“The Clever Machine” – a computational tool for dataset exploration and prediction

Petr Klus

TESI DOCTORAL UPF / ANY 2016

DIRECTOR DE LA TESI

Dr. Gian Gaetano Tartaglia

DEPARTAMENT OF BIOINFORMATICS AND GENOMICS
AT CENTER FOR GENOMIC REGULATION (CRG)

Věnováno mým největším fanouškům,
rodičům Petrovi a Marcele
a prarodičům Vladimíru a Věře,
bez nich bych nebyl tím, kým jsem.

Acknowledgments

I have to thank my lucky stars for helping me find the opening in Barcelona. It was late 2010, I was in the middle of my master's degree and I was not quite sure where will the next chapter of my life take me and being notoriously bad at geography, I pulled up a map of Europe. As it is my usual trait, I went the other way around than most people would – I first cross-referenced places I would have liked, and only then looked if there are any research centres that would peak my interest (and where is the possibility of such feeling being mutual). Barcelona seemed to have ticked all the boxes and I started researching open positions. After exchanging a few emails, I learned about a fellowship and there it was, I was flying in for an interview! Coming from the UK and hardened by frequent barefoot walks in crispy winter countryside, Barcelona in February seemed very nice and warm. During my interview process, I met with Gian and his group – and from the first moment I knew that this is the place for me to be.

Starting in October 2011, everything seemed great – the city was great, people in the centre were incredible and I started learning the proverbial bioinformatical ropes. Little did I know that one of the biggest challenges of my life was yet to come due to sudden change of my personal circumstances. They say that difficult situations show person's true colours and I will be forever grateful for all of the help and understanding extended to me from Gian and the rest of the lab in the months that followed. I would have not been able to continue my PhD without the help I have so unconditionally received. I would like to thank Federico, Davide, Silvia, Mimma, Nieves and Benni for their support with all things science (and for allowing a Computer Scientist into their midst). I would also like to thank those that joined us a little bit later, namely Riccardo, Teresa, Joana, Natalia and, of course, our two latest talents Fernando and Alex. I would also like to thank my compañeros Marcos, Michael, Birgit, Shalu and Thomas who have spent quality time with me, be it during crazy trips to Andorra or just hanging out in Montgat.

As the last and biggest thank you, I would like to extend enormous amount of gratitude to Zuzana for standing next to me during the happy and tough times alike and providing more love than I could ever comprehend.

Abstract

The purpose of my doctoral studies was to develop an algorithm for large-scale analysis of protein sets. This thesis outlines the methodology and technical work performed as well as relevant biological cases involved in creation of the core algorithm, the *clever*Machine (CM), and its extensions *multiClever*Machine (mCM) and *clever*GO. The CM and mCM provide characterisation and classification of protein groups based on physico-chemical features, along with protein abundance and Gene Ontology annotation information, to perform an accurate data exploration. My method provides both computational and experimental scientists with a comprehensive, easy to use interface for high-throughput protein sequence screening and classification.

Resumen

El propósito de mis estudios doctorales era desarrollar un algoritmo para el análisis a gran escala de conjuntos de datos de proteínas. Esta tesis describe la metodología, el trabajo técnico desarrollado y los casos biológicos envueltos en la creación del algoritmo principal –el *clever*Machine (CM) y sus extensiones *multiClever*Machine (mCM) y *clever*GO. El CM y mCM permiten la caracterización y clasificación de grupos de proteínas basados en características físico-químicas, junto con la abundancia de proteínas y la anotación de ontología de genes, para así elaborar una exploración de datos correcta. Mi método está compuesto por científicos tanto computacionales como experimentales con una interfaz amplia, fácil de usar para un monitoreo y clasificación de secuencia de proteínas de alto rendimiento.

Preface

The work carried out during my doctoral studies was focused on developing high-throughput methods for protein dataset analysis. My main aim was to develop a methodology to extract high-level features (such as propensity to aggregate or form secondary structures) using physico-chemical scales[1], which was exploited in the development of the *clever*Machine (Chapter I). Using an ensemble of machine learning techniques, I developed a classification method that offers an innovative way for end-users to build new classifiers with accuracies higher than other methods available in literatures.

Relevant biological applications, as well as original research are presented along with the computational methods, namely secondary structure, solubility, chaperone requirements prediction (*clever*Suite, Chapter I), RNA-binding, aggregation and disorder propensity prediction (Chapter II) and determination of physico-chemical determinants of neurodegenerative diseases and cancer (Chapter III). Furthermore, an extension of the *cat*RAPID suite, called *cat*RAPID *signature* is introduced (Chapter IV). In the last chapter, I will discuss a collaboration focusing on finding biological significance of multiple repetitions of amino acids (Chapter V).

All of the methods presented exploit protein sequences as input data and integrate functional annotation databases, such as expression level databases or physico-chemical scales, to perform data exploration and analysis. Both the *clever*Suite and *cat*RAPID *signature* can be used freely via a web service I built[2].

[1] Numerical mappings between each amino acid and feature of interest
[2] See http://www.tartaglialab.com for links to each of the methods

Contents

INTRODUCTION

In the last decade, there has been an exponential increase high-throughput screenings of experimental data (Hawkins, Hon, and Ren 2010), which has led to a paradigm shift in the analysis – moving from manual interpretation of experimental data to utilising computational algorithms to perform data screening and interpretation. The conceptual change allowed great expansion of already available information, ranging from building a more comprehensive understanding of biological processes to being able to perform new analyses with larger experimental readings. However, the technological advance also meant that researchers who used to be able to independently interpret their findings now rely on other teams to help them sift through the data, requiring both wider expertise as well as increased number of human resources involved.

Many bioinformatics tools are being developed with the aim of automating and simplifying various aspects of data analysis (Bailey et al. 2009), both stand-alone tools (Rost 1996) and online accessible algorithms and web services (Rice, Longden, and Bleasby 2000). Indeed, the latter form of algorithms proves to be more affordable and accessible means of performing computational research (Dudley et al. 2010). Notwithstanding, there still exists a barrier of entry for non-computational scientists that can render large amount of tools unavailable due to complexity or complete lack of graphical user interface. Many of the tools on the market focus on specific features only (Linding et al. 2003; Eden et al. 2009; L. Fu et al. 2012) and no systematic approach has been attempted yet to provide end-users with a general purpose algorithm.

My original contribution is the *clever*Suite – a series of algorithms focused on providing easy to use, graphical user interface for data analysis and classification. In this thesis, I introduce the individual components of the toolkit, as well as their application in the later chapters.

1. Amino-acids and peptides

Amino acids (AAs) are biomolecules consisting of carboxylic acid and amine group bound together at the alpha-carbon of the carboxylic acid. A side chain can be bound to the same alpha-carbon and provides each of the AA's unique features (see Figure 1 for basic Amino Acid structure).

