and environment observed in several aspects of health and disease also extend to drug response and safety. For example, alcohol consumption and smoking are both associated with changes in the expression of the metabolic enzyme CYP2E1, therefore affecting the pharmacokinetics of certain drugs (Howard *et al*, 2003).

From the above paragraphs it is clear that the study of the molecular mechanisms underlying ADR requires achieving a synthesis of information across multiple disciplines. In particular, it requires the integration of information from a variety of knowledge domains, ranging from the chemical to the biological up to the clinical. Different resources cover information about these knowledge areas, and many of them are freely available on the web, such as biological and chemical databases and the literature. On the other side, new knowledge is produced continuously, and the list of the resources and published papers that a researcher interested in ADRs needs to cope with is turning more into a problem than into a solution. It has been already recognized that the adequate management of knowledge is becoming a key factor for biomedical research, especially in the areas that require traversing different disciplines and/or the integration of diverse and heterogeneous pieces of information (Ruttenberg *et al*, 2007). Furthermore, computational approaches are becoming critical for the translation of relevant discoveries into the clinical practice (Butte, 2008).

Knowledge management (KM) is the process of systematically capturing, representing and using information to develop an understanding on how a particular system works (Antezana *et al*, 2009). Most scientific information is present in unstructured format, such as free text in publications, and therefore can only be understood by humans. By manual curation of the literature, several databases have been developed to store information on different aspects of biomedical research, and are continuously updated (e.g. UniProt (Apweiler *et al*, 2004), PharmGKB (Klein *et al*, 2001)). However, the manual curation process is very difficult to scale up in order to cope with the vast amount of publications that are being generated on a daily basis, and in consequence there is a gap between the information stored in these expert curated databases and the information present in publications. Thus, to

be able to systematically extract information on a particular domain approaches for its automatic extraction from knowledge resources and texts are required. In the particular case of scientific publications, computer assisted approaches such as text-mining is needed (Ananiadou *et al*, 2006; Rzhetsky *et al*, 2008).

Once the data has been collected, it has to be organised, structured and stored in order to allow its subsequent analysis and interpretation by computational approaches. Data integration is one of the most important aspects of KM, and at the same time one of the most challenging areas of research in Computer Science (Antezana *et al*, 2009). Data integration in the life sciences has its own difficulties, as already discussed by several authors (Fisher and Henzinger, 2007; Louie *et al*, 2007; Philippi and Kohler, 2006). These problems are rooted in the inherent complexity of the biological domain, its high degree of fragmentation, the data deluge problem, and the widespread incidence of ambiguity in the naming of entities (Antezana *et al*, 2009). The latter is evident in the nomenclature of genes and proteins, which poses a challenge to text mining systems as well.

On the other hand, the computational analysis of several biomedical problems can only be addressed by using a variety of bioinformatic methods. An attractive approach that emerged in the last years is the combination of different bioinformatic analysis modules by means of processing pipelines or workflows (Oinn *et al*, 2004). This technology allows the integration of a variety of computational techniques into a processing pipeline in which data input and outputs are standardized. In the last years, this kind of integration has been greatly facilitated by the use of APIs and webservices allowing programmatic access to data repositories and analysis tools. For instance, PharmGKB, a database specialized on the knowledge about genes that are involved in modulating drug response (Klein *et al*, 2001), allows accessing the data by means of webservice clients. In this context, software has been developed that allows manually creating workflows without the need of programming skills. Taverna is one of such approaches that allows integration of different analysis modules, shared as web services, into a scientific workflow to perform *in silico* experiments (Oinn *et al*, 2004). It was especially developed for bioinformatic applications (Oinn *et al*, 2004). Similar approaches are also used for the processing of free text documents (<http://incubator.apache.org/uima/>) or for combining data mining methods (<http://www.knime.org/>).

Recent studies by several groups highlight the use of disparate data sets in the study of ADRs, enabled by bioinformatics methodologies. Combining the study of protein–drug interactions on a structural proteome-wide scale with protein functional site similarity search, small molecule screening, and protein–ligand binding affinity profile analysis, Xie and colleagues have elucidated a possible molecular mechanism for the previously observed, but molecularly uncharacterised, side effect of selective estrogen receptor modulators (SERMs). The results of this study show that the side effect involves the inhibition of the Sacroplasmic Reticulum Ca²⁺ ion channel ATPase protein transmembrane domain. The prediction provides molecular insight into reducing the adverse effect of SERMs and is supported by clinical and *in vitro* observations (Xie *et al*, 2007). In another series of studies, Huang and colleagues sought to determine the genetic variants

THESIS PUBLICATIONS

associated with the side effects caused by the drugs cisplatin (Huang *et al*, 2007), carboplatin (Huang *et al*, 2008) and daunorubicin (Duan *et al*, 2007). To this end, they analysed gene expression profiles in response to the different drug treatments on lymphoblastoid cell lines, for which there was SNP genotype data available from the HapMap project. The studies pinpointed SNPs associated with the cytotoxicity of the above mentioned drugs. Finally, Campillos and colleagues exploited the side effect information from prescription drug labels to identify novel molecular activities of existing drugs, information that can be used for drug re-purposing (Campillos *et al*, 2008). The Unified Medical Language System (UMLS) Metathesaurus was used as a vocabulary for the side effects, and a weighting scheme to account for the rareness and interdependence of side effects was developed. Since similarity in side effects correlated with shared targets between drugs, side effect similarity was used to predict novel targets between any two “unexpected” drug pairs. By combining side effect similarity with chemical similarity, 13 of 20 novel target predictions were validated, thereby identifying novel anti-target effects that could be used to derive novel indications for these drugs (Campillos *et al*, 2008).

These studies illustrate how computational approaches are paving the way toward elucidating the molecular mechanisms of ADRs. Here we present software tools for investigating the mechanisms of adverse drug reactions. In this regard, different web services were developed that can be combined into data processing workflows to achieve the signal substantiation.

Concept for the substantiation of drug-event pairs

The framework of this work is the EU-ADR project (Trifirò *et al*, 2009), which aims to develop an innovative computerized system to detect adverse drug reactions from electronic health records and biomedical databases to aid in the early detection of adverse drug reactions. The automatic mining of electronic healthcare record (EHR) databases will enable a more proactive alternative to drug safety monitoring, but at the same time, the huge increase in the number of potential signals presents a major challenge to evaluating and confirming significant drug-adverse event associations. Thus, another important goal of the project is, once a signal is detected, to provide a possible biological explanation for each signal. We refer to this as the *signal substantiation* process. The purpose of this substantiation process is to place the signals in the context of the current biomedical knowledge that might explain the signal. Essentially, we are searching for evidence that supports causal inference of the signal. The list of signals is assessed by investigating feasible paths that connect the drug and the adverse reaction involved in the signal. The general strategy is the automatic linkage of biomedical entities (drugs, proteins and their genetic variants, biological pathways, and clinical events) by means of a variety of bioinformatic approaches.

In pharmacovigilance, a signal is defined as an unexpected association of a clinical event with a given drug. According to the WHO definition (Edwards and Biriell, 1994), it refers to reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or

incompletely documented previously. However, for the sake of simplicity, from now on we refer to signals to any drug-event pair irrespective of the knowledge about its causal relationship.

One of the key aspects of the substantiation process is the knowledge about all the targets of the drugs. A drug can be regarded as an environmental factor, upon drug binding the target triggers cellular events that ultimately lead to certain phenotypic changes. Another aspect to be considered is the knowledge about the genetic basis of diseases, that is, all the genes that are known to lead to a disease phenotype. Thus, if the adverse effect elicited by a drug is similar to the phenotype observed in a genetic disease, we can propose that the drug acts on the same molecular processes that are altered in the disease.

Hence, the most simple scenario that can provide a causal inference of the signal is found if one of the targets of the drug is known to be directly associated to the clinical event. Here, we only present this simple scenario, however there are other paths connecting drug and event possible. For instance, it is well established that many adverse drug reactions are caused by altered drug metabolism for which genetic variants in metabolizing enzymes are often responsible. Consequently, another substantiation scenario involves assessing if the metabolites of the drug target proteins are known to be associated to the clinical event. Another scenario includes biological pathways. Here, the drug target is part of a molecular network, such as a signalling pathway, and its activity has an effect on another component of the pathway that is known to be directly associated with a disease phenotype. Thus, it could be argued that the drug triggers a signalling cascade that mimics the altered function of the protein associated to the disease. In the following, we present the webservice developed during this PhD thesis that will be part of an automatic signal substantiation pipeline. Moreover, we present one possible workflow in more detail.

Implementation of the substantiation concept

The view on the substantiation of signals was implemented in different software modules that can be combined in different ways to perform specific tasks for signal filtering and signal substantiation. We have implemented webservices and sub-workflows useful for signal substantiation. Here, we present one workflow in detail (see Figure 2). This workflow tries to establish a direct connection between the clinical event and the drug through a gene or protein, by identifying the proteins that are targets of the drug and are also associated with the event. In such a case it can be argued that the direct binding of the drug to the protein leads to the observed event phenotype. Associations between the event and proteins are found by querying our integrated gene-disease association database (Bauer-Mehren *et al*, 2010). As this database provides annotations of the gene-disease associations to the articles reporting the association and in case of text-mining derived associations even the exact sentence, the article or sentence can be studied in more detail in order to find a mechanistic explanation for the adverse event. It has to be mentioned that our gene-disease association database also contains information about genetic variants or SNPs and their association to diseases or adverse drug events.

Workflows

The workflow for signal substantiation was developed within Taverna, a tool to integrate resources that are shared as web services into a scientific workflow to perform *in silico* experiments (Oinn *et al*, 2004). Taverna provides several built-in functions, such as XML splitters or various local services for file handling, string manipulation, among others. Figure 1 shows the signal substantiation workflow. It uses three webservices that are described below. Moreover, it uses several XML splitters for parsing, as well as local services to remove duplicated entries from lists of entities and to find intersections between two lists of entities. In more detail, as input it uses an ATC code for the drug and the event. The following steps are performed in parallel. The SMILE of the drug is obtained using the *getSmileFromATC* service. Subsequently, the output of this service (the SMILE code of the drug) is passed to the *getUniprotListFromSmile* service in order to obtain all targets of this drug. At the same time, the event is mapped to CUI concept identifiers by the *eventToCuis* service. The CUI concept identifiers are subsequently processed by the *getDiseaseGenes* service, which provides a list of proteins associated to list of CUI concept identifiers representing the event. It is important to mention that both services (*getUniprotListFromSmile* and *getDiseaseGenes*) provide annotations to proteins represented by their UniProt identifiers. This allows intersecting the output of the two services. A XPath query (*XQuery_getTargets*) is used to parse the output of the two webservices for the two lists of proteins (list of drug-targets and list of proteins associated to the event). Using the built-in service *Remove String Duplicates* (namely *Script_removeDuplicatedGenes* and *Script_removeDuplicatedDrugTargets*) duplicated proteins are removed from the two lists. The local service *String List Intersection* is then used to intersect the two lists. The result is stored in an output document called *intersection*. It contains a list of proteins associated to the event and the drug. Moreover, a small beanshell script is used to provide a binary output, “YES” if there is at least one protein annotated to both drug and event, and “NO” if there are no proteins in the intersection list.

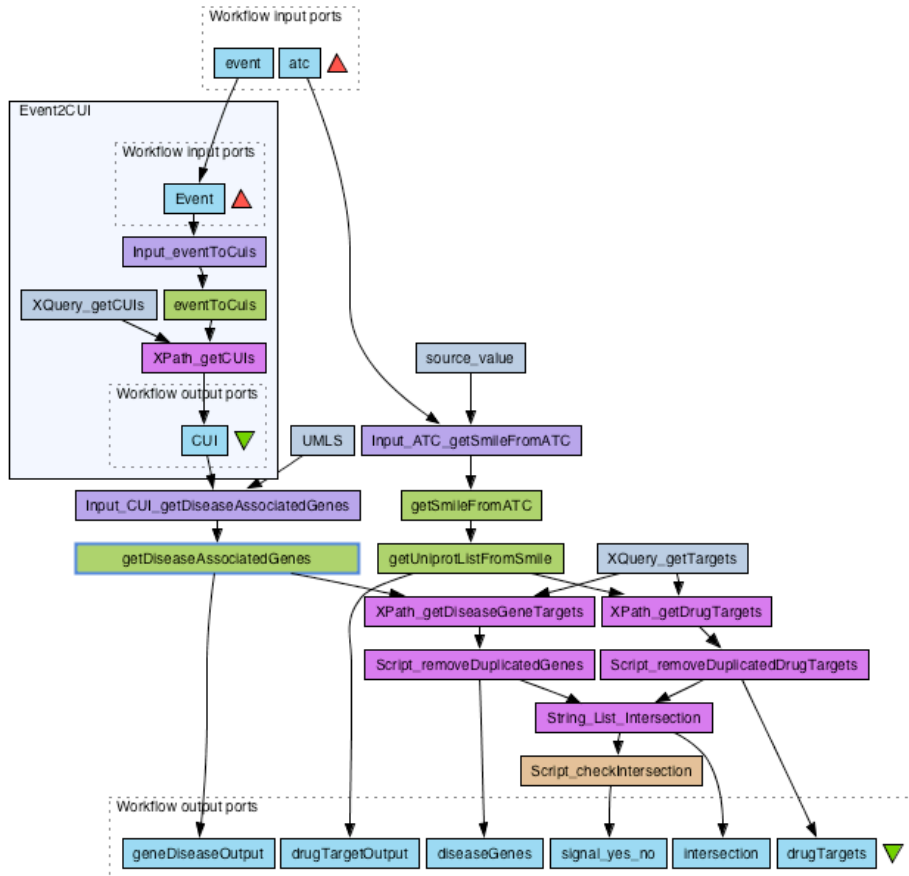


Figure 1: Tavana workflow for signal substantiation

The inputs are the event and the ATC code of the drug. The different modules run in parallel. The event is translated into CUI concept identifiers by the *eventToCuis* service. The CUI concept identifiers are input of the module *getDiseaseAssociatedGenes*, which returns relationships between the event and proteins (stored in the output element *geneDiseaseOutput*). The list of proteins is extracted by means of a XPath query (*XPath_getDiseaseGeneTargets*) and a Taverna built-in function to remove duplicates from string lists (*Script_removeDuplicatedGenes*). The drug ATC code is processed by the module *getSmileFromATC*, which returns the SMILE code of the drug. The SMILE code is further processed by the module *getUniprotListFromSmile*, which return the relationships between the drug and the targets (stored in the output element *drugTargetOutput*). The drug targets can be found in the output element *drugTargets*, obtained by using a XPath query and a Taverna built-in function to remove duplicates from string lists (*XPath_getDurgTargets* and *Script_removeDuplicatedDrugTargets*). The two lists, one containing proteins being associated to the event and one containing the drug-targets are then intersected (*Script_checkIntersection*) in order to find overlapping proteins. The list of intersection proteins is stored in an output element *intersection*. The final output of the workflow is a yes/no (*signal_yes_no*). A yes is returned if the list of intersecting proteins is not empty, a no if the list is empty.

Modules for Signal substantiation

In this section a description of the modules is provided. There are several webservices useful for generating workflows for signal substantiation. The `getDiseaseAssociatedGenes` webservice was implemented as part of this PhD thesis.

`getDiseaseAssociatedGenes`

This webservice provides for a given CUI concept identifier all associated proteins. In particular, it queries a new integrated database of human gene-disease associations, which also contains adverse events. Detailed information about the database can be found in (Bauer-Mehren *et al*, 2010). In brief, the integrated database combines information about gene-disease associations of expert-curated databases and text-mining derived associations. The webservice returns a list of relationships between the source CUI concept identifier and the target protein identifiers. Additionally, information is provided regarding the source repository reporting the association, the type of association (such as “Genetic Variation” and available literature evidence. Herein, the literature evidence is given by means of PubMed identifiers to the articles reporting the associations, and in case of text-mining derived associations the exact sentence stating the association. The WSDL for this service is available at <http://ibi.imim.es/axis2/services/AdrPathService?wsdl>.

`getSmileFromATC` and `getUniprotListFromSmile`

These webservices were developed by the CGL group and annotate a drug given by its ATC code to the protein targets. A detailed description can be found in (García Serna *et al*, 2010).

`eventToCuis`

This webservice was developed by Aveiro University. It annotates an adverse event with the according CUI concept identifiers of the UMLS metathesaurus. A detailed description will be reported elsewhere.

Discussion and conclusions

We have presented an exemplary workflow for the automatic substantiation of drug-event pairs. The system capitalizes on prior knowledge and uses state of the art bioinformatic approaches, such as *in silico* profiling methods, text mining and analysis and advanced data integration approaches. In the future, we plan to validate the framework by applying it to a test set of true positive and true negative signals. Moreover, we are working on the implementation of additional workflows using metabolite and pathway information.

Acknowledgements

This work was generated in the framework of the EU-ADR project co-financed by the European Commission through the contract no. ICT-215847. The Research programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB) and member of the COMBIOMED network. We thank the Departament d'Innovació, Universitat i Empresa (Generalitat de Catalunya) for a grant to author ABM. The authors also wish to thank the NLM for making UMLS available free of charge.

References

- Ananiadou, S., Kell, D.B. & Tsujii, J.-i. Text mining and its potential applications in systems biology. *Trends Biotechnol.* **24**, 571-579 (2006).
- Antezana, E.Z., Kuiper, M. & Mironov, V. Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief. Bioinform.* **10**, 392-407 (2009).
- Apweiler, R. *et al* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115-119 (2004).
- Bau, D.-T., Wang, T.-S., Chung, C.-H., Wang, A.S.S. & Jan, K.-Y. Oxidative DNA adducts and DNA-protein cross-links are the major DNA lesions induced by arsenite. *Environ. Health Perspect.* **110 Suppl 5**, 753-756 (2002).
- Bauer-Mehren, A., Bundschuh, M., Rautschka, M., Mayer, M.A., Sanz, F. & Furlong, L.I. Network analysis of an integrated gene-disease association database reveals functional modules in mendelian, complex and environmental disease. *submitted* (2010).
- Butte, A.J. Translational Bioinformatics: Coming of Age. *J. Am. Med. Inform. Assoc.* **15**, 709-714 (2008).
- Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L.J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263-266 (2008).
- Chiang, A.P. & Butte, A.J. Data-Driven Methods to Discover Molecular Determinants of Serious Adverse Drug Events. *Clin. Pharmacol. Ther.* **85**, 259-268 (2009).
- Chiang, J.H. & Yu, H.C. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics* **19**, 1417-1422 (2003).
- Duan, S., Bleibel, W.K., Huang, R.S., Shukla, S.J., Wu, X., Badner, J.A. & Dolan, M.E. Mapping Genes that Contribute to Daunorubicin-Induced Cytotoxicity. *Cancer Res.* **67**, 5425-5433 (2007).
- Edwards, I.R. & Biriell, C. Harmonisation in pharmacovigilance. *Drug Saf.* **10**, 93-102 (1994).
- Ekins, S. Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discovery Today* **9**, 276-285 (2004).
- Fisher, J. & Henzinger, T.A. Executable cell biology. *Nat. Biotechnol.* **25**, 1239-1249 (2007).
- García Serna, R., Carrascosa, M.C. & Mestres, J. 26, UPF, (2010).

