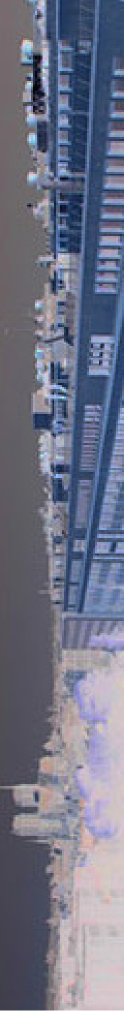
FUNCTIONAL PROTEIN SITES DECIPHERED BY EVOLUTION AND STRUCTURAL DYNAMICS

Elodie Laine



Habilitation à Diriger des Recherches

Sorbonne Université

présentée et soutenue publiquement par

Elodie Laine

le 02 Octobre 2020

FUNCTIONAL PROTEIN SITES DECIPHERED BY EVOLUTION AND STRUCTURAL DYNAMICS

Jury

Dr. Marc Baaden,	Examinateur
Dr. Alexandre de Brevern,	Rapporteur
Dr. Jacques Chomilier,	Examinateur
Prof. Franca Fraternali,	Rapporteur
Dr. Julie Thompson,	Rapporteur

Sorbonne Université
Laboratoire de Biologie Computationnelle et Quantitative
UMR 7238, CNRS-SU, 75005 Paris, France

H
D
R

Contents

Acknowledgements

First, I want to thank Marc Baaden, Alexandre de Brevern, Jacques Chomilier, Franca Fraternali and Julie Thompson for kindly accepting to evaluate my work. It is an honour to have you as jury members.

The works presented in this memoire are the results of collective adventures, and I want to thank all the people who have participated in it over the years, though collaborations, exchanges, and support. Although I do not present any of my PhD work, I am of course thinking about my advisors Arnaud Blondel and Thérèse Malliavin, with whom everything started, and in a very nice way. I want to thank Luba Tchertanov, who guided and supported me during my post-doctorate. I think this transition period is not the easiest one in a scientific carrier, and you helped me go through it. I am very grateful to Alessandra Carbone, who gave me many opportunities to work on interesting questions and participate in exciting projects. Our collaboration has been very fruitful and stimulating, and I have learnt a lot with you. I want to thank Hugues Richard, with whom I shared my office for some time. I am very happy that our endless conversations took the form of a concrete project. I hope we will be able to continue to work together. I want to thank my other collaborators, for making the effort to get involved into new subjects and dedicate time and ressources to make things advance. Although I do not talk about teaching in this memoire, it represents a big part of my activity. I shall thank the people who allow me to do this part of my job in pleasant conditions. In particular, I want to thank Anne Lopes who proposed me to embark on the Meet-U adventure. It has been one of the most exciting experience I have had in the past years. I also thank the students I have had the chance to meet and interact with. Your energy and ideas have fuelled my motivation to keep trying and explore new paths.

Beyond what is visible, there are all the moments, be they contemplative, ordinary or adventurous, spent with my friends and loved ones. I dedicate them this memoire and thank them for being here, accompanying me, sharing thoughts, doubts and experiences, making me discover new worlds and realise we can do things I thought were impossible.

Avant-Propos

Такие дела

Over the past thirteen years, I have been dedicating my research activities to characterise protein sequences, structures, dynamics and interactions, and relate them to protein (mal-)functioning. During my PhD and post-doctoral experiences, I extensively studied two particular systems, namely the edema factor (EF) of anthrax and the receptor tyrosine kinase KIT. I developed and applied computational methods to simulate their behaviour in solution and predict the impact of chemical modifications (ion binding, oncogenic mutations) on their activation. I discovered a new family of inhibitors preventing the interaction between EF and calmodulin, and I designed a rescuing mutation in KIT. I mostly used structural information, generating and exploiting conformational ensembles. Since my recruitment at LCQB in 2012, I have started to look at protein sequences and to include evolutionary aspects in my work. I have also moved to the development of methods that can be applied on a large scale, without requiring expert knowledge on the studied systems.

During the last two decades, biology has been revolutionised by the advent of high-throughput technologies and the rapid increase of computing resources. Our understanding of biological objects has been improved, but also challenged by the massive amounts of data produced and accessible. It would be no exaggeration to say that most popular and basic concepts in Biology lack a clear and unambiguous definition. What is exactly a gene? Shall a protein domain be seen as an autonomous folding unit or as an evolutionary persistent building block? Answering these questions is not a trivial task. As data accumulate, we are becoming more and more aware of the complexity of living systems and of the need to revisit long standing over-simplifying paradigms. Experimental evidence of pervasive translation, tissue-specific transcriptomes and cellular heterogeneity give us a glimpse of the huge function diversification potential that living organisms have evolved. For instance, a given protein may interact with hundreds of partners through distinct or overlapping regions, interconvert between a wide range of conformations, and perform completely different functions in the cell (*e.g.* RNA binding and chemical reaction catalysis). Ideally, one would like to be able to decipher such complexity, toward controlling existing proteins and designing new ones with desired qualities. Unfortunately, directly observing proteins in action remains very challenging and accurately simulating their behaviour and interactions with their environment is computationally prohibitive. Alternatively, the information encoded in protein sequences provides a valuable indirect mean for probing protein functioning. It has the advantages of being accessible in large amount at low cost and of resulting from millions of years of evolution, thus complying with physical and environmental constraints. In this dissertation, I will discuss the relative contributions of sequence- and structure-based information in identifying protein functional sites and in predicting the impact of variations at these sites.

Experimental measurements are generally regarded as the ground truth. Predictions produced by a computational method shall be ultimately validated by wet-lab experiments. However, the later are not exempted from biases coming from the chosen conditions and the instruments of measure. Moreover, there is a theoretical model, and some underlying assumptions, behind any experimental data published in the literature. For instance, solving a protein (or protein complex) structure with X-ray crystallography often implies expressing it in another organism, modifying it, placing it in a physico-chemical environment very different from the physiological one, biasing the final output toward known structures of similar proteins and applying crystal packing constraints. One may argue that computational methods generating *in silico* data are a mean of probing a system, just like an experimental device. Here, I will present molecular simulations at time and length scales that cannot be probed by wet-lab experiments. I will show how medium-scale computational experiments can inform us about the social life of proteins, and how apparently false positive predictions can make us aware of the multiplicity of protein surface usage.

Predictive models have adapted to the increasing body of accessible empirical knowledge. For instance, the description of protein structures has evolved from models assembled from basic pieces according to physical elementary laws, through task-driven statistics to complex artificial intelligent systems that *learn* directly from (possibly) heterogeneous and high-dimensional data. In very recent years, deep neural networks have proven successful in predicting protein structures, assessing mutational outcomes and guiding protein design. They are being applied to an ever higher number of problems in biology. When dealing with them, the human task seems shifted from understanding complex biological objects to understanding complex network architectures, their hyperparameters and their filters. Still, biological and physical priors play an essential role in achieving good performance, by guiding the selection and annotation of the training and testing data, and by imposing some constraints on the architecture. One drawback of these complex systems is that they lack interpretability. In the following, I will report on methods based on human learning and intuition, rather than machine learning. They typically rely on a few biologically and physically meaningful parameters. They are designed to provide a clear readout of the input data and to transform these data into comprehensible objects useful for making predictions and interpreting them. I will show how these methods can help to reason about the origins and functions of protein interactions and can provide mechanistic explanations for observed phenotypes.

Finally, I would like to emphasise that the work presented herein is the result of a collaborative effort. Several colleagues, collaborators and students, have been involved at every step of it, from the formulation of an idea to its validation and exploitation to generate new knowledge. Since my recruitment at LCQB, I have had the opportunity to co-supervise several Master students, PhD students and post-doctoral fellows, mostly in collaboration with A. Carbone, head of the Analytical Genomics team. I have been involved in one of the main research themes developed in the team, which concerns protein-protein interactions. We have also collaborated on the prediction of the effect of mutations on protein structural stability and function. In recent years, I have been co-leading a project team with H. Richard (now at RKI, DE) on the topic of alternative splicing. The aim of the project is to systematically assess the structural impact of alternative splicing in evolution. This dissertation is organised in three chapters reporting some outcomes of these projects. These collaborations have been very enriching, especially because we come from different backgrounds and look at the same biological objects from different points of view. I would also like to acknowledge a big part of my activity, which is devoted to teaching. Elaborating the pedagogical material and interacting with the students strongly influence the research I am doing, and in more than one way.