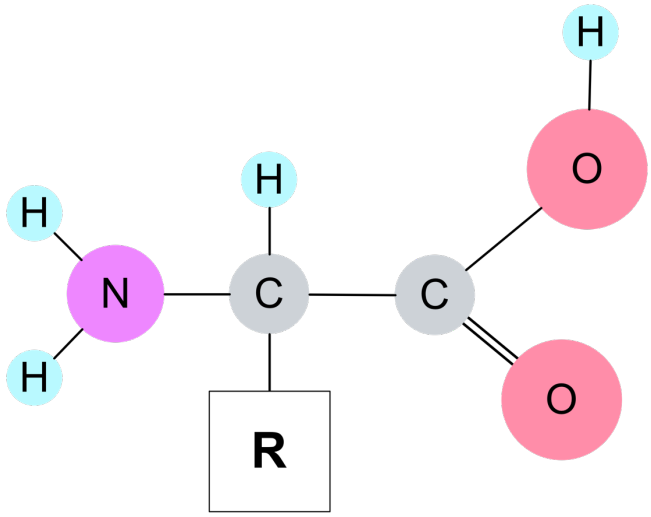

Figure 1. Amino Acid Core Diagram. *The "R" is a placeholder for the side chain, which is the main factor in determining the AAs chemical features. For example, the side chain composition affects polarity, charge, overall size and many other aspects of the AA features. See* ***Figure 2*** *for further details.*

There is wide variety in the side-chain content – ranging from its complete absence in glycine, to attachment of aromatic heterocyclic compounds (indole) found in tryphtophan (see Figure 2 for overview of chemical features of proteogenic AAs). Approximately 500 different AAs have been found in nature (Wagner and Musso 1983) but only 20 are considered proteinogenic (compose proteins) and are directly encoded by the genetic code of eukaryotes. Based on the composition of their side chains, the AAs have different physico-chemical properties and can be classified by their polarity, pH level, presence of aromatic core, presence of hydroxyl or sulphur, etc. As building blocks of proteins, AAs and their interactions have fundamental influence on the protein features, such as protein secondary structure, overall polarity, conformational stability and solubility.

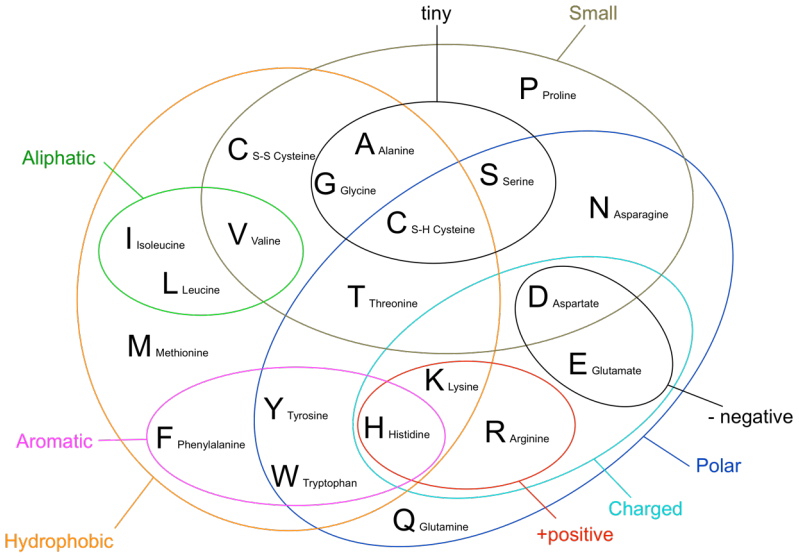

Figure 2. Proteogenic Amino Acids (AAs) categorisation based on chemical features of their side chains. *The AA features vary greatly based on chemical features of the side chain. For example, phenylalanine contains heterocyclic compound at its sidechain, making it aromatic, as well as hydrophobic AA.*

To form proteins, amino acids are linked together via covalent bonds (peptide bonds); basic sequence of amino acid residues is referred to as the primary structure of the protein. For computational purposes, each of the proteinogenic amino acids has a code assigned. This allows mapping of the primary protein structure in to a sequence of letters; akin to the genetic code describing sequence of nucleotides in the genome.

2 Protein feature prediction

2.1 Particle simulations

Any biological structure can be represented as an entity in the physical world, consisting of particles and forces between them. Therefore, in a broad sense, to predict function of biological systems without making any approximations, it would require calculation of the electronic structure for the system – numerical solution of the system's Schrödinger's equation. While there are advantages of using first-principle molecular dynamics - mainly absence of any empirically derived parameters - it has so far only been applied to complex biological problems with smaller system size (Fattebert et al. 2015). There is progress being made in our ability to perform atomistic simulations in greater detail (Kapil, VandeVondele, and Ceriotti 2015), however, it is not yet feasible to perform such simulations on a scale required to simulate complex processes such as determination of protein structure or its function or interactions.

An approach similar to molecular dynamics (MD) is the so called "classical" approach – close to the first principle MD (FPMD) in the sense that it contains model of particles but the forces and relationships between them are estimated by a function of varying complexity and accuracy depending on the computational power and resolution required.

For example, in the CHARMM toolkit (Brooks et al. 2009) the particle trajectories are calculated using classical MD for simulation of an oligomeric peptide system. The following example considers all of the heavy atoms, as well as hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 potential function). The equation 1is used to calculate the free (effective) energy *E*, which corresponds to the force used in the MD:

$$E(r) = E_{vacuum}(r) + G_{solv}(r) \qquad (1)$$

In the equation above, the molecular system has its atom's nuclei at $r = (r_1, \dots, r_N)$. The first part of the free energy equation is the free energy in vacuum:

$$E_{vacuum}(r) = \frac{1}{2}\sum_{bonds} k_b\,(b - b_0)^2 \qquad (2)$$

$$+ \frac{1}{2}\sum_{bond\ angles} k_\theta(\theta - \theta_0)^2$$

$$+ \frac{1}{2}\sum_{dihedral\ angles} k_\phi[1 + \cos(n\phi - \delta)]$$

$$+ \frac{1}{2}\sum_{improper\ dihedrals} k_\omega(\omega - \omega_0)^2$$

$$+ \sum_{i>j} \epsilon_{ij}^{min}\left[\left(\frac{d_{ij}^{min}}{r_{ij}}\right)^{12} - 2\left(\frac{d_{ij}^{min}}{r_{ij}}\right)^{6}\right]$$

$$+ \sum_{i>j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}}$$

The variables in the equation above have following meaning; b is bond length, k_θ is bond angle, k_ϕ is a dihedral angle $k_\omega$is an improper dihedral. The r_{ij}signifies distance between atoms i and j, q_i and $q_j$are partial charges and the optimal van der Waals distance and energy are d_{ij}^{min} and ϵ_{ij}^{min}. To account for effects of an aqueous solvent, the energy function can contain calculation of the solvation energy G_{solv}:

$$G_{solv}(r) = \sum_{i=1}^{N} \sigma_i A_i(r) \qquad (3)$$

where N is the number of heavy atoms with Cartesian coordinates $r = (r_1, \dots, r_{N)}$. There are only two values needed for the σ_i

parameter; one for sulphur and carbon atoms ($\sigma_{C,S} = 0.012\ kcal.mol^{-1}$ Å) and another for oxygen and nitrogen atoms ($\sigma_{N,O} = 0.06\ kcal.mol^{-1}$ Å). The $A_i(r)$ parameter is solvent-accessible surface, which approximated analytically using 1.4 Å probe radius. The model and force field outlined above have been used in literature to simulate aggregation and folding of structured peptides (Paci et al. 2004). Another, more recent example is ALMOST (B. Fu et al. 2014) – an atom molecular simulation toolkit for structure determination and assessment of structural and dynamic properties of complex molecular systems.

There are currently multiple classical MD tools available to detect protein features, ranging from standalone tools, pure MD standalone tools (Pearlman et al. 1995; Brooks et al. 2009) to hybrid methods combining MD with machine learning to both speed up the calculation and improve the result (Khoury et al. 2014). The MD approaches, either FPMD or classical MD, have great advantage of using none or minimal amount or parameters and empirical knowledge, making them theoretically most precise metrics. That said, they are also the most computationally intensive which makes them unsuitable for high-throughput data processing and screening.