THESIS PUBLICATIONS

- Giacomini, K.M., Krauss, R.M., Roden, D.M., Eichelbaum, M., Hayden, M.R. & Nakamura, Y. When good drugs go bad. *Nature* **446**, 975-977 (2007a).
- Giacomini, K.M., Krauss, R.M., Roden, D.M., Eichelbaum, M., Hayden, M.R. & Nakamura, Y. When good drugs go bad. **446**, 975-977 (2007b).
- Gurwitz, D. & Motulsky, A.G. 'Drug reactions, enzymes, and biochemical genetics': 50 years later. *Pharmacogenomics* **8**, 1479-1484 (2007).
- Härmark, L. & Grootheest, A.C. Pharmacovigilance: methods, recent developments and future perspectives. *Eur. J. Clin. Pharmacol.* **64**, 743-752 (2008).
- Howard, L.A., Miksys, S., Hoffmann, E., Mash, D. & Tyndale, R.F. Brain CYP2E1 is induced by nicotine and ethanol in rat and is higher in smokers and alcoholics. *Br. J. Pharmacol.* **138**, 1376-1386 (2003).
- Huang, R.S., Duan, S., Kistner, E.O., Hartford, C.M. & Dolan, M.E. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. *Mol. Cancer Ther.* **7**, 3038-3046 (2008).
- Huang, R.S., Duan, S., Shukla, S.J., Kistner, E.O., Clark, T.A., Chen, T.X., Schweitzer, A.C., Blume, J.E. & Dolan, M.E. Identification of Genetic Variants Contributing to Cisplatin-Induced Cytotoxicity by Use of a Genomewide Approach. *Am. J. Hum. Genet.* **81**, 427-437 (2007).
- Kawai, S. Cyclooxygenase selectivity and the risk of gastro-intestinal complications of various non-steroidal anti-inflammatory drugs: A clinical consideration. *Inflamm. Res.* **47**, 102-106 (1998).
- Klein, T.E. *et al* Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J.* **1**, 167-170 (2001).
- Lazarou, J., Pomeranz, B.H. & Corey, P.N. Incidence of Adverse Drug Reactions in Hospitalized Patients: A Meta-analysis of Prospective Studies. *JAMA* **279**, 1200-1205 (1998).
- Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. & Tarczy-Hornoch, P. Data integration and genomic medicine. *J. Biomed. Inf.* **40**, 5-16 (2007).
- Oinn, T., Addis, M., Ferris, J., Marvin, D., Greenwood, M., Carver, T., Pocock, M.R., Wipat, A. & Li, P. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045 - 3054 (2004).
- Olsson, S. The Role of the WHO Programme on International Drug Monitoring in Coordinating Worldwide Drug Safety Efforts. *Drug Saf.* **19**, 1-10 (1998).
- Peleg, M., Rubin, D. & Altman, R.B. Using Petri Net Tools to Study Properties and Dynamics of Biological Systems. *J. Am. Med. Inform. Assoc.* **12**, 181-199 (2005).
- Philippi, S. & Kohler, J. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat. Rev. Genet.* **7**, 482-488 (2006).
- Ruttenberg, A. *et al* Advancing translational research with the Semantic Web. *BMC Bioinformatics* **8**, S2 (2007).
- Rzhetsky, A., Seringhaus, M. & Gerstein, M. Seeking a New Biology through Text Mining. *Cell* **134**, 9-13 (2008).

- Trifirò, G., Fourrier-Réglat, A., Sturkenboom, M.C., Diaz Acedo, C., van der Lei, J. & group, E.-A. The EU-ADR project: preliminary results and perspective. *Stud. Health Technol. Inform.* **148**, 43-49 (2009).
- van der Hooft, C.S., Sturkenboom, M.C., van Grootheest, K., Kingma, H.J. & Stricker, B.H. Adverse drug reaction-related hospitalisations: a nationwide study in The Netherlands. *Drug Saf.* **29**, 161-168 (2006).
- Wilke, R.A., Lin, D.W., Roden, D.M., Watkins, P.B., Flockhart, D., Zineh, I., Giacomini, K.M. & Krauss, R.M. Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. *Nat. Rev. Drug Discov.* **6**, 904-916 (2007).
- Xie, L., Wang, J. & Bourne, P.E. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comp. Biol.* **3**, e217 (2007).

THESIS PUBLICATIONS

3.5. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks

Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI

submitted. 2010

THESIS PUBLICATIONS

Applications note

DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks

Anna Bauer-Mehren¹, Michael Rautschka¹, Ferran Sanz¹, Laura I. Furlong^{1,*}

¹ Research Programme on Biomedical Informatics (GRIB) IMIM, DCEXS, Universitat Pompeu Fabra, C/Dr. Aiguader 88, 08003 Barcelona, Spain

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

ABSTRACT

Summary: DisGeNET is a plugin for Cytoscape to query and analyze human gene-disease networks. DisGeNET allows user-friendly access to a new gene-disease database that we have developed by integrating data from several public sources. DisGeNET permits queries restricted to (i) the original data source, (ii) the association type, (iii) the disease class or (iv) specific gene(s)/disease(s). It represents gene-disease associations in terms of bipartite graphs and provides gene centric and disease centric views of the data. It assists the user in the interpretation and exploration of the genetic basis of human diseases by a variety of built-in functions. Moreover, DisGeNET permits multicolouring of nodes (genes/diseases) according to standard disease classification for expedient visualization.

Availability: DisGeNET is compatible with Cytoscape 2.6.3 and 2.7.0, please visit <http://ibi.imim.es/DisGeNET/DisGeNETweb.html> for installation guide, user tutorial and download.

Contact: lfurlong@imim.es

1 INTRODUCTION

One of the most challenging problems in biomedical research is to understand the underlying mechanisms of human diseases. Great effort has been spent on determining genes associated to diseases (Botstein and Risch, 2003; Kann, 2010). However, there is more and more evidence that most human diseases cannot be attributed to single genes but arise due to complex interactions between multiple genetic variants and environmental risk factors (Hirschhorn and Daly, 2005). Several databases have been developed storing associations between genes and diseases such as Online Mendelian Inheritance in Man (OMIM) (Hamosh, et al., 2005). Each of these databases focuses on different aspects of phenotype to genotype relationships. For instance, PharmGKB is specialized on how genetic variation is related to drug response (Altman, 2007), whereas the toxicogenomics database CTD stores information about the effect of environmental chemicals on human health (Mattingly, et al., 2006). Hence, integration of different databases is needed to allow a comprehensive view of the state of the art knowledge within this research field. It is widely established in bioinformatics to represent associations between biomedical entities as networks and to analyze their topology to get a

global understanding of underlying relationships (Butts, 2009; Goh, et al., 2007; Yildirim, et al., 2007). Cytoscape is a widely used Java-based, open-source software for networks visualization and analysis (Shannon, et al., 2003). The Cytoscape framework is extendable through the implementation of plugins. Up to now, a vast variety of plugins has been developed ranging from advanced network analysis tools to webservices. In the following, we present DisGeNET, a new Cytoscape plugin to query, integrate and visualize networks of human gene-disease associations.

2 OVERVIEW

2.1 The human gene-disease database

We compiled a comprehensive database of human gene-disease associations by integrating data from various expert curated databases and text-mining derived associations including mendelian, complex and environmental diseases (Bauer-Mehren, et al., 2010). We created bipartite graphs called OMIM, UNIPROT, PHARMGKB, CTD, CURATED (combining data from the curated databases), LHGDN (from the text-mining data only) and ALL (including all available gene-disease associations). Moreover, we calculated two network projections for each bipartite graph in order to generate disease and gene centric data representations. These projections allow an enhanced view on the genetic basis of complex diseases. We furthermore classified all diseases into one of 26 possible disease classes following the MeSH hierarchy (Bauer-Mehren, et al., 2010).

2.2 Gene-disease networks within Cytoscape

The gene-disease networks are bipartite graphs with two types of nodes (gene and disease) (Goh, et al., 2007; Newman, 2003). Gene and disease nodes are connected through edges if the according gene-disease association is covered in the gene-disease database. We allow displaying multiple edges between nodes, each representing a unique association found in the original data sources. Moreover, we colour the edges according to the association type following our gene-disease association ontology (Bauer-Mehren, et al., 2010). The disease and gene projection networks are monopartite graphs only containing either gene or disease nodes. Nodes are connected through edges if the two genes (diseases) share a disease (gene) in the bipartite gene-disease network. Thus, this representation allows studying diseases with similar genetic origin or genes associated to similar diseases. DisGeNET can be started from the plugins menu in Cytoscape. The main panel consists of three tabs, one for the gene-disease networks called "Gene Disease Network" and one for each projection, namely "Disease Projections" and "Gene Projections". The "Gene Disease Network" tab contains three drop-down menus allowing a restriction to (i) source, (ii)

*To whom correspondence should be addressed.

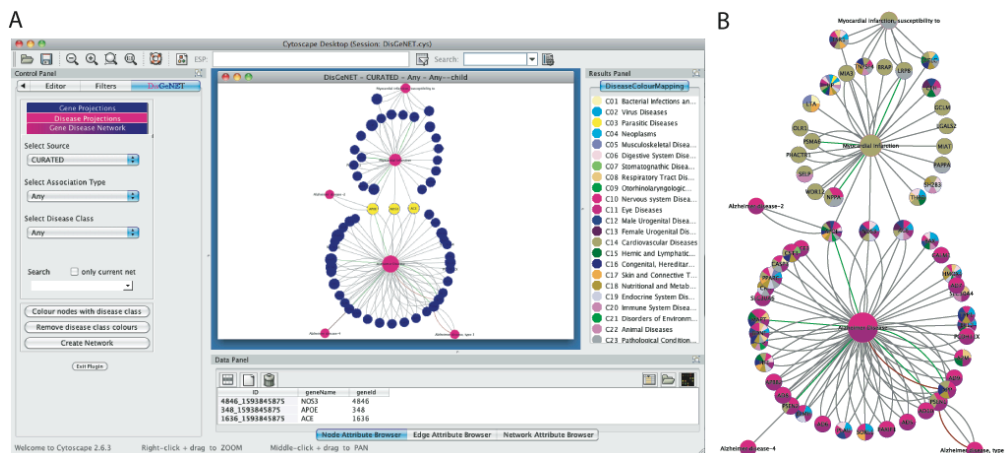


Figure 1: (A) Cytoscape screenshot of DisGeNET. The diseases *Alzheimer Disease* and *Myocardial Infarction* and their shared genes are displayed (in yellow). (B) The same network is shown with nodes colored according to the disease classes of the nodes.

association type and (iii) disease class. The two projection panels only contain two drop-down menu options to restrict the query to source and disease class. DisGeNET incorporates an advanced search function for each of the three network types. The user can search for a gene or a disease of interest and even for any set of diseases or genes by using the wild card symbol (asterisk). The search box can be either used to create new networks or to select nodes of already generated networks. DisGeNET makes use of the Cytoscape VizMapper to create visual styles for the networks. Gene nodes are coloured in blue and disease nodes in magenta (see Figure 1A). The node size increases with increasing number of associated diseases, respectively genes. Edges are coloured corresponding to the association type. Moreover, disease and gene nodes can be coloured according to the disease class by using the “Colour nodes with disease class” button. Since it is possible to have diseases and genes assigned to more than one disease class, multicolour pie charts can be overlaid onto (and removed from) nodes (see Figure 1B).

2.3 Use cases

Some exemplary use cases showing the utility of DisGeNET are available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#UserGuide>. They address problems such as (i) which are the genes annotated to breast neoplasm in expert-curated databases?, (ii) do comorbidities of alzheimer disease and myocardial infarction observed in patients reflect in a common genetic origin? or (iii) which are the diseases that are associated to post-translational modifications such as phosphorylation?

3 CONCLUSION

DisGeNET is a coherent tool for easy analysis and interpretation of human gene-disease networks. It allows user-friendly access to a comprehensive database comprising gene-disease associations for mendelian, complex and environmental diseases. DisGeNET displays gene-disease association networks as bipartite graphs and provides gene centric and disease centric views of the data. It assists the interpretation and exploration of human diseases with respect to their genetic origin. Diverse options for generating subnetworks, as well as an advanced search tool, facilitate not only the analysis of single diseases but also the study of sets of diseases or certain disease classes specified through their associated genes. Herein, the multicolouring of gene and disease nodes offers a convenient visualization of disease classifications in the networks.

We plan regular updates of the underlying gene-disease association database as well as the integration of further data sources.

4 ACKNOWLEDGEMENTS

This work was generated in the framework of the EU-ADR (no. ICT-215847) and the eTOX projects (no. 115002) co-financed by the European. The GRIB is a node of the Spanish National Institute of Bioinformatics and member of the COMBIOMED network. We thank the AGAUR for a grant to author ABM.

Conflict of interest: None declared.

REFERENCES

Altman, R.B. (2007) PharmGKB: a logical home for knowledge relating genotype to drug response phenotype, *Nat. Genet.*, **39**, 426.

Bauer-Mehren, A., Bundschuh, M., Rautschka, M., Mayer, M.A., Sanz, F. and Furlong, L.I. (2010) Network analysis of an integrated gene-disease association database reveals functional modules in mendelian, complex and environmental disease, *submitted*.

Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, *Nat. Genet.*, 228-237.

Butts, C.T. (2009) Revisiting the Foundations of Network Analysis, *Science*, **325**, 414-416.

Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabási, A.-L. (2007) The human disease network, *Proc. Natl. Acad. Sci.*, **104**, 8685-8690.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Res.*, **33**, D514-517.

Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits, *Nat. Rev. Genet.*, **6**, 95-108.

Kann, M.G. (2010) Advances in translational bioinformatics: computational approaches for the hunting of disease genes, *Brief. Bioinform.*, **11**, 96-110.

Mattingly, C.J., Rosenstein, M.C., Davis, A.P., Colby, G.T., Forrest, J.N. and Boyer, J.L. (2006) The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks, *Toxicol. Sci.*, **92**, 587-595.

Newman, M.E.J. (2003) The structure and function of complex networks, *SIAM Review*, **45**, 167-256.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.*, **13**, 2498-2504.

Yildirim, M.A., Goh, K.-I., Cusick, M.E., Barabasi, A.-L. and Vidal, M. (2007) Drug-target network, *Nat. Biotechnol.*, **25**, 1119-1126.



A Cytoscape plugin to visualize,
integrate, search and analyze
gene-disease networks

USER GUIDE

DisGeNET user guide

1. DisGeNET installation guide3

1.1. Download and install DisGeNET.....3

1.2. Troubleshooting5

1.2.1. Allocating more memory5

1.2.2. Download and installation problems.....5

2. DisGeNET database.....6

2.1. Original data sources.....6

2.2. Generation of gene-disease networks.....7

2.3. Mapping of disease vocabularies8

2.4. Gene-disease association ontology8

3. DisGeNET tutorial.....9

3.1. Basic functions9

3.1.1. Generate gene-disease association network9

3.1.2. Generate gene or disease projection network.....11

3.1.3. Restrict the network to a certain association type11

3.1.4. Restrict the network to a certain disease class.....12

3.1.5. Search for a particular gene/disease or set of genes/diseases.....12

3.1.6. DisGeNET LinkOut.....13

3.1.7. DisGeNET Expand.....13

3.1.7.1. Expand DisGeNET networks13

3.1.7.2. Expand foreign networks.....15

3.2. Specific use cases.....21

3.2.1. Which are the genes annotated to breast cancer in expert curated databases?21

3.2.1.1. Do comorbidities observed in patients reflect a common genetic origin of the diseases?.....24

3.2.2. Which are the diseases that are associated to post-translational modifications such as phosphorylation?.....27

3.3. Analyzing DisGeNET data using external tools28

3.3.1. Extract data from DisGeNET database28

3.3.2. Build networks using igraph library.....28

4. Contact.....29

4.1. Biomedical Informatics group.....29

4.2. Citation.....29

4.3. Acknowledgements30

4.4. Contact30

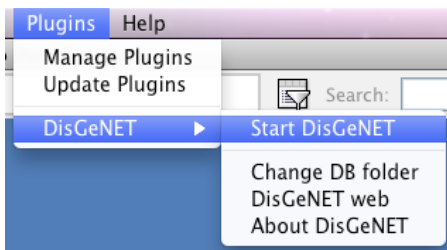
5. Attribute tables.....31

6. References32

1. DisGeNET installation guide

1.1. Download and install DisGeNET

- Download DisGeNET.jar from <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download>
- Put the jar (DisGeNET.jar) in the Cytoscape "plugins" folder.
(The default location in Windows is C:\Program Files\Cytoscape-v2.x\plugins).
The plugin will be automatically loaded the next time Cytoscape is started, and will appear as a menu item in the plugins menu. You can start the plugin by clicking on *Start DisGeNET*.

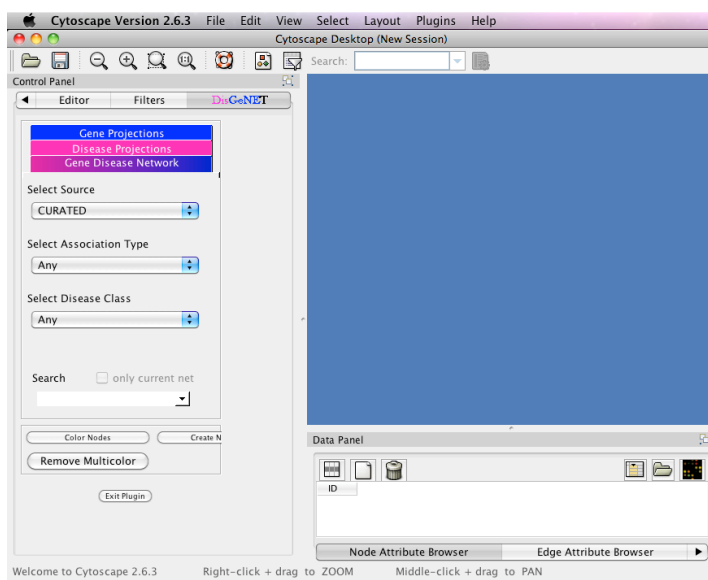
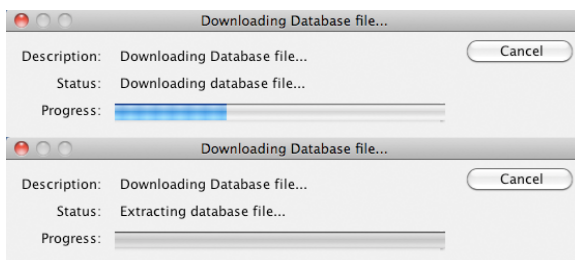


- The first time you start the plugin it will automatically download and unpack the gene-disease database (DisGeNET.db ~326,5MB) into a directory of your choice.