Chapter 1

The fate and effects of protein mutations

1.1 Motivation

Understanding which and how genetic variations affect proteins and their biological functions is a central question for bioengineering, medicine, and fundamental biology. Disease-associated mutations can impair protein function in various ways, by destabilising the protein structure, by shifting the equilibrium of conformation populations, or by modulating the binding affinity of the protein for its cellular partner(s), to name a few. Ideally, one would like to be able to rapidly assess the outcomes of thousands of mutations and to provide mechanistic explanation for these outcomes. This would allow to reach some level of control over proteins, needed to improve the treatment of diseases, the design of new proteins and the synthesis of molecular libraries.

Deep mutational scans[1] or multiplexed assays for variant effects[2] have enabled the full description of the mutational landscapes of a few tens of proteins (see[3] for a list of proteins and associated experiments). They have revealed that a protein contains a relatively small number of positions highly sensitive to mutations, where almost any substitution induces highly deleterious effects[4,5]. Although these methods represent major biotechnological advances, they remain resource intensive and are limited in their scalability. Moreover, the measured phenotype and the way it is measured vary substantially from one experiment to another, making it difficult to compare different measurements and/or proteins[6]. These limitations call for the development of efficient and accurate computational methods for high-throughput mutational scans.

Many computational methods predicting mutational effects exploit information coming from protein sequences observed in nature[3,7–20]. They look at the amino acid frequencies of occurrence in multiple alignments of related sequences. Multiple sequence alignments (MSAs) provide invaluable information about protein evolutionary history, diversity, conservation and function[21,22]. Typically, one expects that rarely occurring mutations are likely deleterious. A straightforward way to estimate frequencies of occurrence is to treat each position in the alignment independently from the others. However, the amino acid residues comprising a protein are inter-dependent, and the effect of a mutation depends on the amino acids present at other positions, a phenomenon referred to as 'epistasis'[23,24]. By leveraging on the increasing wealth of genomic data, recent developments have enabled modelling inter-dependencies between positions and have significantly improved the accuracy of mutational effects predictions[3,7–10,13,14]. Specifically, some statistical methods estimate couplings between pairs of positions[7–14]. They are very accurate in identifying a few strong direct couplings responsible for the whole co-variability observed in homologous sequences and corresponding to physical contacts in protein structures[7,25,26]. In the context of mutational outcome prediction, the ensemble of all pairwise couplings is used as a proxy to capture the influence of the whole sequence context on a particular position. One of the limitations of these methods is that the explicit calculation of higher-order couplings is computationally intractable. To circumvent this issue, a deep latent-variable model was proposed where the global sequence context is implicitly accounted for by coupling the observed positions to latent ('hidden') variables[3]. The model is fully trained on each studied protein family to generate sequences likely to belong to the family. Deviations between outputs and inputs are then used as estimates of the mutational effects. The mutational landscapes of certain protein families are very well captured by this deep learning approach, but the results strongly depend on the variability of the input data. More generally, the statistical inference of a large body of parameters from a finite, and sometimes very low, sequence sampling is a challenging problem[27,28]. It is particularly relevant in the case of viral proteins whose sequences are often highly conserved. Several technical advances employing regularisation terms have improved the accuracy of inter-residue coupling estimation when dealing with viral proteins[7,9,11–14]. Moreover, the usage of very small position-specific amino acid alphabets has reduced the computational cost of the inference. These efforts have allowed achieving very good agreement between the fitness landscapes inferred from patient sequences and *in vitro* experiments for several proteins from HIV and HCV. Nevertheless, the available methods still remain computationally costly and have only been evaluated against low-throughput experimental data.

While sequence-based methods can yield very accurate predictions of mutational phenotypic outcomes, structural approaches provide a unique way to shed light on the molecular mechanisms underlying them. There are few reported cases where crystallised protein mutants provide clear insights on the effects of the mutations (*e.g.* p53 cancer mutations affecting the arginines in contact with DNA[29,30]). However, in the vast majority of cases, the global shape of the protein remains unchanged upon mutations, even when the latter result in deleterious phenotypes[31]. This is very well exemplified by PSD95^{pdz3}: the crystallographic structures of several deleterious mutants were solved and are very similar to that of the wild type[32]. Moreover, mutation signals can propagate across protein structures and affect very distant protein sites, a phenomenon referred to as "allosteric coupling"[33]. In this context, characterising the dynamical behaviour of the protein may reveal internal dynamics changes associated to the mutations, and help assess and interpret their outcomes. Such characterisation can be realised by all-atom molecular dynamics (MD), and there are several examples in the literature where MD simulations, even of only a few tens of nanoseconds, revealed conformational rearrangements upon mutations and brought valuable insights into the molecular mechanisms underlying mutational outcomes[34–45]. The time scales reachable by MD have largely increased and it is now possible to simulate a mutated system for several microseconds[46]. Nevertheless, simulating tens of mutants on such long time periods remains very costly and the complete description of a protein's conformational landscape is still far beyond reach. Another drawback is that identifying the protein properties (inter-residue distance, inter-domain angle, local unfolding, solvent exposure...) that should be recorded along the simulation to guide an automatic detection of mutational effects, usually demands an expert knowledge of the system under study. Even with such knowledge, it may be difficult to determine what matters or not.

For decades, models and methods have been proposed toward rationalising the propagation of a signal, such as a point mutation, across a protein structure[33,47–58]. In recent years, several methods have been developed to identify "communication routes"[59–69], "dynamic domains"[70–76] and/or critical allosteric residues[70,77] in proteins in an automated way (see also methods reviewed in[78]). Most of them construct a graph representing the protein where the nodes are the residues and the edges are determined based on the strength of non-covalent interactions (hydrogen-bonds, hydrophobic contacts, salt bridges...) and/or on correlations between residues displacements. The latter are inferred either from all-atom MD simulations, or from more coarse-grained and computationally efficient approaches like the Elastic Network Model (ENM), where residues close in 3D space are linked by springs. The constructed graph is then analysed to extract paths and communities of residues. Residues identified in the paths and/or playing particular roles (*e.g.* hubs) in the communities have been shown to be important for the protein structural stability and allosteric regulation. However, the agreement of computationally identified paths/communities with experimental data has been mostly assessed qualitatively, and little agreement has been found between different computational approaches or simulations[78].

In the following, I present contributions toward sequence-based large-scale computational mutational scans of proteins[79], the automated extraction of routes of information transmission across protein structures[42,80–83], the systematic prediction of mutation severity based on structural dynamics[83] and the characterisation of the molecular mechanisms underlying mutational outcomes[80,82]. They were published in six research articles, among which three as first author and three as co-corresponding author.

1.2 Contributions

1.2.1 Evolution-based large scale prediction of mutational outcomes

To assess mutational effects at large scale, we have developed a fast, scalable, and simple method that explicitly models the evolutionary history of natural sequences (**Fig. 1.1**). Our two main hypotheses are that mutations occurring rarely in nature are likely deleterious and that the different positions in a protein sequence influence each other. These hypotheses are not new, but to implement them we adopt an orthogonal approach compared to the state of the art. Instead of treating the input sequence data as a 2D matrix (MSA), we exploit the underlying tree structure of the data coming from their evolutionary history.

