Università
degli Studi
di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di **Ingegneria dell'Informazione**

SCUOLA DI DOTTORATO DI RICERCA IN: **INGEGNERIA DELL'INFORMAZIONE**
INDIRIZZO: **SCIENZA E TECNOLOGIA DELL'INFORMAZIONE**
CICLO XXVII

Computational Approaches to Address the Next-Generation Sequencing Era

Direttore della Scuola: Ch.mo Prof. Matteo Bertocco
Coordinatore d’indirizzo: Ch.mo Prof. Carlo Ferrari
Supervisore: Ch.mo Prof. Carlo Ferrari

Dottorando: Manuel Giollo

Ringraziamenti

Guardandomi indietro dopo un percorso della durata di tre anni, mi rendo conto che la strada che ho intrapreso é costellata di eventi fondamentali. Ci sono stati momenti intensi, grandissime soddisfazioni, e sicuramente molto impegno e dedizione. Trovo quindi doveroso ringraziare il Professor Carlo Ferrari per il suo supporto, la sua guida esperta mi ha aiutato a vedere cose che avrei capito solo dopo anni di ricerca. Un doveroso grazie anche al Professor Silvio Tosatto per la collaborazione nel Biocomputing lab di Padova. Ricordo con piacere il lavoro asfissiante del CAFA durante il periodo natalizio con Damiano Piovesan, e le centinaia di ore passate in mensa con Ian Walsh discutendo di come far funzionare le cose. Sarò eterno debitore di Tomás Di Domenico per le discussioni e per il suo supporto anche dopo aver lasciato Padova. E naturalmente, Giovanni Minervini, Alessandro Masiero, Emilio Potenza, Alessandra Gasparini, Simona Aufiero, Lisanna Paladin, Marco Carraro, Layla Martinez, Ivan Micetić, Claudio Bassot, Emanuela Leonardi e Alberto Martin. Con ognuno di loro ho trovato discussioni stimolanti e una tazza di caffé in caso di bisogno.

Un grosso grazie anche alla mia famiglia. Loro ci sono sempre stati e hanno sempre creduto a me e alle mie capacità. Un ringraziamento speciale anche alla mia futura sposa Veronica, che ha saputo aspettarmi anche quando stavo inseguendo i miei sogni. E un grazie ai miei grandissimi amici, sia per le ore passate insieme tirando dadi, sia per le chiacchierate sulle cose più svariate o le lamentele dell'ennesimo ritardo di trenitalia. Porterò sempre tra i miei ricordi gli eoni passati in treno con Mauro e l'entusiasmo di Andrea per il suo lavoro.

In qualche modo, siete stati tutti un punto di riferimento che ha reso possibile questa fantastica avventura. Spero di continuare ad avervi nel mio cammino per lungo tempo.

Sommario

In questa tesi, vengono proposti nuovi algoritmi e modelli per affrontare problemi biologici. L'informatica svolge un ruolo chiave nella proteomica e nella ricerca genetica dovuto alla gestione delle grandi moli di dati biologici. Nel contesto dello studio di proteine, ho sviluppato nuovi metodi per la predizione delle loro funzioni basati su principi di reperimento dell'informazione. Utilizzando fonti eterogenee di conoscenza, come la ricerca su grafi e la similarità di sequenze, ho progettato uno strumento chiamato INGA che può essere utilizzato per annotare interi genomi. Questo è stato valutato imparzialmente dal Critical Assessment of Function Annotation, e ha dimostrato di essere uno degli approcci più efficaci per l'inferenza di funzione.
Per meglio caratterizzare le proteine dal punto di vista strutturale, ho proposto una strategia di rilevamento delle conformazioni delle proteine basata su rete di interazione di residui (RIN). Le reti RIN sono state quindi estese per gestire le fluttuazioni temporali delle coordinate atomiche. Tali grafi sono stati infine generati automaticamente da algoritmi di clustering. Un'implementazione chiamata RING MD ha evidenziato efficacemente i principali amminoacidi noti per essere funzionalmente rilevanti nell'Ubiquitina. Questi aminoacidi sono infatti molto importanti per spiegare la dinamica strutturale della proteina. Con la stessa logica, sono stati usati i grafi RIN anche per prevedere l'impatto delle mutazioni all'interno di una struttura proteica. Combinando informazioni sul nodo mutante in una rete e le sue caratteristiche, una rete neurale artificiale è stata addestrata per stimare la variazione di energia libera di Gibbs all'interno di una proteina. Cambiamenti estremi nell'energia interna potrebbe portare all'unfolding della proteina, ed eventualmente ad una malattia. D'altro canto, anche la riduzione della flessibilità proteica può ostacolare la sua funzione. Ad esempio, le fluttuazioni estreme osservate nelle proteine intrinsecamente disordinate (IDP) sono fondamentali per le loro attività. Per studiare le IDP, ho contribuito

alla raccolta del più grandi dataset di regioni disordinate mai esistito. Nella seguente analisi è stato dimostrato quali sono le funzioni tipiche di queste sequenze e i processi biologici in cui sono coinvolte. Data l'importanza della loro identificazione, una valutazione globale di predittori del disordine è stata eseguita per mostrare quali sono i metodi più efficaci e le loro limitazioni.

Nel contesto della genetica, mi sono concentrato sulla previsione di fenotipi. Durante il Critical Assessment of Genome Interpretation (CAGI), ho proposto nuovi approcci per l'analisi dei dati dell'esoma progettati per valutare il rischio di morbo di Crohn e di ipercolesterolemia. Queste sono spesso definite come malattie complesse, dal momento che il meccanismo alla base della loro insorgenza è ancora sconosciuto. Nel mio studio, i campioni umani con un arricchimento di mutazioni in geni critici sono stati predetti come soggetti a rischio genetico elevato. Oltre ai geni associati alla malattia, le reti di interazione proteiche sono state considerate per valutare l'accumulo di varianti in pathway biologici. Tale strategia ha dimostrato di essere tra le migliori secondo gli organizzatori del CAGI. Nel caso più semplice dei tratti mendeliani, con BOOGIE ho progettato un metodo per la predizione dei gruppi sanguigni umani basata su dati di esoma. Esso utilizza una versione specializzata dell'algoritmo nearest neighbour al fine di far corrispondere le varianti genetiche in un esoma non annotato con quelle disponibili in una base di conoscenza di riferimento. L'esempio più simile è usato per trasferire il gruppo sanguigno. Con una precisione superiore al 90 %, BOOGIE è un prototipo che mostra le potenziali applicazioni della predizione genetica, e può essere facilmente esteso a qualsiasi tratto mendeliano.

Riassumendo, questa tesi è una risposta parziale alla crescita esponenziale di sequenze disponibili che necessitano ulteriori esperimenti. Integrando informazioni eterogenee e la progettazione di nuovi modelli predittivi basati su apprendimento automatico, ho sviluppato nuovi strumenti per l'analisi di dati biologici e per la loro classificazione. Tutte le implementazioni sono liberamente disponibili per la comunità e potrebbero essere utili durante indagini future come in studi di malattie e nella progettazione di farmaci.

Abstract

In this thesis, I propose new algorithms and models to address biological problems. Computer science in fact plays a key role in proteomics and genetics research due to the advent of big datasets. In the context of protein study, I developed new methods for protein function prediction based on information retrieval principles. By using heterogeneous source of knowledge, like graph search and sequence similarity, I designed a tool called INGA that can be used to annotate entire genomes. It has been benchmarked during the Critical Assessment of Function Annotation challenge, and it proved to be one of the most effective approach for function inference.

To better characterize proteins from the structural point of view, I proposed a protein conformers detection strategy based on residue interaction network (RIN) data. RIN graphs were extended to deal with the time-dependent protein coordinate fluctuations, and were generated by clustering algorithms. An implementation called RING MD highlighted effectively the key amino acids known to be functionally relevant in Ubiquitin. These amino acids in fact are very important to explain the protein three-dimensional dynamics. With the same rationale, RIN graphs were used also to predict the impact of mutations within a protein structure. By combining information about a mutant node in the network and its features, an artificial neural network was trained to estimate the free Gibbs energy change of a protein. Extreme changes in the internal energy might lead to the protein unfolding, and possibly to disease. The reduction of a protein flexibility may hamper its function as well. As an example, the extreme fluctuations observed in intrinsically disordered proteins (IDPs) are fundamental for their activities. To better understand IDPs, I contributed in the collection of the largest dataset of disordered regions. In the following analysis, it was shown what are the typical functions of these sequences and the biological processes where they are involved. Due to the importance of their detection, a comprehensive assessment of disorder

predictors was performed to show what are the state-of-the-art methods and their limitations.

In the context of genetics, I focused on phenotype prediction. During the Critical Assessment of Genome Interpretation (CAGI), I proposed new approaches for the analysis of exome data to prioritize the risk of Crohn's disease and abnormal cholesterol levels. These are often defined as complex disease, since the mechanism behind their insurgence is still unknown. In my study, human samples with an enrichment of mutations in critical genes were predicted to have an high genetic risk. In addition to disease associated genes, protein interaction networks were considered to better account for variants accumulation in biological pathways. Such strategy was shown to be among the best approaches by CAGI organizers. In the simpler case of Mendelian traits, with BOOGIE I designed a method for human blood groups prediction based on exome data. It uses a specialized version of nearest neighbor algorithm in order to match the gene variants in an unannotated exome with the ones available in a reference knowledge base. The most similar hit is used to transfer the blood group. With an accuracy above 90%, BOOGIE is a proof-of-concept that shows the potential applications of genetic prediction, and can be easily extended to any Mendelian trait.

To summarize, this thesis is a partial answer to the exponential growth of sequences available that need further experiments. By integrating heterogeneous information and designing new predictive models based on machine learning, I developed novel tools for biological data analysis and classification. All implementations are freely available for the community and might be helpful during future investigations like in drug design and disease studies.

Contents

Chapter 1

Introduction

1.1 The cell machinery: genes and proteins

In the early stages of genomics the study of *single* proteins and genes was the principal topic for molecular biologists. Focused analytical approaches determined as much properties as possible about a research subject, lead to a very deep understanding for a small set of interesting targets. Things changed with the advent of high-throughput instruments like Next-Generation Sequencing (NGS) technologies, which analyze a huge spectrum of molecules in parallel. This observation of thousand to million variables enabled the quantitative analysis of complex systems in the whole, shifting the research from a *single* protein or gene to the *entire* cell metabolism. As a result, modern biology requires very often the use of computer science and statistical methods to handle large volumes of experimental data and generate new knowledge - all tasks typically managed in bioinformatics research.

The main topics addressed in bioinformatics are *genes*, *proteins* and their interactions under different conditions. From the biochemical point of view, a gene is the sequence of nucleotides (adenine, cytosine, guanine, and thymine) containing the information necessary for the production of one or more proteins. More than 20.000 genes are known in the human genome, but the overall number can vary a lot in different organisms. Inside a gene, one can identify (see Fig. 1.1) a set of coding regions (exons), a set of non-coding regions (introns) and a promoter region (UTR). Exons contain the information to produce a protein, while the promoter region and introns are responsible respectively for the gene expression regulation and the alternative splicing.

Whenever a gene sequence is copied by the polymerase enzyme, a sophisticated process transforms the mRNA into a protein.

Proteins are responsible for a vast amount of different roles, like enzymatic,

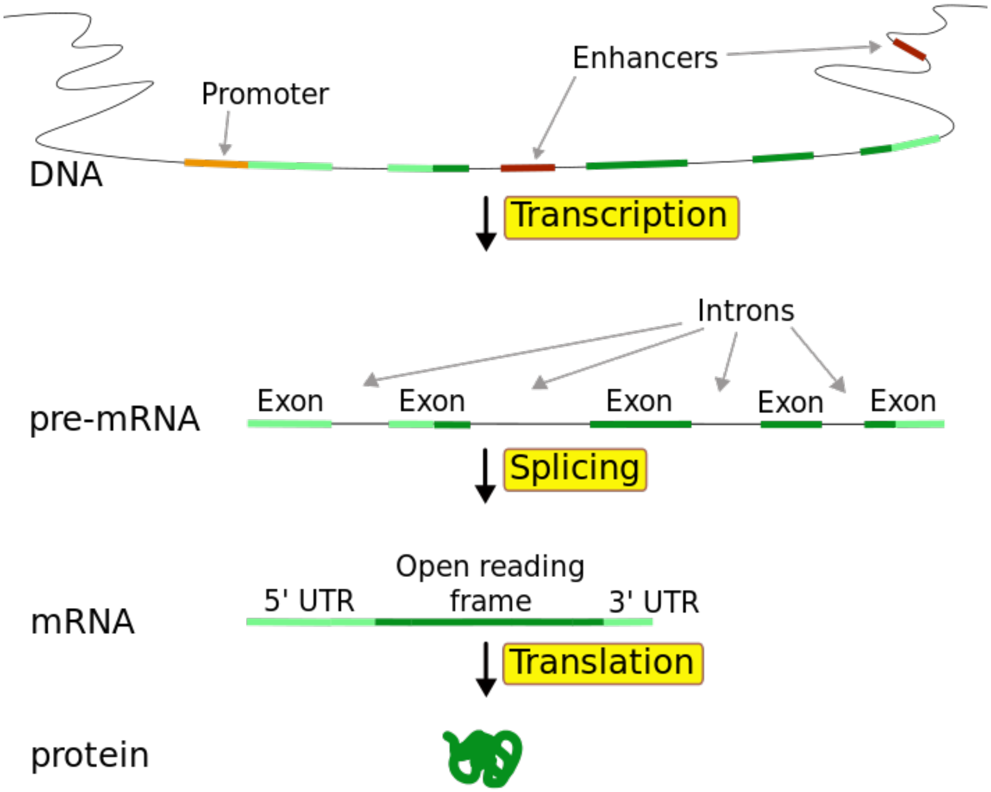


Figure 1.1: Diagram of the gene-to-protein process. The structure of a gene is also shown.

signaling, structural and mechanic activities. They are formed by sequences of 20 amino acids (or residue) which can be determined directly from the nucleotide triplets in the exons. What's important about amino acids is their different atomic composition and structure resulting in different chemical properties. The inter-atomic interactions of residues lead to the protein *folding*, which is the three-dimensional arrangement of atoms with minimal internal energy. From the structural point of view, one can identify four information levels (Fig. 1.2):

Primary structure: the amino acids sequence, which can be encoded as a string with a 20 character alphabet. This information is typically used to emphasize evolutionary information, like genetic similarity among different organisms.

Secondary structure: the first effect of amino acids chemical composition is the local rearrangement of atoms into *α-helix*, *β-sheet*, *coils* and *intrinsic disorder*. Such structure is determined uniquely by the amino acid sequence. Unlike helices and sheets, which show a static structure, the latter two have high flexibility.

Tertiary structure: the tree-dimensional coordinates of the entirely folded protein sequence. In this large structure, one can detect a set of protein *domains*, representing peculiar folding related to well defined protein functions.

Quaternary structure: sometimes two or more proteins are combined together to form a larger complex. This set of proteins might change their tree-dimensional arrangement due to local atomic interactions.

Roughly speaking, one can think that protein sequence determines structure, while protein structure determines function. However, things are more complicated in reality, due to additional factors that perturb the environment of a protein. As an example, proteins interact among themselves and with other molecules (e.g. water, metals and ligands) to perform their activities in the so called *interactome*. Such interactions and natural molecular dynamics may result in structural changes, which have very often a functional relevance. The same gene expressed in different tissue or cellular components can have a diverse protein folding and behaviour, so it is important to evaluate precisely the environmental context in bioinformatics research.

1.1.1 Data banks

Data openness has become more and more important in recent years, as it enables research reproducibility and promotes data reuse. This increasing amount of public information stored in the data banks of National Center for Biotechnology Information (NCBI) and in the European Bioinformatics Institute (EBI) helped bioinformatics development, and is nowadays one of the main sources of data. In terms of genomic data, the Ensembl

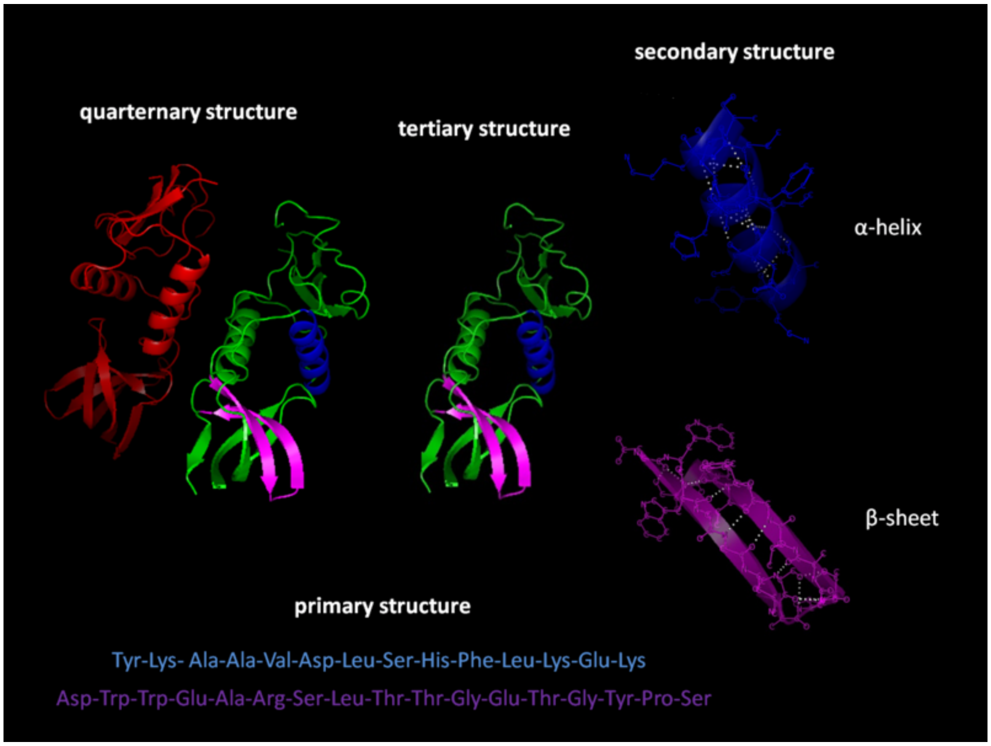

secondary structure
quarternary structure
tertiary structure
α-helix
β-sheet
primary structure
Tyr-Lys- Ala-Ala-Val-Asp-Leu-Ser-His-Phe-Leu-Lys-Glu-Lys
Asp-Trp-Trp-Glu-Ala-Arg-Ser-Leu-Thr-Thr-Gly-Glu-Thr-Gly-Tyr-Pro-Ser

Figure 1.2: The primary, secondary, tertiary and quaternary structure of a protein. Each level emphasizes information ranging from amino acids to protein complexes.

database [1] (`www.ensembl.org`) contains the whole genome for 69 organisms, with meta-data about gene expression levels for different tissue and mutations as well. The 1000 Genome project [2] (`www.1000genomes.org`) also managed to characterize human genetic variants frequency in order to analyze population structure and potential links with disease. Lots of knowledge about human mutations and diseases can also be obtained from dbSNP [3] (`www.ncbi.nlm.nih.gov/SNP/`) and GWAS central database [4] (`www.gwascentral.org`), providing functional annotation for thousands single nucleotide polymorphisms (SNP).

Protein level annotations are collected in the UniProt database [5] (`www.uniprot.org`), where functional terms, pathways, mutations and structural details are available. In the current release 2014_11 there are almost 90 million sequences, with their number growing exponentially due to the fast

development of sequencing technologies. Over 98% of sequences have no existing evidence at the protein or transcript level, and even less contain experimental annotation. This lack of knowledge is the key motivation that lead to the development of bioinformatics in recent years. For example, the use of the *transfer learning* principle can generalize the results of long and expensive experiments among *similar* proteins. Similarity can be measured in countless way, like sequence analysis, structural comparison, data mining in databases and from protein-protein interaction (PPI) networks. The String database [6] (`string-db.org`) is a clear instance of PPI, where proteins are represented as nodes and their interactions are encoded by edges. Different relationship types are possible, like co-expression in a tissue, neighborhood in a organism genome and experimentally validated. There are also examples of interacting proteins in the Protein Data bank [7] (PDB, `www.rcsb.org`), with more than 100,000 protein and DNA structures available. In fact, almost 5,000 structures show protein complexes, providing insights at the molecular level. From the experimental point of view, the most typical approaches to solve a protein structure are x-ray and nuclear magnetic resonance (NMR) crystallography, and in both cases one can obtain three-dimensional coordinates for all protein atoms.
Clearly, these are just few of the hundreds [8] databases useful for bioinformatic research, which demonstrates the ongoing effort in terms of data openness. Additional data banks and specific details will be introduced in the next sections to clarify research goals and methods.

1.2 Data science

The human ability to create accurate tools and measure events was one of the major contributing factors for the understanding of key natural principles. With the advent of high throughput technologies and big data, the detection of relevant signal is no more an easy task. The combination of heterogeneous skills becomes a critical factor to succeed in modern research, and bioinformatics lies exactly in the very important intersection between computer science, statistics and biology. As a data scientist, one has to start from a raw dataset and perform several operations to find out meaningful knowledge. The *knowledge discovery in database* [9] (KDD) process summarize the main steps that needs to be performed (see Fig. 1.3).

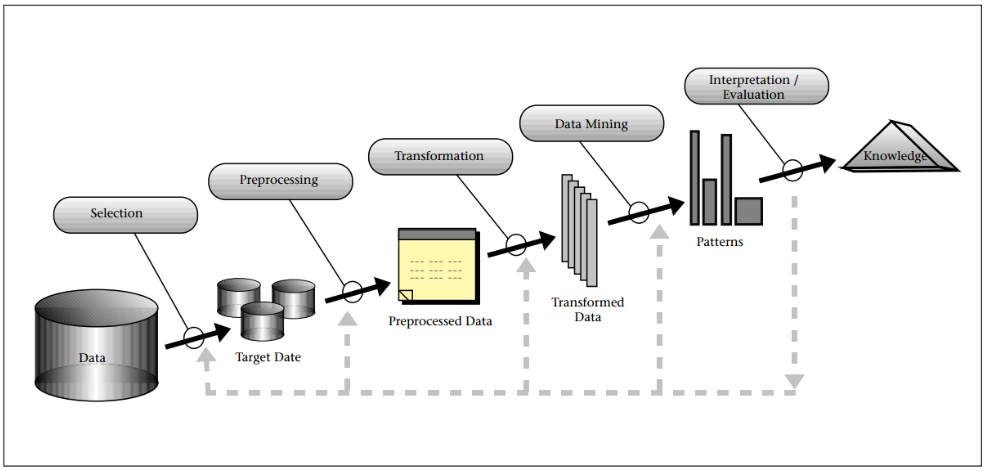


Figure 1.3: The KDD workflow, as shown in [9]. Five steps can transform raw data in useful knowledge. A very important concept is the iterative structure of the process, because any novel piece of information can suggest new scientific hypothesis that deserve analysis.

Selection given a scientific hypothesis, one should select from biological databases variables of interest and as much samples as possible. Sample size is strictly related to the specific research, and might be limited to a very small number. At this early stage, it is convenient to evaluate if statistical challenges can be addressed with the available data. As an example, in the context of pathogenic SNP analysis, one might try to compare the distribution of rare alleles in human UTRs and exons (CDS) using data from the 1000 genome (1000G) project. UTRs are the untranslated regions of a gene, while CDS are its coding part.

Preprocessing and transformation data sources can be very heterogeneous and hard to manage like web pages, text documents and images. As a rule of thumb, the "garbage in, garbage out" principle applies well in KDD, so one should be very careful during an initial data cleaning. Typical steps are the elimination of noisy samples and outliers, the management of missing variables and the normalization of observed attributes. The cleaning process can lead to a dramatic reduction in dataset size, so one should reconsider if samples are enough in statistical terms. Finally, all data have to be mapped to a unique and convenient format which is composed by the sample *fea-*

tures (the attributes). Features can be very complex and might require a combination of multiple data sources to help in the research question. To continue with the SNPs distribution example, one has to use a gene definition database like GENCODE [10] to map chromosome coordinates to either UTRs and CDSs (integration). Introns and intergenic SNPs are removed, as they are not important for the initial goal. Few genes have overlapping coordinates in the chromosome, leading to an apparent SNPs duplication in these regions that needs to be managed.
Variables of interest in bioinformatics research are generally related to gene and proteins sequences. In section 1.3 they will be introduced in depth.

Data mining seeking knowledge from complex data requires the use of sophisticated data mining and statistical methods. Common activities are *cluster analysis* [11], which is the discovery of unknown data subclasses (e.g. find automatically organism populations in a study), *association rule mining* [12], which variable relationships is found (e.g. find SNPs related to a disease) and *outlier detection* [13], which is the identification of unexpected behaviours (e.g. eco-system disturbances). In the SNPs example, one has to generate a probability distribution function of variants frequency for UTRs and CDSs (see Fig. 1.4) and compare them. The comparison can be tested by the null hypothesis of "identical frequency distribution for UTR and CDS variants" and the alternative hypothesis "CDS variants have lower frequencies than UTR SNPs" with the Wilcoxon rank-sum test. The p-value is close to 0, suggesting that there is a statistically significant difference in the variants distribution.

Interpretation the statistical significance of extracted patterns can be used to give confidence about the automatic data processing. However, the functional implication of correlated variables must be clear from the biological point of view. In the SNPs example, it seems that the distribution of variants is quite different for UTR and CDS variants. The former type is the most common for SNPs with frequency greater than 0.002. On the other hand, CDS are the most common when looking at lower frequencies. This is likely to be the result of strong evolutionary constraints in CDS regions, which reduces such regions variability .

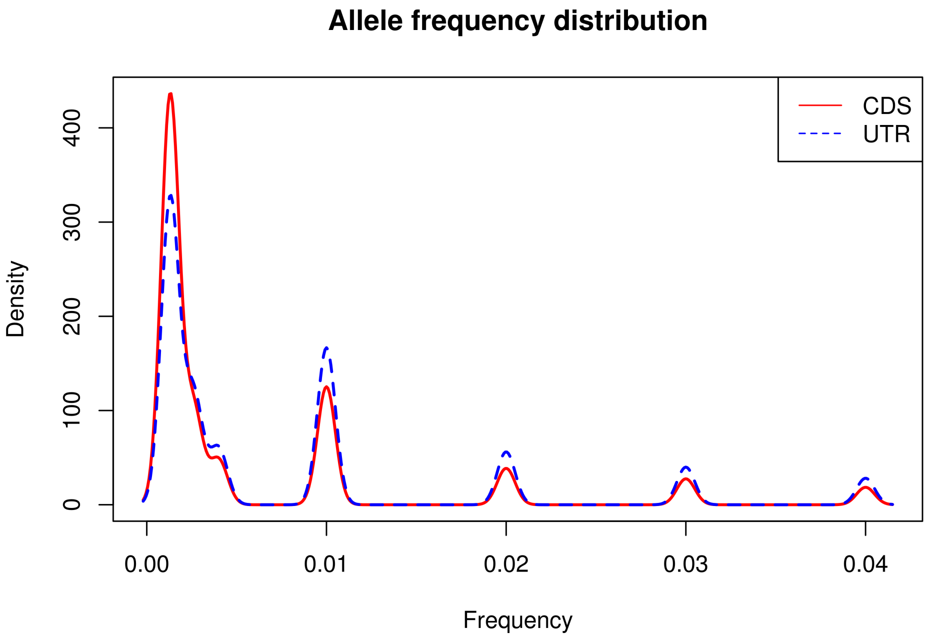


Figure 1.4: Probability distribution function of UTR and CDS variants with frequency below 5%. Exons contains a lower amount of common polymorphisms.

1.2.1 Machine learning

In order to use any knowledge for decision-making, one can replace the data mining step of KDD with machine learning (ML). A ML project will probably start from the results of a KDD research, so any feature proven to be discriminative for a biological process will be used to build computational models using *supervised* [14], *unsupervised* [15] or *reinforcement learning* [16]. Examples of algorithms are neural networks [17], support vector machines [18] and Bayesian networks [19]. To continue with the SNP frequency example, it is possible to predict whether a genome region is either a UTR or CDS. By looking at the consecutive variants frequency in a sequence, the enrichment of rare SNPs (minor allele frequency lesser than 0.01) can be detected using Hidden Markov Models just like in the *unfair casino* problem [20]. This very simple approach can be inaccurate, thus additional features like the sequence nucleotides can be considered for the prediction. UTRs in fact have an high GC-content, but this is not the sole property of these sequences: many attributes can be used to build a multivariate observation, which will be automatically managed by algorithms designed to compute optimized classification models. Classification is always related to *thresholds* that separates data in a optimal way. In the SNPs case a specific variant frequency could be used as a threshold as part of the learning strategy. Rather than using points, the classification of multidimensional data is achieved by hyperplanes (see Fig. 1.5).

Classification [21] is the process of predicting values into categories with the most used being simple binary categories. When the goal in machine learning becomes the estimation of a continuous variable, like the amount of an expressed gene, *regression* algorithms [21] are adopted. Hyperplanes can be once again the solution, even though non-linear functions can be preferred to minimize the error. Algorithms can have different classification performance depending on how they compute thresholds, thus a fair comparison will help to pick the best one for the scientific problem at hand. To do it properly, data have to be divided in three equal chunks called *training set*, *validation set* and *test set* [21]. The former two are used respectively to parametrize the model (e.g. the hyperplane) and to perform its quality assessment. As

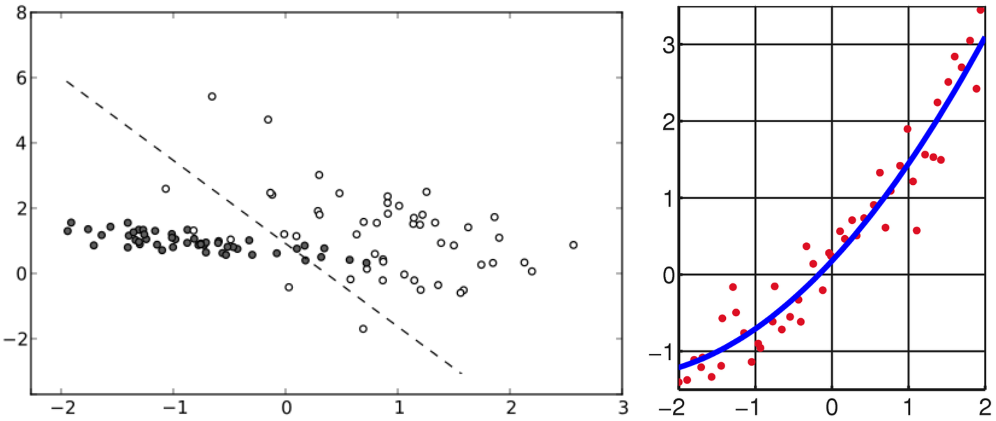


Figure 1.5: Linear classifier and regression. During the development of a machine learning method, hyperplanes and non-linear functions can be used to model data. Left: black and white points represents the categories. Classification and regression errors are common, even though they are minimized by means of an optimization process.

an example, neural networks optimize each neuronal unit weight by minimizing the classification error on the training set using a stochastic process. Different neural networks are trained in parallel, and the one whose random optimization lead to best performance on the validation set is selected. The test set is used at the end to evaluate fairly the best neural network, and to perform a comparison with other published methods. *K-fold cross validation* is a similar model validation approach where data division, training and testing are repeated multiple times in order to minimize the risk of bad data partitioning.

Quality measure Classification of an unseen target can be imprecise, so it is important to adopt performance measures which clarify the expected behaviour of a ML tool. In a test set it is possible to count *true positives* TP, *true negatives* TN (correct prediction for the two classes of interest), *false positives* FP and *false negatives* FN (wrong model prediction). These four value are sufficient to calculate all quality measures.

$$\begin{aligned} SENS &= \frac{TP}{TP+FN} & PREC &= \frac{TP}{TP+FP} \\ ACC &= \frac{TP+TN}{TP+FP+TN+FN} & SPEC &= \frac{TN}{TN+FP} \\ F &= 2 \times \frac{PREC \times SENS}{PREC + SENS} \end{aligned}$$

These formulas are respectively *sensitivity*, *precision*, *accuracy*, *specificity*, and *F-score*. They can give an idea about the overall performance of a classifier. These measures can be affected negatively by unbalanced classes, so the *Matthews correlation coefficient* can be considered as well:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Correlation is an important concept, which can be generalized for the regression problem. However, all these quality measures should be considered together, as they describe important aspects of a ML method performance.

1.3 Protein features and representation

Currently the distinct three-dimensional coordinates of proteins amount to 35 thousand. This was accumulated over a 50 year period and usually each can be found in the Protein Data Bank [7] (PDB). This data is of vital importance for medical research and biotechnology, where it plays a key role in drug and enzyme design. Research projects in structural bioinformatics typically aim to predict structural and functional information using as much data as possible, but experimental noise and missing observations require the use of statistical approaches to generate knowledge. Determinism is in fact an exception when dealing with complex datasets, so the use of machine learning, data mining and statistical methods is the only way to pick up significant signal. As an example, the goal of protein modeling is the prediction of a protein structure from its amino acid sequence.

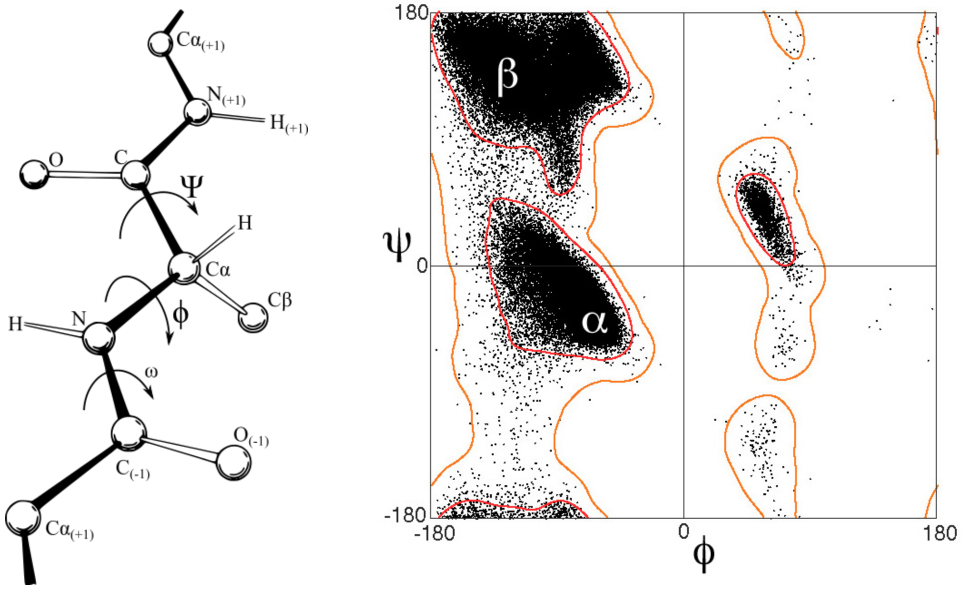
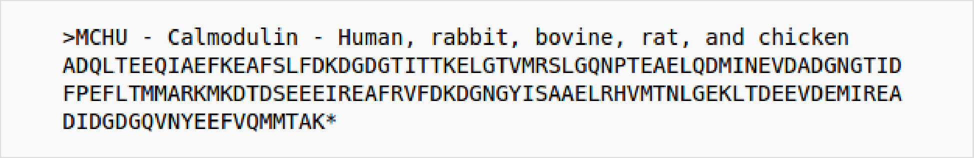


Figure 1.6: Left: the three-dimensional structure of an amino acid. The backbone is composed by nitrogen, carbon, oxygen and hydrogen. Right: Ramachandran plot. The backbone atoms in an amino acid have three degree of freedom from the geometrical point of view. The angles with most variability are ϕ and ψ, and their arrangement is associated with secondary structure.

The *ab initio* [22] approach exploits physical laws to generate a protein structure with minimal internal energy. On the other hand, the *homology modeling* [23] approach takes advantage of the knowledge in the PDB, and tries to use sequence-structure relationships to create a protein structure. These relationships are extracted from the data, so one can measure directly their likelihood, and use them to generate the most probable structure. The *Ramachandran plot* [24] (see Fig. 1.6) is a clear example of knowledge that can be obtained from protein structure data, providing a probability density function that guides the placement of an amino acids in a predicted structure. Best implementations use a mixture of concepts from ab initio and homology modeling, since their joint combination lead to the idea of efficient conformation exploration and proper quality evaluation of the final model.
In general, sequence and structural information are the most commonly used in structural bioinformatic projects due to their availability, so they will be introduced properly in the next paragraphs. It is important to note that such information becomes useful when it is used to predict unseen variables and protein properties that are hard to obtain without expensive and time-consuming experiments. Last but not least, predictions are a guess, so one should always try to evaluate with well established quality measure the accuracy of a computational method.

1.3.1 Sequence features

With an alphabet of just 20 residues, one can encode all protein genetic information exactly in FASTA files (see Fig. 1.7). A first macroscopic difference among proteins is primary sequence length, which ranges from few dozen residues to the $\sim 30,000$ amino acids of Titin protein, with an average of 317 according to UniProt 2014_11 [25]. The main challenge is therefore the interpretation of these protein “recipes”. With an alphabet size roughly similar to the one used in the western countries, bioinformatics keeps working on the process of understanding the basic words of protein language, but their semantic and relationships remain still unclear. *Sequence similarity* [26] is the key concept developed in bioinformatics research that motivated the use of computer science approach to meet biologist needs: given a protein sequence, a single amino acid random mutation probably will have no impact in terms of protein folding and function. Moreover, there are amino acids with similar chemical properties, which makes them interchangeable. Stochastic sequence variations are very common in the daily life of any organism, and form the

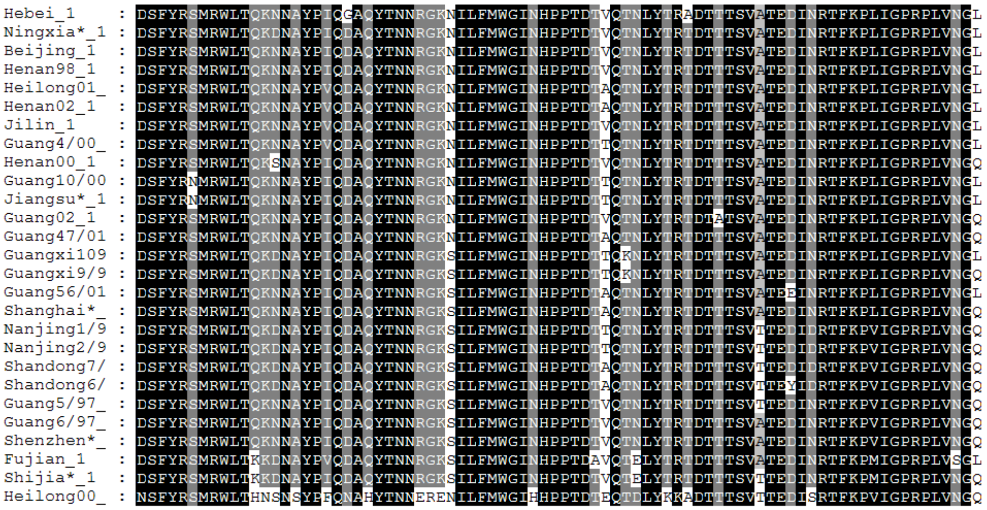


```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

Figure 1.7: FASTA file example. The header line, starting with the > character, denotes protein meta-data. The sequence is reported from the second line onward.

basis of evolution. Any sequence change with equal or increasing fitness will be kept in a population, while the ones with pathogenic effect tend to be removed. Lots of work was done to measure the maximum amount of residues that can be mutated while keeping the folding and function unchanged, resulting in the development of *protein alignment* tools [27]. These methods are designed to match the residues of two protein sequences, and therefore highlight similar sequence regions related to a common ancestry. From a practical point of view, in case of sequence similarity it is possible to transfer evidence observed in previously studied sequences to uncharacterized ones. This generalization comes with error risks, which can still be estimated by the sequence similarity score. This possibility to extend the knowledge gained from expensive experiments to million sequences automatically is the main strength of machine learning and statistical approaches.

Another important concept often used is *sequence modularity*. As an example, local sequence composition is critical for secondary structure (see Section 1.1), especially in the case of α-helix, and intrinsic disorder. In order to use protein sequence information to make structural and functional prediction, sequence modularity can be exploited by means of m-sized *sliding windows*. These windows are vectors in $\mathbb{R}^m$ describing a consecutive sequence region, where the first dimension encode for the residue observed at the i-th position of the protein sequence, the second dimension for the $i+1$-th position and so on. Therefore, one can compute $n-m+1$ different sliding windows from a protein with length n. As a potential application of these multivariate vectors, it is possible to find statistical associations between short sequences and secondary structure. However, a concrete analysis of data collected using sliding windows needs a way to compare these vectors. Each element is an amino acid encoded using characters, so the fundamental mathematical operations are not defined (e.g. the concept of distance between residue is unclear, and the lexicographic distance is not helpful). To overcome this issue, characters must be converted to meaningful values in $\mathbb{R}$. Possible solutions are the *one-hot encode*, where each residue is represented as a 20-dimensional binary vector. Just a single element equals one in such vectors, while all other elements are set to zero. As an example, the first amino acid in lexicographic order is encoded with the one in the first position, while the second residue according to the alphabet set to one the second position of the vector and so on. According to this definition, non-identical amino acids are orthogonal. In the *Atchley encode* [28], each residue is represented in a 5-dimensional space, which accounts respectively for polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. One can use just a subset of these features, if the research topic does not require some particular information. An effective practice adopted in the pathogenicity prediction problem [29] (i.e. predict whether a SNP will cause a protein loss of function) is the use of *residue conservation* value to encode amino acids. Conservation is obtained from a multiple alignment by evaluating the frequency of an amino acids in a given position (see Fig. 1.8). Conserved residues have very often a functional role, so this value is often considered an important attribute.



```
Hebei_1     : DSFYRSMRWLTQKNNAYPIQGAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRADTTTSVATEDINRTFKPLIGPRPLVNGL
Ningxia*_1  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Beijing_1   : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Henan98_1   : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Heilong01_  : DSFYRSMRWLTQKNNAYPVQDAQYTNNRGKNILFMWGINHPPTDTAQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Henan02_1   : DSFYRSMRWLTQKNNAYPVQDAQYTNNRGKNILFMWGINHPPTDTAQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Jilin_1     : DSFYRSMRWLTQKNNAYPVQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Guang4/00_  : DSFYRSMRWLTQKNNAYPVQDAQYTNNRGKNILFMWGINHPPTDTTQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Henan00_1   : DSFYRSMRWLTQKSNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGQ
Guang10/00  : DSFYRNMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTTQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Jiangsu*_1  : DSFYRNMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTTQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Guang02_1   : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTATSVATEDINRTFKPLIGPRPLVNGQ
Guang47/01  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTAQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGQ
Guangxi109  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTTQKNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Guangxi9/9  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTTQKNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGQ
Guang56/01  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTAQTNLYTRTDTTTSVATEEINRTFKPLIGPRPLVNGL
Shanghai*_  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTAQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGQ
Nanjing1/9  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTTQTNLYTRTDTTTSVTTEDIDRTFKPVIGPRPLVNGQ
Nanjing2/9  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTTQTNLYTRTDTTTSVTTEDIDRTFKPVIGPRPLVNGQ
Shandong7/  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTAQTNLYTRTDTTTSVTTEDIDRTFKPVIGPRPLVNGQ
Shandong6/  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTAQTNLYTRTDTTTSVTTEYIDRTFKPVIGPRPLVNGQ
Guang5/97_  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTVQTNLYTRTDTTTSVTTEDINRTFKPVIGPRPLVNGQ
Guang6/97_  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPVIGPRPLVNGQ
Shenzhen*_  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPVIGPRPLVNGQ
Fujian_1    : DSFYRSMRWLTKKDNAYPVQDAQYTNNRGKSILFMWGINHPPTDAVQTELYTRTDTTTSVATEDINRTFKPMIGPRPLVSGL
Shijia*_1   : DSFYRSMRWLTKKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTVQTELYTRTDTTTSVTTEDINRTFKPMIGPRPLVNGQ
Heilong00_  : NSFYRSMRWLTHNSNSYPFQNAHYTNNERENILFMWGIHHPPTDTEQTDLYKKADTTTSVTTEDISRTFKPVIGPRPLVNGQ
```

Figure 1.8: A protein alignment. Each row corresponds to an organism specific protein sequence. Sequence positions where amino acids does not change are conserved (black columns), while the lighter colors highlights lack of conservation.

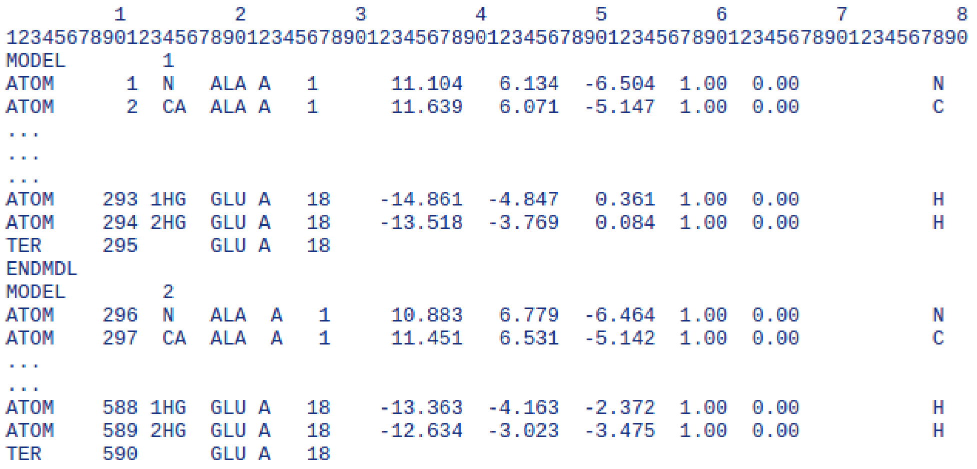
All the one-hot, Atchley and conservation encoding preserve the amino acids order, but they are strongly related to the size of the sliding window. When considering a large window size, one incurs in statistical and computational issue called *curse of dimensionality*. The *sequence composition* is an alternative encoding immune to this issue which simply stores in a 20-dimensional vector the frequency of each amino acids in a given sequence. As a results, a whole protein can be encoded using few aggregated variables, at the cost of amino acids ordering.

Many other attributes exists and can be conceived to emphasize information relevant for a particular research. As a rule of thumb, having a valid hypothesis about the important sequence-based features for protein prediction problems is the best starting point to succeed in machine learning and biostatistics.

1.3.2 Structural features

Shifting the focus from one-dimensional data to three-dimensional data pose great challenges. Sequences encoded in a simple textual format with no missing information are replaced with protein structure files [7] (PDB), which contain three-dimensional coordinates for each atoms in the protein (see Fig. 1.9), plus additional data for ligands and DNA. As a natural consequence of proteins large size and flexibility uncertainty may arise even in wet lab experiments.For example, missing atom coordinates or low quality values are often observed in protein structure data. In fact, solving a crystal structure requires by itself the use of complex statistical methods to interpret a electron density maps for x-ray PDBs, thus noise is an expected drawback. NMR structures by definition allow multiple model in different conformation, emphasizing flexibilities.

Despite these challenges in using this data, structural information can easily address problems otherwise unsolvable. As an example, secondary structure is an information directly available, and can be extracted with automatic tools [30]. In addition, missing amino acids coordinates in x-ray or highly fluctuating residues in NMR can be interpreted as disorder. One can distinguish buried and exposed amino acids by looking at their position, and assign them a *solvent accessibility* score [31]. The presence of additional elements like metals, molecules and post-translational modifications (PTM) can be identified, which is clearly unpredictable from a sequence. Cavities and holes, along with the orientation of amino acid side chains and surface elec-