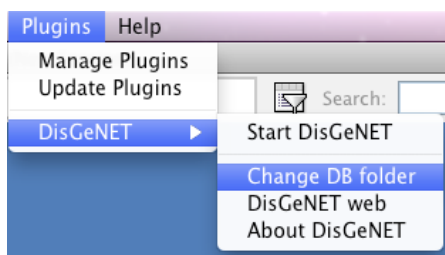


DisGeNET installation guide

- The download might take several minutes. When the download is finished, the plugin starts automatically.



- Now the plugin is ready to be used.



- The database folder can be changed at any time. Please restart the plugin to activate the changes.

1.2. Troubleshooting

1.2.1. Allocating more memory

Some of the networks are very large, especially when using LHGDN or ALL as source databases. In order to visualize large networks, you need to allocate more memory to Cytoscape. Memory usage depends on the number of nodes/edges and number of attributes. For detailed information check the Cytoscape manual available at <http://www.cytoscape.org/>. For Cytoscape version 2.7.0, you can find the information here: http://www.cytoscape.org/manual/Cytoscape2_7Manual.html#How%20to%20increase%20memory%20for%20Cytoscape

1.2.2. Download and installation problems

- Make sure you have writing permission for the Cytoscape subfolders
- Download is interrupted with NullPointerException (in Linux or Mac OSX)
 - ➔ Instead of starting Cytoscape via the icon, try to start it via command line from the installation folder, e.g.:
`sh /Applications/Cytoscape-6.x/cytoscape.sh`

2. DisGeNET database

The DisGeNET database integrates human gene-disease associations from various expert curated databases and text-mining derived associations including mendelian, complex and environmental diseases (Bauer-Mehren, et al., 2010). The integration is performed by means of gene and disease vocabulary mapping and by using a gene-disease association ontology as described below.

2.1. Original data sources

OMIM: Online Mendelian Inheritance in Man (OMIM) focuses on inherited or heritable diseases (Hamosh, et al., 2005). Gene-disease associations were obtained by parsing the mim2gene file for associations of type “phenotype” (data was downloaded from <ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene> on June, 6th 2009). All associations were labelled “phenotype” as provided in the mim2gene file and classified as Marker in our gene-disease association ontology. In total, we obtained 2198 distinct genes and 2473 distinct disease terms resulting in 3432 gene-disease associations. After mapping of disease vocabularies, the OMIM network contained 2417 distinct diseases.

UNIPROT: UniProt/SwissProt is a database containing curated information about protein sequence, structure and function (Apweiler, et al., 2004). Moreover, it provides information on the functional effect of sequence variants and their association to disease. We extracted this information from UniProt/SwissProt release 57.0 (March 2009) as described in (Bauer-Mehren, et al., 2009). All protein identifiers were converted to Entrez Gene identifiers in order to allow integration with the other data sources. All gene-disease associations were classified as GeneticVariation. UniProt provided 1746 distinct gene-disease associations for 1240 distinct genes and 1475 distinct diseases.

PHARMGKB: The Pharmacogenomics Knowledge Base (PharmGKB) is specialized on the knowledge about pharmacogenes, genes that are involved in modulating drug response. Genes are classified as pharmacogenes because they are (i) involved in the pharmacokinetics of a drug (how the drug is absorbed, distributed, metabolized and eliminated) or (ii) the pharmacodynamics of a drug (how the drug acts on its target and its mechanisms of action) (Altman, 2007). Hence, it covers less broadly human gene-disease associations but was found to be complementary to the other sources, as it contains some gene-disease associations not present in the other repositories. We downloaded the genes.zip, diseases.zip and relationships.zip from http://www.pharmgkb.org/resources/downloads_and_web_services.jsp on June 6th 2009 and parsed the files to extract gene-disease associations. We furthermore made use of the perl webservices to obtain all available annotations and supporting information. We included 1772 associations for 79 distinct genes and 261 distinct diseases. PharmGKB associations were classified as Marker if the original label was “Related” and as RegulatoryModification if the original label was “Positively Related” or “Negatively Related”.

CTD: The Comparative Toxicogenomics Database (CTD) contains manually curated information about gene-disease relationships with focus on understanding the effects of environmental chemicals on human health (Mattingly, et al., 2006). We downloaded the CTD_gene_disease_relations.tsv file from <http://ctd.mdibl.org/downloads/> on June 2nd 2009 and parsed it for gene-disease associations of type “marker” or “therapeutic” (see <http://ctd.mdibl.org/help/glossary.jsp> for description of the original labels). CTD includes associations from OMIM but with some differences (i) for some associations extra information such as cross-links to PubMed are available and (ii) some associations are missing in either of the two databases. Hence, we kept all available gene-disease associations from both sources. All CTD gene-disease associations were classified as Marker if the original label was “marker” and as Therapeutic if the original label was “therapeutic”. All cross-links to PubMed were kept. In total CTD data provided 6469 associations for 2702 distinct diseases and 3345 distinct genes.

LHGDN: The literature-derived human gene-disease network (LHGDN) is a text mining derived database with focus on extracting and classifying gene-disease associations with respect to several biomolecular conditions. It uses a machine learning based algorithm to extract semantic gene-disease relations from a textual source of interest. The semantic gene-disease relations were extracted with F-measures of 78 (see (Bundschus, et al., 2008) for further details). More specifically, the textual source utilized here originates from Entrez Gene's GeneRIF (Gene Reference Into Function) database (Mitchell, et al., 2003). This database represents a rapidly growing knowledge repository and consists of high-quality phrases created or reviewed by MeSH indexers. Hereby, the phrases refer to a particular gene in the Entrez Gene database and describe its function in a concise phrase. Using this textual repository for text mining has recently gained increasing attention, due to the high quality of the provided textual data in the GeneRIF database (Bundschus, et al., 2008; Lu, et al., 2007; Rubinstein and Simon, 2005). LHGDN was created based on a GeneRIF version from March 31st, 2009, consisting of 414241 phrases. These phrases were further restricted to the organism *Homo sapiens*, which resulted in a total of 178004 phrases. We extracted all data from LHGDN and classified the original associations using our ontology. In total, LHGDN provided 59342 distinct gene-disease associations for 1850 diseases and 6154 distinct genes. The LHGDN is also available in the Linked Life Data Cloud (<http://linkedlifedata.com/sources/>).

2.2. Generation of gene-disease networks

Gene-disease associations were collected from several sources. The source databases use two different disease vocabularies (MIM and MeSH). Entrez Gene identifiers are used for genes (except for UniProt/SwissProt which uses UniProt identifiers). Moreover, the kind of association differs among the databases and ranges from the generic term "related" to more specific terms such as "altered expression". In order to merge all gene-disease associations and to present them in one comprehensive gene-disease network, we (i) mapped UniProt identifiers to EntrezGene identifiers if necessary, (ii) mapped MIM to MeSH vocabulary if possible (see Mapping of disease vocabularies) and (iii) integrated associations through our gene-disease association ontology (see Gene-disease association ontology). We furthermore constructed different gene-disease networks for each source (OMIM, UNIPROT, PHARMGKB, CTD, LHGDN), as well as two integrated networks CURATED (containing gene-disease associations of OMIM, UNIPROT, PHARMGKB or CTD) and ALL (containing all gene-disease associations). Our comprehensive database is also available as SQLite database (DisGeNET.db). All gene-disease networks are represented as bipartite graphs. A bipartite graph has two types of vertices and the edges run only between vertices of un-like types (Newman, 2003). The bipartite graphs are multigraphs in which two vertices can be connected by more than one edge. In our networks, the multiple edges represent the multiple data sources reporting the gene-disease association. We generated two projections, one for the diseases and one for the genes using the igraph library in R (Gabor and Tamas, 2006). The projected graphs contain only vertices of the same kind (monopartite) and two nodes are connected if they share a neighbour in the original bipartite graph. Before projecting the networks, we simplified the graphs and removed multiple edges. Hence, nodes that are connected by multiple edges are only connected by one edge in the simplified graph. This simplification is needed in order to correctly run the projection as implemented in the igraph library. Moreover, the node degree in the simplified graphs represents the number of first neighbours.

2.3. Mapping of disease vocabularies

We used the MeSH hierarchy for disease classification. The repositories of gene-disease associations use two different disease vocabularies, MIM terms for OMIM diseases (used by OMIM, UniProt, CTD) and MeSH terms (used by CTD, PharmGKB, LHGDN). We used the UMLS metathesaurus to map from MIM to MeSH vocabularies. This step was performed to merge disease terms representing the same disorder, thus reducing redundancy. We were able to map 497 MIM terms directly to MeSH using UMLS and we additionally mapped 23 MIM terms by using a string mapping approach. Briefly, we searched the UMLS metathesaurus for MeSH terms for which there is at least one synonym exactly matching one of the synonyms describing the MIM term of interest. The resulting 63 matched terms were manually checked and reduced to 23 terms. For disease classification, we considered all 23 upper level concepts of the MeSH tree branch C (Diseases), plus two concepts (“Psychological Phenomena and Processes” and “Mental Disorders”) of the F branch (Psychiatry and Psychology). Moreover, we added one disease class “Unclassified” for all disease terms for which a classification was not possible. We categorized all diseases into one or more of the 26 possible disease classes. For MeSH disease terms we directly used its position in the MeSH hierarchy, for MIM disease terms that were not mapped to MeSH, we used the disease classification of (Goh, et al., 2007). Then, we mapped their disease classification to the MeSH hierarchy and extended the mapping using a disease classification available at CTD (CTD_disease_hierarchy.tsv downloaded August, 8th 2009). In total, we were able to classify 3980 (98.39 %) diseases. The disease classification allows filtering and searching of the network restricted to disease class.

2.4. Gene-disease association ontology

For a correct integration of gene-disease association data, we developed a gene-disease association ontology. We classified all association types as found in the original source databases into Association if there is a relationship between the gene/protein and the disease, and into NoAssociation if there is no association between a gene/protein and a certain disease (in other words, if there is evidence for the independence between a gene/protein and a disease). The different association types from the original databases were mapped to the ontology for a seamless integration. In this study, we only considered gene-disease associations of type Association. The ontology is available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download>.

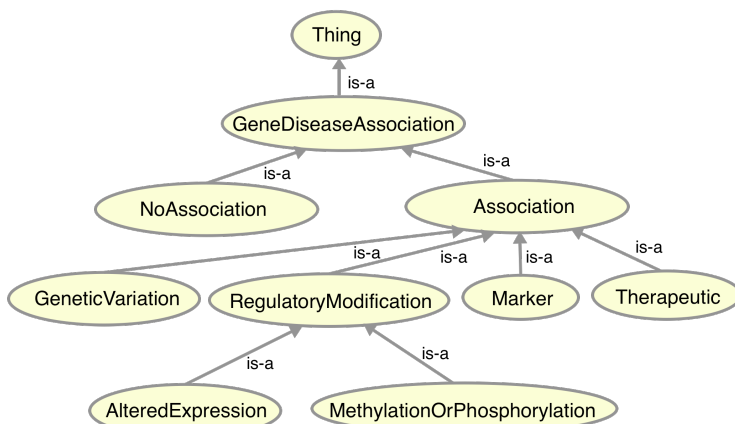


Figure 1: Gene-disease association ontology

3. DisGeNET tutorial

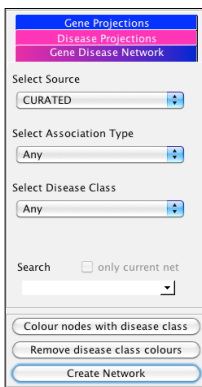
DisGeNET is a plugin for Cytoscape (Shannon, et al., 2003) to query and analyze human gene-disease networks. For this purpose, we have developed a new gene-disease association database integrating information from several expert curated databases and a resource containing text-mining derived associations (Bauer-Mehren, et al., 2010).

3.1. Basic functions

By selecting different data sources, association types and/or disease classes from their respective drop-down menus, you can generate different gene-disease association networks. In addition, gene-disease association networks can be generated around a specific disease or gene of interest using the search box provided with the plugin. Most of these functionalities are also available to generate disease and gene monopartite networks.

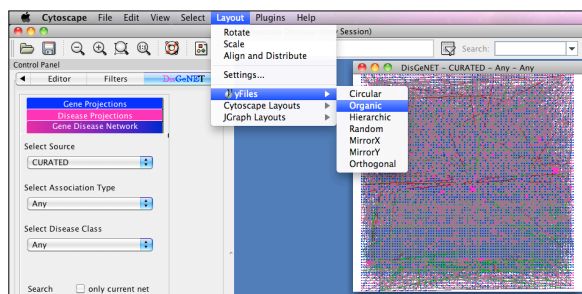
3.1.1. Generate gene-disease association network

In order to obtain a gene-disease association network without any restrictions on association type and disease class follow the next steps:



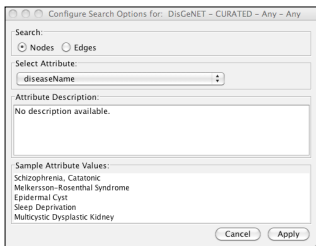
- Select the source of interest, e.g. CURATED containing information from all expert curated databases in our database (OMIM, PHARMGKB, UNIPROT and CTD).
- Set Association Type and Disease Class *Any*
- Press *Create Network*

- Apply a Cytoscape layout algorithm to generate the view of choice, e.g. select the layout *Organic*



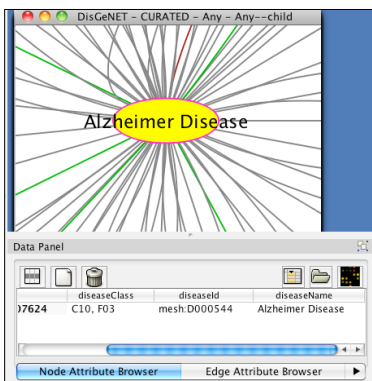
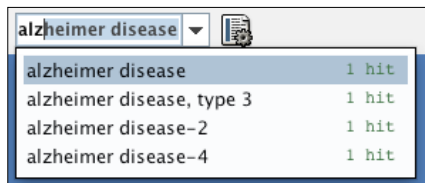
DisGeNET tutorial

Once the network is obtained, specific information on the nodes and their relationships can be explored as detailed below:

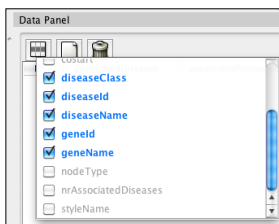


- Select nodes and edges and check their attributes.
- For example, use the Cytoscape search function to query for *Alzheimer Disease*. For this purpose, modify the search options and select the attribute *diseaseName*.

- Search for a particular disease, e.g. *Alzheimer Disease*

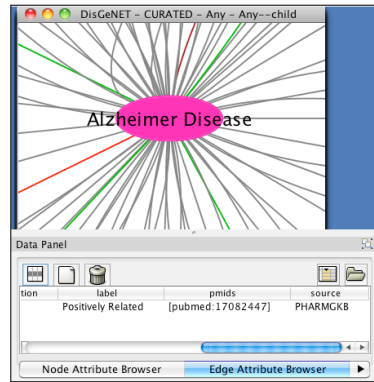


- Zoom into the network and select the *Alzheimer Disease* node
- More information about this node is found in the *Node Attribute Browser*
- All available node and edge attributes are listed in Tables 1 and 2.



- For this purpose you might want to select attributes to be displayed in the *Node Attribute Browser* or *Edge Attribute Browser* of the Cytoscape *Data Panel*.

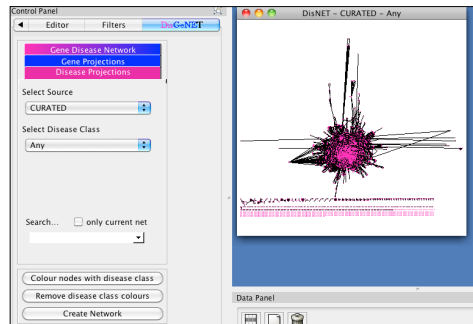
- Select an edge to display information about a particular gene-disease association such as associationType, data source providing this association, supporting evidence (PubMed identifiers), etc.



3.1.2. Generate gene or disease projection network

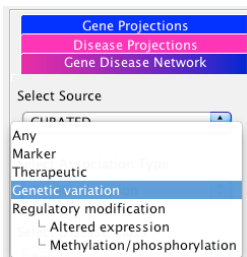
In addition to bipartite graphs representing gene-disease associations, DisGeNET allows generating monopartite networks representing the gene or the disease projection of the gene-disease association network. In order to obtain the disease projection of the network generated from CURATED source (described in 2.1.1) follow the instructions detailed below:

- Select the *Disease Projection* tab in the DisGeNET main panel.
- Select the source, e.g. CURATED
- Press *Create Network*



3.1.3. Restrict the network to a certain association type

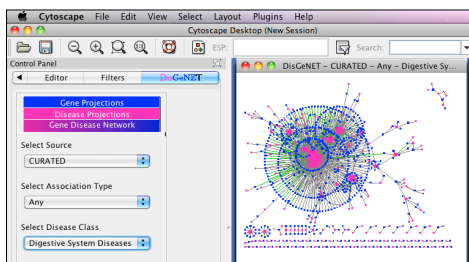
Note: This option is only available for Gene Disease Networks.



- Select the Source, e.g. CURATED
- Select the Association Type, e.g. *Genetic variation*
- Press *Create Network*

3.1.4. Restrict the network to a certain disease class

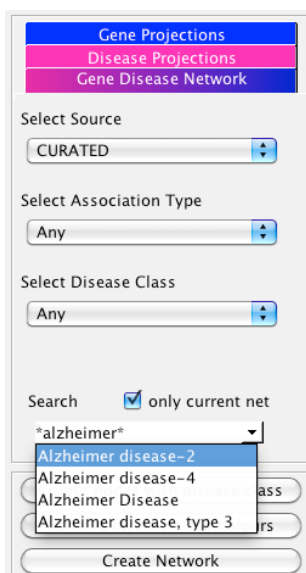
Note: This option is available for all types of networks. The classification is based on the disease branch of the MeSH hierarchy.



- Select the Source, e.g. CURATED
- Select the Disease Class, e.g. Digestive System Diseases
- Press *Create Network*

3.1.5. Search for a particular gene/disease or set of genes/diseases

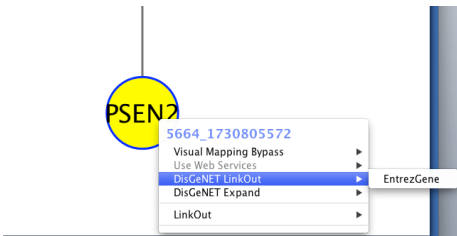
The search option included in the DisGeNET tab can be used to generate networks around a disease or gene of interest. In addition, it can be used to search for a given disease or gene of interest in a network already generated.