A global epistatic model accounting for protein evolutionary history

Specifically, to evaluate the impact of a given mutation at a given position in a query sequence, we look at an ensemble of sequences homologous to that query (**Fig. 1.1a**) and at the topology of the tree reflecting their evolutionary relationships (**Fig. 1.1b**). Our main contribution is to extract conservation patterns in line with the topology of the tree and use them to determine the extent to which a mutation will be deleterious for the function of the query protein. For this, the first step of our approach consists in estimating the biological importance of each residue in the query by computing its evolutionary conservation (**Fig. 1.1c**). Our measure of conservation is computed by the Joint Evolutionary Trees (JET) method[84] and is inspired from the notion of evolutionary trace[85,86]. It is markedly different from measures that quantify frequencies of occurrence at single positions (columns) of an alignment. Indeed, for each position in the query, we look at the level in the tree where the amino acid at that position appeared and remained conserved thereafter (**Fig. 1.1b**, see gray rectangles). Since the tree is inferred from global similarities between entire sequences, the conservation degree of a given position embeds the covariations between this position and all other positions in the sequence. Hence, two positions can have the same distribution of letters but different conservation levels. For example, this will happen if one position displays all occurrences of the most represented letter in a subtree of ancient origin while the other displays them in several subtrees. Here, we deal with a potentially large number of sequences, and the reconstruction of a unique tree relating all of them may lead to an unreliable topology. To cope with this issue, we construct many small trees from subsets of sequences and average conservation levels over all trees.

We use the computed conservation degrees to weight positions. Specifically, to compare different substitutions occurring at a given position, we combine two quantities. The first one is the relative frequency of occurrence of the mutation, relying on physicochemical similarities rather than amino acid identities. It depends only on the position of interest (*independent* contribution). The second one is the minimum evolutionary distance one has to go in the evolutionary tree to observe a natural sequence displaying the mutation (*epistatic* contribution). This evolutionary distance between the query q and some sequence s is expressed as

$$\mathrm{D}_{evol}(q,s) = \sum_{k=1}^{n} \mathrm{T}_{\mathrm{JET}}(k)^2 * \mathbf{1}_{\mathrm{X}_k^q \neq \mathrm{X}_k^s}, \tag{1.1}$$

where n is the length of q, X_k^q is the amino acid of q at position k, and $\mathbf{1}_{\mathrm{X}_k^q \neq \mathrm{X}_k^s}$ is the indicator function. It explicitly accounts for the conservation degrees of all variable positions between the query q and the closest sequence s bearing the mutation. Then, to compare mutations occurring at different positions, we rely on the hypothesis that more conserved positions will be more sensitive to any mutation than less conserved positions. To implement this idea, we re-weight the predicted mutational effects by the evolutionary conservation degrees. The normalised predicted effect (NPE) of a mutation X-to-Y at position i is expressed as

$$\mathrm{NPE}(\mathrm{Y}_i) = \mathrm{T}_{\mathrm{JET}}(i) * \mathrm{PE}(\mathrm{Y}_i). \tag{1.2}$$

where PE(Y_i) is the predicted effect. As a result, highly deleterious mutations will be mainly found at highly conserved positions (**Fig. 1.1c**, second matrix, dark squares are mainly localised at conserved positions, highlighted by arrows, while the other columns have been "whitened"). We have only two hyper-parameters, namely the relative weight of the independent and epistatic contributions and a reduced amino acid alphabet used for computing the frequencies of occurrence. The method is implemented as a fully automated tool, Global Epistatic Model for predicting Mutational Effects (GEMME), available as a downloadable package and as a web server at: www.lcqb.upmc.fr/GEMME/. GEMME is applicable to single site mutations and also to combinations of mutations.

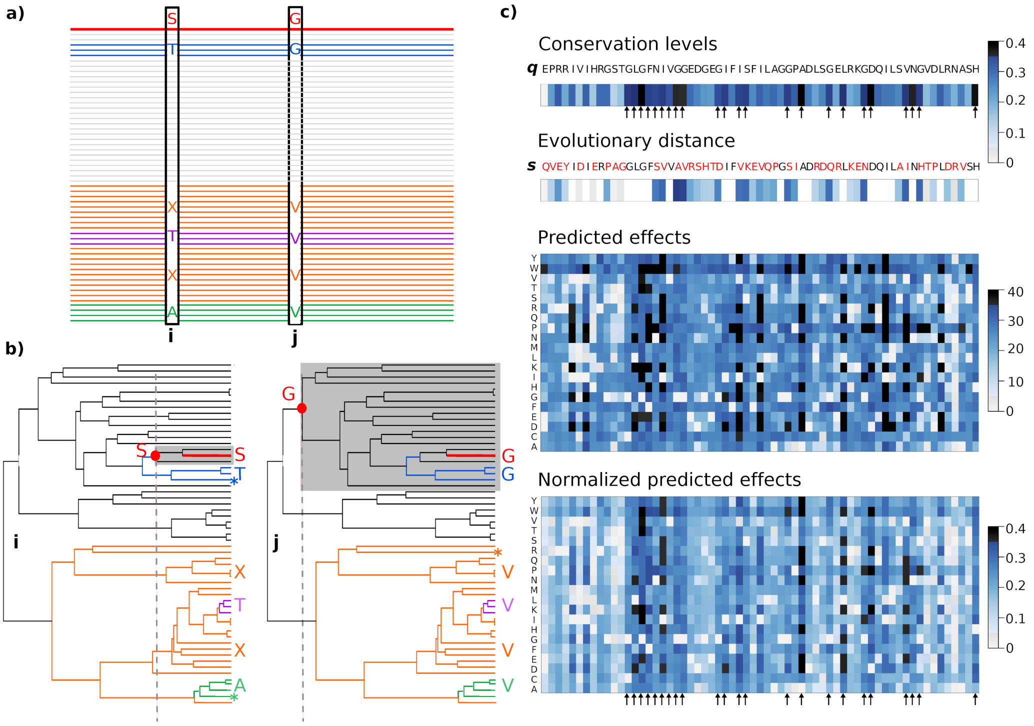

a)
S
G
T
G
X
V
T
V
X
V
A
V
i
j
b)
S
S
T
G
G
G
i
j
X
T
X
A
V
V
V
V
c)
Conservation levels
q EPRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSVNGVDLRNASH
Evolutionary distance
s QVEYIDIERPAGGLGFSVVAVRSHTDIFVKEVQPGSIADRDQRLKENDQILAINHTPLDRVSH
0.4
0.3
0.2
0.1
0
Predicted effects
Y W V T S R Q P N M L K I H G F E D C A
40
30
20
10
0
Normalized predicted effects
Y W V T S R Q P N M L K I H G F E D C A
0.4
0.3
0.2
0.1
0

Figure 1.1: **Principle of GEMME. (a)** Ensemble of sequences related to a query sequence, on top and in red. The query displays a serine (S) at position *i* and a glycine (G) at position *j*. Some sequences are coloured according to the amino acids they display at the two positions: T-G in blue, T-V in purple, X-V in orange (X stands for any amino acid, except for T and A) and A-V in green. **(b)** Tree representing the evolutionary relationships between the related sequences with position-specific information. The colour code is the same as in (a). The red dots and dotted grey lines indicate the levels where S and G appeared at positions *i* and *j* and remained conserved thereafter. The associated subtrees are highlighted by grey rectangles. The stars indicate the closest sequences to the query displaying the S-to-T mutation at *i* (left, in blue), the S-to-A mutation at *i* (left, in green) and the G-to-V mutation at *j* (right, in orange). Mutation S-to-A et *i* will be predicted as more deleterious than S-to-T because it more distant from the query. But it will be predicted as less deleterious than mutation G-to-V at *j* because position *j* is much more conserved than position *i*. **(c)** Workflow of the method applied on the third PDZ domain of PSD95 (DLG4). The colour strip on top gives conservation levels for the query *q*. Positions highlighted by arrows are highly conserved. A homologous sequence *s* is displayed below. The changes are highlighted in red and the colour strip indicates the associated computed values (squared conservation levels). The two matrices give the predicted effects and normalised predicted effects, respectively, for all possible substitutions at all positions in *q*.