```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
MODEL        1
ATOM      1  N   ALA A   1      11.104   6.134  -6.504  1.00  0.00           N
ATOM      2  CA  ALA A   1      11.639   6.071  -5.147  1.00  0.00           C
...
...
...
ATOM    293 1HG  GLU A  18     -14.861  -4.847   0.361  1.00  0.00           H
ATOM    294 2HG  GLU A  18     -13.518  -3.769   0.084  1.00  0.00           H
TER     295      GLU A  18
ENDMDL
MODEL        2
ATOM    296  N   ALA  A   1      10.883   6.779  -6.464  1.00  0.00           N
ATOM    297  CA  ALA  A   1      11.451   6.531  -5.142  1.00  0.00           C
...
...
ATOM    588 1HG  GLU A  18     -13.363  -4.163  -2.372  1.00  0.00           H
ATOM    589 2HG  GLU A  18     -12.634  -3.023  -3.475  1.00  0.00           H
TER     590      GLU A  18
```

Figure 1.9: The coordinates section of a PDB. The format is positional, and encode for 15 protein attributes. The former 8 are atom identifiers, while the following 5 describe respectively 3D coordinates and crystallization quality (defined in terms of occupancy and temperature factor).

trostatic charge are the key players in the development of drugs that could target a protein [32].
All this information is readily available, and can be used to train machine learning methods. The approach is similar to what was introduced in the previous paragraph: one can enrich the protein sequence vectorized representation by adding residue attributes about secondary structure, solvent accessibility and electrostatic charge. In parallel with respect to sequence composition encoding, there is a structure-based corresponding representation. Given a reference amino acids, the frequency of all residues within a sphere radius r is computed [33]. In other words, rather than counting the amino acids inside a sliding window, a virtual three-dimensional sphere is considered. Despite the similarities of the two encodings, the main difference lies in the new definition of *residue neighborhood*. This idea gained relevance with PDBs, since distant residues in the sequence (more than 25 positions) can be very close in a three-dimensional space, and might interact with each other. Hydrogen bonds, ionic bonds and Van-Der-Waals interactions are just few of the most common chemical bonds, and their existence are essential for the protein folding and function. From the practical point of view, *contact maps* [34] are matrices emphasizing these interactions (see Fig. 1.10).

They can also be transformed in *residue interaction networks* [35] (RINs), a protein graph where nodes correspond to amino acids, while edges represent interactions. By working directly in these interaction-centric structures, the concept of neighborhood is highlighted. One can also try to exploit the solid theoretical background of networks to extract protein structural features. In fact, high node centrality is very often a common feature for functionally relevant amino acids [36], so it can be used as a discriminative attribute for protein analysis.

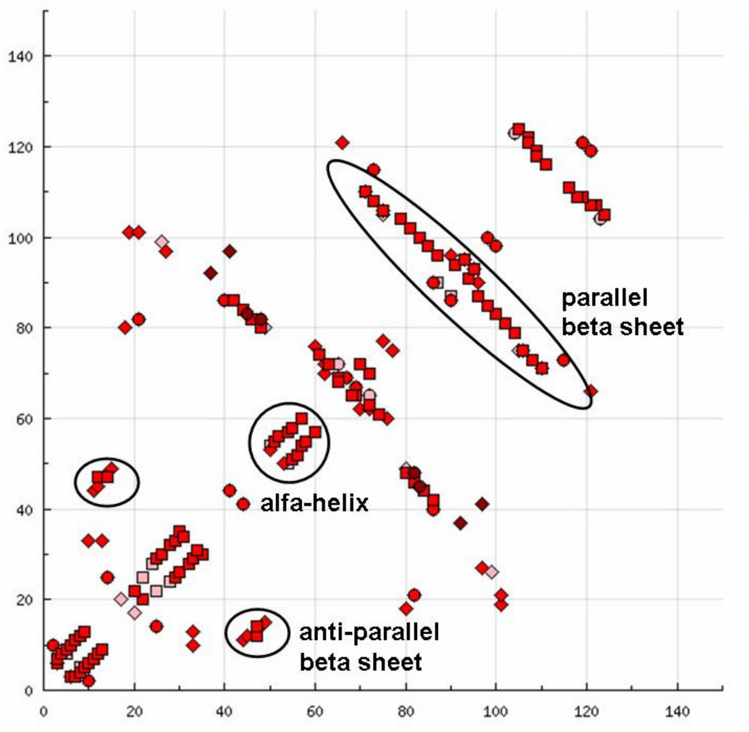


Figure 1.10: A sample contact map. A point at $\langle x, y\rangle$ coordinate shows that the x-th and the y-th residues share an hydrogen bond. Other type of interactions are possible, like hydrogen bonds, π-cation interactions, π-π stacks, ionic bonds, disulfide bonds and Van-Der-Waals interactions.

An interesting sequence-based interaction between residue is *mutual information* [37], a statistical measure that estimates the co-evolution of amino acids. It is computed from a multiple alignment on sequence position pairs by looking at mutual residues change in different organisms. To put it sim-

ple, mutual information shows from the statistical point of view how much the presence of an amino acid influence a second one in a different position. Evolution is the key actor, as structurally and functionally related residues are expected to co-evolve.
To conclude, contact maps, RINs and PDBs are all equivalent protein representations, but each of them is very useful to emphasize a particular type of information. All the structural data that they encode can be either the objective function of a sequence-based predictor (like in the case of secondary structure prediction or protein structure modeling), or a new feature that will be considered by a machine learning predictor. Such addition will lead obviously to higher prediction accuracy, at the cost of extra prior knowledge. When users expect high predictive power, like in drug design, the use of three-dimensional data is critical.

1.3.3 Protein annotation

The PDB reports primary amino acid sequence and their corresponding structural information. A third level of heterogeneous meta-data called *annotation* is used very often to associate additional knowledge to proteins. Unlike the information introduced in the previous paragraphs, annotation in general is not a property of an amino acid, but an attribute of the whole protein. As an example, the UniProt database [25] reports that NOD2 protein interacts with another 29 proteins, shows few mutations associated with Crohn's disease and is highly expressed in bone marrow and leukocytes. Very often, the goal in bioinformatics is the prediction of this information for uncharacterized proteins, even though in some cases it can be used as feature to build a predictor. Predicting the protein function from sequence and interactome data is just one of the possible applications, so it is important to be aware of the available knowledge:

Organism and phylogenetics the taxonomy lineage, ortholog proteins and phylogenetics trees can specify organism relationships with other species and improve sequence alignments. As an example, one might decide to use only sequences of closely related organisms, or try to find all the sequence paralogs in mammals of a given protein.

Protein function the *gene ontology* (GO) [38] defines formally all the known molecular function, biological process and cellular components. It is a vocabulary more than 40,000 terms, which are organized in tree-like data structures where nodes can have multiple parents (a directed acyclic graph). This formal and hierarchical representation of specific concepts is of great importance, but most of this annotation comes from predictions with no experimental validation. The Critical Assessment of Function Annotation [39] (CAFA) challenge is a scientific experiment covered in this manuscript that aims to assess the state of the art in terms of GO prediction. The problem will be discussed with more details in section 2.2.

Protein domains a whole protein sequence is like a sentence, and domains are equivalent to words. These regions of various length have a specific functional role and structural topology. The *PFAM* database [40] (`pfam.xfam.org`) reports almost 15,000 different domains, which are detected using Hidden Markov Models on aligned sequences. Even though 3,000 have no functional characterization yet, nearly 80% of proteins in UniProt are annotated with a PFAM domain, enabling a quick and simple categorization of sequences. In the context of the CAFA challenge described in section 2.2, this information was of great relevance for protein function prediction.

Variants and disease biology is very often a human-centric research. Therefore, protein variants observed in humans and the corresponding genetic disease are reported in order to clarify the protein role in medical terms. Recently, this information was used effectively to train methods for pathogenicity prediction [29] and disease gene prioritization [41]. This data source was a starting point during the Critical Assessment of Genome Interpretation (CAGI) challenge (described in detail in section 5.2), where human exomes were ranked according to their genetic risk to develop specific diseases.

Interactions proteins are the main actors inside a cell, but their activity is not meaningful if considered outside their interaction network [6]. Proteins in fact can create larger complexes and might alter each other to transmit a signal, like in the case of post-translational modifications. Research on protein protein interaction (PPI) networks showed that biological networks are scale-free, so they are structurally tolerant to the presence of faulty proteins due to the limited amount of central nodes. PPI databases are very important to interpret biological processes, but the available knowledge graph is incomplete and imprecise. This source of information was used both in the CAFA and CAGI challenges to contextualize the analyzed proteins.

Expression when considering a specific tissue, the overall amount of a certain protein can express its function in the cell. There are ubiquitous proteins, which are important for cell survival, while proteins specific to few tissues are the ones with a differential role. Comparative gene expression studies can provide insight about interacting proteins and disease risk.

Literature references with more than 20 millions published papers in PubMed, a huge amount of knowledge encoded in natural language is in reality buried, waiting to be indexed effectively in a database. Manual curators populate many datasets with high quality annotation, and report citations to the original source material. In the next future, text mining techniques for the extraction of published knowledge are likely to be the only sustainable option, and computer science will play a key role with the development of modern natural language processing techniques.

There is clearly much more in public databases, but the vast majority of knowledge needed for this manuscript is covered by these concepts. Lots of this information is very useful in medicine and biology, and therefore become the objective function for a machine learning tool. As already pointed out, no one expects deterministic knowledge due to the huge complexities related to the prediction of molecular mechanisms or disease. However, the development of computational methods that prioritize experimental choices is a rational approach to investigation, which motivates bioinformatics research.

1.4 The scientific contribution of this thesis

To summarize, my research was focused on protein annotation and phenotype prediction. My contribution on each project was mainly related to the integration of heterogeneous information and the design of new predictive models useful to analyze biological data, which can be labeled as data science. With the background introduced in the previous sections it will be easier to understand the scientific topics addressed in this thesis, as they are all biological problems where machine learning and data mining represent the state-of-the-art approach.

Molecular dynamics

Molecular Dynamics (MD) simulations [42] are well known to provide good insight about the possible three-dimensional conformations of a protein using force fields that emulates the classical physics of molecules in water. They are commonly used for drug design, as they can predict the affinity between a protein and a drug. Unfortunately, these simulations produce a overwhelming amount of data, which is often hard to interpret. Moreover, a MD user might aim to get an overview of the simulation before starting any investigation. During my work I focused on a MD case study (the Ubquitin protein), and I proposed a statistical model of RINs [35] for protein dynamics.Time-dependent RINs were constructed, and their relationships were captured using k-means algorithm and Hidden Markov Models (HMM). As a result, protein conformational states were detected, and their representative structures could be obtained to simplify the analysis of a MD. In addition, state transitions were parametrized, providing insight about the relationship among states. This kind of analysis was performed for six types of residue interactions, like hydrogen bonds and ionic bonds, allowing a detailed study of a target protein. The model itself can be used also for the comparison of different MD of the same protein, like in the case where wild-type and mutant proteins are evaluated for the study of a disease. The work was awarded as one of the best posters in Proteine 2014 conference.

Protein disorder

Large protein flexibility is often a product of *protein disorder* [43]. This is a relatively new research topic and it is providing new insights about the dynamic behavior of proteins. I contributed in the construction of a new collection of disordered sequences using missing coordinates in x-ray PDBs. By doing this, we extracted the largest dataset of disordered sequence regions with one order of magnitude more samples with respect to previous studies. Over 1,700 proteins contained long disordered regions of at least 30 residues, suggesting that they could play a functional role. By means of such unique dataset, I evaluated the accuracy of state-of-the-art tools for disorder prediction. Best tools and practices for disorder prediction were highlighted, even though it seems that current methods still lack the desired precision for a proper protein annotation. Moreover, there is a clear bias in the training of most methods (like the use of short disorder fragments and the focus on PDB chains rather than protein sequences), which might affect future investigation on the field.

Protein function prediction

With the unprecedented increase of sequences deposited in public database, automatic function prediction is the only feasible solution to characterize millions of proteins. Protein function space is defined by Gene Ontology (GO) [38], which describes the actual *Molecular Function*, the *Cellular Component* where the function is performed and the *Biological Process*. From the machine learning perspective, this is a multi-class problem, since the whole ontology contains 41,947 classes that have to be associated to new sequences. Given a query protein, three automatic information retrieval systems were built to extract potentially useful annotation from sequence databases, PFAM protein domains database and STRING-db protein interaction networks. The three predictions were integrated in a consensus based on *generalized additive models* which resulted in higher predictive power compared to the single ones. An implementation of this consensus called Interaction Network GO Annotator (INGA) is freely available for the community (url: `protein.bio.unipd.it/inga`). INGA has been tested in the *Critical Assessment of Function Annotation* [39] experiment (CAFA, url: `biofunctionprediction.org`), a community scientific challenge where about 50 research group predicted GO terms for 100,000 protein sequences. In that event, my approach was recognized as a top performer.

Protein stability

Given a protein sequence, a residue change (i.e. a *non-synonymous Single Nucleotide Polymorphism*) can lead to marked differences in its final structure due to increased compactness or flexibility. This might cause severe damage for to a protein, and eventually cause a loss of function and possibly a disease. The variation in the *Gibbs free energy* $\Delta\Delta$G is a physical measure of such structural change [44]. In the context of *protein stability* prediction, the goal is the evaluation of $\Delta\Delta$G upon mutations. To estimate this value in the presence of a mutation, I combined sequence and structural information according to a *data fusion* paradigm. In particular, the novelty of the work lies in the use of *Residue Interaction Networks* [35] (RIN) for protein structure representation. In RIN graphs, nodes are amino acids, while edges represent residue interactions. RINs can be used to evaluate the importance of a residue using graph centrality measures, which can be thought as descriptors of the overall protein structure. I trained a regression model of the data using a neural network, and tested it in well a established dataset. The NEtwork Enthalpic MOdelling (NeEMO) is the publicly available implementation of this work (url: `protein.bio.unipd.it/neemo`), which was shown to have higher accuracy compared to other state-of-the-art tools, mainly due to RIN features contribution.

Phenotypes and Genetics

The sharp reduction of genome sequencing cost is opening new possibilities for disease studies and medical diagnostics. This huge opportunity raised nevertheless complex scientific and technological question, which are still largely unsolved. When profiling the genetic information, the big amount of information is the first challenge that has to be addressed. Gigabites of data per sample must be managed to reconstruct a human genome, and the following analysis is even more complex [45]. In the context of disease risk prediction, the *Critical Assessment of Genome Interpretation* (CAGI) is an international challenge where participants are asked to detect the health status for the samples given their genetic data. As an example, in the Crohn's Disease challenge the exomes of 66 people had to be classified for disease risk. Similarly, the Familial Combined Hyperlipidemia challenge (FCH) goal was the prediction of the cholesterol and triglycerides level for a family compostedby 5 individuals. Potentially dangerous mutations had to be found and prioritized. Given the general absence of training data in CAGI, my approach was mainly related to the KDD process. All coding variants placed in genes associated with the disease of interest were taken into account. To increase the coverage of genes potentially involved in the disease onset, protein interaction networks were used to retrieve pathway variants. Several features were extracted for the selected mutations in order to evaluate their disease association. In agreement with *burden tests* idea [46], the presence of many overrepresented variants in a single individual was used to rank the samples in the dataset. In other words, the higher the number of mutations, the

stronger the disease risk evidence. Using this simple but powerful method, we ended up among the top performers among 30 groups for the two challenges mentioned above.
In the context of CAGI phenotype prediction, I also developed *BOOGIE*, a tool for the prediction of blood groups (like ABO or Rh) using genotype data. Mutations relevant for the blood type were collected from the BGMUT public database [47], and processed in order to match them with current gene definition database. Genetic knowledge about a given blood group was represented as binary rules, and constituted the reference database where I could map the DNA of a new sample. The matching method uses jointly a k-nearest neighbor algorithm and a *haplotype phasing* technique to match a patient allele set to the database just described. Even though accuracy for ABO and Rh is high (higher than 90%), it is not as good as the experimental serological tests. However, BOOGIE is a proof-of-concept: it shows that important traits can be detected from genetic data. It also predicts minor blood groups (like Vel, and other 30 systems) which are normally ignored in the medical practice during transfusions. The tool is freely available for the community (url: `protein.bio.unipd.it/download`), and can be parametrized to predict additional phenotypes.
Last but not least, I contributed in the development of a Petri network [48] for the Von Hippel-Lindau (VHL) cancer pathway, which is useful to simulate the effect of gene knockout and the resulting effects in terms of changes in the substrate concentration.

Bioinformatic infrastructure and technologies

As pointed out in the early pages of this manuscript, IT infrastructure and technologies are central for bioinformatics research. Fresh data is stored in public databases and open source libraries for of biological data analysis can speed up the data processing. With *Repeats DB* (url: `repeatsdb.bio.unipd.it`), I contributed in the development of a tandem repeat protein annotations database. Hundreds of repeated proteins were analyzed and classified semi-automatically to highlight potentially new domains. Almost 20 protein repeat categories were employed with additional manually curated annotation for significant cases.
To simplify the management of PDBs, the C++ Victor library (Virtual Construction Toolkit for Proteins) was also published and is currently maintained (url: `protein.bio.unipd.it/victor`). It provides advanced functions for

sequence alignment, PDB manipulation and statistical estimation of physical energy. This library contributed in nearly half works presented in this thesis, thus it can be considered a basic software implementation for bioinformatics data management.

Chapter 2

Protein Function Prediction

Protein function prediction from sequence using the Gene Ontology (GO) classification is useful in many biological problems. Its has recently attracted increasing interest, due to the overwhelming amount of unannotated sequences in public databases and to the CAFA (Critical Assessment of Function Annotation) [39] challenge.

In this thesis I introduce two computational methods for GO terms prediction. GAS (Guilty by Association on STRING) is a first tool designed to predict protein function exploiting protein-protein interaction networks without sequence similarity. The assumption is that whenever a protein interacts with other proteins, it is part of the same biological process and located in the same cellular compartment. GAS retrieves interaction partners of a query protein from the STRING database [6] and measures enrichment of the associated functional annotations to generate a sorted list of putative functions. A performance evaluation based on CAFA metrics and a fair comparison with optimized BLAST [27] similarity searches are provided. The consensus of GAS and BLAST (GAS-C) is proven to improve overall performance. The PPI approach is shown to outperform similarity searches for biological process and cellular compartment GO predictions. Moreover, an analysis of the best practices to exploit protein-protein interaction networks is also provided. GAS description and its details are reported in Section 2.1.

A second tool called INGA extends GAS-C. It exploits information encoded in PFAM domains [49] to better prioritize sequence patterns. The method was used to predict GO terms and their likelihood for more than 100,000 sequences in CAFA 2014 edition, and proved to be one of the most effective strategies among over 50 predictors. All details about INGA are shown in section 2.2.

2.1 Protein Function Prediction Using Guilty by Association from Interaction Networks

The large amount of available protein sequences requires usage of in silico methods for automatic large-scale function prediction. The annotation process, assigning functions to target proteins, is generally based on the transfer by homology principle [50]. Protein space can be partitioned in subsets (families) that groups proteins with a common ancestor and, possibly, the same function. Whenever evolutionary relationships between two different proteins are available, all features from one protein are transferred to the other. Sequence comparison is used to infer homology and collect evidence about membership in a given family. However it requires to properly choose similarity measures and related cut-off values in order to avoid false positives (and, conversely, false negatives). As each family has its own story and is the result of different and complex evolutionary phenomena, available data are usually not sufficient to trace an unambiguous phylogenetic tree [51]. Any time two sequences appear to greatly diverge, it becomes impossible to find annotated homologs. On the other hand, the same protein can perform different functions when placed in a different organism, and sequence information alone cannot distinguish such situations. Within the CAFA (*Critical Assessment of protein Function Annotation*) experiment [39], it has been stated that the currently best methods to predict protein function rely on sequence similarity searches for conserved regions or homologous proteins [52, 53]. Moreover, it has been recommended to extend standard homology search with new methods that use different sources of information on protein function [54, 55]. The CAFA experiment also provided standard criteria for the evaluation of the predictions, e.g. the dataset used for the blind test and the definition of function space through Gene Ontology (GO) terms [38]. The scoring metrics for comparing function predictions in CAFA are mainly based on precision-recall curves.

New effective experimental techniques to find genome-wide interactions make protein-protein interaction data widely available and ready to be used for functional annotation [56, 57, 58]. Approaches exploiting interaction networks have been widely used for annotation of the Yeast genome [59, 60, 61, 62, 63]. At the same time, many tools which analyze biological network properties are already available. Some of them use interaction networks to prioritize genes that are part of disease pathways. These applications use enriched functional terms to describe clusters of interacting proteins or genes. The STRING interaction database [6] itself provides tools to compute GO term enrichment in selected sub-networks. To the best of our knowledge, functional enrichment in protein-protein interaction networks has never been used effectively as a tool for predicting function of unknown proteins.

For example, the χ^2 test has been used to rank the functional terms associated to a group of interacting partners by comparing the frequency of the terms within the group and with the expected distribution in the whole network [62]. Another work, PRODISTIN [59], focuses on the clusterization of the entire Yeast interaction graph by means of a distance measure to define groups associated with the same functional class. A Bayesian approach [61] has been applied to calculate the posterior probability that a given protein has the function of interest. This method takes into account the prior probability of the entire network but it does not consider the dependencies among terms. Another method, FunctionalFlow [63], treats annotated nodes as “sources” and propagates the associated annotation through the connecting edges following some simple rules. These rules take into account the distance between two nodes and the number of alternative paths connecting them to produce a score.

All of these methods are based on a single model organism and cannot easily be compared with other state-of-the-art methods like those participating in CAFA. Moreover, they used a very small ad hoc ontology for Yeast which is two orders of magnitude smaller than the full GO. It is also difficult to evaluate their impact on the coverage of genome annotation, as the number of interactions available today is not comparable with networks available a few years ago.

In this section we introduce GAS (Guilty by Association on STRING) to predict protein function exploiting protein-protein interaction networks without sequence similarity measures. We also provide an analysis on the implementation details and parameters necessary to maximize accuracy as well as important considerations about best practices to exploit protein-protein interaction networks.

2.1.1 Methods

GAS

Protein-protein interaction (PPI) networks provide relevant information about protein function. The aim GAS is to exploit the annotation of the neighborhood of a protein to transfer the function. The choice of the network, the definition of the set of interacting partners, the strategy to transfer annotation and the method to build the consensus represent key factors to improve accuracy and implement an effective prediction tool. The idea at the basis of GAS arises from the analogy with the "Guilty-by-association" principle. This concept asserts that qualities of one object are inherently qualities of another, merely by an independent association. In our case it means that if a protein physically interacts (*association*) with other proteins it should share a similar function (*quality*). Proteins in a living cell have many physical interactors, each group of interacting proteins is expected to participate in the same biological process and to operate in the same subcellular compartment. Given a protein with unknown function, GAS uses the STRING [6] network to collect the set N of directly interacting nodes. All experimental GO terms are then associated to the annotated proteins retrieved from SwissProt [64] and ranked by a measure representing their specificity in the collected set. We estimated this specificity, by measuring enrichment with respect to the entire training set (i.e. the remaining STRING nodes). The P-value associated to the enrichment is computed according to Fisher's exact test, which represents the probability that a specific term, GO_I, is associated to a given set by chance (null hypothesis). For each collected term GO_I, the contingency table is shown in Table 2.1. The P-value is generated with the following standard formula:

$$P - value(GO_i) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

GAS was evaluated on two different protein interaction networks. In the first case we focused on highly confident STRING interactions (edge score $\leq$ 900). In the second one, we selected all STRING interactions with edge score $\leq$ 500. All nodes in the STRING network were mapped to the UniRef90 database to extend the number of interacting nodes and increase the

		Categories	
		GO_i	*Not* GO_i
Node Set	*Cluster*	a	b
	DB	c	d

Table 2.1: Contingency table. *Cluster* indicates the set N of directly interacting nodes and *DB* represents the rest of nodes in STRING associated to experimental GO terms.

chance of collecting experimental GO terms as well as to make GAS comparable with our version of BLAST (see GAS-C).

GAS-C

GAS-C, where "C" indicates Consensus, is an extended version of the algorithm that merges GAS with BLAST predictions. For each input protein, GAS-C first computes GAS and BLAST predictions independently and then combines them. BLAST hits are retrieved running the program with default parameters discarding hits with e-value higher than 10^{-3} sorted by bit-score (default output), since it maximizes performances [39]. The presence of large groups of homologous proteins with high sequence similarity in the sequence database may affect a BLAST prediction. The UniRef90 [65] database was used to address the redundancy issue. For each hit corresponding to the representative sequence of a UniRef90 cluster, all experimental GO terms associated to all cluster members are transferred. This strategy increases sensitivity, allowing to retrieve hits with lower sequence identity but possibly richer annotations. To make GAS comparable with BLAST, we mapped UniRef90 clusters to the interacting nodes and transferred functional annotation from all members belonging to these clusters. Since the F-score was found to be poorly correlated with the native output score (Bit-score and P-value), the F-score computed on the rank position was considered instead. For BLAST, the rank corresponds to the hit position in the output list, e.g. at rank 1 we find GO terms (plus ancestors) transferred from the first hit, the one with best Bit-score. For GAS, the rank is given by the P-value, e.g. at rank 1 we find terms (plus ancestors) with the lowest P-value. The values f_{tr}, converted to the F-measure of the r-th (or higher) ranked terms for target t. From these data, the expected rank-dependent performance was evaluated through an exponential curve emphasizing the correlation between the ranking r of the predicted term and the F-measure:

$$E[F(r)] = e^{a+b\cdot r} + c$$

where a, b and c were estimated through a nonlinear least square on the predictions and corresponding F-measures. Next, the corresponding rank-dependent score S_I and S_B were assigned to each GO term predicted by GAS and BLAST respectively. Whenever the same GO_I was predicted by both approaches, its score was updated as follows:

$$S_{combined} = 1 - (1 - S_B(GO_i)) \cdot (1 - S_I(GO_i))$$

Finally, all scores are propagated to the root of the ontology guaranteeing that each ancestor node always inherits the maximum probability from its children.

Training and test sets

The evaluation set for the prediction models is made of protein sequences with experimental annotation from SwissProt. It includes previously unannotated proteins that accumulated GO terms annotation in one year, accounting for 8,976 proteins from 283 organisms. 4,432, 3,931 and 3,194 sequences were counted for the MF, BP and CC sub-ontologies respectively. It was obtained as the difference between the SwissProt releases v2012_07 and v2013_07, applying a filtering criterion for automatically predicted terms. Experimental ("trusted") annotation were considered as those terms which are associated to the evidence codes EXP, IDA, IMP, IGI, IEP, TAS and IC. Figure 2.1 shows that the organism distribution of new annotated sequences differ strongly. 1,940 new sequences (22% of the entire test set) come from "other" organisms. The training set was obtained by randomly sampling 10,000 targets from the experimentally annotated sequences in SwissProt v2012_07.

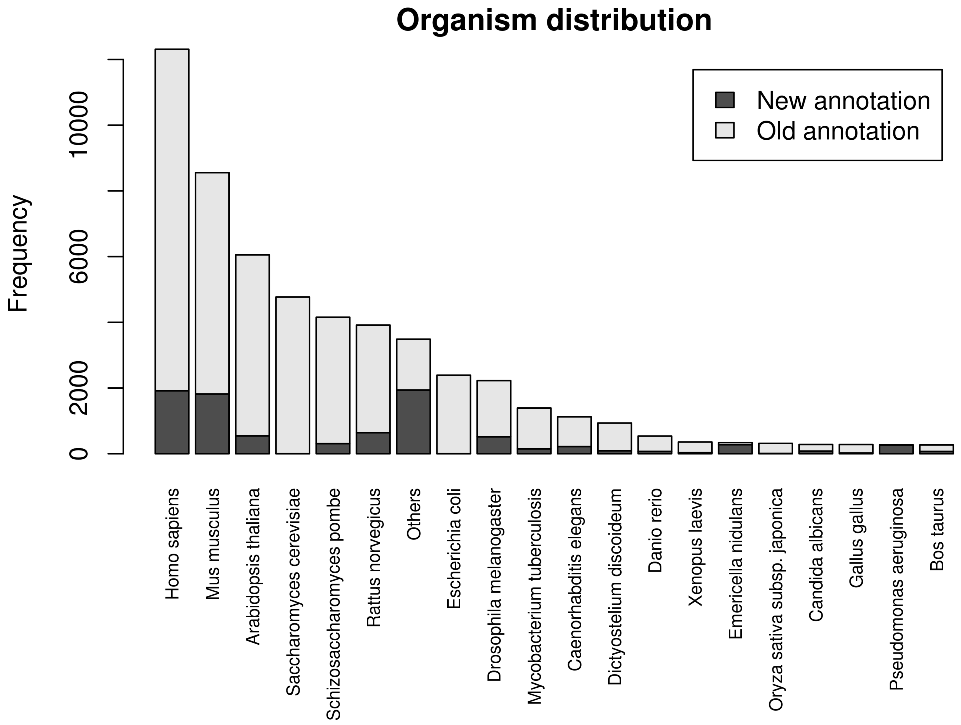


Figure 2.1: Distribution of SwissProt entries annotated experimental GO terms and categorized by organism. Dark bars ("New annotation") represent the number of sequences that accumulated experimental annotation in one year and that were used as test set.

Performance evaluation

Two different strategies to evaluate GAS and GAS-C models, one based on a *target-by-target* comparison and the other based on the *whole dataset* were adopted. In the former approach, for each target protein, predicted GO terms by their ranking position were evaluated (as described in the GAS-C paragraph) and then the mean on the entire test set for all possible ranking r computed. In the *whole dataset* strategy all targets were considered together and performance was calculated for all possible score thresholds t. The scores in this case correspond to the P-value, Bit-score and $S_{combined}$ respectively for GAS, BLAST and GAS-C predictors. For FANN-GO we used the score as it is provided by the tool and for the Naïve method the frequency in the SwissProt database. We used the following well established measures adopted in CAFA to evaluate performance:

$$precision(r) = \frac{|GO_t \cap GO_p(r)|}{|GO_p(r)|}$$
$$recall(r) = \frac{|GO_t \cap GO_p(r)|}{|GO_t|}$$

where GO_t represents the set of true terms associated to a protein in the test set, while GO_p is the set of predicted terms. Precision and recall are measures of correctness and completeness for a method respectively. They both depend on r, which corresponds to the *ranking* in the *target-by-target* approach and to the score threshold in the *whole dataset* strategy. A third useful metric is the F-measure, which is obtained by calculating the harmonic mean of precision and recall:

$$F(r) = 2 \cdot \frac{precision(r) \cdot recall(r)}{precision(r) + recall(r)}$$

2.1.2 Results

We introduce GAS (Guilty by Association on STRING), a tool to predict protein function exploiting protein-protein interaction networks without sequence similarity measures. The assumption is that whenever a protein interacts with other proteins, it is part of the same biological process and located in the same cellular compartment. Two proteins exhibiting the same interaction partners can reasonably be inferred to have the same function. Given the sequence of an unknown target protein, GAS is able to retrieve its interacting partners from the STRING network and measures the enrichment of the associated functional annotations to generate a sorted list of putative functions. In the following, we will present some experiments that explain how protein interaction networks can contribute to solve the problem of protein function prediction. We will start with an analysis of the STRING network and then we will provide a comparison with some methods. The list of evaluated tools includes BLAST [27], which is known as the standard baseline tool for function prediction based on homology inference, the Naïve method implemented as described in CAFA and FANN-GO that was the only one available as standalone software, and trained on an old dataset. Finally, we show that the GAS-C consensus model can increase performance accuracy using both GAS and BLAST predictions.

Experimental annotation in STRING

We implemented GAS using STRING as the reference interaction network. STRING is the largest database of protein-protein interactions including experimental derived data, third party information coming from other databases and predicted interactions [6]. However, GAS does not use the entire network but only a portion composed by only those nodes that can be mapped to SwissProt entries and annotated with experimental terms. Exploitation of functional information coming from protein protein interaction networks requires minimization of false positive interactions. STRING provides a score representing an edge quality estimate that also tracks the information source. Figure 2.2 shows the distribution of STRING edge scores for different interaction types coming from different sources.

Most of the STRING edges connecting SwissProt entries have low quality values and come from text-mining and co-expression data, 53% and 40% of the total interactions respectively, while only 5.7% are confirmed experimentally (Table 2.2). When multiple sources of information support the existence of an interaction, they result in a higher global score. We evaluated GAS performances by filtering the STRING network for different edge confidence values. An edge cut off of 900 on one hand guarantees the selection of reliable protein interactions, often confirmed in third party databases, but also reduces the amount of available interacting partners and therefore the annotation that can be transferred. On the contrary, a relaxed threshold yields a higher chance of collecting experimental GO terms useful for the prediction. One of the major limitations of function prediction from interactome data is coverage (see Fig. 2.3). In fact, when no restriction in terms of alignment coverage and identity is applied, BLAST is capable of generating new GO terms in almost the totality of targets. For GAS, we are able to find experimentally annotated interacting nodes for our target protein in 29% to 47% of the cases, depending on the ontology. We tested GAS performance by filtering edges for different cut off values. All tables and figures in the chapter refer to the GAS predictions coming from a high confidence

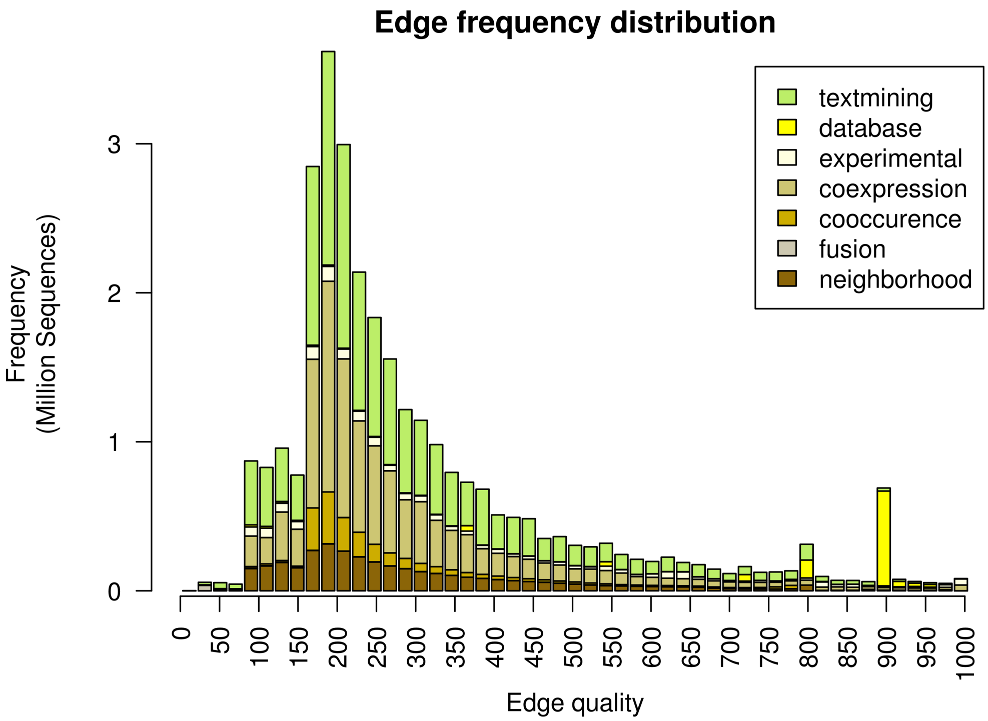


Figure 2.2: Distribution of STRING edge types by edge quality. Frequencies are calculated using entries with experimental GO terms in SwissProt. Text mining and coexpression edges are the most common among the low qualities, while interactions from database are the most reliable.

STRING sub network with a cut off of 900. In Figure 2.3 we reported the same comparison relaxing the edge filtering at a cut off of 500. The ability to predict new potential functions increases greatly, ranging between 57% to 68% coverage. Moreover filtering out low confidence edges significantly decreases false positive interactions resulting in a slightly greater accuracy (data not shown).

Target-by-target performance

We compared GAS, GAS-C and the other tools by evaluating their performance on the test set using the same approach adopted by the CAFA assessors. To clarify the evaluation procedure, in Figure 2.4 we provides two

Edge Type	Edges (%)
Text mining	53.3
Co-expression	40.1
Neighborhood	14.3
Co-occurrence	7.4
Experimental	5.7
Database	4.3
Fusion	0.3

Table 2.2: Edge types distribution of experimentally annotated proteins in STRING.

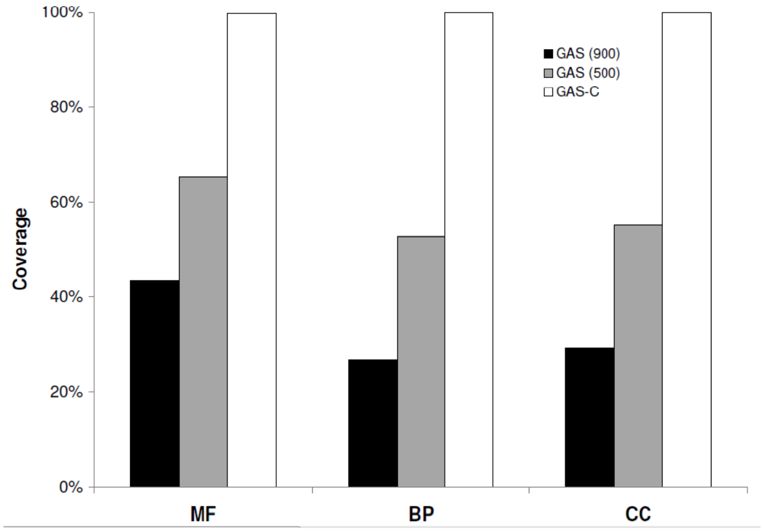


Figure 2.3: Prediction coverage on the dataset for GAS at STRING edge weight cutoffs 900 and 500, as well as GAS-C.

examples of GAS and BLAST predictions. On the left the GAS enrichment with the target connected with its direct interactors in STRING (grouped in the blue box). Red and gray circles nearby interactors represent GO terms and their size is correlated with the corresponding P-value. On the right is represented the BLAST output, the green box represents the best experimentally annotated hit. Precision recall measures are calculated by comparing the truth graph (red circles) with the predicted graphs (GAS predictions in blue and BLAST in green).

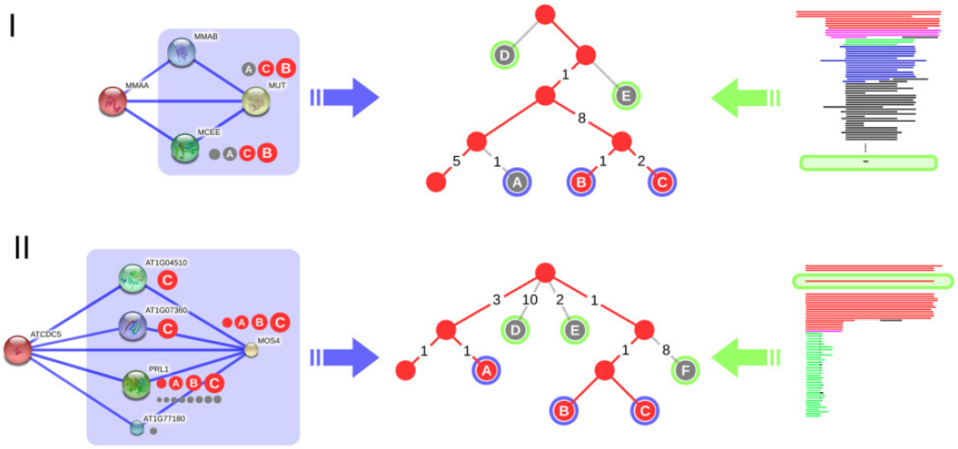
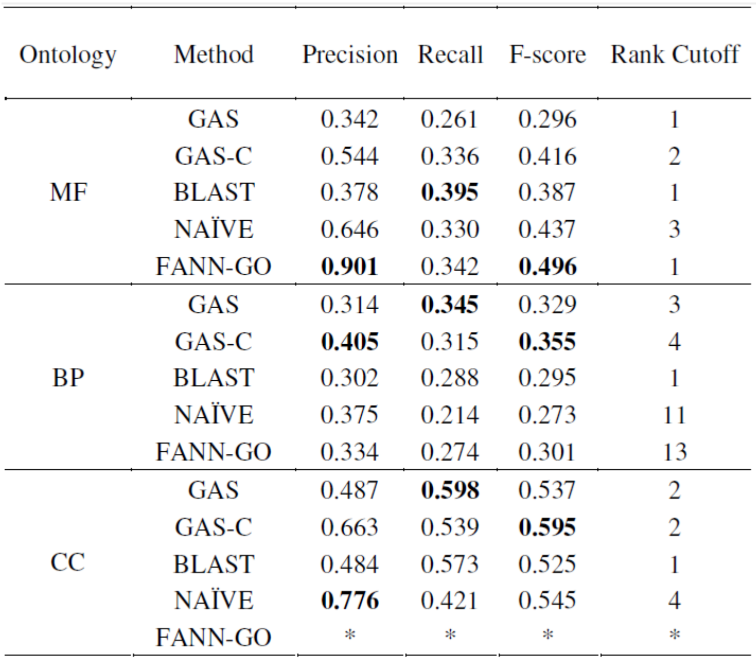


Figure 2.4: GAS prediction on the left, BLAST on the right and the comparison with the real target annotation in the middle (GAS predictions in blue and BLAST in green).

Table 2.3 reports the *target-by-target* maximum F-score (see Section 2.1.1), computed on the test set targets where all listed methods are able to make a prediction. The first consideration is that different methods behave differently for the three ontologies. Protein protein interaction networks contain useful information about the biological process (BP) and the cellular compartment (CC) of a target. GAS does not produce good results for molecular function (MF). This is not surprising, since interacting proteins, even if they participate in the same biological process, usually carry out different biochemical reactions. For example two proteins may be involved in the regulation of the cell cycle, but the first can be a regulatory protein performing phosphorylation and the second a transcription factor with a completely different biochemical attitude.

The second observation is about the difference in terms of performance observed for the three ontologies in general. The BP terms are definitely the hardest to predict due to the more complex structure of the sub ontology. Focusing the attention to the BP ontology is possible to observe the effectiveness of combining GAS and BLAST in the GAS-C consensus, that



Ontology	Method	Precision	Recall	F-score	Rank Cutoff
MF	GAS	0.342	0.261	0.296	1
	GAS-C	0.544	0.336	0.416	2
	BLAST	0.378	**0.395**	0.387	1
	NAÏVE	0.646	0.330	0.437	3
	FANN-GO	**0.901**	0.342	**0.496**	1
BP	GAS	0.314	**0.345**	0.329	3
	GAS-C	**0.405**	0.315	**0.355**	4
	BLAST	0.302	0.288	0.295	1
	NAÏVE	0.375	0.214	0.273	11
	FANN-GO	0.334	0.274	0.301	13
CC	GAS	0.487	**0.598**	0.537	2
	GAS-C	0.663	0.539	**0.595**	2
	BLAST	0.484	0.573	0.525	1
	NAÏVE	**0.776**	0.421	0.545	4
	FANN-GO	*	*	*	*

Table 2.3: Target-by-target performance. Performances are computed for entries where methods can make a prediction (Targets). The maximum F-score is used to select the corresponding precision and recall. The cutoff explain the number of top rank scores that should be considered to achieve the best f-score. The best performance for each ontology is highlighted in bold. * FANN-GO does not predict cellular component.

rewards those terms predicted by both methods (see Section 2.1.1). GAS-C obtains the maximum F-score over all methods even if the recall is penalized compared to GAS itself. To better appreciate the predictor performance we plotted the precision-recall curves for all methods (Figure 2.5). Another important observation is the good performance obtained by the Naïve method for the MF terms in Table 2.3. This behavior was already observed during the CAFA experiment and is due to the very high frequency of proteins annotated with some shallow leaf terms very close to the root of the ontology. Naïve reaches a very high accuracy since it always predicts two ancestors of these leaf terms in the first positions ("protein binding" and "catalytic activity"). FANN-GO is subjected to the same phenomenon but it achieve better results since the machine learning approach overcomes Naïve limitations.

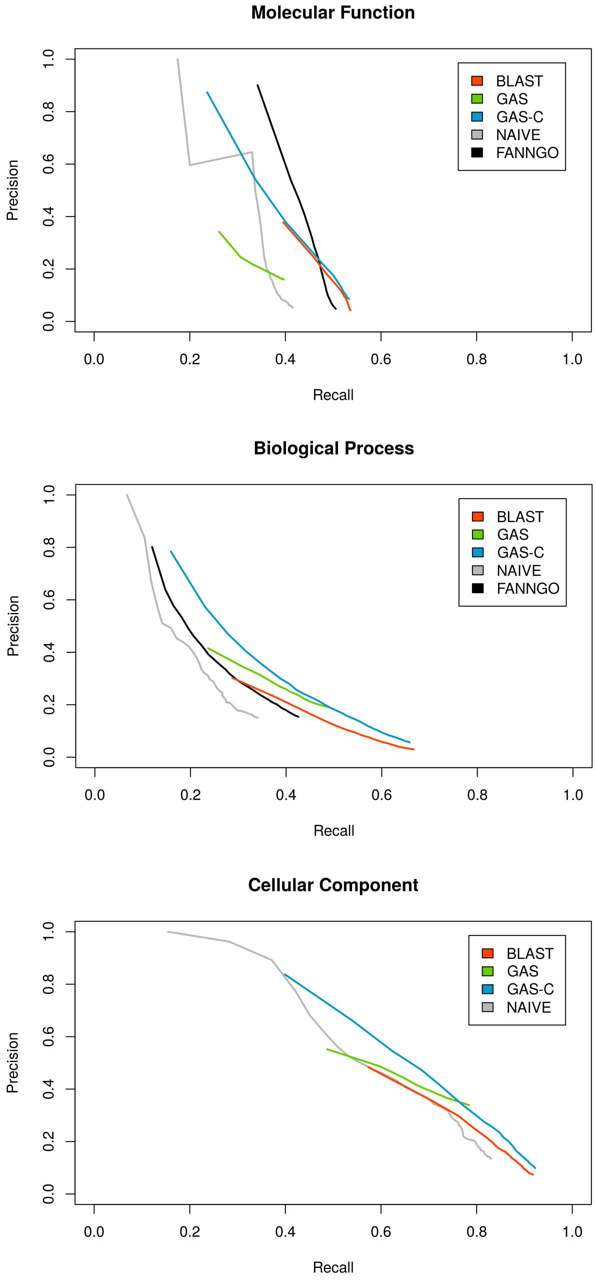


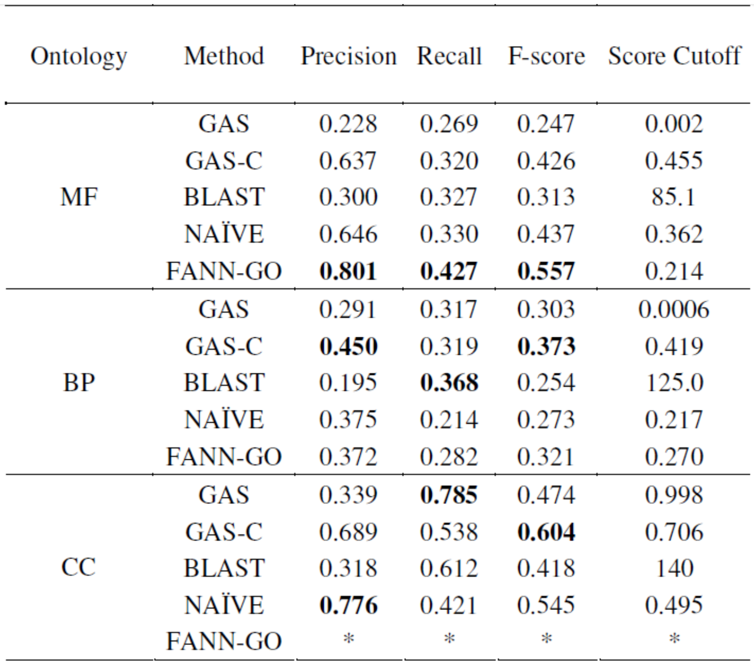
Figure 2.5: Precision-recall curves. FANN-GO is missing in the cellular component chart because the tool does not provide prediction for that ontology.

Whole dataset performance

One important aspect about the different predictors can be highlighted by a correlation analysis. For BLAST, we observed a limited relationship between the Bit-score and the F-measure for each target, with values below 0.290 for the three ontologies. Surprisingly, the same low correlation is also observed when considering sequence identity (not shown), suggesting that is very difficult to find a specific identity threshold useful for discriminating a good source of annotation. For GAS the same result holds for the enrichment P-value (correlation below 0.225). Such a limited correlation between F-score and the predictor confidence score suggested the use of ranks to improve results. We observed that BLAST generally achieved best results by just picking the GO terms associated to the first hit, i.e. sequence with the highest Bit-score. GAS ranks GO terms rather than sequences and has to consider up to the first three predicted terms, depending on the sub-ontology, to achieve optimal performance (see Table 2.3, column Rank Cutoff). This is likely to be the reason why GAS shows a higher maximum precision in general, while BLAST has a higher maximum recall (Table 2.3). Interestingly, the GAS-C score is strongly correlated with the expected F-score (correlation higher than 0,406), and outperforms the rank-based strategy as shown in Table 2.4. This is likely to be a consequence of a good fitting procedure. The increased predictive power shows that GAS and BLAST generate different knowledge. The consensus enables a better prioritization of predicted terms by using two orthogonal sources of information jointly and can truly guide a user to select GO terms depending on the expected annotation quality.

2.1.3 Discussion

In this chapter we presented a novel strategy to predict protein function exploiting protein-protein interaction (PPI) networks, developing a statistical significance estimation to rank GO terms. To the best of our knowledge, this is the first attempt to fairly evaluate the contribution of network interaction data to predict protein function. GAS is based on the "Guilty-by-association" principle applied in the contest of PPI networks. If a protein physically interacts with other proteins it should share a similar function. For example, when all interacting partners operate inside the nucleus, it is reasonable to believe that the subcellular localization of a given target will be the nucleus itself.



Ontology	Method	Precision	Recall	F-score	Score Cutoff
MF	GAS	0.228	0.269	0.247	0.002
	GAS-C	0.637	0.320	0.426	0.455
	BLAST	0.300	0.327	0.313	85.1
	NAÏVE	0.646	0.330	0.437	0.362
	FANN-GO	**0.801**	**0.427**	**0.557**	0.214
BP	GAS	0.291	0.317	0.303	0.0006
	GAS-C	**0.450**	0.319	**0.373**	0.419
	BLAST	0.195	**0.368**	0.254	125.0
	NAÏVE	0.375	0.214	0.273	0.217
	FANN-GO	0.372	0.282	0.321	0.270
CC	GAS	0.339	**0.785**	0.474	0.998
	GAS-C	0.689	0.538	**0.604**	0.706
	BLAST	0.318	0.612	0.418	140
	NAÏVE	**0.776**	0.421	0.545	0.495
	FANN-GO	*	*	*	*

Table 2.4: Whole dataset performance. The performance is calculated over the same target set of Table 2.3 but for all possible thresholds for the score provided by the tools themselves. Score Cutoff indicates the score threshold where the tool get the best F-score. For every tool the scores are: GAS = P-value; GAS-C = Tool score; BLAST = Bit-score; NAÏVE = Frequency; FANN-GO = Tool score. The best performance for each ontology is highlighted in bold. * FANN-GO does not predict cellular component.

However, even if the principle is very simple, some details need to be considered to implement an effective tool. Some aspects are related to the PPI network and others to the scoring function applied for ranking the predicted terms. The size of the PPI network and the reliability of the interactions affect the prediction in two different ways. A big network increases the probability of finding interacting partners endowed with GO annotation while filtering low quality interactions corresponds to a gain in the precision of the prediction. The other key factor is the method used to sort and prioritize the transferred GO terms. We found that the P-value generated by measuring the enrichment of each collected annotation can be conveniently used to sort terms but there is not a linear relationship between the P-value and the