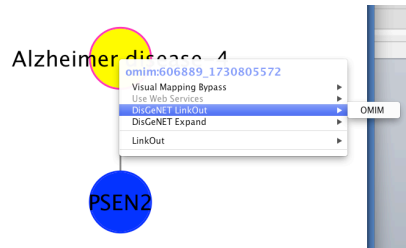
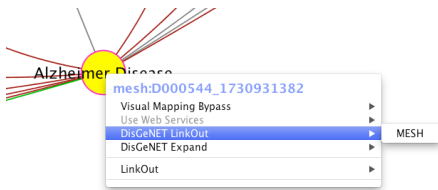
- If *only current net* is **not** ticked, a network only containing associations related to the query will be created (using *Create Network*).
- If *only current net* is **ticked**, the according node will be selected (highlighted yellow) in the current network (with active view) when pressing [Enter].
- The search is restricted to *Source*, *Association Type* and *Disease Class* as selected.
- In this example, we are searching for any kind of Alzheimer Disease (there are four different types) in the CURATED dataset without any restriction of association type or disease class.
- Note: The DisGeNET search allows the use of the wildcard symbol (*). For performance reasons only the first 50 matching terms are listed in the drop-down box but all are included in the generated network.

3.1.6. DisGeNET LinkOut

In order to get more information about a gene or a disease node, you can linkout to the according website (Entrez Gene, OMIM or MeSH) using the *DisGeNET LinkOut* function. It is available in the node context menu, which can be accessed by right-clicking a selected node.



- For gene nodes, a linkout to Entrez Gene is given.
- For disease nodes, linkouts to MeSH or OMIM (depending on the type of disease node) are given.



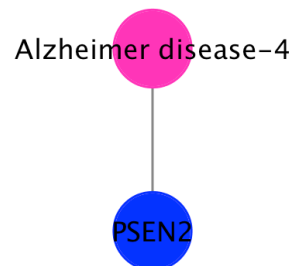
3.1.7. DisGeNET Expand

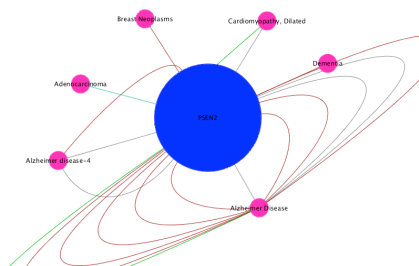
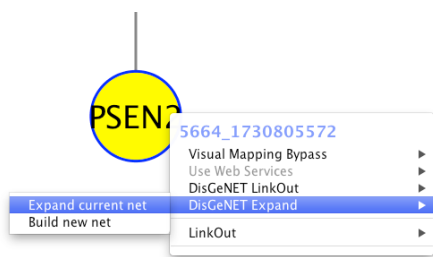
In order to find all diseases/genes that are associated to a gene/disease node in an existing network you can use the *DisGeNET Expand* function. It can either be used to create new DisGeNET networks using the selected nodes for the query or to expand the existing nodes with edges found in DisGeNET.

Note: the function works with one or more selected nodes. To call the function, select one or more nodes, then click the right mouse button. This will open the node context menu containing the *DisGeNET LinkOut* and *DisGeNET Expand* functions. You can then choose between *DisGeNET Expand -> Expand current net* and *DisGeNET Expand -> Build new net*.

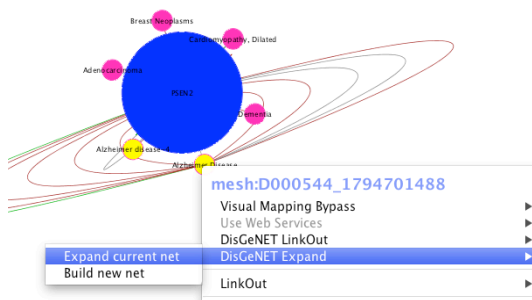
3.1.7.1. Expand DisGeNET networks

- This is a network generated with DisGeNET using as source *OMIM*, as AssociationType and DiseaseClass *Any* and as search term *PSEN2*.
- In OMIM, there is only one disease (*Alzheimer disease-4*) annotated to the gene *PSEN2*.

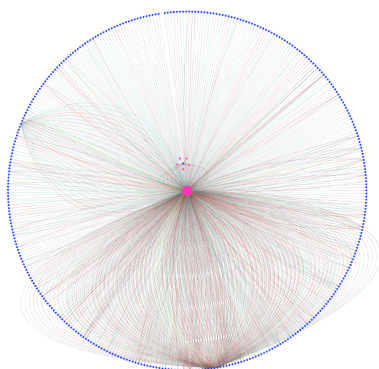




- The expansion can be repeated various times. For instance, in a next step, we can expand this network by querying for more genes associated to *Alzheimer disease - 4* and *Alzheimer Disease*.



- This results in a large network with 373 nodes and 893 edges. It is visible that there are many more genes associated to *Alzheimer Disease*.



3.1.7.2. Expand foreign networks

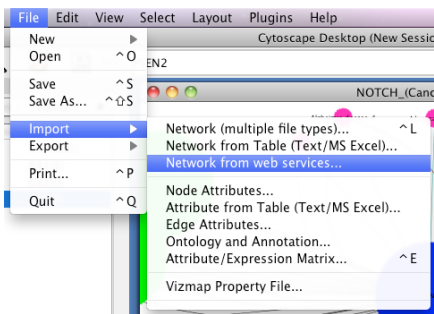
The same functionality to expand gene or disease nodes with more associations found in DisGeNET can be used to expand foreign network that were not created with DisGeNET but contain gene or disease nodes. In order to use the *DisGeNET Expand* function on nodes that were not built within DisGeNET, the node label needs to contain a valid Entrez Gene identifier or valid disease identifiers that are allowed by DisGeNET.

Note: DisGeNET only contains human gene-disease associations and hence can only be queried with human gene identifiers.

Examples for valid identifiers:

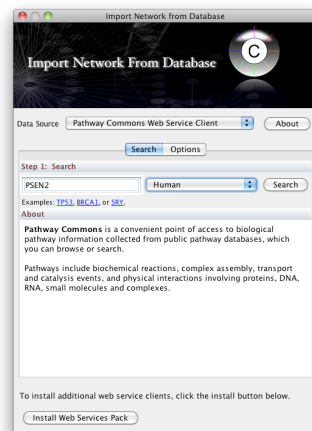
- **5080** for PAX 6 gene
- **mesh:D000544** for Alzheimer Disease
- **omim:217700** for Corneal endothelial dystrophy 2

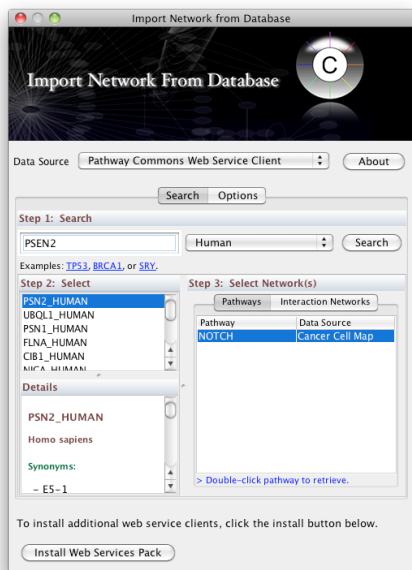
In the following example, we show how a network not generated with DisGeNET can be expanded with DisGeNET gene-disease associations.



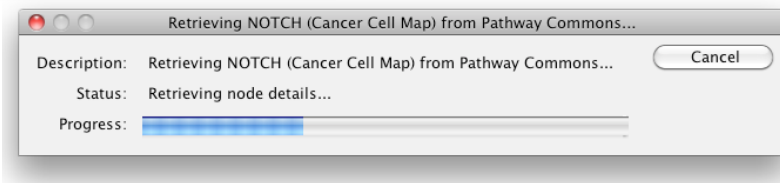
- First, we generate a network using the *File->Import->Network from webservices* function within Cytoscape.

- We query the Pathway Commons database for pathways containing the human gene *PSEN2*. For this, first set the Data Source to Pathway Commons Web Service Client, enter PSEN2 in the Search field and select the organism Human. Press search.

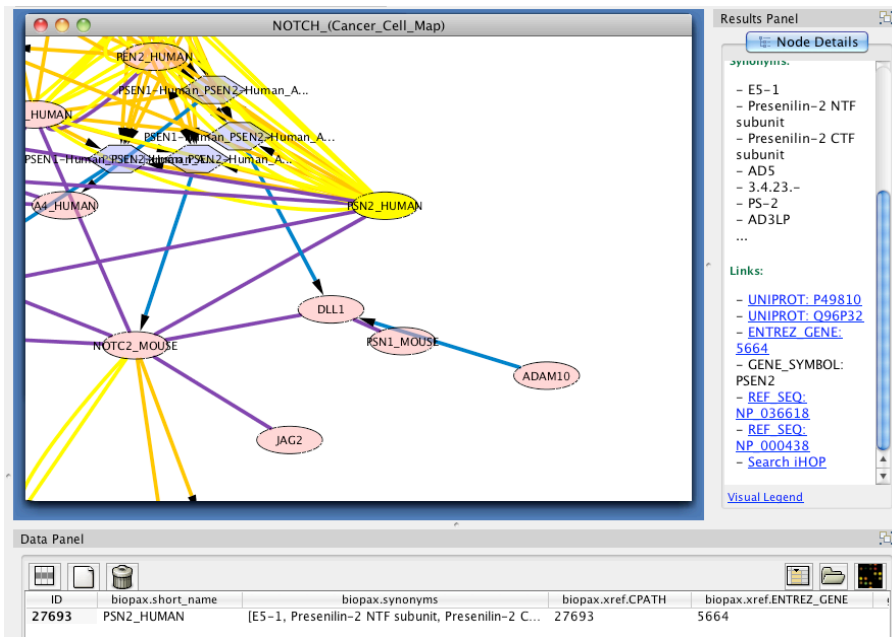




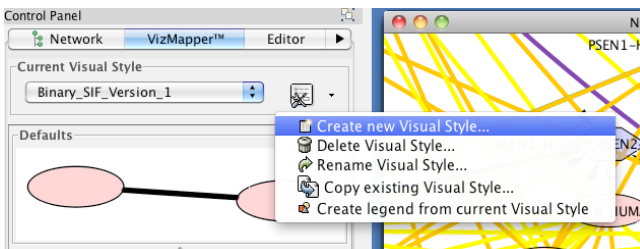
- We select a pathway we are interested in, for instance the NOTCH signalling pathway from the Cancer Cell Map database. Double-click the pathway to retrieve it.



- This results in a network with 113 nodes and 272 edges.
- The network contains the *PSEN2* gene (PSN2_HUMAN).
 Moreover, there are various node attributes available among them the Entrez Gene identifier (biopax.xref.ENTREZ_GENE)



- In order to use *DisGeNET expand*, we need to ensure that the node labels contains the Entrez Gene identifier since DisGeNET uses node labels to query the database.



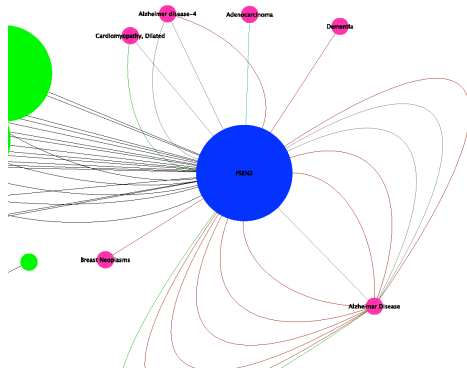
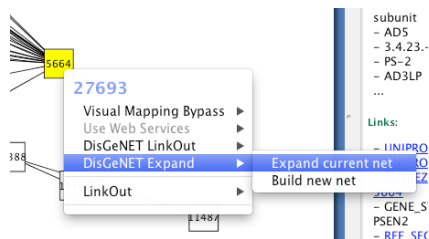
DisGeNET tutorial

- To do so, we first create a new visual style in the VizMapper, for example called "ExpandDisGeNETStyle"

| Node Visual Mapping | |
|---------------------|--------------------------|
| Node Label | biopax.xref.ENTREZ_GE... |
| Mapping Type | Passthrough Mapping |
| 57081 | 7528 |

- Then, we set the node label to the attribute containing the Entrez Gene identifiers, here to biopax.xref.ENTREZ_GENE and use the Passthrough Mapping.

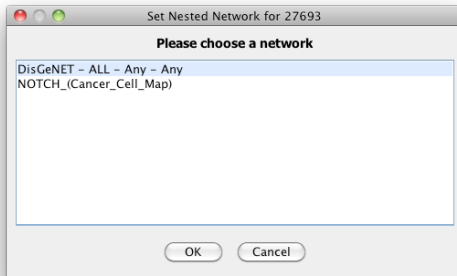
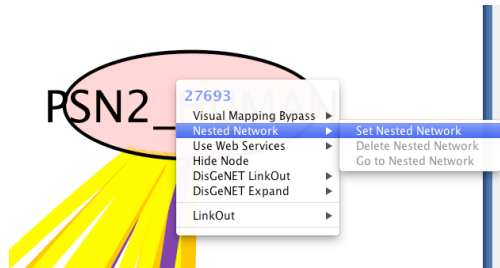
- Now, we can use the *DisGeNET Expand* function to search for gene-disease associations containing the selected node. Using the function for the *PSEN2* node, we can search for all associated diseases in DisGeNET. We can either add the found associations to this net or create a new net.



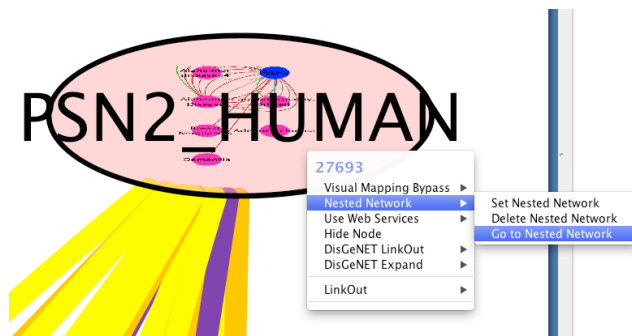
- In the resulting network all diseases associated to *PSEN2* are added

- For Cytoscape 2.7.0 users: You can make use of the *Nested networks* functionality to add the gene-disease association networks as nested networks to the nodes.

- To add the gene-disease association network as nested network to the *PSEN2* gene node, right click on the node and select *Nested Network* -> *Set Nested Network*

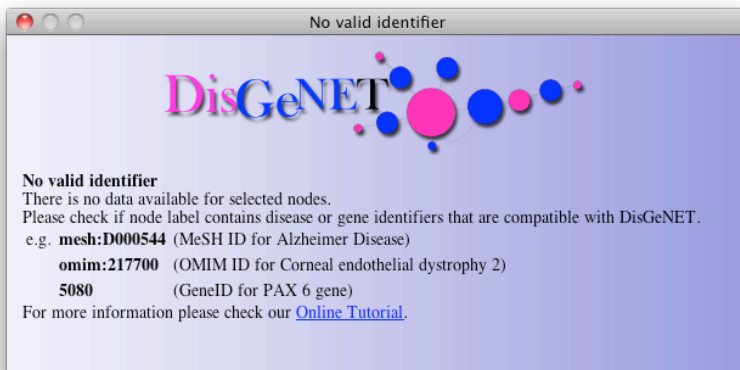


- Now select the gene-disease association network for *PSEN2* as created before using *DisGeNET Expand*



- The *PSEN2* node now contains the gene-disease association network as nested network, which can directly be opened by using the *Nested Network*->*Go to Nested Network* function.

- If the node label does not contain valid identifiers for DisGeNET, an error message is shown.



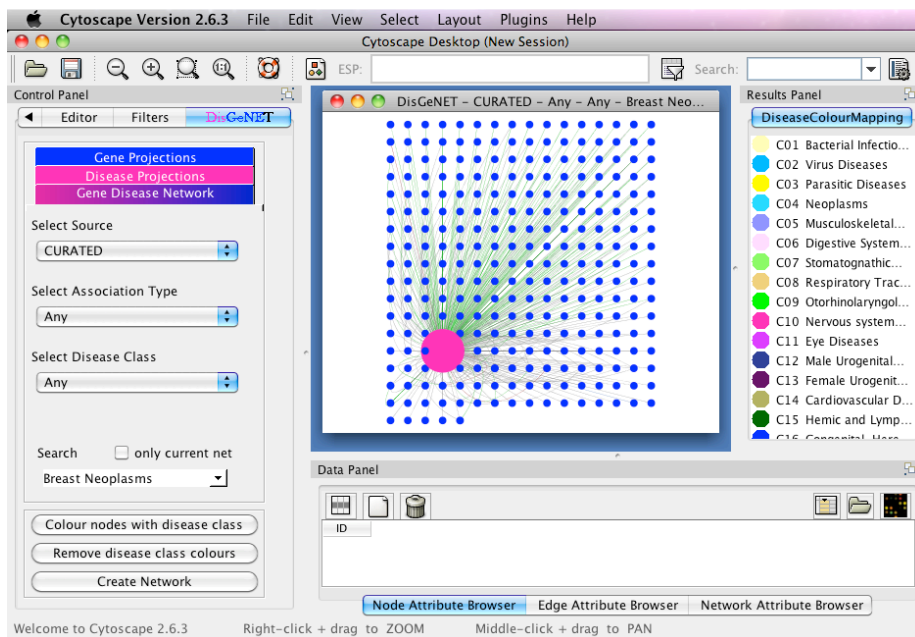
3.2. Specific use cases

In this section some examples that illustrate the kind of questions that can be answered using DisGeNET are presented.

3.2.1. Which are the genes annotated to breast cancer in expert curated databases?

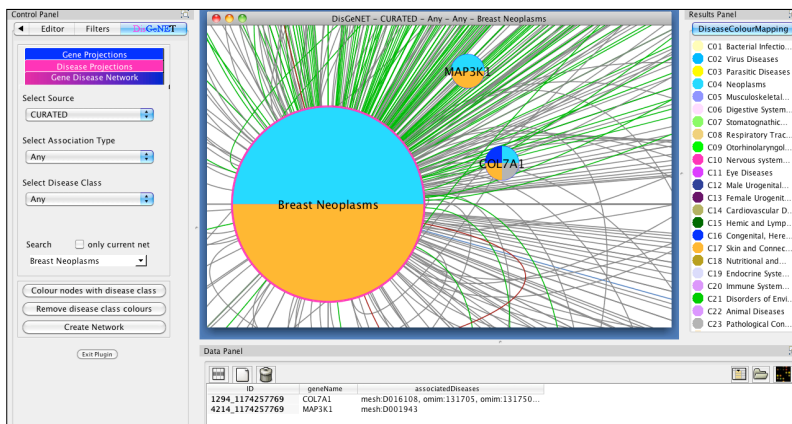
This is an example of a more general question that can be phrased as “Give me all the genes known to be associated to disease x from a given data source”.