Application at large scale and comparison with the state of the art

Assessed against experimental measures collected from 41 high-throughput mutational scans representing 657,840 mutations, GEMME achieved an average Spearman rank correlation $\bar{\rho} = 0.53 \pm 0.13$ (**Fig. 1.2**). Despite its apparent simplicity, it achieves similar or better performance compared to the state of the art methods DeepSequence[3] and EVmutation[8]. Importantly it largely outperforms them when the diversity of the input sequence alignment is low, as is the case for viral proteins. GEMME was also assessed against 128 experimental measures coming from 9 low-throughput mutational studies of two HIV proteins, with sequences coming from patients. With a weighted average correlation of 0.70, the results are similar to those obtained by two coevolution-based computational frameworks[7,9]. Hence, while GEMME was designed to treat any protein family, its performance on viral proteins are similar to recent computational frameworks well suited to treat these proteins. Besides predictive performance, a major advantage of GEMME is its computational efficiency. Our algorithm is faster than state-of-the-art methods by several orders of magnitude **Fig. 1.2b**. It takes less than 10 minutes to treat any protein from the high-throughput dataset. A thorough analysis of revealed that the epistatic contribution is useful to discriminate between equally frequent mutations **Fig. 1.2c** and that the results are robust to parameter changes.

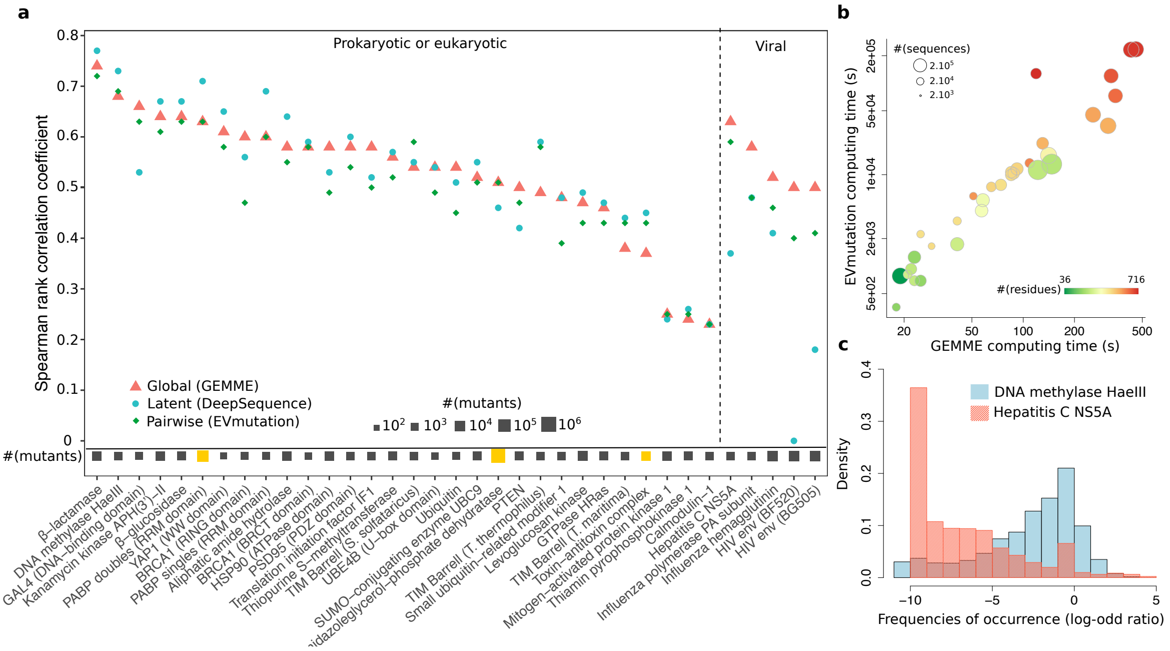


Figure 1.2: **GEMME's predictive and computing performances.** **(a)** Spearman rank correlation coefficients ρ between predicted and experimental measures for 35 high-throughput experiments corresponding to 33 proteins and 1 protein complex. Scans comprising multiple mutations are highlighted by a gold rectangle. **(b)** Computing times of EVmutation and GEMME (in seconds). **(c)** Examples of distributions for the mutations' site-independent relative frequencies of occurrence. For each mutation, the reported value is the log-odd ratio between the number of sequences displaying the mutation over the number of sequences displaying the wild-type amino acid. The epistatic contribution is helpful in the case of DNA methylase HaeIII (in red) but not in the case of Hepatitis C NS5A (in blue).

Contrary to statistical inference-based methods, GEMME does not estimate a joint probability distribution. Instead, it directly exploits the information encoded in natural sequences in a query-centred way. Specifically, each tree constructed to compute conservation levels contains the query, and the evolutionary distances are computed with respect to the query. Hence, our predictions estimate deviations from the query, while predictions from statistical inference-based methods correspond to ratios of probabilities of belonging to the protein "family" represented by the input alignment. On the one hand, this constitutes a limitation of our method. For instance, in case of a mutant with a higher fitness than the wild type and evolutionary far away from it, GEMME will simply predict a strong mutational effect. On the other hand, this can be an advantage in the presence of "subfamilies" performing different functions and displaying different functionally relevant sequence patterns (as is the case of the cryptochrome/photolyase family for instance).

1.2.2 Conformational dynamics-based prediction of mutational effects

We have developed several measures and algorithms to extract information relevant to the propagation of mutational signals across protein structures from conformational ensembles. Their design was inspired by experimental studies suggesting that protein residues "communicate" in different ways. On the one hand, signals may be propagated via stable non-covalent interactions across the protein structure. A classical example is given by hemoglobin, where the binding of oxygen to one subunit induces conformational changes relayed to the other subunits, increasing their binding affinity for oxygen[33]. On the other hand, signals may be transmitted without requiring physical interactions as a support, but simply via local changes in atomic fluctuations[87,88]. In our approach, we aimed at accounting for this dual nature of allosteric coupling by computing different types of correlations and links between residues. We refer to it as "infostery", from “info” – information – and “steric” – arrangement of residues in space.

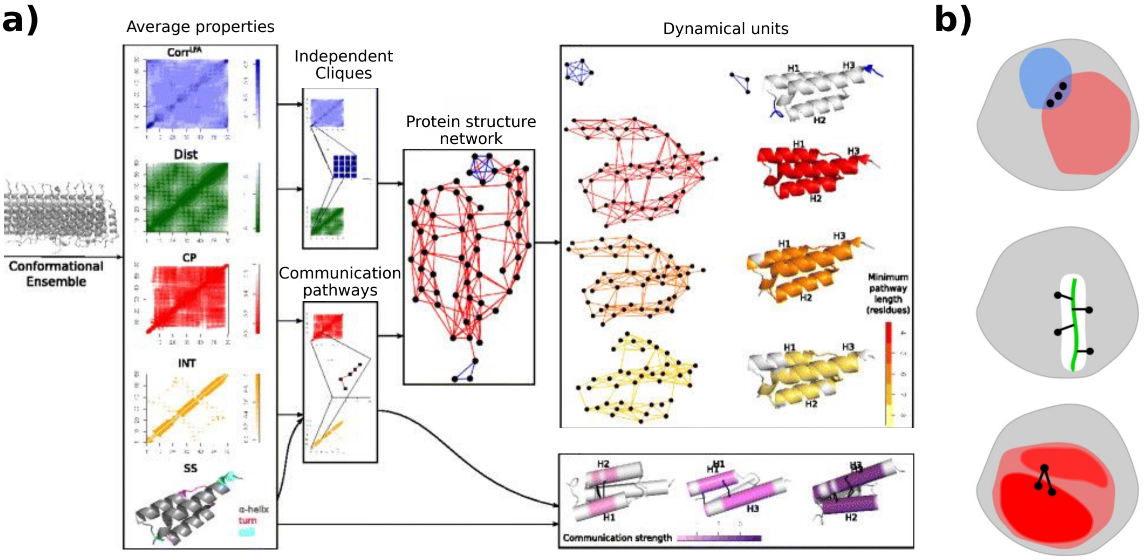


Figure 1.3: **Principle of the infostery analysis.** **(a)** Starting from a set of conformations, we compute residue-based properties: local dynamical correlations (Corr_{LFA}), minimum distances (Dist), communication propensities (CP), non-covalent interaction strengths (INT) and secondary structures (SS). By combining them, we group residues in independent cliques and in communication pathways. A coloured graph is then constructed and connected components, called dynamical units, are extracted. Communication pathways are also used to detect pairs of communicating segments, which are portions of secondary structure elements. **(b)** Top: 3 residues belonging to different types of dynamical units. Middle: 4 protein residues in direct communication with the protein’s ligand (green thick segment). Bottom: 2 pairs of residues bridging two sub-regions of a dynamical unit. The more pronounced colour of the two subregions indicate that they contain many pathways (dense communication).