F-score that measures the quality of a prediction. In other words it means that is not possible to say which could be an optimal P-value threshold that guarantees a good annotation. This is also true for BLAST where the Bit-score provided by the tool correlates very poorly with the F-score (Table 2.4). Conversely both the Bit-score and the P-value provide a good sorting of GO terms and we found a good correlation between the F-score and the position (ranking) in the output list (compare GAS and BLAST F-score in Table 2.3 and Table 2.4). The comparison between GAS and BLAST highlighted important differences among the three GO sub ontologies. As expected, PPI data is very effective for the CC and BP cases. On the other hand, evolutionary inference from sequence similarity represents a better discriminative approach for MF. This fact is consistent with the idea that network prediction can infer knowledge from the local neighborhood. Conversely, the molecular function cannot be directly inferred from the interactome, since the interacting proteins participating in a given biological process contribute themselves with different specific activities and biochemical reactions.

As shown in the first CAFA experiment the performance of consensus methods is generally higher than standard tools. We implemented GAS-C that is able to generate a consensus prediction by combining both BLAST and GAS results. The implemented consensus strategy is extremely fast and simple, consisting in a score transformation, which can be generated in linear time with respect to the number of predictions. GAS-C achieves better results for all the three ontologies compared to BLAST and GAS themselves (Table 2.3 and Table 2.4). It also outperform FANN-GO for the BP ontology. A particular discussion has to be done for the the Naïve performance in the MF ontology. The good F-score was already observed during the CAFA experiment and is due to a bias in SwissProt of some shallow leaf terms very close to the root of the ontology (see results).

In general all presented results in terms of F-score, precision and recall are slightly underestimated compared to the numbers provided by the first CAFA experiment. This happened because we evaluated all the predictions without filtering those terms in the test set not yet available one year before in 2012. However that does not affect the validity of this work since all methods were affected equally by this problem. At the moment, sequence similarity approaches outperform GAS in terms of target coverage, but we believe that good quality interaction data is going to increase consistently, resulting in a better capacity to generate new hypotheses. Moreover, PPI networks represent a complementary source of knowledge compared to evolutionary information, and will be even more effective in the future, when entire organism interactomes will become available.

2.2 The Critical Assessment of Function Annotation[1]

A newly sequenced genome contains several thousands genes, but a thorough functional characterization is unfeasible due to high experimental costs. The Critical Assessment of Function Annotation (CAFA) is a community experiment where tools designed for protein function prediction are benchmarked. In 2014 edition, more than 50 submissions were compared in the same test set by the CAFA assessors, providing an unbiased overview of the state-of-the-art in this field. The challenge starts with the publication of a large sequence dataset (more than 100,000 in 2014), where few of them are completely uncharacterized (*scenario 1*), while others have existing GO annotation (*scenario 2*). This latter situation is interesting for proteins with partial characterization, but requiring further efforts for a comprehensive understanding of their activity. When targets get published by the organizers, research groups can submit within a deadline the predicted GO terms for all sequences and a confidence score. After six months, the new annotations gathered in databases like UniProt [25] enable the assessment for few sequences. In other words, the ground truth is not known at the beginning of the challenge, but it is obtained from the natural growth of public databases. As a reference baseline, BLAST [27] is used to show the performance of a well-established algorithm. In addition, a ranking scheme called *Naive* assigned to all GO terms a score based on their prior probability. Therefore, common terms are given high score for all CAFA targets, while rare GO terms are associated with a lower score, proportional to the probability distribution in public databases. This represents the background "random" distribution of terms. To evaluate submissions performance, *precision*, *recall*, *F-score* (see Section 1.2.1) and *Semantic Distance* (S) [66] are used. All quality measures require to take into account the prediction score when computing aggregated results. Given a threshold t, one can compute recall (or other measures) in whole CAFA dataset (*mode 1*), or just on the subset of proteins where a

[1] Figures and results are taken from CAFA online presentations (URL: biofunctionprediction.org).

submitter gave a prediction score $\geq t$ (*Mode 2*). In other words, mode 1 labels as misclassified targets with no high quality predictions. This penalizes methods unable to predict confidently new terms for all proteins.

2.2.1 Methods

We decided to participate with an extended version of GAS (see Section 2.1) called INGA, where we combined BLAST homology search [27], knowledge of PFAM domains [49] and STRING [6] protein-protein interaction network to build an effective information retrieval (IR) method. The key addition is PFAM domains, which are sequence regions expected to have a functional role especially in terms of GO Molecular Function [38]. Likewise in GAS, the statistical association of a GO term with respect to a given PFAM sequence was estimated using Fisher's exact test in UniProt database. We also implemented two versions of STRING network search based on high (≥ 900) and medium (≥ 700) edge quality. The two models were designed to have different capabilities in terms of *accuracy* and *coverage*. High coverage is very important for computational methods, since the ability to predict GO terms for all targets make them good baseline approaches. With the same rationale, we used BLAST search to detect homologue sequences to CAFA targets using a strict cut-offs in terms of minimal identity similarity ($\geq 40\%$) and coverage ($\geq 80\%$). The two versions of BLAST search used respectively UniProt and UniRef databases. UniRef is the non-redundant version of UniProt, which avoids the detection of nearly identical sequences and solves statistical issue related to very large sample size at the cost of lower coverage. The accurate version of network-based and BLAST search used just the most reliable GO terms (with evidence codes EXP, IDA, IMP, IGI, IEP, TAS and IC), while the implementation focused on coverage exploited any type of annotation.
The combination of these strategies is based on *generalized additive models*, similarly to the idea introduced by GAS. Overall, we achieved 95% coverage, showing that we managed to predict GO terms for the largest part of CAFA targets.

2.2.2 Results

An early assessment of CAFA predictions was presented during the Automated Function Prediction meeting in July 2014. After six months from the submissions, 232 proteins obtained annotation for MF, 410 for BP and 608 for CC (*scenario* 1). As can be seen in Fig. 2.6, INGA (labeled as Tosatto) performed very well in terms of precision-recall curve, with a maximum F-score of 0.667. Sequence information is very significant for MF, suggesting that PFAM domains probably played a central role for such result. Just a single method did slightly better than our consensus, while baseline strategies (BLAST and Naive) were largely outperformed. It is interesting to note that BLAST achieved higher accuracy when used in UniProt 2011 database rather than in 2014 version. This is the result of a recent data curation process in UniProt, which increased the quality of information at the cost of removing available entries. When MF performance are compared by *semantic distance*, there is a mild rearrangement of methods that penalizes our approach (Fig. 2.7). However, our information retrieval system proved once more to be one of the best methods on unseen data.

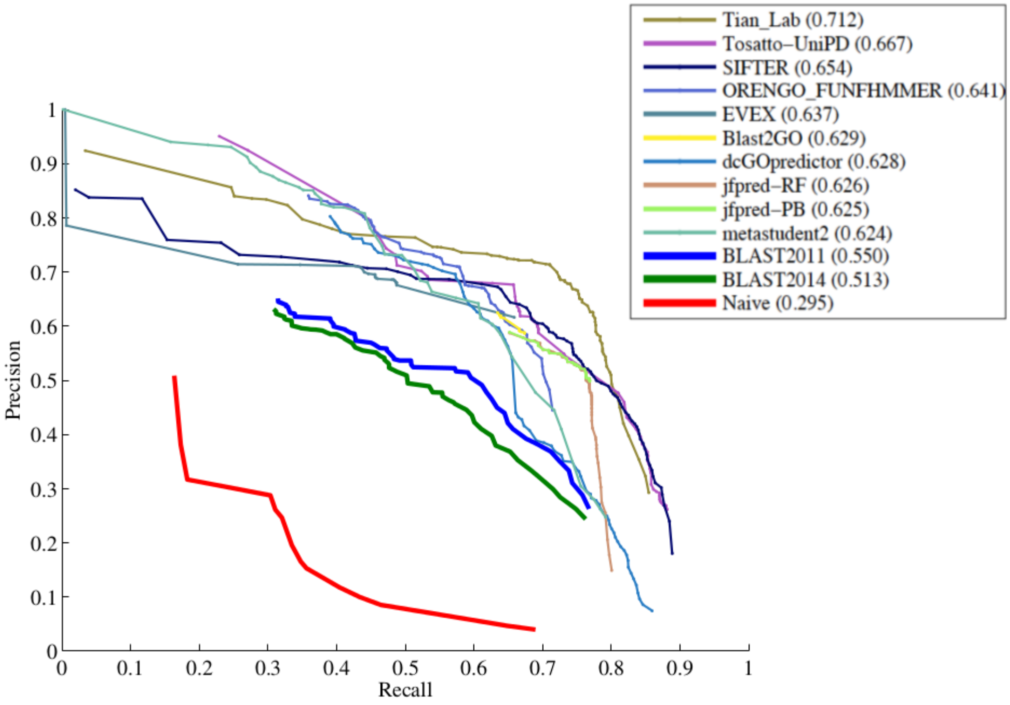


Figure 2.6: Precision-recall curve for the top ten methods in the MF ontology.

Semantic distance performance of INGA is very high also for BP and CC 2.8. It is important to note that there are limited S value changes between

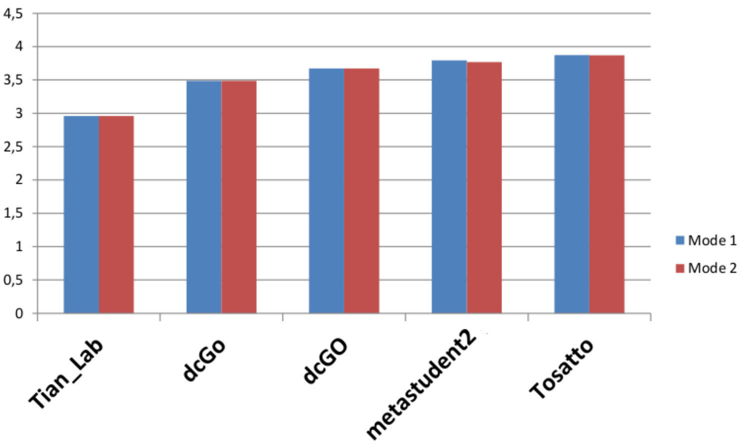


Figure 2.7: Semantic distance for the top five methods in the MF ontology.).

mode 1 and 2, proving that there are few targets with no high quality prediction. On the other hand, there are better methods when evaluating the targets by means of F-score. Surprisingly, none did well with this measure in CC: Naive achieved among the highest results, showing that a nearly random prediction is comparable to state-of-the-art methods.

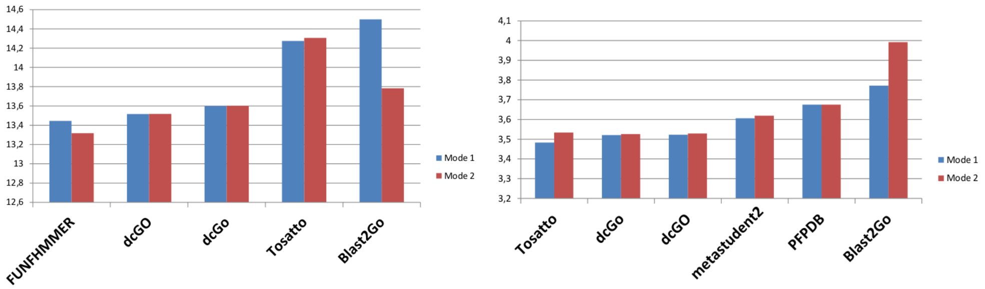


Figure 2.8: Semantic distance for the top five methods in the BP and CC ontology.

In the case of *scenario* 2, there are better methods as well. This was expected, since we did not exploited GO covariance to perform better predictions. Our submission is focused on the common case of newly sequenced genomes where no annotation is available at all. For the future, the exploitation of priorly available GO terms in a given target will surely increase performance.

To conclude, CAFA is a very challenging experiment, where a very large volume of data has to be managed. Targets represented very often the entire genome of model organisms, which ranged from humans to plants. The evaluation task of submissions is clearly complex due to frequent updates of the Gene Ontology that might remove and create new terms. The graph-like structure of the ontology also poses big problems, which motivated the introduction of S performance measure. Despite these issue, CAFA succeed in the goal of creating a community focused in protein function prediction. Results comparison with the previous edition proves an increased accuracy for all methods, showing that there is still room for improvement and a strong commitment to solve this scientific challenge.

Chapter 3

Protein Structures: Dynamics and Repeats

Tree-dimensional structures are among the most useful source of information to study proteins. As pointed out in section 1.3.2, protein sequence and structure representation and manipulation requires dedicated software libraries to support methods of increasing complexity. In section 3.1 I introduce the Victor (VIrtual Constrution TOol for pRoteins) C++ library, an open source platform dedicated to enabling inexperienced users to develop advanced tools and gathering contributions from the community. The provided application examples cover statistical energy potentials, profile-profile sequence alignments and ab initio loop modeling. Victor was used over the last fifteen years in several publications and optimized for efficiency. It is provided as a GitHub repository (URL: `protein.bio.unipd.it/victor/`) with source files and unit tests, plus extensive on-line documentation, including a Wiki with help files and tutorials, examples and Doxygen documentation. Victor was also used as a starting point to design a method for the analyses and interpretation of Molecular Dynamics (MD) simulations. MD simulations in fact have gained increasing relevance over the last years, but the complexity related to their study is still one of the major challenges for most users. With RING MD, I propose a new approach to identify the most important frames (PDB structures) and key residues that cause different conformers to be observed, providing a simple interpretation useful for non-expert users. Section 3.2 shows implementation details and compare RING MD results with the classical analysis of a MD simulation.

In addition to new methods and libraries for protein structure manipulation and analysis, with RepeatsDB I contributed in the development of a database of annotated tandem repeat protein structures. Tandem repeats pose a difficult problem for the analysis of protein structures, as the underlying sequence can be highly degenerate. Using state-of-the-art repeat detection methods and manual curation, the Protein Data Bank was systematically annotated, predicting 10,745 repeat structures. In all, 2797 structures were classified according to a recently proposed classification schema, which was expanded to accommodate new findings. In addition, detailed manual annotations were performed in a subset of 321 proteins. RepeatsDB is an ongoing effort to systematically classify and annotate structural protein repeats in a consistent way. In section 3.3, RepeatsDB database and its implementation are presented. To conclude, this chapter describes few answers to the most typical needs in bioinformatics - tools and technologies for structural analysis. All implementations can be downloaded and used freely, thus contributing directly in the sharing of open applications for the community.

3.1 The Victor C++ library for Protein Structure Management[1]

Structural bioinformatics methods require valid software libraries to represent and manipulate proteins efficiently. A number of widely used tools have been developed over the years to visualize proteins, e.g. Chimera [67], Swiss-PdbViewer [68], MolIDE [69] and VMD [70] to name a few. Software libraries to manipulate proteins efficiently provide basic data representation and more advanced functionality with a different focus each. ESBTL [71] is mainly a PDB file parser. Biskit [72] additionally provides functionality for analysis of molecular dynamics simulations, while PTools [73] focuses on molecular docking. OpenStructure [74] places more attention on structure visualization and energy calculation. The latter is also supported by MSL [75] and Tinker [76], while BALL [77] in addition provides many advanced optimization algorithms. Finally, StrBioLib [78] extracts sequence information from the protein structure and can be used as an interface to several available third-party tools.

[1] The results of this chapter have been published in Hirsh, L., Piovesan, D., Giollo, M., Ferrari, C., Tosatto, S. C. (2014). The Victor C++ library for protein representation and advanced manipulation. *Bioinformatics*, btu773.

The Critical Assessment of techniques for protein Structure Prediction (CASP) series of experiments [79] demonstrates that structure prediction is increasingly becoming an engineering problem, where sophisticated methods have to be combined into extensive pipelines in order to provide state-of-the-art results [80]. This has raised the barrier for entry into the field to a point where little new developments are possible, considering that most software libraries used in CASP are proprietary and not available as open source. Here, we propose the open source Victor (VIrtual Constrution TOol for pRoteins) C++ library as a way to mitigate this problem. Victor is both an efficiently designed C++ library, able to manipulate protein structures with minimal CPU time, and a collection of advanced components for protein sequence and structure manipulation. In particular, Victor provides three sample applications: profile-profile sequence alignments [81], statistical potentials [82] and loop modelling [83]. Each of these three applications has been extensively described in the literature and is beyond the scope of this writing. To the best of our knowledge, neither is available as an open source C++ library yet. Profile-profile sequence alignments, in particular, have been widely used to improve target-template alignment in CASP [84]. Victor is composed of over 60,000 lines of code and still expanding as it is used in the main authors teaching. It was developed in-house over the last fifteen years with the contribution of tens of developers and has reached a high level of maturity. Victor is released to provide a platform for contributions from the interested community. It provides extensive online material in the form of a Wiki with help files, tutorials, Doxygen documentation and a list of applications built using Victor can be accessed from the URL: `http://protein.bio.unipd.it/victor/`. The actual GitHub repository with C++ source files, a precompiled Ubuntu 64-bit version and unit tests are available from URL: https://github.com/BioComputingUP/Victor.

3.1.1 Core Library

The Victor C++ library currently contains two components for data representation and manipulation in separate directories: tools and Biopool. Tools provides basic manipulation methods, e.g. vector coordinates and file I/O. The core of the library is provided by the Biopool module, which defines all relevant data structures and algorithms to represent protein structures and manipulate them at a higher level of abstraction. The core data structures were carefully developed using design patterns [85], in order to provide an elegant and simple, yet powerful set of C++ classes. In order to allow the simple manipulation of protein structure through the more intuitive torsion angles, automating low level geometric transformations, atom positions are coded both explicitly in 3D coordinates and as a position relative to the previous atom on a graph structure. This ensures consistency in the structure, while allowing the programmer to change the protein conformation rotating a torsion angle with a single line of code. Computational efficiency is guaranteed by updating the corresponding Cartesian coordinates only when necessary. All low level geometrical transformations remain transparent to the user. Biopool is able to read properly all existing PDB files. Additional tools are also provided, such as protein secondary structure automatic assignment with an ad hoc implementation of the original DSSP algorithm [30]. Extensive online documentation allows the interested programmer to learn how to manipulate the Biopool data structures.

3.1.2 Applications

The Victor library provides three main examples to demonstrate the range of possible applications, which are included as separate subdirectories: Energy, Align and Lobo. Extensive documentation, including detailed tutorials, is provided online in order to allow users to become familiar with the software and build on existing knowledge. Energy contains everything that is necessary to develop statistical potentials to evaluate protein structures. Two sample implementations of published methods included in the library, FRST [82] and TAP [86], can serve as a guide to develop additional methods. Both are contained in the Energy subdirectory and functioning code is provided both to generate the statistical potential itself as well as to use it on a PDB structure to calculate the potential energy. The interested user can thus easily develop additional statistical potentials.

The Align directory provides basic sequence alignment algorithms [87] augmented with secondary structure element [88]. Many different profile-profile scoring schemes [81] are implemented, which have been extensively used in CASP to detect remotely homologous protein sequences. Code is also provided for variable gap penalties with additional terms for sequence to structure fit [89] and advanced weighting schemes such as PSIC [90]. Alignment parameters have been extensively benchmarked and the default parameters are optimized for performance.

Last but not least, the Lobo directory contains an application of ab initio loop modeling using a fast divide and algorithm [83]. This makes extensive use of the functions to construct novel amino acids and manipulate the protein structure locally, providing sample code for more complex structural manipulations. It can easily be extended for ab initio structure prediction in combination with statistical potentials as target function.

3.1.3 Conclusions

The Victor library is an open source project devoted to the structural bioinformatics community. It provides a unique combination of methods for sequence and structure manipulation. Expansion is ongoing both through in-house development, as it is the basis for several more publications (e.g. RING [35] and NeEMO [91]), and as part of the authors teaching activities, which include software development projects for students. We hope that the Victor library will contribute towards an easier development of advanced methods for structural bioinformatics.

3.2 RING MD: Gathering Time Into Structures

Molecular dynamics (MD) is nowadays a key toolbox for disease studies [92, 93, 94, 95] and protein characterization [96, 97, 98, 99]. MD gained further relevance after a Nobel Prize award for Chemistry was assigned for insights in this field [42]. Despite its proven effectiveness, the knowledge required to interpret and gather information from classical MD output still represents the greatest issue for the scientific community. RING MD is a tool designed to simplify the analysis and interpretation of MD experiments, which are otherwise dependent on human expertise. Networks [100, 101] are increasingly used in biology, but little work was done to analyze MD simulations [102, 103]. Residue Interaction Networks (RINs) forms the basis of this work, thanks to their proven ability in the representation of key protein features and contacts [100, 101]. Every MD simulation can be seen as multiple snapshots of the same structure, changing its conformation and fluctuating around the so-called equilibrium state. As the conformation changes, the respective RIN changes as well. The RING web server [100] computes a RIN where 6 different chemical interactions are identified, namely hydrogen bonds, π-cation interactions, π-π stacks, ionic bonds, disulfide bonds and van-der-Waals interactions. Generating a network from a single experimental PDB is nevertheless a rough approximation of the true RIN, as the stochastic molecular fluctuations can modify the edges in the protein network. Ideally, RIN interactions should be extended to highlight (i) possible or (ii) time-dependent edges (see Fig. 3.1). We performed RING MD analysis on 50ns MD simulations [104, 70, 105, 106] on Ubiquitin protein [107] to validate the tool. For each snapshot of the simulation, RING MD computes different conformer networks called Time-Dependent Contact-Maps (TDCM)[102, 108, 109], which describes the possible states explored during time. These networks are different, as the intrinsic protein flexibility highlights different protein states. To address the high sparcity of amino acid interactions, we encoded RINs to a set of binary values associated with each residue. Amino acids involved in a chemical bond are the only active variables (see Fig 3.1). We captured protein variability using k-means clustering algorithm [108], which grouped RINs into k similar protein conformers that model part of the MD. The amount of clusters k is a tradeoff between model likelihood and number of parameters as defined by the Bayesian information criterion (see Section 3.2.1).

3.2.1 Methods

A protein r can be represented as a $\mathbb{R}^{n\times 3}$ matrix, where n is the number of amino-acids that need a 3-dimensional representation. Let r_i be the i-th residue, then we can compute the so-called contact map $m \in [0,1]^{n\times n}$, which is defined as follows:

$$m_{ij} = \begin{cases} 1 & \text{if } \|r_i - r_j\| < c \\ 0 & \text{otherwise} \end{cases}$$

Where $\|.\|$ is the well known euclidean norm and c is a user-defined cutoff (typically set to 6-12 Å). Contact maps are intimately related to RINs, and any result obtained in the former can be immediately extended to the latter. In our work, we introduced the concept of Time-Dependent Contact-Map (TDCM), which is based on the following set of contact matrices:

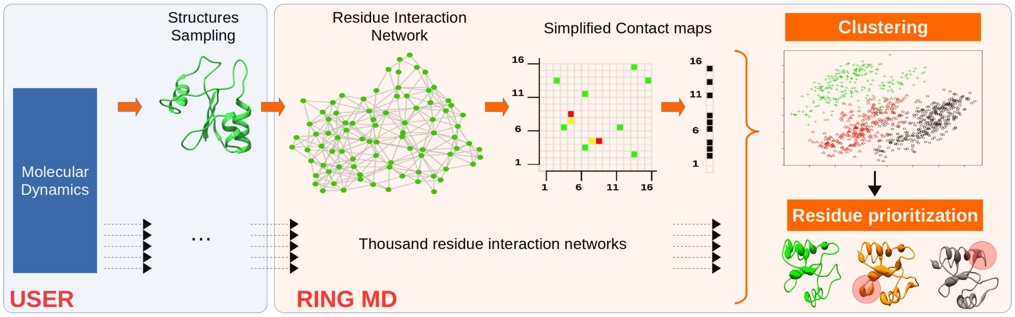


Figure 3.1: RING MD Workflow. Starting from a MD, a user can easily sample representative PDB structures of the whole simulation. With RING MD, the corresponding Residue Interaction Networks and contact maps are computed, and converted into a novel data structure called contact vector. This representation emphasizes at the residue level the interacting amino-acids by removing the sparse information about the contact partners. A clustering procedure is then used to select different conformers based on contact vectors. Structures belonging to highly diverse protein states are provided to the user, along with a prioritization of the amino-acids explaining conformers diversity. Finally, the interaction frequency for each residue is reported in the selected structures, which summarize the intra-cluster contacts variability.

$$m_{ij}(t) = \begin{cases} 1 & \text{if } \|r_i(t) - r_j(t)\| < c \\ 0 & \text{otherwise} \end{cases}$$

where $p(t)$ is the set of 3-dimensional coordinates of the protein r at discrete time $t \in T$ of MD simulation. The first straightforward implementation of TDCM is proposed in [102], where authors proposed to calculate the mean of $m(t)$ matrices to easily summarize the MD. In order to generalize this concept, we represent each contact-map $m(t)$ as a vector by means of the vectorization operator vec():

$$\text{vec}(m) = [m_{1,1}, \ldots, m_{n,1}, m_{1,2}, \ldots, m_{n,2}, \ldots m_{1,n}, \ldots, m_{n,n}]^T$$

and use the concept of convex hull to denote the space of possible contact-maps:

$$\text{conv}(m(t_1), \ldots, m(t_{|T|})) = \left\{ \sum_{t \in T} a_t \text{vec}(m(t)) \middle| \forall t : a_t \geq 0 \land \sum_{t \in T} a_t = 1 \right\}$$

In other words, the space of conformers observed during the MD is described by a linear combination of the contact-maps. This is an interesting formulation, as it consider explicitly all the possible states of the protein. In particular, if we set $a_t = |T|^{-1} \forall t \in T$, we obtain the formulation proposed in [102]. On the other hand, a given protein of interest explores different conformers during the MD, and none of them will be described accurately by the mean contact map proposed in [102]. To overcome this limitation, we propose to use clustering methods to better represent TDCM. In our work, we used k-means algorithm [108] to identify k similar contact maps:

$$\text{argmin}_S \sum_{i=1}^{k} \sum_{m(t)_j \in S_i} ||m(t_j) - \mu_i||$$

Thus, contact-maps are assigned to different clusters in S, and their means $1, ..., k$ can be used to summarize states-dependent interactions. In other words, the convex hull is basically segmented by assigning a coefficient of $a_j = |S_i|^{-1}$ to all contact-maps that belonging to certain cluster S_j and 0 otherwise. As results, we obtain a set of contact maps that are likely to be generated from similar conformers, and therefore good representative of a that protein state. The number k of conformers has been estimated by means of the Bayesian information criterion (BIC) [108], which is a well-known model selection technique. It is defined as follows:

$$-2 \ln(P(m(t)|\mu, S)) + k \cdot n^2 \cdot \ln(|T|)$$

The objective function is dependent on two aspects: the likelihood of the k-means fitting, and a penalty related to the $k \cdot n^2$ number of parameters needed for k clusters. Thus, BIC aims to select a model representing a compromise between fitting quality and its complexity. In order to speed up the computations, we simplified the contact map representation using a map function $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^n$. This function transform the matrix $m(t)$ into a a vector $m'(t)$ where:

$$m'_i(t) = \begin{cases} 1 & \text{if} \sum_{j=1}^{n} m_{i,j}(t) > 0 \\ 0 & \text{otherwise} \end{cases}$$

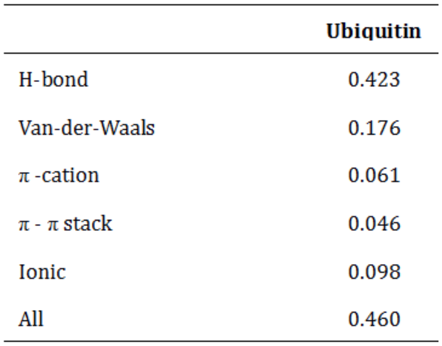
Indeed, for each residue, we set to 1 those ones that are currently involved in an interaction. This is a critical improvement, since it is practically infeasible to run clustering algorithms on thousands of protein frames where the number of residues is high (e.g. greater than 1,000). In addition, it is common to focus only on contact maps with main chain-side chain and side chain-side chain interactions, as they are the most informative bonds [109]. These contact maps are highly sparse matrices, so the information lost by means of our function f is very limited. Finally, this has also an impact in the BIC model selection process, as it reduces the penalty term related to the model complexity and allowing the discovery of additional useful conformers. It should be noted that the parameters obtained from clustering can be interpreted as a mixture model. Therefore, one can use such a model to test if a different MD of the same protein is generated from the same stochastic process. A clear instance of this idea is the evaluation of mutations impact in the protein dynamic or in its stability.

3.2.2 Results

Clustering validation

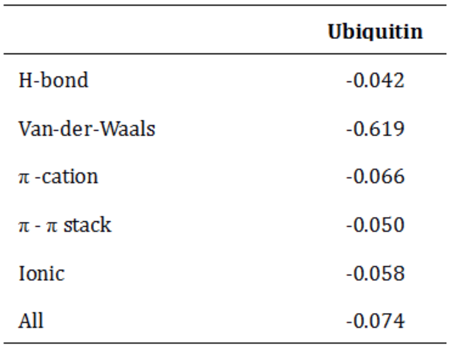
In order to validate RING MD clustering methodology, we used and idea from CoDNaS [110], a database of known experimental conformations for thousands of proteins. In their work, the authors suggested to use hierarchical clustering on the protein structure coordinates to determine PDB similarities. Such an approach enables the detection of largely different conformers describing diverse native states. Ideally, we expect to observe agreement between RING MD and CoDNaS-based clustering for MD simulated conformers. In agreement to CoDNaS definition, we computed the pairwise distance of MD frame coordinates using the Bio3D package [111], and calculated a dendrogram. Using in-house code, we obtained the conformers dendrogram represented from simplified contact maps (SCM). In Table 3.1,

the clustering similarity is evaluated by means of Cophenetic correlation. For Ubiquitin, hydrogen bonds seems the main cause of tridimensional conformation switch. The joint combination of multiple types of interactions (denoted as all) has a correlation value of almost 0.5, and demonstrates the high explanatory power of RINs in the detection of different protein native states. A second appealing strength of RINs is their ability to spot the type of interactions promoting conformational changes. This is a clear abstraction of the whole dynamics, and enables the quick detection of key factors for the protein stability and its fluctuations. A last important consideration arises when we compare the average number of interactions for a given residue during the MD with its root mean square fluctuation. As shown in Table 3.2, the inter-atomic contacts explain a significant part of the amino-acid degree of freedom. This is obvious in practice, and is also well modeled by our RIN-based MD representation.
We also explored the use of Hidden Markov Models (HMM) to better summarize the dynamics. However, the increasing number of parameters and higher computational burden and this approach provided a similar data separation, confirming the effectiveness of simpler clustering methods.



	Ubiquitin
H-bond	0.423
Van-der-Waals	0.176
π -cation	0.061
π - π stack	0.046
Ionic	0.098
All	0.460

Table 3.1: Cophenetic correlation of PDB clustering and RING MD average clustering. There is good agreement between the dendrograms produced from the different type of protein representations. The “All” row represents the sum of all the interactions but the inter-atomic contacts, and describes the existence of any bond.



	Ubiquitin
H-bond	-0.042
Van-der-Waals	-0.619
π -cation	-0.066
π - π stack	-0.050
Ionic	-0.058
All	-0.074

Table 3.2: Correlation of RING MD mean number of interactions of a residue with respect to amino acid RMSF. Total amount of Van-der-Waals contacts correlates with the residue fluctuations.

In depth analysis

A comparison between classical MD analysis and RING MD output is shown in Fig. 3.2. The RMSD [Fig.3.2a] of two Ubiquitin [107] ionic interactions (residues K27-D52 and D39-R72) is compared with the RING MD summary structure [Fig. 3.2d-e]. We focused on R72 as it has widely been demonstrated to be involved in several ubiquitin interactions and cleavage phenomena [110, 112, 113]. Particularly, in 2011 Ralat and coworkers demonstrated that Ubiquitin is a target of Insulin-degrading-enzyme (IDE). The catalytic activity of this enzyme passes through a first cleavage, involving the terminal glycine dimer and forming the 1-74 Ubiquitin segment, being R74 exposed to solvent. A second and slower cleavage follows the first one, and targets R72 forming the 1-72 Ubiquitin segment [110]. In RING MD summary structure, residues are highlighted based on their interaction frequencies (see Fig. 3.2). The interaction between residues D39 and R72 is less frequent compared to K27-D52, which instead occurs almost in every frame. Therefore, according to RING MD, the slower IDE activity towards R72 could be explained by the formation for the 37% of the frames of R72-D39 ionic interaction, not allowing R72 to participate in the cleavage reaction. Indeed, IDE catalytic chamber has been proven to show protease activity towards highly flexible chains [114]. Furthermore, RING MD provides an interaction-based trajectory clustering, with the conformers observed during the MD described by a linear combination of the contact-maps. In the reported case study, 4 clusters were obtained (see Fig. 3.2c). Particularly, structural clusters 3 and 4 represent molecular conformers not suitable for the cleavage reaction, containing R72 tightly locked with D39. On the contrary, in clusters 1 and 2 R72 is not involved in ionic interactions, and free to participate in other biochemical mechanisms. These protein states can clearly coexist, with the probability of observing them changing during time. Figure 3.2c represents the stochastic dynamic process, and facilitate in finding relationships among conformer transitions and periodicity. Extending the concept of RINs over a time dependent scenario proved that the joint combination of multiple simplified contact maps contains all the initial RIN information. In fact, RING MD can reconstruct automatically the initial contacts with no information loss (see automatically coloured interacting residues in [Fig. 3.2b]). Moreover, the integration of the clustering results with the structural view can explain in detail the reasons of conformational changes. The output of the entire analysis is provided in supplementary information Fig. 3.1, highlighting changes in ionic, h-bond, and overall interactions. Conformational changes identified by the RING MD clustering show a good agreement with structure-based analysis (Tab. 3.1)[110], and can be used to highlight the type of interactions relevant for these changes. It is important to note that classical clustering is based on 3D coordinates comparison of the structure, while TDCM compare graphs. Highly fluctuating residues, like N and C termini have high impact in terms of coordinates change, leading to a potential overestimation of structural distance. On the other hand, contact maps are robust to these irrelevant fluctuations and therefore more likely to provide a good final clustering. In this context, RING MD can play a key role in Conformational Ensembles for the analysis of intrinsic disorder proteins [112], since the classical structural comparison is known to be ineffective. We also observed that Inter-atomic interaction probabilities calculated with RING MD correlate significantly with Root Means Squared Fluctuations (RMSF) (Tab. 3.2), thus explaining well the degrees of freedom of protein residues.

Figure 3.2 description

a) Root mean square deviations (RMSDs) of ASP52-LYS27 (red) and ASP39-ARG72 (black), and ASP58-ARG54 (green) interactions. The distance in the plot is expressed in angstroms. b) Time Dependent Contact Map (TDCM) of the ionic interactions occurring in Ubiquitin MD simulation. Highlighted in red ASP52 and LYS27, in black ASP39-ARG72 and in green ASP58 and

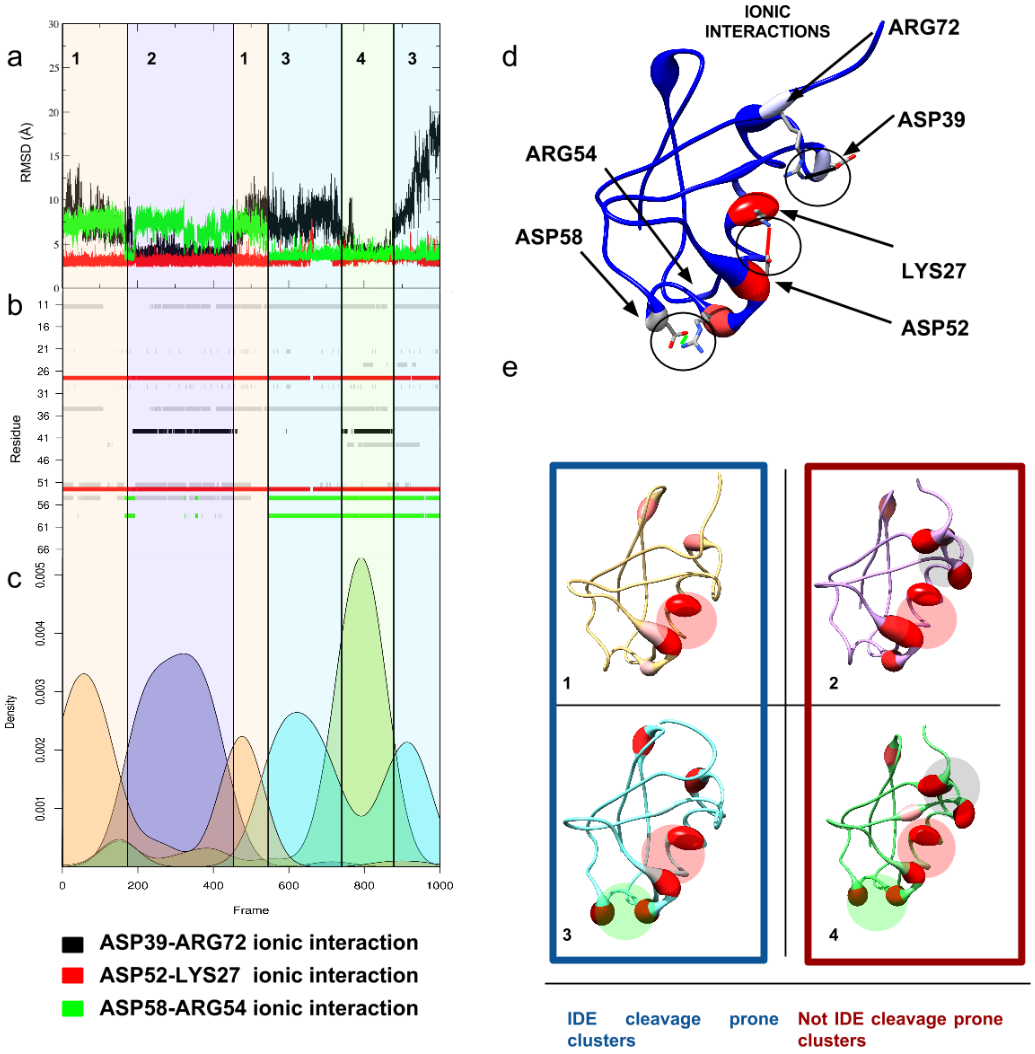


Figure 3.2: Results of Ubiquitin RING MD analysis.

ARG54. c) Clusters deriving from k-means analysis of 1000 frames. a, b and c share the x axis, which is time (frames) of the MD simulation. The colouring scheme is the following: cluster 1 is light orange, cluster 2 is violet, cluster 3 is light blue and cluster 4 is light green. d) Summary structure reporting the frequencies of interaction between ASP39-ARG72 (light blue), LYS27-ASP52 (red) and ASP58-ARG54 (light blue and red, respectively). The thicker regions of this structure depict where ionic interactions (not highlighted) are occurring more frequently. e) Clusters structures. In cluster 1 ASP52-LYS27 is the only interaction occurring frequently. In cluster 2 ASP39-ARG72 and ASP52-LYS27 are occurring. In cluster 3 ASP52-LYS27 and ASP58-ARG54 are interacting and no interaction is found for ASP39 and ARG52. In cluster 4 all the chosen ionic couples are interacting.

3.2.3 Conclusion

The capability of molecular force-fields in simulating macromolecular systems provided a plethora of useful insights to scientific development. RING MD fits perfectly in this background, providing the user speed and reliability in analysing key features of a given system. The tool allows to quickly recognize structural hot spots in protein structures MD simulations, providing a fertile ground to improve protein biophysics knowledge in a faster and easier way. The results obtained with RING MD, compared to the literature, revealed that RING MD analysis could match hot spot residues with a simple trajectory analysis [113, 114, 115].

3.3 RepeatsDB: a database of tandem repeat protein structures[2]

A large portion of proteins contain repetitive motifs, which are generated by internal duplications and frequently correspond to structural and functional units of proteins. Many repetitions in protein sequences can be identified by using different approaches [116, 117, 118, 119]. A more difficult problem for identification is however posed by repeats in protein structure, which can be highly degenerate [120, 121]. In fact, it is possible for a protein to maintain a repetitive structure even in the presence of massive amounts of point mutations [122]. Several repeat families have been studied so far due to their relevance in different biological processes such as health [123],

[2] The results of this chapter have been published in Di Domenico, T., Potenza, E., Walsh, I., Parra, R. G., Giollo, M., Minervini, G., Piovesan D, Ihsan A, Ferrari C, Kajava A, Tosatto, S. C. (2013). RepeatsDB: a database of tandem repeat protein structures. *Nucleic acids research*, gkt1175.

neurodevelopment [124] and protein engineering [125, 126, 127], to name just a few.

Repeats have been previously divided into five broad classes, primarily as a function of repeat length [128, 129]. At the lower end of the repeat length spectrum, i.e. less than five residues, very short repeats can either form insoluble aggregates (crystallites, class I) or long and winding helices of fibrous structures like collagen and α-helical coiled-coils (class II). At the other end of the spectrum, repeats containing $>\sim$50 residues appear to fold mostly as domains forming beads-on-a-string structures (class V). In between, for unit lengths of 540 residues, the known repeats can form either open elongated solenoids (class III) or closed toroids (class IV). Due to their fundamental functional importance, classes III and IV contain the most studied types of tandem repeat proteins. Solenoid folds appear to follow the distribution of repeat lengths rather closely, from all-beta (e.g. anti-freeze proteins) [130] to mixed alpha/beta (e.g. leucine-rich repeats) [131, 132] to all-alpha structures (e.g. Armadillo and HEAT repeats) [133, 134, 135]. They are characterized by some of the largest known autonomously folding domains, with 500 or more residues forming a single structure [136]. Rapid addition or deletion of repeat units even between close homologs is of particular note for solenoid structures [137]. Toroids on the other hand are restricted in overall size by their closed circular nature. Known toroid structures include the highly versatile TIM barrel and large outer membrane beta-barrels [138]. Perhaps a more interesting fold is the beta propeller (e.g. WD repeats), which can accommodate variable numbers of repeat units while maintaining a closed circular structure [139, 140].

An open question regarding repeat proteins is the existence of other common structures that may have gone undetected. After all, the most common way to detect repeat families so far was to manually annotate the sequence family first and only afterwards visually recognize their structural repetitiveness. Such an approach is obviously difficult when dealing with the entire Protein Data Bank (PDB) [141], especially considering the many uncharacterized protein structures deposited by the main structural genomics consortia [142]. The systematic description of repeat structures becomes a question of using automated methods to detect them in protein structures. This field is relatively new, with only few available methods. One of the first attempts was made by the Thornton group [143], but is unfortunately no longer available. Some methods [119, 144, 145, 146, 147, 148] were developed to detect internal symmetries in proteins, but these may be difficult to adapt to the systematic classification of repeats. Recently, our group has developed RAPHAEL [149] in an attempt to fill the gap for repeat detection from structure. Widely used structural classifications such as CATH [150] and SCOP [151] also do not explicitly annotate repeats in protein structures, although it may be possible to leverage individual annotations to find similar repeats. Some databases exist for the detection of repeats from sequence [152, 153], but usually these are limited to short tandem repeats and do not take into account divergent repeats, such as solenoids or toroids. The main domain sequence databases such as Pfam [49] and SMART [154] do not excel at the annotation of these repeat types either, as coverage is rather low and many repeat units go undetected. For Pfam most of the largest clusters of human sequence regions not covered were recently found to be repeats [155]. To the best of our knowledge, no database or classification is currently available for repeat structures. This is the motivation for our present work, and we introduce RepeatsDB as a way to fill this gap. The database was developed to provide a central resource for the systematic annotation and classification of repeats. Given the fact that the structure-based search and classification of repeat proteins is more complete than on the basis of sequences or key words, our database will allow more accurate assignment of proteins with repeats to the corresponding families. For example, it will be used to suggest a better subdivision of alpha-solenoid proteins where at present the boundaries between the structures with Armadillo, HEAT, TPR and other repeat types are frequently blurred.

3.3.1 Database description

Data curation

The initial dataset for RepeatsDB was extracted from the PDB [7]. Repeat candidates were identified from the reduced PDB dataset with RAPHAEL [149], which uses a geometric approach imitating the work of a human curator (score cutoff ≥ 1). The resulting dataset consisted of >10,000 repeat candidates, stored in the database as predicted entries, which underwent a classification and curation process.

The dataset of predicted repeats was manually curated using a two-level annotation system. The first manual annotation level (manually classified) classifies an entry into structural repeat class and subclass. This classification is based on previous work [129], where five classes of repeat structures are proposed, which are then further divided into subclasses. Class assignment is based mainly on repeat unit length and subclass assignment on secondary and tertiary structure features. The second manual annotation level (detailed) consists in providing information about the start and end positions of the repeat units, repeat regions and/or insertions. We define a repeat unit as the smallest structural building block that is repeated to form a repeat region. A repeat region is a group of at least three repeat units. Inclusion of proteins with two repeat units would significantly complicate classification because many typical globular domains have this type of architecture. Insertions are non-repeated segments of structure that occur either inside a repeat unit or between two of them. These are particularly interesting because they break the repeat symmetry, and represent a challenge both for automatic detection and for the analysis of repeat structures [149].

Several curators annotated each protein undergoing manual classification by consensus. For first-level annotations, at least 75% of the curators had to agree in order for a protein to be included, otherwise it would be excluded and placed on a reserve list for future annotation. The rationale for this choice is that ambiguous cases are generally difficult to classify but may occasionally represent a novel repeat class. For second-level annotations, the threshold for consensus was at least 65% agreement (typically two of three curators). In case of discrepancy, an expert would arbitrate the final annotation based on the alternative proposals. Proteins with detailed annotations were also used to search for similar sequences in proteins from the PDB. Any PDB chain with at least 40% sequence identity and a coverage of at least 80% of the classified protein, belonging to the initial list of predicted entries, is added to the classified by similarity annotation level. The similarity thresholds were selected to exclude possible false-positives (data not shown).

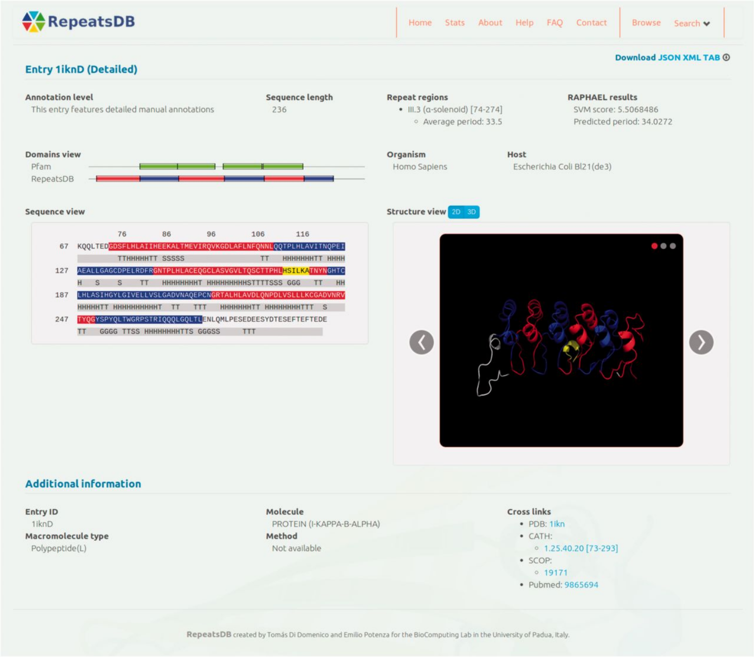
Implementation

RepeatsDB was designed with a multi-tier architecture, using separate modules for data management, data processing and presentation functions. To simplify development and maintenance, all tiers handle the common JSON (JavaScript Object Notation) format, thereby eliminating the need for data conversion. The MongoDB database engine is used for data storage and Node.js as middleware between data and presentation. RepeatsDB exposes its resources through RESTful web services, by using the Restify library for Node.js. The Angular.js framework and Bootstrap library were selected to provide the overall look-and-feel. Angular.js to Bootstrap integration is available through the angular-ui project. A customized version of the BioJS [156] sequence component is used as sequence visualizer. Additional information is added to entries by querying the PDB web services at the structure and chain level. At the structure level, annotations like organism and experimental method used when resolving the structure are provided. At the chain level, secondary structure and links to other databases, among others. RepeatsDB offers users both graphical web interface access and RESTful web services from URL: `http://repeatsdb.bio.unipd.it/`.

3.3.2 Using RepeatsDB

The user interface presents an intuitive tree-based browsing mechanism, where the root of the tree is the full database, second-level nodes repeat classes and third-level nodes subclasses. When clicking on a node, the user is presented with the list of RepeatsDB entries corresponding to the selected category. Each row of the list shows basic information about the entry, like its entry ID, title and organism. All annotated chains corresponding to an entry are displayed in a single page. The user interface presents a structure and sequence visualization widget (Figure 3.3). The user may choose to visualize the structure in four static images, or by using the 3D visualizer. If the entry features detailed annotations, the repeat regions, units and/or insertions are displayed using a combination of colours. The sequence visualization widget displays the sequence and secondary structure corresponding to the structure. It displays the same colour coding as the structure visualization widget, associating repeat annotations in the structure and sequence views. Additional information at the structure and chain levels is also provided.

The RepeatsDB search toolbar, available on top of every page, allows to search for entries either by database IDs or UniProt text query. The database ID search allows comma-separated PDB or UniProt IDs. The UniProt text search query uses the full UniProt search engine, see online documentation. RESTful web services are directly accessible through HTTP URLs. All data available on RepeatsDB are also available for programmatic access. Please refer to the Help section of the website for details on using the RepeatsDB web services. Datasets can be downloaded in JSON, XML or text format using the browse function or RESTful web services.



Entry 1iknD (Detailed)

Annotation level
This entry features detailed manual annotations

Sequence length
236

Repeat regions
- III.3 (α-solenoid) [74-274]
  - Average period: 33.5

RAPHAEL results
SVM score: 5.5068486
Predicted period: 34.0272

Domains view
Pfam
RepeatsDB

Organism
Homo Sapiens

Host
Escherichia Coli Bl21(de3)

Sequence view