2.2 Motif-based methods

Primary structure is often used to predict a number of features, such as secondary structure and hydrophobicity. Some methods focus on known functional parts of the protein sequence, such as domains, annotated functional sites or other sequence patterns (in general referred to as motifs) to identify features of interest. One example of such database, called PROSITE (Sigrist et al. 2002), contains curated set of motifs along with rich descriptions of the origin of the motif, experimental characterization supporting its existence, as well as known proteins associated with the database entry. To use existing models to describe new sequences, the database is searched for known patterns (de Castro et al. 2006) and matches against the provided sequence are reported. The use of database-backed methods allows identification of known motifs in the query sequence, which is linked to relevant data and literature. However,

The *clever*Suite (Chapter I) is an example of an integrated approach that combines physico-chemical feature generation and model building in a single package, combining features of, for example, PROFEAT with custom subsequent analysis.

4 Signal detection in protein datasets

As described in Chapter I, the physico-chemical scales can be used to detect differentially-enriched features of proteins sets. Introducing binary comparisons of datasets is an important concept, as it enables discrimination of a background signal from the true signal of interest. For example, simple single-set analysis may reveal that majority of an experimental sample contains alpha-helical proteins. However, when we cross-reference this result with another sample coming from a similar population (but not exhibiting the phenotype of interest), we may find that alpha helix is not an important feature as it is found in both sets. To quantify the differential enrichment between sets, we introduce the concept of coverage (Klus et al. 2014, Chapter I):

$$coverage(P,N) = \frac{1}{P_{tot}} \sum_p \vartheta \left(\frac{1}{N_{tot}} \sum_n \vartheta(\pi_p - \pi_n) - \alpha \right) \quad (8)$$

In the equation above, π is the signal extracted from the protein profile, the counter function $\vartheta(x-y)$ is 1 if $x > y$ and 0 otherwise. The P_{tot} and N_{tot} are total numbers of sequences in the datasets P and N. The additional parameter α is used as a counting cut-off (see Chapter I. for further details). Using the coverage parameter provides accurate measure to establish differential strength of each of the predictors in use. Furthermore, as the equation takes into account the calculated property strength and not an identifier or sequence, it discriminates based on the real feature enrichment. Allowing submission of both negative and positive sets is an important concept, as it completely removes any dependence on built-in data. While there exist efforts to build databases of negative/non interacting biomolecules (Blohm et al. 2014), it will never provide the per-case flexibility that is frequently needed. The *clever*Suite approach covered in Chapter I has an advantage over other methods due to it being scale-neutral. It finds the best-covering scales and their combinations (only combinatorial/computational complexity is the limit), either from built-in set or even from user-provided scales. It also builds *ad hoc* scale that could be used for discrimination.

There are similar approaches employed to interpret gene expression data - Gene Set Enrichment Analysis (Subramanian et al. 2005), or to assess over representation-based enrichment in protein or gene sets (Glaab et al. 2012).

5 Physico-chemical determinants of neurological diseases

5.1 Solubility and saturation

Maintaining solubility in the cell is an important part of its homeostasis as aggregation has been linked to a wide variety of diseases. One of the main outstanding questions in the field of protein aggregation is a clear determination whether protein is aggregation-prone based on its sequence alone without knowing its usual expression levels or further detail. This observation is further supported by the fact that only some of the proteins seem to be aggregating *in vivo* – suggesting a link between their concentration and solubility (Ciryam et al. 2013). Similar trend can be found when looking at proteins in *S. cerevisiae* with high and low solubility (Albu et al. 2015), where the less soluble proteins were found less abundant under normal conditions (see Chapter II). Intriguingly, proteins linked to genes up-regulated in Central Nervous System (CNS) disorders have also been found less abundant under physiological conditions (see Chapter III).

5.2 Role of protein expression

A strong anti-correlation has been observed between *in-vitro* aggregation rates of human proteins and *in-vivo* expression rates of human genes, as is illustrated in Figure 4 (Tartaglia et al. 2007). The explanation for this trend is that an evolutionary pressure acts to decrease the risk of aggregation by reducing expression levels of aggregation-prone proteins (Bolognesi and Tartaglia 2013). The same trend has also been observed when investigating gene products involved in neurodegenerative diseases, which were consistently less abundant under physiological conditions (see Chapter III for more details).

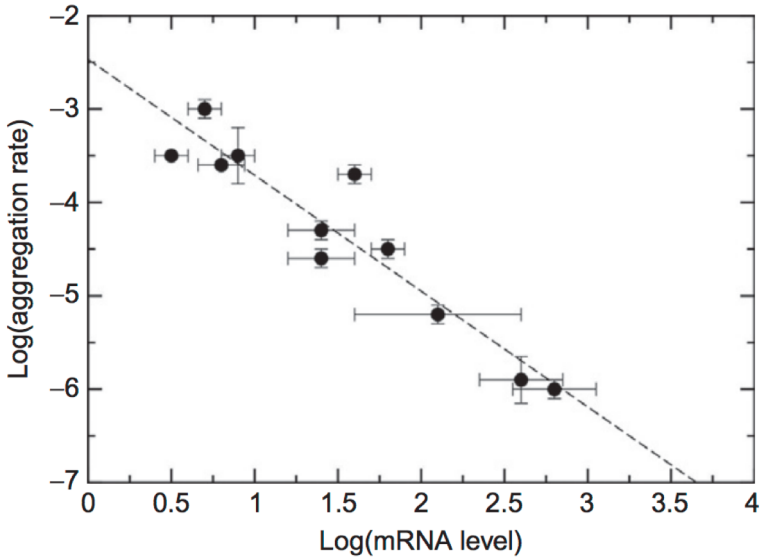

Figure 4. Aggregation rates correlate negatively with expression levels. (Tartaglia et al. 2007, Bolognesi and Tartaglia 2013).

CHAPTER I

The cleverSuite approach for protein characterization

The main focus of my PhD work is the *clever*Suite (CS). The original idea of the project was to provide a tool that would describe protein datasets and allow creation of ad-hoc classifiers, enabling non-computational users to create their own predictors without deep knowledge of machine learning algorithms. This publication represents the core of my work, a building block that is later user throughout my other publications. The first part of the CS is the *clever*Machine (CM), a tool that performs statistical analysis on protein sequences by comparing their physico-chemical propensities. The CM contains a curated set of 80 predictors selected both from existing databases (Kawashima et al. 2008) and by performing statistical analysis on computational tools. The second part is the *clever*Classifier, which applies model created with CM to classify other datasets. The CS has been validated with proteomic data from existing experiments – secondary structure, solubility, chaperone requirements and RNA-binding abilities were benchmarked and the CS achieved similar results to existing tools or outperformed them. Descriptions of the datasets generated by the CM were in great agreement with experimental findings.

Klus P, Bolognesi B, Agostini F, Marchese D, Zanzoni A, Tartaglia GG. The cleverSuite approach for protein characterization: predictions of structural properties, solubility, chaperone requirements and RNA-binding abilities. Bioinformatics. 2014 Jun 1;30(11):1601–8. DOI: 10.1093/bioinformatics/btu074

CHAPTER II

Protein aggregation, structural disorder and RNA-binding ability

This chapter covers the extensions of the *clever*Suite work described in the Chapter I, as well as other relevant example datasets. I introduced two new algorithms, the multiCleverMachine (mCM) and cleverGO (cGO). The mCM builds on top of the cleverMachine (CM) and extends it functionality by allowing easy submission and comparison of multiple datasets, which was the most common use case of the CM. The tool simplifies upload of multiple files and provides additional, high-level visualisation of properties found in the submitted sets. The result can be used on its own and the users can also "drill-down" into the individual comparisons to see full CM models or to launch further investigation using the second algorithm – cGO and Boxplotter, which are also described in this chapter: cGO is a tool that visualises semantic similarity between enriched GO terms. Boxplotter visualises protein trends performing statistical analysis. As sample use cases, this chapter investigates RNA-binding abilities of *S. cerevisiae* chaperone substrates and provides links between aggregation and structural disorder in *S. cerevisiae*, *C. elegans*, *M. musculus* and *H. sapiens*. The results of our investigation are in strong agreement with experimental evidence.
.