A method to describe protein dynamical architectures

To describe inter-residue communication, we exploit average quantities computed from conformational ensembles and identify *independent cliques* and *communication pathways* (**Fig. 1.3a**, on the left). We define an independent clique as a cluster of residues close to each other in 3D space displaying high concerted atomic fluctuations. They typically correspond to solvent-exposed loops which are highly flexible and also move independently from the rest of the protein. We define a communication pathway as a chain of residues, where all the residues "communicate" efficiently with each other and any pair of residues adjacent in the pathway are linked by stable non-covalent interactions. *Communication efficiency* is computed from the conformational ensemble as the inter-residue distance variance. Hence, residues that move together (small variance) will be considered to communicate efficiently. Two residues adjacent in a pathway are said to be in *direct* communication, as opposed to *indirect* communication when the residues are in the same pathway but not adjacent in it. The notion of direct communication is more refined than that of physical contact and should not be confounded with it: accounting for inter-residue displacements correlations enables discriminating among physical contacts. By linking residues belonging to the same clique or to the same pathway, one can define a protein structure network (**Fig. 1.3a**, in the middle) and extract connected components from it. These components can be thought of as the *dynamical units* of the protein. Intuitively, residues in a pathway-based unit move together in a rather rigid way, while residues in a clique-based unit are more flexible (**Fig. 1.3a**, on the right).

This description of the protein structure can be used to assess the effects of mutations, by comparing either individual communication pathways or statistics on the pathways between wild-type and mutated proteins. It can also be useful to identify residues playing a key role in the stability and intra-communication of the protein structure and thus likely sensitive to mutations. Namely, we devised three strategies to identify residues that serve as *communication bridges* (**Fig. 1.3b**). The first class is comprised of residues belonging to both a pathway-based unit and a clique-based unit (**Fig. 1.3b**, on top). The second class contains residues establishing direct communication with the ligand, in case of a protein-ligand complex (**Fig. 1.3b**, in the middle). The third class comprises pairs of residues that are in direct communication while their neighboring residues are not (**Fig. 1.3b**, at the bottom). The intuition is that because their communication signal is isolated, disrupting these pairs should have an impact on the overall communication of the unit. The approach is implemented as a fully automated tool, COMmunication MApping v2.0 (COMMA2, http://www.lcqb.upmc.fr/COMMA2/).

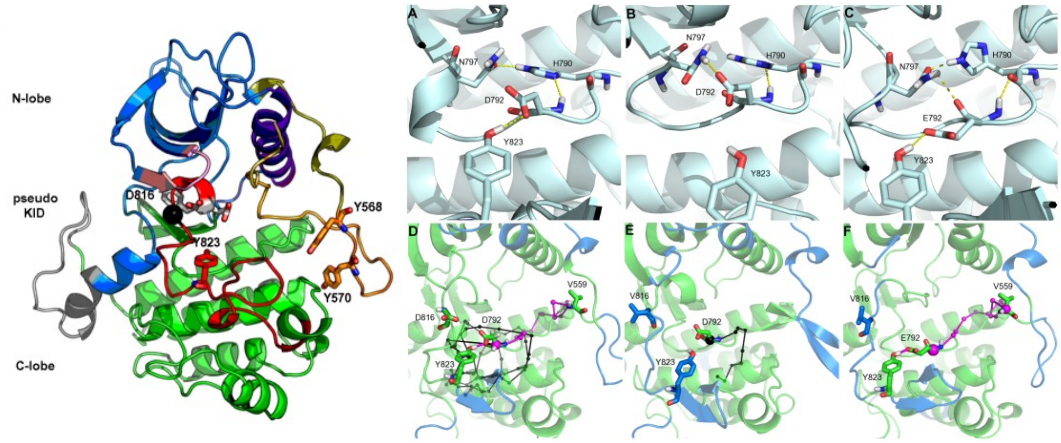


Figure 1.4: **Infostery analysis of KIT cytoplasmic region. Left:** Structure of KIT cytoplasmic region, in the auto-inhibited inactive state (PDB code: 1T45). Some sub-regions documented in the literature are highlighted with colours. The A-loop is in red and the JMR is in orange tones. The mutation site D816 is highlighted by a black sphere. **Right:** Average MD conformations: **(A,D)** wild type KIT, **(B,E)** D816V mutant, **(C,F)** D816V/D792E double mutant. The interaction network between the A-loop residues Y823 and the catalytic loop residues H790, D792 and N797 is depicted on top. The communication pathways starting from the A-loop toward the JMR are displayed at the bottom. Each pathway is displayed as a chain of small black spheres connected together by black lines. The path linking the A-loop and the JMR through the catalytic loop is highlighted in magenta. Residues D/V816 and Y823 in the A-loop, D/E792 in the catalytic loop and V559 in the JMR are highlighted in licorice and labeled.

Long-range mutational effects and rescue mutant design

As a proof of concept, we applied our method to the receptor tyrosine kinase KIT and its oncogenic mutant, $\text{KIT}^{\text{D816V}}$. The D816V mutation takes place in the activation loop (A-loop) of the protein kinase region and makes the protein constitutively active. We performed 50-ns MD simulations of the auto-inhibited inactive state of the protein and observed that the mutation had a short-range destabilising effect on a small helix, and also a long-range stabilising effect on the juxta-membrane region (JMR). The JMR is located 15 Å away from the mutation site (**Fig. 1.4**, on the left) and plays a crucial role in the auto-inhibition of the protein. By using our formalism, we could rationalise this long-range effect in terms of communication across the protein structure. In the wild type, we identified a communication pathway linking the A-loop and the JMR through the catalytic loop (**Fig. 1.4D**). This pathway likely serves as a physical support for information transmission between the two sites. In the mutant, the local hydrogen-bond (H-bond) network around the mutation is disturbed and the communication pathway is disrupted (**Fig. 1.4E**). This observation provides a mechanistic explanation for the structural reorganisation of the JMR in the mutant. This interpretation is consistent with the results obtained by principal component analysis revealing a decoupling of A-loop and JMR motions in the mutant. By comparing the H-bond networks in the wild type and the mutant, we spotted a key H-bond in the communication pathway of the wild type, namely $823\cdots792$ that was lost in the mutant (**Fig. 1.4A-B**). We hypothesised that a way to re-establish the communication between the A-loop, the catalytic loop and the JMR in the D816V mutant would be to restore this H-bond. To this aim, we substituted the aspartate (D) in position 792 by a glutamate (E), bearing the same charge and displaying a longer side chain. We anticipated that such replacement would facilitate the formation and stabilisation of an H-bond with Y823 that adopted in the mutant an orientation unfavourable to such interaction. 50-ns MD simulations of this double mutant confirmed our hypothesis. The H-bond was indeed restored (**Fig. 1.4**C), and so was the communication pathway (**Fig. 1.4**F). Moreover, the position and the conformation of the JMR were similar to those observed in the simulation of the wild type. Binding free energy calculations further showed that the similar structural properties and dynamical behaviours displayed by the JMR in the wild type and D816V/D792E-mutated KIT forms correspond to nearly equivalent thermodynamic landscapes.