```
                76        86        96        106       116
 67  KQQLTEDGDSFLHLAIIHEEKALTMEVIRQVKGDLAFLNFQNNLQQTPLHLAVITNQPEI
              TTHHHHHTT SSSSS                   TT   HHHHHHHTT HHHH
127  AEALLGAGCDPELRDFRGNTPLHLACEQGCLASVGVLTQSCTTPHLHSILKATNYNGHTC
     H   S    S    TT   HHHHHHHHT HHHHHHHHHSTTTTSSS GGG   TT   HH
187  LHLASIHGYLGIVELLVSLGADVNAQEPCNGRTALHLAVDLQNPDLVSLLLKCGADVNRV
     HHHHHTT HHHHHHHHHHT  TT   TTT   HHHHHHHTT HHHHHHHHTTT  S
247  TYQGYSPYQLTWGRPSTRIQQQLGQLTLENLQMLPESEDEESYDTESEFTEFTEDE
     TT   GGGG TTSS HHHHHHHHTTS GGGSS     TTT
```

Structure view

Additional information

Entry ID
1iknD
Macromolecule type
Polypeptide(L)

Molecule
PROTEIN (I-KAPPA-B-ALPHA)
Method
Not available

Cross links
- PDB: 1ikn
- CATH:
  - 1.25.40.20 [73-293]
- SCOP:
  - 19171
- Pubmed: 9865694

RepeatsDB created by Tomás Di Domenico and Emilio Potenza for the BioComputing Lab in the University of Padua, Italy.

Figure 3.3: Screenshot of a sample RepeatsDB entry results page (PDB entry 1ikn). The sequence viewer and the structure viewer are shown in the middle of the page, towards the left and the right, respectively. Additional annotations at the structure and chain level are displayed, including links to other databases (above) and classifications (below).

Statistics

Analysis of the full PDB dataset yielded 10 745 repeats predicted by RAPHAEL, of which 2797 were finally classified into the RepeatsDB schema. Table 3.3 shows the distribution between classes and subclasses. The bulk of the annotations (<90%) consist of entries belonging to classes III and IV. No effort was made to balance the distribution of entries between classes in this initial release. As coverage increases in the future, we expect the balance to approximate the real distribution more closely, although it may be necessary to fine-tune RAPHAEL. Of the classified entries, 321 representatives of the entire dataset were annotated in detail with information about the start and end of repeat regions, repeat units and/or insertions (Table 3.3). It is interesting to note the different distribution of insertions between classes. Apparently, some classes such as β-solenoid (class III.1) or TIM barrels (class

IV.1) have stronger propensity to accommodate insertions.

3.3.3 Conclusions and future work

RepeatsDBs goal is to provide the community with a resource for high-quality tandem repeat protein structure annotations. The user can either interactively analyse his proteins of interest via the user interface, or create and download datasets for offline use. Far from being a static classification process, the annotation effort for the initial RepeatsDB dataset alone already motivated the extension of the original classification schema [129]. Some of the curated structures, while clearly representing structural repeats, did not belong to any of the pre-defined subclasses. To allow them to be classified, subclasses IV.5 (α/β prism) and IV.6 (α-barrel) were added to the initial schema [129]. Class V also underwent a re-classification according to the secondary structure content of the single domain repeats (beads) to allow a broader classification range beyond individual repeat families, as the list of possible beads-on-a-string folds may be considerably larger than currently appreciated. The other subclass was also added to allow collection of repeats that do not fit into the current classification scheme. RepeatsDB provides the community with a previously unavailable opportunity to easily create datasets of tandem repeat proteins. The detailed annotation subset further presents a unique opportunity to better understand the nature of tandem repeat proteins.

Beyond its initial release, RepeatsDB is a continuous effort to expand, revise and improve tandem protein repeat annotations. Predictions for new PDB structures are simple and fully automated, allowing regular database updates every 3 months. Manual curation of new entries for inclusion is also ongoing, aiming at regular and steady updates. Options to involve the community into the annotation process through crowd-sourcing tools are currently being analysed. A main goal for future versions is the extension of the annotation of repeats at the sequence level, starting from annotation for intrinsically disordered regions from MobiDB [157]. We anticipate that RepeatsDB should prove valuable towards the understanding of the sequencestructure relationship in tandem repeat proteins and their evolutionary relationship.

Subclass	Name	Detailed	Classified (manually)	Classified (by similarity)	Predicted
I.1	Poly-alanine β structure	0	0	0	0
II.1	Collagen triple-helix	0	5	0	0
II.2	α helical coiled coil	23	38	69	0
III.1	β-solenoid	43	113	21	0
III.2	α/β solenoid	21	43	27	0
III.3	α-solenoid	48	246	631	0
III.4	Trimer of β spirals	7	0	13	0
III.5	Single layer anti-parallel β	4	3	0	0
IV.1	TIM-barrel	84	118	626	0
IV.2	β-barrel	8	1	8	0
IV.3	β-trefoil	20	0	29	0
IV.4	β-propeller	40	182	227	0
IV.5	α/β prism	0	17	0	0
IV.6	α-barrel	6	0	0	0
V.1	α-beads	2	1	0	0
V.2	β-beads	29	12	71	0
V.3	α/β-beads	3	3	1	0
V.other	Unknown subclass	3	0	4	0
UA	Unassigned	0	0	0	7948
	Total	321	749	1727	7948

Table 3.3: Statistics for RepeatsDB. The subclass name is shown together with the number of entries on each of the four annotation levels. Note that Unassigned entries are automatically predicted by RAPHAEL and therefore not assigned to a specific class.

Chapter 4

Intrinsic protein disorder

Intrinsically disordered regions are key for the function of numerous proteins. They are commonly defined from missing electron density in x-ray structures. Due to the difficulties in experimental disorder characterization, many computational predictors have been developed with various disorder flavors. Their performance is generally measured on small sets mainly from experimentally solved structures, e.g. Protein Data Bank (PDB) chains. MobiDB [157] annotates disorder for UniProt sequences, and enabled the retrieval of 25,833 different sequences with X-ray crystallographic structures. From this unique dataset, section 4.1 reports the first large-scale assessment of fast disorder predictors. In addition to a comprehensive ranking of methods, the analysis produced interesting observations about bias and limitations of computational tools. Experimental evidence for long disorder regions (LDRs) of at least 30 residues in the dataset deserved the specialized analysis of section 4.2. Here, I describe the first comprehensive and fully automated large-scale study of experimental LDRs for 1,758 unique proteins, demonstrating a recent increasing coverage of intrinsic disorder in the Protein Data Bank. Overall, the chapter is mainly related to the characterization of disorder and its features. The *knowledge discovery in database* process (section 1.2) will be the main actor from the methodological point of view, but its results will represents the basis for future predictors. In fact, all findings presented here will be critical for the next generation of machine learning tools and for disorder pattern recognition in proteins.

4.1 Comprehensive Large Scale Assessment of Protein Disorder[1]

The rigid structure of proteins has been considered the determinant of function for many years. Recently, an alternative view is emerging with respect to non-folding regions, suggesting a reassessment of the structure-to-function paradigm [158] [159] [160]. Flexible segments lacking a unique native structure, known as intrinsic disordered regions [161], are widespread in nature, especially in eukaryotic organisms [162]. These regions have been shown to play important roles in various biological processes such as cell signaling or regulation [163], DNA binding and molecular recognition [164]. Their malleable properties allow multiple binding partners [165] with the flexible region often becoming folded on binding [166].

Despite an emerging consensus regarding their existence, there is no single definition of disorder. As a result, various flavors of disorder have been proposed [167]. These disorder flavors have become diverse with some based on amino acid composition [167], flexibility [168] and functional roles coupled with conservation [169]. Perhaps the simplest flavor distinction is the length of a disordered region, separated into short and long. Long regions seem to behave differently [170] and are difficult in structural determination, causing them to be underrepresented in the Protein Data Bank (PDB) [171]. The PDB contains structural information from X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, which can be used indirectly to study disorder. A plethora of computational predictors have also appeared, with special efforts to capture different flavors. Available methods can be broadly divided into three classes: biophysical, machine learning and consensus based. Biophysical methods [172] [173] [174] [175] derive pseudo-energy functions from residue pairings in rigid structures (i.e. non-disorder) to recognize sequence regions with high energy as disordered. Machine learning, especially neural networks, has been widely used to predict protein disorder [176] [177] [178] [179] [180], [181] [182] [183]. Many are tuned for the disorder style used in the Critical Assessment of techniques for protein Structure Prediction (CASP), where the goal is to detect missing residues in the X-ray

[1] The results of this chapter have been published in Walsh, I., Giollo, M., Di Domenico, T., Ferrari, C., Zimmermann, O., Tosatto, S. C. (2014). Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics*, btu625. For Supplementary Material, check the on-line version of the paper.

crystal [184]. Others attempting to move away from this disorder style measure some form of protein backbone flexibility. For example, ESpritz [181] can predict mobile NMR regions and DisEMBL [179] loops regions with high B-factor (high flexibility). The most recent disorder predictor category uses a consensus of various biophysical and machine learning methods [185] [186] [187] [180] [188]. Consensus approaches are frequently more accurate, but at the computational cost of running several predictors in parallel and averaging their output. Because there is no consensus on how to define disorder, predictors often vary in their parameter setting and disorder output. In nearly all cases, disorder is defined at the residue level, and the goal of the predictor is to maximize recovery of correct residues.

Among applications of disorder prediction, we can distinguish at least two different scenarios. The first is the CASP experiment [184], i.e. methods are used to predict a relatively small number of proteins with maximum accuracy and consensus predictors aiming for maximum accuracy should clearly excel. A more practical scenario is represented by high-throughput analysis of protein disorder on entire genomes [159]. Over the years, most prediction methods have addressed the first problem, with comparatively little attention to the practicalities of large-scale predictions [181]. MobiDB [189] is a large-scale disorder database containing experimental information on the entire PDB and predictions for all UniProt [5] sequences. Here, we use the vast quantity of disorder data for a first large-scale assessment. While most assessments are performed with hundreds or a few thousand examples, we have analyzed $>$ 25,000 UniProt sequences combining all available X-ray crystallographic structures. All disorder assessments so far are carried out on single PDB chains, whereas here the UniProt sequence is the final target. The UniProt annotation is unique and we compare it with standard PDB chain analysis for further insights.

4.1.1 Methods

Datasets and classifications

All UniProt [5] sequences with at least one X-ray annotation in MobiDB [189] were downloaded on the May 13, 2013 (25,833 entries). Where more than one MobiDB annotation was available, a majority vote was used (Fig. 4.1) to produce a more stable disorder definition, filtering rare conflicts due to experimental conditions. Where MobiDB cannot find annotation for part of the UniProt sequence, residues are annotated as unknown and ignored. Each PDB [171] chain that covered UniProt entries was also extracted for comparison and identical chains majority voted (Fig. 4.1). Similar chains were removed at 90% pairwise sequence identity using CD-HIT (101,338 chains reduced to 24,669). See Table 4.1 for statistics. Each UniProt entry was assigned to CATH using SIFTS [190]). Gene Ontology (GO) terms [38] were downloaded from UniProt and expanded to the ontology root. For a deeper analysis, the UniProt dataset was further split according to the following rules (Supplementary Table S1): removing short (<30 residue) PDB fragments, excluding conflicting residues, up to 10 non-consecutive disordered residues, >10 disordered residues.

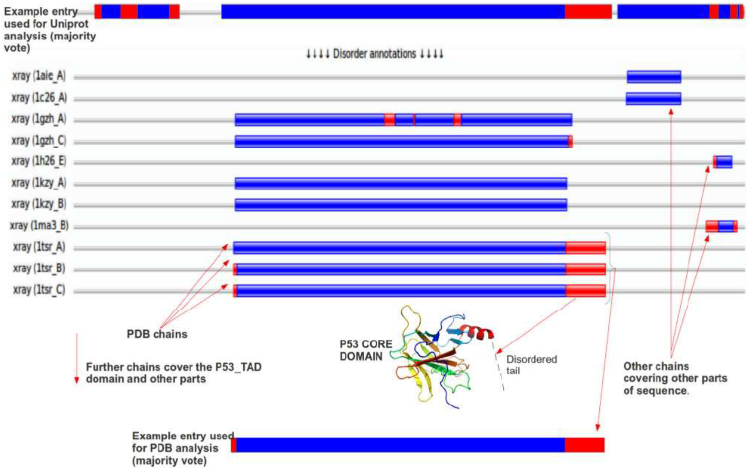
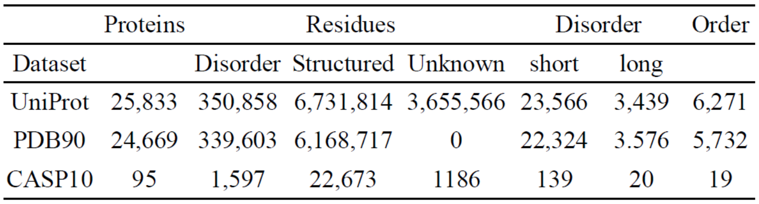

Example entry used for Uniprot analysis (majority vote)
↓↓↓↓ Disorder annotations ↓↓↓↓
xray (1aie_A)
xray (1c26_A)
xray (1gzh_A)
xray (1gzh_C)
xray (1h26_E)
xray (1kzy_A)
xray (1kzy_B)
xray (1ma3_B)
xray (1tsr_A)
xray (1tsr_B)
xray (1tsr_C)
PDB chains
Further chains cover the P53_TAD domain and other parts
P53 CORE DOMAIN
Disordered tail
Other chains covering other parts of sequence.
Example entry used for PDB analysis (majority vote)

Figure 4.1: Human P53 (UniProt ID: P04637) disorder annotation. The top bar shows the majority voting scheme, with blue for order and red for disorder. For simplicity, only a subset of PDB hits was shown. Missing regions are not considered. The bottom bar shows an example majority voted chain used in the PDB chain analysis.

Predictors

Predictors were selected with the condition that they must be available as an executable and fast, ideally returning predictions in <1 min. The following 11 programs were used (disorder definition used in parenthesis): ESpritz (X-ray, NMR and DisProt; [181], IUPred (short and long; [172], DisEMBL (hot



	Proteins	Residues			Disorder		Order
Dataset		Disorder	Structured	Unknown	short	long	
UniProt	25,833	350,858	6,731,814	3,655,566	23,566	3,439	6,271
PDB90	24,669	339,603	6,168,717	0	22,324	3.576	5,732
CASP10	95	1,597	22,673	1186	139	20	19

Table 4.1: Number of proteins, residues and region size.

loops and remark 465; [179], RONN (X-ray; [183] and VSL2b (combination of X-ray and Disprot; [191], GlobPlot (globularity; [174] and FoldIndex (folding; [175]. This resulted in a total of 11 predictors with different disorder flavors. A short description of the predictors is given in the Supplementary Material. Predictor similarity was calculated on their residue scores (e.g. probability of disorder) was shown as a dendogram based on SOV performance. Low-complexity regions are parts of the sequence with strongly biased compositions (e.g. polyQ), which are thought to correlate with intrinsic disorder [192]. The low-complexity predictors SEG [193] and Pfilt [194] were used both as disorder predictors and to analyze disorder predictor performance in low-complexity regions.

Performance assessment

Disorder prediction is a binary classification problem. As such, the standard measures accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC) and area under the curve (AUC) are used (see Supplementary Material). All these measures are calculated both per residue and as average on a per protein basis. MCC and AUC were replaced by SOV and FPreg in the per protein analysis. SOV is the mean of the segment overlap for disorder and structure, in analogy to secondary structure [195]. FPreg counts the number of predicted false-positive disordered regions. Disorder content measures the ability to recover the fraction of disordered residues in a protein independent of residue position. We adopted two previously used measures [196], root mean square error (RMSE) and Pearson Correlation Coefficient (PCC), with predicted and observed disorder content normalized by the number of annotated residues. As a large number of measures hinders a global view of performance, we established an overall ranking as the average over all 12 quality measures. The Welch t-test was used to compute statistical significance.

4.1.2 Results

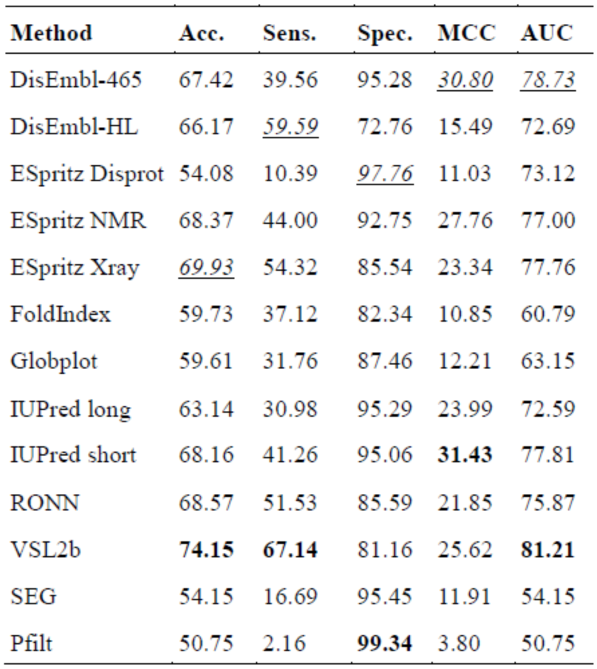
First large-scale disorder assessment

We report the first large-scale assessment of disorder predictions on UniProt sequences through a comprehensive assessment of 11 fast predictors with new performance measures and a statistical evaluation. With respect to diversity, the UniProt set has 15,942 unique clusters at 40% identity cutoff. The 24,699 PDB chains are non-redundant by design (see Methods). Therefore, in both sets there is no large cluster of similar sequences, guaranteeing no bias in the analysis.

Figure 4.1 shows human P53 sequence (UniProt ID: P04637) covered by different disorder and structure definitions from the PDB. Using the majority voting approach, all structural and disorder information is combined and a more reliable global picture of the full p53 complex is constructed. Table 4.1 shows the number of proteins and residues using this annotation strategy. A total of 25,833 UniProt entries are annotated and a dataset of unique PDB chains is constructed for comparison purposes. For all sets, there is a clear imbalance between disorder and structured residues. Similar to the CASP10 experiment [184] where 20% of the data was completely ordered, 6271 sequences 24.3% of the UniProt dataset are completely ordered. Long disordered regions (>20 residues) are also abundant with 3439 examples.

Table 4.2 shows the per-residue performance on the UniProt dataset. Supplementary Tables S3 and S4 show similar results when excluding short PDB fragments and conflicting positions. The same trends are also found when separating the UniProt dataset for disorder content (see Supplementary Tables S5 and S6 and Section 3.4 below).

Table 4.3 shows the per protein and disorder content performance. Most predictors have disorder scores significantly above random (AUCs >70). Depending on the prediction style, e.g. high coverage (overprediction) or highly confident (underprediction), one could argue for and against different predictors. For example, VSL2b has a lower residue specificity (81.16) predicting many false positives (1,268,274 residues), yet its AUC is the highest. On the contrary, IUPred-short has the best MCC (31.43) due to its high specificity. Table 4.2 can reveal detailed future objectives such as the need to retune



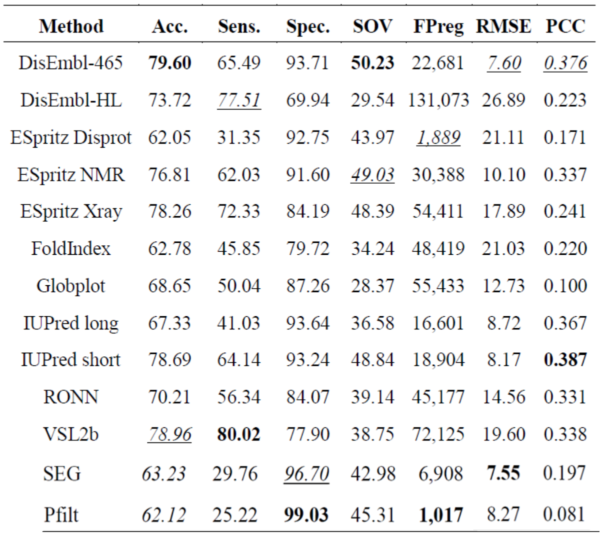
Method	Acc.	Sens.	Spec.	MCC	AUC
DisEmbl-465	67.42	39.56	95.28	*30.80*	*78.73*
DisEmbl-HL	66.17	*59.59*	72.76	15.49	72.69
ESpritz Disprot	54.08	10.39	*97.76*	11.03	73.12
ESpritz NMR	68.37	44.00	92.75	27.76	77.00
ESpritz Xray	*69.93*	54.32	85.54	23.34	77.76
FoldIndex	59.73	37.12	82.34	10.85	60.79
Globplot	59.61	31.76	87.46	12.21	63.15
IUPred long	63.14	30.98	95.29	23.99	72.59
IUPred short	68.16	41.26	95.06	**31.43**	77.81
RONN	68.57	51.53	85.59	21.85	75.87
VSL2b	**74.15**	**67.14**	81.16	25.62	**81.21**
SEG	54.15	16.69	95.45	11.91	54.15
Pfilt	50.75	2.16	**99.34**	3.80	50.75

Table 4.2: UniProt per-residue performance. All values are shown as percentages. The top performing method in each category is shown in bold and the second best underlined.

the VSL2b decision threshold for higher specificity. For the SOV measure, DisEmbl-465 has the best performance (50.23). FPreg measures overprediction on segments as opposed to single residues. Again VSL2b clearly over predicts compared with DisEmbl-465.

Similarity between measures and predictors

While the evaluation complexity arises due to predictor variability and the quantity of performance measures, it is useful to understand the deeper predictor behavior. Although many more observations could be made, for the sake of brevity a summarized ranking was chosen to give a clearer performance summary. Before ranking, it is important that the measures are evaluating different aspects of the predictions. Figure 4.2 shows that no measure correlates highly and most are diverse (pairwise correlation value

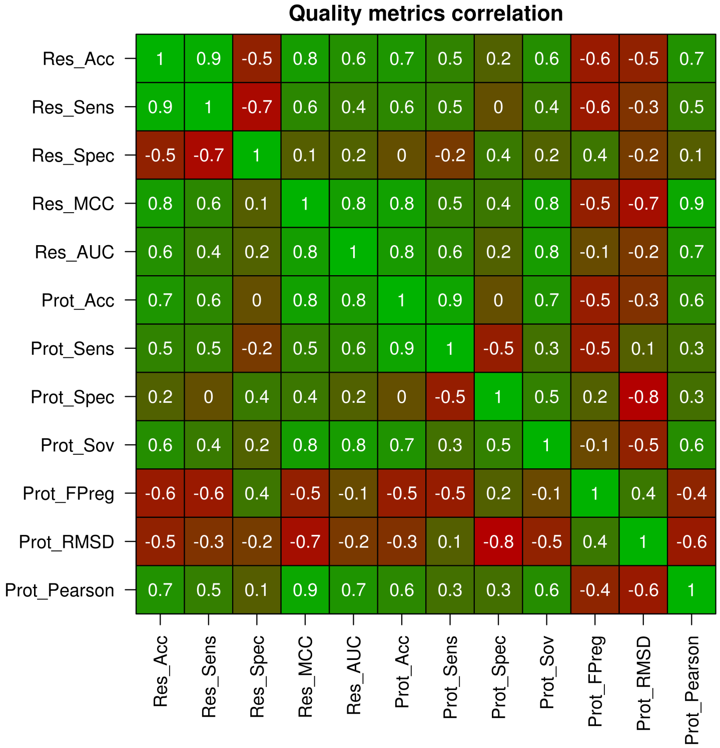


Method	Acc.	Sens.	Spec.	SOV	FPreg	RMSE	PCC
DisEmbl-465	**79.60**	65.49	93.71	**50.23**	22,681	*7.60*	*0.376*
DisEmbl-HL	73.72	*77.51*	69.94	29.54	131,073	26.89	0.223
ESpritz Disprot	62.05	31.35	92.75	43.97	*1,889*	21.11	0.171
ESpritz NMR	76.81	62.03	91.60	*49.03*	30,388	10.10	0.337
ESpritz Xray	78.26	72.33	84.19	48.39	54,411	17.89	0.241
FoldIndex	62.78	45.85	79.72	34.24	48,419	21.03	0.220
Globplot	68.65	50.04	87.26	28.37	55,433	12.73	0.100
IUPred long	67.33	41.03	93.64	36.58	16,601	8.72	0.367
IUPred short	78.69	64.14	93.24	48.84	18,904	8.17	**0.387**
RONN	70.21	56.34	84.07	39.14	45,177	14.56	0.331
VSL2b	*78.96*	**80.02**	77.90	38.75	72,125	19.60	0.338
SEG	*63.23*	29.76	*96.70*	42.98	6,908	**7.55**	0.197
Pfilt	*62.12*	25.22	**99.03**	45.31	**1,017**	8.27	0.081

Table 4.3: UniProt per-protein and content performance. All values are shown as percentages, except RMSE and PCC. The top performing method in each category is shown in bold and the second best underlined.

>0.7 or <0.7 in only 9 of 66 cases). This diversity ensures the ranking procedure is fair. Interestingly, per-residue MCC correlates highly with both disorder content measures (0.7 with RMSD and 0.9 with PCC), suggesting they could be captured effectively by residue-level MCC. Both were kept because they are not completely redundant and the content measures do not depend on residue positions. From Figure 4.3, the top-ranked predictors are DisEmbl-465 and IUPred-short, with no statistically significant difference between performance (P-value 0.64). A second group consists of Espritz-NMR, VSL2b and Espritz-X-ray. In Supplementary Figure S2, these top five predictors are analyzed using receiver operating characteristic curves at low 05% false-positive rates (FPR). VSL2b starts outperforming the rest at around 2% FPR, again suggesting high-quality residue scores but a need for recalibration of its decision threshold.

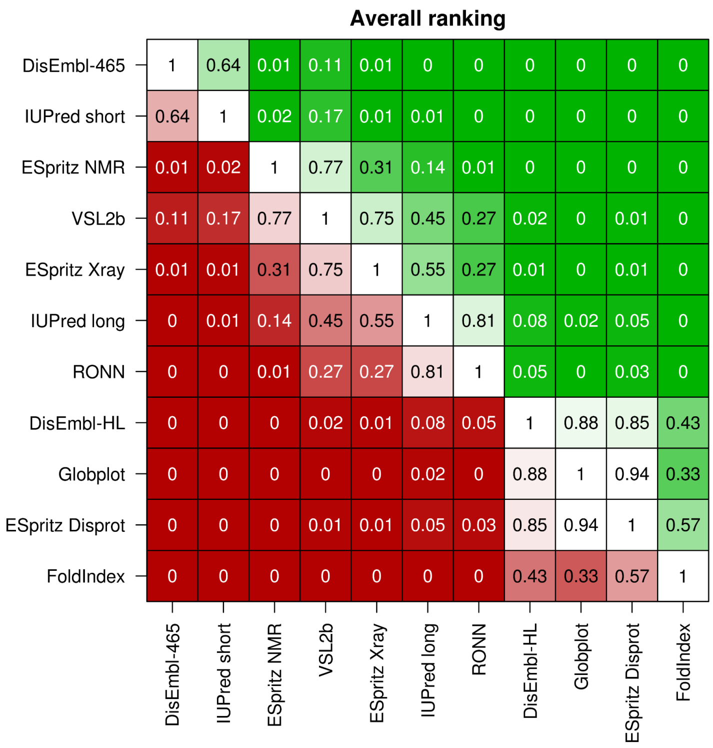
Combining several good but complementary predictors is the heart of most consensus methods. Figure 4.4 shows the Pearson correlation between predictors. Examining similarity and performance is the first step in designing a consensus. Both are not necessarily related, e.g. IUPred-short, IUPred-long and RONN form a group of highly correlated predictors (PCC



	Res_Acc	Res_Sens	Res_Spec	Res_MCC	Res_AUC	Prot_Acc	Prot_Sens	Prot_Spec	Prot_Sov	Prot_FPreg	Prot_RMSD	Prot_Pearson
Res_Acc	1	0.9	-0.5	0.8	0.6	0.7	0.5	0.2	0.6	-0.6	-0.5	0.7
Res_Sens	0.9	1	-0.7	0.6	0.4	0.6	0.5	0	0.4	-0.6	-0.3	0.5
Res_Spec	-0.5	-0.7	1	0.1	0.2	0	-0.2	0.4	0.2	0.4	-0.2	0.1
Res_MCC	0.8	0.6	0.1	1	0.8	0.8	0.5	0.4	0.8	-0.5	-0.7	0.9
Res_AUC	0.6	0.4	0.2	0.8	1	0.8	0.6	0.2	0.8	-0.1	-0.2	0.7
Prot_Acc	0.7	0.6	0	0.8	0.8	1	0.9	0	0.7	-0.5	-0.3	0.6
Prot_Sens	0.5	0.5	-0.2	0.5	0.6	0.9	1	-0.5	0.3	-0.5	0.1	0.3
Prot_Spec	0.2	0	0.4	0.4	0.2	0	-0.5	1	0.5	0.2	-0.8	0.3
Prot_Sov	0.6	0.4	0.2	0.8	0.8	0.7	0.3	0.5	1	-0.1	-0.5	0.6
Prot_FPreg	-0.6	-0.6	0.4	-0.5	-0.1	-0.5	-0.5	0.2	-0.1	1	0.4	-0.4
Prot_RMSD	-0.5	-0.3	-0.2	-0.7	-0.2	-0.3	0.1	-0.8	-0.5	0.4	1	-0.6
Prot_Pearson	0.7	0.5	0.1	0.9	0.7	0.6	0.3	0.3	0.6	-0.4	-0.6	1

Figure 4.2: PCC among performance measures. Each cell shows the PCC for the corresponding measures, with colors varying from green (+1) to red (1). Res denotes per residue and Prot per protein measures.

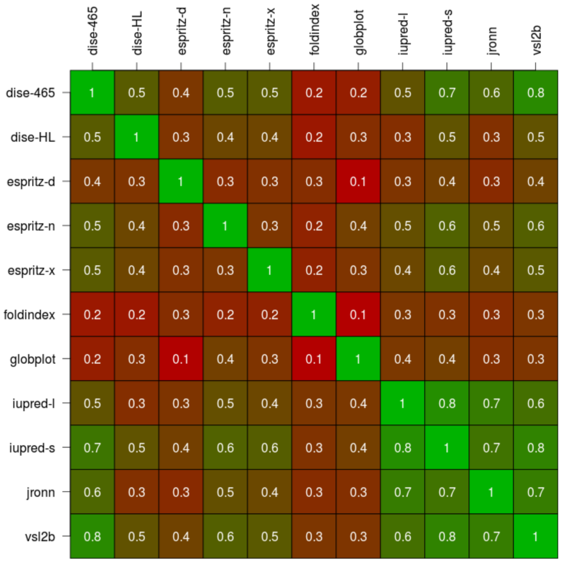
range 0.70.8) with different performances. Figure 4.5 shows a dendrogram of predictors grouped by SOV. This SOV difference of correlated predictors is mainly due to their selected decision threshold with residue scores remaining similar. Three methods, FoldIndex, GlobPlot and DisEmbl-HL correlate poorly with all others (PCC < 0.5) and also perform poorly on SOV. They nevertheless define different disorder flavors, which may be useful in certain situations. IUPred-short and DisEmbl-465 have high correlation (>0.7 PCC) in conjunction with similarly high SOV, suggesting they detect the same disordered regions well. For a consensus approach, however, it is more interesting to combine, for example, Espritz-NMR/X-ray with DisEmbl-465, as they have low correlation (PCC 0.5) and quality SOV. To investigate consensus further, we measure agreement among predictors. Figure 4.6 shows the residues split into three equal groups: consensus structure, consensus dis-



	DisEmbl-465	IUPred short	ESpritz NMR	VSL2b	ESpritz Xray	IUPred long	RONN	DisEmbl-HL	Globplot	ESpritz Disprot	FoldIndex
DisEmbl-465	1	0.64	0.01	0.11	0.01	0	0	0	0	0	0
IUPred short	0.64	1	0.02	0.17	0.01	0.01	0	0	0	0	0
ESpritz NMR	0.01	0.02	1	0.77	0.31	0.14	0.01	0	0	0	0
VSL2b	0.11	0.17	0.77	1	0.75	0.45	0.27	0.02	0	0.01	0
ESpritz Xray	0.01	0.01	0.31	0.75	1	0.55	0.27	0.01	0	0.01	0
IUPred long	0	0.01	0.14	0.45	0.55	1	0.81	0.08	0.02	0.05	0
RONN	0	0	0.01	0.27	0.27	0.81	1	0.05	0	0.03	0
DisEmbl-HL	0	0	0	0.02	0.01	0.08	0.05	1	0.88	0.85	0.43
Globplot	0	0	0	0	0	0.02	0	0.88	1	0.94	0.33
ESpritz Disprot	0	0	0	0.01	0.01	0.05	0.03	0.85	0.94	1	0.57
FoldIndex	0	0	0	0	0	0	0	0.43	0.33	0.57	1

Figure 4.3: Average ranking over all performance measures with statistical tests. The predictors to the left and bottom are ranked with P-value separating groups (Welch t-test) in each cell. Colors for the left predictor range from green (better) to red (worse), passing through white (tied). P-values >0.05 mean the performance distributions are similar and the difference between two predictors is not statistically significant.

order and uncertain. Uncertain is defined when there is disorder agreement for 47 predictors because the accuracy continually decreased below 61.2% (see Supplementary Fig. S3). When there is confident agreement, accuracy increases as expected, i.e. both tails of Supplementary Figure S3: 03 structure and 811 disorder agreement. In these regions, a consensus can recover 88.8 and 24.5% highly confident structured and disordered residues, respectively. Applying a simple majority vote in analogy to secondary structure [197] produced 43.3% sensitivity, 95.6% specificity and an AUC of 78.8 per residue (see Table 4.2 for comparison).



	dise-465	dise-HL	espritz-d	espritz-n	espritz-x	foldindex	globplot	iupred-l	iupred-s	jronn	vsl2b
dise-465	1	0.5	0.4	0.5	0.5	0.2	0.2	0.5	0.7	0.6	0.8
dise-HL	0.5	1	0.3	0.4	0.4	0.2	0.3	0.3	0.5	0.3	0.5
espritz-d	0.4	0.3	1	0.3	0.3	0.3	0.1	0.3	0.4	0.3	0.4
espritz-n	0.5	0.4	0.3	1	0.3	0.2	0.4	0.5	0.6	0.5	0.6
espritz-x	0.5	0.4	0.3	0.3	1	0.2	0.3	0.4	0.6	0.4	0.5
foldindex	0.2	0.2	0.3	0.2	0.2	1	0.1	0.3	0.3	0.3	0.3
globplot	0.2	0.3	0.1	0.4	0.3	0.1	1	0.4	0.4	0.3	0.3
iupred-l	0.5	0.3	0.3	0.5	0.4	0.3	0.4	1	0.8	0.7	0.6
iupred-s	0.7	0.5	0.4	0.6	0.6	0.3	0.4	0.8	1	0.7	0.8
jronn	0.6	0.3	0.3	0.5	0.4	0.3	0.3	0.7	0.7	1	0.7
vsl2b	0.8	0.5	0.4	0.6	0.5	0.3	0.3	0.6	0.8	0.7	1

Figure 4.4: PCC among predictors. Each cell shows the PCC for the corresponding measures with colors varying from green (+1) to red (1).

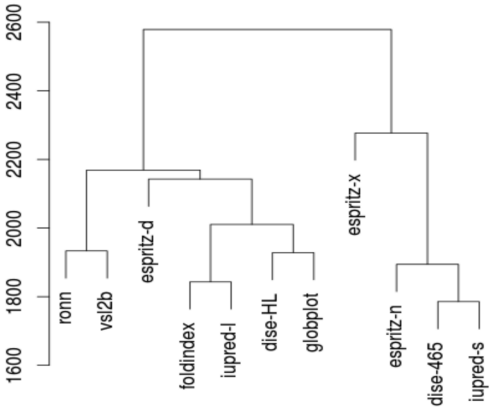


Figure 4.5: Dendogram for disorder residue score and performance using the SOV measure. On the Y axis, the cumulative SOV score difference is plotted.

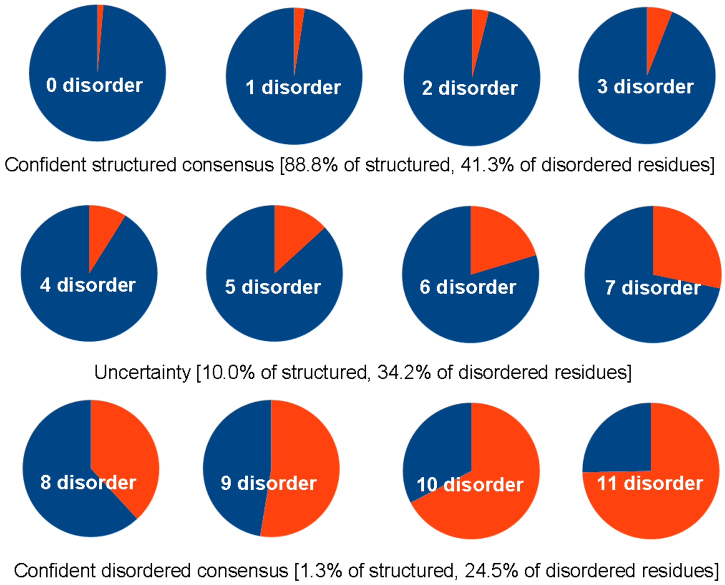


Figure 4.6: Proportion of data that can be assigned confidently using a consensus. The pie chart shows each of the 11 possible scenarios (i.e. from 0 to 11 disorder predictions) and the corresponding fraction of truly disordered residues (in red). Each row corresponds to a situation (structure, uncertain, disorder) for which the percentages of occurrence are summarized.

Uniprot versus PDB chains

The most surprising result is the large decrease in performance compared with previously published performances. As most of the assessments in the literature are based on PDB chains, we examine whether assessments on PDB chains behave differently from UniProt sequences. Figure 4.7 shows the per-residue AUC differences between the UniProt and PDB chain datasets. Most predictors perform better on PDB chains and start approaching their published values (e.g. ESpritz-X-ray AUC 86.58 on CASP9). This is possibly due to the fact that predictor parameters are optimized on PDB chains. Another possible reason may be the positional dependence in PDB sequences, i.e. missing atoms or disordered residues in solved structures are often located at the N and C termini. This effect was recently noted for CASP-10 [184]. Given that most methods encode the sequence context (e.g. using sliding windows in neural networks) they will implicitly learn the position of the termini. This information is lost when the PDB sequence is assigned to a part, often the middle, of the UniProt sequence. Moreover, the definition is different in UniProt because it is a majority combination of multiple experimental sources. Supplementary Table S7 shows the full set of performance measures on the PDB chain set.

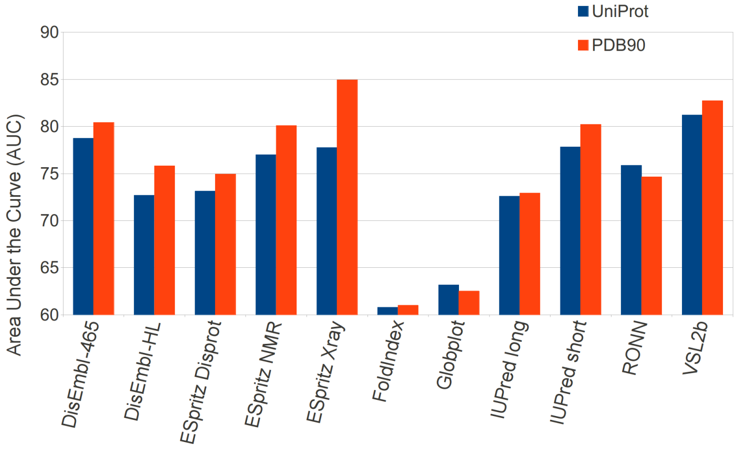


Figure 4.7: Comparison between PDB chains and UniProt.

Sequence and structure variability

Fluctuations in performance given different protein properties are often overlooked. To our knowledge, this has never been examined comprehensively. In all cases, performance is assessed using SOV and there are indeed some striking performance differences. Proteins are grouped into bins of low complexity content (1% intervals), and the top five predictors are analyzed for performance changes in Supplementary Figure S4. Increasing low complexity content increases SOV logarithmically with the largest gains in the 520% low complexity range. A similar concept was already investigated [192], suggesting that disorder predictors are generally using low complexity patterns to predict unstructured regions. At first glance, this seems to contradict Table 4.2, which shows SEG and Pfilt producing almost random predictions. It can be explained by the fact that no low complexity regions were detected for 40.1% of the disordered proteins. From Supplementary Figure S4 it is clear that low complexity has a significant relationship with disorder performance whenever present.

Disorder region length is perhaps the most obvious sequence property, and the majority (75.8%) of sequences in our dataset contain at most 10 disordered residues (Supplementary Fig. S5). The performance separated on this threshold is shown in Supplementary Tables S5 and S6. This skewed distribution may not reflect the truth in nature, especially for long proteins where long disordered regions may be missed owing to lack of evidence, but is nevertheless interesting, as we are using a common disorder definition [184]. Figure 4.8 shows the performance of each method separated into two sets. Proteins containing at least one long disorder stretch (i.e. >20 residues) or not. Detection of disorder with long regions had a decreased performance in 5 of 11 methods, but 4 of these were the top ranked ones. Conversely, methods trained to take into account long disorder (e.g. IUPred-long, Espritz-Disprot, RONN and VSL2b) and the folding predictors (GlobPlot and Foldindex) showed better performance on proteins with long disorder. This suggests that the top-ranked methods can be improved by taking into account long disorder regions in their training.

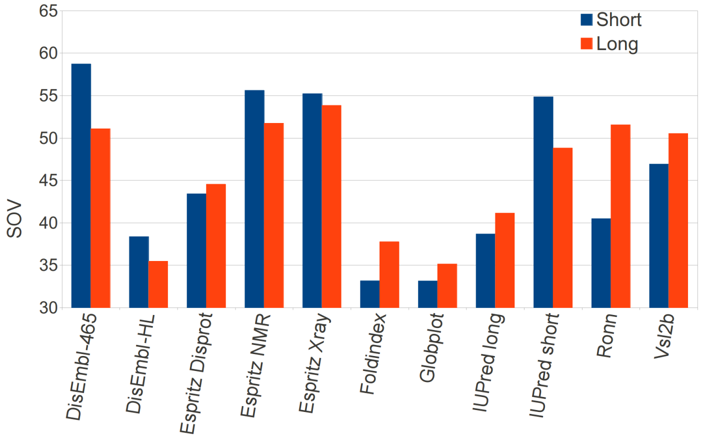


Figure 4.8: Comparison between predictors on long and short disorder proteins.

Using CATH, 9378 proteins with disorder were extracted form the UniProt set. Supplementary Figure S6 shows the performance of the top five methods on the four main CATH classes. The CATH few secondary structure class is predicted considerably below average. This could be due to the high quantity of disorder in this class (Supplementary Fig. S7). Mainly alpha structures are clearly easier to detect, perhaps due to alpha helices being dependent on local sequence. Conversely, mainly beta structures are harder to predict perhaps because beta sheet hydrogen bonds are dependent on distant residues. This difficulty in capturing distant sequential dependencies is a common problem in secondary structure prediction [198] and likely also true for disorder.

Functional variability

The top five predictor SOV performances for proteins with at least one disordered residue are separated into the three GO classes [38]. Cellular Component covers 3458 proteins, Biological Process 4696 and Molecular Function 5260. Figure 4.9 shows how SOV varies significantly for different cellular component terms. Virion-related proteins have the most interesting performance drop, probably due to an increased level of disorder in these proteins (Supplementary Fig. S8). Perhaps more interestingly, performance for membrane and extracellular proteins is generally lower than average, even though the amount of disorder was not enriched (Supplementary Fig. S8). This may be a consequence of disorder having different amino acid composition in these proteins [199]. For GO Molecular Function (Supplementary Fig. S9), binding and transporter activity are predicted well but the activity relationships structural molecule and receptor have the lowest performance. Disorder performance also varies with Biological Process (Supplementary Fig. S10). Signaling and regulation were easier to predict while biological adhesion has a glaringly poor performance. In each of the three GO classes, SOV performance varies by up to 10% for different GO terms.

4.1.3 Discussion

Efficient disorder predictions are vital for understanding large collections of proteins and entire proteomes. In this work, 11 predictors are evaluated on 25,833 UniProt sequences with disorder annotations from X-ray crystallographic structures. The evaluation procedure consists in measuring performances using 12 different scores and ranking the predictors while highlighting statistically similar groups. Although in some cases the disorder definition used will not represent true functional mobility, we feel that it should cap-

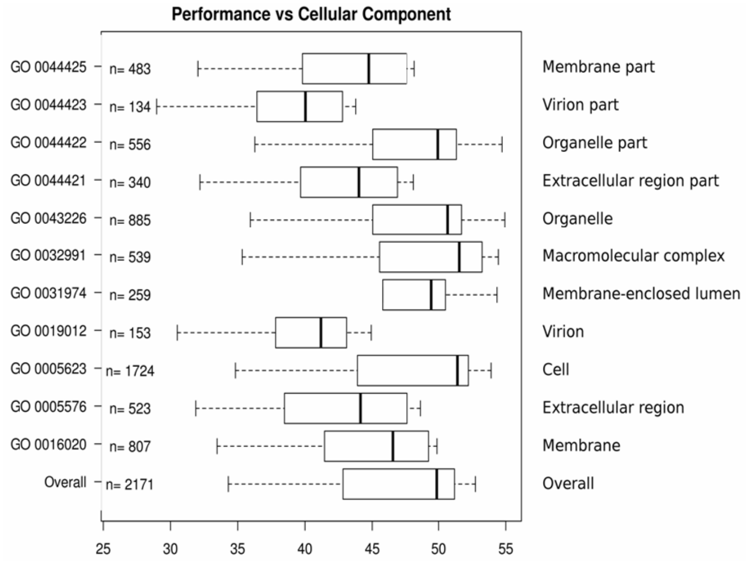


Figure 4.9: Relationship between GO cellular component and disorder performance of top five predictors. All proteins had at least some disorder. GO terms only considered if the number of proteins is >50.

ture most aspects of intrinsic disorder. The assumption is that in most cases, missing backbone atoms in PDB structures correlate with intrinsic disorder defined in DisProt [200]. Our definition is also arguably more stable because it is based on a majority vote on all PDB structures covering a UniProt entry. Thus experimental errors in X-ray crystallography (e.g. missing residues due to low resolution) should be removed, as disorder is only considered if it occurs most frequently in the PDB.

The evaluation reveals a strong variability in predictors across the 12 measures, indicating different prediction styles (e.g. overprediction or confident underprediction). Ranking each predictor with the 12 measures shows both DisEmbl-465 and IUPred-short performing consistently well on each measure. The ranking was robust because the 12 measures show little correlation (Fig. 4.2). Predictors that ranked poorly still contain a good signal across our disorder definition. In most cases, they fall behind because they offer a different interpretation, which may be useful in alternative settings. At CASP, the best disorder predictors are widely known to be meta-predictors combining orthogonal information (i.e. unique predictors performing well). A correlation analysis on the predictors produced similar clusters as well as unique predictors. Some predictors showed both uniqueness and good performance, indicating a consensus predictor may be beneficial. Using predictor combinations, 88.8% of structured and 24.5% of disordered residues are found with highly confident agreement. The remaining 10.0% structured and 34.2% disordered residues classified as uncertain may be decided with more sophisticated heuristics (e.g. high residue scores).

Highly accurate predictors on UniProt sequences are vital, considering that users are invariably trying to understand disorder properties of unannotated proteins and not the PDB, which is already annotated with quality structural information. Despite this, the literature largely concentrates on PDB chain assessments. The performance on the UniProt disorder definition is substantially lower than the equivalent evaluation on PDB chains. A similar effect was recently noted in the CASP-10 assessment, where database predictions were worse than the direct submissions by the same methods [184]. In general, increases in PDB chains are observed across all measures (Supplementary Table S7), suggesting that the prediction of the more desirable UniProt disorder may be worth considering for training new predictors. There are large performance variations when splitting the data into groups of proteins. As expected, predictors prefer large amounts of low sequence complexity, but the performance seems to plateau after 20% low complexity. On the other hand, long disorder detection seems to be more difficult, especially for the predictors we find to be accurate. While both trends are somewhat expected, the dependence of performance on structure and function are less obvious. At the structural level, beta-only proteins seem to be more difficult to predict compared with alpha-only or mixed alpha/beta. The few secondary structure class is certainly the poorest, but this may be due to long disordered regions being poorly detected. Given that functional disorder analysis using predictors is gaining attention [182] [201], prediction error is shown relative to GO. The analysis shows that the average error rate is not universal across all functions. It is possible that enriched functions found in genome analysis may have a slight bias. For example, the association with disorder and binding, signaling and regulation is known, but here we found that they are more easily detected, possibly inferring enrichment. Compared with virion sequences, which are more abundant in experimental disorder (Supplementary Fig. S8) and supported by the literature [201], the error rates on their predictions are higher than average. One of the main reasons for performance variation could be the distribution of protein types in predictor training sets. Binding, signalling and regulation proteins together constitute a large fraction of known disorder datasets, and it is reasonable to assume that the same distributions are used in each predictor. It is therefore possible that predictors are optimized for these common families. Optimistically, the use of this prior knowledge could enhance predictor training or motivate the development of specific tools.

To our knowledge, this is not only the first large scale analysis of disorder predictions from X-ray crystallographic structures but also the first attempt to provide error rates on sequence, structure and functional protein types. We are in the process of developing this evaluation into an automatic evaluation server and plan to integrate it in the new version of MobiDB [189].

4.2 Experimentally determined long intrinsically disordered regions in proteins are now abundant in the Protein Data Bank

Despite a recent consensus regarding the existence of intrinsic disorder in proteins[43], its classification is still quite ambiguous[202]. As a result, various flavors of disorder have been proposed, some based on amino acid composition[167], flexibility[168] and functional roles coupled with conservation[169]. Perhaps the simplest distinction is between proteins with short and long disordered regions. Proteins with long disorder regions (LDRs) are special, since they seem to behave differently in function[182] and evolution[203]. Structurally, disorder can range from regions that in solution are totally unfolded to those that present two or more different, but defined, conformations. Unfortunately, in an x-ray crystal structure the two cases are often difficult to distinguish, in particular at low or medium resolution. In fact, if in a structure at resolution better than 1.5 Å it is in general possible to observe loops or short areas present in two (sometimes three) different conformations, at lower resolution this is generally not the case as these areas are not visible in the electron density map. Consequently, the corresponding residues are left out from the molecular model. In the rest of the chapter, portions of the polypeptide chain that were present in the protein which has been crystallized but are absent from the crystal structure will be generally defined as disordered, without any attempt to distinguish between disordered, flexible or mobile regions. On the contrary, regions characterized by very high thermal parameters, generally an indication of flexibility, will not be taken into account here, but will be left for future developments. As different crystal structures of the same protein may contain varying amounts of disorder, a clear definition is necessary. Here, we use two different rules. In the first, called majority rule, a segment is considered LDR if it is disordered for at least 30 residues in the majority of the crystal structures of the same polypeptide chain. The second, defined zero rule, applies to the subset of majority cases where only disorder is present for a given LDR. Figure 4.10 shows an example of the difference 2 between both rules. In total, extraction with the majority rule removing sequence pairs above 40% identity from MobiDB[189] provided 1,758 unique proteins with at least one LDR (Table 4.4 for dataset composition). Of these, 1,567 are confirmed with the stricter zero rule and 717 have a single Protein Data Bank (PDB)[171] entry. The percentage of disorder in this dataset (13.2%) is higher by design than other datasets. The fraction of absent residues (28.3%) could harbour a further source of LDRs resisting crystallization, making our dataset a lower estimate for disorder.

		Residues			
Data set	**Proteins**	**Disordered**	**Structured**	**Unknown**	**LDRs**
Long disorder proteins	1,519	115,847	517,339	254,822	1,679
Completely structured	4,391	0	979,238	641,321	0

Table 4.4: **Dataset composition**. The number of proteins, residues and LDRs is shown for both the long disorder dataset from x-ray sources in MobiDB and a set of completely structured proteins extracted for comparison. Residues are unknown if there is no PDB structure assigned to those residues in the UniProt entry. More than one LDR per protein may be present. Maximum pairwise sequence identity in both sets is 40

One possible factor contributing to LDRs could be the resolution of the x-ray experiments (see Figure 4.11). As resolution increases, the length of the disordered regions decreases slightly on average. Lower resolution seems to favor the presence of LDRs or, more likely, the presence of LDRs in the protein decreases the diffracting power of the crystal, reducing resolution. The deposition dates for the PDB structures used in the dataset (see Figure

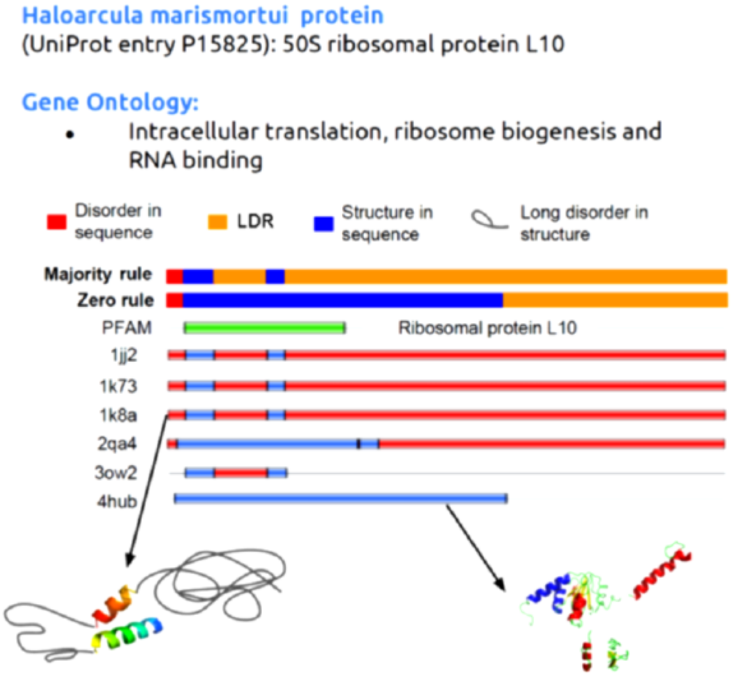

Haloarcula marismortui protein
(UniProt entry P15825): 50S ribosomal protein L10
Gene Ontology:
• Intracellular translation, ribosome biogenesis and RNA binding
Disorder in sequence
LDR
Structure in sequence
Long disorder in structure
Majority rule
Zero rule
PFAM
Ribosomal protein L10
1jj2
1k73
1k8a
2qa4
3ow2
4hub

Figure 4.10: Example of the alternative majority and zero rule LDR definitions on the same protein with one Pfam domain and numerous x-ray structures shown by PDB code. Each line corresponds to a disorder annotation. Notice how the C-terminus is always disordered, while the first half has some x-ray structures suggesting order

4.12) shows that most LDR structures have been deposited in the last five years, suggesting that the dataset will automatically increase over the next years.

The distribution of proteins grouped by their largest LDR is shown in Figure 4.13, with DisProt[200] and IDEAL[204] for comparison, showing an exponential decay with increasing length. 94.4% have at least one region between 30-119 amino acids and this decrease is consistent with IDEAL and DisProt. The 99 proteins with extreme LDRs (up to 500 amino acids) are a very unusual part of the PDB. Although each protein may contain more than one LDR, one region is the norm (1,587 proteins), with two being somewhat

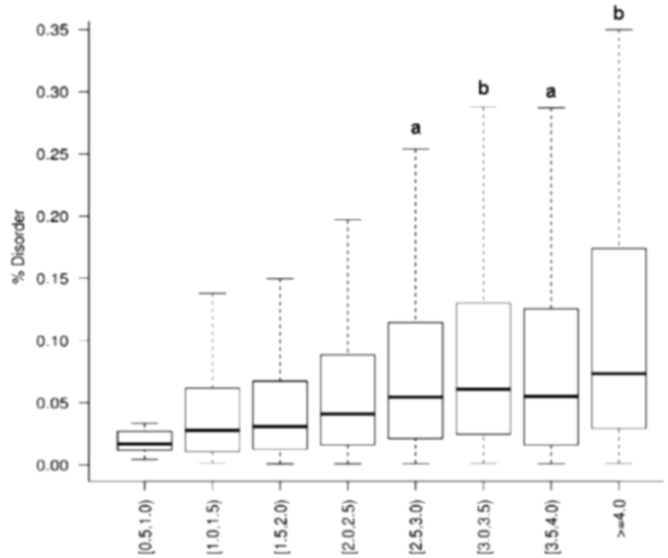


Figure 4.11: Disordered region length as a function of x-ray resolution. The box plot show median (dark bar) and upper/lower quartiles as box boundaries, with the 1.5x interquartile range as dashed lines. Outliers are not shown for clarity. Each bin is statistically different(Wilcoxon test, p-value $<$ 0.05) than the others, with exception of the two groups marked with a and b.

common (157 proteins) but only rarely up to four. Somewhat unexpectedly, the majority of LDRs (52%) are not at the commonly flexible N and C termini (see Figure 4.14).

The collected data is more abundant than LDR proteins in the DisProt and IDEAL databases. While all three are derived from experimental sources, this dataset is over three times bigger than DisProt and over 14 times IDEAL (see Figure 4.15), with only small intersections of similar proteins. Combining the three sets will allow the construction of a larger set of 2,079 unique proteins. Our data does not replace the excellent work of IDEAL and Disprot, but rather offers a much larger complementary experimental LDR source.

While some of the LDRs may be the result of poor diffraction quality, it is now well established that the majority of them have functional

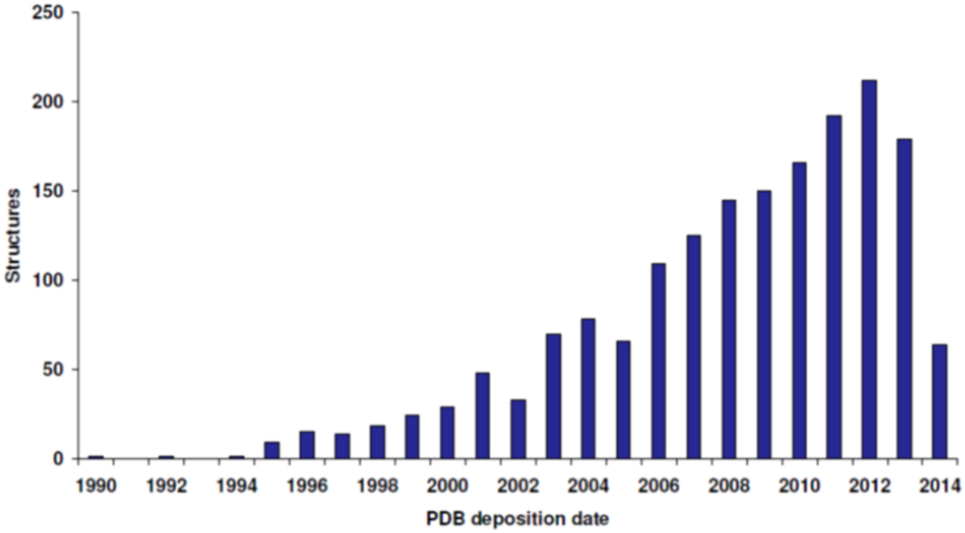


Figure 4.12: Deposition dates of the PDB structures from the LDR set. The first LDR structure was deposited in 1990. Notice that 2014 is lower because it was analyzed with data from July 2014 and there still may be proteins deposited in 2013 on hold in the PDB.

roles[43, 202, 200, 205, 170]. To further support this, Supplementary Table 2 shows that among the 33 proteins with unusual LDRs ($\geq$ 200 residues), 20 have literature defining the region to be disordered or unstructured, 4 mention disorder but not for a particular region, 5 were missing possibly to help crystallization, whilst for the remaining 4 no specific reasons are mentioned. Even the largest LDRs in x-ray structures are likely due not to specific or accidental 3 experimental conditions, but to functional disorder, yielding a dataset of the highest available quality.

The size of the dataset allows function and domain enrichments to be calculated (Figure 4.16,4.17,4.18). Intrinsically disordered regions do not have the same structural constraints as globular regions and are likely to evolve more rapidly[206], making functional domain assignments difficult at the sequence level. In Pfam (version 27.0), 5.6% of the sequence residues in the database were in predicted disordered regions of 50 amino acids or more[40]. Previous observations based on predictions showed a substantially increased number of disordered residues when domain assignment is missing[202]. In 69% (1,338/1,945) of the cases there is no Pfam domain assignment inside

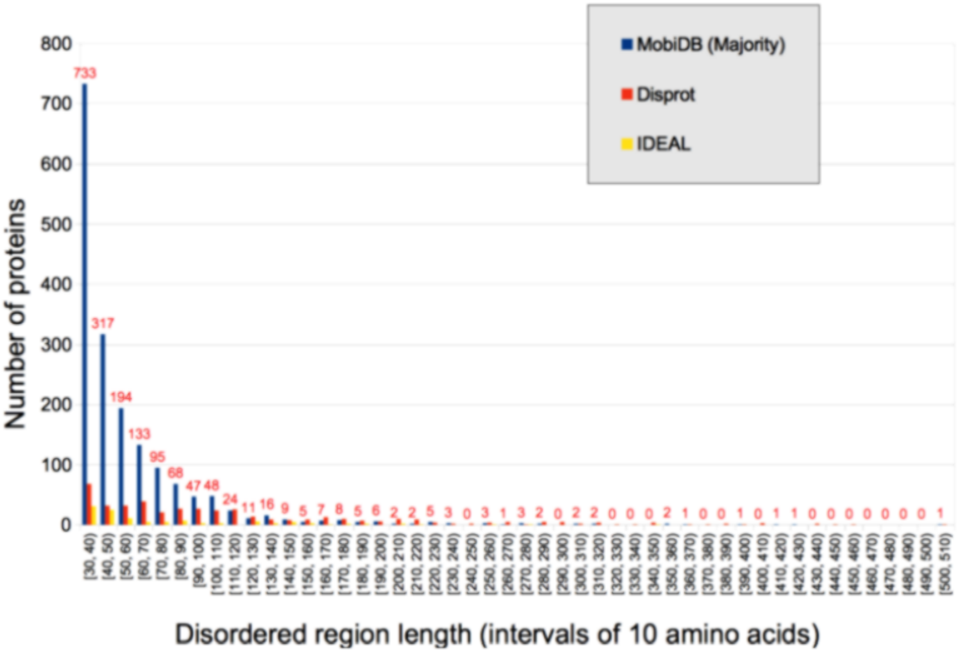


Figure 4.13: Length distribution of the disorder regions found in the LDR set and the DisProt and IDEAL databases of manually curated disorder. The data is grouped in bins of ten residues.

LDRs. This absence was recently noted for predicted residues[202], but here the ratios are confirmed on experimental LDRs. There are 611 Pfam domains located in LDRs, with no clear trend about the type. The most frequent domain associated to LDRs is the LacI domain, with only 9 hits. At the Pfam clan level (i.e.superfamilies), only a slight consensus emerges with the most frequent being helix-turn-helix (25hits).

To analyze functional enrichment, the completely structured proteins are compared to LDRs separately for the three Gene Ontology (GO) classes[207]. The three most enriched terms in molecular functions of LDR proteins are protein binding, carbohydrate derivative binding and receptor activity. The terms catalytic, oxidoreductase activity and cofactor binding, common in highly structured enzymes, are instead largely absent in LDRs. Although DNA/RNA binding is commonly considered prevalent in disorder, it is less represented in the LDR set. Proteins binding DNA or RNA are generally

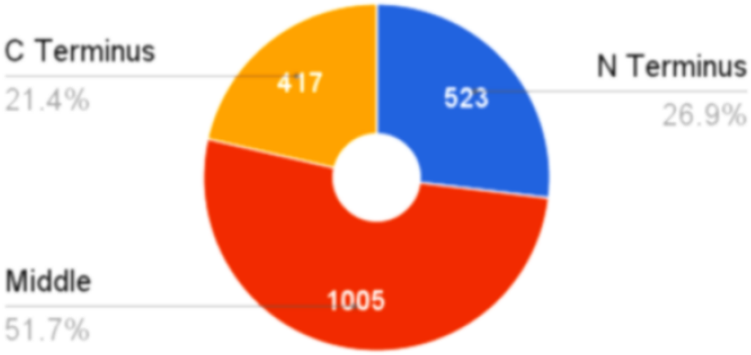


Figure 4.14: Distribution of LDRs between N- and Ctermini and the middle of the protein.

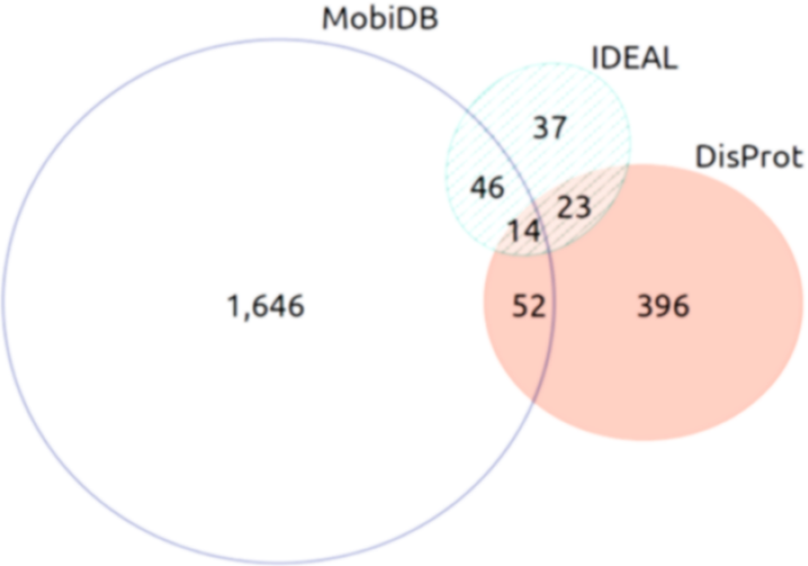


Figure 4.15: Venn diagram showing the overlap between the LDR dataset and the DisProt and IDEAL databases of manually curated disorder.

crystallized with bound nucleotides, which stabilize the folded conformation.Chromatin binding, which is a related DNA/RNA binding term, is however significantly involved in LDRs. In terms of biological processes, LDR proteins are practically absent in metabolism, but are instead enriched in biogenesis, organization and regulation. Finally, for cellular components, protein complex, membrane-enclosed lumen and its child term organelle lumen are enriched in LDRs. In summary, many GO terms previously associated with disorder have been confirmed[182, 208, 209] e.g.binding and regulation.

We found 1,758 diverse LDR protein sequences from missing backbone atoms in x-ray structures.In case of relatively short regions, the missing electron density is often a consequence of alternative conformations in highly flexible areas, whilst for very long regions it most likely corresponds to unstructured portions of the polypeptide chain. The use of x-ray crystal structure in this study deserves a specific comment, since it is empirically well known that macromolecules with long flexible parts will tend to resist crystallization. It is common practice among crystallographers to produce different constructs of the same protein in order to reduce the flexible portions and, in doing so, favor crystal growth. In this sense, we would expect that our analysis underestimates the 4 fraction of disordered regions present in the protein world. Most likely a larger fraction of disorder is present in the proteins that have not been crystallized yet. Most of the collective conclusions regarding long (and short) disordered regions have until now been based on predictions. One of the main reasons for developing computational approaches was the scarcity of experimental data to make hypotheses. Despite this, predictors have given some interesting hypotheses with respect to LDRs, such as a functional analysis in full proteomes [182] and biological processes[210].However, although predictors have good accuracy and can generate large quantities of data, they still contain systematic errors. For instance, on LDR proteins the predictor ESpritz[211] achieves 54.3% sensitivity and 91.2% specificity, while IUPred-long[172] produces a 38.9/93.6% sensitivity/specificity. While both prove a performance considerably above random, nevertheless substantial errors remain. Our experimental LDR set is also significantly different from the currently available curated databases DisProt and IDEAL.It is important to stress that our data are not simply PDB entries, but rather multiple x-ray experiments assigned to frequently multi-domain UniProt sequences. Different x-ray experiments may be assigned to the same sequence with the final disorder/structure decision based either on majority evidence or complete lack of structure. This should produce a more stable definition since it will remove noise, e.g. missing residues arising from low resolution data or not well refined crystal structures. The majority definition also includes folding upon binding events (mixtures of disorder and structure), but we found these to be a rare occurrence as shown by the comparatively small difference (ca. 10%) between both definitions. However, more work is required to generalize the domains further, for example classifying them as wobbly domains or not. For a good example of a wobbly domain see Glutamine-tRNA ligase[212] from D.radiodurans (UniProt entry P56926) in our data set. It must also be considered that the presence of a domain whose orientation is flexible with respect to the rest of the protein may have two effects. One is to drastically reduce the probability of growing suitable crystals, the other is that the position of the wobbly domain becomes artificially frozen by the crystallization process, hampering its identification in the structure. Finally, for the first time a large dataset of diverse LDR proteins are available for new prediction techniques. Training a novel predictor on this large amount of quality data using state-of-the-art machine learning algorithms can only enhance our understanding of the phenomenon. We have provided the first attempt at functionally classifying experimental LDRs. A clearer picture will emerge as more x-ray structures are deposited each year in the PDB and annotated in our LDR collection as part of MobiDB[189].