In order to answer this question, query the Gene Disease network selecting CURATED as source, no restriction on the association type or disease class, but specifying *Breast Neoplasms* in the search field to restrict the search to the genes annotated to this disease term. This will generate a network with 277 nodes (one disease and 276 gene nodes) and 417 edges. The edges are coloured according to the association type.



Many genes associated to *Breast Neoplasms* are also annotated to other diseases. We can inspect these diseases by exploring the node attributes *associatedDiseases* in the Node Attribute browser and also by colouring the nodes according to MeSH disease classification. For this purpose, use the function *Colour nodes with disease class*.

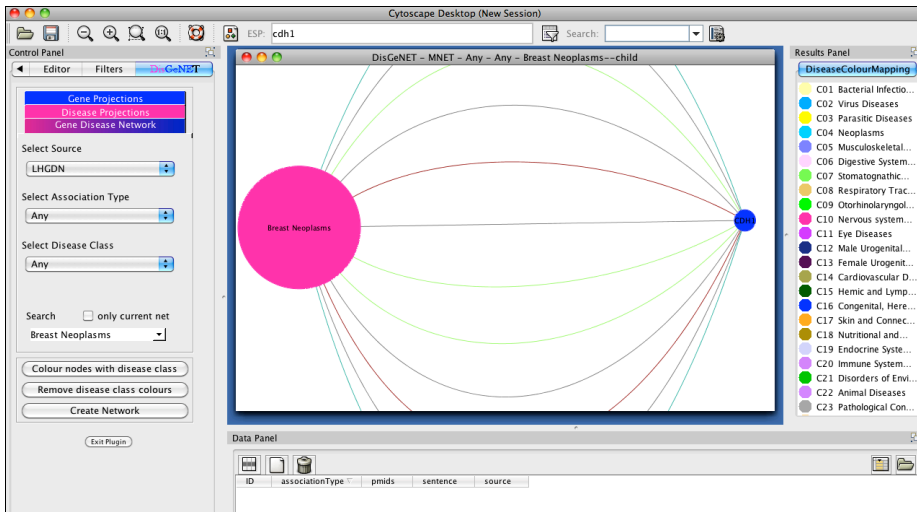
DisGeNET tutorial



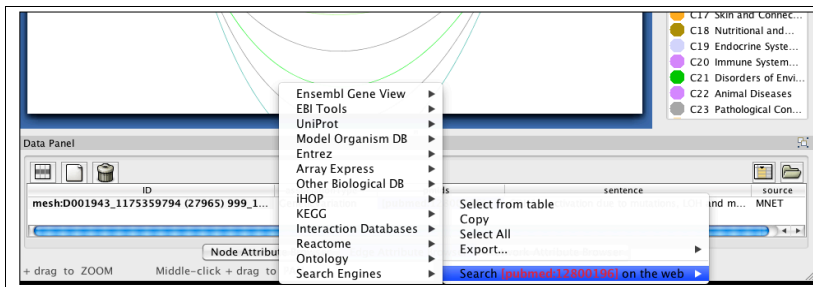
Breast Neoplasms is classified as *Neoplasms* and *Skin and Connective Tissue Disease*. *MAP3K1* is a gene annotated only to *Breast Neoplasms* in the *CURATED* data set, while *COL7A1* is annotated to 8 different diseases belonging to 4 different disease classes.

In order to know if there are other genes described in the literature but not recorded in the set of curated databases considered, we perform the query on the LHGDN set. This query will retrieve annotations derived from text-mining. The result is a network composed of 1099 genes annotated to *Breast Neoplasms* (1100 nodes and 3321 edges).

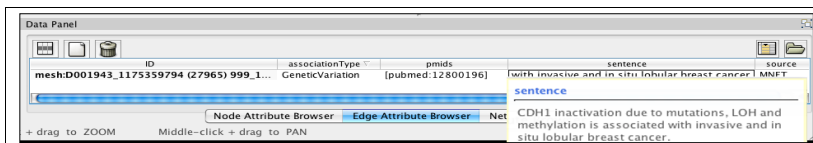
If we inspect the association between gene *CDH1* and *Breast Neoplasms*, we see that there are 12 edges connecting the two nodes. The associations belong to different classes (Marker, GeneticVariation, etc.), hence they are coloured differently.



Furthermore, we can explore the supporting evidence for each gene-disease association by inspecting the edge attribute browser. We can examine the associations by either linking out to the according publication (using the Cytoscape function *Search on the web*).



Or we can view the sentence that was found by text-mining that supports the association between the gene and the disease (using the node attribute *sentence*).



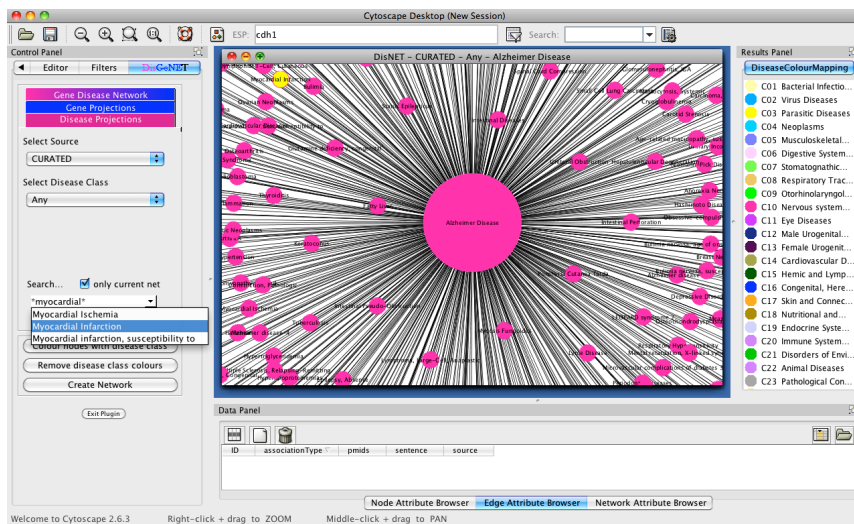
This example illustrates the value of incorporating information from literature, since the curated databases currently don't cover all knowledge about gene-disease association available in the literature.

3.2.1.1. Do comorbidities observed in patients reflect a common genetic origin of the diseases?

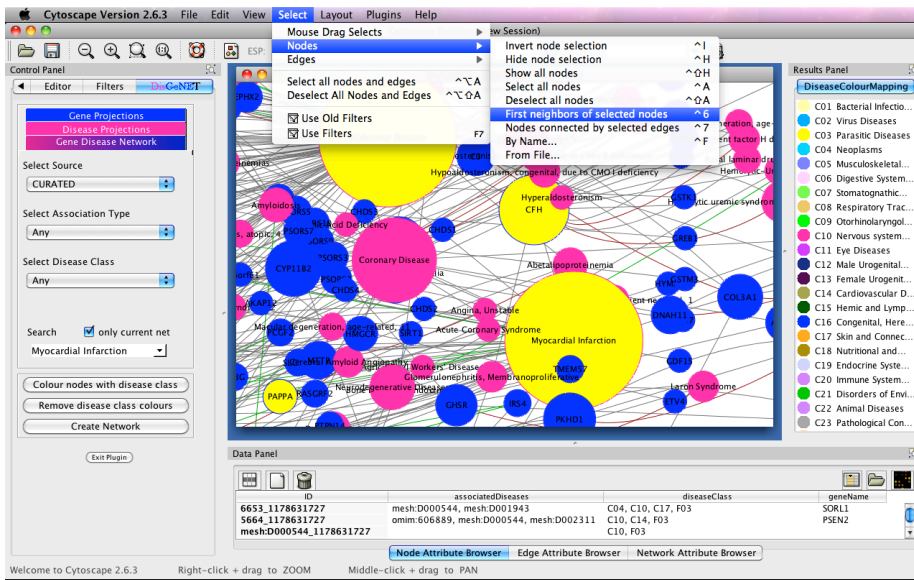
This is a specific example of a more general question such as: Are diseases x and y related by genetic origin?

Some diseases are known to co-occur in a patient, a process known as disease comorbidities (Park, et al., 2009). Disease comorbidities can be studied considering the common genetic origin of both diseases. *Alzheimer Disease* and *Myocardial Infarction* are one example of comorbidity. By querying the Gene Disease network we can answer the question if these two diseases share a common genetic origin.

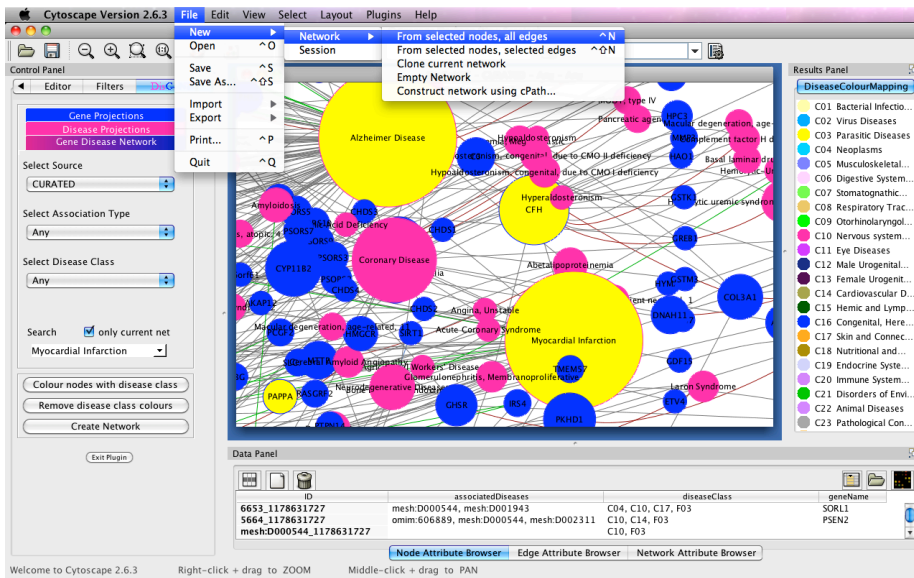
First, we query the Disease projection (CURATED) for *Alzheimer Disease* with no restriction to *Disease class* and create the network. Then, we search this network for *Myocardial Infarction* using the DisGeNET search function with the option *only current net* ticked. We immediately see that both diseases are connected.



Once we know that there is at least one gene shared between both diseases, we can go back to the CURATED Gene Disease Network (or create it) and then create a subnetwork containing the two diseases and their associated genes. For this purpose, we first select the four nodes representing subtypes of *Alzheimer Disease* (Alzheimer Disease, Alzheimer disease-2, Alzheimer disease-4, Alzheimer disease, type 3) and the two nodes for *Myocardial Infarction* (Myocardial Infarction and Myocardial infarction, susceptibility to) and their associated genes using the Cytoscape function *Select -> Nodes -> First neighbours of selected nodes*.



Then, we create a subnetwork containing all selected nodes and all edges using the Cytoscape function *File -> New -> Network -> From selected nodes, all edges*.

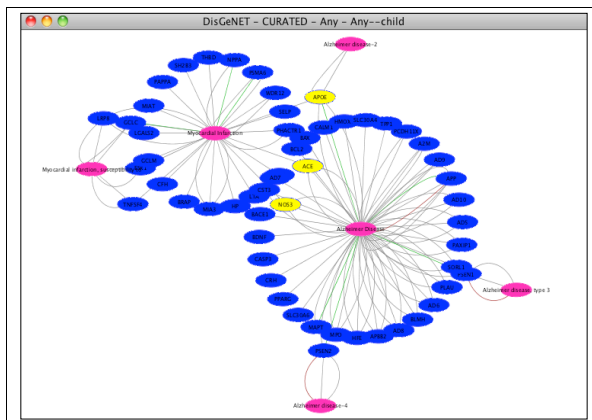


THESIS PUBLICATIONS

DisGeNET tutorial

As a result we obtain a network containing 62 nodes and 126 edges. We can see that *Alzheimer disease* and *Myocardial Infarction* are both annotated to the genes NOS3, ACE and APOE, supporting the hypothesis that alterations in the function of these genes can result in the development of both diseases in the same patient.

The same result can be obtained using the Cytoscape plugin “Advanced Network Merge”. For this purpose, we can create two separate gene-disease networks for *Alzheimer Disease* and *Myocardial Infarction* and then use the “Advanced Network Merge” to merge these networks using for instance the geneId as matching attribute.



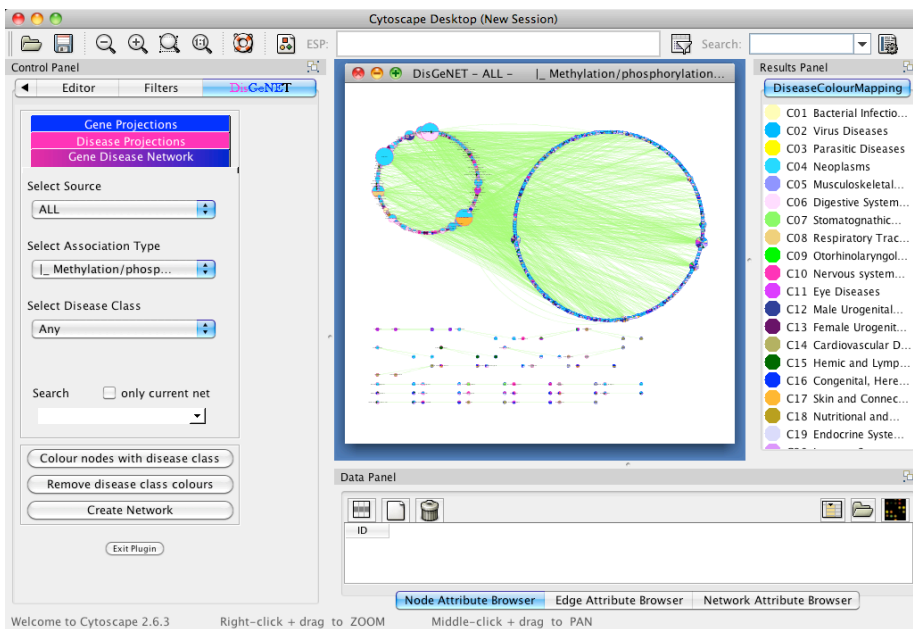
3.2.2. Which are the diseases that are associated to post-translational modifications such as phosphorylation?

This is an example of a more general search query that can be expressed as: Give me all the diseases for which there are alterations in post-translational modifications such as x.

This type of use case might be of interest for drug discovery projects in which the identification of disease genes able to be targeted by drugs interfering with phosphorylation is needed.

First, we create a network querying the complete Gene Disease network (ALL) and restricting the *Association Type* to *Methylation/phosphorylation* (no restriction to any *Disease class*). The query results in a network composed of 621 nodes (157 disease nodes) and 1117 edges.

By exploring the diseases (use *Colour nodes with disease class*), it can be observed that most of them belong to the class *Neoplasms*, but there are other diseases such as those belonging to *Nervous systems Diseases*, *Hemic and Lymphatic diseases*, *Immune Systems Diseases*. The supporting evidence for each gene-disease association can be explored using the edge attribute browser as explained in section 2.1.



3.3. Analyzing DisGeNET data using external tools

Some of the networks can get very large, especially when using LHGDN or ALL and for this reason the plugin will not create gene projections with the LHGDN or ALL setting.

In order to analyze DisGeNET data with external network analysis tools such as the igraph library for complex network research (Gabor and Tamas, 2006), we provide all networks and attributes in a sqlite database available at <http://ibi.imim.es/DisGeNET/DisGeNETweb.html#Download>.

3.3.1. Extract data from DisGeNET database

Sqlite can be downloaded from <http://www.sqlite.org/download.html>. Please also check the sqlite documentation for more information.

Connect to the DisGeNET database using the following command (call from the folder containing the DisGeNET database, e.g. ../cytoscape-v2.6.3/plugins/DB/):

```
sqlite3 DisGeNET.db
```

Use the following commands to extract the whole (ALL) gene-disease network and to write them into a tab delimited text file named DisGeNET_ALL.txt:

```
sqlite> .mode tab
sqlite> .output ./DisGeNET_ALL.txt
sqlite> select * from geneDiseaseNetwork where source="ALL";
```

3.3.2. Build networks using igraph library

Once you have saved the network, you can access and visualize it with any external tools for network analysis. Many tools can read tab delimited text files such as the igraph library for R. The igraph library can be downloaded from <http://igraph.sourceforge.net>.

Start R and use the igraph library.

```
R
R> library(igraph)
```

Read in the network and build a graph object.

```
R> edges -> read.csv(file="./DisGeNET_ALL.txt", sep="\t", header=F)
R> graph -> graph.data.frame(edges, directed=F)
```

Now you can make use of a variety of graph manipulation functionalities available in igraph. For further information check the igraph documentation.

4. Contact

4.1. Biomedical Informatics group

We are interested in the understanding of the mechanisms underlying biomedical related problems at the molecular scale. This involves the study of the network of interactions between molecules that underlay, for instance, the etiology of a complex disease. In addition to the study of diseases of complex origin, we are also interested in the mechanisms underlying the appearance of side effects after drug treatments.

One part of our research is focused on strategies to, once a network of molecular interactions is obtained, characterize the network and model its behavior in order to gain insight into the etiology of the disease phenotype. In particular, we are interested in the application of qualitative modeling approaches, such as Petri Nets and Boolean networks.

Another line of research involves strategies for obtaining the networks that are relevant for the biomedical related problems already mentioned. For this, we are developing software for the retrieval and analysis of data from public network repositories (databases of signaling pathways, gene regulatory networks and metabolic reactions). Although the publicly available network databases contain valuable information, we are aware that their coverage is not complete: a lot of information regarding interaction between biomedical entities (genes, proteins, phenotypes, chemicals, drugs, etc) still lies in the biomedical literature as free text.

This is where our third line of research comes in, which involves the use of text mining approaches for the extraction of relationships between biomedical entities from the biomedical literature. In the past years we have developed NER systems for the identification of mentions of gene sequence variants from MEDLINE abstracts, and linkage of the mentions found in text to the corresponding database identifiers (in this case dbSNP). In addition, we have developed a corpus with annotations for variation mentions for the evaluation of this kind of NER systems. Currently, we are working on the application of NLP approaches for the identification and extraction of different types of relationships between biomedical entities.