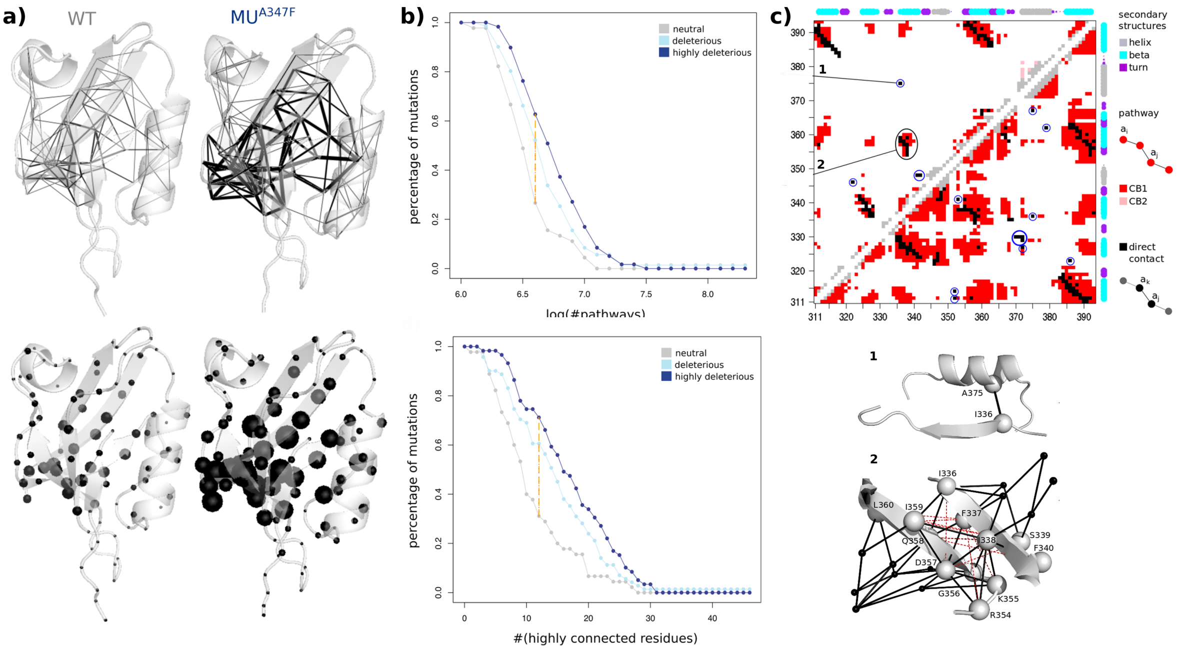


Figure 1.5: **Infostery analysis of the PSD95^{pdz3}-CRIPT complex.** **(a)** Pathway properties mapped onto averaged MD conformations. Top: Communication pathways are displayed as segments linking residues' C-α atoms. The thickness is proportional to the number of pathways. Bottom: Pathway concentration is displayed as spheres whose sizes are proportional to the number of pathways crossing the residue. **(b)** Inverse cumulative distributions of the number of pathways (on top) and highly connected residues (at the bottom) for 45 neutral (in grey), 71 deleterious (in light blue) and 59 highly deleterious (in dark blue) mutations. Each y value gives the percentage of mutations with a number higher than the x value. The orange and red lines indicate the largest differences between the grey and dark blue curves and between the grey and light blue curves, respectively. **(c)** Dotplot representing direct and indirect communication between PSD95^{pdz3} residues. The upper and lower triangles correspond to different thresholds. Each dot indicates a communication pathway linking 2 residues. Grey: the 2 residues are close in the sequence. Black: direct communication. Red/Pink: indirect communication. Isolated direct communications are encircled in blue. The secondary structures are also indicated. At the bottom, two communication motifs are mapped onto the 3D structure of PDZ. The pathways linking the residues in the motifs are displayed as black solid lines. The C-α atoms of the residues belonging to the motif are represented as grey spheres (black smaller spheres outside the motif). Dashed red lines indicate indirect communications.

Medium-scale application on hundreds of mutants

As a more quantitative application, we used our formalism to study the wild-type complex between PSD95^{pdz3} and its cognate substrate (CRIPT peptide) and 175 single-point mutants. We addressed two questions: (*i*) Is a particular substitution at a given position deleterious? (*ii*) What are the positions highly sensitive to mutations? We used data obtained from a deep mutational scanning experiment[4] to assess the predictive power of our approach.

To answer to the first question, the wild-type and mutated complexes were simulated for 100 ns (5 replicates of 20 ns), leading to a total of 17.6 µs. Over this time scale, we did not observe any significant difference in terms of shapes and motions between the different systems. Nevertheless, our analysis revealed a clear and statistically significant correlation between mutational phenotypic outcome and pathway concentration (**Fig. 1.5a-b**). Specifically the more deleterious the mutants, the higher the number of communication pathways and of highly connected residues (*i.e.* crossed by many pathways). We further showed that the number of highly connected residues can be used as a predictor of the severity of the mutations. We obtained an MCC of 47% on a balanced set of 15 highly deleterious mutations and 15 neutral ones.

To answer to the second question, we identified communication bridges (**Fig. 1.3b**) from the simulation of the wild-type complex only, and compared them with 20 experimentally identified 'hot spots'. These 'hot spots' are positions that display a very high sensitivity to virtually any substitution. In total, we detected 18 communication bridges, among which 16 were 'hot spots'. This corresponds to a sensitivity of 80% and a precision of 89%. Importantly, we could describe the role of these positions in the inter-residue communication and dynamical architecture of the complex, thanks to our rationale for picking them up. To assess the robustness and transferability of our results, we extended the MD simulations by several tens of nanoseconds, and we applied the same analysis on two unrelated systems, namely the the β-lactamase TEM-1 and the complex between growth hormone and its receptor. We found that our results were not affected by the duration of the MD simulations, and that our approach was pertinent and useful on the two other systems. Moreover, we carefully compared our results with those obtained from other structure-based methods (ENcoM[89], STRESS[70], PRS[62], RIP[60] and CARDS[77]) and sequence-based methods (JET[84], SCA[90], DCA[26] and MST[91]). We found that the predictive power of our approach is similar or higher than those methods. Moreover, there is a very good overlap between the set of infostery-detected residues, the set of conserved/coevolved and buried residues, and the set of residues highly sensitive to mutations. This clearly indicates a link between evolutionary constraints and structural constraints. One key ingredient of our infostery analysis is the usage of relatively short (tens of ns) MD simulations. This ensures the applicability of the method on a large scale.

1.3 Conclusions and perspectives

We have proposed and compared different approaches to predict the protein residues sensitive to mutations and the effect of single or multiple mutations on protein structure, dynamics and/or function. We have shown, along with other studies, that the massive amounts of sequence data available nowadays can be leveraged to predict mutational outcome on a large scale with high accuracy. Our main contribution was to propose a method that is much faster than the state of the art and more robust to variability in the input data. It also allows interpreting the predictions in terms of the evolutionary history of the sequences observed today in nature. We have also proposed a new formalism to automatically extract pertinent information from conformational ensembles, typically generated by relatively short MD simulations. By using this formalism, we could rationalise long-range mutational effects and guide mutation design. We quantitatively assessed its predictive power on several hundreds of single-point mutants of a protein complex. We demonstrated that the positions highly sensitive to mutations can be identified using information from the wild type only and we described their role in maintaining the structural stability of the protein.

Perspectives for GEMME consist in applying it to entire proteomes, investigating how it behaves on sequences coming from a population, versus sequences from different species. We have started a collaboration with M. Rera (IBPS, SU) on this matter, with a Master student (spring 2020). We will look at several hundreds of lineages in Drosophila and will focus on a specific phenotype, namely longevity. Another direction would be to extend the method to deal with non-canonical amino acids (NCAAs). Predicting which and where NCAAs can be incorporated shall be very useful for chemists. I have started to discuss about this with I. Coin (University of Leipzig). Perspectives for COMMA consist in reducing its computational cost by replacing the simulations with normal mode analysis or else. As a case study, we are considering successive mutations occurring in bacterial resistance to antibiotics. Another potential application of COMMA is the prediction of disordered regions, or regions with ambiguous secondary structure.