4.2.1 Methods

Long disorder data and clustering

UniProt[25] sequences with at least one x-ray annotation in MobiDB[189] were downloaded in July 2014 producing a list of 25,833 entries. Two strategies were used to define disorder. In the first majority rule, where more than one MobiDB annotation was available, a majority vote was used, with structure preferred if equal (see Figure 4.104.114.12). The second zero rule is more restrictive and considers the subset of majority rule regions where only disorder and no structure can be found. Proteins were extracted if there was at least one long disordered region ≥ 30 amino acids. Where MobiDB cannot find annotation for part of the UniProt sequence, residues are annotated as unknown. All proteins with long disorder were forced to be non-redundant using BLASTClust[213] with less than 40% pairwise sequence identity and 70% coverage. Long disordered regions of at least 30 residues from DisProt release 6.02[200] and IDEAL available in July 2014[204] were used for com-

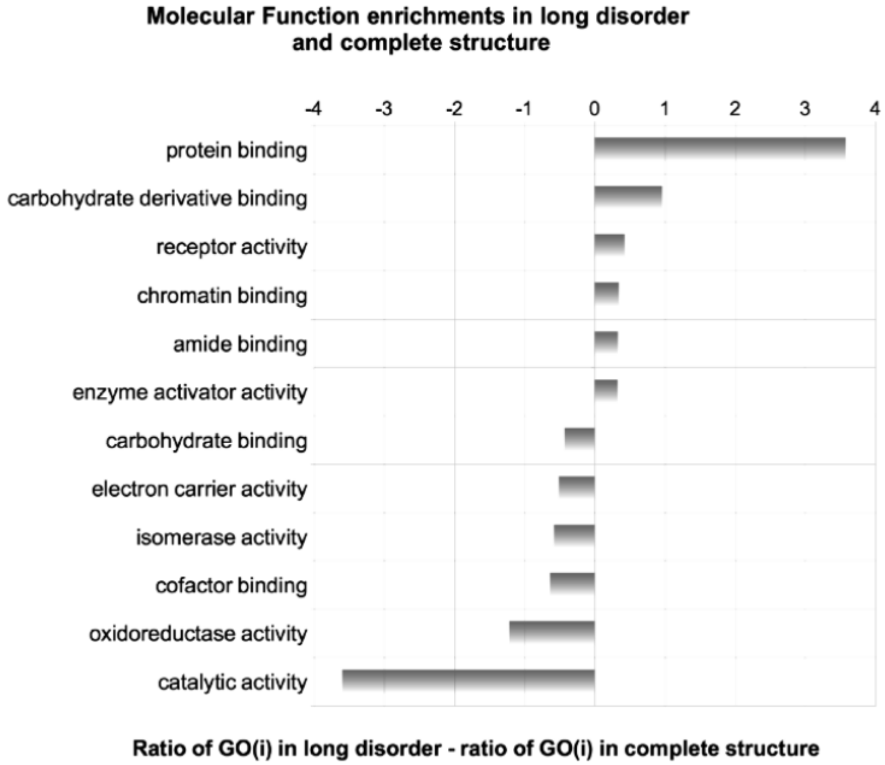


Figure 4.16: Enrichment plot for Molecular Function GO terms.

parison. The same majority rule was applied for IDEAL on their disorder definitions, with tags disorder and high_rmsd used for disordered residues. Sequences between the three sets were considered overlapping with at least 90% sequence identity and 90% coverage using BLASTClust. The data can be accessed online from URL: `http://mobidb.bio.unipd.it/long`.

Pfam and function

Only high quality Pfam-A annotations were used for domain annotations. Disordered regions having an overlap of at least 50% of residues with a Pfam domain were considered hits (see Figure 4.104.114.12 for some examples). Functional enrichment was calculated for the first two levels of the Gene Ontology (GO)[207] graph as available in March 2014. Fishers exact statistical tests were carried out for the enrichment analysis using the positive

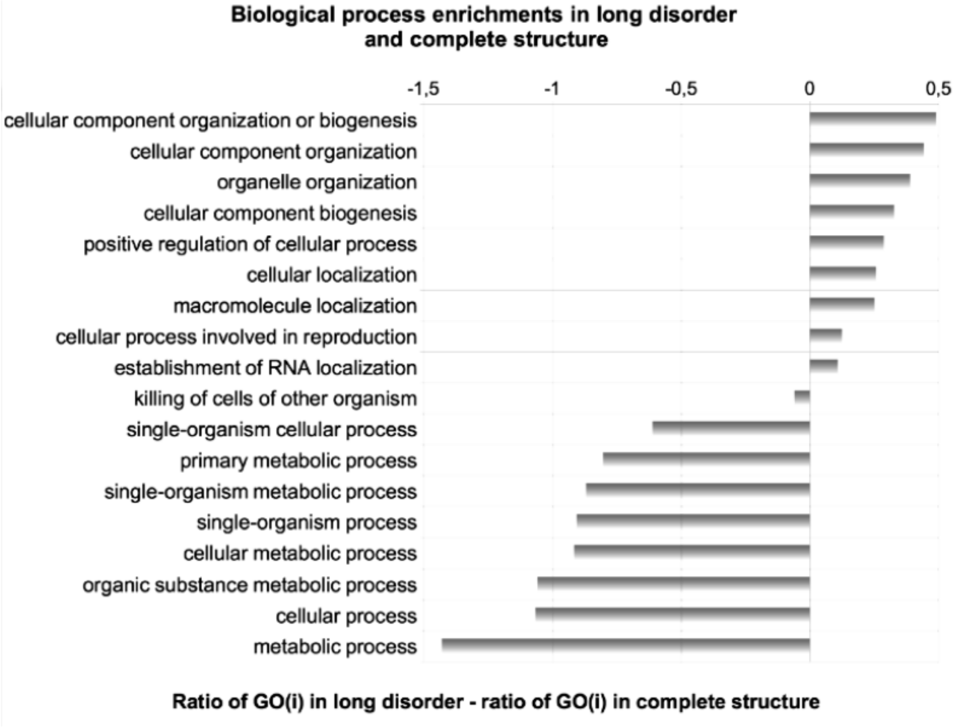


Figure 4.17: Enrichment plot for Biological Process GO terms.

set defined above and a non-redundant set of 4,391 UniProt entries with no disorder. A function was considered enriched if the p-value with Bonferroni correction was outside the 95% confidence interval of the mean.

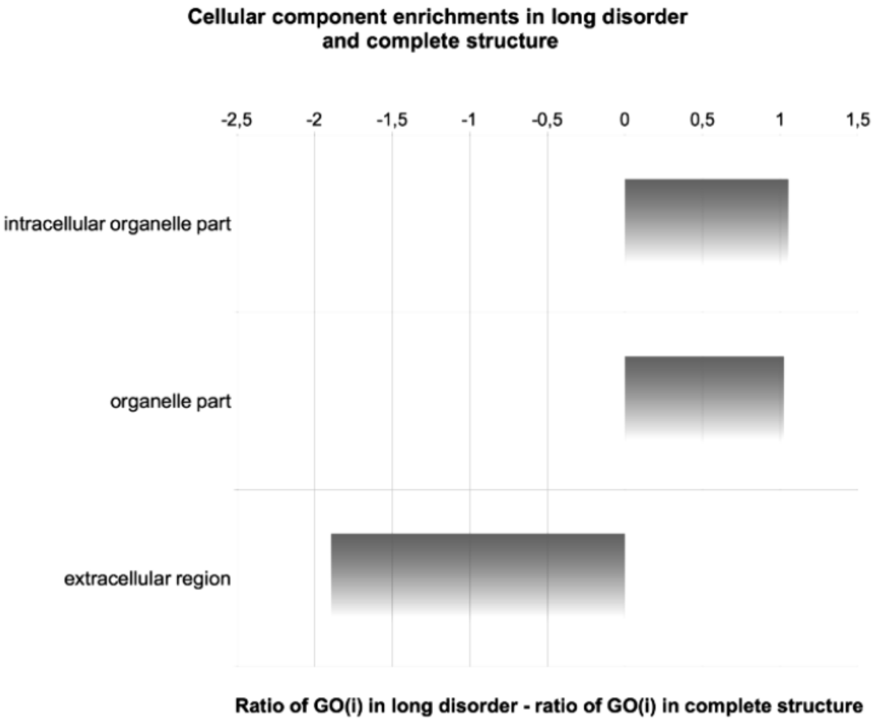


Figure 4.18: Enrichment plot for Cellular Component GO terms. Terms enriched in LDR are > 0, while terms enriched in structure are < 0. Only statistical significant enrichments (Fisher test with Bonferroni correction, p-value < 0.05).

Chapter 5

Evaluating the Impact of a Genetic Mutation

The rapid growth of unannotated missense variants poses challenges requiring novel strategies for their interpretation. From the thermodynamic point of view, amino acid changes can lead to a variation in the internal energy of a protein and induce structural rearrangements. This is of great relevance for the study of diseases. In section 5.1 I introduce NeEMO, a tool for the evaluation of stability changes using an effective representation of proteins based on residue interaction networks (RINs). RINs are used to extract useful features describing interactions of the mutant amino acid with its structural environment. Benchmarking shows NeEMO to be very effective, allowing reliable predictions in different parts of the protein such as β-strands and buried residues. The ability to predict the impact of a mutation can be very important also for disease prediction. As an example, the Critical Assessment of Genome Interpretation (CAGI) is a scientific challenge that aims to assess the state-of-the-art of genetic variants analysis for disease risk prediction. As a successful participant of CAGI, in section 5.2 I report the methodologies that were proposed to estimate Crohn's disease probability in human samples where genetic data was provided. In addition to the analysis of known disease genes, I considered variants in the context of protein-protein interaction networks. The idea of guilty by association was the main diver of this strategy: the degradation of disease pathway is probably exacerbated by an accumulation of variants in its network. Despite the good performance of the method in CAGI datasets, more research has to be done to achieve confidently clinical needs in the general case. However, there are simpler phenotypes where the genetic background is well understood. Section 5.3 introduces BOOGIE, a tool for the annotation of all known blood groups, like ABO and Rh, from DNA sequences. This has the potential to improve healthcare, improving detection of rare blood types and avoiding common blood typization errors. The simplicity of the method serves as a proof-of-principle about the potential application of genetic testing. In fact, BOOGIE can be easily extended to predict other genetically encoded phenotypic traits for personalized medicine.

Genetic factors are not sufficient to explain disease. Protein-protein interactions play a central role as well. In section 5.4 I introduce a Petri network designed to model pVHL functional pathways. The germline inactivation of this protein cause the Von Hippel-Lindau (VHL) syndrome, a condition predisposing to the development of cancer. The model was built using functional information derived from the literature, and demonstrated the ability to reproduce VHL syndrome at the molecular level. The reliability of the Petri Network model also allowed in silico knockout experiments, useful to simulate the evolution of a disease.

To conclude, there are many problems related to phenotypes and disease. With NeEMO, I tried to look at the impact of variants at the single mutation level. This works well for Mendelian traits, as shown in the context of blood groups. In the case of complex cases like Crohn's disease, protein-protein interactions has to be modeled, and Petri networks can be an effective tool. Big challenges are still unsolved in the phenotype prediction field, but the development of new computational approaches exploiting protein interaction networks seem a good direction to interpret complex disease datasets.

5.1 NeEMO: A Method Using Residue Interaction Networks to Improve Prediction of Protein Stability upon Mutation[1]

The development of Next Generation Sequencing technologies has a tremendous impact on the discovery of missense variants. In humans, dbSNP [3]

[1] The results of this chapter have been published in Giollo, M., Martin, A. J., Walsh, I., Ferrari, C., Tosatto, S. C. (2014). NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC genomics*, 15(Suppl 4), S7. For Supplementary Material, check the online version of the paper.

reports more than one million such variants, while only 1% of them have functional annotation or are referenced in the literature. This gap represents a problem for understanding disease development [214], as the proper characterization of variant effects may require expensive experiments. This is not only important for healthcare, but also for biotechnology, where alanine-scanning mutagenesis is commonly used to study the effect of amino acid variants on protein function and interactions [215]. Finally, designing mutants for protein design [216] and to evaluate their effects on function requires a deeper understanding of the mechanisms by which single variants affect stability. The Gibbs free energy (ΔG) defines the thermodynamic energy of folding compared to the denatured state. The difference between wild type and mutant polypeptide energy ($\Delta\Delta$G = ΔGwt - ΔGmut) is a measure of how the amino acid change affects protein stability. Polypeptide chains are held together by non-covalent interactions between the residues forming them. The most relevant factors affecting protein folding and stability are hydrogen bonds, van der Waals, electrostatic and hydrophobic interactions, backbone angle preferences and protein chain entropy [217]. Interestingly, the assessment of stability changes has been shown to be critical for the interpretation of variants in key proteins like TP53 [218], which is known to have a strong connection with cancer development. In order to help understand the impact of amino acid changes, the ProTherm database [44] collects the free Gibbs energy for thousands wild type and mutant proteins. This source of information is critical for the development of new methods that try to fill the gap of unannotated variants. For the last 15 years, a number of computational tools have been developed for the prediction of stability changes in mutant proteins. Energy-based methods are based on two main approaches [219]. The first type is based on the use of molecular (or quantum mechanic) force fields that try to reflect the physical energy of molecules [220] [221]. The second type, also known as knowledge-based potential functions (KBPFs), are energy functions based on statistics computed on sets of experimental or artificially generated protein structures. Most KBPFs rely on a weighted combination of several statistical terms, as in Eris [222] or FoldX [223]. In particular, the latter considers nine different terms like van-der-Waals contributions, solvation energy, hydrogen bonds and the entropy cost. All terms are linearly combined after fitting to experimental data [223].

A completely different approach is adopted by machine learning algorithms (ML). Rather than trying to explicitly describe complex models of thermodynamic energy, they are trained by minimizing the classification error on a reference dataset. A number of ML tools have been proposed for stability prediction of variants, like AutoMute [224], I-Mutant [33] [225], MuPro [226] and PoPMuSiC 2.0 [227]. Most of these simulate the change by replacing the side chain of the mutated residue, disregarding possible structural rearrangements in the backbone. As an example, I-Mutant 2.0 [33] represents variants as a vector with 42 dimensions: two for pH and temperature, 20 for encoding the wild-type and mutant residues, and 20 to describe the residue frequency in the environment surrounding the amino acid. Similarly, two versions of MuPro [226] use vectors with 140 elements to encode the residue in a sliding window that considers 3 positions on the left and right of the mutant amino acid. Both methods trained a Support Vector Machine for classification and regression purposes with the radial basis function kernel [228]. This is a general trend of ML-based approaches for stability prediction: non-linear functions are preferred due to their increased ability to detect patterns in the data, leading to better performance. In addition, all methods try to encode explicitly information about the protein of interest using either structure or sequence information. Both information can be described effectively using residue-residue interaction networks (RINs), as suggested by RING [35]. RINs are a graph description of protein structures where nodes represent amino acids and edges represent different types of physico-chemical bonds (e.g. hydrogen bonds, salt bridges, hydrophobic contacts). Using RINs can be of interest for stability estimation due to their implicit detailed representation of different chemical interactions in proteins. These interactions play a central role for the internal folding energy, so they may introduce new discriminative variables for the analysis of mutants [229]. Using this insight, in our work we trained a non-linear neural network for the prediction of stability changes based on RINs. We will show that using this effective protein representation there is an improvement in the prediction of protein stability. We believe that NeEMO can contribute significantly for the characterization of un-annotated missense variants and for protein mutagenesis studies, increasing the knowledge in this challenging field.

5.1.1 Methods

Dataset

For machine learning methods, the construction of a dataset is a critical process requiring a meticulous selection and curation of the starting data. The

ProTherm database [44] represents a reference dataset describing the effects of amino acid mutations in terms of thermodynamic energy changes, currently containing information on 647 different proteins. Roughly one third of the 22,713 entries represent the Gibbs free energy of the wild type protein, while the reminder report the ΔG of a mutant. It is clear that there is a remarkable redundancy of information that needs to be managed. Here, we decided to focus on the curated version of ProTherm used to train PoPMuSiC 2.0 [227]. In order to avoid bias, we evaluated sequence similarity on the 131 proteins of this training dataset. Using PANADA [230], clustering at 90% and 40% identical sequences produces 129 and 119 different clusters respectively. In particular, none of these clusters had more than three sequences in it. This high diversity is therefore a key factor for the machine learning procedure, as it is likely to provide an effective estimation of the data model.
This dataset is particularly informative because it corrects misinterpretations of the original papers and considers only single-site protein variants with known structures that are meaningful for mutation prediction. It should be noted that none of the variants involves either prolines or mutations that destabilise the structure by more than 5 kcal/mol, as these variants tend to alter protein folding significantly. Due to limitations of RING [35] for the management of PDB files with multiple chains, we focused on 113 proteins and 2,399 mutations. Figure 5.1 shows the training set $\Delta\Delta$G distribution, highlighting how destabilizing variants are the most frequent ones and proving that the filtering procedure preserves the correct data distribution.

To perform additional tests, we created a second dataset (IM_631) from the training data used in MuPro [226] and I-Mutant [225], containing 631 new mutations in 30 different proteins, to be used as independent samples providing indication of overfitting. The dataset distribution is quite different from the PoPMuSiC data (Supplementary Figure S1), especially in the frequency of highly destabilizing variants ($\Delta\Delta$G $>$ 5 kcal/mol). The latter dataset explicitly removed strong mutants likely to yield significant changes to the protein structure, which may represent a threat during the learning process. On the other hand, the IM_631 dataset collects real variants with no prior filtering, so these mutations can be used to evaluate NeEMO without bias. Last but not least, the S350 dataset contains further mutations which are typically used to compare the performances of different methods [227]. This data will be considered to obtain a fair comparison of NeEMO

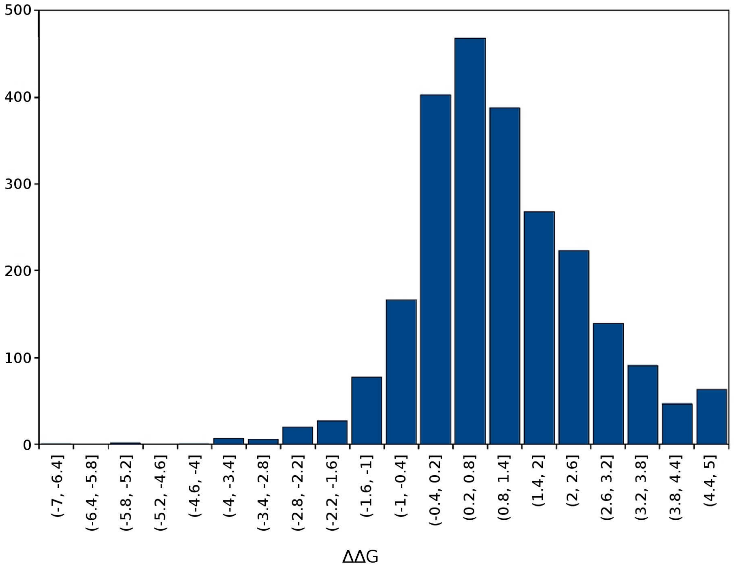


Figure 5.1: ΔΔG distribution on the training set.

performance with other stability prediction tools.

Relevant features

Our objective is to investigate how useful RINs are in the context of stability prediction. RINs are potentially interesting because they can be used to detect informative amino-acids in a target protein using standard graph algorithms like Dijkstra's shortest path or PageRank [231]. These networks have been generated by RING [35] with default parameters, i.e. closest atom networks where interactions are reported for residues that have atoms at less than 5 Å. There are four main features that we obtain with this tool, which will be briefly described in the following. For a more detailed description of the features see Supplementary Table S1.

Evolutionary information: The overall idea is that evolutionary information can discriminate key residues in the protein, either for stability or functional reasons. NeEMO considers conservation, Mutual Information and its correction Average Cluster Purity as a feature for stability prediction. These values are generated by RING, which generates a multiple sequence alignment using PSI-BLAST [232] on the UniRef90 sequence database and computes several measures reflecting evolutionary information of each residue.

Residue conformational propensities: The impact of variants strongly depends on the local environment of each residue in the structure. Classical tools for the evaluation of protein structures can highlight residues with high structural constraints that should not be mutated. In the current implementation, RING uses TAP [86], FRST [82], and QMEAN [233] to estimate the amino acid energy contribution. In particular, these tools evaluate statistical potentials such as all atom distance-dependent pairwise, torsion angle, and solvation potentials. All these numerical terms are included in NeEMO for an accurate description of the mutant context.
Amino acid information: The wild type, the mutant and its two adjacent residues in the sequence (left and right) are used to describe protein changes. One-hot encoding is used to represent the sequence information, as it was previously shown to be effective [181]. I.e. the 20 standard residues $r, i \in \{1, ..., 20\}$ are translated into a 20-dimensional vector where the i-th element is 1, and the others are 0. In addition, secondary structure and relative solvent accessibility (RSA) defined by DSSP [30] are used to describe the local context.
Network topology: Using RING it is possible to distinguish between H-bond, inter-atomic contacts, π-cation, π-π stacks, salt bridges and the atoms involved in these interactions [35]. The standard node parameters described in NetworkAnalyzer [234] are computed on that information and used to describe the mutant and its sequence neighbor (left and right) for stability prediction. Centralities are computed by considering multiple sub-network that consider a single chemical bond at a time. In addition, the network size and frequency of each amino acid type in contact with the mutation position in the RIN were also counted. Neighboring residues are defined as those which have any atom at ≤ 5 Å to any of the atoms from the other residue. The overall idea is to comprehensively assess the network connections, and measure if the mutant is central in the protein graph topology. This information was critically discriminative in previous work [225] [229], so we expect it to be also effective in the context of stability prediction.
Last but not least, pH and temperature are considered during the prediction. All information is stored in 184 dimensional vectors for each mutation. Almost half of the features are needed to describe *amino acid information*, due to the one-hot encoding sparsity with 20 descriptors for every residue.

Training

Using the encoding described in the previous section, the 2,399 examples were transformed in vectors for training a three-level neural network, with the goal to predict variant ΔΔG values. As shown in Figure 5.2, the input layer uses RIN information, a single hidden layer is used for non-linear projection of the input data, and a third level is used to estimate the mutation effect in terms of thermodynamic energy. After initial assessment, 5 hidden layer neurons were found sufficient to encode the model data, meaning that the neural network was able to detect a limited number of patterns during the training process which can effectively explain the mutations impact on stability. We used 10 fold cross-validation as implemented in WEKA [235] to estimate the method parameters, i.e. the dataset was randomly split into 10 parts, where 9 were used to train the model and the tenth used as test set. To increase the robustness of the method, 15% of the training data were used as a validation set. During model optimization, the training is stopped once the performance on the validation set does not improve for five iterations. All starting features have non-zero coefficients, so we expect them to be relevant for the final prediction. Three different neural networks we trained. NeEMO uses all 184 features. NeEMO_NOCC does not use network topology and centrality information. Finally, NeEMO_NORING uses only amino acid information, pH, temperature, conservation, QMEAN potential and protein length.

Performance measures

Several regression and classification measures are computed for a detailed comparison of NeEMO with other methods. Real value ΔΔG predictions are evaluated using standard Pearson correlation r:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where x and y represent the real and the predicted energy value. The correlation of the actual ΔΔG ranking is measured against a ranking order induced by the predictions using Kendall's tau (τ) and Spearman rank (ρ), reporting how predictors differentiate smaller stability changes from bigger ones. Both

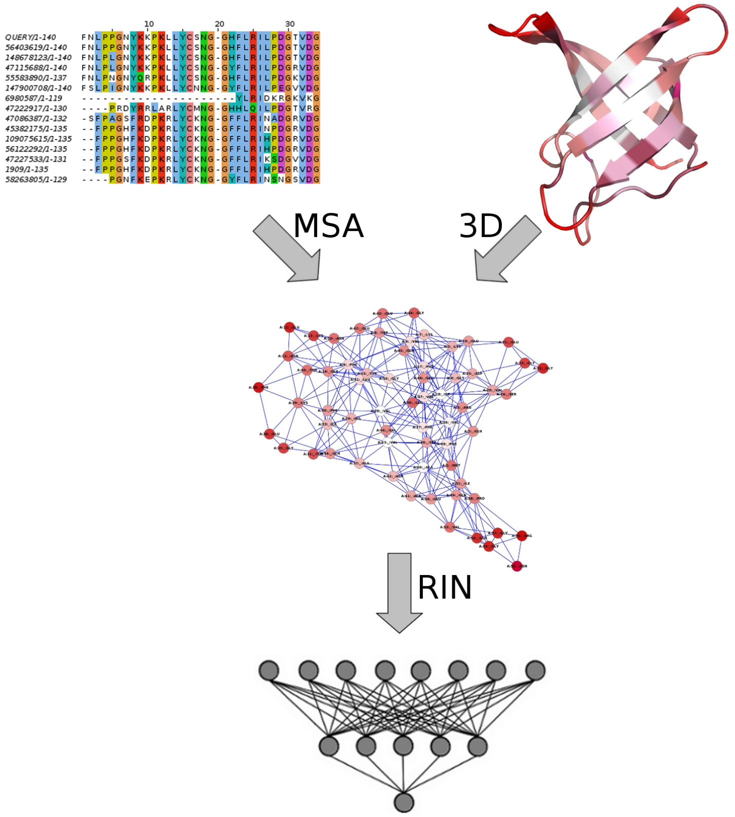


Figure 5.2: Representation of the NeEMO pipeline. The 3D structure of a protein is transformed into a RIN. Node centralities are then computed and combined with other RIN features, such as secondary structure, relative solvent accessibility and sequence conservation. All numerical descriptors are then fed into a neural network that predicts the ΔΔG for the chosen mutation.

statistics are calculated as follow:

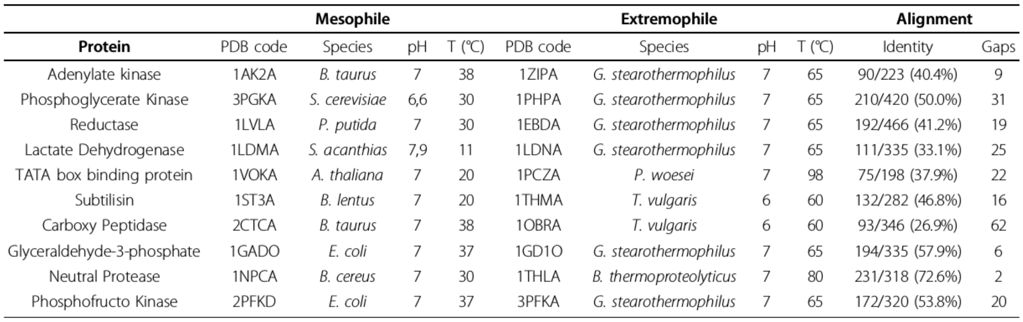
$$\rho = \frac{\sum(r(x_i) - r(\bar{x}))(r(y_i) - r(\bar{y}))}{\sqrt{\sum(r(x_i) - r(\bar{x}))^2 \sum(r(y_i) - r(\bar{y}))^2}} \qquad \tau = \frac{CP - DP}{0.5n(n-1)}$$

where ρ is identical to Pearson correlation applied to the rank of the predictions, while τ accounts for the number of prediction pairs having correct order

(CP) or wrong order (DP) with respect to the real ΔΔG for the n dataset examples. Finally, the standard error σ is used to report the expected distance of the prediction from the real ΔΔG of the mutation.

Termophile case study

As an additional test, we consider ten pairs of mesophilic proteins with their termophilic counterparts presented in [236]. In order to compare the stability changes with NeEMO, each sequence pair was first aligned using the Needleman-Wunsch algorithm from the EMBOSS package [237]. The NeEMO energy was then calculated for each aligned residue pair from the mesophilic to thermophilic mutation (MtoT) and vice versa (TtoM). Table 5.1 lists the 10 pairs of protein structures, their similarity and the pH and temperature at which the ΔΔG was predicted.



	Mesophile				**Extremophile**				**Alignment**	
Protein	PDB code	Species	pH	T (℃)	PDB code	Species	pH	T (℃)	Identity	Gaps
Adenylate kinase	1AK2A	*B. taurus*	7	38	1ZIPA	*G. stearothermophilus*	7	65	90/223 (40.4%)	9
Phosphoglycerate Kinase	3PGKA	*S. cerevisiae*	6,6	30	1PHPA	*G. stearothermophilus*	7	65	210/420 (50.0%)	31
Reductase	1LVLA	*P. putida*	7	30	1EBDA	*G. stearothermophilus*	7	65	192/466 (41.2%)	19
Lactate Dehydrogenase	1LDMA	*S. acanthias*	7,9	11	1LDNA	*G. stearothermophilus*	7	65	111/335 (33.1%)	25
TATA box binding protein	1VOKA	*A. thaliana*	7	20	1PCZA	*P. woesei*	7	98	75/198 (37.9%)	22
Subtilisin	1ST3A	*B. lentus*	7	20	1THMA	*T. vulgaris*	6	60	132/282 (46.8%)	16
Carboxy Peptidase	2CTCA	*B. taurus*	7	38	1OBRA	*T. vulgaris*	6	60	93/346 (26.9%)	62
Glyceraldehyde-3-phosphate	1GADO	*E. coli*	7	37	1GD1O	*G. stearothermophilus*	7	65	194/335 (57.9%)	6
Neutral Protease	1NPCA	*B. cereus*	7	30	1THLA	*B. thermoproteolyticus*	7	80	231/318 (72.6%)	2
Phosphofructo Kinase	2PFKD	*E. coli*	7	37	3PFKA	*G. stearothermophilus*	7	65	172/320 (53.8%)	20

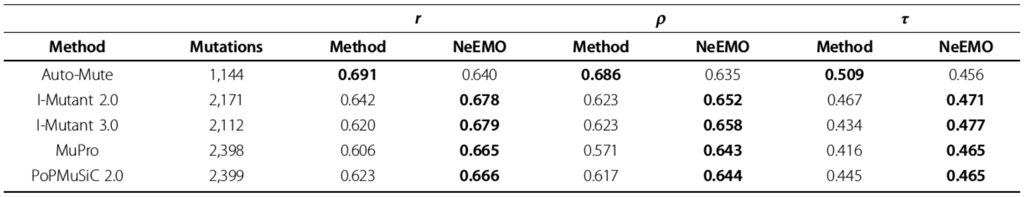
Table 5.1: Summary of the 10 pairs of mesophilic and thermophilic proteins used in the case study, their similarity and the environmental conditions (pH and Temperature) used to perform the test [236].

5.1.2 Results

We developed NeEMO, a machine learning method that uses RIN information, to evaluate the impact of amino acid changes in protein stability. Using a curated ProTherm dataset, 10-fold cross validation was used for training and performance evaluation. Finally, the tool is tested on two independent sets of protein variants, providing an unbiased evaluation of its reliability and a fair comparison with other methods.

Training and cross-validation

NeEMO was trained on a large dataset previously used by PoPMuSiC 2.0 [227]. The results of 10-fold cross validation on this dataset are shown together with a preliminary comparison to other methods in Table 5.2. Our goal was to assess if the features and the mathematical model of our method are able to fit effectively into the traning data. Several state-of-the-art methods were used, namely Auto-Mute, I-Mutant 2.0 and 3.0, MuPro and PoPMuSiC 2.0. The comparison was not straightforward, as most predictors were occasionally not able to make a prediction for some variants due to their inability to manage certain PDB files. We decided to compare NeEMO only on the mutations where all tools were executed successfully. In many cases the variants of this test set are part of the training dataset of other methods. For this reason, this performance comparison cannot be considered unbiased, and therefore it is just a mean to measure if the fitting procedure is as good as the one used on other methods. As shown in Table 2, NeEMO performs consistently well compared to other state-of-the-art tools. Auto-Mute is the only method providing comparable results, but seems very poor in the input and mutation management, as the method cannot make a reliable prediction for half of the examples (e.g. NMR solved proteins, or in case if atoms with repeated coordinate sets). In view of the good performance in the cross-validation, we expect that the fitting process was overall good.



		r		*ρ*		*τ*	
Method	**Mutations**	**Method**	**NeEMO**	**Method**	**NeEMO**	**Method**	**NeEMO**
Auto-Mute	1,144	**0.691**	0.640	**0.686**	0.635	**0.509**	0.456
I-Mutant 2.0	2,171	0.642	**0.678**	0.623	**0.652**	0.467	**0.471**
I-Mutant 3.0	2,112	0.620	**0.679**	0.623	**0.658**	0.434	**0.477**
MuPro	2,398	0.606	**0.665**	0.571	**0.643**	0.416	**0.465**
PoPMuSiC 2.0	2,399	0.623	**0.666**	0.617	**0.644**	0.445	**0.465**

Table 5.2: Regression performance comparison of NeEMO with other methods on the ten-fold cross-validation test. The evaluation is performed only on mutations where both methods were able to make a prediction. In addition, the other methods are likely to have used the test samples in their training.

Interestingly, it seems that NeEMO performs particularly well for amino-acids on β strands (Supplementary Table S3). This improvement is of particular interest, as it suggests that our method can capture and model accurately long range interactions that typically occur in these secondary structures. In addition, performance on buried residues and on coils (see Supplementary Tables S4 and S6) indicate that the method performs very well compared to other methods, confirming that network topology contributes significantly to a proper description of the local amino acid context. On the other hand, NeEMO performance for α helices and exposed residues (Supplementary Tables S2 and S5) are comparable to other methods. As a results, we believe that the training process was successful, suggesting that chosen features and neural networks are a good model of the data.

NeEMO in-depth analysis

In order to test the contribution of the 184 mutation descriptors, we compare the performance of NeEMO, NeEMO_NOCC and NeEMO_NORING on the IM_631 dataset. As can be seen in Figure 5.3, NeEMO regression has a steep slope that confirms the effectiveness of the training. NeEMO_NOCC and NeEMO_NORING decrease performance (Supplementary Figure S2), showing larger errors for mutations producing a higher stability increase. As expected, the quality of the $\Delta\Delta$G estimation decreases when less information is provided, suggesting the need of RIN data for good predictions.

To study how performance varies for mutations in different conditions, we divided the cross-validation test set into subsets containing only mutations in each of the three secondary structure states (α, β, coil) and computed the class-specific performance (see Table 5.3). While mutations on α helices show a similar performance compared to the entire dataset, larger differences are found for mutations in β-strands and coils. The correlation for mutations in β-strands is much higher than for any other subset, while performance on mutations occurring in coils shows the lowest results. This was expected, because coil residues tend to be on the surface of globular proteins. Having a tendency towards mobility, they are believed to be regions where unfolding begins. Increased coil mobility facilitates solvent exposure, leading to a reduced number of interactions and hence a lower contribution in RINs. Table 3 shows a similar result for solvent exposed (E, RSA $> 25\%$) and buried (B, RSA $\leq 25\%$) mutations. In this case, despite NeEMO working better on buried mutations, the difference is less marked. This suggests that secondary structure context is probably the most important feature for stability prediction upon mutations. Last but not least, use of RING network information significantly improves prediction quality in all experiments. On

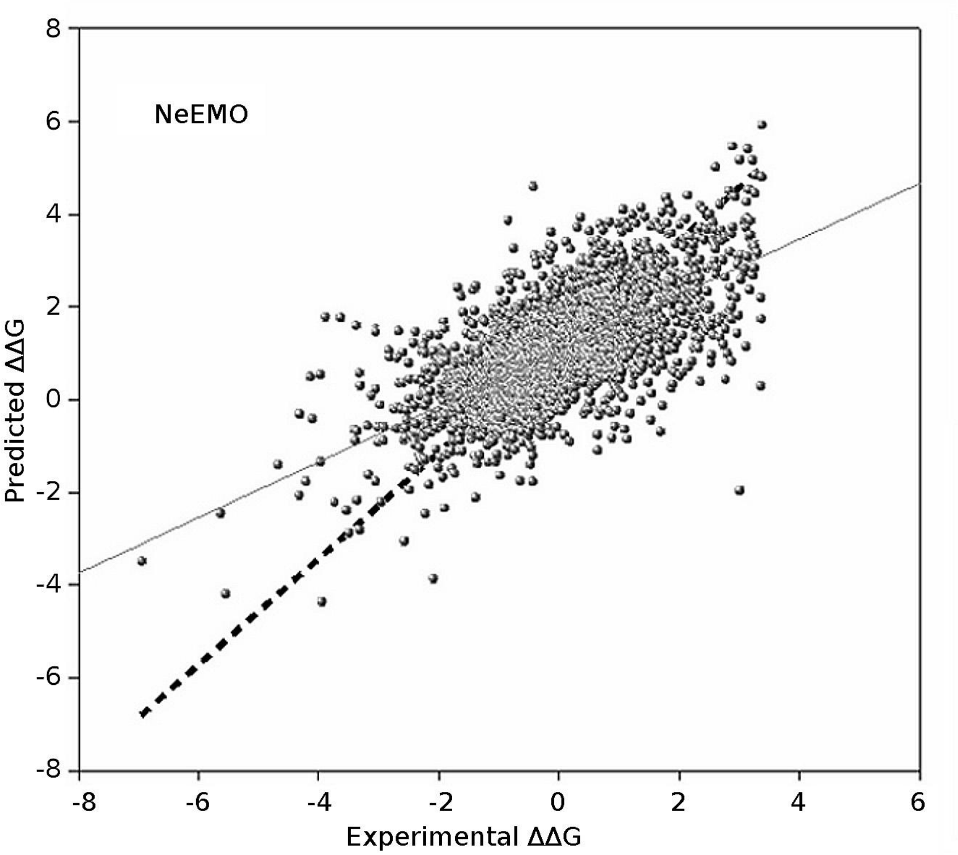
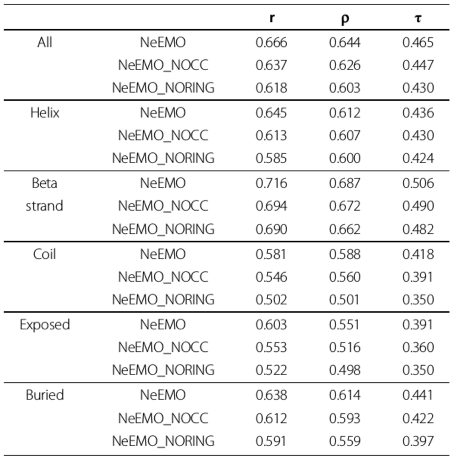


Figure 5.3: Regression results of NeEMO versions on the training set.

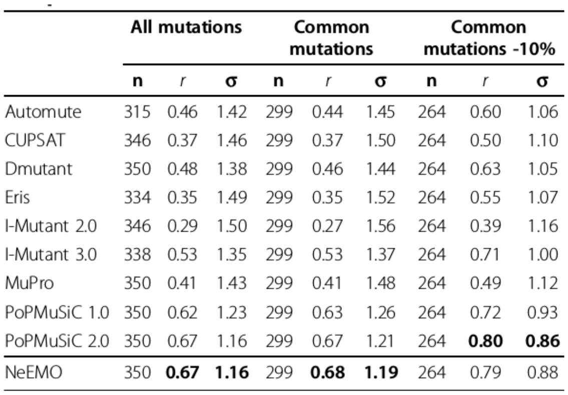
the IM_631 dataset the r, ρ and τ correlations are 0.63, 0.60 and 0.43 respectively. Considering how the dataset contains unseen mutations, a small drop in performance is expected. It shows that there is no overfitting, and that our network features describe the effect on stability of single amino acid variants. We expect that NeEMO can also perform well in other datasets with very different proteins.

Comparison with other methods

We compare the performance of NeEMO with several state-of-the-art methods, namely MuPro [226], two versions of I-Mutant [33], [225], PoPMuSiC



		r	ρ	τ
All	NeEMO	0.666	0.644	0.465
	NeEMO_NOCC	0.637	0.626	0.447
	NeEMO_NORING	0.618	0.603	0.430
Helix	NeEMO	0.645	0.612	0.436
	NeEMO_NOCC	0.613	0.607	0.430
	NeEMO_NORING	0.585	0.600	0.424
Beta strand	NeEMO	0.716	0.687	0.506
	NeEMO_NOCC	0.694	0.672	0.490
	NeEMO_NORING	0.690	0.662	0.482
Coil	NeEMO	0.581	0.588	0.418
	NeEMO_NOCC	0.546	0.560	0.391
	NeEMO_NORING	0.502	0.501	0.350
Exposed	NeEMO	0.603	0.551	0.391
	NeEMO_NOCC	0.553	0.516	0.360
	NeEMO_NORING	0.522	0.498	0.350
Buried	NeEMO	0.638	0.614	0.441
	NeEMO_NOCC	0.612	0.593	0.422
	NeEMO_NORING	0.591	0.559	0.397



	All mutations			**Common mutations**			**Common mutations -10%**		
	n	*r*	**σ**	**n**	*r*	**σ**	**n**	*r*	**σ**
Automute	315	0.46	1.42	299	0.44	1.45	264	0.60	1.06
CUPSAT	346	0.37	1.46	299	0.37	1.50	264	0.50	1.10
Dmutant	350	0.48	1.38	299	0.46	1.44	264	0.63	1.05
Eris	334	0.35	1.49	299	0.35	1.52	264	0.55	1.07
I-Mutant 2.0	346	0.29	1.50	299	0.27	1.56	264	0.39	1.16
I-Mutant 3.0	338	0.53	1.35	299	0.53	1.37	264	0.71	1.00
MuPro	350	0.41	1.43	299	0.41	1.48	264	0.49	1.12
PoPMuSiC 1.0	350	0.62	1.23	299	0.63	1.26	264	0.72	0.93
PoPMuSiC 2.0	350	0.67	1.16	299	0.67	1.21	264	**0.80**	**0.86**
NeEMO	350	**0.67**	**1.16**	299	**0.68**	**1.19**	264	0.79	0.88

Table 5.3: Left: Correlation measure performance of different NeEMO versions on the IM_631 dataset. NeEMO uses all input features, NeEMO_NOCC does not use node centralities, NeEMO_NORING does not use any RIN feature. Comparisons are shown for the entire dataset, on each of the 3 different secondary structure states and occurring in amino acids exposed to the solvent (e, RSA $>$ 25%) or buried (b, RSA $\leq$ 25%). Right: Performance of different methods on the independent S350 dataset. The comparison is reported (a) for all the mutations in the dataset, (b) the maximal subset of mutations where each tool is able to make a prediction and (c) the maximal subset where 10% of outliers are removed. The number of mutations (n) is shown together with the Pearson correlation (r) and distance from the real $\Delta\Delta$G values (σ). The best prediction in each column is shown in bold.

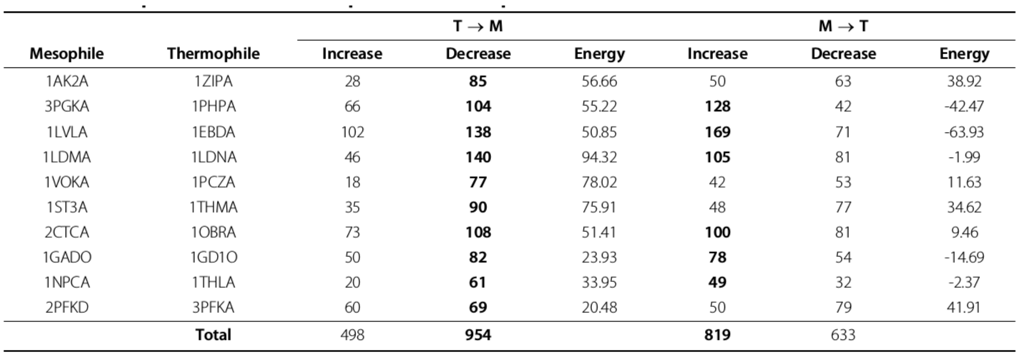
1.0 and 2.0 [238][227], Automute [224], Eris [239], CUPSAT [240] and Dmutant [241]. In order to provide an unbiased evaluation of effectiveness, the S350 dataset [227] was used, as it contains mutations unseen to each method. NeEMO was re-trained in order to exclude examples that overlap with the training set. Performance in Table 5.3, are reported (a) for all the mutations that a single tool can evaluate, (b) for the maximal set of 299 mutations where all predictors are able to make a prediction and (c) for the maximal dataset where we additionally remove 10% of the outliers (leading to 264 mutations). In the latter dataset, outliers are selected automatically for each method as those having the largest residuals in the regression of predicted-observed $\Delta\Delta$G values. As can be seen, performance clearly suggests that

NeEMO is able to outperform most methods, proving the validity of the training strategy and the strong impact of using residue-residue interaction network data as a tool to study the mutation impact on protein stability. PopMusic2.0 is the only tool with comparable performance, but the unbiased cross validation correlation shown in Table 2 suggests that NeEMO is considerably better on a larger set. The comparison is suitable as both methods trained on exactly the same dataset, so it should give a fair comparison of the predictors. It is also interesting to note that the NeEMO performance is basically the same in both the S350 and cross validation sets, while PoPMusic2.0 has a drop in performance. NeEMO is overall reliable and shows a very good performance on different structure types, like β strands or buried residues (data not shown). RINs seem a clear contribution for the $\Delta\Delta$G prediction of variants, and NeEMO can be useful for variant annotation.

Termophile analysis

Effective stability predictors can be used to investigate aspects of biology ranging from protein design to organism evolution. As a proof of principle for NeEMO, we analyzed ten proteins from mesophilic organisms and the correspondent homologs in thermophilic organisms [236]. The simple hypothesis to test is that variants found in termophilic proteins increase stability, while mesophilic variants have the opposite effect. Performing these experiments is complicated by the presence of insertions and deletions in the amino acid sequences which cannot be easily interpreted. NeEMO was used to predict the stability changes upon termophile to mesophile (T→M) and mesophile to termophile (M→T) for each alignable residue. As shown in Table 5.4, the results are encouraging. In 66% of T→M variants our simple hypothesis seems confirmed (53% of exposed and 76% of buried positions), leading to an expected stability decrease. Overall, the sum of predicted $\Delta\Delta$G also confirms the mutant tendency to reduce stability.

In the M→T dataset, the expected change in folding energy is not as marked, but there is still an interesting signal. For 6 of the 10 proteins there is a majority of variants predicted to increase stability. This is also confirmed in the sum of predicted $\Delta\Delta$G, where 56% of the mutations support the hypothesis of increased stability. Surprisingly, 68% of exposed positions seem to reduce protein stability, while just 44% of buried residues increase stability. This is in contrast with the T→M dataset, and could be due to the highly



		T → M			M → T		
Mesophile	**Thermophile**	**Increase**	**Decrease**	**Energy**	**Increase**	**Decrease**	**Energy**
1AK2A	1ZIPA	28	**85**	56.66	50	63	38.92
3PGKA	1PHPA	66	**104**	55.22	**128**	42	-42.47
1LVLA	1EBDA	102	**138**	50.85	**169**	71	-63.93
1LDMA	1LDNA	46	**140**	94.32	**105**	81	-1.99
1VOKA	1PCZA	18	**77**	78.02	42	53	11.63
1ST3A	1THMA	35	**90**	75.91	48	77	34.62
2CTCA	1OBRA	73	**108**	51.41	**100**	81	9.46
1GADO	1GD1O	50	**82**	23.93	**78**	54	-14.69
1NPCA	1THLA	20	**61**	33.95	**49**	32	-2.37
2PFKD	3PFKA	60	**69**	20.48	50	79	41.91
	Total	498	**954**		**819**	633	

Table 5.4: NeEMO predictions on the mesophilic and thermophilic mutations. Amount of reciprocal variants in mesophilc and thermophilic predicted to increase or decrease the stability of the 10 proteins, and their cumulative energy. Cases where predictions support our hypothesis of symmetric ΔΔG behavior of variants are highlighted in bold.

divergent structures of some proteins. The well predicted Phosphoglycerate Kinase (Figure 5.4) shows little divergence in the two PDBs. In contrast, protein pairs with unclear support for our hypothesis tend to have divergent 3D structures. Overall, NeEMO seems to be useful in this proof of principle, evaluating a simple hypothesis on stability change in termophiles. Although a more thorough investigation will be necessary to confirm the generality of these observations, it nevertheless provides evidence that NeEMO can be used to prioritize mutagenesis experiments and may be used to support protein design studies.

Web server

The NeEMO web server is freely available to the scientific community from URL: `http://protein.bio.unipd.it/neemo/`. Once a PDB file is specified by the user, the service computes the RIN in a few minutes, and provides a user-friendly interface for variant prediction. Multiple amino acid changes can be tested at a time, including different pH and temperature parameters. The tool is also very fast. Once the multiple alignment is computed, the effect of a residue change on the protein structure can be predicted in few seconds, making it scalable for large-scale usage.

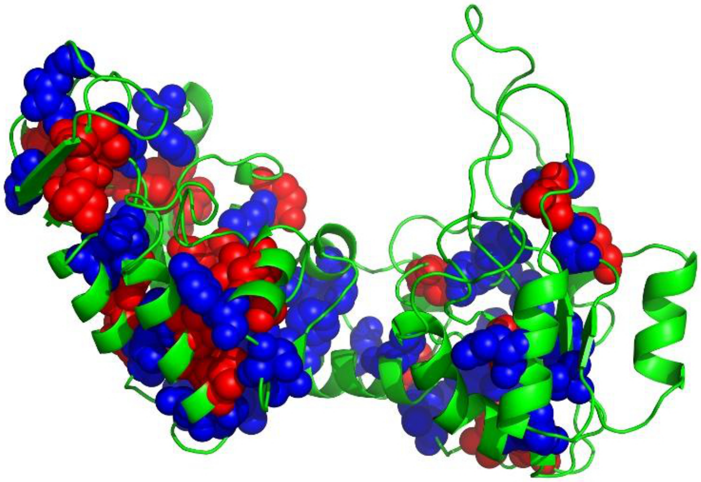


Figure 5.4: 3D structure of the 3PGKA, showing well predicted buried residues (blue) and mispredictions (red). For this protein, the mesophile and thermophile core amino acids share a similar structure.

5.1.3 Conclusions

NeEMO represents a novel approach to predict ΔΔG changes after point mutations in protein structures. It takes advantage of RINs created by our previous work RING [35] to describe protein structures and interactions between the amino acids forming them. In RINs each residue is described by several features, including secondary structure, solvent accessibility, conservation and a number of residue-specific energy potentials. RING also provides detailed information about interactions found between different amino acids, including their occurrence and types. The interactions present in the RIN are used to compute node centralities that encode the relevance of each RIN node in a protein structure. Inclusion of RINs and information derived from them was shown to improve mutation stability prediction performance. Overall, NeEMO seems able to significantly outperform all other tested methods, and shows very good accuracy across different secondary structures and in classification. It also seems good in terms of reliability, as it can manage and produce a prediction for nearly all PDB files of the PoPMuSiC 2.0 dataset. For the near future, we are planning to extend NeEMO to map multiple chains directly into an integrated RIN.

Another advantage of our approach is that it does not rely on 3D models for the mutant proteins. Instead the RIN for the wild type protein is used to predict the stability change. Other methods have to model the mutant structure first, which may be computationally expensive and in some cases can introduce errors that our protocol avoids. In addition, RINs are very comprehensive data structures that help the management of heterogeneous information sources like evolutionary and topological data. In fact, experiments show that network data improves prediction quality without exception. Finally, it is interesting to note that the evaluation on unseen examples in IM_631 results in basically unchanged performance. This is a nice result, because the $\Delta\Delta$G distribution of the training data was significantly different. The overall results also prove no overfitting was introduced in NeEMO, and confirm that it can be used effectively for the assessment of mutation impact. As the number of known variants and PDB structures in different organisms is rapidly increasing, we believe that the tool can be important for variant assessment. Finally, NeEMO can also play a role for pathogenicity prediction as shown in [242]. It is well known that stability loss in proteins like TP53 [218] is associated with disease development. The ability of RINs to describe proteins and their variants effectively can play a role for the detection of deleterious protein changes, and may also contribute to pathogenicity prediction.

5.2 The Critical Assessment of Genome Interpretation[2]

Over the last few years, Next-Generation Sequencing (NGS) has provided insights and generated knowledge about the role of single nucleotide variants (SNVs) in common and Mendelian diseases. Exploiting this NGS data can contribute significantly to clinical practice. The Critical Assessment of Genome Interpretation (CAGI) is a scientific challenge that aims to evaluate the state-of-the-art in terms of disease prediction. Similarly to CASP, it can be divided in three moments. In the *prediction season*, the participants can make blind predictions on novel datasets. In the *assessment season*, all submissions are evaluated against the ground truth. Finally, results are shown

[2] All figures are results are taken from CAGI online presentations (URL: `genomeinterpretation.org`).

during a *conference* to help discussion.
The Crohn's disease (CD), Hypoalphalipoproteinemia (HA) and Familial Combined Hyperlipidemia (FCH) datasets in CAGI are clear examples where entire families were sequenced, and specialized bioinformatics algorithms can play a key role for disease risk prediction. In these challenges, we used a two stage approach. First, we clustered family members (when unknown), and used these family groups to generate a highly personalized disease likelihood assessment. In the second step, we collected all SNVs relevant for disease pathways and used them to predict risk. To select SNVs several features were used, including: conservation, interaction network centrality and expression data. We did not stick to a predetermined set of mutations suggested by past Genome-Wide Association Studies, since their limited power could not handle the weak within-family genotype variability. Conversely, using SNVs in relevant loci and in the interaction network it is easier to capture even the smallest variance in the SNVs distribution, thus allowing a more robust evaluation of pathogenic mutations in family members. Broadly speaking, population stratification was used directly to identify highly penetrant variants in NGS data, while the use of candidate loci improved significantly the accuracy of SNP selection. Finally, the disease risk score is predicted from the SNP overall count.

5.2.1 Datasets

CAGI proposed ten challenges in 2013 related to genotype interpretation. Given their high heterogeneity, we decided to focus on exome analysis with the goal of predicting genetic disease risk. There were four datasets of this type, namely the *Crohn's disease (CD)*, the *Familial combined hyperlipidemia (FCH)*, the *Hypoalphalipoproteinemia (HA)* and the *Personal Genome Project (PGP)*:

Crohn's disease in a dataset of 66 exomes, 15 samples are healthy and rest are Crohn's patient. Given the VCF file containing the variants, participants had to predicting the disease probability for each sample. Data was sampled from 28 different German pedigrees, and included one monozygous discordant twin pair.

Familial combined hyperlipidemia (FCH) the dataset contained 5 exomes in VCF format from a family, where the mother and 2 daughters had high level of LDL cholesterol (LDL-C), while the father and one daughter were healthy. There were three questions: (a) detect the mutations causing high LDL-C, (b) predict the individual with abnormal triglycerides (TG) and HDL cholesterol (HDL-C), and (c) predict TG and HDL-C values for each family member. For each challenge, a real value and its standard deviation were required.

Hypoalphalipoproteinemia (HA) predict the probability of hypoalphalipoproteinemia disease given the 4 exomes of a familiar group (VCF format).

Personal Genome Project starting from 291 medical profiles publicly available and the genome of 77 anonymous individuals, the goal was matching the two information automatically. A total of 214 medical profiles were given as decoys, to increase the challenge difficulty. For each profile, the matching probability with respect to all genomes had to predicted.

In this thesis we will focus on the CD dataset, as it contained the larger sample size from a well designed study. Further considerations about other challenges will be given at the end of next section.

5.2.2 Results

Crohn's disease is a heterogeneous syndrome, whose genetic causes are still largely unknown. Such complexity requires a highly specific analysis for each individual of the challenge, due to different familiar background of participants. In this work, we addressed our analysis at a family-wise level. Using an implementation of agglomerative clustering, we uncovered hidden relationships between participants, obtaining a family dendrogram. This was particularly important, especially for the identification of discordant twins and outliers (see Fig. 5.5). Clustering was repeated multiple times according to bootstrap concept and including different SNVs (either common variants or rare variants) during the analysis. Clustering also proved that few mutations are peculiar of each family (black blocks in Fig. 5.5), while samples with few shared SNVs are outliers.

We decided to use CAGI 2011 Crohn's dataset to detect interesting mutations or patterns that may be shared in the CAGI 2013 dataset. The most striking result was the high similarity among few patient exomes in these

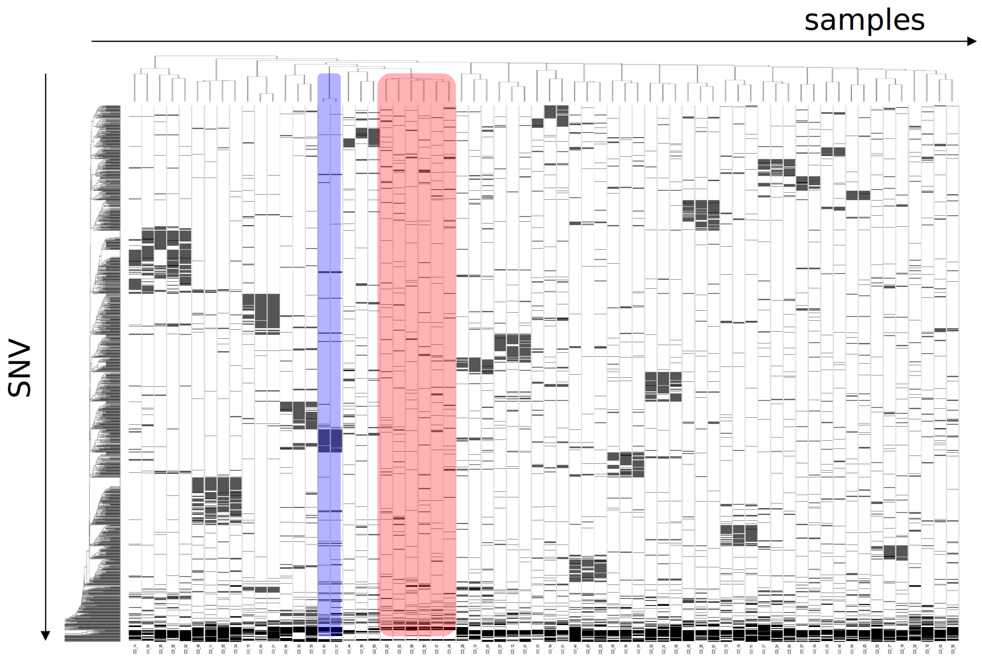


Figure 5.5: Heatmap of CAGI samples generated from genetic data. Families are automatically detected by means of agglomerative clustering. In blue, twins are correctly shown as the most similar patients. In red, few of the control samples grouped together as outliers.

challenges, which seems relevant for the identification of healthy individuals. In fact, outliers in Fig. 5.5) are strongly related to CAGI 2011 Crohn's challenge controls. According to knowledge transfer principle, we decided to annotate as "healthy" these samples.
As a second step, we tried to detect important SNVs for disease risk assessment in all samples. Variants annotation and filtering were performed using ANNOVAR [243]. We discarded from our analysis SNVs with allele frequency > 2%, low conservation score in mammals, and reported in dbSNP v132 [3]. Also, SNVs in low entropy regions were removed, along with the ones observed in more than 15 samples in the CD dataset. This process was expected to select just rare variants in critical regions of the genome. We also considered SIFT pathogenicity prediction [29] to better rank mutations. Finally, we adopted a *filter*, *expand* and *match* approach for disease risk evaluation (Fig. 5.6).

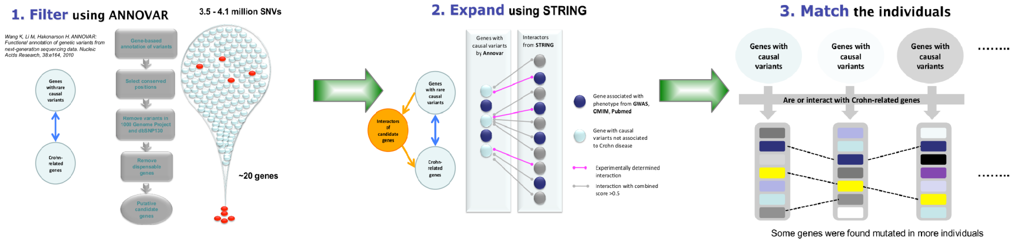

1. Filter using ANNOVAR
3.5 - 4.1 million SNVs
Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Research, 38:e164, 2010
Gene-basaed annotation of variants
Select conserved positions
Remove variants in 1000 Genome Project and dbSNP130
Remove dispensable genes
Putative candidate genes
Genes with rare causal variants
Crohn-related genes
~20 genes
2. Expand using STRING
Genes with rare causal variants
Interactors of candidate genes
Crohn-related genes
Genes with causal variants by Annovar
Interactors from STRING
Gene associated with phenotype from GWAS, OMIM, Pubmed
Gene with causal variants not associated to Crohn disease
Experimentally determined interaction
Interaction with combined score >0.5
3. Match the individuals
Genes with causal variants
Genes with causal variants
Genes with causal variants
....
Are or interact with Crohn-related genes
....
Some genes were found mutated in more individuals

Figure 5.6: Disease risk prioritization process.

Filter We selected SNVs placed in genes associated with CD. The main source of information were literature reviews, OMIM [244] and genome wide association study (GWAS) databases [4]. It is important to note that GWAS databases report just SNVs and their disease association p-values. These variants were kept into account in our analysis. In addition, we assumed that any other mutation "close" to the associated SNVs (in the same gene) could play a role for disease prioritization.

Expand The heritability expected in complex disease is still not completely explained by genetics. This is due to the huge statistical challenges related to genetic studies, like noise and confounders not taken into account by researchers. Protein-protein interaction network knowledge is typically exploited in pathway-based association methods, like *burden tests*. These approaches look for SNV enrichment within a group of genes involved in a biological process [46]. Borrowing this idea, besides the variants in CD genes we collected SNVs in their direct interactors in STRING PPI network [6]. This database uses different strategies to report interacting genes, like experimentally validated interactions, co-expression in tissue and text mining. In addition, a reliability score for each interaction is provided. In our selection process, text mining edges of the graph were removed. To further increase the quality of interactors, we focused on genes having interaction quality ≥ 0.5x. To avoid irrelevant genes, we selected just the interactors expressed in colon, small intestine or stomach (the main CD tissue) [245]. Finally, we used a consensus of two disease gene prioritization tools (ToppGene [41] and Endeavour [246]) to select the top-ranked (≤ 150) genes that are expected to play a role based on CD features.

Match After the previous filtering, we ended up with 113 variants in CD genes and 44 among their interactors. In both cases, SIFT classified roughly half of them as pathogenic (respectively 57 and 26). We used this data (a) to separate cases from controls and (b) to assign a disease risk score. By construction, healthy individuals were given a disease risk ranging from 0 to 49, which was generated from the amount of SNVs (scaled appropriately). Controls instead were given a disease risk scaled between 50 and 100 using the mutation count once again. To predict a *disease score*, we assigned manually a weight to each mutation category. SNVs associated to CD in GWAS were given higher score (3 in case of allele frequency $\geq 5\%$, 6 otherwise). Other variants in CD genes were given a weight of 1 or 2, depending if they were predicted to be pathogenic by SIFT. Finally, SNVs placed in interacting genes and predicted to be pathogenic had weight 1. Ideally, these coefficients should be estimated from real datasets, but we had no training data. Thus, parameter values just reflected our prior understanding of the problem.

Healthy individual selection was done semi-automatically. Eight of them were detected using CAGI 2011 data. The twin with less mutations was predicted to be healthy (exploiting the prior knowledge of discordant phenotype). All of these decisions resulted to be correct in the final evaluation stage. To detect other control individual, we tried two strategies:

Clustering using bi-clustering in the original data, we separated samples. Surprisingly, this ended up with a first group of 15 individuals containing the outliers of Fig. 5.5 (assumed to be healthy) and a second cluster with 51 samples (labeled as cases). This was shown to be one of the best submission in CAGI (see Fig. 5.7).

Family based assessment we assumed that no more than a healthy individual was placed into a familiar group. We selected the individual with minimal *disease score* for each pedigree. Among them, we classified as control the ones with overall lower *disease score*. Table 5.5 shows the performance gain of using *all SNVs* or removing variants in the PPI genes (*no network*).

5.2.3 Discussion

Overall, Table 5.5 shows that clustering is very effective both in terms of classification and risk value prediction. In fact, there are just four misclassified samples, and the predicted *disease score* has a remarkable AUC value.