Klus P, Ponti RD, Livi CM, Tartaglia GG. Protein aggregation, structural disorder and RNA-binding ability: a new approach for physico-chemical and gene ontology classification of multiple datasets. BMC Genomics. 2015 Dec 16;16(1):1071. DOI: 10.1186/s12864-015-2280-z

CHAPTER III

Neurodegeneration and cancer: Where Disorder prevails

It has been reported in medical literature that genes up-regulated in cancer are often down-regulated in Central Nervous System CNS diseases and vice versa, which suggests a strong link between these two pathologies on a molecular level. In this chapter, we employ the tools introduced in Chapter I and II to investigate physico-chemical features distinguishing proteins associated with cancer and CNS diseases. The cancer group includes prostate, colorectal and lung cancers and in the CNS group included Parkinson's and Alzheimer's diseases and Schizophrenia. Analysis of protein abundances using the boxplotter[3] algorithm reveals that CNS disease genes code for proteins that are less abundant under physiological conditions, suggesting that the disease state has strong effect on gene expression. The opposite trend is apparent for cancer-related genes, whose products show increased expression under physiological conditions. Using the multicleverMachine, we found that structural disorder is a key feature to differentiate cancer types and CNS diseases. More specifically, disorder is significantly enriched in the up-regulated protein groups for each of the CNS diseases and anti-correlates with order-promoting features (alpha-helix and beta-sheet), as well as burial propensity. Also, in agreement with existing studies (Liu et al. 2006), we observed enrichment in nucleic-acid binding propensity of proteins up-regulated in lung and colorectal cancer. The second part of the analysis was performed using *clever*GO to investigate Gene Ontology features of the groups under investigation.

Klus P, Cirillo D, Botta Orfila T, Tartaglia GG.
Neurodegeneration and Cancer: Where the Disorder Prevails.
Sci Rep. 2015 Dec 23;5(1):15390. DOI: 10.1038/srep15390

[3] See Chapter II for further details or access http://www.tartaglialab.com/boxplotter/submit directly

CHAPTER IV

catRAPID signature

Current RNA-binding tools often exploit sequence similarity between query sequences and known RNA-binding domains (RD). The approach relies on annotation of RDs only and cannot be employed to identify proteins that bind to RNA although lacking known RD. The *cat*RAPID *signature* addresses this issue by considering the physico-chemical features of known RDs for training, and computing and comparing the same features for new query sequences. Such approach abstracts away from actual sequence and allows discovery of binding domains without knowledge of their exact sequence. My main contribution to *cat*RAPID signature was the generation of a library of physico-chemical predictors and associated features (profile generation, smoothing, signal extraction, etc.).

Livi CM, Klus P, Delli Ponti R, Tartaglia GG. cat RAPID signature : identification of ribonucleoproteins and RNA-binding regions. Bioinformatics. 2016 Mar 1;32(5):773–5. DOI: 10.1093/bioinformatics/btv629

CHAPTER V

Non-random distribution of homo-repeats: links with biological functions and human diseases

This chapter covers results of collaboration between our lab and Laboratory of Protein Physics at Institute of Protein Research (RAS), Russia.

The number of homo-repeats in eukaryotic and bacterial proteomes is significantly larger than expected from theoretical estimates. Our calculations indicate that the minimal length that is statistically significant varies with amino acid type and proteome. In *H. sapiens*, occurrence of homo-repeats is associated with high content of structurally disordered regions and enhanced RNA-binding potential, which is in line with recent experimental findings. We also observed that proteins containing homo-repeats have a large number of interactions, which can promote perturbation of protein networks and cause dysfunction. Although the functional roles of homo-repeats are unknown, we found that their occurrence is associated with pathology.

Lobanov MY, Klus P, Sokolovsky I V, Tartaglia GG, Galzitskaya O V. Non-random distribution of homo-repeats: links with biological functions and human diseases. Sci Rep. 2016 Jul 3;6(1):26941. DOI: 10.1038/srep26941

DISCUSSION

In the previous chapters, I have presented a series of biological problems that I have addressed through algorithms and other theoretical principles. The Chapters I to III, my first-author publications, were focused on the individual algorithms and their scientific significance. Although some applications are presented in Chapters IV and V, I will describe the future plans in the following text, along with a discussion on the technical implementations needed to make the work possible. Therefore, this final section is divided into two parts – "What does the future hold?" which covers work in progress and future plans of the *clever*Suite. The second part, aptly titled "Technical challenges", describes the work needed to implement the algorithms.

1. What does the future hold?

1.1. *clever*Suite improvements

As we can see in Table 1 below, the *clever*Suite in its current version has performances on par with purpose-built tools and predictors. However, *clever*Suite and its capabilities described in Chapter I are just a starting point in the lifecycle of the project. New features, both on technical and scientific side are currently in progress and will be released when they reach maturity.

	cleverSuite			Reference		
	ACC[1] (%)	TPR[2] (%)	TNR[2] (%)	Method	TPR[3] (%)	TNR[3] (%)
Alpha-beta	97.9	90.4	93.2	*RePROF*	92.6	72.0
Disorder	86.1	84.5	73.6	*FoldIndex*	62.9	64.7
Solubility	89.8	84.7	60.5	*PROSO II*	78.5	74.0
Chaperones	81.6	75.4	60.0	*Limbo*	100.0	22.5
mRNA	84.3	72.9	79.2	*RNApred*	82.5	52.8

Table 1 cleverSuite performances *10-fold cross-validation accuracy for cleverMachine (CM) models (ACC is accuracy). 2. Independent validation performances for cleverClassifier (CC). 3. Performance comparison with algorithms reported in literature. TPR (true positive rate) and TNR (true negative rate) are calculated on the same sets used to validate CC (2).*

a) *clever*Machine's signal extraction

The original *clever*Machine uses averages of physico-chemical properties to determine overall strength of the signal, which means a single value for each input sequence. My approach provides good discrimination with both experimentally and computationally derived scales. Nevertheless, there are areas in which this approach is lacking:

- Protein contains specific sites that are neglected by the averaging approach;
- The enrichment is based on full-length features.

As presented in Chapter I, we are evaluating methods to extract specific regions from the profiles. One of the proposed solutions is to consider areas that are one standard deviation away from the local (protein) or global (dataset) mean. Another approach would exploit fragmentation of the profiles into smaller parts to consider them as individual entities for the purposes of the enrichment calculation. There are other improvements that are currently under way, for example detection of feature changes for single-point mutations and integration of protein networks to consider interaction sites. Lastly, we are exploring new approaches to directly integrate other algorithms. The *clever*Suite currently contains computationally derived scales that were produced by linking AA frequencies to features detected by the tools but does not integrate the tools directly. This leads to inevitable loss of information that could be avoided by integrating the third-party tools directly.

b) Classification feature enhancements

The *clever*Classifier currently supports binary classification of incoming set, with an addition of signal strength. Calculations are done through multiple samplings of the input data and by creating of different models where the score is effectively an agreement between them. Apart from using the models to provide binary classifications of the input set, there is a potential in extending the confidence percentage in to a score metric that could better evaluate the input data.

c) Database(s) integration

One of the primary goals of *clever*Suite (CS) was to be organism "agnostic". To achieve this requirement, it was designed without any links to existing libraries or protein sets. While this requirement stays the same for the CS core, there is a benefit in enriching the calculation by using information from existing databases of knowledge – for example, it may be helpful to only consider

specific sets (Bateman et al. 2000) for information extraction, or, alternatively, use them as another independent information source.

d) Pair-wise prediction

Another feature that is in the works is to expand on which information goes in to the profiles. For example, we want to be able to create protein-protein or protein-RNA interaction profile for further data mining, similar to approach taken in PAnDA (Cirillo, Botta-Orfila, and Tartaglia 2015).