Chapter 2

The multiplicity of protein interactions

2.1 Motivation

Proteins regulate virtually all biological processes through a complex network of dynamical interactions[92]. A detailed description of these interactions is expected to provide direct information on the way to interfere with them and more generally on the (mal-)functioning of the cell[93]. However, the experimental assessment of all possible interactions of a protein is very challenging[94,95]. This has motivated the development of integrative approaches combining experimental and computational techniques to probe the 'molecular sociology of the cell'[96]. They have proven successful in determining the structures of several macromolecular assemblies[96–100] and have improved our understanding of protein interactions at the genome scale[101–107]. These efforts have revealed the complexity and multiplicity of protein interactions (**Fig.** 2.1). A protein may interact with several partners at the same time – each partner binding to a different site, or may present a shared binding region that will be used by different partners at different moments of its lifetime. It is estimated that as much as 75% of the surface could potentially be used for interactions[108]. Partners include other proteins and also nucleic acids. Determining whether a site is competent to bind only one type of partners or several types of partners is very challenging. For example, some proteins were shown to be able to bind both proteins and DNA via the same region[109,110], and there are more and more evidence that proteins accommodate indifferently DNA and RNA[111]. Another layer of complexity is added by conformational changes which may be substantial, especially upon binding to nucleic acids. In this context, there is a need for the development of tools able to decrypt protein surfaces at the residue level and to precisely estimate binding affinities on large ensemble of potential partners. Ideally, one would like to infer the number of interactions for a protein, identify precisely the borders of each interaction site possibly overlapping other sites, discriminate between strong and weak binders and identify the locations on a protein surface where artificial molecules (e.g. drugs) could best interfere with protein partners.

A long standing problem relevant to protein interactions has been to determine the 3D arrangement formed between two protein partners. This implies determining the position and orientation of the two proteins relative to each other and predicting binding-associated conformational changes. To address this problem, molecular docking algorithms have been developed, stimulated by the CAPRI competition[112]. Candidate conformations are evaluated based on properties reflecting the strength of the association, *e.g.* shape complementarity, electrostatics, desolvation, conformational entropy. Experimental data and evolutionary information (conservation or coevolution signals) can be included to improve the selection of candidate conformations[113–115]. Although docking methods have shown great improvements over the years, they are still far from perfect in correctly ranking near-native complex conformations and in modeling the conformational rearrangements associated to binding[116,117]. A related problem is that of identifying the surface regions involved in interactions. Evolutionary, physico-chemical and geometrical properties have been shown to be relevant to this issue[84,86,118–125], and a number of predictive tools have been developed based on them[126–132] (see[133,134] for surveys). Although some of these tools achieve very high accuracy against subsets of known experimental binding sites, their predictions are generally much smaller than the expected interacting surface size[108]. Moreover, many tools do not propose sites but rather evaluate the probability of a residue to be involved in interactions. Molecular docking calculations can also be employed for detecting binding sites, as it has been observed that the latter display a high propensity to be targeted by partners and also by non-interactors[135]. Last but not the least, a highly challenging problem is the identification of a protein's 'true' partners. In the context of a very crowded cellular environment, how can a protein distinguish its dedicated partners from non-interactors? While *in vitro*/*in vivo* experiments can suggest and test putative partners, computations provide a unique way to characterise interactions at very large scale and to explore the space of negatives, *i.e.* of what does not occur in the cell. Due to the limitations of scoring functions, the identification of interaction partners has generally been regarded as beyond the scope of molecular docking[136–138]. It is only recently that molecular docking-based strategies have been devised to this problem. The first proof-of-principle for such an approach applied to the reconstruction of biological networks was reported in[139]. In 2007, the first large-scale cross-docking study[140] for the prediction of interaction partners was launched on 168 proteins[141] whose interactions were known. This study highlighted the importance to develop appropriate concepts and tools for improving the discriminative power of molecular docking.

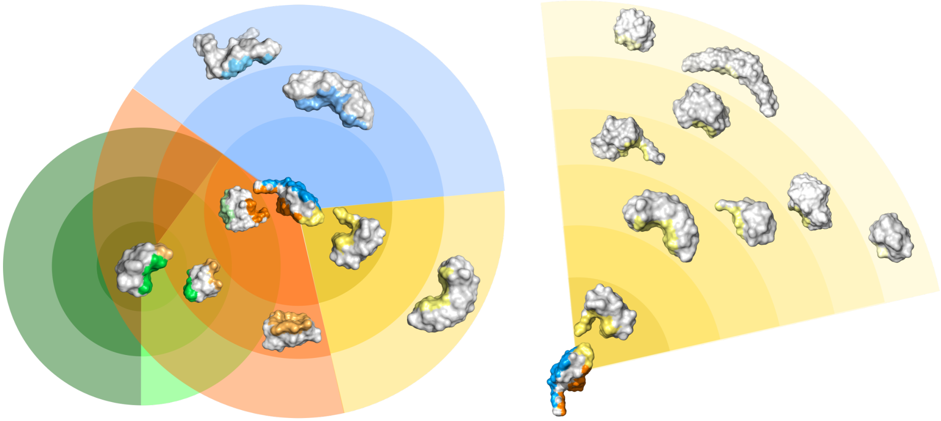


Figure 2.1: **Schematic representation of the complexity of protein interactions.** Proteins are represented as opaque surfaces, where the colours highlight the binding sites. Each circle corresponds to a given protein (in the center) and its potential partners. The concentric layers indicate different degrees of affinity (from strong to low when going away from the center). The sectors are coloured according to the corresponding binding sites. On the left, two cercles of affinity associate to 2 proteins are shown, for clarity, but a cercle could be drawn for each displayed protein. On the right, a sector is zoomed in and detailed.

In the following, I present contributions toward predicting protein interfaces with other proteins[142–144] and with nucleic acids[145], deciphering the complexity associated to protein surface usage by protein partners[146], efficiently detecting interfaces generated by high-throughput docking[147] and discriminating cognate partners from non-interactors[107,148,149]. They were published in eight research articles, among which three as first author and three as co-corresponding author.

2.2 Contributions

2.2.1 What is an interface?

Given a protein complex structure, interacting residues can be identified based on interatomic distances, changes in residue solvent accessible surface area (SASA) upon binding[150] or a Voronoi model of the interface[151]. Different views on protein interfaces are underlaid by these different criteria and they may complement each other. The ensemble of protein residues interacting with the partner form an *interacting site* (IS). This classical notion of IS is very restrictive and does not account for the interface variability that may come from structure ensembles. Indeed, the definition of the interface between two given proteins may vary from one structure to another, depending on the crystallization conditions, on the quality of the data/model and/or on the inherent flexibility of the assembly. What is more, the notion of IS masks the complexity of protein surface usage by multiple partners. This motivated us to define the new concept of *interacting region* (IR), obtained by merging overlapping ISs (typically, ≥ 60% overlap). We used this notion to describe the surface usage of 262 protein chains (P-262). Based on the observation that functional interfaces are conserved across closely related homologs[152], we collected all functional ISs involving these query proteins or their close homologs (≥ 90% sequence identity) from the Protein Data Bank (PDB)[153]. This amounted to 23 642 ISs, which were merged into 370 IRs (1.4 per protein chain). The notion of IR captures the multiplicity of protein surface usage by several partners (**Fig. 2.2a**) and also the interface variability coming from molecular flexibility (**Fig. 2.2b**). On average, about 50% of the protein surface is covered by functional interactions, in line with a previous study[108]. Moreover, a significant number of proteins have their surface completely or almost completely covered. If one were to consider only one PDB structure for each protein, the estimation would drop down to one third of the surface. This finding challenges the role of specificity in the evaluation of protein interface prediction methods and rather put the emphasis on precision.