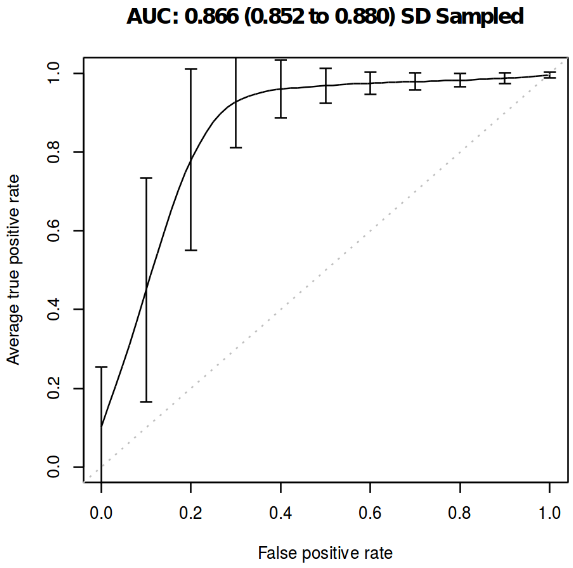
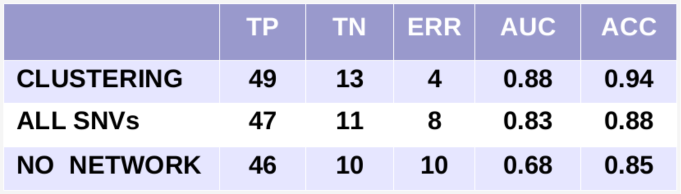


Figure 5.7: Prediction ROC curve for the Crohn's disease dataset. Vertical bars show the variance obtained from bootstrapping. The average AUC equals 0.88.



	TP	TN	ERR	AUC	ACC
CLUSTERING	**49**	**13**	**4**	**0.88**	**0.94**
ALL SNVs	**47**	**11**	**8**	**0.83**	**0.88**
NO NETWORK	**46**	**10**	**10**	**0.68**	**0.85**

Table 5.5: Comparison of three submission in terms of true positives, true negatives, misclassification, AUC and accuray. In the best scenario (clustering), just four samples were misclassified. When no clustering data is taken into account, performance decrease. Interestingly, mutations on interacting proteins seems associated with disease onset.

The *family based assessment* also proves that by just looking at the *disease score* one still obtain very competitive AUC, even though with an increased number of misclassifications. Categorization is however a harder problem, since the slightest change in the disease score (e.g. change of coefficients) has strong impact in terms of separation. This is why AUC is a better quality measure for this challenge, as pointed out by the CAGI organizers. What is very important from Table 5.5 is the contribution of variants gathered from interacting genes: when such SNVs are removed, we can observe a dramatic decrease of AUC value. This suggest that the PPI mutations contribute effectively for disease risk prioritization. We adopted a similar approach also in CAGI 2011 Chron's disease challenge, where we achieved very high performance compared to other participants. Likewise CAGI 2013, we exploited clustering and a disease score to classify patients.

We also performed well in the familial combined hyperlipidemia challenge, where we detected the causing mutation in the LDLR gene. In the HA and PGP we achieved mild results, showing that there is room for improvement. In general, none of the participants did well on each challenge, and many decisions required the human intervention. CAGI in fact is posing research questions far from being solved, so any idea could be fundamental in terms of future computational methods. Genotype to phenotype research is still a flourishing topic which is getting more attention due to the potential medical applications. We expect that the availability of additional public data will help the tuning of better tools which could be truly helpful for the scientific community. Finally, it is important to note that we are still in the game stage, where there are strong bias and high variability in terms of results. The development of next-generation sequencing is the basis of CAGI, and we hope that future editions will establish gold standard methodologies for genotype data interpretation.

5.3 BOOGIE: Predicting blood groups from high throughput sequencing data.[3]

Advances in genome sequencing due to high throughput sequencing (HTS) over the last years have detected a huge amount of new Single Nucleotide Variants (SNVs) [247], producing a tremendous growth of variation databases. Finding genotype-phenotype correlations is a critical topic in personalized medicine, as personal genome sequencing is expected to become increasingly common over the next few years [248]. One of the more interesting develop-

[3] An early version of this work appeared in Giollo, M., Minervini, G., Scalzotto, M., Leonardi, E., Ferrari, C., Tosatto, S. C. (2012). In silico blood genotyping from exome sequencing data. *AIMM*.

ments in this field is the Personal Genome Project (PGP), collecting genome sequences and clinical phenotypes of participants who have signed an informed consent with the goal to make genome information freely available for research for thousands of participants [249],[250]. Using this data, a number of methods have been used to predict particular phenotypes thought to be genetically determined, among which blood groups may provide a good test case.

The first discovery of blood groups dates back to the early years of the 20th century, and represent an important moment for medicine. Karl Landsteiners ABO system was nevertheless just a first step in the characterization of blood. Forty years later the Rh was shown to play an important role, and today dozens of different blood systems have been reported [47]. As an example, the Dombrock group [251] is known to be relevant from the clinical point of view, as it can lead to severe hemolytic transfusion reactions or hemolytic disease of the newborn. Similar problems can also be caused by many other blood systems, like Lan [252], Gerebich [253] and Junior [254]. Compared to the 25 groups reported in 2000 [251], public databases currently report more than 30 blood systems [47], so it can be thought that in the close future new ones will be detected. All blood groups are determined by the presence of specific proteins on the surface of red blood cells and body fluids [255]. Their expression is fully genetically determined and cannot vary during a persons life span. Agglutination tests have been used extensively for the identification of the ABO and Rh groups, with an error rate below 1 out of 250,000 tests [256]. On the other hand, it has been reported that non-ABO groups are among the main cause of death after blood transfusion [257], mainly due to systems like Scianna [258] where serological tests are not accurate. Sequencing techniques have been widely used as the main tool for the identification of molecular differences among systems [254],[259],[260],[261]. Several opinions suggest to introduce this kind of experiment as an additional blood typing test [262],[263]. Interestingly, HTS technologies have led to the widespread availability of sequencing in the medical and scientific community [264], making genetic tests more and more common for diagnostic purposes. It can be expected that genotype information will be useful for blood typing. Personal genome sequencing is expected to become increasingly used over the next few years [248], as it can predict patient response to drugs or diseases [265], and help in the diagnosis and cure. The BLOODchip ® system [266] is a first example of a commercial solution using genotype data to detect blood types, showing that modern sequencing techniques can be used for the identification of six different blood groups. Even though this is an appealing idea, a number of issues must still be solved, since the three billion human nucleotides are difficult to manage [267]. In addition, experimental difficulties related to sequencing and population variability [268] led to the versioning of the reference genome, and the requirement of an accurate management of SNVs linked to old genome definitions (i.e. hg18 or lesser). Last but not least, the one mutation-one phenotype paradigm used in many public databases [3],[269] is clearly unable to explain many traits of clinical relevance like the ABO or Rh blood systems, where multiple co-occurring variants determine the blood group [261] [259]. The situation is finally complicated by heterozygous variants, as they require inference of the correct patient haplotype [270]. The Rh trait is just an example with good genotype knowledge and a complicated basis, since it is encoded by two different genes resulting in the two proteins RhD and RhCE [271]. The former is the determinant of the most common Rh antigen while the latter is responsible for a large part of weak inter medium Rh traits. Patients are routinely typed for D antigen and both the common terms Rh positive and Rh negative refer to the presence or absence of this antigen. The antigens C, c, E, e coded by the RhCE gene [272] are typed routinely only in patients which have developed an atypical immunological response to long-term transfusions. Furthermore, Rh is characterized by the presence of numerous hybrids resulting from genetic rearrangement of both the RhD and RhCE genes [271]. These hybrids exhibit their condition with weak serological response to the routine test. This may result in a dangerous misclassification due to both false positive and negative blood typing [273]. To complicate this scenario further, 50 other Rh system antigens are known. While these are not well studied and poorly understood, such genetic complexity well explains the importance of large human variant databases for blood groups such as BGMUT [47].

In order to improve the quality of blood typing we present BOOGIE, a method that predicts blood groups by means of a user-defined knowledge base of Boolean rules. Starting from HTS data, the tool solves the haplotype phasing problem using information stored in the user database, and uses the same data to infer the most likely blood type. BOOGIE was tested on PGP data for the prediction of 30 blood groups and their traits reported in BGMUT. This is novel, as it shows that a set of Boolean rules can predict real phenotypes, which may be complex and present a large degree of heterogeneity. Therefore, a broad range of applications can be conceived, while the increasing amount of genetic studies will significantly enhance the power of its inference engine.

Results

BOOGIE is designed to predict phenotypes using HTS data using explicit tables that describe the correlation of SNVs with traits. These are extracted from the BGMUT database [47] which stores information about experimentally validated variants known to be relevant for determination of currently 34 different blood groups. The predictor is built to infer the closest haplotype to the known variant combinations, implicitly producing a haplotype phase. BOOGIE blood group predictions thus represent the most probable scenario of how the patient variants interact to yield a given blood group phenotype. In the following we will describe how the predictor works with the ABO group example, followed by validation on public PGP data for the ABO and Rh groups and finally presenting the distribution of predicted rarer blood groups.

ABO case study

In order to clarify how the algorithm works, we describe the example of ABO prediction for PGP sample hu604D39 (see Figure 5.8), which is known to have an AB blood group. In haplotype phasing, homozygous variants are easy to manage, as they are known to appear in both chromatids. On the other hand, heterozygous ABO group SNVs can be distributed among 32 distinct allele configurations, which may produce very different traits during the final decision. Ranking the configurations by variant similarity shows that there are 3 optimal choices. All three infer a first chromatid expressing the Ax02 blood group, while the second should express either B101 or Bw19 (see Table 5.6). In particular, the B group allele shifts all heterozygous mutations towards a single chromatid in the best scoring configuration. This fact is well known in the literature [274] and confirms the validity of our decision strategy. Finally, the 13 SNVs in Figure 5.8 are fully annotated in BGMUT, yielding high confidence for our prediction. It is also interesting to note that the database contains just 99 SNVs for ABO blood group characterization. Our focus on this small set of variants seems to describe effectively even the most uncommon ABO blood groups. This is important for the efficiency as using just few decision variables rather than the full ABO sequence decreases prediction time.

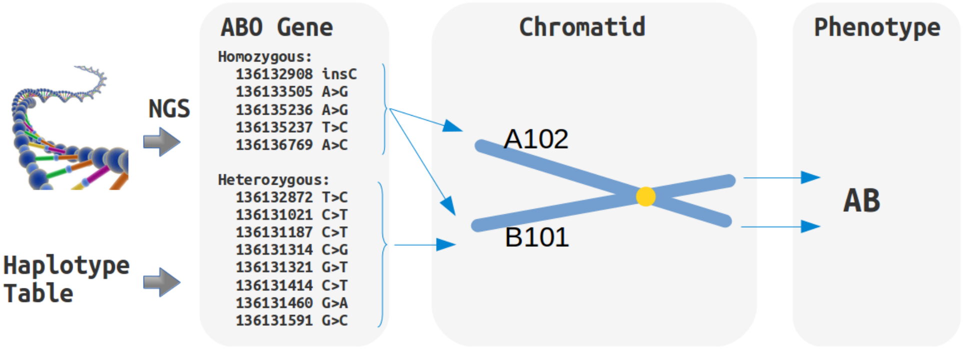


Figure 5.8: ABO gene exonic mutations. Given the input variants, BOOGIE select all the key mutations specified in the Haplotype table for ABO group that can play a role (in chromosome 9). The tool assigns all heterozygous SNVs to the same chromatid, as this represents the most likely haplotype. As result, the corresponding proteins will express either A102 or B101 blood group. Genetic analysis suggests that both antigens will be present in sample hu604D39, which is confirmed by a serological test. It should be noted that the SNVs reported use hg19 as reference, thus differing from BGMUT definition of A and B blood groups. In addition, not all the variants reported are necessary for the final phenotype, as few of them are common for a number of different groups.

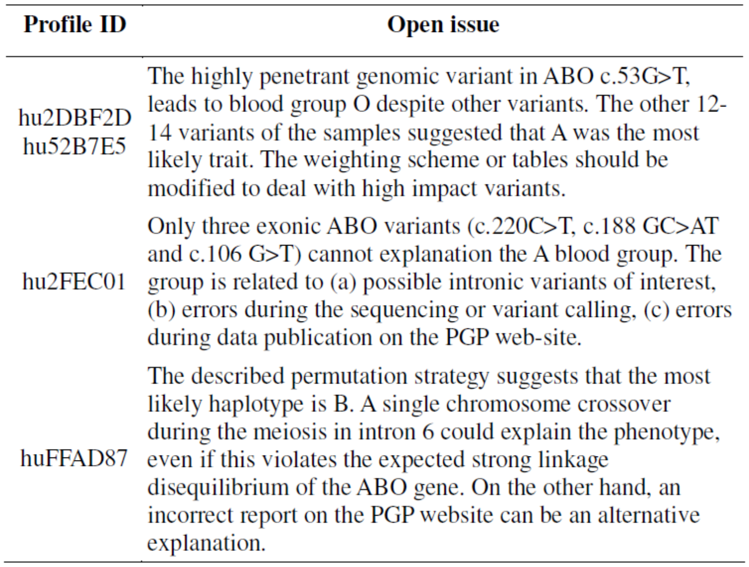
Phenotype	Chr9:136132908	Chr9:136131650	Chr9:136131414
A101	GG	C	G
A102	GG	T	G
O02	G	C	G
B101	GG	C	A

Table 5.6: Sample haplotype table extracted for the ABO system.

ABO performance

We tested BOOGIE on the PGP full genome dataset using ABO data from BGMUT. The overall accuracy is 94.2%, recognizing correctly 25/27 cases for A, 11/11 for B, 1/1 for AB and 30/32 for O group. In light of the generality of the method this is a very good result. Being able to identify the correct traits without systematic preference suggests that there is enough data in the literature to pave the way towards genetic tests for the ABO blood group. In addition, weak antigens and subgroups were identified with no effort. This is clearly an advantage compared to common serological tests and may also yield good immunological results, due to the increased ability to classify blood diversity. Regarding the amount of SNVs contained in HTS data, we observed between 4 and 18 SNVs in the genome dataset samples.Despite this large heterogeneity, BOOGIE correctly assigned heterozygous variants to the chromatid and obtained good recognition performance. It is worth to note that B typed samples typically show a higher number of exonic mutations, as also shown in past studies [274]. Interestingly, the few misclassification cases can be explained (see Table 5.7). In particular, profiles hu2DBF2D and hu52B7E5 were not detected as O. This is due to the ABO c.53G>T mutation, which has complete penetrance for the determination of the O group [274]. Even if this position is reported in our ABO haplotype table, our scoring system uses the same weight for all SNVs. In these three profiles, identification of the A haplotype, which was supported by 12-14 variants, was predicted to be the more likely. Of course, this could be solved from the technical point, either by adding weights to the mutations or entries to the haplotype table. Conversely, samples huFFAD87 and hu2FEC01 are quite unexpected, and can be explained by the relevance of intronic variants, and occurrence of a chromosome crossover during meiosis in the 6th intron respectively. This latter scenario is quite strange, and violates our implicit assumption of linkage disequilibrium (LD) in the coding region. It is possible that wrong variant calling or experimental errors are present, or the reported PGP participant data may simply not be correct. Three important loci for A, B or O blood group determination are in the minus strand of chromosome 9. The hg19 chromosome positions of interest are reported in the first row. The first one is a frameshift insertion, while the latter two are SNVs. With this information, we can distinguish between the B, O and two A group variants as well.

On the 23andMe dataset, BOOGIE has 91.43% accuracy. It should be noted that we removed 6 cases from the starting pool, because the experimental data had missing values for important positions. We correctly recalled 46/51 for group O, 44/44 for A, 18/21 for B and 4/5 for AB. The problem at hand is very different from the first one, due to the amount of observed SNVs (ranging from 12 to 43)and data quality. In fact, it seems that a major issue for this data is identification of indel c.261delG, which is missing a in



Profile ID	Open issue
hu2DBF2D hu52B7E5	The highly penetrant genomic variant in ABO c.53G>T, leads to blood group O despite other variants. The other 12-14 variants of the samples suggested that A was the most likely trait. The weighting scheme or tables should be modified to deal with high impact variants.
hu2FEC01	Only three exonic ABO variants (c.220C>T, c.188 GC>AT and c.106 G>T) cannot explanation the A blood group. The group is related to (a) possible intronic variants of interest, (b) errors during the sequencing or variant calling, (c) errors during data publication on the PGP web-site.
huFFAD87	The described permutation strategy suggests that the most likely haplotype is B. A single chromosome crossover during the meiosis in intron 6 could explain the phenotype, even if this violates the expected strong linkage disequilibrium of the ABO gene. On the other hand, an incorrect report on the PGP website can be an alternative explanation.

Table 5.7: Explanation for the ABO blood group mispredictions. In two cases the solution requires an improved weighting scheme, while in the latter two there is no definitive answer.

few cases, and possibly wrong in others. This would explain most of the misclassifications for this validation set. It is interesting to note that having only localized SNVs rather than complete exome data does not significantly affect performance. In fact, the ABO group is mainly determined by an accumulation of mutations related to known haplotypes. This confirms how using few localized SNVs can be effective in ABO prediction, as also assumed in previous genome-wide association studies [275]. This may have important implications for the development of commercial genetic tests for the ABO blood group, as the test of localized hot-spots can be significantly cheaper than whole exome sequencing.

Rh performance

On the full genome data, BOOGIE has 94.2% accuracy. We recalled correctly 57/57 for Rh+ and 8/12 for Rh-. This is a good result, considering that we marked profile huC14AE1 as misclassified, where the Rh+ and Rh- scores were the same due to uncertainty in the prediction. Profile huFE71F3 and huC30901 are quite hard to explain, because they have just two SNVs in genomic coordinates 25617282 and 25634204 of chromosome 1, which seem of no impact according to the literature. A similar situation happens for profile hu025CEA.

It is interesting to note that individuals with few SNVs in the RhD gene are likely to show an Rh+ trait. In fact, almost 80% of the samples have less than five variants. Conversely, a large number of variations is related to the D-CE-D RhD hybrid group leading to a Rh- trait. CE 5-9DBT and CE 7-8-9DIVb are just two of the groups that we observed [276]. This result cannot be obtained with classical serological tests and may be relevant for highly specific blood transfusions.

On the 23andMe dataset, there are 93 Rh+ and 18 Rh- samples and our method cannot discriminate between different RhD chromatid configurations in many situations. This is mostly due to the small number of SNVs reported in 23andMe experiments which leads BOOGIE to predict Rh+ for all cases. E.g., five samples (hu3B89BD, hu8B4E43, hu25BD97, huB2C416, huF7E042) report no mutation in RhD but are marked as Rh-. On the other hand, it is well known in the literature that this trait is connected to few mutations. These 23andMe samples are clear examples where scarcity of information makes correct trait prediction based on genotype data impossible.

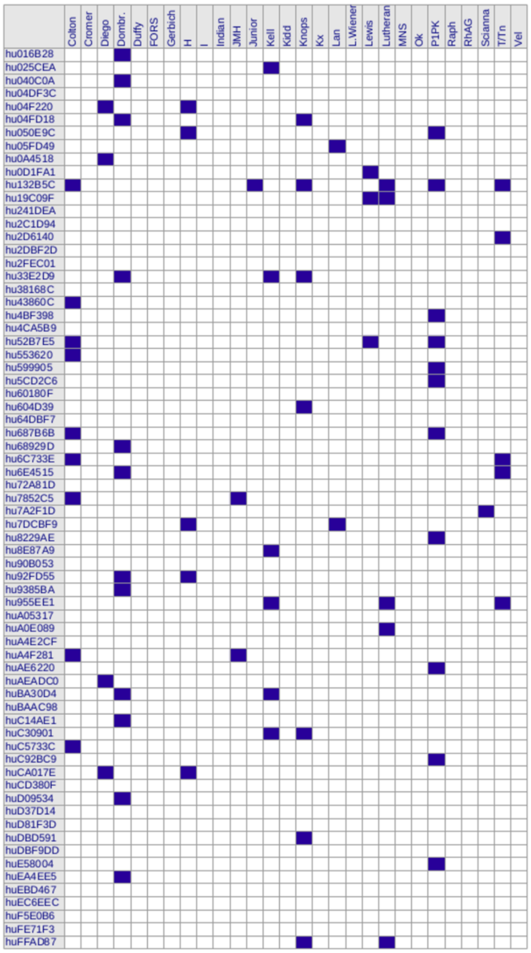
The coordinates reported by 23andMe experiments (c.329T>C, c.676G>C, c.712G>A, c.787G>A, c.933C>A) seem to be designed for the recognition of RhD-RhCE hybrids, but the existence of minor Rh groups (like CE 5 Va 4) makes the recognition of the correct allele configuration impossible. These positions may distinguish 32 distinct situations, while there are 67 known subgroups for Rh, leading to a non-universal assignment which cannot produce reliable results. A possible solution for this ambiguity may be to use population frequency of traits to rank tied scores. This could effectively suggest the maximum likelihood estimate in such a weakly informed context.

Other blood groups

We tested the PGP genome dataset for the other 28 blood groups with BOOGIE, with striking results. As shown in Figure 5.9, a considerable amount of samples has uncommon traits, which may be important during blood transfusions or in related situations. For example, in the Junior system we reported the case of a weak trait. This could be relevant for hemolytic disease of the fetus and newborn, as already reported in previous studies [254]. For the Lewis and John Milton Hagen systems we observed samples with opposite traits. This is not likely to be of interest from the clinical point. Nevertheless, these groups need more study in order to completely characterize their medical relevance. It should also be noted that in the PGP web-site, no information is available for the 28 blood groups, so no validation is possible. More in general, experimental characterization of these minor blood-groups may currently be non trivial due to the lack of an unambiguous clinical test [263]. Routine typization is conducted with genetic analysis in patients who manifest an atypical immunological response, such as long-term transfusion patients. Methods such as BOOGIE may help to address the right strategy to adopt. Clearly, the use of highly specific annotations can be considered an additional tool to predict blood transfusion compatibility. On the other hand, it is interesting to note that HTS data can provide detailed blood group information compared to common serological tests. In fact, testing only highly specific SNVs may result in good performance, as already proven in the Rh and ABO group tests.

Amount of non-considered mutations

Genetic variations are in many situations not relevant for the determination of a phenotype, as can be clearly seen in dbSNP. Very few mutations report a clinical relevance tag or reference to PubMed. In order to understand how much is known in the context of blood groups, we report the amount of SNVs in BGMUT and in dbSNP for each blood system, and use these two parameters to measure the completeness of our haplotype tables. Half the blood groups considered in this work seem to cover at least 20% of the SNVs in dbSNP (see Figure 5.10). At first glance this may seem a mixed result, but we are not interested in full coverage. We reasonably assume that (a) a small set of variants can be used as representative of a full haplotype, thanks to linkage disequilibrium and (b) a large amount of SNVs are unlikely to cause phenotypic changes. For the latter point, data from BGMUT is frequently updated, and represents the state of art for blood group characterization, so it is clear that we are using as much information as possible. As shown for the ABO and Rh systems, we argue that BOOGIE can be a valuable tool for phenotype characterization. The haplotype phase quality is also interesting and confirms that crossover recombination hot-spots are typically localized in non-coding regions [277]. This is also observed in HapMap data, where linkage disequilibrium is often observed in coding DNA. It is important in



	Colton	Cromer	Diego	Dombr.	Duffy	FORS	Gerbich	H	I	Indian	JMH	Junior	Kell	Kidd	Knops	Kx	Lan	L.Wiener	Lewis	Lutheran	MNS	Ok	P1PK	Raph	RhAG	Scianna	T/Tn	Vel
hu016B28				●																								
hu025CEA													●															
hu040C0A				●																								
hu04DF3C																												
hu04F220			●					●																				
hu04FD18				●											●													
hu050E9C								●															●					
hu05FD49																	●											
hu0A4518			●																									
hu0D1FA1																			●									
hu132B5C	●											●			●					●			●				●	
hu19C09F																			●	●								
hu241DEA																												
hu2C1D94																												
hu2D6140																											●	
hu2DBF2D																												
hu2FEC01																												
hu33E2D9				●									●		●													
hu38168C																												
hu43860C	●																											
hu4BF398																							●					
hu4CA5B9																												
hu52B7E5	●																		●				●					
hu553620	●																											
hu599905																							●					
hu5CD2C6																							●					
hu60180F																												
hu604D39															●													
hu64DBF7																												
hu687B6B	●																						●					
hu68929D				●																								
hu6C733E	●																										●	
hu6E4515				●																							●	
hu72A81D																												
hu7852C5	●										●																	
hu7A2F1D																										●		
hu7DCBF9								●									●											
hu8229AE																							●					
hu8E87A9													●															
hu90B053																												
hu92FD55				●				●																				
hu9385BA				●																								
hu955EE1													●							●							●	
huA05317																												
huA0E089																				●								
huA4E2CF																												
huA4F281	●										●																	
huAE6220																							●					
huAEADC0			●																									
huBA30D4				●									●															
huBAAC98																												
huC14AE1				●																								
huC30901													●		●													
huC5733C	●																											
huC92BC9																							●					
huCA017E			●					●																				
huCD380F																												
huD09534				●																								
huD37D14																												
huD81F3D																												
huDBD591															●													
huDBF9DD																												
huE58004																							●					
huEA4EE5				●																								
huEBD467																												
huEC6EEC																												
huF5E0B6																												
huFE71F3																												
huFFAD87															●					●								

Figure 5.9: Predicted blood group distribution for the PGP full genome dataset. For each PGP sample, the possible occurrence of uncommon blood groups is highlighted, based on coding variants. The prediction is based on the observation of SNVs known to be associated with uncommon blood groups. When no known variant is found, the phenotype is assumed to be the reference one. Clearly, the existence of non-coding or new uncharacterized variants relevant for a blood system can influence BOOGIE, leading to a false negative prediction.

our decision as we focus only on SNVs reported in our haplotype tables, providing greater flexibility. We can either report the full gene sequence and known allele, or just describe the most relevant loci. In both situations, genes in strong linkage disequilibrium are very likely to work very well with our algorithm, as shown with ABO group tests. We are aware that BOOGIE will not be very effective for the less studied systems. As the amount of data available for trait prediction keeps growing, we expect our tool will become increasingly valuable over the next years. It is striking that the most of the SNVs of genes relevant for blood transfusion (RhD and ABO) are already well characterized (see Figure 5.10), suggesting that BOOGIE can be useful for these important blood groups.

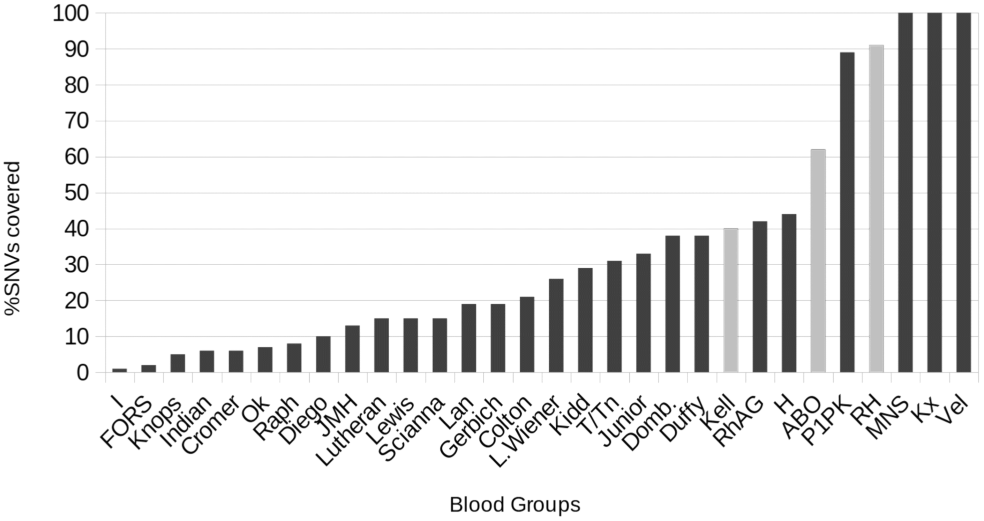


Figure 5.10: Fraction of dbSNP variants annotated for each blood group. The most important groups for the transfusion (grey bars) have good annotation in BGMUT, explaining the quality of our results. Conversely, most variants in the less studied systems still need further research for a proper automatic annotation.

5.3.1 Discussion

In our work, we developed the tool BOOGIE for blood group prediction. The main idea is to focus on few genetic locations with annotation from the literature for a set of common phenotypes. Using this information, we resolve the haplotype phasing problem and infer the most likely trait. From the theoretical point of view, linkage disequilibrium is the key factor leading to high accuracy even in presence of a small number of observed mutations.

BOOGIE was tested for human blood group prediction using PGP data and obtained very good results. Even if performance is not as good as serological tests and not yet suited for direct medical application, the few misclassification cases were easily detected. These provided interesting indications for the need of good quality SNVs in relevant spots and the importance of high penetrance mutation management, which can be easily dealt with in haplotype tables. In fact, lack of informative observations can result in the impossibility to properly classify sequencing data, as shown for the Rh group in the 23andMe dataset. Nevertheless, there is a number of advantages in our approach to be considered. The ability to directly detect ABO and Rh blood subgroups can be useful in ordinary blood transfusions. Whenever HTS data is available for some reason, it will be possible to test rare blood groups. For the 28 minor blood groups, we checked the PGP genomes for uncommon traits finding interesting results. As the experimental techniques for antigen detection are poorly sensitive in these blood systems, e.g. the Dombrock and Scianna system [258], genetic tests can be particularly useful. There is evidence of consistent discrepancies for the well-known ABO and Rh blood systems, estimated at 3.7% between real blood type and the one reported in identity documents [278]. Genetic tests could therefore represent an additional, albeit not exclusive, tool for checking compatibility during transfusions. Trait assignments in this work is based on PGP data donated by volunteers not directly connected with our laboratory. For this reason, no further experimental validation is possible for the other considered blood groups. The dataset nevertheless is clearly representative, as it contains samples of different ethnicity obtained with different experimental techniques.

BOOGIE uses a multivariate strategy based on maximum parsimony, which is particularly meaningful for proper characterization of non-trivial phenotypes, and is well explored in the context of sequence analysis. The method is very fast and produced blood group annotation for all 30 systems in a few seconds on a desktop PC. In light of the increasing amount of available HTS data, this is of great practical relevance for the quick and scalable annotation of genomes. Exponential growth of allele configurations is not a real danger, due to the limited number of heterozygous exonic SNVs typically observed in single genes. As long as the focus is on relevant hot spots, the number of permutations will be strongly reduced, leading to high computational speed. A key aspect of the system is flexibility. Simply adding more entries to the haplotype table will allow detection of new traits, while creation of new haplotype tables allows to tackle other genotype to phenotype problems. This is an important step towards the automation of trait detection in personalized medicine, in view of the constant growth of discovered phenotypes. In the context of multifactorial diseases this may be unfeasible, as we are still far from a clear description of key SNVs. Nevertheless, an increasing amount of studies may clarify these complex phenotypes and will suggest putative loci to test for mutations, as shown for cholesterol level models [279]. It should also be noted that no interpretation is possible for variants with no annotation in our knowledge base. Hence, proper reasoning will be possible only when all relevant variants are fully annotated. Dominance so far is also not considered, because it strongly depends on the context (e.g. X chromosome inactivation in women). BOOGIE computes two separate predictions, one for each chromatid. Proper interpretation of the resulting trait is left to the user and was straightforward in the ABO and Rh context. Despite these limitations, blood group traits are clearly relevant from the clinical point of view and may be effectively detected by our strategy. In addition to phenotype detection, the approach can also be a valuable tool for population studies. Some anthropological marker genes are important due to ethnicity-specific polymorphisms of certain human populations. The versatility of the tool allows us to imagine different scenarios where similar methods may be used for detection of rare diseases or in forensic medicine. In addition, BOOGIE can be downloaded (URL:`http://protein.bio.unipd.it/download/`) and customized for any trait prediction. Thanks to its effectiveness in HTS data interpretation, it can be of benefit for the clinical community and may help to develop of a new generation of tools for personalized medicine.

5.3.2 Methods

Data collection

In order to construct BOOGIE we had to extract as much information as possible about blood system classification from the literature. BGMUT [47] stores information about experimentally validated mutations known to be relevant for blood group determination. 34 known blood systems are described and are included in BOOGIE, with ABO and Rh of interest for method validation. The prediction strategy relies on an explicit definition of the ground truth, to be used during classification of new human samples. All loci of interest for genome classification are grouped in haplotype tables for each blood group. For each known phenotype, the table defines all expected

SNVs that should be observed for its determination. E.g. definition of the 177 known ABO blood groups uses 99 explicitly reported SNVs. Whenever BGMUT reports no data about a SNV, it is assumed to be the reference hg19 gene in our tables. Only exonic mutations were used, since they cover the largest part of the database and are more easily measurable during sequencing experiments. Obtaining the correspondence tables required meticulous manual curation, as many blood groups and traits assumed old reference genes. This is an important issue related to recent improvement of sequencing techniques, as they provided a number of different DNA sequence versions of increasing quality that cannot be easily combined. E.g. the ABO reference gene in BGMUT corresponds to the A group, while the reference gene in hg19 corresponds to the O group. This leads to a shift and re-labeling of most mutations. In many cases, like in the Indian system, SNVs simply report a wrong reference sequence, see variations in [280]. In addition, HTS data after variant calling uses genomic coordinates rather than gene-based coordinates. For this reason we used the BiomaRt [281] R package for quick coordinate translation, and manually fixed all mismatches in case of discrepancies with hg19. All conflicting cases were solved using the public databases dbSNP [3], UCSC reference genes [282], Phencode [283] and the original publications from PubMed. We decided to drop traits and mutations when the manual conversion failed. After this process, we obtained annotation for more than 800 different traits based on almost 1,000 unique coding variants for 30 blood groups.

Phenotype prediction

In HTS experiments, data obtained after variant calling provides an invaluable source of information for phenotype prediction. On the other hand, heterozygous mutations by themselves do not explicitly specify the genomic sample haplotype. This is known as the haplotype phasing problem in the literature. It can be solved effectively using information from HapMap [284] and expectation maximization approaches with Hidden Markov Models [285]. These methods are computationally expensive and based on low resolution data, which cannot distinguish rare haplotypes or uncommon SNVs [284]. Due to these limitations, we developed a new fast tool that could deal with (a) haplotype phasing and (b) phenotype decisions for the resulting predicted chromatid. For a particular patient gene, we assumed without loss of generality that there are N homozygous and M heterozygous mutations. With no prior assumption, this would lead to 2M-1 possible allele configurations C in the two human chromatids. As a first step, we enumerated all configurations, where exactly one of them represents reality. In the second step, we solved jointly the haplotype phasing and phenotype decision problem using our haplotype tables H. Given a particular configuration c, we determined its phenotype by means of a K-nearest neighbour strategy using data in the haplotype table H. In other words, we measured the sequence distance of c to those in H, and transferred by similarity the phenotype of the closest. Similarity is measured as Hamming distance, i.e. number of substitutions and insertions between two sequences, which is also important for determination of the most likely haplotype phase. The higher the value, the lesser the chance to observe c in nature. This is clearly linked to the maximum parsimony principle, which is well established in phylogenetic reconstruction [286]. We ranked all allele configurations by Hamming distance and selected the best as real allele configuration. The overall idea is that we are trying to find an example in our haplotype table which has the same mutations, so we can be reasonably confident that phenotype and haplotype will be the same. Conversely, a wrong allele configuration will look very different from the previously published ones and is not likely to exist at all.

BOOGIE system

BOOGIE was written with Java JDK 1.7, making it portable to all major operating systems. As shown in Figure 5.11, BOOGIE requires two files for its execution. The haplotype table for the phenotype of interest and the target genotype file. Format details are provided in a README file. Variants contained in the genotype file will be considered if and only if they are part of the haplotype table, i.e., they are useful for phenotype prediction. Once the variants have been selected, all the possible assignment to the two human chromatids are enumerated. The phenotype of each permutation is predicted by means of the 1-nearest neighbour algorithm, and the corresponding score of the most similar haplotype stored. The two assignments with overall maximal score become the predicted phenotypes. Note that dominance is not taken into account, so it is up to the user to determine the final traits. E.g., if the two alleles A and 0 are predicted for the ABO blood group, the expert user should infer that the A group is dominant.

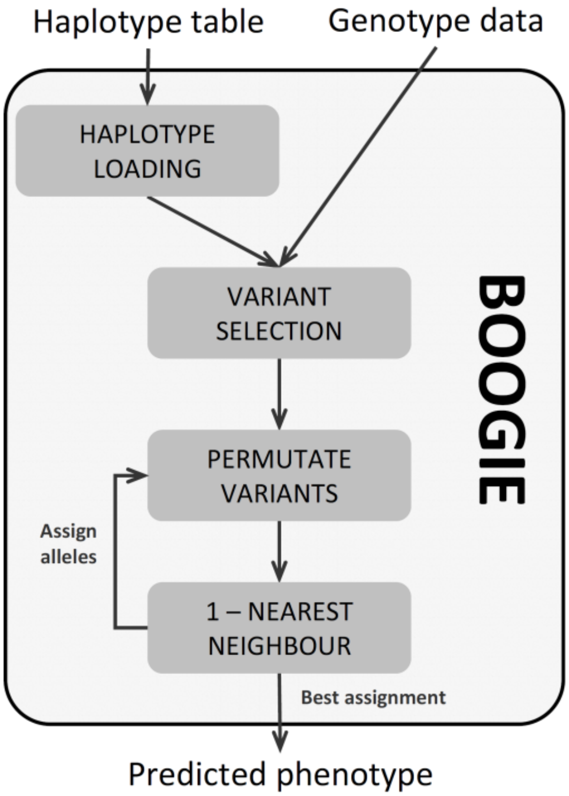


Figure 5.11: Schematic BOOGIE overview. The tool requires just genotype data and an haplotype table for its execution. Genotype must be specified in a tabular file similar to VCF format, where chromosome, genomic position, nucleotides ad zigosity are specified. Haplotypes are defined in a tabular file, and each row specify the expected SNVs of a target phenotype. See README file of the application for format details. BOOGIE search for key variants in the input genotype, and optimize their assignment to a haplotypes with known phenotype according to the 1-nearest neighbour algorithm. The SNV permutation with best score is the one with highest phenotype likelihood.

PGP dataset

The first goal for testing our method was the collection of samples with phenotype annotation and genetic data. We chose to work with public data from the PGP [249],[250], mainly due to the richness of the clinical profiles. From the 2,651 profiles accessed on 13 May 2013, entries having ABO and Rh annotation with sequencing data were extracted. We obtained two datasets with 69 samples with full genome sequences from Complete Genomics and 111 with 23andMe SNP data. The nature of the 23andMe data set is very heterogeneous in array size (ranging from 570k to 1000k SNVs) and chip type (customized Illumina Hap550+ or HumanOmniExpress BeadChip Kit). The reference genome used was either hg18 or hg19. For a detailed description of the data, please refer to the PGP website (URL:`http://personalgenomes.org/`). The full genome dataset was used for benchmarking the accuracy of the tool for ABO and Rh when full data is available. The 23andMe dataset was used as a further set to evaluate our prediction strategy when only partial information is available.

5.4 Von Hippel-Lindau (VHL) Pathway Modeling using Petri Net[4]

Pathological deregulation of cellular pathways often results in a family of complex and correlated diseases commonly termed cancer [287]. Cancer is a multi factorial disease where different causes contribute to its development. Several computational methods have been developed to explore the functional pathways involved in tumorigenesis. Some of them focus on differential gene expression between healthy and pathologic tissues [288, 289], on protein-protein interaction network analysis [290, 291] or on molecular dynamics simulations [292]. Other methods approach the disease through discretization of pathological components that result in tumor [293]. All of these approaches are very powerful when the variables related to the disease, although complex, are well known and studied. A multi-factorial disease can be approached by means of mathematical theory, building a theoretical model where cell components are connected with each other. In biology, several problems were dealt with network theory [294, 295]. A network is a group of objects strongly inter-connected with each other (e.g. proteins and enzymes

[4] The results of this chapter have been published in Minervini, G., Panizzoni, E., Giollo, M., Masiero, A., Ferrari, C., Tosatto, S. C. (2014). Design and Analysis of a Petri Net Model of the Von Hippel-Lindau (VHL) Tumor Suppressor Interaction Network. PloS one, 9(6), e96986. For Supplementary Material, check the online version of the paper.

of a pathway or animals belonging to interacting populations). Their construction and subsequent simulation is made via mathematical analysis of the connections between nodes found in the system and their time-dependent behavior [296]. A biological network is generally composed of proteins, nucleic acids and cofactors connected by biological reactions such as protein complex formation or enzyme activity regulation [296]. Von Hippel-Lindau syndrome (VHL) [297] is a good study case to test the network theory applied to cancer due to the similar medical history and pathological phenotype that patients share. While hereditary cancers represent only a small part of all human tumors, their investigation represents a challenge to understand the pathway leading to tumor formation. In 2010, Heiner et al. first approached VHL using the so-called Petri Net (PN) simulation networks [298]. Their work, inspired by a previous theoretical model of cellular oxygen-related pathways [299, 300], was a preliminary investigation of the core oxygen sensing system and its connection with VHL onset. Heiner and coworkers proposed three different functional modules responsible for hypoxia network control and for HIF-1α degradation [298]. In other words, they theorized that hereditary forms of cancer, such as different manifestations of VHL, are the result of different and concomitantly compromised metabolic pathways.

Von Hippel-Lindau Disease

Von Hippel-Lindau protein (pVHL) is the product of the von Hippel-Lindau gene, located in the short arm of 3rd chromosome, and constantly transcribed in both fetal and adult tissues [301]. Mutations of pVHL are related to a pathological outcome termed VHL syndrome, an inherited form of cancer [302]. VHL syndrome is characterized by cysts and tumors growing in specific parts of the organism [302, 303]. It is considered a severe autosomal dominant genetic condition with inheritance of one person in over 35,000 [304]. The tumor injuries, which can be either benign or malign, are usually located in the retina, adrenal glands, epididymis, central nervous system, kidneys and pancreas [305]. As a genetic disorder, VHL syndrome follows Knudsons two hit principle. A copy of the gene is mutated in the germ line, but the other gene copy still produces a functional protein. Complete protein inactivation appears during life due to somatic inactivation of the remaining functional copy [306]. On the contrary, mutations occurring during early fetal formation result in unsuccessful development [307]. The pVHL gene has 11,213 base pairs including three exons [304] and the final transcript is a protein commonly present in two isoforms: pVHL30 and pVHL19, of 213 and 160 residues respectively. Neither isoform contains a known enzymatic domain, but rather appears to serve as a multipurpose adapter protein engaging in multiple protein-protein interactions [308]. pVHL structure is organized in an α- and β-domain and its stability was demonstrated to be ensured by direct interaction with other proteins such as Elongins B and C [309]. Both Elongin B and C are also required for the best characterized function of pVHL, the ubiquitination dependent degradation of Hypoxia Inducible Factor (HIF) via the proteasome [310]. However, pVHL is considered a multipurpose protein due to its high number of known interactors. At the time of writing, the IntAct database [311] presents more than 200 different interaction partners, with some of them competing for the same Elongin binding site. Indeed, pVHL was found in different cellular compartments and seems to be involved in many different cellular processes such as apoptosis, cell proliferation, survival and motility [312]. Considering the huge number of interactors and multiple cellular localizations, many different functions have been described or hypothesized, such as regulation of cytoplasmic microtubules during mitosis [313] and endothelial extracellular matrix deposition [314]. On the other hand, considering the huge number of players involved in VHL syndrome and the lack of reliable kinetic data, a PN based approach may be a preferable option for an entire VHL pathway simulation.

Petri Net for Interaction Pathways

Since their invention, by Carl Adam Petri in the early sixties, PNs were mostly used to describe technical systems, but later the utility in describing biological and biochemical functions has also been demonstrated [315]. PNs were successfully used in many studies to describe biological networks [48], such as the regulation and etiopathology in human Duchenne Muscular Dystrophy [316] and the hypoxia response network [298]. PNs are qualitative mathematical models that can graphically represent many object types, not only metabolites but also different protein states and are useful to simulate networks where not only metabolites are involved. Indeed, PNs can be a powerful tool to study all concurrent interactions in a specific pathway, even if the proteins or kinetics are not well-known. Due to the large number of different pVHL functions involved in VHL disease progression, we decided to extend the PN based analysis of [298] increasing the number of considered protein-protein interactions. We generated a novel manually curated PN model of the entire VHL regulation system collecting data from the literature and including the signaling pathways and glucidic metabolism. In order to build a realistic network, literature from both biochemical experiments and in silico predictions were used as source. It was decided to build a PN with only confirmed pVHL interactions whose function was also known. The resulting PN was validated using an analysis of specific properties as suggested by previous studies using the same method [315]. After validating the PN structure, in silico knock outs of specific proteins were done in order to observe the different network behaviors and the resulting biological effect.

5.4.1 Methods

The network was designed in the Snoopy PN framework (version 2, revision 1.13) [317], respecting the mathematical PN formalism as described in [298, 318]. PN were demonstrated to be useful in describing discrete and concurrent processes in a simple graphical representation [48] and have been used to describe biomedical processes due to their capacity of representing sequential steps in a process. PN modeling methods are actively used to describe, simulate, analyze, and predict the behavior of biological systems. The Snoopy PN framework provides an extensible multi-platform framework to design, animate, and simulate Petri nets [317]. We chose Snoopy to facilitate future extensions of the VHL pathway presented here. Among different available PN types a standard PN was chosen to limit the number of variables. Both Charlie and PInA analyzers were used for PN analysis and validation [319]. Further, in silico knock out experiments were used to test the biological reliability of the model. Structural model validation was made by analysis of the T-invariants to demonstrate whether the system was covered by T-invariants and to confirm the biological meaning of each invariant. The use of T- and P-invariants is given by their own properties: they are a set (of transitions or places, respectively) that allow the reproduction of the same state after n transformations. A P-invariant represents a set of places where the number of tokens is constant and independent on the firing rate. A T-invariant instead represents a set of transitions that cyclically comes back to show the same initial set. Biologically a P invariant can represent the process of regulating a protein, whereas T invariants can represent cyclical biochemical transformations such as metabolic reactions. To this end, the computed invariants were grouped in Maximal Common Transition Sets (MCTS) and Clusters, the former based on occurrence of specific sets of transition inside the various T-invariants, and the latter based on similarities between T-invariants. Different numbers of clusters will be defined depending on the resulting square matrix. Where MCTS create disjunctive nets, Clusters merge together similar T-invariants. Behavioral validation was made by selectively deleting tokens inside the model, imitating possible biological disruptions such as disease-causing mutations. The resulting network behavior was compared to what is reported in the literature. Total runtime for invariants computation were less than ten seconds on a mainstream Linux x86 workstation. Literature sources used to build the model are reported in Table S1. The Snoopy framework for PN construction, Charlie and PInA tools for analysis are available at the website (URL:`http://www-dssz.informatik.tu-cottbus.de/DSSZ/Software`). Finally, the model was used to simulate the network behavior through visual inspection of both token movement and accumulation in specific parts of the network. For a visual explanation of token movement in a PN refer to Video S1.

Model Availability

The resulting VHL disease PN model is available in File S1.

5.4.2 Results

Notations and Assumptions

The PN built here focuses on pVHL interactions that were already proven by biochemical experiments and reported in the literature. We chose to model a realistic VHL disease pathways based on confirmed literature data, including all known VHL functions, VHL related signal pathway and glucidic metabolism. All bibliographic sources used to design the model are presented in Table S1. The final PN is composed of 323 places and 238 transitions, connected by 801 arcs. Tables S1 and S2 show all places and transitions and the related biological correspondence. Places are mainly proteins and enzymes, while some represent DNA or small molecular substrates such as glucose and cofactors (e.g. ATP). Notation for both pre- and post-places and their biological meaning are explained in Table S1. In a few cases, places are used to represent a whole group of changes generated by DNA transcription, (e.g. p_32 and p_33 or Et_eff1 and Et_eff2). Transitions instead symbolize complex formation between two proteins or post-translational modifications. Output transitions stand for degradation or movement to other parts of the cell or organism to complete their functions (e.g. degrad_1 and degrad_2) whereas input transitions show the generation of a substrate or protein. In order to simplify the design of such a large network, we decided to use macro nodes to group reactions representing complex molecular pathways such as signaling pathways or secondary signal cascades. The whole process is merged into a single node with a given name to allow visual inspection only in case of need. From the top level all transitions can still be found in a hierarchical lower layout level. Logic nodes were used for places participating in many reactions throughout the network such as ATP and ADP (7 logical copies each) or NAD and NADH (4 logical copies each). A total nesting depth of two was chosen to model macro nodes. Special arcs were not used while we chose to model the permanent presence of some objects using double arcs (e.g. for elob, eloc and places standing for enzymatic activity). In case of proteins which are actively degraded, it was preferred to create an input transition simulating constant production (or synthesis) and an output for consumption. This is the case for pkcz2, Jade1, pVHL and HIF-1α. As can be seen from Figures 5.12 to 5.14, which represent the entire model, two major nodes can be immediately identified: pVHL and vcb, the complex made by pVHL and the two elongins. Another relevant part is the glucidic metabolism, modeled due to its hypoxia induced regulation. It is represented in detail in Figure 5.13.