4.2. Citation

If you are using DisGeNET for your own research, please cite:

Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI: **Network analysis of an integrated gene-disease association database reveals functional modules in mendelian, complex and environmental diseases.** *Submitted.*

Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI: **DisGeNET - a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks.** *Submitted.*

Contact

4.3. Acknowledgements

This work was generated in the framework of the EU-ADR project co-financed by the European Commission through the contract no. ICT-215847 and the eTOX project from the European Community's Seventh Framework Program (FP7/2007-2013) for the Innovative Medicine Initiative under grant agreement no. 115002. The Research Unit on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB) and member of the COMBIOMED network. We thank the Departament d'Innovació, Universitat i Empresa (Generalitat de Catalunya) for a grant to author ABM.

4.4. Contact

If you have question, comments or suggestions, please contact us.

[Laura Furlong \(lfurlong@imim.es\)](mailto:lfurlong@imim.es)

Tel: (+34) 93 316 0521

Fax: (+34) 93 316 0550

<http://ibi.imim.es/>

Integrative Biomedical Informatics Laboratory
Research Group on Biomedical Informatics (GRIB) - IMIM/UPF
Parc de Recerca Biomèdica de Barcelona
Dr. Aiguader, 88
E-8003 Barcelona

5. Attribute tables

Table 1: Edge attributes in the gene-disease network

| Name | Description |
|-----------------|---|
| associationType | Association type of the gene-disease association according to the gene-disease association ontology (see Gene-disease association ontology}). |
| interaction | Unique identifier for this association. |
| label | Association type as originally assigned by the source database |
| pmids | List of PubMed identifiers of publications supporting the reported gene-disease association, if available. |
| sentence | The actual sentence in which the gene-disease association was detected (only available for LHGDN). |
| source | Database in which this gene-disease association was reported (OMIM, UNIPROT, PHARMGKB, CTD, CURATED, LHGDN, ALL) |

Table 2: Node attributes in the gene-disease network

| Name | Description |
|--|---|
| associatedDiseases | List of disease identifier associated to a gene node. |
| associatedDiseaseNames | List of disease names which are associated to a gene node. |
| associatedPathwayNames | List of KEGG and Reactome pathways the gene is annotated to (only for gene nodes). |
| associatedPathways | List of KEGG and Reactome pathway identifiers the gene is annotated to (only for gene nodes). |
| nrAssociatedDiseases/ nrAssociatedGenes | Number of associated diseases or genes (number of first neighbours of the node). |
| diseaseClass | List of disease class identifiers (disease classes according to MeSH hierarchy). |
| diseaseId | MIM or MeSH identifier for the disease node |
| diseaseName | Name of the disease according to MeSH or OMIM morbidmap. |
| geneId | Entrez Gene identifier of the gene. |
| geneName | Name of the gene. |
| nodeType | The type of node (gene or disease). |
| styleName | Name of gene or disease, needed for the DisGeNET visual style. |
| styleSize | Number of first neighbours of the node, needed for the DisGeNET visual style. |

6. References

- Altman, R.B. (2007) PharmGKB: a logical home for knowledge relating genotype to drug response phenotype, *Nat. Genet.*, **39**, 426.
- Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L.-S.L. (2004) UniProt: the Universal Protein knowledgebase, *Nucleic Acids Res.*, **32**, D115-119.
- Bauer-Mehren, A., Bundschus, M., Rautschka, M., Mayer, M.A., Sanz, F. and Furlong, L.I. (2010) Network analysis of an integrated gene-disease association database reveals functional modules in mendelian, complex and environmental disease, *submitted*.
- Bauer-Mehren, A., Furlong, L., Rautschka, M. and Sanz, F. (2009) From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways, *BMC Bioinformatics*, **10**, S6-S6.
- Bundschus, M., Dejori, M., Stetter, M., Tresp, V. and Kriegel, H.-P. (2008) Extraction of semantic biomedical relations from text using conditional random fields, *BMC Bioinformatics*, **9**, 207-207.
- Gabor, C. and Tamas, N. (2006) The igraph Software Package for Complex Network Research, *InterJournal, Complex Systems*, 1695.
- Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabási, A.-L. (2007) The human disease network, *Proc. Natl. Acad. Sci.*, **104**, 8685-8690.
- Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Res.*, **33**, D514-517.
- Lu, Z., Cohen, K.B. and Hunter, L. (2007) GeneRIF QUALITY ASSURANCE AS SUMMARY REVISION, *Pac. Symp. Biocomput.*, 269-280.
- Mattingly, C.J., Rosenstein, M.C., Davis, A.P., Colby, G.T., Forrest, J.N. and Boyer, J.L. (2006) The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks, *Toxicol. Sci.*, **92**, 587-595.
- Mitchell, J.A., Aronson, A.R., Mork, J.G., Folk, L.C., Humphrey, S.M. and Ward, J.M. (2003) Gene Indexing: Characterization and Analysis of NLM's GeneRIFs, *AMIA Annu. Symp. Pro.*, **2003**, 460-464.
- Newman, M.E.J. (2003) The structure and function of complex networks, *SIAM Review*, **45**, 167-256.
- Park, J., Lee, D.-S., Christakis, N.A. and Barabasi, A.-L. (2009) The impact of cellular networks on disease comorbidity, *Mol. Syst. Biol.*, **5**.
- Rubinstein, R. and Simon, I. (2005) MILANO - custom annotation of microarray results using automatic literature searches, *BMC Bioinformatics*, **6**, 12.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.*, **13**, 2498-2504.

3.6. From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways

Bauer-Mehren A, Rautschka M, Furlong LI, Sanz F

BMC Bioinformatics. 2009;**10**(Suppl 8):S6-S.

Bauer-Mehren A, Rautschka M, Furlong LI, Sanz F. [From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways.](#) BMC Bioinformatics. 2009; 10 (Suppl 8): S6.

4. DISCUSSION

This thesis is focused on understanding the molecular basis of human diseases and adverse drug reactions by means of novel integrative analysis methodologies. Clearly, there is an urgent need for such new methods considering the fact that we are on the one hand able to generate vast amounts of “omics” data but on the other hand we are not fully able to automatically combine and analyse all that information in a meaningful way, yet. Although for many diseases there is successful treatment, we are still not able to prevent and cure common human diseases, such as cancer. In addition, the number of drug candidates failing in the last phase due to lack of efficacy or severe side effects is still dramatically high, despite the recent advances in experimental and computational technologies. Hence, it becomes obvious that the knowledge about the molecular basis of human diseases and drug adverse reactions can often not be directly translated into clinical practice. In many cases, the problem is related to the fact that the knowledge is fragmented and difficulties in data retrieval and integration are hindering the direct use of the available information. In other cases, the molecular mechanisms underlying the disease or the adverse drug reaction are partly understood but need to be studied in more detail.

In this regard, the main objective of this PhD thesis was the development of novel integrative analysis approaches for investigating the molecular basis of human diseases and adverse drug reactions. In the previous chapter (chapter 3), the results of the research carried out during this PhD thesis have been presented. These results were published or are under review for their publication in peer-reviewed journals, or in case of chapter 3.1.4. in preparation. In this section, the main results and outcomes of this thesis are critically discussed. Herein, the results are ordered by the methodological approach they are based on ranging from the classical multivariate statistical methods to the recent systems biology approaches.

4.1. Multivariate statistical approaches to study adverse drug reactions

In the introduction of this PhD thesis we have explained the difficulty of advancing in the understanding of the molecular basis of human diseases and adverse drug reactions. Due to the extreme complexity of this task we need new methodological approaches able to shed light on the relationships between the involved entities and their descriptors. In this regard, we developed a new method for exploiting molecular and pharmacological information for a set of drugs known to interact with multiple receptors in order to investigate, for instance, undesired side effects of the drugs. This method introduces a multilevel approach, which is based on the sequential building of linked multivariate statistical models, all based on PCA or PLS where each model introduces a different level of drug description. These levels comprise 1) *clinical/in vivo* observations related to the therapeutic and side effects of the drugs, 2) *in vitro* binding affinities of the drugs to receptors and 3) the molecular structures of the ligand-receptor complexes (see section 3.1). The method was described and validated using as example a set of antipsychotic drugs for which data at all three levels was available. Herein, it is important to mention

DISCUSSION

that the selected set of drugs should consist of structurally diverse compounds since the underlying statistical models work by grouping together objects with similar properties. The results of the method applied to the set of antipsychotic drugs strongly support the usefulness of the proposed methodology and literature evidence was found for most findings (see section 3.1). The clear advantage of this method is that instead of producing models that behave like “black boxes”, the obtained models can be inspected in different ways. This is possible due to the use of the statistical methods PCA and PLS, which allow inspecting on the one hand the objects clustered by their similarity and on the other hand the factors responsible for the clustering. All in all, the developed methodology for the first time uses multivariate statistical models to study the properties of drugs at different levels, facilitating the design of more efficient and safer drugs. Nevertheless, it has to be kept in mind that the found relationships are of statistical nature, which have not to be causal relationships. Hence, in order to gain a mechanistic understanding of the established relationships, a more detailed approach has to be followed up. Nonetheless, the clear advantage of the proposed method is that it requires few data regarding *clinical/in vivo*, *in vitro* and structural properties of the studied drugs, all typically available during drug development and after drug marketing. Moreover, finding a statistical association between, for instance, drug properties and adverse drug reactions could be seen as first step guiding mechanistic subsequent analysis approaches.

4.2. Systems biology approaches to investigate the molecular basis of diseases and adverse drug reactions

A promising research area for a more detailed analysis of the molecular basis of human traits is the recent systems biology as introduced in section 1.3.2. It has been mentioned several times that in order to gain a mechanistic understanding of the biological processes implicated in diseases and adverse drug reactions, the interactions between the involved key players have to be taken into account. In this respect, within the field of systems biology, network analysis approaches have evolved to study the static and dynamic properties of network models for biological pathways. This has immensely improved our understanding of important cell signalling processes. However, up to now the building of such network models is typically done manually by inspecting all literature related to the processes of study. Consequently, in this PhD thesis we extensively studied public pathway databases with respect to their suitability to provide data ready to be used for automatic building of network models. As shown in section 3.2, a smooth automatic integration of pathway data is currently not possible although the data quality would in principle allow using the data directly to answer specific biological questions. We extensively discussed the current major limitations and illustrated them in several examples. Mostly they are related to lack of data, misuse of standard data formats and incorrect or missing annotations. Moreover, our analysis revealed a lack of communication between the systems biologists building network models to study specific biological processes and the database owners and curators. We

therefore proposed a workflow for an automatic merging of publicly available pathway data and the subsequent automatic network model building (see section 3.2). We also suggested that model builder and database curator should engage in collaborative projects to take advantage of the data already available in public databases and work together on representations that fit the needs of both communities. We believe that a closer collaboration will yield in more accurate data, improved automatic access and integration, which will eventually allow the fast and automatic generation of network models useful to answer important biological questions. Future directions in this regard include the development of more advanced computational tools allowing the merging of biological pathways that overcome the shortcomings of current approaches and that cope with the annotation issues we have discussed. Moreover, automatic building of network models would require the collection of additional information, for instance on the kinetics of the involved processes, for which to the best of our knowledge no standard databases exist at the moment. We are hoping to advance in this direction by building qualitative models, which are mainly based on the network structure and do not require information about the kinetics. Hence, in the future we plan to work extensively on automatic pathway integration and qualitative network modelling with the ultimate goal of accomplishing the proposed scenario for automatically building network models useful to answer biomedical questions.

Throughout working on this PhD thesis, it has become evident that in order to solve biomedical questions related to disease mechanisms and adverse drug reactions, an integrative analysis approach is the most promising for the task. In the introduction it has been pointed out that there are, however, currently several limitations to such integrative analysis approach. Biomedical data is typically dispersed over various databases each covering different aspects of the biomedical entities they are addressing. This has become especially visible when analysing public pathway repositories, as discussed above. Moreover, much of the information is still locked in the literature and has not made its way into such expert-curated databases, yet. Another issue is the lack of data standards allowing smooth and automatic integration. Hence, most work in this PhD thesis was directed towards the development of bioinformatics analysis tools for integration, visualization and analysis of biomedical data in order study the molecular basis of diseases and drug adverse reactions.

One example for which integration of different kinds of data is required is the study of disease mechanisms. In the last years, it has become evident that most human diseases arise due to complex interactions among multiple genetic variants and environmental factors variants (Hirschhorn *et al*, 2005). Hence, in order to study a particular disease all accessible information about the involved genetic and environmental variables has to be combined. Thus, we compiled a comprehensive gene-disease association database through integration of data from various expert-curated sources and text-mining derived associations. To the best of our knowledge this resource combines mendelian, complex and environmental diseases for the first time, and hence allowed us to study all these diseases at a global scale (see section

DISCUSSION

3.3). This repository serves as a suitable framework to study diseases and also adverse drug reactions because it incorporates the whole range of human diseases including those of environmental origin, which are caused by the exposure to xenobiotics including drugs. Moreover, it has to be mentioned here that a global analysis of human diseases has so far only been done for mendelian diseases and hence our integrated database provides a new framework for studying human diseases at a broader scale.

Our global analysis shows that the integration of data from various sources is required to obtain a complete picture of the genetic origin of human diseases. In particular, we demonstrate how text-mining derived associations close knowledge gaps existing in the expert-curated databases. Hence, the presented unified gene-disease association database can provide important biological insights that might not be discovered when considering each of the data sources independently. Clearly, this integration was not trivial. Next to mapping of diverse controlled vocabularies (i.e. for diseases), we developed a gene-disease association ontology for a smooth integration of the gene-disease associations. The use of this ontology on the one hand allowed us to correctly integrate different types of relationships found in the original databases. On the other hand it will facilitate future population of the ontology with further data, for instance derived from other text-mining approaches or additional online databases. Moreover, the mapping of disease vocabularies allowed the use of a standard disease classification schema to classify all diseases in our database. This is of particular interest for studies in which whole disease classes are investigated. In addition, by performing the mapping through the UMLS Metathesaurus, it is possible to represent the diseases with any other vocabulary covered by this resource. For example, vocabularies used in the clinical practice, such as SNOMED-CT, ICD-9 or ICD-10, could be used in order to allow an integration of data coming from patient records, for example. Furthermore, it allows using vocabularies in different languages, hence facilitating the integration with clinical data.

The global analysis also points out that human diseases have many gene associations in common indicating a highly shared genetic origin. In a next step, we extracted disease-related modules by means of clustering. Our findings confirm that the concept of modularity, which had been shown for mendelian diseases by other authors (Goh *et al*, 2007; Lim *et al*, 2006; Oti *et al*, 2007; van Driel *et al*, 2006), also applies to complex and environmental diseases. We furthermore demonstrate that most diseases are associated to a core set of biological processes indicating the importance of cross-talks between pathways in disease development. We believe that this has significant implications for disease treatment and drug development. A therapy that considers the diversity of biological processes related to a disease might be of advantage. More strikingly, similar findings are obtained when studying groups of diseases. Even for clusters of diseases, there is a set of core biological processes associated. This suggests that the diseases in these groups, which can be very similar but also very unrelated, might arise due to dysfunction of the same biological processes in the cell. This again has direct implications for disease treatment and drug development. If a set of diseases is related to the same

pathways, a treatment or drug already successful for one of the diseases could also be applied to the other diseases (Berger *et al*, 2009). Hence, we went a step further and identified these core biological processes. Moreover, our analysis shows that only few diseases are solely caused by defects in direct interactions between proteins and that for most diseases a set of core biological processes need to be studied in detail.

All in all, our integrated database serves as a suitable starting point for individual researchers studying a particular disease in detail because it combines all available gene-disease associations for this disease, even including associations only reported in the literature and not available in curated databases, yet. As introduced in section 1, there is a major need to investigate the interactions among environmental carcinogens and genetic factors (Sankaranarayanan *et al*, 2010). Our database allows such analysis since it integrates mendelian, complex and environmental diseases. Hence, the comprehensive gene-disease association database provides a valid source for the extraction of disease-related modules, which can be studied in detail to understand how they respond to genetic and environmental perturbations (see figure 1c). It can also serve as a framework to study disease relationships, for instance, when trying to understand comorbidities in patients. This is possible by studying disease clusters or neighbours in the disease projection network because their shared genetic origin could explain why they co-occur in patients. We are planning to further investigate in this direction by comparing data of comorbidities in patients derived from patient record databases with the found disease relationships in our networks. This is greatly facilitated by the use of disease vocabularies being part of the UMLS metathesaurus, which allows the direct integration with clinical data. Moreover, we showed in several use-case scenarios various other applications of the database. For instance, we used the gene clusters in conjunction with pathway data to predict new candidate disease genes in the networks derived solely from expert-curated databases. We then confirmed the prediction by checking the whole databases, which also includes text-mining derived data, and found the according literature evidence for the prediction. In another example, we established a connection between environmental exposure with arsenic compounds and cancer at the genetic level stressing the usefulness of the database to study the effect of environmental factors on human health. Finally, we showed how the database could be used to study adverse drug reactions.

In this regard, we have been working more deeply on a framework for the automatic substantiation of signals, which are unexpected associations of a clinical event with a given drug. For this purpose, we have implemented a webservice to query our integrated gene-disease association database allowing easy integration within analysis pipelines or workflows. Moreover, we have implemented an exemplary workflow for signal substantiation. It can be used to investigate biological explanations for adverse drug reactions. In this regard, the use of webservices allows the development of a variety of workflows for similar applications. In the near future we are planning to develop additional workflows and to submit our work regarding the automatic signal substantiation to a peer-

DISCUSSION

reviewed journal for publication. In addition, the workflow is currently used within the EU-ADR project for the analysis of signals.

In summary, we believe that our integrated gene-disease association database is of immense value for other biomedical researchers. In several example applications we have validated not only the usefulness of the database per se but the use of network analysis tools to uncover disease-related modules and to study disease mechanisms in detail. Hence, next to providing programmatic access to the database by means of a webservice, we developed DisGeNET, a plugin for Cytoscape (see section 3.5). DisGeNET represents a coherent tool for easy analysis and interpretation of human gene-disease networks. It allows user-friendly access to our integrated gene-disease association database by querying the data and creating gene-disease association networks within Cytoscape. DisGeNET assists the user in the interpretation and exploration of human diseases with respect to their genetic origin. Diverse options for generating subnetworks, as well as an advanced search tool, facilitate not only the analysis of single diseases but also the study of sets of diseases or certain disease classes specified through their associated genes.

It was explained before that in order to understand the molecular basis of diseases and adverse drug reactions, networks of the involved key players have to be investigated in detail. In addition, it is crucial to study how these networks respond to environmental influences including drugs. In this regard, static and dynamic network analysis approaches have been presented in the introduction. The advantage of such network models is clearly that they can be used not only to simulate the behaviour of biological processes in the cell but also to predict the effect of disruption or perturbation of these processes. This is of extreme value for studies of disease mechanisms or adverse drug reactions. Here the perturbation originates from genetic variations and environmental factors, including drugs, and modelling and predicting its effect might have direct implications for clinical practice.