1.2. Easier sharing of ideas and code

As mentioned earlier in this chapter, parts of my effort during the PhD pertained to making reuse of ideas and algorithms easier. We have seen creation of shared code repositories, which lead to fruitful collaboration within the lab. However, due to the nature of the work and licenses bound to part of the algorithms, it was not yet possible to make the repositories completely public. Furthermore, some of the algorithms have very specific dependencies that can alter the result if not set up exactly as they should be.

During my PhD, a software solution became industry-ready that could alleviate the second point – managing dependencies and requirements of the algorithms. The solution is containerisation, with Docker (Merkel 2014) being the most notable example. In a nutshell, it allows packaging of algorithms including their dependencies in a pre-configured container. Using containerisation to encapsulate the algorithm and all of its dependencies is only the first part of the equation in simplifying the distribution. Akin to github (Dabbish et al. 2012), the Docker project also maintains an open repository of existing containers, making the task of code re-use a matter of issuing a few simple commands. During the last year of my PhD, I have started a shift of all of our software to Docker, with the plan of including binary versions of algorithms where source code could not be disclosed (see more on this below). The containerization is the ideal solution in terms of space-efficiency and portability, however, there is a learning curve associated with it for the algorithm developers and users alike. Furthermore, the Docker project is extremely fast growing, which means that the

documentation does not always keep up and it does require specific version of the operating system to run in an optimal manner.

Therefore, for users who simply wants to run the algorithm on their machine, I have collaborated with other members of our lab to build set of virtual machines containing the algorithms and all of their dependencies, as well as running webserver for easy result review. Using a set of open standards, the virtual machine can be used by variety of desktop virtualization solutions (VMWare Player/Workstation/Fusion, Parallels Desktop, Oracle Virtual Box and any other supporting the Open Virtualization Format).

Distributing the algorithm inside of a container or a virtual machine brings the computation to the user's control and they can easily integrate our tools into their pipelines and workflows. While this is a great step forward in dissemination and sharing of our work, there is still much more that could be done to foster collaboration on even deeper level. The future plan for all of the algorithms mentioned in my thesis is to make them fully open-source. This does not mean simply flipping the switch and making the repository public, a process needs to be undertaken to refactor some of the source code to make it more accessible, as well as to provide documentation on how to setup development environment and create robust compilation and deployment scripts.

Last days of my time at the CRG were devoted to making the above process a reality. I handed over the work on "dockerization" of algorithms, as well as put plans down for making the *clever*Suite repositories public and accessible to the scientific community. I have also been involved in patent disclosure process[4] for the public release of the *cat*RAPID suite of algorithms, which is now approved the software is now being distributed.

[4] Software Disclosure Form (SDF) with identifier "ID 0056_CRG_SDF"

2. Technical challenges

Apart of the scientific challenges presented in the previous chapters, a several technical advancements had to be made to make algorithms work together efficiently and to distribute the calculation across resources.

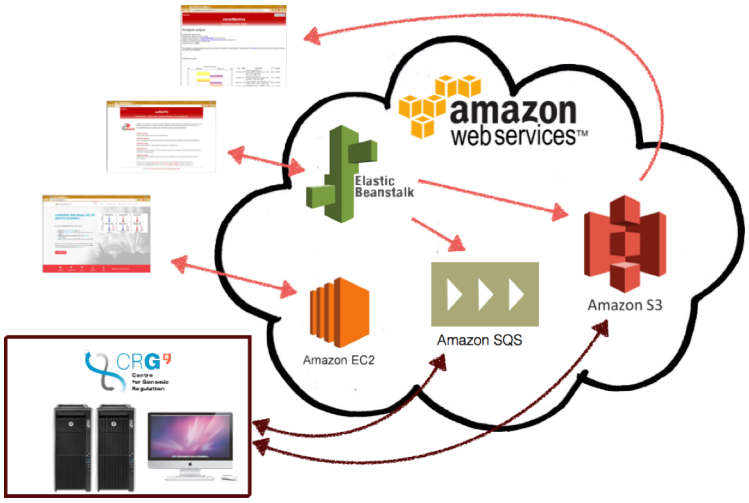

Figure 3. Computational infrastructure overview. *To deliver results to users of our algorithms, we employ hybrid infrastructure consisting of local servers for the heavy computation (under CRG logo) and cloud infrastructure to store the results and manage the computational pipeline.*

2.1 Computational infrastructure overview

The *clever*Suite (CS) algorithms introduced in Chapter I inspired the first step of computational infrastructure development I have performed as part of my PhD. First of all, the amount of calculation required meant that CS jobs had to be distributed across multiple server resources and the execution properly coordinated. Also, we had to enable other lab members to easily re-use the system and introduce new algorithms. As there was no easy off-the-shelf solution we could use, I have developed a simple system for job submission, distribution, result storage and other "glue" functionality needed to complete the submission lifecycle. The

challenge was to come up with a complete system that would fit all the requirements in very modest amount of time there was. The solution comprises of following parts:

a) Algorithm definition

While not necessarily a software component, the algorithm definition is an important part of the system that describes each of the algorithms, their inputs and outputs, and allows exchange of information between all of the components.

Algorithm definition is a text file that outlines inputs of the algorithm and technical details needed for its execution. We have considered multiple formats as one of the requirements was minimizing the user's learning curve when trying to setup new tools online. We have chosen YAML file format as it is both a human and machine readable format and does not require any special tools or knowledge for its creation or maintenance. The algorithm definition has following parts:

- Algorithm name
- Algorithm description
    - Shown to end-users on the website on submission pages.
- Field definition
    - To allow automatic form-generation on the server.
- Command structure
    - Command line arguments and path to the algorithm including relevant placeholders for input data and output directories.
- Supporting data
    - Additional information for users to show before/after the automatically generated form, links to documentation and tutorial, etc.

b) Submission interface

The job submission, status checking and results database can be accessed using the website component, which creates an user

interface based on the algorithm definitions. The interface caters for new job creation, status checking and related administrative tasks and can be ran locally, as well as deployed to remote servers. This makes it easier for other lab members to prototype new algorithms and makes them available faster. To cater for public usage of our algorithms, a website has been created at http://s.tartaglialab.com using the website component.

c) Job queue and execution

The submission system had to be able to cater for multiple independent entities submitting and processing work. This is addressed by using a per-algorithm global queue (Amazon SQS[5]), which ensures that no two entities claim job for processing, jobs get processed in first come first served manner and that creation of new jobs is possible no matter what other activity there is. Furthermore, the system also caters for computational nodes failing and not completing a job. This is achieved by adopting a "renewal" mechanism where computational nodes need to periodically inform the queue manager that they are still busy working on a job. Should this notification fail to arrive in a timely manner, job is returned to the queue and is made available for processing to other machines.