HIF-1α Transcription Activity

The HIF-1α transcription factor stimulates proliferation of endothelial cells to create new blood vessels during localized or broad hypoxia. In human, it is present as three different paralogs: HIF-1α, HIF-2α and HIF-3α. The sequence is quite conserved between the former two, whereas the latter is slightly shorter and seems to have completely different functions compared to the other two [318, 319]. Both HIF-1α and -2α stimulate DNA transcription but the exact products of this activity are still poorly understood. In our model, only HIF-1α in vivo activity was considered. It cannot be excluded

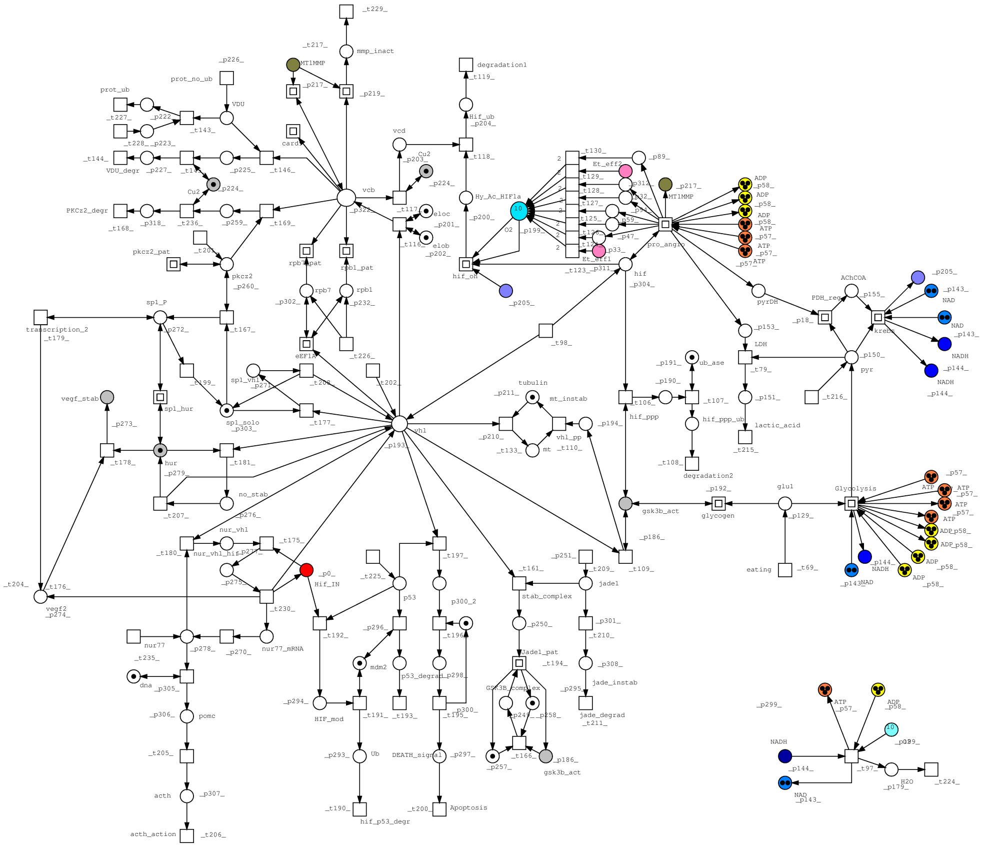


Figure 5.12: **Top level model.** The colors of some tokens were arbitrarily chosen to give a clearer identification of the central nodes (ATP, Vcb and oxygen) or for nodes involved in more reactions such as GSK3β. The group of nodes in the bottom left is not disconnected from the central body of the network thanks to the presence of logic nodes for ATP synthesis (t_97).

that other biological effects depend on the second paralog. Indeed, both have a pro-angiogenetic function and are degraded by pVHL via proline-directed hydroxylation. HIF is a heterodimer of HIF-1α and HIF-1β, the latter being also termed Aryl hydrocarbon Receptor Nuclear Translocator (ARNT). We started from the transcription activity of HIF due to its regulation is the most studied pVHL function. Our model, as expected from literature data, shows that HIF-1α enters the nucleus when not degraded by pVHL. It subsequently binds HIF-1β to form the HIF heterocomplex which interacts with

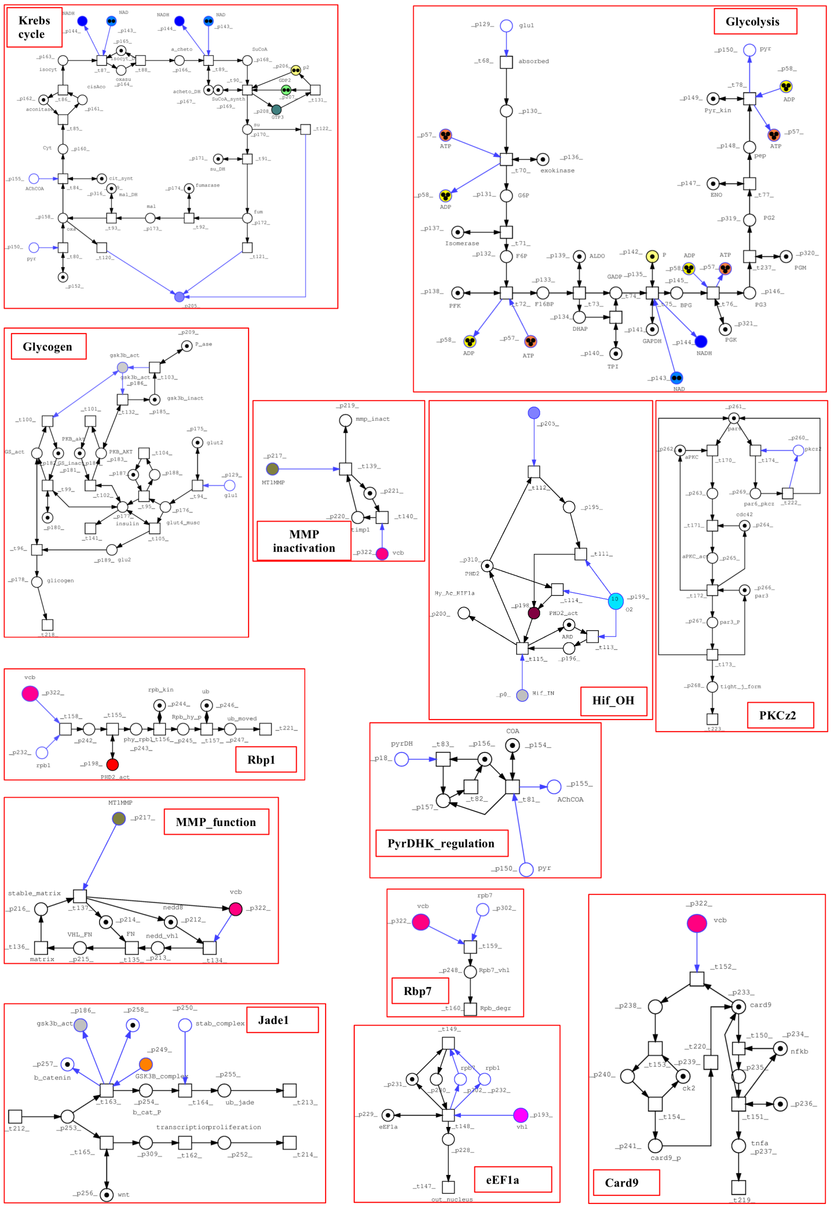


Figure 5.13: **Lower hierarchical PN levels.** Pathways from the top level are grouped in macro-nodes (functional subordinated layer), in particular glucidic metabolism and various VHL functions.

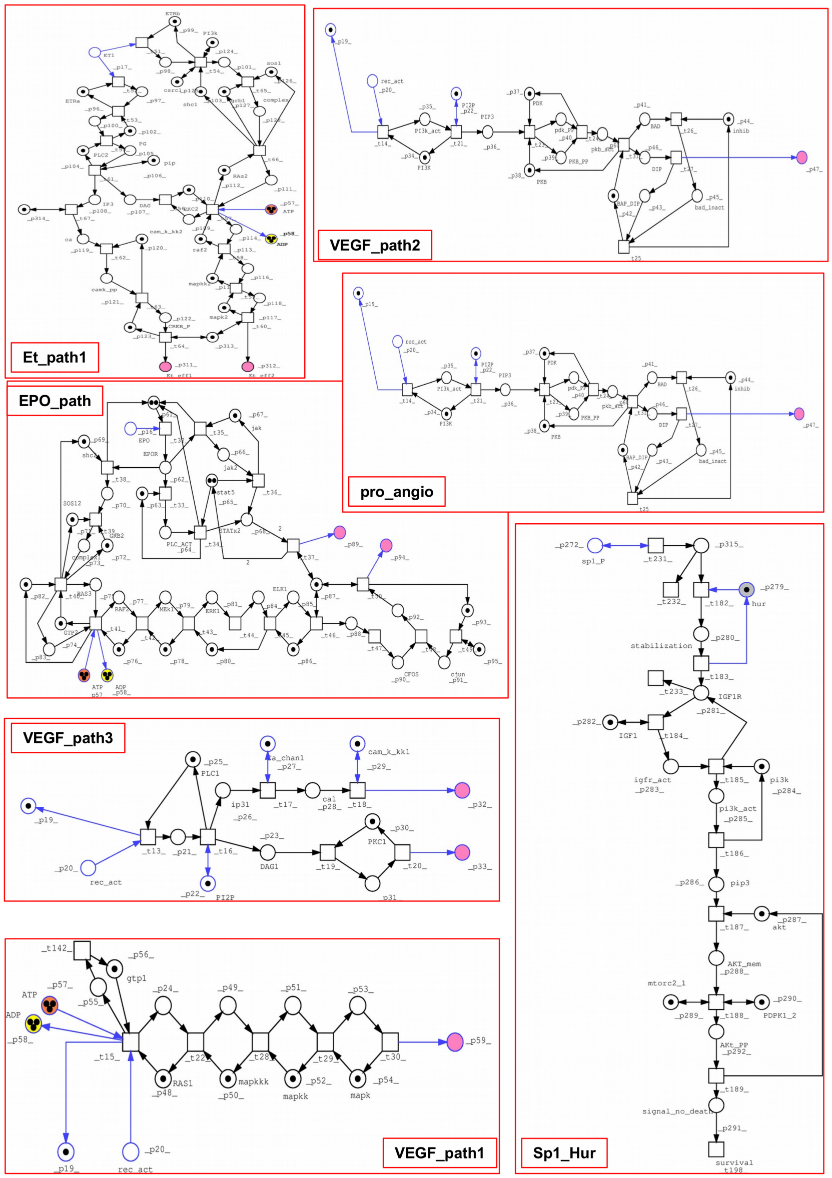


Figure 5.14: **Lower hierarchical PN levels, in particular HIF-1α regulation and HIF-1α-dependent pro-angiogenic signaling.** VEGF and EPO pathways are at a lower hierarchical level than the pro_angio macro-node.

DNA. Our model correctly simulates the increased affinity of HIF towards DNA. Transcription is enhanced by some co-factors binding both subunits of HIF and other proteins such as p300, Creb and cjun. This takes place in a specific DNA promoter sequence termed Hypoxia Response Element (HRE). Furthermore, during transcription some pro-angiogenic factors are produced: Vascular Endothelial Growth Factor (VEGF), Endothelin (ET) and Erythropoietin (EPO). All described pathways are in agreement with previous observations reported in [320]

Metabolic Processes

HIF-1α transcription activity includes some proteins which are dependent on oxygen but involved in other pathways (e.g. oxidative metabolism) or completely independent (e.g. metallo-proteinase MT1MMP). Further, HIF-1α stimulates production of proteins involved in the glucidic pathway. The final product of the metabolism is adenosine triphosphate (ATP), a molecular form of energy, composed by adenosine, an adenine ring connected to a ribose sugar, and three phosphate moieties. When a phosphate moiety is hydrolyzed it releases energy, used by cells for enzymatic reactions. The glucidic metabolism is composed of glycolysis, Krebs cycle, glycogen formation and respiratory chain with ATP synthesis. Glucose is absorbed in cells by enzymatic glucose transporters (GLUT), which carry the molecule to the location inside the cell where the metabolism takes place [321]. There are many isoforms of these transporters: GLUT1 is present in all cells and in particular in erythrocytic membranes, neurons and glia [322]. GLUT2, located in both liver and pancreatic beta cells, is characterized by low affinity for glucose, hence it requires a higher glucose concentration to be activated [323]. Right after eating, glucose concentration increases, thereby quickly activating them. GLUT2 stimulates production of insulin, a hormone regulating the plasmatic glucose concentration. Glucose plasmatic concentration can also increase due to an opposite pathway, originating from liver glycogen being decomposed into glucose and reaching systemic circulation. GLUT3 is mostly present in neurons, whereas GLUT4 is the insulin activated transporter located in myocytes, adipocytes and cardiomyocytes [321, 324]. In our model, we chose to exclude GLUT3 due to its specific role in neuronal cells. Glycolysis occurs in the cytoplasm and during this process each glucose molecule is phosphorylated, consuming two molecules of ATP, then divided

into two smaller molecules. Further modifications of these two molecules result in new ATP production. The molecule obtained at the end of glycolysis is pyruvate, which can be again modified through three different pathways. It can be decarboxylated and linked to Co-enzyme A to form acetyl-Co-enzyme A. It can then be carboxylated to obtain oxalacetate, or transformed through lactate dehydrogenase into lactic acid. Pyruvate can also be generated by other metabolic pathways, like protein or fatty acid disruption and amino-acid modifications. Acetyl-CoA and oxalacetate are the molecules used in the following glucidic metabolism process, the Krebs cycle, taking place in the mitochondrial matrix. The Krebs cycle starts with acetyl-CoA and oxalacetate merging to create citric acid, which continues undergoing modifications until oxalacetate is formed again. During the process some co-enzymes are modified. Decarboxylation of pyruvate to form acetyl-CoA already transforms a NAD+ (Nicotinamide Adenine Dinucleotide) in NADH (reduced form), afterwards obtaining one more of ATP, GTP, FADH2 (Flavin Adenine Dinucleotide) and three more NADH per pyruvate molecule entering the Krebs cycle. The redox co-enzymes are considered electron transporters. During metabolic reactions they reduce themselves and get electrons (and protons) to oxidize the substrate of the enzymatic reaction. Electrons taken during the glucose metabolism are then used in the respiratory chain taking place in the internal mitochondrial membrane. The respiratory chain consists in transporting electrons through enzymes called cytochromes and others co-enzymes, characterized by the capability to receive and donate electrons.
NADH (FADH2) is oxidized again by cytochromes going back to the form of NAD (or FAD). Electrons gained through oxidation are used to reduce half a molecule of oxygen into water, releasing more energy. The FADH2 and NADH redox chain establishes a chemical potential causing the push of protons outside the internal membrane towards the inter-membrane space, which stays between the mitochondrial inner and outer membrane. This also causes a higher concentration of protons outside the inner membrane. The resulting gradient causes the tendency of protons to enter the cell. The final step is ATP-synthetase, formed by a channel that allows protons to enter, pushed by the gradient, allowing the enzyme to change conformation and make its reaction. This kinetic energy is converted into ATP. In our model, the glycolytic and Krebs cycles were described in detail, represented at the hierarchical second level by the coarse transition Glycolysis. The respiratory chain was instead merged into a single node (t_97). We chose to represent creation and consumption of ATP in order to show the effects of lower and

higher oxygen concentration on the network. On the other hand, oxygen consumption for ATP synthesis during the respiratory chain creates a flow of oxygen in the model. Oxygen is not the only connection between glucidic metabolism and hypoxia. Indeed, HIF-1α transcription activity enhances the transcription of many GLUT isoforms (such as 1, 3 and 9) and the pyruvate dehydrogenase kinase, which determines the pyruvate dehydrogenase (PyrDH) inactivation and consequent Acetyl-CoA formation from pyruvate. Finally, Lactate dehydrogenase is also produced, to ensure an alternative compound, creating energy needed for cell survival [321, 322, 323, 324, 325]

pVHL-dependent Processes

Some interactors can bind pVHL in regions interacting with Elongin C. These are HuR, Nur77, p53 and Jade1. Nur77 has a complex function and its role in pVHL tumor suppressor activity is still not entirely clear. Nur77 can bind pVHL, inhibiting Elongin binding while allowing HIF-1α binding. Its transcription is stimulated by HIF-1α itself, and pVHL-HIF-1α-Nur77 complex formation stabilizes the transcription activity of HIF-1α by inhibiting the pVHL-dependent degradation [326]. Another Nur77 function is the stimulation of proopiomelanocortin (POMC) transcription, which is a precursor for adrenocorticotropic hormone (ACTH) formation. This hormone has an important stress response function, stimulating cortisol production and other neurotransmitters from the adrenal glands, to enhance the organism reaction to danger and stress stimuli e.g. increase of gluconeogenesis and muscle mass. An excess of this hormone can cause desensitization of its receptors for feedback down-regulation and thus muscular weakness, tiredness, hyperglycemia and osteoporosis [327]. p53 can bind to pVHL avoiding the degradation of this tumor suppressor. Instead, it stimulates the apoptotic signal cascade via the p300 co-activator, which stimulates production of proteins enhancing the cell programmed death. If p53 cannot bind pVHL, two more mechanisms are described in the model. One is its modification and degradation by Mdm2 and the other is the pVHL-independent degradation of HIF-1α. Interaction with Mdm2 is needed in both cases [328, 329]. Jade1 is a short-lived protein whose main function is to stimulate the phosphorylation-dependent degradation of β-catenin. This is a subunit of the cadherin protein complex acting as an intracellular signal transducer in the Wnt signaling pathway. It seems that β-catenin is able to stop cell division via a contact-dependent inhibition signal, whereas in Wnt signaling it is also involved in proliferative transcription. When Wnt is not present, β-catenin can be phosphorylated by Glycogen Synthetase Kinase, type 3β (GSK3β) in complex with APC (Adenomatous Polyposis Coli) and Axin. β-catenin can interact with Jade1 and be only successfully degraded after this interaction [330]. Related functions are represented in the macro node Jade1_pat. GSK3β seems to be a protein involved in many different pathways. GSK3β is involved in Glycogen Synthetase deactivation and can even phosphorylate pVHL and HIF-1α. In the case of HIF-1α, it generates a pVHL-independent degradation pathway, where phosphorylation allows ubiquitination, whereas in the case of pVHL, it inhibits pVHL stabilization of microtubules [331].

Structural Model Analysis

Based also on previous observations of Heiner et al., [332], in 2008 Grunwald et al., demonstrated that PN can be used to describe large and complex metabolic pathways [316]. They postulated the following set of minimal rules that a PN should satisfy to be considered biologically reliable: (1) the network should be entirely connected, (2) the network should be covered by T-invariants, and (3) each T-invariant and P-invariant should have a biological meaning. The model described here was tested with respect to what previously done by Grunwald and co-workers [316] and resulted to be covered by T-invariants, connected, homogeneous and each place has a pre-transition and a post-transition. Transitions without pre- or post-places were used to simulate the system interface to the surroundings. The network is alive, in other words, it continues to work forever, with all transitions contributing to the net behavior forever, and no dead transitions. The MCTS and Cluster analysis were used due to the large number of T- and P-invariants included in the model. Both methods are used in PN theory to reduce the complexity connected with such a large network and to reduce the errors connected with manual investigation. From the 238 transitions present at the beginning in the model, 393 T-invariants were computed without considering 10 trivial invariants. The latter consist in a pair of transitions that usually represent a forward and backward reaction, such as the active and inactive state of a protein. Trivial invariants could be erased to reduce the dimension of the network without disturbing the overall system when the interest is focused on the steady state behavior [298]. T-invariants were grouped into 44 Clusters using the Tanimoto coefficient with similarity threshold of 65%, as described in [316]. Only 11 of these 44 comprised more than one T-invariant. The three biggest Clusters are C9, composed of 144 T-invariants, C8 of 72 and C11 of 64 T-invariants. Separation into clusters allows easier analysis of networks pathways represented by each T-invariant, since they are grouped by similarity, specifically the common transitions by which they are composed. T-invariants named in the text are shown in Table S3, while T-invariants grouped in C8, C9, C10, C11 are explained in Table S4 and described as follows.

Cluster C8

Cluster C8 groups all transitions included in HIF-1α pathways, including transcription, signaling cascades, degradation via pVHL, p53 and GSK3β, and eventually the Krebs cycle. For the EPO signaling pathway, two transitions (t_35 and t_36) are not included which cause Jak activation and consequent Stat5 activation to stimulate DNA transcription. Matrix stability regulation is also part of the cluster due to the destabilization induced by HIF-1α transcription of metallo-proteinase (MMP), transitions from t_134 to t_140. The largest T-invariant in C8 is Inv_280 (93 transitions) while the smallest is Inv_377 (81 transitions). The differences between T-invariants show the possibility of alternative pathways inside the model. For example, the VEGF dependent signal cascade can proceed in three different ways: t_13, t_14 and t_15, which lead to the pathways being merged in the coarse nodes Vegf_path3, Vegf_path2 and Vegf_path1, respectively. The occurrence rate in C8 is 24 transitions for each path. The Endothelin, VEGF and Erythropoietin pathways are not in conflict and occurring together. Disaggregation of the matrix via MMPs is present in 18 T-invariants, whereas inhibition of these proteins, i.e. matrix stabilization, is present in the remaining 54 transitions. Regarding the Krebs cycle, 47 T-invariants have t_91, of which only 24 reach t_92 and t_93, representing the last three steps of the cycle: succinate to fumarate, fumarate to malate, and malate to oxalacetate. All the malate being produced is used to regenerate oxalacetate. Degradation of HIF-1α occurs in any T-invariant of the cluster. The pVHL-dependent degradation of HIF-1α is always present (transitions t_116 to t_119). In 19 T-invariants degradation takes place via p53 (t_191 to t_193) or, alternatively, via phosphorylation by GSK3β in another 17 T- invariants. Two of the three pathways can be present in the same T-invariant, as in Inv_227, where degradation via pVHL and degradation via p53 are both present. This was considered as the

HIF-1α dependence on the lack of degradation by these proteins. All three degradation pathways never appear in the same T-invariant. The p53 and GSK3β paths are never present together but each of them is accompanied by pVHL-dependent proteasomal degradation. Inv_377 lacks the EPO signaling pathway but is the only one in this cluster to have t_34, t_33 and t_37. These invariants have all input and output transitions. For example, t_202 the second input for pVHL, is present in only 18 invariants. Other inputs are t_98, always present, leading to formation of HIF-1α and pVHL, t_192, producing p53 and t_216, representing other pyruvate generating metabolic pathways. The latter is also present in each invariant allowing formation of the pyruvate needed for Krebs cycle progression.

Cluster C9

Cluster C9 is the largest cluster in our model and includes 144 T-invariants. It is characterized by complete EPO pathway abrogation which goes through formation of the Shc-Grb-Sos complex and the consequent mapk-dependent phosphorylation cascade. Transition t_127, representing EPO effects on oxygen production, is absent. In its place, t_35 and t_36 are considered, which are present in 72 T-invariants. In cluster C9, the largest T-invariants are Inv_278 and Inv_279 (74 transitions) while the shortest ones are Inv_101, Inv_105, Inv_144 and Inv_148 with 65 transitions each.

Cluster C10

Cluster C10, composed of 52 T-invariants, is characterized by the presence of glycolysis between many transitions grouped in the cluster. This is also the cluster containing the most populated T-invariant of all computed 393 non-trivial T-invariants. This is Inv_245, including 101 transitions and covering almost half of the whole model. Cluster C10 also includes Inv_125, the shortest invariant of this model, composed by 85 transitions due to lack of the Krebs cycle. Another difference with the other three major clusters is that here both EPO paths are present, specifically, the Jak pathway belongs to 4 T-invariants and Shc-Grb-Sos is observed throughout the cluster. Vegf_path1 seems to be more common in this cluster, being present in 36 T-invariants, whereas the other two are present 12 times each. This time they are present even in the same invariant, as for Inv_60, Inv_129, Inv_172 and Inv_215, with both t_13 and t_15, and Inv_142, Inv_185 and Inv_228 with t_14 and t_15 and all subsequent signaling appearing at the same time. Despite glycolysis being present in all cluster invariants, the Krebs cycle appears only in 11 cases. p53-dependent degradation of HIF-1α occurs in 11 cases while the phosphorylation-dependent one appears in 13. An input transition has been added with respect to the other major clusters so far analyzed (i.e. t_69_eating) without which glycolysis could never take place.

Cluster C11

Cluster C11 is composed of 64 T-invariants. Only part of the EPO pathway is described here, with the major difference that the Krebs cycle is completely abrogated while Prolyl Hydroxylase type 2 (PHD2) regulation by oxalacetate is included. HIF-1α interaction with Nur77 and transcription of VEGF by Sp1 are also present. t_80 (transformation of pyruvate in oxalacetate) is not present in the first 42 cluster T-invariants. Nur77 interaction with HIF-1α is present only in 8 T-invariants, specifically Inv_87 to Inv_94. Sp1 transcription activity is appearing in twice the amount, including the same 8 invariants just mentioned. VEGF transcription via Sp1 activity is aPKC2 phosphorylation dependent, which does however not appear in the cluster. When VEGF is synthesized, it is subsequently stabilized by Hur, followed by t_178 and Hur is recreated to allow other functions. Indeed, it is one of the few places without input transition but with a token that goes forward and backward again. Compared to the other clusters, C11 also shows one less transition in the coarse PHD regulation node, specifically t_81, which shows the transformation of pyruvate by pyrDH into acetyl-Coenzyme A, needed for the Krebs cycle. The four clusters C-8 to C12 are very similar to each other, as can be seen from the distance tree in Figure 5.15. They all contain the HIF-1α transcription activity and signaling pathways caused by EPO, VEGF and the HIF-1α degradation options. They include the effects of other transcription activity products, like metallo-proteinase and pyruvate dehydrogenase kinase, which regulate the activation state of PyrDH. All include part of the glucidic metabolism but not Glycogen formation itself. Other five clusters from C12 to C16 have a smaller number of T-invariants and fewer transitions present in each invariant. They do not include transcription activity but are only formed by the VEGF and glycolytic pathways. The information contents of these clusters turned out to be uninformative and their analysis was not included. The same applies to clusters composed by 13 T-invariants.

Finally, some transitions are not present in the clusters and not listed in the T-invariants because trivial invariants were excluded from cluster analysis. These transitions are shown in Table 5.8 with their respective biological meaning.

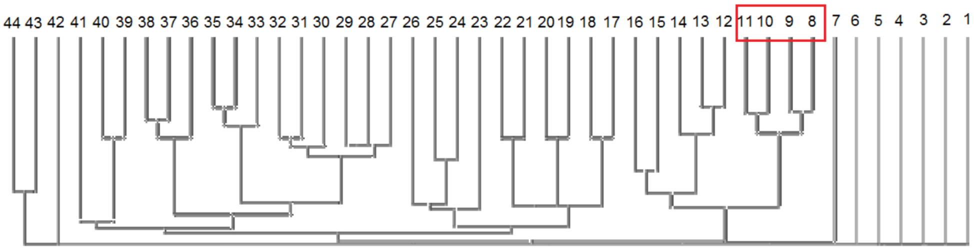
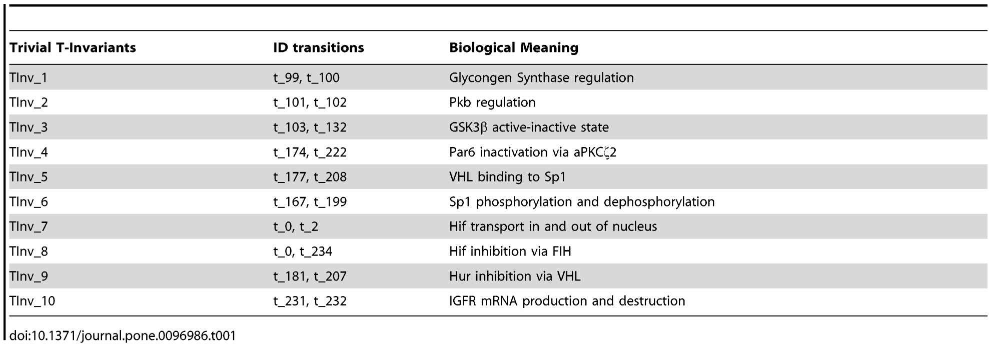


Figure 5.15: **PinA Distance Matrix clustering, using Tanimoto coefficient and 65% threshold of.** The numbers indicates clusters. In C8,C9, C10, C11 are highlighted a red square.



Trivial T-Invariants	ID transitions	Biological Meaning
TInv_1	t_99, t_100	Glycongen Synthase regulation
TInv_2	t_101, t_102	Pkb regulation
TInv_3	t_103, t_132	GSK3β active-inactive state
TInv_4	t_174, t_222	Par6 inactivation via aPKCζ2
TInv_5	t_177, t_208	VHL binding to Sp1
TInv_6	t_167, t_199	Sp1 phosphorylation and dephosphorylation
TInv_7	t_0, t_2	Hif transport in and out of nucleus
TInv_8	t_0, t_234	Hif inhibition via FIH
TInv_9	t_181, t_207	Hur inhibition via VHL
TInv_10	t_231, t_232	IGFR mRNA production and destruction

doi:10.1371/journal.pone.0096986.t001

Table 5.8: List of Trivial T-invariants excluded from calculation with their associated biological meaning.

MCTS Analysis

Another way to group invariants is by the amount of single transitions present in them. Maximal common transition set (MCTS) analysis provides a PN decomposition into non-overlapping subnets, sharing parts of the same T-invariants [315]. In a biochemical network, MCTS could be interpreted as enzyme subsets operating together under steady state conditions, computed based on the support of a T-invariant. MCTS computation does not consider stoichiometric relations, describing exclusively sets of reactions present in a maximal number of T-invariants resulting shared by different signaling pathways [333]. A total of 40 non-trivial MCTS were identified, with results and related biological means shown in Table 5.9. Some transitions do not belong to a non-trivial MCTS, because their occurrence has no similarity with other transitions and they create separate MCTS (specifically: t_69, t_82, t_91, t_94, t_98, t_114, t_116, t_120, t_121, t_122, t_179, t_202, t_209, t_212, t_216, t_225 and t_229). MCTS define transitions that always take place together, but are not necessarily connected, thus representing disjunct building blocks constituting the network. Considering both analyses, a table was automatically built in PInA [315] showing a correlation between clusters and MCTS. Transitions (t) or MCTS (M) are compared to evaluate how many T-invariants clusters cover the selected M or t (if the transition is not already part of the MCTS, as listed above). The more covered a transition or set is, the more central it could be considered for the network behavior. Recently, a network coarsening method based on abstract dependent transition sets (ADT) was presented [334]. It is formulated without the requirement of pre-computation of the T-invariants and is a tool commonly used for the decomposition of large biochemical networks into smaller subnets. Due to the manually designed nature of our model, we preferred to maintain a logic hierarchy based on metabolic pathways in order to maintain the network centered on pVHL and its interaction. The MCTS calculation results shows that the most covered set by cluster T-invariants is M20 with 358 T-invariants covering all transitions in the set, indicating that this MCTS corresponds to more T-invariants than the others. All transition sets are an important link to the others, as tokens pass through these transitions more often. A transition not present in any set but most covered by T-invariants is t_98, which is also the most frequently occurring transition, see Figure 5.16. The 10 most occurring transitions are listed in the Table 5.10.

P-invariant Analysis

Although the network is not covered by P-invariants, it has 130 P-invariants. 47 of these are trivial P-invariants, comprising a single place, connected with double arcs to imitate an activator arc function. Another object represented

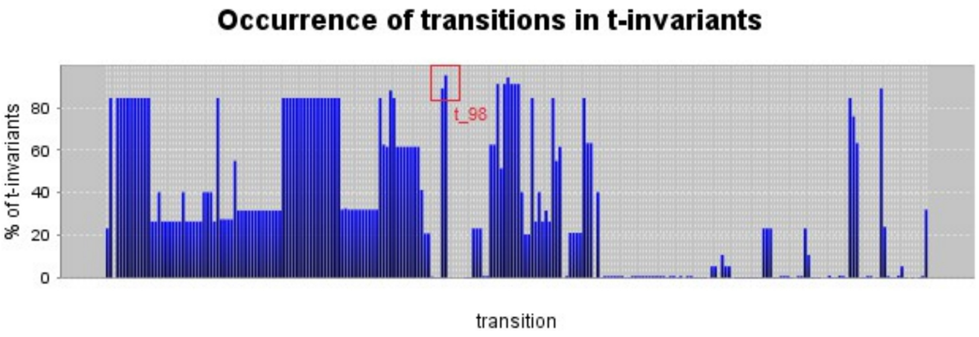
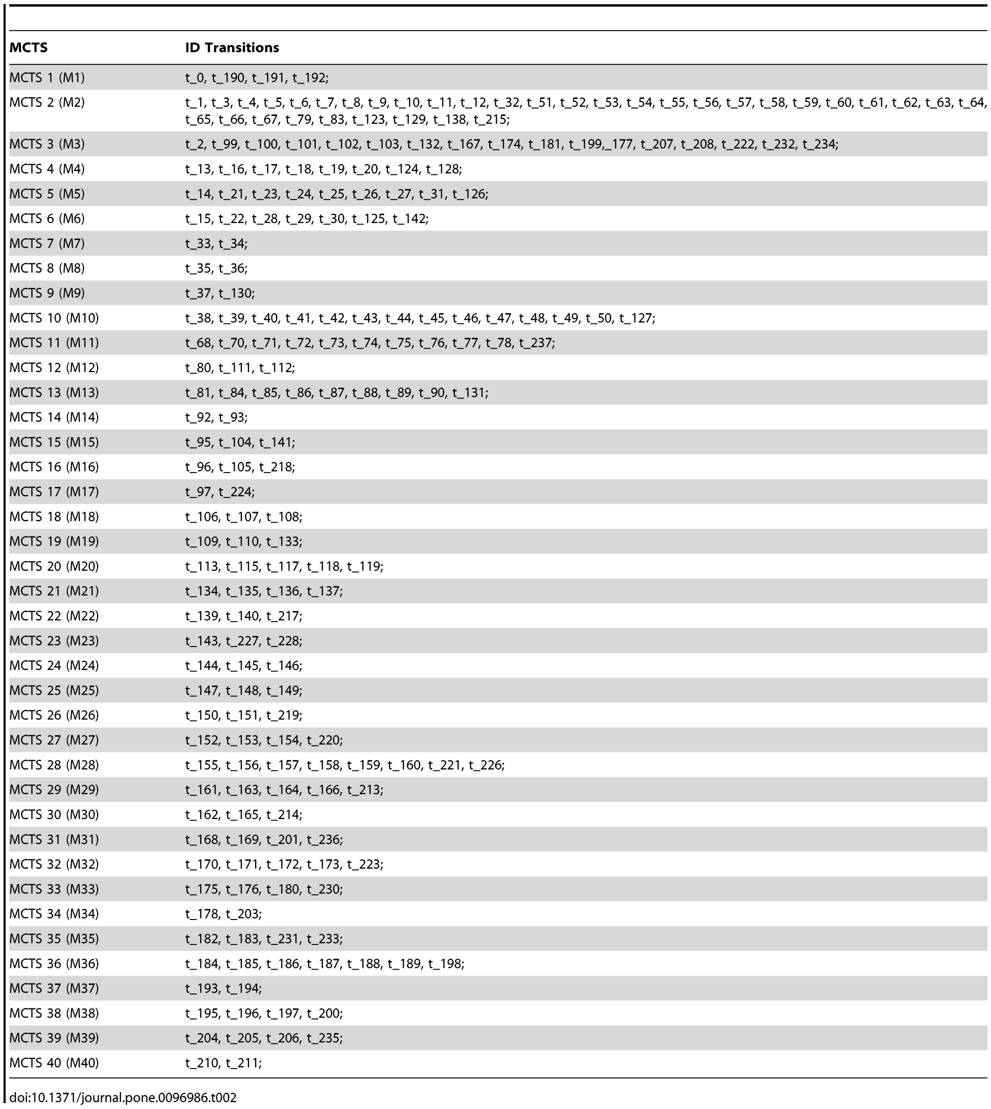


Figure 5.16: **Transitions occurrence T-invariants.** Transitions are ordered by name and t_98 is highlighted in red.

with double arcs is the enzymatic activity catalyzing a reaction and immediately going back to the steady state. P-invariants show places or sets of places where token numbers always remain equal and do not move outside the subnetwork induced by the P-invariant in the initial marking. In other words, they do not grow nor diminish. The remaining P-invariants are mostly located in signal transduction pathways, such as situations in which a protein is sequestered from its function and then goes back after a second reactivation mechanism. This scenario is present in p_41, p_42 and p_45 located in invariant P_58. It is important to notice that ATP and ADP, as well as NAD and NADH, are modeled as P-invariants. P_90, P_91 and t_97 are able to transform ATP and ADP. More in general, all energy consuming transitions are considered to be backward transitions of invariants. Invariants not related to signal transduction are places located in the Hur system, where Hur is removed from its function by pVHL. This is a good approximation for sequential modifications that momentarily activate proteins. Afterwards, Hur can go back and stabilize VEGF to increase its transcription activity.

In Silico Knock Out Experiments

The previously described clustering and MCTS analysis for T-invariants allowed us to identify the most common transitions and to understand which transitions can be depleted in our knock out experiments in order to get the most important biological effect. The knock out experiments were performed erasing selected transitions or tokens and observing which transitions



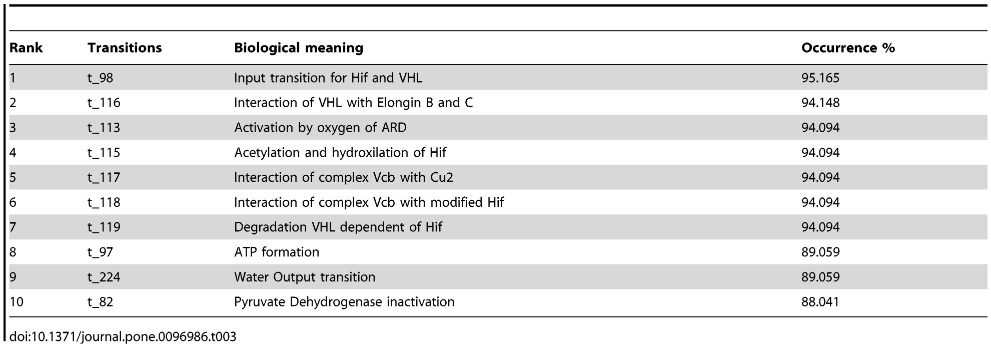
MCTS	ID Transitions
MCTS 1 (M1)	t_0, t_190, t_191, t_192;
MCTS 2 (M2)	t_1, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_10, t_11, t_12, t_32, t_51, t_52, t_53, t_54, t_55, t_56, t_57, t_58, t_59, t_60, t_61, t_62, t_63, t_64, t_65, t_66, t_67, t_79, t_83, t_123, t_129, t_138, t_215;
MCTS 3 (M3)	t_2, t_99, t_100, t_101, t_102, t_103, t_132, t_167, t_174, t_181, t_199,_177, t_207, t_208, t_222, t_232, t_234;
MCTS 4 (M4)	t_13, t_16, t_17, t_18, t_19, t_20, t_124, t_128;
MCTS 5 (M5)	t_14, t_21, t_23, t_24, t_25, t_26, t_27, t_31, t_126;
MCTS 6 (M6)	t_15, t_22, t_28, t_29, t_30, t_125, t_142;
MCTS 7 (M7)	t_33, t_34;
MCTS 8 (M8)	t_35, t_36;
MCTS 9 (M9)	t_37, t_130;
MCTS 10 (M10)	t_38, t_39, t_40, t_41, t_42, t_43, t_44, t_45, t_46, t_47, t_48, t_49, t_50, t_127;
MCTS 11 (M11)	t_68, t_70, t_71, t_72, t_73, t_74, t_75, t_76, t_77, t_78, t_237;
MCTS 12 (M12)	t_80, t_111, t_112;
MCTS 13 (M13)	t_81, t_84, t_85, t_86, t_87, t_88, t_89, t_90, t_131;
MCTS 14 (M14)	t_92, t_93;
MCTS 15 (M15)	t_95, t_104, t_141;
MCTS 16 (M16)	t_96, t_105, t_218;
MCTS 17 (M17)	t_97, t_224;
MCTS 18 (M18)	t_106, t_107, t_108;
MCTS 19 (M19)	t_109, t_110, t_133;
MCTS 20 (M20)	t_113, t_115, t_117, t_118, t_119;
MCTS 21 (M21)	t_134, t_135, t_136, t_137;
MCTS 22 (M22)	t_139, t_140, t_217;
MCTS 23 (M23)	t_143, t_227, t_228;
MCTS 24 (M24)	t_144, t_145, t_146;
MCTS 25 (M25)	t_147, t_148, t_149;
MCTS 26 (M26)	t_150, t_151, t_219;
MCTS 27 (M27)	t_152, t_153, t_154, t_220;
MCTS 28 (M28)	t_155, t_156, t_157, t_158, t_159, t_160, t_221, t_226;
MCTS 29 (M29)	t_161, t_163, t_164, t_166, t_213;
MCTS 30 (M30)	t_162, t_165, t_214;
MCTS 31 (M31)	t_168, t_169, t_201, t_236;
MCTS 32 (M32)	t_170, t_171, t_172, t_173, t_223;
MCTS 33 (M33)	t_175, t_176, t_180, t_230;
MCTS 34 (M34)	t_178, t_203;
MCTS 35 (M35)	t_182, t_183, t_231, t_233;
MCTS 36 (M36)	t_184, t_185, t_186, t_187, t_188, t_189, t_198;
MCTS 37 (M37)	t_193, t_194;
MCTS 38 (M38)	t_195, t_196, t_197, t_200;
MCTS 39 (M39)	t_204, t_205, t_206, t_235;
MCTS 40 (M40)	t_210, t_211;

doi:10.1371/journal.pone.0096986.t002

Table 5.9: List of MCTS and transitions from PInA.

or MCTS become inactivated. Considering our results and the literature, we decided to knock out the following pathway elements: (i) pVHL, (ii) HIF1α alone and with Sp1, (iii) t_98, (iv) PHD2, (v) MCTS1, (vi) t_97 and (vii) GSK3β. In the following, we describe the effect of each knock out scenario on our model.

(i) pVHL knock out. Degradation of HIF-1α is not completely depleted



Rank	Transitions	Biological meaning	Occurrence %
1	t_98	Input transition for Hif and VHL	95.165
2	t_116	Interaction of VHL with Elongin B and C	94.148
3	t_113	Activation by oxygen of ARD	94.094
4	t_115	Acetylation and hydroxilation of Hif	94.094
5	t_117	Interaction of complex Vcb with Cu2	94.094
6	t_118	Interaction of complex Vcb with modified Hif	94.094
7	t_119	Degradation VHL dependent of Hif	94.094
8	t_97	ATP formation	89.059
9	t_224	Water Output transition	89.059
10	t_82	Pyruvate Dehydrogenase inactivation	88.041

doi:10.1371/journal.pone.0096986.t003

Table 5.10: Ranking of the 10 most occurring transitions with biological meaning and percentage of occurrence.

due to presence of both p53- and GSK3β-dependent alternative degradation pathways. All other processes usually inhibited by pVHL take place in an uncontrolled way, including creation of VEGF via Sp1 transcription activity and increased matrix regulation due to lack of fibronectin crosslinking. Hur resulted constantly activated and nur77 can stimulate synthesis of Proopiomelanocortin, precursor for the Adrenocorticotropic hormone. Card9 increases release of tumor necrosis factor, and NF-kB when not inhibited by pVHL. Instead, Jade1 is unable to survive long enough to inhibit βcatenin, generating a proliferation signal with Wnt. Lactic acid is also not produced due to LDH enzyme production being HIF-1α transcription activity dependent.

(ii) HIF-1α knock out. VEGF is still created thanks to Sp1, thus oxygen is still generated even if in lower proportion. If HIF-1α and Sp1 are both knocked out at the same time, oxygen is quickly consumed and the metabolism is soon unable to proceed. Lactic acid is not produced due to LDH enzyme production being HIF-1α transcription activity dependent. Glycolysis and glycogen are produced normally and the metabolism is not inhibited by PyrDH negative regulation and lactic acid formation. Since pVHL is present, other tumor suppressor activities are enabled, except for proteasomal degradation of HIF-1α due to the substrate being non-existent.

(iii) HIF-1α and pVHL double knock out. This generates a situation where the metabolism is normal but oxygen regeneration is less productive, with only Sp1 acting for transcription. Due to absence of pVHL, all proliferation-stimulating processes are active, causing an unbalanced consumption of resources. Our model shows that this condition is compatible with cell growth and multiplication, but new blood vessel generation is consistently slower and glucidic metabolism appears principally based on the glycolysis reaction. Similar activity reduction applies to both tight junction and cellular external matrix (ECM) pathway regulation. It cannot be excluded that some observed effects could be mitigated by both HIF-2α and HIF-3α activity in vivo.

(iv) PHD2 knock out. The protein is involved in pVHL mediated and oxygen dependent degradation of HIF-1α. Further, PHD2 is involved in hydroxylation of the RNA polymerase II subunit Rpb1 to allow its translocation to less chromatin-concentrated areas of the nucleus. When it is knocked out, HIF-1α degradation can continue via alternative pathways as seen in the pVHL knock out experiment and there is more RNA polymerase II activity, even if rpb7 can still be inactivated by pVHL.

(v) MCTS1 knock out. MCTS1 groups some reactions involved in the HIF-1α p53-dependent degradation pathway (Table 5.9). To perform this knock out, we erased the necessary token in mdm2, making the precondition insufficient to enable the MCTS transitions. p53 is not degraded and can continue its proapoptotic signal. On the other hand, a HIF-1α degradation mechanism is also knocked out resulting in an increased HIF-1α transcription activity.

(vi) t_97 knock out. This is the ATPase transition, allowing the model to imitate oxygen consumption for ATP synthesis. If this transition is inactive, oxygen accumulates infinitely and ATP is not regenerated after few simulation steps. At the beginning, ATP is formed during the first step of glycolysis but afterwards it is consumed again. At some point, these reactions do not have any ATP available to allow the system to re-balance the consumed ATP. After few simulation steps, oxygen reaches a high level due to slower consumption in the PHD2 regulation process. Biologically, this means that the metabolism stops and the cell is not able to create energy to survive. There is no accumulation other than glucose in the model. A few oxygen creation processes are blocked as well due to absence of ATP, e.g. t_15, t_41 and t_57.

(vii) GSK3β knock out. This enzyme is involved in negative glycogen synthetase (GS) regulation and is inactivated when phosphorylated. When GSK3β is knocked out, glycogen is continuously produced due to the enzyme remaining in an active state. In a real organism there are alternative forms of GSK3β which can inactivate GS, hence the effect will be less sharp. GSK3β is also involved in the degradation of HIF-1α, causing its phosphorylation and following ubiquitination. It is also involved in the degradation of β-catenin, where it is responsible for primary phosphorylation. If knocked out, even if Jade1 can be stabilized by pVHL, the effect will be similar to a knock out of Jade1, where β-catenin is free to continue proliferation stimulating transcription activity.

5.4.3 Discussion

We started from a core model of hypoxia response [298] and extended the original network with functional data derived from the literature in order to represent a complete description of the pVHL interaction pathway according to current knowledge. VHL syndrome is characterized by the formation of tumors and cysts affecting different organism districts and tissues. Indeed, pVHL is a tumor suppressor whose functions are connected to inhibition of proliferation and survival, growth and stability of extracellular matrix and microtubules, as well as cell polarity and migration. The IntAct database reports more than 200 suspected pVHL interactors and for most of them interaction and function details remain largely unknown. We chose to model the pVHL interactions in a credible cellular context with many protein activities occurring at the same time. The main idea was to create a novel manually curated PN description of the entire VHL disease pathway, including glucidic metabolism and signaling pathways. The model was designed as a standard PN and is composed of 238 transitions and 323 places, connected by 801 edges. A biologically realistic PN model needs to be covered by T-invariants, meaning each transition in the model has to be included in a T-invariant, and each invariant needs to have a biological meaning [316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332]. We used the T-invariant analysis to validate the reliability of the model. We computed a total of 393 T-invariants, plus 10 trivial invariants, which were excluded from analysis. These were grouped into 44 Clusters and, through use of T-invariants, transitions were grouped into 40 MCTS. The model obtained is connected, covered by T-invariants with each invariant holding a biological meaning. MCTS analysis was used to identify the most frequent crucial transitions occurring in the model. This specific subset was further used to plan in silico knock out experiments and for the model validation and analysis of expected biological behavior. The model was then used to perform in silico knock out experiments inactivating specific transitions during qualitative network analysis. Our results showed that the model is able to represent important transitions reflecting real biological outcomes, i.e. transitions involving species such as oxygen or ATP are correctly inactivated under certain circumstances as expected from the bibliographic data. Biological energy-related reactions (e.g. ATP production from ADP) were modeled as P-invariants. Although the network is intentionally not covered by P-invariants, P-invariant analysis was used to verify all modeled energy consuming transitions. Both the ATP and NADH balances appeared constant during the simulation, with irrelevant P-invariants located in the Hur system. This approximation was used to verify the Hur-dependent regulation of VEGF, with results in accordance with [335]. The specific pVHL knock out suggests that this protein alone is not sufficient for complete HIF-1α inactivation. Indeed, other concurrent HIF-1α degradation pathways promote a sort of cell cycle regulation backup. On the contrary, simple deletion of pVHL turned out to be sufficient to increase all its other inhibitory functions, showing similar effects to pathological VHL symptoms. Indeed, ECM destabilization increases cell migration to other areas, promoting metastasis outbreak in case of tumor cells. Further, pVHL-dependent inhibition of tight junction formation by aPKCII participates in an easier cellular detachment. The interactions of Nur77 could be considered a good example for pathological effects. It is a stimulator of Proopiomelanocortin production, a precursor for the Adrenocorticotropic hormone. If excessively released, it promotes an overproduction of adrenergic neurotransmitters by adrenal glands. Coming at clinical condition known as Cushing syndrome. On the very long term, Nur77 deregulation is known to cause tumors of the pituitary and adrenal glands [336, 337]. This happens in pheochromocytoma, which is one of the main VHL disease manifestations. We speculate that continuous VEGF transcription, even in situations where HIF-1α (but not Sp1) is knocked out, could be the explanation for clinical studies where VEGF-targeting drugs have turned out to be effective in kidney cancer treatment as reported in [338]. Although we used only confirmed data from the literature, Nur77 may be involved in other regulation systems which were not considered in our model. The transitions for pVHL fibronectin stabilization show a behaviour which is coherent with biochemical experiments, illustrating a complete abrogation of ECM stabilization and an increased matrix metallo-proteinase action. Although the results are encouraging, the presented model will need further improvements since standard PNs do neither allow a complete transition control nor enzymatic activity modulation.

Nevertheless, thanks to its manual curation our model can be used to plan new in vitro and in vivo experiments. The results are convincing enough to suggest our model as a comprehensive pathway model to simulate the main pVHL functions.

Chapter 6

Conclusions

In this thesis, I explored the contribution of data science to meet biologists need. With millions of protein sequences and thousands of PDBs, knowledge discovery requires strong efforts in order to combine different data sources and learn natural laws. My research was mainly divided into two activities: the *analysis* of data and the *development* of computational methods. These two process are intrinsically related, as one's output is the input for the other. I developed four machine learning methods for gene and protein automatic classification and annotation called BOOGIE, NeEMO, RING MD and INGA. These tools are radically diverse in terms of data sources (exome data, PDB structures and on-line database), since they address completely different problems. In BOOGIE, gene variants are used to predict human blood groups using an optimized version of nearest neighbour algorithm. RING MD and NeEMO are designed to study a protein residue interaction network. The former detects key amino acids and interactions using k-means algorithm and Hidden Markov Models on molecular dynamics data, while the latter predicts stability changes due to mutations by means of neural networks. INGA is a consensus of three information retrieval systems that exploit public databases to predict protein function. It uses a combination of generalized additive models and Bayesian methods to achieve higher predictive accuracy than previously published tools, as shown during the Critical Assessment of Function Annotation.

In terms of data analysis, I focused on flexible protein regions called intrinsic disorder and on their characterization. Firstly, a dataset with one order of magnitude more samples than anyone previously published was gathered. Such a collection was used to assess the disorder prediction accuracy for state-of-the-art methods. Their limitations and bias were highlighted especially for long disordered regions, which were well represented and deserved themselves a second analysis.

In the midst between exome data analysis and genetic disease risk prioritization lays the Critical Assessment of Genome Interpretation (CAGI). CAGI organizers published datasets of human genetic studies where cases and controls were taken into account. The goal was the automatic classification of samples with Crohn's disease, familial combined hyperlipidemia and other illnesses. After a first analysis and selection of relevant mutations, a semi-automatic prioritization approach based on patients clustering was proposed, which was shown to be among the top methods.

To conclude, data science seems today a very consolidated approach spanning different areas of biology. It promotes data reuse and can simplify complex tasks, enabling previously unthinkable analysis. The biggest challenges for the future are surely data openness and integration, mainly due to the coordination of thousands research groups and institutes. The gap between database content and the knowledge encoded using natural language inside papers is also an unsolved problem where lots of work is done by human curators. In terms of statistical issues, correlation and causality are different concepts that might be confused when dealing with an high volume of data. Statistical significance and effect size became also a big limitation in the context of big data, since research is trying to address very complex topics with noisy data. Thus, computer science and statistics will be once again among the main players for science development in the coming years, but their efforts must address inter-disciplinary challenges where the only solution is a skill mixture. Bioinformatics seems a natural example for biology, and will become a central area of expertise for life science applied research.

Bibliography

[1] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp, "The Ensembl genome database project," *Nucleic Acids Res.*, vol. 30, pp. 38–41, Jan 2002.

[2] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean, D. M. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, and et al, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, pp. 56–65, Nov 2012.

[3] S. Sherry, M. Ward, M. Kholodov, J. Baker, L. Phan, E. Smigielski, and K. Sirotkin, "dbsnp: the ncbi database of genetic variation," *Nucleic Acids Res*, vol. 29, no. 1, pp. 308–311, 2001. Jan.

[4] T. Beck, R. K. Hastings, S. Gollapudi, R. C. Free, and A. J. Brookes, "GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies," *Eur. J. Hum. Genet.*, vol. 22, pp. 949–952, Jul 2014.

[5] R. Apweiler, M. Jesus Martin, C. O'onovan, M. Magrane, Y. Alam-Faruque, R. Antunes, E. Barrera Casanova, B. Bely, M. Bingley, L. Bower, B. Bursteinas, W. Mun Chan, G. Chavali, A. Da Silva, E. Dimmer, R. Eberhardt, F. Fazzini, A. Fedotov, J. Garavelli, L. G.

Castro, M. Gardner, R. Hieta, R. Huntley, J. Jacobsen, and et al, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Res.*, vol. 40, pp. D71–75, Jan 2012.

[6] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen, "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Res.*, vol. 41, pp. D808–815, Jan 2013.

[7] P. W. Rose, C. Bi, W. F. Bluhm, C. H. Christie, D. Dimitropoulos, S. Dutta, R. K. Green, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, A. G. Ramos, J. D. Westbrook, J. Young, C. Zardecki, H. M. Berman, and P. E. Bourne, "The RCSB protein data bank: new resources for research and education," vol. 41, pp. D475–482. PMID: 23193259.

[8] X. M. Fernandez-Suarez, D. J. Rigden, and M. Y. Galperin, "The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection," *Nucleic Acids Res.*, vol. 42, pp. 1–6, Jan 2014.

[9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[10] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Res.*, vol. 22, pp. 1760–1774, Sep 2012.

[11] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[12] C. Zhang and S. Zhang, *Association rule mining: models and algorithms*. Springer-Verlag, 2002.

[13] V. J. Hodge and J. Austin, “A survey of outlier detection methodologies,” *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[14] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: A review of classification techniques,” 2007.

[15] H. Barlow, “Unsupervised learning,” *Neural computation*, vol. 1, no. 3, pp. 295–311, 1989.

[16] A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 1998.

[17] S. Haykin, “A comprehensive foundation,” *Neural Networks*, vol. 2, no. 2004, 2004.

[18] J. A. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[19] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[20] R. Durbin, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[21] C. M. Bishop *et al.*, *Pattern recognition and machine learning*, vol. 1. springer New York, 2006.

[22] R. Bonneau and D. Baker, “Ab initio protein structure prediction: progress and prospects,” *Annual review of biophysics and biomolecular structure*, vol. 30, no. 1, pp. 173–189, 2001.

[23] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch, “Swiss-model: an automated protein homology-modeling server,” *Nucleic acids research*, vol. 31, no. 13, pp. 3381–3385, 2003.

[24] G. Ramachandran, C. t. Ramakrishnan, and V. Sasisekharan, “Stereochemistry of polypeptide chain configurations,” *Journal of molecular biology*, vol. 7, no. 1, pp. 95–99, 1963.

[25] R. Apweiler, A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, E. Alpi, R. Antunes, J. Arganiska, E. Barrera Casanova, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, W. Mun Chan, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, F. Fazzini, P. Gane, L. G. Castro, P. Garmiri, and et al, "Activities at the Universal Protein Resource (UniProt)," *Nucleic Acids Res.*, vol. 42, pp. D191–198, Jan 2014.

[26] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 89, pp. 10915–10919, Nov 1992.

[27] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, Oct 1990.

[28] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Druke, "Solving the protein sequence metric problem," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 6395–6400, May 2005.

[29] P. C. Ng and S. Henikoff, "Sift: Predicting amino acid changes that affect protein function," *Nucleic acids research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[30] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983. Dec.

[31] M. L. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids," *Science*, vol. 221, no. 4612, pp. 709–713, 1983.

[32] T. Lengauer and M. Rarey, "Computational methods for biomolecular docking," *Current opinion in structural biology*, vol. 6, no. 3, pp. 402–406, 1996.

[33] E. Capriotti, P. Fariselli, and R. Casadio, "I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure," *Nucleic Acids Res*, vol. 33, no. suppl 2, pp. W306–W310, 2005. Jul.

[34] L. Holm and C. Sander, "Mapping the protein universe," *Science*, vol. 273, no. 5275, pp. 595–602, 1996.

[35] A. Martin, M. Vidotto, F. Boscariol, D. Domenico, I. Walsh, and S. Tosatto, “Ring: networking interacting residues, evolutionary information and energetics in protein structures,” *Bioinformatics*, vol. 27, no. 14, pp. 2003–2005, 2011. Jul.

[36] A. del Sol, H. Fujihashi, D. Amoros, and R. Nussinov, “Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families,” *Protein Sci.*, vol. 15, pp. 2120–2128, Sep 2006.

[37] G. B. Gloor, L. C. Martin, L. M. Wahl, and S. D. Dunn, “Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions,” *Biochemistry*, vol. 44, no. 19, pp. 7156–7165, 2005.

[38] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,” *Nat. Genet.*, vol. 25, pp. 25–29, May 2000.

[39] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Toronen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, and et al, “A large-scale evaluation of computational protein function prediction,” *Nat. Methods*, vol. 10, pp. 221–227, Mar 2013.

[40] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta, “Pfam: the protein families database,” *Nucleic Acids Res.*, vol. 42, pp. D222–230, Jan 2014.

[41] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, “Toppgene suite for gene list enrichment analysis and candidate gene prioritization,” *Nucleic acids research*, vol. 37, no. suppl 2, pp. W305–W311, 2009.

[42] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, “All-atom empirical potential for molecular modeling and dynamics studies of proteins,” *J Phys Chem B*, vol. 102, pp. 3586–3616, Apr 1998.

[43] J. Habchi, P. Tompa, S. Longhi, and V. N. Uversky, “Introducing protein intrinsic disorder,” *Chem. Rev.*, vol. 114, pp. 6561–6588, Jul 2014.

[44] K. Bava, M. Gromiha, H. Uedaira, K. Kitajima, and A. Sarai, “Protherm, version 4.0: thermodynamic database for proteins and mutants,” *Nucleic Acids Res*, vol. 32, no. Database, pp. D120–121, 2004. Jan.

[45] E. R. Mardis, “The $1,000 genome, the$100,000 analysis?,” *Genome Med*, vol. 2, no. 11, p. 84, 2010.

[46] K. Wang, M. Li, and M. Bucan, “Pathway-based approaches for analysis of genomewide association studies,” *Am. J. Hum. Genet.*, vol. 81, pp. 1278–1283, Dec 2007.

[47] S. K. Patnaik, W. Helmberg, and O. O. Blumenfeld, “BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems,” *Nucleic Acids Res.*, vol. 40, pp. D1023–1029, Jan 2012.

[48] C. Chaouiya, “Petri net modelling of biological networks,” *Briefings in Bioinformatics*, vol. 8, no. 4, pp. 210–219, 2007.

[49] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, “The pfam protein families database,” vol. 40, pp. D290–D301.

[50] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, “Assigning protein functions by comparative genome analysis: protein phylogenetic profiles,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, pp. 4285–4288, Apr 1999.

[51] B. E. Engelhardt, M. I. Jordan, J. R. Srouji, and S. E. Brenner, "Genome-scale phylogenetic function annotation of large and diverse protein families," *Genome Res.*, vol. 21, pp. 1969–1980, Nov 2011.

[52] D. Cozzetto, D. W. Buchan, K. Bryson, and D. T. Jones, "Protein function prediction by massive integration of evolutionary analyses and multiple data sources," *BMC Bioinformatics*, vol. 14 Suppl 3, p. S1, 2013.

[53] D. Piovesan, P. L. Martelli, P. Fariselli, A. Zauli, I. Rossi, and R. Casadio, "BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences," *Nucleic Acids Res.*, vol. 39, pp. 197–202, Jul 2011.

[54] F. Minneci, D. Piovesan, D. Cozzetto, and D. T. Jones, "FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences," *PLoS ONE*, vol. 8, no. 5, p. e63754, 2013.

[55] W. T. Clark and P. Radivojac, "Analysis of protein function and its prediction from amino acid sequence," *Proteins*, vol. 79, pp. 2086–2096, Jul 2011.

[56] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, and et al, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, pp. 180–183, Jan 2002.

[57] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, pp. 1497–1502, Jun 2007.

[58] H. Zhu and M. Snyder, "Protein chip technology," *Curr Opin Chem Biol*, vol. 7, pp. 55–63, Feb 2003.

[59] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq, "Functional classification of proteins for the prediction of cellular func-

tion from a protein-protein interaction network," *Genome Biol.*, vol. 5, no. 1, p. R6, 2003.

[60] H. N. Chua, W. K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions," *Bioinformatics*, vol. 22, pp. 1623–1630, Jul 2006.