It is known that many human diseases and also many adverse drug reactions are associated to genetic variants. For instance, it has been found that impaired drug metabolism can result from genetic variations in metabolizing enzymes. In most of the cases, however, the exact role of the genetic variants in disease development or the progression of the undesired side effects of the drug is not fully understood. Nevertheless, the advances in experimental technologies have allowed us to define the functional effect of such genetic variations onto the gene products. It is very likely that this effect on the function of the encoded protein will also affect further downstream processes, in which the protein is involved. Hence, in principle the integration of this information with the biological processes, in which the affected genes or proteins play a role, could yield a better understanding of the effect of the sequence variations on the whole process. In this context, in section 3.6, we have presented a general strategy for the integration of pathway and sequence variation data, towards their use for network visualization and analysis, as well as for the modelling of signalling pathways. The detailed understanding of the effect of sequence variations on biological processes in the human body will give important

information about the mechanisms underlying development or homeostasis. Moreover, if the sequence variation is known to be associated to a disease or individual drug response, the underlying mechanisms could also be uncovered. In this context, the integration of gene-disease association data is useful. In the gene-disease association database described before, we combine information about multiple kinds of relationships between genes and diseases including genetic variations. Moreover, information about biological processes is available by means of biological pathways in the aforementioned pathway databases. We succeeded in integrating the information about the functional effect of the sequence variations with the biological pathways by means of mapping it onto the nodes in the network representations of the pathways. Nevertheless, in order to really assess its affect on the involved biological reactions, a mapping onto the edges, which represent the actual reactions in a biological pathway, would be required. Herein, the basic problem is that the functional effect description and the biological reaction often differ from their level of granularity. For protein-protein interactions we achieved the mapping of the functional effect onto the network edges. In detail, if the textual description of the functional effect of a sequence variation contained both protein mentions, the functional effect was mapped onto the edge between the two proteins. However, for other interaction types, such as in signalling pathways, this mapping was more difficult. While the textual description of the functional effect is rather general (i.e. “leads to decreased protein activity”), the interactions in a pathway are described at the biochemical level. A careful evaluation showed that the fully automatic mapping goes beyond tasks that any current text mining system would be able to handle and that hence, manual intervention for a correct mapping would be required. Nevertheless, we proposed some strategies, which would make use of the fact that biological pathways are typically represented in the BioPAX format, an ontology providing detailed information that in principle could be used to allow such mapping. In this regard, future work is planned, for which the detailed analysis of pathway databases and pathway representation formats as presented in section 3.2 is of extreme value.

However, even considering the current limitations in fully automatically mapping the functional effects onto the biological reactions, the integration we achieved clearly aids the development of dynamic network models studying the effect on a systems level. Moreover, for working with protein-protein interaction networks, the approach provides a complete mapping. Moreover, we presented an example, in which we assessed the functional effect of a sequence variation onto the dynamics of a cell-signalling pathway. Such analysis can have practical applications for biomedical research. If the sequence variation has a clinical phenotypic effect (i.e. the sequence variation is associated with colon cancer) and the functional phenotypic effect is known (i.e. the sequence variation produces a decrease of enzymatic activity), the effect of the sequence variation can be evaluated in the context of the affected reactions and processes. This is a very important issue as it provides information about the functional effect of mutations at the cellular level that are relevant in the clinical practice. Moreover, in principle it would be possible to assess the effect of different sequence variations in the same model, an approach

DISCUSSION

particularly relevant to consider the polygenic character of complex diseases. Clearly, this can have significant consequences for understanding the mechanisms of diseases and adverse drug reactions. Overall, this approach addresses the identification of disease-related molecular networks that can then be studied with respect to genetic and environmental perturbations and how these perturbations affect the disease risk (see figure 1c).

4.3. Summary and outlook

In summary, in this PhD thesis we have developed several integrative analysis approaches that address typical biomedical problems regarding the understanding of the molecular basis of human diseases and adverse drug reactions. We have extensively studied data sources providing the needed information, as well as analysis approaches and concepts. In this regard, both, the more classic statistical approaches and the more recent systems biology approaches have been considered. We unravelled important limitations of current repositories and methods and developed several approaches that overcome these limitations. The developed methodologies and tools are all of immense value for biomedical researchers since they represent user-friendly applications that can be directly used to address typical biomedical problems. In addition, several new important biological findings were presented that were derived from applying the novel approaches with respect to studying the mechanisms underlying human diseases and adverse drug reactions. The practical applications of the novel methodologies developed during this PhD thesis confirm their practical usefulness for other biomedical researchers.

Moreover, we have pointed out some key points to be addressed in future work regarding the automatic generation of network models to answer specific questions. Hence, we would like to proceed in this direction by investigating in more detail network modelling techniques and by making use of all the methodologies developed throughout this PhD thesis.

It is greatly hoped that the here presented methodologies and tools will eventually aid to improve our understanding of the molecular basis of diseases and adverse drug reactions and that this will bring us a step closer to an improved clinical practice being personalized and preventive.

5. CONCLUSIONS

CONCLUSIONS

1. A novel multilevel statistical method for its application in drug discovery projects has been developed. It is based on the sequential building of linked multivariate statistical models, where each model introduces a different level of drug description. By using antipsychotic drugs and metabolic side effects as example, the usefulness of the method was demonstrated.
2. Public pathway repositories have been critically evaluated. We assessed the suitability of the pathways to be used to automatically build network models useful to answer biomedical questions. This evaluation showed that in principle the available data is accurate enough to be directly used, though currently not fully automatically. All limitations have been critically discussed and possible solutions have been proposed.
3. The integration of gene-disease association data from various sources including expert-curated online databases and text-mining derived associations resulted in a comprehensive gene-disease association database. It has been shown that this new resource closes existing knowledge gaps in the original databases combining mendelian, complex and environmental diseases.
4. The development of a gene-disease association ontology made possible an accurate integration of gene-disease association data from diverse sources. The ontology will be useful for future projects, for which the ontology can be populated with additional data from literature or other sources.
5. The detailed analysis of the global properties of a network representation of the integrated gene-disease association database supports the concept of modularity for mendelian, complex and environmental human diseases. The combination of network and pathway analysis approaches allowed the identification of core biological processes related to human diseases and adverse drug reactions, which will aid future studies on both of them.
6. We have shown several applications of the newly developed gene-disease association database. They include the discovery of novel gene-disease associations, the identification of shared mechanisms of different diseases, the study of the relationship of environmental and genetic factors in disease development, and the investigation of adverse drug reactions. Moreover, future applications for studies on the molecular basis of comorbidities in patients and for drug repurposing have been proposed.

CONCLUSIONS

7. A webservice allowing programmatic access to our integrated gene-disease association has been developed. It has been integrated into a workflow that combines information about associations between genes and adverse drug reactions with data about drugs and their targets. The proposed workflow will be of great usefulness for investigating the molecular mechanisms of adverse drug reactions.
8. The implementation of DisGeNET, a Cytoscape plugin for user-friendly access to our integrated gene-disease association database, results in a useful tool for the biomedical community, supporting the studies on the molecular basis of human diseases and adverse drug reactions. The detailed user guide incorporating several use-cases has been developed to further support the user.
9. The integration of information about sequence variations and their functional effect with biological pathway data allows the development of network models, which can directly assess the effect of the variations on the dynamics of the biological processes. This is of particular value if the sequence variations are known to be associated to a disease or an adverse drug reaction, because in such cases the effect of the genetic variations on the biological processes can give mechanistic explanations for the disease or the adverse drug reactions.

6. LIST OF PUBLICATIONS

Articles

1. **Bauer-Mehren A**, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. "Network analysis of an integrated gene-disease association database reveals functional modules in human disease." *submitted*. (2010).
2. **Bauer-Mehren A**, Rautschka M, Sanz F, Furlong LI. DisGeNET - a "Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks." *submitted*. (2010).
3. Bundschus M, **Bauer-Mehren A**, Furlong L, Tresp V, Kriegel H-P, editors. "Digging for Knowledge with Information Extraction: A Case Study on Human Gene-Disease Associations." *19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*, 2010 26-30 October 2010, Toronto, Canada.
4. Selent J, **Bauer-Mehren A**, López L, Loza MI, Sanz F, Pastor M. "A novel multilevel statistical method for the study of the relationships between multireceptorial binding affinity profiles and in vivo endpoints." *Molecular Pharmacology*. (2010), 77(2):149-58.
5. Risselada R, Lingsma HF, **Bauer-Mehren A**, Friedrich CM, Molyneux AJ, Kerr RSC, et al. "Prediction of 60 day case-fatality after aneurysmal subarachnoid haemorrhage: results from the International Subarachnoid Aneurysm Trial (ISAT)." *European Journal of Epidemiology*. (2010).
6. **Bauer-Mehren A**, Furlong L, Rautschka M, Sanz F. "From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways." *BMC Bioinformatics*. (2009), 10(Suppl 8):S6-S.
7. **Bauer-Mehren A**, Furlong LI, Sanz F. "Pathway databases and tools for their exploitation: benefits, current limitations and challenges." *Molecular Systems Biology*. (2009), 5:290.

Oral communications

8. **Bauer-Mehren A**, “The EBI Bioinformatics roadshow – Perth”, Invited speaker/trainer for the topic “Integration of biological annotations and networks using Cytoscape”, Murdoch University, Perth, Australia, November 15-19, 2010
9. Avillach P, Coloma PM, **Bauer-Mehren A**, Gini R, Thiessard F, Trifiro G, Schuemie MJ, Sturkenboom M, Oliveira JL, Sanz F, Molero E, Diaz C, Van der Lei J on behalf of the EU-ADR group, “Design, development and validation of a computerised system that exploits data from electronic health records and biomedical information for the early detection of adverse drug reaction. The EU-ADR project: Preliminary results”, *Workshop at the Medinfo 2010 conference*, Cape Town, South Africa, September 12-15, 2010
10. **Bauer-Mehren A**, van Mullingen E, Avillach P, Coloma P, Garcia-Serna R, Herings R, Sturkenboom M, Boyer S, Sanz F, Mestres J, Kors J, and Furlong LI, “Substantiation of drug safety signals: taking automated signal detection a step further”, *26th International Conference on Pharmacoepidemiology and Therapeutic Risk Management*, Brighton, UK, August 18-22, 2010
11. **Bauer-Mehren A**, “EMBO Practical Course ‘In silico systems biology: network reconstruction, analysis and network based modelling’”, Invited speaker/trainer for the network analysis workshop, EMBL-EBI, Hinxton, Cambridge, UK, April 10-13, 2010
12. **Bauer-Mehren A**, Furlong LI, Sanz F, “From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways”, *10th Annual BioPathways Meeting*, special interest group meeting within the 17th ISMB and 8th ECCB, Stockholm, Sweden, June 27-July 2, 2009
13. Risselada R, Friedrich CM, Ebeling C, Klinger R, **Bauer-Mehren A**, “Workflows for Data Mining in Integrated multi-modal Data of Intracranial Aneurysms using KNIME”, *useR! Conference*, Rennes, France, July 8-10, 2009
14. Selent J, López L, **Bauer-Mehren A**, Fernández C, Loza MI, Sanz F, Pastor M, “Integrated Approach for the Multireceptorial Design of Antipsychotic Agents”, *XXth International Symposium on Medicinal Chemistry*, Vienna, Austria, August 31 – September 4, 2008

Poster communications

15. **Bauer-Mehren A**, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI, “Combination of network topology and pathway analysis to reveal functional modules in human disease”, *11th International Conference on Systems Biology*, Edinburgh, Scotland, UK, October 11-14, 2010 and *9th European Conference on Computational Biology (ECCB)*, Ghent, Belgium, September 26-29, 2010
16. **Bauer-Mehren A**, Rautschka M, Sanz F, Furlong LI, “DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks”, *9th European Conference on Computational Biology (ECCB)*, Ghent, Belgium, September 26-29, 2010
17. Friedrich CM, Ebeling C, Furlong LI, Fornes O, **Bauer-Mehren A**, Oliva B, Sanz F, Engelbrecht S, Yilmaz S, McGregor J, Cambien F, Hofmann-Apitius M, “@neuLink – Knowledge Discovery from Structured and Unstructured Data Sources”, *Virtual Physiological Human (VPH) Initiative day*, Brussels, Belgium, September 9, 2009
18. **Bauer-Mehren A**, Sanz F, Furlong LI, “Comprehensive view of pathways related to complex diseases and their cross-talks”, *17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 8th European Conference on Computational Biology (ECCB)*, Stockholm, Sweden, June 27-July 2, 2009
19. **Bauer-Mehren A**, Loza MI, Sanz F and Pastor M, “Application of multilevel multivariate data analysis in drug design. The future of QSAR?”, *17th European Symposium on Quantitative Structure-Activity Relationships & Omics Technologies and Systems Biology*, Uppsala, Sweden, September 21-16, 2008
20. **Bauer-Mehren A**, Loza MI, Sanz F and Pastor M, “Application of multilevel multivariate data analysis in drug design”, *XXth International Symposium on Medicinal Chemistry EFMC-ISMC*, Vienna, Austria, August 31 – September 4, 2008
21. **Bauer-Mehren A**, Selent J, López L, Sanz F, Pastor M, “Combination of direct and indirect approaches to study the D2/5-HT2A selectivity of antipsychotic drugs”, *22nd "Darmstadt" Molecular Modelling Workshop*, Erlangen, Germany, April 29-30, 2008

LIST OF PUBLICATIONS

7. REFERENCES

REFERENCES

- Adriaens, M.E., Jaillard, M., Waagmeester, A., Coort, S.L.M., Pico, A.R. & Evelo, C.T.A. The public road to high-quality curated biological pathways. *Drug Discov. Today* **13**, 856-862 (2008).
- Aittokallio, T. & Schwikowski, B. Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.* **7**, 243-255 (2006).
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. *Molecular Biology of the Cell*. Garland, (2007).
- Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450-461 (2007).
- Altman, R.B. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat. Genet.* **39**, 426 (2007).
- Alves, R., Antunes, F. & Salvador, A. Tools for kinetic modeling of biochemical networks. *Nat. Biotechnol.* **24**, 667-672 (2006).
- Ananiadou, S., Freidman, C. & Tsujii, J.i. Introduction: named entity recognition in biomedicine. *J. Biomed. Inf.* **37**, 393-395 (2004).
- Ananiadou, S., Pyysalo, S., Tsujii, J.i. & Kell, D.B. Event extraction for systems biology by text mining the literature. *Trends Biotechnol.* **28**, 381-390 (2010).
- Antezana, E.Z., Kuiper, M. & Mironov, V. Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief. Bioinform.* **10**, 392-407 (2009).
- Apweiler, R. *et al* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115-119 (2004).
- Armstrong, K., Calzone, K., Stopfer, J., Fitzgerald, G., Coyne, J. & Weber, B. Factors Associated with Decisions about Clinical BRCA1/2 Testing. *Cancer Epidemiol. Biomarkers Prev.* **9**, 1251-1254 (2000).
- Arrell, D.K. & Terzic, A. Network Systems Biology for Drug Discovery. *Clin. Pharmacol. Ther.* **88**, 120-125 (2010).
- Ashburner, M. *et al* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25-29 (2000).
- Auffray, C., Chen, Z. & Hood, L. Systems medicine: the future of medical genomics and healthcare. *Genome Med.* **1**, 2 (2009).
- Bader, G.D., Cary, M.P. & Sander, C. Pathguide: a pathway resource list. *Nucleic Acids Res.* **34**, D504-506 (2006).
- Bagley, S.C., White, H. & Golomb, B.A. Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain. *J. Clin. Epidemiol.* **54**, 979-985 (2001).
- Barabási, A.-L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101-113 (2004).
- Bauer-Mehren, A., Bundschuh, M., Rautschka, M., Mayer, M.A., Sanz, F. & Furlong, L.I. Network analysis of an integrated gene-disease association database reveals functional modules in mendelian, complex and environmental disease. *submitted* (2010a).
- Bauer-Mehren, A., Furlong, L., Rautschka, M. & Sanz, F. From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. *BMC Bioinformatics* **10**, S6-S6 (2009a).

REFERENCES

- Bauer-Mehren, A., Furlong, L.I. & Sanz, F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol.* **5**, 290 (2009b).
- Bauer-Mehren, A., Rautschka, M., Sanz, F. & Furlong, L.I. DisGeNET - a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *submitted* (2010b).
- Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. & Morissette, J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inf.* **41**, 706-716 (2008).
- Bentele, M., Lavrik, I., Ulrich, M., Stößer, S., Heermann, D.W., Kalthoff, H., Krammer, P.H. & Eils, R. Mathematical modeling reveals threshold mechanism in CD95-induced apoptosis. *J. Cell Bio.* **166**, 839-851 (2004).
- Berger, S.I. & Iyengar, R. Network analyses in systems pharmacology. *Bioinformatics* **25**, 2466-2472 (2009).
- Bergman, R.N., Ider, Y.Z., Bowden, C.R. & Cobelli, C. Quantitative estimation of insulin sensitivity. *Am. J. Physiol. Gastrointest. Liver Physiol.* **236**, G667 (1979).
- Birtwistle, M.R., Hatakeyama, M., Yumoto, N., Ogunnaike, B.A., Hoek, J.B. & Kholodenko, B.N. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. *Mol. Syst. Biol.* **3**, 144 (2007).
- Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucl. Acids Res.* **32**, D267-270 (2004).
- Bodenreider, O. & Stevens, R. Bio-ontologies: current trends and future directions. *Brief. Bioinform.* **7**, 256-274 (2006).
- Borisov, N. *et al* Systems-level interactions between insulin-EGF networks amplify mitogenic signaling. *Mol. Syst. Biol.* **5**, 256-256 (2009).
- Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, 228-237 (2003).
- Boulesteix, A.-L. & Strimmer, K. Predicting transcription factor activities from combined analysis of microarray and CHIP data: a partial least squares approach. *Theor. Biol. Med. Model* **2**, 23 (2005).
- Brazma, A. *et al* Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat. Genet.* **29**, 365-371 (2001).
- Bundschuh, M., Bauer-Mehren, A., Furlong, L., Tresp, V. & Kriegel, H.-P. in 19th ACM International Conference on Information and Knowledge Management CIKM Toronto, Canada, (2010).
- Bundschuh, M., Dejori, M., Stetter, M., Tresp, V. & Kriegel, H.-P. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics* **9**, 207-207 (2008).
- Butcher, E.C., Berg, E.L. & Kunkel, E.J. Systems biology in drug discovery. *Nat. Biotechnol.* **22**, 1253-1259 (2004).
- Butts, C.T. Revisiting the Foundations of Network Analysis. *Science* **325**, 414-416 (2009).