d) Result storage and submission database

Our services use cloud-hosted database (Amazon SimpleDB[6]) and cloud-based storage (Amazon Simple Storage Service[7]) to synchronise data across its instances, providing both ease of use and data durability.

e) Real world usage

The information in this section provide brief overview of the computational infrastructure I have built during my PhD to support the work presented in this thesis, as well as creation and evolution of multiple algorithms at our lab. After the initial release, the system was quickly adopted for both new and legacy algorithms. In

[5] https://aws.amazon.com/sqs/
[6] https://aws.amazon.com/simpledb/
[7] https://aws.amazon.com/s3/

a span of mere 2 years, following new algorithms were developed on top of the new system:

- catRAPID omics (Agostini et al. 2013)
- ccSOL omics (Agostini, Cirillo, Livi, et al. 2014)
- catRAPID signature (Livi et al. 2015)
- cleverSuite (Klus et al. 2014)
- multiCleverMachine (Klus et al. 2015)
- SeAMotE (Agostini, Cirillo, Ponti, et al. 2014)
- PAnDA (Cirillo, Botta-Orfila, and Tartaglia 2015)

2.2 Interoperability

a) Code re-use

The job distribution and result storage system described in this chapter provides a way to create jobs and reuse their results, however, it still treats each of the algorithms as independent "black boxes" without any restrictions on their internal structure. This makes it very easy to create new algorithms and make them available to potential users but it does not necessarily make it easier for developers to re-use parts of algorithms and create derivative work.

To address this, I have pioneered use of distributed source code version control (using git) in our centre to open up all of our work to collaboration and sharing of ideas on the raw code level. This started transition from individual, independent tools to collaborative approach where everybody can easily contribute and improve each of the group's algorithms, as well as cherry-pick and re-use interesting features. Results of this level of collaboration can be seen in Chapter IV – algorithm *cat*RAPID signature. To create *cat*RAPID signature, parts of the cleverSuite (Chapter I) have been re-used and adapted to suit a new purpose.

b) Result re-use and custom "server" creation

The unified result storage opened up an option to create customised version of each of the algorithms based on the user submitted data.

For example, one can obtain a re-usable classification algorithm after training the cleverMachine (Chapter I). This can be easily shared between users. Example of this approach can be seen in Figure 4.

CM	CC	Description	
Alpha-helix vs. beta-sheet (Bernstein)			by Petr Klus
link	-	Alpha-helix vs. beta-sheet model/training	run classifier
-	link	Alpha-helix rich proteins vs. alphabeta model	
-	link	Beta sheet rich proteins vs. alphabeta model	
mRNA-binding interactome in H. sapiens (Castello)			by Petr Klus
link	-	mRNA-binding interactome in H. sapiens (Castello) training/model	run classifier
-	link	mRNA binding proteins (Baltz) vs. mRNA binding proteins (Castello)	
-	link	proteins not binding mRNA (Shazman) vs. mRNA binding proteins (Castello)	
Structurally disordered proteins			by Petr Klus
link	-	Structurally disordered proteins training/model (Sickmeier)	run classifier
-	link	Prions vs. disorder model (Alberti)	
-	link	Structured proteins vs. disorder model (Tartaglia)	
E. coli solubility			by Petr Klus
link	-	E. coli solubility model/training (Niwa)	run classifier
-	link	Independently folding proteins (Tartaglia) vs. solubility model	
-	link	Chaperone-dependent proteins vs. solubility model	
Chaperone-dependent proteins			by Petr Klus
link	-	Chaperone-dependent proteins model/training (Kerner)	run classifier
-	link	DnaK/GroEL dependent proteins vs. chaperone model	
-	link	Independently folding proteins (Tartaglia) vs. chaperone model	
E. Coli and H. Sapiens full proteome analysis			by Petr Klus
link	-	E. Coli and H. Sapiens full proteome analysis	

Figure 4 Featured models and derived classifiers *Ad-hoc page featuring classifiers built with cleverMachine that are available to the community (full page available at http://service.tartaglialab.com/static_files/algorithms/clever_machine/featured_submissions.html)*

Conclusions

This thesis, titled "The Clever Machine – a computational tool for dataset exploration and prediction" provides description of all of my activities during my PhD at the Centre for Genomic Regulation (CRG) in Barcelona, Spain. It describes both scientific and technical work performed to achieve publications in chapters I to V, as well as work outside of the publications. The chapters are organised in a chronological manner, with the exception of Chapter V, which spanned through almost entire duration of my PhD.

This thesis, titled "The Clever Machine – a computational tool for dataset exploration and prediction" provides description of all of my activities during my PhD at the Centre for Genomic Regulation (CRG) in Barcelona, Spain. It describes both scientific and technical work performed to achieve publications in chapters I to V, as well as work outside of the publications. The chapters are organised in a chronological manner, with the exception of Chapter V, which spanned through almost entire duration of my PhD.

Chapters:

I. The *clever*Suite (CS), a toolkit designed to simplify data extraction and classification of protein sequences. At its core, it uses physico-chemical scales to extract features from protein sequences. The CS contains curated list of 80 physico-chemical scales, which predict protein secondary structures; aggregation, disorder, membrane protein and nucleic acid binding propensities, as well as solubility. Information is extracted for all of the included scales (and any custom scales the user provides) and enrichment is calculated and shown to the user on an interactive webpage. Furthermore, the submitted data is used to train a classification model that can be then used for predictions. To sum up, the CS serves two purposes – it's a tool for feature detection and visualisation, as well as an ad-hoc classifier creation service.

II. Extension of the *clever*Suite algorithm to allow for more efficient exploration of larger number of datasets manifested as a new algorithm – *multi*CleverMachine (mCM). The mCM greatly simplifies evaluation of multiple datasets by allowing batch-execution of cleverMachine jobs and

providing unified user interface to interpret the results. It also serves as a central "launchpad" for running other integrated tools on the already submitted data. Second part of the chapter describes second tool, the *clever*GO (cGO). The cGO algorithm is a new, integrated approach to visualisation of Gene Ontology enrichment information by providing semantic similarity visualisation of the enriched terms.

III. Application of the tools described in Chapter I and II to investigate physico-chemical features that distinguish CNS diseases and cancer. Using the boxplotter algorithm, we have found that CNS disease genes code for proteins that are less abundant under physiological conditions, suggesting that the disease has an effect on gene expression. The opposite is true for cancer-related gene products, which show increased expression under physiological conditions. Furthermore, we have found the structural disorder to be the most distinguishing feature (enriched in up-regulated for each of the CNS diseases) between the two disease groups.

IV. Result of a re-use of the *clever*Suite core to create predictor of RNA-binding domains (RD) which is not based on known sequences but uses features of the sequences to predict its propensity to be RNA-binding. This approach allows discovery of novel RDs as it's not based on sequence similarity. My contribution was the creation of the re-useable library of physico-chemical predictors and their associated function library.

V. Investigation of biological function and prevalence of homo-repeats in 122 bacterial and eukaryotic genomes. Discrepancy between expected number of homo-repeats and total number is investigated, as well as interaction profiles of proteins containing homo-repeats. Lastly, the chapter covers physico-chemical features of highly-interacting proteins and their relation to human diseases.

Bibliography

Agostini, Federico, Davide Cirillo, Carmen Maria Livi, Riccardo Delli Ponti, and Gian Gaetano Tartaglia. 2014. ‘ccSOL Omics: A Webserver for Large-Scale Prediction of Endogenous and Heterologous Solubility in E. Coli’. *Bioinformatics*, July, btu420. doi:10.1093/bioinformatics/btu420.