[61] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," *J. Comput. Biol.*, vol. 10, no. 6, pp. 947–960, 2003.

[62] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi, "Assessment of prediction accuracy of protein function from protein–protein interaction data," *Yeast*, vol. 18, pp. 523–531, Apr 2001.

[63] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh, "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps," *Bioinformatics*, vol. 21 Suppl 1, pp. i302–310, Jun 2005.

[64] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M. C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, and et al, "The UniProt-GO Annotation database in 2011," *Nucleic Acids Res.*, vol. 40, pp. D565–570, Jan 2012.

[65] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "UniRef: comprehensive and non-redundant UniProt reference clusters," *Bioinformatics*, vol. 23, pp. 1282–1288, May 2007.

[66] W. T. Clark and P. Radivojac, "Information-theoretic evaluation of predicted ontological annotations," *Bioinformatics*, vol. 29, pp. 53–61, Jul 2013.

[67] C. C. Huang, E. C. Meng, J. H. Morris, E. F. Pettersen, and T. E. Ferrin, "Enhancing UCSF Chimera through web services," *Nucleic Acids Res.*, vol. 42, pp. W478–484, Jul 2014.

[68] N. Guex, M. C. Peitsch, and T. Schwede, "Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective," *Electrophoresis*, vol. 30 Suppl 1, pp. S162–173, Jun 2009.

[69] A. A. Canutescu and R. L. Dunbrack, "MollDE: a homology modeling framework you can click with," *Bioinformatics*, vol. 21, pp. 2914–2916, Jun 2005.

[70] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *J Mol Graph*, vol. 14, pp. 33–38, Feb 1996.

[71] S. Loriot, F. Cazals, and J. Bernauer, "ESBTL: efficient PDB parser and data structure for the structural and geometric analysis of biological macromolecules," *Bioinformatics*, vol. 26, pp. 1127–1128, Apr 2010.

[72] R. Grunberg, M. Nilges, and J. Leckner, "Biskit–a software platform for structural bioinformatics," *Bioinformatics*, vol. 23, pp. 769–770, Mar 2007.

[73] A. Saladin, S. Fiorucci, P. Poulain, C. Prevost, and M. Zacharias, "PTools: an opensource molecular docking library," *BMC Struct. Biol.*, vol. 9, p. 27, 2009.

[74] M. Biasini, V. Mariani, J. Haas, S. Scheuber, A. D. Schenk, T. Schwede, and A. Philippsen, "OpenStructure: a flexible software framework for computational structural biology," *Bioinformatics*, vol. 26, pp. 2626–2628, Oct 2010.

[75] D. W. Kulp, S. Subramaniam, J. E. Donald, B. T. Hannigan, B. K. Mueller, G. Grigoryan, and A. Senes, "Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL)," *J Comput Chem*, vol. 33, pp. 1645–1661, Jul 2012.

[76] Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P. Ren, "The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins," *J Chem Theory Comput*, vol. 9, no. 9, pp. 4046–4063, 2013.

[77] A. Hildebrandt, A. K. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N. C. Toussaint, A. Moll, D. Stockel, S. Nickels, S. C. Mueller, H. P.

Lenhof, and O. Kohlbacher, "BALL–biochemical algorithms library 1.3," *BMC Bioinformatics*, vol. 11, p. 531, 2010.

[78] J. M. Chandonia, "StrBioLib: a Java library for development of custom computational structural biology applications," *Bioinformatics*, vol. 23, pp. 2018–2020, Aug 2007.

[79] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)–round x," *Proteins*, vol. 82 Suppl 2, pp. 1–6, Feb 2014.

[80] G. A. Khoury, A. Liwo, F. Khatib, H. Zhou, G. Chopra, J. Bacardit, L. O. Bortot, R. A. Faccioli, X. Deng, Y. He, P. Krupa, J. Li, M. A. Mozolewska, A. K. Sieradzan, J. Smadbeck, T. Wirecki, S. Cooper, J. Flatten, K. Xu, D. Baker, J. Cheng, A. C. Delbem, C. A. Floudas, C. Keasar, M. Levitt, Z. Popovi?, H. A. Scheraga, J. Skolnick, and S. N. Crivelli, "WeFold: a coopetition for protein structure prediction," *Proteins*, vol. 82, pp. 1850–1868, Sep 2014.

[81] G. Wang and R. L. Dunbrack, "Scoring profile-to-profile sequence alignments," *Protein Sci.*, vol. 13, pp. 1612–1626, Jun 2004.

[82] S. Tosatto, "The victor/frst function for model quality estimation," *J Comput Biol J Comput Mol Cell Biol*, vol. 12, no. 10, pp. 1316–1327, 2005. Dec.

[83] S. C. Tosatto, E. Bindewald, J. Hesser, and R. Manner, "A divide and conquer approach to fast loop modeling," *Protein Eng.*, vol. 15, pp. 279–286, Apr 2002.

[84] A. Kryshtafovych, K. Fidelis, and J. Moult, "CASP10 results compared to those of previous CASP experiments," *Proteins*, vol. 82 Suppl 2, pp. 164–174, Feb 2014.

[85] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-oriented Software*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1995.

[86] S. Tosatto and R. Battistutta, "Tap score: torsion angle propensity normalization applied to local protein structure evaluation," *BMC Bioinformatics*, vol. 8, p. 155, 2007.

[87] S. C. Tosatto, A. Albiero, A. Mantovan, C. Ferrari, E. Bindewald, and S. Toppo, "Align: a C++ class library and web server for rapid sequence alignment prototyping," *Curr Drug Discov Technol*, vol. 3, pp. 167–173, Sep 2006.

[88] P. Fontana, E. Bindewald, S. Toppo, R. Velasco, G. Valle, and S. C. Tosatto, "The SSEA server for protein secondary structure alignment," *Bioinformatics*, vol. 21, pp. 393–395, Feb 2005.

[89] M. S. Madhusudhan, M. A. Marti-Renom, R. Sanchez, and A. Sali, "Variable gap penalty for protein sequence-structure alignment," *Protein Eng. Des. Sel.*, vol. 19, pp. 129–133, Mar 2006.

[90] S. R. Sunyaev, F. Eisenhaber, I. V. Rodchenkov, B. Eisenhaber, V. G. Tumanyan, and E. N. Kuznetsov, "PSIC: profile extraction from sequence alignments with position-specific counts of independent observations," *Protein Eng.*, vol. 12, pp. 387–394, May 1999.

[91] M. Giollo, A. J. Martin, I. Walsh, C. Ferrari, and S. C. Tosatto, "NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation," *BMC Genomics*, vol. 15 Suppl 4, p. S7, 2014.

[92] G. Minervini, A. Masiero, S. Moro, and S. C. Tosatto, "In silico investigation of PHD-3 specific HIF1- proline 567 hydroxylation: a new player in the VHL/HIF-1 interaction pathway?," *FEBS Lett.*, vol. 587, pp. 2996–3001, Sep 2013.

[93] M. C. Scaini, G. Minervini, L. Elefanti, P. Ghiorzo, L. Pastorino, S. Tognazzo, S. Agata, M. Quaggio, D. Zullato, G. Bianchi-Scarra, M. Montagna, E. D'Andrea, C. Menin, and S. C. Tosatto, "CDKN2A unclassified variants in familial malignant melanoma: combining functional and computational approaches for their assessment," *Hum. Mutat.*, vol. 35, pp. 828–840, Jul 2014.

[94] P. Das, S. G. Kang, S. Temple, and G. Belfort, "Interaction of amyloid inhibitor proteins with amyloid Beta peptides: insight from molecular dynamics simulations," *PLoS ONE*, vol. 9, no. 11, p. e113041, 2014.

[95] M. S. Liu, D. Wang, H. Morimoto, H. C. Yim, A. T. Irving, B. R. Williams, and A. J. Sadler, "Molecular dynamics reveal a novel kinase-substrate interface that regulates protein translation," *J Mol Cell Biol*, vol. 6, pp. 473–485, Dec 2014.

[96] A. Masiero, S. Aufiero, G. Minervini, S. Moro, R. Costa, and S. C. Tosatto, "Evaluation of the steric impact of flavin adenine dinucleotide in Drosophila melanogaster cryptochrome function," *Biochem. Biophys. Res. Commun.*, vol. 450, pp. 1606–1611, Aug 2014.

[97] R. M. Betz and R. C. Walker, "Paramfit: Automated optimization of force field parameters for molecular dynamics simulations," *J Comput Chem*, vol. 36, pp. 79–87, Jan 2015.

[98] B. H. Varkuti, Z. Yang, and A. Malnasi-Csizmadia, "Structural model of weak-binding actomyosin in the prepowerstroke state," *J. Biol. Chem.*, Nov 2014.

[99] J. M. Karp, E. Erylimaz, and D. Cowburn, "Correlation of chemical shifts predicted by molecular dynamics simulations for partially disordered proteins," *J. Biomol. NMR*, vol. 61, pp. 35–45, Jan 2015.

[100] A. J. Martin, M. Vidotto, F. Boscariol, T. Di Domenico, I. Walsh, and S. C. Tosatto, "RING: networking interacting residues, evolutionary information and energetics in protein structures," *Bioinformatics*, vol. 27, pp. 2003–2005, Jul 2011.

[101] N. T. Doncheva, Y. Assenov, F. S. Domingues, and M. Albrecht, "Topological analysis and interactive visualization of biological networks and protein structures," *Nat Protoc*, vol. 7, pp. 670–685, Apr 2012.

[102] M. Pasi, M. Tiberti, A. Arrigoni, and E. Papaleo, "xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures," *J Chem Inf Model*, vol. 52, pp. 1865–1874, Jul 2012.

[103] M. Tiberti, G. Invernizzi, M. Lambrughi, Y. Inbar, G. Schreiber, and E. Papaleo, "PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins," *J Chem Inf Model*, vol. 54, pp. 1537–1551, May 2014.

[104] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, "Scalable molecular dynamics with NAMD," *J Comput Chem*, vol. 26, pp. 1781–1802, Dec 2005.

[105] I. G. Tsoulos and A. Stavrakoudis, "Eucb: A c++ program for molecular dynamics trajectory analysis," *Computer Physics Communications*, vol. 182, no. 3, pp. 834 – 841, 2011.

[106] N. M. Glykos, "Software news and updates. Carma: a molecular dynamics analysis program," *J Comput Chem*, vol. 27, pp. 1765–1768, Nov 2006.

[107] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, "Structure of ubiquitin refined at 1.8 A resolution," *J. Mol. Biol.*, vol. 194, pp. 531–544, Apr 1987.

[108] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651 – 666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR) 19th International Conference in Pattern Recognition (ICPR).

[109] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio, "Prediction of contact maps with neural networks and correlated mutations," *Protein Eng.*, vol. 14, pp. 835–843, Nov 2001.

[110] A. M. Monzon, E. Juritz, M. S. Fornasari, and G. Parisi, "CoDNaS: a database of conformational diversity in the native state of proteins," *Bioinformatics*, vol. 29, pp. 2512–2514, Oct 2013.

[111] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. Caves, "Bio3d: an R package for the comparative analysis of protein structures," *Bioinformatics*, vol. 22, pp. 2695–2696, Nov 2006.

[112] M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A. K. Dunker, I. C. Felli, J. D. Forman-Kay, R. W. Kriwacki, R. Pierattelli, J. Sussman, D. I. Svergun, V. N. Uversky, M. Vendruscolo, D. Wishart, P. E. Wright, and P. Tompa, "pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins," *Nucleic Acids Res.*, vol. 42, pp. D326–335, Jan 2014.

[113] M. Mandal and C. Mukhopadhyay, “Microsecond molecular dynamics simulation of guanidinium chloride induced unfolding of ubiquitin,” *Phys Chem Chem Phys*, vol. 16, pp. 21706–21716, Oct 2014.

[114] L. A. Ralat, V. Kalas, Z. Zheng, R. D. Goldman, T. R. Sosnick, and W. J. Tang, “Ubiquitin is a novel substrate for human insulin-degrading enzyme,” *J. Mol. Biol.*, vol. 406, pp. 454–466, Feb 2011.

[115] D. Komander, “The emerging complexity of protein ubiquitination,” *Biochem. Soc. Trans.*, vol. 37, pp. 937–953, Oct 2009.

[116] J. C. Wootton, “Non-globular domains in protein sequences: Automated segmentation using complexity measures,” vol. 18, pp. 269–285.

[117] J. Jorda and A. V. Kajava, *Protein Homorepeats: Sequences, Structures, Evolution, and Functions*, vol. Volume 79 of *Advances in Protein Chemistry and Structural Biology*, pp. 59–88. Academic Press.

[118] M. Gribskov, A. D. McLachlan, and D. Eisenberg, “Profile analysis: detection of distantly related proteins,” vol. 84, pp. 4355–4358. PMID: 3474607.

[119] A. Biegert and J. Sding, “De novo identification of highly diverged protein repeats by probabilistic consistency,” vol. 24, pp. 807–814. PMID: 18245125.

[120] E. Schaper, A. V. Kajava, A. Hauser, and M. Anisimova, “Repeat or not repeat?–statistical validation of tandem repeat prediction in genomic sequences,” vol. 40, pp. 10005–10017. PMID: 22923522 PMCID: PMC3488214.

[121] J. Buard and G. Vergnaud, “Complex recombination events at the hypermutable minisatellite CEB1 (D2S90).,” vol. 13, pp. 3203–3210. PMID: 8039512 PMCID: PMC395212.

[122] M. A. Andrade, C. Perez-Iratxeta, and C. P. Ponting, “Protein repeats: Structures, functions, and evolution,” vol. 134, pp. 117–131.

[123] A. V. Kajava and A. C. Steven, *β-Rolls, β-Helices, and Other β-Solenoid Proteins*, vol. Volume 73 of *Fibrous Proteins: Amyloids, Prions and Beta Proteins*, pp. 55–96. Academic Press.

[124] J. de Wit, W. Hong, L. Luo, and A. Ghosh, "Role of leucine-rich repeat proteins in the development and function of neural circuits," vol. 27, pp. 697–729. PMID: 21740233.

[125] E. R. Main, A. R. Lowe, S. G. Mochrie, S. E. Jackson, and L. Regan, "A recurring theme in protein engineering: the design, stability and folding of repeat proteins," vol. 15, pp. 464–471.

[126] N. Stefan, P. Martin-Killias, S. Wyss-Stoeckle, A. Honegger, U. Zangemeister-Wittke, and A. Plckthun, "DARPins recognizing the tumor-associated antigen EpCAM selected by phage and ribosome display and engineered for multivalency," vol. 413, pp. 826–843.

[127] Y. Javadi and L. S. Itzhaki, "Tandem-repeat proteins: regularity plus modularity equals design-ability," vol. 23, pp. 622–631.

[128] E. M. Marcotte, M. Pellegrini, T. O. Yeates, and D. Eisenberg, "A census of protein repeats," vol. 293, pp. 151–160.

[129] A. V. Kajava, "Tandem repeats in proteins: From sequence to structure," vol. 179, pp. 279–288.

[130] A. Bateman, A. G. Murzin, and S. A. Teichmann, "Structure and distribution of pentapeptide repeats in bacteria," vol. 7, p. 14771480.

[131] J. Bella, K. L. Hindle, P. A. McEwan, and S. C. Lovell, "The leucine-rich repeat structure," vol. 65, pp. 2307–2333.

[132] B. Kobe and A. V. Kajava, "The leucine-rich repeat as a protein recognition motif," vol. 11, pp. 725–732.

[133] R. Tewari, E. Bailes, K. A. Bunting, and J. C. Coates, "Armadillo-repeat protein functions: questions for little creatures," vol. 20, pp. 470–481.

[134] A. V. Kajava, C. Gorbea, J. Ortega, M. Rechsteiner, and A. C. Steven, "New HEAT-like repeat motifs in proteins regulating proteasome structure and function," vol. 146, pp. 425–430.

[135] M. A. Andrade, C. Petosa, S. I. ODonoghue, C. W. Mller, and P. Bork, "Comparison of ARM and HEAT protein repeats," vol. 309, pp. 1–18.

[136] B. Kobe and A. V. Kajava, “When protein folding is simplified to protein coiling: the continuum of solenoid protein structures,” vol. 25, pp. 509–515.

[137] . K. Bjrklund, D. Ekman, and A. Elofsson, “Expansion of protein domain repeats,” vol. 2, p. e114.

[138] M. Remmert, A. Biegert, D. Linke, A. N. Lupas, and J. Sding, “Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin,” vol. 27, pp. 1348–1358. PMID: 20106904.

[139] Z. Jawad and M. Paoli, “Novel sequences propel familiar folds,” vol. 10, pp. 447–454.

[140] I. Chaudhuri, J. Sding, and A. N. Lupas, “Evolution of the β-propeller fold,” vol. 71, p. 795803.

[141] H. M. Berman, G. J. Kleywegt, H. Nakamura, and J. L. Markley, “The future of the protein data bank,” vol. 99, p. 218222.

[142] B. H. Dessailly, R. Nair, L. Jaroszewski, J. E. Fajardo, A. Kouranov, D. Lee, A. Fiser, A. Godzik, B. Rost, and C. Orengo, “PSI-2: structural genomics to cover protein domain family space,” vol. 17, pp. 869–881.

[143] K. B. Murray, W. R. Taylor, and J. M. Thornton, “Toward the detection and validation of repeats in protein structure,” vol. 57, p. 365380.

[144] R. G. Parra, R. Espada, I. E. Snchez, M. J. Sippl, and D. U. Ferreiro, “Detecting repetitions and periodicities in proteins by tiling the structural space,” vol. 117, pp. 12887–12897.

[145] L. Marsella, F. Sirocco, A. Trovato, F. Seno, and S. C. E. Tosatto, “REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete fourier transform,” vol. 25, pp. i289–i295. PMID: 19478001.

[146] R. Szklarczyk and J. Heringa, “Tracking repeats using significance and transitivity,” vol. 20 Suppl 1, pp. i311–317. PMID: 15262814.

[147] A. Heger and L. Holm, “Rapid automatic detection and alignment of repeats in protein sequences,” vol. 41, p. 224237.

[148] A.-L. Abraham, E. P. C. Rocha, and J. Pothier, "Swelfe: a detector of internal repeats in sequences and structures," vol. 24, pp. 1536–1537. PMID: 18487242 PMCID: PMC2718673.

[149] I. Walsh, F. G. Sirocco, G. Minervini, T. Di Domenico, C. Ferrari, and S. C. E. Tosatto, "RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures," vol. 28, pp. 3257–3264. PMID: 22962341.

[150] I. Sillitoe, A. L. Cuff, B. H. Dessailly, N. L. Dawson, N. Furnham, D. Lee, J. G. Lees, T. E. Lewis, R. A. Studer, R. Rentzsch, C. Yeats, J. M. Thornton, and C. A. Orengo, "New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures," vol. 41, pp. D490–D498. PMID: 23203873 PMCID: PMC3531114.

[151] A. Andreeva, D. Howorth, J.-M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "Data growth and its impact on the SCOP database: new developments," vol. 36, pp. D419–425. PMID: 18000004.

[152] J. Jorda, T. Baudrand, and A. V. Kajava, "PRDB: protein repeat DataBase," vol. 12, p. 13331336.

[153] H. Luo, K. Lin, A. David, H. Nijveen, and J. A. M. Leunissen, "ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins," vol. 40, pp. D394–399. PMID: 22102581.

[154] I. Letunic, T. Doerks, and P. Bork, "SMART 7: recent updates to the protein domain annotation resource," vol. 40, pp. D302–305. PMID: 22053084.

[155] J. Mistry, P. Coggill, R. Y. Eberhardt, A. Deiana, A. Giansanti, R. D. Finn, A. Bateman, and M. Punta, "The challenge of increasing pfam coverage of the human proteome," vol. 2013. PMID: null PMCID: PMC3630804.

[156] J. Gmez, L. J. Garca, G. A. Salazar, J. Villaveces, S. Gore, A. Garca, M. J. Martn, G. Launay, R. Alcntara, N. Del-Toro, M. Dumousseau, S. Orchard, S. Velankar, H. Hermjakob, C. Zong, P. Ping, M. Corpas,

and R. C. Jimnez, “BioJS: an open source JavaScript framework for biological data visualization,” vol. 29, pp. 1103–1104. PMID: 23435069.

[157] T. Di Domenico, I. Walsh, A. J. M. Martin, and S. C. E. Tosatto, “MobiDB: a comprehensive database of intrinsic protein disorder annotations,” vol. 28, pp. 2080–2081.

[158] P. E. Wright and H. J. Dyson, “Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm,” *J. Mol. Biol.*, vol. 293, pp. 321–331, Oct 1999.

[159] A. Schlessinger, C. Schaefer, E. Vicedo, M. Schmidberger, M. Punta, and B. Rost, “Protein disorder–a breakthrough invention of evolution?,” *Curr. Opin. Struct. Biol.*, vol. 21, pp. 412–418, Jun 2011.

[160] A. K. Dunker and Z. Obradovic, “The protein trinity–linking function and disorder,” *Nat. Biotechnol.*, vol. 19, pp. 805–806, Sep 2001.

[161] P. Tompa, “Intrinsically unstructured proteins,” *Trends Biochem. Sci.*, vol. 27, pp. 527–533, Oct 2002.

[162] A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown, “Intrinsic protein disorder in complete genomes,” *Genome Inform Ser Workshop Genome Inform*, vol. 11, pp. 161–171, 2000.

[163] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovi?, “Intrinsic disorder and protein function,” *Biochemistry*, vol. 41, pp. 6573–6582, May 2002.

[164] P. Tompa, M. Fuxreiter, C. J. Oldfield, I. Simon, A. K. Dunker, and V. N. Uversky, “Close encounters of the third kind: disordered domains and the interactions of proteins,” *Bioessays*, vol. 31, pp. 328–335, Mar 2009.

[165] Z. Dosztanyi, J. Chen, A. K. Dunker, I. Simon, and P. Tompa, “Disorder and sequence repeats in hub proteins and their implications for network evolution,” *J. Proteome Res.*, vol. 5, pp. 2985–2995, Nov 2006.

[166] P. E. Wright and H. J. Dyson, “Linking folding and binding,” *Curr. Opin. Struct. Biol.*, vol. 19, pp. 31–38, Feb 2009.

[167] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Flavors of protein disorder," *Proteins*, vol. 52, pp. 573–584, Sep 2003.

[168] A. J. Martin, I. Walsh, and S. C. Tosatto, "MOBI: a web server to define and visualize structural mobility in NMR protein ensembles," *Bioinformatics*, vol. 26, pp. 2916–2917, Nov 2010.

[169] J. Bellay, S. Han, M. Michaut, T. Kim, M. Costanzo, B. J. Andrews, C. Boone, G. D. Bader, C. L. Myers, and P. M. Kim, "Bringing order to protein disorder through comparative genomics and genetic interactions," *Genome Biol.*, vol. 12, no. 2, p. R14, 2011.

[170] A. Mohan, V. N. Uversky, and P. Radivojac, "Influence of sequence changes and environment on intrinsically disordered proteins," *PLoS Comput. Biol.*, vol. 5, p. e1000497, Sep 2009.

[171] P. W. Rose, C. Bi, W. F. Bluhm, C. H. Christie, D. Dimitropoulos, S. Dutta, R. K. Green, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, A. G. Ramos, J. D. Westbrook, J. Young, C. Zardecki, H. M. Berman, and P. E. Bourne, "The RCSB Protein Data Bank: new resources for research and education," *Nucleic Acids Res.*, vol. 41, pp. D475–482, Jan 2013.

[172] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, "The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins," *J. Mol. Biol.*, vol. 347, pp. 827–839, Apr 2005.

[173] O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov, "Prediction of amyloidogenic and disordered regions in protein chains," *PLoS Comput. Biol.*, vol. 2, p. e177, Dec 2006.

[174] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson, "GlobPlot: Exploring protein sequences for globularity and disorder," *Nucleic Acids Res.*, vol. 31, pp. 3701–3708, Jul 2003.

[175] J. Prilusky, C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman, and J. L. Sussman, "FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded," *Bioinformatics*, vol. 21, pp. 3435–3438, Aug 2005.

[176] J. Eickholt and J. Cheng, "DNdisorder: predicting protein disorder using boosting and deep networks," *BMC Bioinformatics*, vol. 14, p. 88, 2013.

[177] S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, and T. Noguchi, "POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions," *Bioinformatics*, vol. 23, pp. 2046–2053, Aug 2007.

[178] T. Ishida and K. Kinoshita, "PrDOS: prediction of disordered protein regions from amino acid sequence," *Nucleic Acids Res.*, vol. 35, pp. W460–464, Jul 2007.

[179] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, "Protein disorder prediction: implications for structural proteomics," *Structure*, vol. 11, pp. 1453–1459, Nov 2003.

[180] I. Walsh, A. J. Martin, T. Di Domenico, A. Vullo, G. Pollastri, and S. C. Tosatto, "CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs," *Nucleic Acids Res.*, vol. 39, pp. W190–196, Jul 2011.

[181] I. Walsh, A. Martin, D. Domenico, and S. Tosatto, "Espritz: accurate and fast prediction of protein disorder," *Bioinforma Oxf Engl*, vol. 28, no. 4, pp. 503–509, 2012. Feb.

[182] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *J. Mol. Biol.*, vol. 337, pp. 635–645, Mar 2004.

[183] Z. R. Yang, R. Thomson, P. McNeil, and R. M. Esnouf, "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins," *Bioinformatics*, vol. 21, pp. 3369–3376, Aug 2005.

[184] B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano, and K. Fidelis, "Assessment of protein disorder region predictions in CASP10," *Proteins*, vol. 82 Suppl 2, pp. 127–137, Feb 2014.

[185] T. Ishida and K. Kinoshita, "Prediction of disordered regions in proteins based on the meta approach," *Bioinformatics*, vol. 24, pp. 1344–1348, Jun 2008.

[186] M. J. Mizianty, W. Stach, K. Chen, K. D. Kedarisetti, F. M. Disfani, and L. Kurgan, "Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources," *Bioinformatics*, vol. 26, pp. i489–496, Sep 2010.

[187] A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, and B. Rost, "Improved disorder prediction by combination of orthogonal approaches," *PLoS ONE*, vol. 4, no. 2, p. e4433, 2009.

[188] B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker, and V. N. Uversky, "PONDR-FIT: a meta-predictor of intrinsically disordered amino acids," *Biochim. Biophys. Acta*, vol. 1804, pp. 996–1010, Apr 2010.

[189] T. Di Domenico, I. Walsh, A. J. Martin, and S. C. Tosatto, "MobiDB: a comprehensive database of intrinsic protein disorder annotations," *Bioinformatics*, vol. 28, pp. 2080–2081, Aug 2012.

[190] S. Velankar, J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M. J. Martin, and G. J. Kleywegt, "SIFTS: Structure Integration with Function, Taxonomy and Sequences resource," *Nucleic Acids Res.*, vol. 41, pp. D483–489, Jan 2013.

[191] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein intrinsic disorder," *BMC Bioinformatics*, vol. 7, p. 208, 2006.

[192] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, "Sequence complexity of disordered protein," *Proteins*, vol. 42, pp. 38–48, Jan 2001.

[193] J. C. Wootton, "Non-globular domains in protein sequences: automated segmentation using complexity measures," *Comput. Chem.*, vol. 18, pp. 269–285, Sep 1994.

[194] D. T. Jones and M. B. Swindells, "Getting the most from PSI-BLAST," *Trends Biochem. Sci.*, vol. 27, pp. 161–164, Mar 2002.

[195] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost, "A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment," *Proteins*, vol. 34, pp. 220–223, Feb 1999.

[196] M. J. Mizianty, T. Zhang, B. Xue, Y. Zhou, A. K. Dunker, V. N. Uversky, and L. Kurgan, "In-silico prediction of disorder content using hybrid sequence representation," *BMC Bioinformatics*, vol. 12, p. 245, 2011.

[197] M. Albrecht, S. C. Tosatto, T. Lengauer, and G. Valle, "Simple consensus procedures are effective and sufficient in secondary structure prediction," *Protein Eng.*, vol. 16, pp. 459–462, Jul 2003.

[198] B. Rost, "Review: protein secondary structure prediction continues to rise," *J. Struct. Biol.*, vol. 134, no. 2-3, pp. 204–218, 2001.

[199] B. Xue, L. Li, S. O. Meroueh, V. N. Uversky, and A. K. Dunker, "Analysis of structured and intrinsically disordered regions of transmembrane proteins," *Mol Biosyst*, vol. 5, pp. 1688–1702, Dec 2009.

[200] M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker, "DisProt: the Database of Disordered Proteins," *Nucleic Acids Res.*, vol. 35, pp. D786–793, Jan 2007.

[201] B. Xue, D. Blocquel, J. Habchi, A. V. Uversky, L. Kurgan, V. N. Uversky, and S. Longhi, "Structural disorder in viral proteins," *Chem. Rev.*, vol. 114, pp. 6880–6911, Jul 2014.

[202] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu, "Classification of intrinsically disordered regions and proteins," *Chem. Rev.*, vol. 114, pp. 6589–6631, Jul 2014.

[203] C. J. Brown, S. Takayama, A. M. Campen, P. Vise, T. W. Marshall, C. J. Oldfield, C. J. Williams, and A. K. Dunker, "Evolutionary rate

heterogeneity in proteins with long disordered regions," *J. Mol. Evol.*, vol. 55, pp. 104–110, Jul 2002.

[204] S. Fukuchi, T. Amemiya, S. Sakamoto, Y. Nobe, K. Hosoda, Y. Kado, S. D. Murakami, R. Koike, H. Hiroaki, and M. Ota, "IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners," *Nucleic Acids Res.*, vol. 42, pp. D320–325, Jan 2014.

[205] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nat. Rev. Mol. Cell Biol.*, vol. 6, pp. 197–208, Mar 2005.

[206] C. J. Brown, A. K. Johnson, A. K. Dunker, and G. W. Daughdrill, "Evolution and disorder," *Curr. Opin. Struct. Biol.*, vol. 21, pp. 441–446, Jun 2011.

[207] J. A. Blake, M. Dolan, H. Drabkin, D. P. Hill, N. Li, D. Sitnikov, S. Bridges, S. Burgess, T. Buza, F. McCarthy, D. Peddinti, L. Pillai, S. Carbon, H. Dietze, A. Ireland, S. E. Lewis, C. J. Mungall, P. Gaudet, R. L. Chrisholm, P. Fey, W. A. Kibbe, S. Basu, D. A. Siegele, B. K. McIntosh, D. P. Renfro, A. E. Zweifel, and et al, "Gene Ontology annotations and resources," *Nucleic Acids Res.*, vol. 41, pp. D530–535, Jan 2013.

[208] A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman, "Function and structure of inherently disordered proteins," *Curr. Opin. Struct. Biol.*, vol. 18, pp. 756–764, Dec 2008.

[209] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovi?, and A. K. Dunker, "Intrinsic disorder in cell-signaling and cancer-associated proteins," *J. Mol. Biol.*, vol. 323, pp. 573–584, Oct 2002.

[210] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic, and V. N. Uversky, "Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins," *J. Proteome Res.*, vol. 6, pp. 1917–1932, May 2007.

[211] I. Walsh, A. J. Martin, T. Di Domenico, and S. C. Tosatto, "ESpritz: accurate and fast prediction of protein disorder," *Bioinformatics*, vol. 28, pp. 503–509, Feb 2012.

[212] M. Deniziak, C. Sauter, H. D. Becker, C. A. Paulus, R. Giege, and D. Kern, "Deinococcus glutaminyl-tRNA synthetase is a chimer between proteins from an ancient and the modern pathways of aminoacyl-tRNA formation," *Nucleic Acids Res.*, vol. 35, no. 5, pp. 1421–1431, 2007.

[213] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, Sep 1997.

[214] T. Manolio, F. Collins, N. Cox, D. Goldstein, L. Hindorff, D. Hunter, M. McCarthy, E. Ramos, L. Cardon, A. Chakravarti, J. Cho, A. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. Rotimi, M. Slatkin, D. Valle, A. Whittemore, M. Boehnke, A. Clark, E. Eichler, G. Gibson, J. Haines, T. Mackay, S. McCarroll, and P. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009. Oct.

[215] F. VanPetegem, K. Duderstadt, K. Clark, M. Wang, and D. Minor, "Alanine-scanning mutagenesis defines a conserved energetic hotspot in the cavalpha1 aid-cavbeta interaction site that is critical for channel modulation," *Struct Lond Engl 1993*, vol. 16, no. 2, pp. 280–294, 2008. Feb.

[216] J. Bryson, S. Betz, H. Lu, D. Suich, H. Zhou, K. O'Neil, and W. DeGrado, "Protein design: A hierarchic approach," *Science*, vol. 270, no. 5238, pp. 935–941, 1995. Nov.

[217] K. Dill and J. MacCallum, "The protein-folding problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012. Nov.

[218] A. Bullock, J. Henckel, B. DeDecker, C. Johnson, P. Nikolova, M. Proctor, D. Lane, and A. Fersht, "Thermodynamic stability of wild-type and mutant p53 core domain," *Proc Natl Acad Sci USA*, vol. 94, no. 26, pp. 14338–14342, 1997. Dec.

[219] T. Lazaridis and M. Karplus, “Effective energy functions for protein structure prediction,” *Curr Opin Struct Biol*, vol. 10, no. 2, pp. 139–145, 2000. Apr.

[220] A. Benedix, C. Becker, B. deGroot, A. Caflisch, and R. Bockmann, “Predicting free energy changes using structural ensembles,” *Nat Methods*, vol. 6, no. 1, pp. 3–4, 2009. Jan.

[221] N. Pokala and T. Handel, “Energy functions for protein design: Adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity,” *J Mol Biol*, vol. 347, no. 1, pp. 203–227, 2005. Mar.

[222] S. Yin, F. Ding, and N. Dokholyan, “Modeling backbone flexibility improves protein stability estimation,” *StructLond Engl 1993*, vol. 15, no. 12, pp. 1567–1576, 2007. Dec.

[223] R. Guerois, J. Nielsen, and L. Serrano, “Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations,” *JMol Biol*, vol. 320, no. 2, pp. 369–387, 2002. Jul.

[224] M. Masso and I. Vaisman, “Auto-mute: web-based tools for predicting stability changes in proteins due to single amino acid replacements,” *Protein Eng Des Sel PEDS*, vol. 23, no. 8, pp. 683–687, 2010. Aug.

[225] E. Capriotti, P. Fariselli, I. Rossi, and R. Casadio, “A three-state prediction of single point mutations on protein stability changes,” *BMC Bioinformatics*, vol. 9, no. Suppl 2, p. S6, 2008. Mar.

[226] J. Cheng, A. Randall, and P. Baldi, “Prediction of protein stability changes for single-site mutations using support vector machines,” *Proteins*, vol. 62, no. 4, pp. 1125–1132, 2006. Mar.

[227] Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, and M. Rooman, “Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: Popmusic-2.0,” *Bioinforma Oxf Engl*, vol. 25, no. 19, pp. 2537–2543, 2009. Oct.

[228] M. Musavi, W. Ahmed, K. Chan, K. Faris, and D. Hummels, "On the training of radial basis function classifiers," *Neural Netw*, vol. 5, no. 4, pp. 595–603, 1992. Jul.

[229] T. Cheng, Y. Lu, M. Vendruscolo, P. Lio, and T. Blundell, "Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms," *PLoS Comput Biol*, vol. 4, no. 7, p. e1000135, 2008. Jul.

[230] A. Martin, I. Walsh, T. Domenico, I. Mi010Deti0107, and S. Tosatto, "Panada: protein association network annotation, determination and analysis," *PloS One*, vol. 8, no. 11, p. e78383, 2013.

[231] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," 1999. 11 Nov 23 Oct 2013.

[232] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, 1997. Sep.

[233] P. Benkert, S. Tosatto, and D. Schomburg, "Qmean: A comprehensive scoring function for model quality assessment," *Proteins*, vol. 71, no. 1, pp. 261–277, 2008. Apr.

[234] Y. Assenov, F. Ramirez, S. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282–284, 2008. Jan.

[235] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *SIGKDD Explor Newsl*, vol. 11, no. 1, pp. 10–18, 2009. Nov.

[236] K. Brinda and S. Vishveshwara, "A network representation of protein structures: Implications for protein stability," *Biophys J*, vol. 89, no. 6, pp. 4159–4170, 2005. Dec.

[237] P. Rice, I. Longden, and A. Bleasby, "Emboss: The european molecular biology open software suite," *Trends Genet*, vol. 16, no. 6, pp. 276–277, 2000. Jun.

[238] D. Gilis and M. Rooman, "Popmusic, an algorithm for predicting protein mutant stability changes: application to prion proteins," *Protein Eng*, vol. 13, no. 12, pp. 849–856, 2000. Dec.

[239] S. Yin, F. Ding, and N. Dokholyan, "Eris: an automated estimator of protein stability," *Nat Methods*, vol. 4, no. 6, pp. 466–467, 2007. Jun.

[240] V. Parthiban, M. Gromiha, and D. Schomburg, "Cupsat: prediction of protein stability upon point mutations," *Nucleic Acids Res*, vol. 34, no. Web Server, pp. W239–242, 2006. Jul.

[241] H. Zhou and Y. Zhou, "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction," *Protein Sci Publ Protein Soc*, vol. 11, no. 11, pp. 2714–2726, 2002. Nov.

[242] A. Olatubosun, J. Valiaho, J. Harkonen, J. Thusberg, and M. Vihinen, "Pon-p: integrated predictor for pathogenicity of missense variants," *Hum Mutat*, vol. 33, no. 8, pp. 1166–1174, 2012. Aug.

[243] K. Wang, M. Li, and H. Hakonarson, "Annovar: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic acids research*, vol. 38, no. 16, pp. e164–e164, 2010.

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

[245] X. Liu, X. Yu, D. J. Zack, H. Zhu, and J. Qian, "TiGER: a database for tissue-specific gene expression and regulation," *BMC Bioinformatics*, vol. 9, p. 271, 2008.

[246] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, *et al.*, "Gene prioritization through genomic data fusion," *Nature biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[247] D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler, R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, L. Peltonen, E. Dermitzakis, P. E. Bonnen, D. M. Altshuler, R. A. Gibbs, P. I.

de Bakker, P. Deloukas, S. B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, and et al, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, pp. 52–58, Sep 2010.

[248] A. von Bubnoff, "Next-generation sequencing: the race is on," *Cell*, vol. 132, pp. 721–723, Mar 2008.

[249] M. P. Ball, J. R. Bobe, M. F. Chou, T. Clegg, P. W. Estep, J. E. Lunshof, W. Vandewege, A. Zaranek, and G. M. Church, "Harvard Personal Genome Project: lessons from participatory public research," *Genome Med*, vol. 6, no. 2, p. 10, 2014.

[250] G. M. Church, "The personal genome project," *Mol. Syst. Biol.*, vol. 1, p. 2005.0030, 2005.

[251] A. N. Gubin, J. M. Njoroge, U. Wojda, S. D. Pack, M. Rios, M. E. Reid, and J. L. Miller, "Identification of the dombrock blood group glycoprotein as a polymorphic member of the ADP-ribosyltransferase gene family," *Blood*, vol. 96, pp. 2621–2627, Oct 2000.

[252] V. Helias, C. Saison, B. A. Ballif, T. Peyrard, J. Takahashi, H. Takahashi, M. Tanaka, J. C. Deybach, H. Puy, M. Le Gall, C. Sureau, B. N. Pham, P. Y. Le Pennec, Y. Tani, J. P. Cartron, and L. Arnaud, "ABCB6 is dispensable for erythropoiesis and specifies the new blood group system Langereis," *Nat. Genet.*, vol. 44, pp. 170–173, Feb 2012.

[253] Y. Colin, C. Rahuel, J. London, P. H. Romeo, L. d'Auriol, F. Galibert, and J. P. Cartron, "Isolation of cDNA clones and complete amino acid sequence of human erythrocyte glycophorin C," *J. Biol. Chem.*, vol. 261, pp. 229–233, Jan 1986.

[254] C. Saison, V. Helias, B. A. Ballif, T. Peyrard, H. Puy, T. Miyazaki, S. Perrot, M. Vayssier-Taussat, M. Waldner, P. Y. Le Pennec, J. P. Cartron, and L. Arnaud, "Null alleles of ABCG2 encoding the breast cancer resistance protein define the new blood group system Junior," *Nat. Genet.*, vol. 44, pp. 174–177, Feb 2012.

[255] G. A. Denomme, "Molecular basis of blood group expression," *Transfus. Apher. Sci.*, vol. 44, pp. 53–63, Feb 2011.

[256] L. T. Goodnough, “Autologous blood donation,” *Crit Care*, vol. 8 Suppl 2, pp. 49–52, 2004.

[257] L. T. Goodnough, J. H. Levy, and M. F. Murphy, “Concepts of blood transfusion in adults,” *Lancet*, vol. 381, pp. 1845–1854, May 2013.

[258] C. E. van der Schoot, B. Veldhuisen, and M. de Haas, “Will Genotyping Replace Serology in Future Routine Blood Grouping? - Opinion 5,” *Transfus Med Hemother*, vol. 36, no. 3, pp. 234–235, 2009.

[259] A. Doscher, C. Vogt, R. Bittner, I. Gerdes, E. K. Petershofen, and F. F. Wagner, “RHCE alleles detected after weak and/or discrepant results in automated Rh blood grouping of blood donors in Northern Germany,” *Transfusion*, vol. 49, pp. 1803–1811, Sep 2009.

[260] D. M. Lublin, S. Kompelli, J. R. Storry, and M. E. Reid, “Molecular basis of Cromer blood group antigens,” *Transfusion*, vol. 40, pp. 208–213, Feb 2000.

[261] S. P. Yip, “Sequence variation at the human ABO locus,” *Ann. Hum. Genet.*, vol. 66, pp. 1–27, Jan 2002.

[262] D. J. Anstee, “Red cell genotyping and the future of pretransfusion testing,” *Blood*, vol. 114, pp. 248–256, Jul 2009.

[263] J. R. Storry and M. L. Olsson, “Will Genotyping Replace Serology in Future Routine Blood Grouping? - Opinion 4: Personalized versus Universal Blood Transfusions - Combining the Efforts,” *Transfus Med Hemother*, vol. 36, no. 3, pp. 232–233, 2009.

[264] E. R. Mardis, “The impact of next-generation sequencing technology on genetics,” *Trends Genet.*, vol. 24, pp. 133–141, Mar 2008.

[265] W. Burke and B. M. Psaty, “Personalized medicine in the era of genomics,” *JAMA*, vol. 298, pp. 1682–1684, Oct 2007.

[266] N. D. Avent, A. Martinez, W. A. Flegel, M. L. Olsson, M. L. Scott, N. Nogues, M. Pis?cka, G. L. Daniels, E. Muniz-Diaz, T. E. Madgett, J. R. Storry, S. Beiboer, P. M. Maaskant-van Wijk, I. von Zabern, E. Jimenez, D. Tejedor, M. Lopez, E. Camacho, G. Cheroutre, A. Hacker, P. Jinoch, I. Svobodova, E. van der Schoot, and M. de Haas,

"The Bloodgen Project of the European Union, 2003-2009," *Transfus Med Hemother*, vol. 36, no. 3, pp. 162–167, 2009.

[267] G. M. Cooper and J. Shendure, "Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data," *Nat. Rev. Genet.*, vol. 12, pp. 628–640, Sep 2011.

[268] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nat. Genet.*, vol. 36, pp. 949–951, Sep 2004.

[269] O. Horaitis, C. C. Talbot, M. Phommarinh, K. M. Phillips, and R. G. Cotton, "A database of locus-specific databases," *Nat. Genet.*, vol. 39, p. 425, Apr 2007.

[270] K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, and et al, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, pp. 851–861, Oct 2007.

[271] W. A. Flegel, "Molecular genetics and clinical applications for RH," *Transfus. Apher. Sci.*, vol. 44, pp. 81–91, Feb 2011.

[272] P. Bugert, E. A. Scharberg, C. Geisen, I. von Zabern, and W. A. Flegel, "RhCE protein variants in Southwestern Germany detected by serologic routine testing," *Transfusion*, vol. 49, pp. 1793–1802, Sep 2009.

[273] T. J. Legler, S. W. Eber, M. Lakomek, R. Lynen, J. H. Maas, A. Pekrun, M. Repas-Humpe, W. Schroter, and M. Kohler, "Application of RHD and RHCE genotyping for correct blood group determination in chronically transfused patients," Aug 1999.

[274] M. Amado, E. P. Bennett, F. Carneiro, and H. Clausen, "Characterization of the histo-blood group O(2) gene and its protein product," *Vox Sang.*, vol. 79, no. 4, pp. 219–226, 2000.

[275] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand,

N. J. Samani, J. A. Todd, P. Donnelly, J. C. Barrett, P. R. Burton, D. Davison, P. Donnelly, D. Easton, D. Evans, H. T. Leung, J. L. Marchini, A. P. Morris, C. C. Spencer, M. D. Tobin, and et al, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–678, Jun 2007.

[276] C. H. Huang, Y. Chen, M. E. Reid, and Y. Okubo, "Evidence for a separate genetic origin of the partial D phenotype DBT in a Japanese family," *Transfusion*, vol. 39, no. 11-12, pp. 1259–1265, 1999.

[277] L. Kauppi, A. J. Jeffreys, and S. Keeney, "Where the crossovers are: recombination distributions in mammals," Jun 2004.

[278] F. J. Rentas and P. A. Clark, "Blood type discrepancies on military identification cards and tags: a readiness concern in the U.S. Army," *Mil Med*, vol. 164, pp. 785–787, Nov 1999.

[279] N. C. van de Pas, R. A. Woutersen, B. van Ommen, I. M. Rietjens, and A. A. de Graaf, "A physiologically based in silico kinetic model predicting plasma cholesterol concentrations in humans," *J. Lipid Res.*, vol. 53, pp. 2734–2746, Dec 2012.

[280] M. J. Telen, M. Udani, M. K. Washington, M. C. Levesque, E. Lloyd, and N. Rao, "A blood group-related polymorphism of CD44 abolishes a hyaluronan-binding consensus sequence without preventing hyaluronan binding," *J. Biol. Chem.*, vol. 271, pp. 7147–7153, Mar 1996.

[281] A. Kasprzyk, "BioMart: driving a paradigm change in biological data management," *Database (Oxford)*, vol. 2011, p. bar049, 2011.

[282] L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent, "The UCSC Genome Browser database: extensions and updates 2013," *Nucleic Acids Res.*, vol. 41, pp. D64–69, Jan 2013.

[283] B. Giardine, C. Riemer, T. Hefferon, D. Thomas, F. Hsu, J. Zielenski, Y. Sang, L. Elnitski, G. Cutting, H. Trumbower, A. Kern, R. Kuhn,

G. P. Patrinos, J. Hughes, D. Higgs, D. Chui, C. Scriver, M. Phommarinh, S. K. Patnaik, O. Blumenfeld, B. Gottlieb, M. Vihinen, J. Valiaho, J. Kent, W. Miller, and R. C. Hardison, "PhenCode: connecting ENCODE data with mutations and phenotype," *Hum. Mutat.*, vol. 28, pp. 554–562, Jun 2007.

[284] G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein, "The International HapMap Project Web site," *Genome Res.*, vol. 15, pp. 1592–1593, Nov 2005.

[285] S. R. Browning and B. L. Browning, "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering," *Am. J. Hum. Genet.*, vol. 81, pp. 1084–1097, Nov 2007.

[286] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Mol. Biol. Evol.*, vol. 28, pp. 2731–2739, Oct 2011.

[287] T. Soga, "Cancer metabolism: Key players in metabolic reprogramming," *Cancer Science*, vol. 104, no. 3, pp. 275–281, 2013.

[288] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[289] J. Thomas, J. Olson, S. Tapscott, and P. Zhao, "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles," *Genome Research*, vol. 11, no. 7, pp. 1227–1236, 2001.

[290] E. Leonardi, A. Murgia, and S. Tosatto, "Adding structural information to the von hippel-lindau (vhl) tumor suppressor interaction network," *FEBS Letters*, vol. 583, no. 22, pp. 3704–3710, 2009.

[291] E. Leonardi, M. Martella, S. Tosatto, and A. Murgia, "Identification and in silico analysis of novel von hippel-lindau (vhl) gene variants

from a large population," *Annals of Human Genetics*, vol. 75, no. 4, pp. 483–496, 2011.

[292] G. Minervini, A. Masiero, S. Moro, and S. Tosatto, "In silico investigation of phd-3 specific hif1- proline 567 hydroxylation: A new player in the vhl/hif-1 interaction pathway?," *FEBS Letters*, vol. 587, no. 18, pp. 2996–3001, 2013.

[293] M. Wu, L. Liu, H. Hijazi, and C. Chan, "A multi-layer inference approach to reconstruct condition-specific genes and their regulation," *Bioinformatics*, vol. 29, no. 12, pp. 1541–1552, 2013.

[294] E. Almaas, "Biological impacts and context of network theory," *Journal of Experimental Biology*, vol. 210, no. 9, pp. 1548–1558, 2007.

[295] J. Weitz, P. Benfey, and N. Wingreen, "Evolution, interactions, and biological networks.," *PLoS biology*, vol. 5, no. 1, p. e11, 2007.

[296] S. Proulx, D. Promislow, and P. Phillips, "Network thinking in ecology and evolution," *Trends in Ecology and Evolution*, vol. 20, no. 6 SPEC. ISS., pp. 345–353, 2005.

[297] P. Mahon, K. Hirota, and G. Semenza, "Fih-1: A novel protein that interacts with hif-1 and vhl to mediate repression of hif-1 transcriptional activity," *Genes and Development*, vol. 15, no. 20, pp. 2675–2686, 2001.

[298] M. Heiner and K. Sriram, "Structural analysis to determine the core of hypoxia response network," *PLoS ONE*, vol. 5, no. 1, 2010.

[299] K. Kohn, J. Riss, O. Aprelikova, J. Weinstein, Y. Pommier, and J. Barrett, "Properties of switch-like bioregulatory networks studied by simulation of the hypoxia response control system," *Molecular Biology of the Cell*, vol. 15, no. 7, pp. 3042–3052, 2004.

[300] Y. Yu, G. Wang, R. Simha, W. Peng, F. Turano, and C. Zeng, "Pathway switching explains the sharp response characteristic of hypoxia response network," *PLoS Comput Biol*, vol. 3, no. 8, p. e171, 2007.

[301] C. Stolle, G. Glenn, B. Zbar, J. Humphrey, P. Choyke, M. Walther, S. Pack, K. Hurley, C. Andrey, R. Klausner, and W. Marston Linehan,

"Improved detection of germline mutations in the von hippel-lindau disease tumor suppressor gene," *Human Mutation*, vol. 12, no. 6, pp. 417–423, 1998.

[302] J. Gnarra, K. Tory, Y. Weng, L. Schmidt, M. Wei, H. Li, F. Latif, S. Liu, F. Chen, F.-M. Duh, I. Lubensky, D. Duan, C. Florence, R. Pozzatti, M. Walther, N. Bander, H. Grossman, H. Brauch, S. Pomer, J. Brooks, W. Isaacs, M. Lerman, B. Zbar, and W. Linehan, "Mutations of the vhl tumour suppressor gene in renal carcinoma," *Nature Genetics*, vol. 7, no. 1, pp. 85–90, 1994.

[303] F. Latif, K. Tory, J. Gnarra, M. Yao, F.-M. Duh, M. Orcutt, T. Stackhouse, I. Kuzmin, W. Modi, L. Geil, L. Schmidt, F. Zhou, H. Li, M. H. Wei, F. Chen, G. Glenn, P. Choyke, M. Walther, and Y. Weng, "Identification of the von hippel-lindau disease tumor suppressor gene," *Science*, vol. 260, no. 5112, pp. 1317–1320, 1993.

[304] W. Kim and W. Kaelin, "Role of vhl gene mutation in human cancer," *Journal of Clinical Oncology*, vol. 22, no. 24, pp. 4991–5004, 2004.

[305] A. Vortmeyer, S. Huang, S. Pack, C. Koch, I. Lubensky, E. Oldfield, and Z. Zhuang, "Somatic point mutation of the wild-type allele detected in tumors of patients with vhl germline deletion," *Oncogene*, vol. 21, no. 8, pp. 1167–1170, 2002.

[306] A. Knudson Jr., "Mutation and cancer: statistical study of retinoblastoma.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 68, no. 4, pp. 820–823, 1971.

[307] J. Gnarra, J. Wards, F. Porter, J. Wagner, D. Devor, A. Grinberg, M. Emmert-Bucki, H. Westphal, R. Klausner, and W. Linehan, "Defective placental vasculogenesis causes embryonic lethality in vhl- deficient mice," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 17, pp. 9102–9107, 1997.

[308] I. Frew and W. Krek, "Pvhl: A multipurpose adaptor protein," *Science Signaling*, vol. 1, no. 24, p. pe30, 2008.

[309] A. Schoenfeld, E. Davidowitz, and R. Burk, "Elongin bc complex prevents degradation of von hippel-lindau tumor suppressor gene prod-

ucts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 15, pp. 8507–8512, 2000.

[310] G. Semenza, "Regulation of mammalian o2 homeostasis by hypoxia-inducible factor 1," *Annual Review of Cell and Developmental Biology*, vol. 15, pp. 551–578, 1999.

[311] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob, "The intact molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, no. D1, pp. D841–D846, 2012.

[312] I. Frew and W. Krek, "Multitasking by pvhl in tumour suppression," *Current Opinion in Cell Biology*, vol. 19, no. 6, pp. 685–690, 2007.

[313] C. Thoma, A. Toso, K. Gutbrodt, S. Reggi, I. Frew, P. Schraml, A. Hergovich, H. Moch, P. Meraldi, and W. Krek, "Vhl loss causes spindle misorientation and chromosome instability," *Nature Cell Biology*, vol. 11, no. 8, pp. 994–1001, 2009.

[314] N. Tang, F. Mack, V. Haase, M. Simon, and R. Johnson, "pvhl function is essential for endothelial extracellular matrix deposition," *Molecular and Cellular Biology*, vol. 26, no. 7, pp. 2519–2530, 2006.

[315] A. Sackmann, M. Heiner, and I. Koch, "Application of petri net based analysis techniques to signal transduction pathways," *BMC Bioinformatics*, vol. 7, 2006.

[316] S. Grunwald, A. Speer, J. Ackermann, and I. Koch, "Petri net modelling of gene regulation of the duchenne muscular dystrophy," *BioSystems*, vol. 92, no. 2, pp. 189–205, 2008.

[317] C. Rohr, W. Marwan, and M. Heiner, "Snoopy-a unifying petri net framework to investigate biomolecular networks," *Bioinformatics*, vol. 26, no. 7, pp. 974–975, 2010.

[318] M. Maynard, A. Evans, T. Hosomi, S. Hara, M. Jewett, and M. Ohh, “Human hif-34 is a dominant-negative regulator of hif-1 and is down-regulated in renal cell carcinoma,” *FASEB Journal*, vol. 19, no. 11, pp. 1396–1406, 2005.

[319] Q. Li, X. Wang, Y. Yang, and H. Lin, “Hypoxia upregulates hypoxia inducible factor (hif)-3 expression in lung epithelial cells: Characterization and comparison with hif-1,” *Cell Research*, vol. 16, no. 6, pp. 548–558, 2006.

[320] T. Smith, P. Robbins, and P. Ratcliffe, “The human side of hypoxia-inducible factor,” *British Journal of Haematology*, vol. 141, no. 3, pp. 325–334, 2008.

[321] S. Richardson, R. Knowles, J. Tyler, A. Mobasheri, and J. Hoyland, “Expression of glucose transporters glut-1, glut-3, glut-9 and hif-1 in normal and degenerate human intervertebral disc,” *Histochemistry and Cell Biology*, vol. 129, no. 4, pp. 503–511, 2008.

[322] S. Vannucci, F. Maher, and I. Simpson, “Glucose transporter proteins in brain: Delivery of glucose to neurons and glia,” *GLIA*, vol. 21, no. 1, pp. 2–21, 1997.

[323] J. Takeda, T. Kayano, H. Fukomoto, and G. Bell, “Organization of the human glut2 (pancreatic -cell and hepatocyte) glucose transporter gene,” *Diabetes*, vol. 42, no. 5, pp. 773–777, 1993.

[324] L. Heather, K. Pates, H. Atherton, M. Cole, and D. Ball, “Differential translocation of fat/cd36 and glut4 coordinates changes in cardiac substrate metabolism during ischemia and reperfusion,” *Circ Heart Fail*, 2013.

[325] J.-W. Kim, I. Tchernyshyov, G. Semenza, and C. Dang, “Hif-1-mediated expression of pyruvate dehydrogenase kinase: A metabolic switch required for cellular adaptation to hypoxia,” *Cell Metabolism*, vol. 3, no. 3, pp. 177–185, 2006.

[326] B.-Y. Kim, H. Kim, E.-J. Cho, and H.-D. Youn, “Nur77 upregulates hif-by inhibiting pvhl-mediated degradation,” *Experimental and Molecular Medicine*, vol. 40, no. 1, pp. 71–83, 2008.

[327] J.-W. Choi, S. Park, G. Kang, J. Liu, and H.-D. Youn, "Nur77 activated by hypoxia-inducible factor-1 overproduces proopiomelanocortin in von hippel-lindau-mutated renal cell carcinoma," *Cancer Research*, vol. 64, no. 1, pp. 35–39, 2004.

[328] J.-S. Roe and H.-D. Youn, "The positive regulation of p53 by the tumor suppressor vhl," *Cell Cycle*, vol. 5, no. 18, pp. 2054–2056, 2006.

[329] D. Fels and C. Koumenis, "Hif-1 and p53: The odd couple?," *Trends in Biochemical Sciences*, vol. 30, no. 8, pp. 426–429, 2005.

[330] J. Berndt, R. Moon, and M. Major, "-catenin gets jaded and von hippel-lindau is to blame," *Trends in Biochemical Sciences*, vol. 34, no. 3, pp. 101–104, 2009.

[331] A. Hergovich, J. Lisztwan, C. Thoma, C. Wirbelauer, R. Barry, and W. Krek, "Priming-dependent phosphorylation and regulation of the tumor suppressor pvhl by glycogen synthase kinase 3," *Molecular and Cellular Biology*, vol. 26, no. 15, pp. 5784–5796, 2006.

[332] M. Heiner, I. Koch, and J. Will, "Model validation of biological pathways using petri nets - demonstrated for apoptosis," *BioSystems*, vol. 75, no. 1-3, pp. 15–28, 2004.

[333] R. Bortfeldt, S. Schuster, and I. Koch, "Exhaustive analysis of the modular structure of the spliceosomal assembly network: a petri net approach," *Stud Health Technol Inform*, vol. 162, pp. 244–247, 2011.

[334] M. Heiner, "Understanding network behavior by structured representations of transition invariants," *Algorithmic Bioprocesses*, pp. 367–389, 2009.

[335] J. Zhu, E. Kaytor, C.-I. Pao, X. Meng, and L. Phillips, "Involvement of sp1 in the transcriptional regulation of the rat insulin- like growth factor-1 gene," *Molecular and Cellular Endocrinology*, vol. 164, no. 1-2, pp. 205–218, 2000.

[336] T. Okabe and H. Nawata, "Functional role of nur77 family in t-cell apoptosis and stress response," *Nippon rinsho. Japanese journal of clinical medicine*, vol. 56, no. 7, pp. 1734–1738, 1998.

[337] E. Murphy and O. Conneely, “Neuroendocrine regulation of the hypothalamic pituitary adrenal axis by the nurr1/nur77 subfamily of nuclear receptors,” *Molecular Endocrinology*, vol. 11, no. 1, pp. 39–47, 1997.

[338] J. Tabernero, “The role of vegf and egfr inhibition: Implications for combining anti-vegf and anti-egfr agents,” *Molecular Cancer Research*, vol. 5, no. 3, pp. 203–220, 2007.