- Celik, I. *et al* Arsenic in drinking water and lung cancer: a systematic review. *Environ. Res.* **108**, 48-55 (2008).
- Cerami, E., Demir, E., Schultz, N., Taylor, B.S. & Sander, C. Automated Network Analysis Identifies Core Pathways in Glioblastoma. *PLoS ONE* **5** (2010).
- Chaouiya, C., Remy, E. & Thieffry, D. Petri net modelling of biological regulatory networks. *Journal of Discrete Algorithms* **6**, 165-177 (2008).
- Chaumont, S., Compan, V., Toulme, E., Richler, E., Housley, G.D., Rassendren, F. & Khakh, B.S. Regulation of P2X2 Receptors by the Neuronal Calcium Sensor VILIP1. *Sci. Signal.* **1**, ra8 (2008).
- Chen, W.W., Schoeberl, B., Jasper, P.J., Niepel, M., Nielsen, U.B., Lauffenburger, D.A. & Sorger, P.K. Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol.* **5** (2009).
- Chin, L. & Gray, J.W. Translating insights from the cancer genome into clinical practice. *Nature* **452**, 553-563 (2008).
- Chiou, H.Y., Hsueh, Y.M., Liaw, K.F., Horng, S.F., Chiang, M.H., Pu, Y.S., Lin, J.S., Huang, C.H. & Chen, C.J. Incidence of internal cancers and ingested inorganic arsenic: a seven-year follow-up study in Taiwan. *Cancer Res.* **55**, 1296-1300 (1995).
- Cochrane, G.R. & Galperin, M.Y. The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Res.* **38**, D1-4 (2010).
- Cohen, K.B., Palmer, M. & Hunter, L. Nominalization and Alternations in Biomedical Language. *PLoS ONE* **3** (2008).
- Cokol, M., Iossifov, I., Weinreb, C. & Rzhetsky, A. Emergent behavior of growing knowledge about molecular interactions. *Nat. Biotechnol.* **23**, 1243-1247 (2005).
- Cote, R.A. & Robboy, S. Progress in Medical Information Management: Systematized Nomenclature of Medicine (SNOMED). *JAMA* **243**, 756-762 (1980).
- Dezi, C., Brea, J., Alvarado, M., Raviña, E., Masaguer, C.F., Loza, M.I., Sanz, F. & Pastor, M. Multistructure 3D-QSAR Studies on a Series of Conformationally Constrained Butyrophenones Docked into a New Homology Model of the 5-HT_{2A} Receptor. *J. Med. Chem.* **50**, 3242-3255 (2007).
- Donaldson, I. *et al* PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* **4**, 11 (2003).
- Dumontier, M. & Villanueva-Rosales, N. Towards pharmacogenomics knowledge discovery with the semantic web. *Brief. Bioinform.* **10**, 153-163 (2009).
- Edwards, R. & Aronson, J.K. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet* **356**, 1212 (2000).
- Fundel, K., Kuffner, R. & Zimmer, R. RelEx--Relation extraction using dependency parse trees. *Bioinformatics* **23**, 365-371 (2007).

REFERENCES

- Furlong, L., Dach, H., Hofmann-Apitius, M. & Sanz, F. OSIRISv1.2: A named entity recognition system for sequence variants of genes in biomedical literature. *BMC Bioinformatics* **9**, 84-84 (2008).
- Fussenegger, M., Bailey, J.E. & Varner, J. A mathematical model of caspase function in apoptosis. *Nat Biotech* **18**, 768-774 (2000).
- Futreal, P.A. *et al* BRCA1 mutations in primary breast and ovarian carcinomas. *Science* **266**, 120-122 (1994).
- Goble, C. & Stevens, R. State of the nation in data integration for bioinformatics. *J. Biomed. Inf.* **41**, 687-693 (2008).
- Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. & Barabási, A.-L. The human disease network. *Proc. Natl. Acad. Sci.* **104**, 8685-8690 (2007).
- Grant, S.F.A. *et al* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320-323 (2006).
- Gruber, T.R. A translation approach to portable ontology specifications. *Knowl. Acquis.* **5**, 199-220 (1993).
- Gurwitz, D. & Motulsky, A.G. 'Drug reactions, enzymes, and biochemical genetics': 50 years later. *Pharmacogenomics* **8**, 1479-1484 (2007).
- Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P. & Kasprzyk, A. BioMart Central Portal--unified access to biological data. *Nucl. Acids Res.* **37**, W23-27 (2009).
- Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-517 (2005).
- Han, J.-D.J. Understanding biological functions through molecular networks. *Cell Res.* **18**, 224-237 (2008).
- Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R. & Fluck, J. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics* **6 Suppl 1**, S14 (2005).
- Harre, F.E., Lee, K.L. & Pollock, B.G. Regression Models in Clinical Studies: Determining Relationships Between Predictors and Response. *J. Natl. Cancer Inst.* **80**, 1198-1202 (1988).
- Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47-52 (1999).
- Hatakeyama, M. *et al* A computational model on the modulation of mitogen-activated protein kinase (MAPK) and Akt pathways in heregulin-induced ErbB signalling. *Biochem. J.* **373**, 451-463 (2003).
- Helgadottir, A. *et al* A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. *Science* **316**, 1491-1493 (2007).
- Hendriks, B.S. Functional pathway pharmacology: chemical tools, pathway knowledge and mechanistic model-based interpretation of experimental data. *Curr. Opin. Chem. Biol.* **14**, 489-497 (2010).
- Hendriks, B.S. *et al* Decreased internalisation of erbB1 mutants in lung cancer is linked with a mechanism conferring sensitivity to gefitinib. *Syst. Biol. (Stevenage)* **153**, 457-466 (2006).

- Herbert, A. *et al* A Common Genetic Variant Is Associated with Adult and Childhood Obesity. *Science* **312**, 279-283 (2006).
- Hersh, W. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. *Brief. Bioinform.* **6**, 344-356 (2005).
- Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95-108 (2005).
- Hoops, S. *et al* COPASI--a COMplex PATHway Simulator. *Bioinformatics* **22**, 3067-3074 (2006).
- Hornberg, J.J., Bruggeman, F.J., Binder, B., Geest, C.R., de Vaate, A.J.M.B., Lankelma, J., Heinrich, R. & Westerhoff, H.V. Principles behind the multifarious control of signal transduction. ERK phosphorylation and kinase/phosphatase control. *The FEBS Journal* **272**, 244-258 (2005).
- Howe, D. *et al* Big data: The future of biocuration. *Nature* **455**, 47-50 (2008).
- Howe, T.J., Mahieu, G., Marichal, P., Tabruyn, T. & Vugts, P. Data reduction and representation in drug discovery. *Drug Discov. Today* **12**, 45-53 (2007).
- Hucka, M. *et al* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524-531 (2003).
- Hunter, L. & Cohen, K.B. Biomedical language processing: what's beyond PubMed? *Mol. Cell* **21**, 589-594 (2006).
- Ideker, T., Galitski, T. & Hood, L. A NEW APPROACH TO DECODING LIFE: Systems Biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343-372 (2001).
- Jensen, L.J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* **7**, 119-129 (2006).
- Jones, S. *et al* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science* **321**, 1801-1806 (2008).
- Jørgensen, J.T. & Winther, H. The new era of personalized medicine: 10 years later. *Personalized Medicine* **6**, 423-428 (2009).
- Joshi-Tope, G. *et al* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428-432 (2005).
- Kanehisa, M. *et al* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480-484 (2008).
- Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).
- Kann, M.G. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief. Bioinform.* **11**, 96-110 (2010).
- Kansal, A.R. Modeling approaches to type 2 diabetes. *Diabetes Technol. Ther.* **6**, 39-47 (2004).
- Kauffman, S.A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437-467 (1969).
- Kennedy, R.L., Fraser, H.S., McStay, L.N. & Harrison, R.F. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. *Eur. Heart J.* **17**, 1181-1191 (1996).

REFERENCES

- Kerrien, S. *et al* IntAct--open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561-565 (2007).
- Kholodenko, B.N., Demin, O.V., Moehren, G. & Hoek, J.B. Quantification of Short Term Signaling by the Epidermal Growth Factor Receptor. *J. Biol. Chem.* **274**, 30169-30181 (1999).
- Klamt, S., Saez-Rodriguez, J. & Gilles, E. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology* **1**, 2 (2007).
- Klein, T.E. *et al* Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J.* **1**, 167-170 (2001).
- Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L. & Valencia, A. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol.* **9**, S1 (2008).
- Lander, E.S. *et al* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
- Lange, J.H.M., Reinders, J.H., Tolboom, J.T.B.M., Glennon, J.C., Coolen, H.K.A.C. & Kruse, C.G. Principal Component Analysis Differentiates the Receptor Binding Profiles of Three Antipsychotic Drug Candidates from Current Antipsychotic Drugs. *J. Med. Chem.* **50**, 5103-5108 (2007).
- Le Novère, N. *et al* BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* **34**, D689-691 (2006).
- Le Novère, N. *et al* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* **23**, 1509-1515 (2005).
- Lee, D.S., Park, J., Kay, K.A., Christakis, N.A., Oltvai, Z.N. & Barabási, A.L. The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci.* **105**, 9880-9885 (2008).
- Lee, T., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D., Tenenbaum, J. & Karp, P. BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics* **7**, 170 (2006).
- Li, H., Ung, C.Y., Ma, X.H., Li, B.W., Low, B.C., Cao, Z.W. & Chen, Y.Z. Simulation of crosstalk between small GTPase RhoA and EGFR-ERK signaling pathway via MEKK1. *Bioinformatics* **25**, 358-364 (2009a).
- Li, Y. & Agarwal, P. A Pathway-Based View of Human Diseases and Disease Relationships. *PLoS ONE* **4**, e4346 (2009b).
- Liebler, D.C. & Guengerich, F.P. Elucidating mechanisms of drug-induced toxicity. *Nat. Rev. Drug Discov.* **4**, 410-420 (2005).
- Lim, J. *et al* A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration. *Cell* **125**, 801-814 (2006).
- Lloyd, C.M., Halstead, M.D.B. & Nielsen, P.F. CellML: its future, present and past. *Prog. Biophys. Mol. Biol.* **85**, 433-450.
- Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **35**, D26-31 (2006).

- Mattingly, C.J., Rosenstein, M.C., Davis, A.P., Colby, G.T., Forrest, J.N. & Boyer, J.L. The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.* **92**, 587-595 (2006).
- McPherson, R. *et al* A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science* **316**, 1488-1491 (2007).
- Miki, Y. *et al* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66-71 (1994).
- Mishra, G.R. *et al* Human Protein Reference Database - 2006 Update. *Nucleic Acids Res.* **34**, D411-414 (2006).
- Nahler, G. in Dictionary of Pharmaceutical Medicine. 96(2009).
- Nigsch, F., Macaluso, N.J.M., Mitchell, J.B.O. & Zmuidinavicius, D. Computational toxicology: an overview of the sources of data and of modelling methods. *Expert Opin. Drug Metab. Toxicol.* **5**, 1-14 (2009).
- Oda, K., Kim, J.-D., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y. & Tsujii, J.i. New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinformatics* **9**, S5 (2008).
- Oinn, T., Addis, M., Ferris, J., Marvin, D., Greenwood, M., Carver, T., Pocock, M.R., Wipat, A. & Li, P. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045 - 3054 (2004).
- Oti, M. & Brunner, H.G. The modular nature of genetic diseases. *Clin. Genet.* **71**, 1-11 (2007).
- Oti, M., Snel, B., Huynen, M.A. & Brunner, H.G. Predicting disease genes using protein-protein interactions. *J. Med. Genet.* **43**, 691-698 (2006).
- Park, S., Bae, J., Nam, B.-H. & Yoo, K.-Y. Aetiology of cancer in Asia. *Asian Pac. J. Cancer Prev.* **9**, 371-380 (2008).
- Pérez-Enciso, M. & Tenenhaus, M. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum. Genet.* **112**, 581-592 (2003).
- Pirmohamed, M. & Park, B.K. Genetic susceptibility to adverse drug reactions. *Trends Pharmacol. Sci.* **22**, 298-305 (2001).
- Przulj, N., Wigle, D.A. & Jurisica, I. Functional topology in a network of protein interactions. *Bioinformatics* **20**, 340-348 (2004).
- Radosavljević, V. & Jakovljević, B. Arsenic and bladder cancer: observations and suggestions. *J. Environ. Health* **71**, 40-42 (2008).
- Risselada, R. *et al* Prediction of 60 day case-fatality after aneurysmal subarachnoid haemorrhage: results from the International Subarachnoid Aneurysm Trial (ISAT). *Eur. J. Epidemiol.* (2010).
- Roth, B.L., Sheffler, D.J. & Kroeze, W.K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3**, 353-359 (2004).
- Saez-Rodriguez, J. *et al* A Logical Model Provides Insights into T Cell Receptor Signaling. *PLoS Comp. Biol.* **3** (2007).
- Samaga, R., Saez-Rodriguez, J., Alexopoulos, L.G., Sorger, P.K. & Klamt, S. The Logic of EGFR/ErbB Signaling: Theoretical Properties and Analysis of High-Throughput Data. *PLoS Comp. Biol.* **5** (2009).

REFERENCES

- Sankaranarayanan, R. & Boffetta, P. Research on cancer prevention, detection and management in low- and medium-income countries. *Ann. Oncol.* (2010).
- Schadt, E.E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218-223 (2009).
- Schaefer, L. An Introduction to the NCI Pathway Interaction Database. NCI-Nature Pathway Interaction Database. (2006).
- Schoeberl, B., Eichler-Jonsson, C., Gilles, E.D. & Muller, G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.* **20**, 370-375 (2002).
- Schoeberl, B. *et al* Therapeutically Targeting ErbB3: A Key Node in Ligand-Induced Activation of the ErbB Receptor-PI3K Axis. *Sci. Signal.* **2**, ra31 (2009).
- Schuemie, M.J., Jelier R & Kors, J.A. in Second BioCreative Challenge Evaluation Workshop. 131-133, (2007).
- Scriver, C.R. & Waters, P.J. Monogenic traits are not simple: lessons from phenylketonuria. *Trends Genet.* **15**, 267-272 (1999).
- Searls, D.B. Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* **4**, 45-58 (2005).
- Selent, J., Bauer-Mehren, A., López, L., Loza, M.I., Sanz, F. & Pastor, M. A novel multilevel statistical method for the study of the relationships between multireceptorial binding affinity profiles and in vivo endpoints. *Mol. Pharmacol.* **77**, 149-158 (2010).
- Sewell, W. MEDICAL SUBJECT HEADINGS IN MEDLARS. *Bull. Med. Libr. Assoc.* **52**, 164-170 (1964).
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498-2504 (2003).
- Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3** (2007).
- Sharma, S.V., Bell, D.W., Settleman, J. & Haber, D.A. Epidermal growth factor receptor mutations in lung cancer. *Nat. Rev. Cancer* **7**, 169-181 (2007).
- Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64-68 (2002).
- Sladek, R. *et al* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-885 (2007).
- Slater, T., Bouton, C. & Huang, E.S. Beyond data integration. *Drug Discov. Today* **13**, 584-589 (2008).
- Smith, A.H., Hopenhayn-Rich, C., Bates, M.N., Goeden, H.M., Hertz-Picciotto, I., Duggan, H.M., Wood, R., Kosnett, M.J. & Smith, M.T. Cancer risks from arsenic in drinking water. *Environ. Health Perspect.* **97**, 259-267 (1992).
- Suderman, M. & Hallett, M. Tools for visually exploring biological networks. *Bioinformatics* **23**, 2651-2659 (2007).

- Suthram, S., Dudley, J.T., Chiang, A.P., Chen, R., Hastie, T.J. & Butte, A.J. Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. *PLoS Comp. Biol.* **6** (2010).
- Tan, Y., Shi, L., Tong, W., Gene Hwang, G.T. & Wang, C. Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Comput. Biol. Chem.* **28**, 235-243 (2004).
- Taniguchi, C.M., Armstrong, S.R., Green, L.C., Golan, D.E. & Tashjian, A.H., Jr. in *Pharmacology: The Pathophysiologic Basis of Drug Therapy*, Edn. 2. Lippincott Williams & Wilkins, Philadelphia, PA, (2007).
- Tarbell, J.M. & Ebong, E.E. The Endothelial Glycocalyx: A Mechano-Sensor and -Transducer. *Sci. Signal.* **1**, pt8 (2008).
- The UniProt, C. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **38**, D142-148 (2010).
- Thun, M.J., DeLancey, J.O., Center, M.M., Jemal, A. & Elizabeth, M.W. The global burden of cancer: priorities for prevention. *Carcinogenesis* **31**, 100-110 (2010).
- Tsuda, T., Babazono, A., Yamamoto, E., Kurumatani, N., Mino, Y., Ogawa, T., Kishi, Y. & Aoyama, H. Ingested arsenic and internal cancer: a historical cohort study followed for 33 years. *Am. J. Epidemiol.* **141**, 198-209 (1995).
- Utrecht, J. Idiosyncratic Drug Reactions: Current Understanding. *Annu. Rev. Pharmacol. Toxicol.* **47**, 513-539 (2007).
- van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G. & Leunissen, J.A.M. A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* **14**, 535-542 (2006).
- Venter, J.C. *et al* The Sequence of the Human Genome. *Science* **291**, 1304-1351 (2001).
- Wilke, R.A., Lin, D.W., Roden, D.M., Watkins, P.B., Flockhart, D., Zineh, I., Giacomini, K.M. & Krauss, R.M. Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. *Nat. Rev. Drug Discov.* **6**, 904-916 (2007).
- Wold, H. Soft modeling: the basic design and some extensions. *Systems under indirect observation* **2**, 589-591 (1982).
- Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics Intellig. Lab. Syst.* **58**, 109-130 (2001a).
- Wold, S., Trygg, J., Berglund, A. & Antti, H. Some recent developments in PLS modeling. *Chemometrics Intellig. Lab. Syst.* **58**, 131-150 (2001b).
- Yaffe, M.B. Signaling Networks and Mathematics. *Sci. Signal.* **1**, eg7 (2008).
- Yang, C.-Y., Chang, C.-C. & Chiu, H.-F. Does arsenic exposure increase the risk for prostate cancer? *J. Toxicol. Environ. Health A* **71**, 1559-1563 (2008).
- Yeung, K.Y. & Ruzzo, W.L. Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763-774 (2001).
- Yildirim, M.A., Goh, K.-I., Cusick, M.E., Barabasi, A.-L. & Vidal, M. Drug-target network. *Nat. Biotechnol.* **25**, 1119-1126 (2007).

REFERENCES

- Zaghloul, N.A. & Katsanis, N. Functional modules, mutational load and human genetic disease. *Trends Genet.* **26**, 168-176 (2010).
- Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. & Cesareni, G. MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135-140 (2002).
- Zweigenbaum, P., Demner-Fushman, D., Yu, H. & Cohen, K.B. Frontiers of biomedical text mining: current progress. *Brief. Bioinform.* **8**, 358-375 (2007).