Agostini, Federico, Davide Cirillo, Riccardo Delli Ponti, and Gian Gaetano Tartaglia. 2014. ‘SeAMotE: A Method for High-Throughput Motif Discovery in Nucleic Acid Sequences’. *BMC Genomics* 15 (1): 925. doi:10.1186/1471-2164-15-925.

Agostini, Federico, Andreas Zanzoni, Petr Klus, Domenica Marchese, Davide Cirillo, and Gian Gaetano Tartaglia. 2013. ‘catRAPID Omics: A Web Server for Large-Scale Prediction of Protein-RNA Interactions’. *Bioinformatics (Oxford, England)*, September. doi:10.1093/bioinformatics/btt495.

Albu, Razvan F., Gerard T. Chan, Mang Zhu, Eric T. C. Wong, Farnaz Taghizadeh, Xiaoke Hu, Arya E. Mehran, James D. Johnson, Jörg Gsponer, and Thibault Mayor. 2015. ‘A Feature Analysis of Lower Solubility Proteins in Three Eukaryotic Systems’. *Journal of Proteomics* 118 (April): 21–38. doi:10.1016/j.jprot.2014.10.011.

Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. ‘MEME Suite: Tools for Motif Discovery and Searching’. *Nucleic Acids Research* 37 (suppl 2): W202–8. doi:10.1093/nar/gkp335.

Bateman, Alex, Ewan Birney, Richard Durbin, Sean R. Eddy, Kevin L. Howe, and Erik L. L. Sonnhammer. 2000. ‘The Pfam Protein Families Database’. *Nucleic Acids Research* 28 (1): 263–66. doi:10.1093/nar/28.1.263.

Blohm, Philipp, Goar Frishman, Pawel Smialowski, Florian Goebels, Benedikt Wachinger, Andreas Ruepp, and Dmitrij Frishman. 2014. ‘Negatome 2.0: A Database of Non-Interacting Proteins Derived by Literature Mining, Manual Annotation and Protein Structure Analysis’. *Nucleic Acids Research* 42 (Database issue): D396–400. doi:10.1093/nar/gkt1079.

Bolognesi, Benedetta, and Gian Gaetano Tartaglia. 2013. 'Physicochemical Principles of Protein Aggregation'. *Progress in Molecular Biology and Translational Science* 117: 53–72. doi:10.1016/B978-0-12-386931-9.00003-9.

Brooks, B. R., C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, et al. 2009. 'CHARMM: The Biomolecular Simulation Program'. *Journal of Computational Chemistry* 30 (10): 1545–1614. doi:10.1002/jcc.21287.

Castello, Alfredo, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M Beckmann, Claudia Strein, Norman E Davey, et al. 2012. 'Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins'. *Cell* 149 (6): 1393–1406. doi:10.1016/j.cell.2012.04.031.

Chou, P Y, and G D Fasman. 1978. 'Prediction of the Secondary Structure of Proteins from Their Amino Acid Sequence'. *Advances in Enzymology and Related Areas of Molecular Biology* 47: 45–148.

Cirillo, Davide, Teresa Botta-Orfila, and Gian Gaetano Tartaglia. 2015. 'By the Company They Keep: Interaction Networks Define the Binding Ability of Transcription Factors'. *Nucleic Acids Research*, June, gkv607. doi:10.1093/nar/gkv607.

Ciryam, Prajwal, Gian Gaetano Tartaglia, Richard I. Morimoto, Christopher M. Dobson, and Michele Vendruscolo. 2013. 'Widespread Aggregation and Neurodegenerative Diseases Are Associated with Supersaturated Proteins'. *Cell Reports* 5 (3): 781–90. doi:10.1016/j.celrep.2013.09.043.

Dabbish, Laura, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. 'Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository'. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 1277–86. CSCW '12. New York, NY, USA: ACM. doi:10.1145/2145204.2145396.

de Castro, Edouard, Christian J. A. Sigrist, Alexandre Gattiker, Virginie Bulliard, Petra S. Langendijk-Genevaux, Elisabeth Gasteiger, Amos Bairoch, and Nicolas Hulo. 2006. 'ScanProsite: Detection of PROSITE Signature Matches and ProRule-Associated Functional and Structural Residues in Proteins'. *Nucleic Acids Research* 34 (Web Server issue): W362–65. doi:10.1093/nar/gkl124.

Deléage, G., and B. Roux. 1987. ‘An Algorithm for Protein Secondary Structure Prediction Based on Class Prediction’. *Protein Engineering* 1 (4): 289–94. doi:10.1093/protein/1.4.289.

Drozdetskiy, Alexey, Christian Cole, James Procter, and Geoffrey J. Barton. 2015. ‘JPred4: A Protein Secondary Structure Prediction Server’. *Nucleic Acids Research* 43 (W1): W389–94. doi:10.1093/nar/gkv332.

Dudley, Joel T., Yannick Pouliot, Rong Chen, Alexander A. Morgan, and Atul J. Butte. 2010. ‘Translational Bioinformatics in the Cloud: An Affordable Alternative’. *Genome Medicine* 2 (8): 51. doi:10.1186/gm172.

Eden, Eran, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. 2009. ‘GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists’. *BMC Bioinformatics* 10 (1): 48. doi:10.1186/1471-2105-10-48.

Eisenberg, D, E Schwarz, M Komaromy, and R Wall. 1984. ‘Analysis of Membrane and Surface Protein Sequences with the Hydrophobic Moment Plot’. *Journal of Molecular Biology* 179 (1): 125–42.

Fattebert, Jean-Luc, Edmond Y. Lau, Brian J. Bennion, Patrick Huang, and Felice C. Lightstone. 2015. ‘Large-Scale First-Principles Molecular Dynamics Simulations with Electrostatic Embedding: Application to Acetylcholinesterase Catalysis’. *Journal of Chemical Theory and Computation* 11 (12): 5688–95. doi:10.1021/acs.jctc.5b00606.

Fu, Biao, Aleksandr B. Sahakyan, Carlo Camilloni, Gian Gaetano Tartaglia, Emanuele Paci, Amedeo Caflisch, Michele Vendruscolo, and Andrea Cavalli. 2014. ‘ALMOST: An All Atom Molecular Simulation Toolkit for Protein Structure Determination’. *Journal of Computational Chemistry* 35 (14): 1101–5. doi:10.1002/jcc.23588.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. ‘CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data’. *Bioinformatics* 28 (23): 3150–52. doi:10.1093/bioinformatics/bts565.

Glaab, Enrico, Anaïs Baudot, Natalio Krasnogor, Reinhard Schneider, and Alfonso Valencia. 2012. ‘EnrichNet: Network-Based Gene Set Enrichment Analysis’.

Bioinformatics 28 (18): i451–57. doi:10.1093/bioinformatics/bts389.

Hawkins, R. David, Gary C. Hon, and Bing Ren. 2010. 'Next-Generation Genomics: An Integrative Approach'. *Nature Reviews. Genetics* 11 (7): 476–86. doi:10.1038/nrg2795.

Jacob A. Stekol. 1964. *Amino Acids and Serum Proteins*. Vol. 44. Advances in Chemistry 44. AMERICAN CHEMICAL SOCIETY. http://pubs.acs.org/doi/book/10.1021/ba-1964-0044.

Jones, Daniel D. 1975. 'Amino Acid Properties and Side-Chain Orientation in Proteins: A Cross Correlation Approach'. *Journal of Theoretical Biology* 50 (1): 167–83. doi:10.1016/0022-5193(75)90031-4.

Kabsch, Wolfgang, and Christian Sander. 1983. 'How Good Are Predictions of Protein Secondary Structure?' *FEBS Letters* 155 (2): 179–82. doi:10.1016/0014-5793(82)80597-8.

Kapil, Venkat, Joost VandeVondele, and Michele Ceriotti. 2015. 'Accurate Molecular Dynamics and Nuclear Quantum Effects at Low Cost by Multiple Steps in Real and Imaginary Time: Using Density Functional Theory to Accelerate Wavefunction Methods'. *arXiv:1512.00176 [cond-Mat, Physics:physics]*, December. http://arxiv.org/abs/1512.00176.

Kawashima, Shuichi, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. 2008. 'AAindex: Amino Acid Index Database, Progress Report 2008'. *Nucleic Acids Research* 36 (Database issue): D202–5. doi:10.1093/nar/gkm998.

Khoury, George A., Phanourios Tamamis, Neesha Pinnaduwage, James Smadbeck, Chris A. Kieslich, and Christodoulos A. Floudas. 2014. 'Princeton_TIGRESS: Protein Geometry Refinement Using Simulations and Support Vector Machines'. *Proteins: Structure, Function, and Bioinformatics* 82 (5): 794–814. doi:10.1002/prot.24459.

Klus, Petr, Benedetta Bolognesi, Federico Agostini, Domenica Marchese, Andreas Zanzoni, and Gian Gaetano Tartaglia. 2014. 'The cleverSuite Approach for Protein Characterization: Predictions of Structural Properties, Solubility, Chaperone Requirements and RNA-Binding Abilities'. *Bioinformatics (Oxford, England)* 30 (11): 1601–8. doi:10.1093/bioinformatics/btu074.

Klus, Petr, Riccardo Delli Ponti, Carmen Maria Livi, and Gian Gaetano Tartaglia. 2015. 'Protein Aggregation, Structural Disorder and RNA-Binding Ability: A New Approach for Physico-Chemical and Gene Ontology Classification of Multiple Datasets'. *BMC Genomics* 16 (1): 1071. doi:10.1186/s12864-015-2280-z.

Linding, Rune, Lars Juhl Jensen, Francesca Diella, Peer Bork, Toby J Gibson, and Robert B Russell. 2003. 'Protein Disorder Prediction: Implications for Structural Proteomics'. *Structure (London, England: 1993)* 11 (11): 1453–59.

Liu, Jiangang, Narayanan B. Perumal, Christopher J. Oldfield, Eric W. Su, Vladimir N. Uversky, and A. Keith Dunker. 2006. 'Intrinsic Disorder in Transcription Factors'. *Biochemistry* 45 (22): 6873–88. doi:10.1021/bi0602718.

Livi, Carmen Maria, Petr Klus, Riccardo Delli Ponti, and Gian Gaetano Tartaglia. 2015. 'catRAPID Signature: Identification of Ribonucleoproteins and RNA-Binding Regions'. *Bioinformatics (Oxford, England)*, October. doi:10.1093/bioinformatics/btv629.

Merkel, Dirk. 2014. 'Docker: Lightweight Linux Containers for Consistent Development and Deployment'. *Linux J.* 2014 (239). http://dl.acm.org/citation.cfm?id=2600239.2600241.

Paci, Emanuele, Jörg Gsponer, Xavier Salvatella, and Michele Vendruscolo. 2004. 'Molecular Dynamics Studies of the Process of Amyloid Aggregation of Peptide Fragments of Transthyretin'. *Journal of Molecular Biology* 340 (3): 555–69. doi:10.1016/j.jmb.2004.05.009.

Pawar, Amol P, Kateri F Dubay, Jesús Zurdo, Fabrizio Chiti, Michele Vendruscolo, and Christopher M Dobson. 2005. 'Prediction of "Aggregation-Prone" and "Aggregation-Susceptible" Regions in Proteins Associated with Neurodegenerative Diseases'. *Journal of Molecular Biology* 350 (2): 379–92.

Pearlman, David A., David A. Case, James W. Caldwell, Wilson S. Ross, Thomas E. Cheatham III, Steve DeBolt, David Ferguson, George Seibel, and Peter Kollman. 1995. 'AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules'. *Computer*

Physics Communications 91 (1–3): 1–41. doi:10.1016/0010-4655(95)00041-D.

Pollastri, Gianluca, and Aoife McLysaght. 2005. 'Porter: A New, Accurate Server for Protein Secondary Structure Prediction'. *Bioinformatics* 21 (8): 1719–20. doi:10.1093/bioinformatics/bti203.

Rao, H B, F Zhu, G B Yang, Z R Li, and Y Z Chen. 2011. 'Update of PROFEAT: A Web Server for Computing Structural and Physicochemical Features of Proteins and Peptides from Amino Acid Sequence'. *Nucleic Acids Research* 39 (Web Server issue): W385–90. doi:10.1093/nar/gkr284.

Rice, P, I Longden, and A Bleasby. 2000. 'EMBOSS: The European Molecular Biology Open Software Suite'. *Trends in Genetics: TIG* 16 (6): 276–77.

Rost, Burkhard. 1996. '[31] PHD: Predicting One-Dimensional Protein Structure by Profile-Based Neural Networks'. http://www.sciencedirect.com/science/article/pii/S0076687996660339.

Russell, Robert B., and Geoffrey J. Barton. 1993. 'The Limits of Protein Secondary Structure Prediction Accuracy from Multiple Sequence Alignment'. *Journal of Molecular Biology* 234 (4): 951–57. doi:10.1006/jmbi.1993.1649.

Schein, C H. 1990. 'Solubility as a Function of Protein Structure and Solvent Components'. *Bio/technology (Nature Publishing Company)* 8 (4): 308–17. doi:1369261.

Sigrist, Christian J. A., Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, and Philipp Bucher. 2002. 'PROSITE: A Documented Database Using Patterns and Profiles as Motif Descriptors'. *Briefings in Bioinformatics* 3 (3): 265–74.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. 'Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles'. *Proceedings of the National Academy of Sciences* 102 (43): 15545–50. doi:10.1073/pnas.0506580102.

Tartaglia, Gian Gaetano, Sebastian Pechmann, Christopher M Dobson, and Michele Vendruscolo. 2007. 'Life on the Edge: A Link between Gene Expression Levels and Aggregation

Rates of Human Proteins'. *Trends in Biochemical Sciences* 32 (5): 204–6. doi:10.1016/j.tibs.2007.03.005.

Tartaglia, Gian Gaetano, and Michele Vendruscolo. 2008. 'The Zyggregator Method for Predicting Protein Aggregation Propensities'. *Chemical Society Reviews* 37 (7): 1395–1401. doi:10.1039/b706784b.

van den Berg, Bastiaan A, Marcel JT Reinders, Johannes A Roubos, and Dick de Ridder. 2014. 'SPiCE: A Web-Based Tool for Sequence-Based Protein Classification and Exploration'. *BMC Bioinformatics* 15 (March): 93. doi:10.1186/1471-2105-15-93.

Wagner, Ingrid, and Hans Musso. 1983. 'New Naturally Occurring Amino Acids'. *Angewandte Chemie International Edition in English* 22 (11): 816–28. doi:10.1002/anie.198308161.

Wilkins, M R, E Gasteiger, A Bairoch, J C Sanchez, K L Williams, R D Appel, and D F Hochstrasser. 1999. 'Protein Identification and Analysis Tools in the ExPASy Server'. *Methods in Molecular Biology (Clifton, N.J.)* 112: 531–52.

Wilkinson, D L, and R G Harrison. 1991. 'Predicting the Solubility of Recombinant Proteins in Escherichia Coli'. *Bio/technology (Nature Publishing Company)* 9 (5): 443–48.