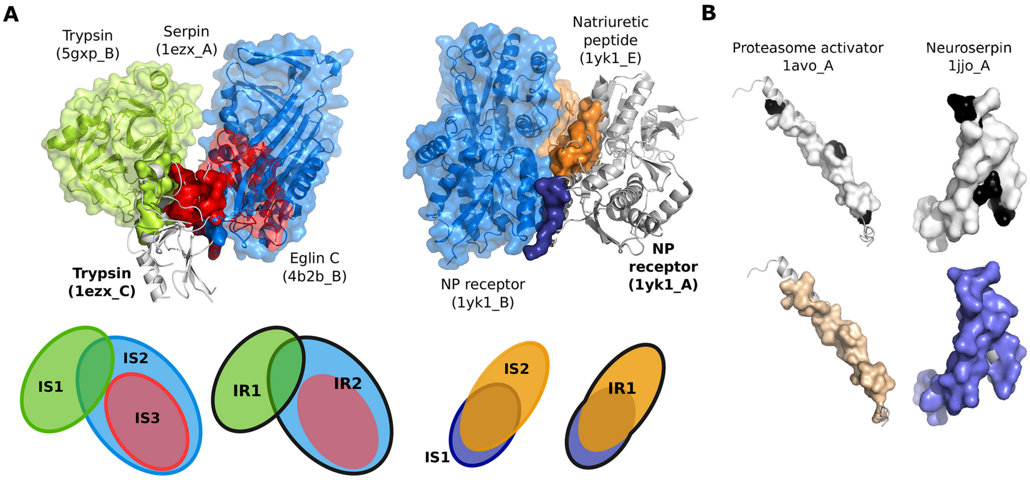


Figure 2.2: **Examples and schema illustrating the notions of interacting site and interacting region.** **(a)** The query proteins are displayed as grey cartoons, their interacting sites as opaque coloured surfaces and their partners as coloured cartoons and transparent surfaces. Left: trypsin (1ezx_C, in grey) interacts with itself (5gxp_B, in green), serpin (1ezx_A, in blue) and eglin C (4b2b_B, in red). The 3 corresponding ISs lead to the definition of 2 IRs, as depicted on the schema at the bottom, where each IR is contoured by a thick black line. Right: the natriuretic peptide receptor forms a homodimer (1yk1_A, in grey, and 1yk1_B, in blue) to bind its substrate (1yk1_E, in orange). The 2 ISs detected at the surface of one receptor monomer (1yk1_A, in grey) are merged into an IR. **(b)** On top, the IS defined from one PDB structure is coloured in white and the additional residues belonging to the IR are in black. The differences reflect the interface variability between different crystallographic structures of the same complex. At the bottom, the patches predicted by JET^2 are coloured in wheat (SC$_{cons}$) and purple (SC$_{notLig}$). The precision increases from 79 to 91% for 1avo_A and from 76 to 92% for 1jjo_A.

2.2.2 Protein-protein interface prediction

We predict interacting patches at the surface of the proteins from by relying on four biologically and physically meaningful residue properties (**Fig. 2.3a**): evolutionary sequence conservation inferred from the analysis of homologous sequences (T$_{\mathrm{JET}}$), physico-chemical properties expected at the interface based on experimentally known complex structures (PC), local geometry computed on the protein 3D structure (CV), and propensities to be found at docked interfaces inferred from high-throughput docking calculations (NIP). The calculation of the later requires to be able to efficiently treat millions of conformations generated by docking. To do so, we have developed INTBuilder(http://www.lcqb.upmc.fr/INTBuilder/), a fast, easy-to-use software whose complexity scales linearly with the number of atoms/residues. The four sequence- and structure-based properties, T$_{\mathrm{JET}}$, PC, CV and NIP, are used to feed a clustering algorithm according to several scoring strategies specifically aimed at detecting the *support*, the *core* and the *rim* of a protein interface (**Fig. 2.3b**, on the left). These three layers are defined for known experimental interfaces by comparing their solvent accessibilities in the presence and absence of the partner. Support residues are buried with and without the partner, core residues become buried upon binding to the partner and rim residues are exposed in the presence and absence of the partner[154]. A threshold of 25%

relative solvent accessibility is used to determine whether a residue is buried or not. To approximate these layers, our algorithm first identifies a small cluster of highly scored residues, called the *seed*. Seeds closer than 5 Å are merged. Then, the detected seeds are progressively extended, and the resulting residue clusters are merged if they are in contact (< 5 Å away). Importantly, the way residues are picked up based on their scores differs between seed and extension, such that the detected signal is very strong in the seed and progressively fades away as the extension is grown. Finally, an outer layer is added to form what we call a *predicted patch* (**Fig. 2.3b**, on the right). By running several iterations of the algorithm, a confidence score is assigned to each residue within each patch.

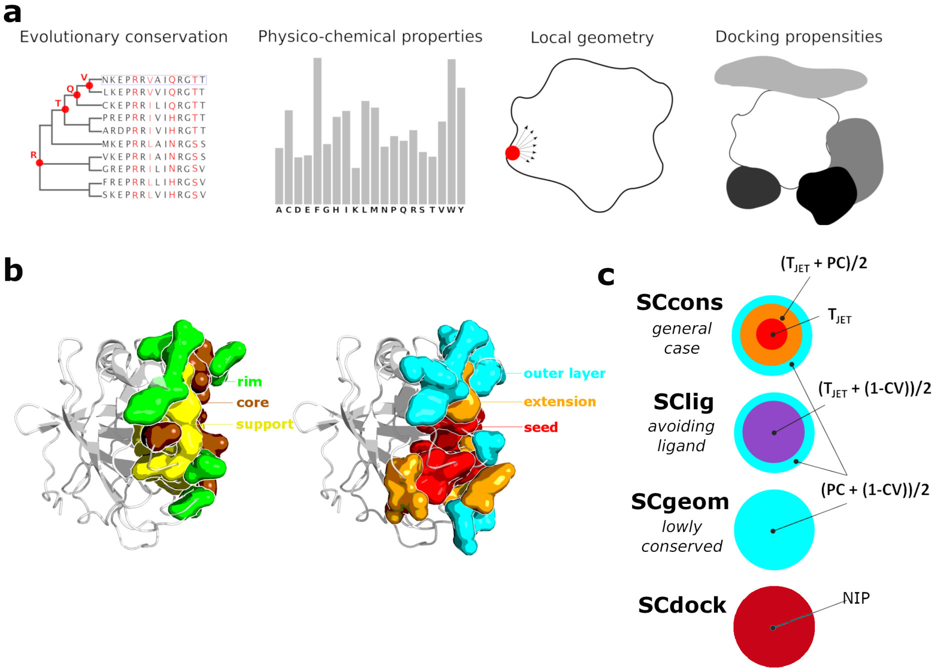


Figure 2.3: **Ingredients of JET^2 and dynJET^2 methods.** (**a**) Sequence and structural residue-based properties. (**b**) Section of an experimental interface (on the left, PDB code: 1N8O) and of the corresponding prediction using SC$_{cons}$ (on the right). The experimental and predicted interface residues are displayed in opaque surface: support, core and rim are in yellow, brown and green; cluster seed, extension and outer layer are in red, orange and cyan. (**c**) Schematic icons picturing the four scoring schemes. T$_{\text{JET}}$: conservation level, PC: interface propensity, CV: circular variance, NIP: docking propensity.

The first three properties, T$_{\text{JET}}$, PC and CV, are used to derive three different scoring strategies (**Fig. 2.3b**) designed to identify different types of protein-protein interfaces (**Fig. 2.3c**): **SC**$_{cons}$ targets very conserved residues (identified by the T$_{\text{JET}}$ score) to form a seed which is then extended using both T$_{\text{JET}}$ and PC scores, and complemented with an outer layer of protruding residues; **SC**$_{notLig}$ differs from SC$_{cons}$in that it not only accounts for conservation to detect the seed, but also for local geometry (CV); **SC**$_{geom}$ disregards evolutionary information and solely employs PC and CV scores for detecting all three layers of the interface. While SC$_{cons}$ is intended to detect diverse protein binding sites, SC$_{notLig}$ specifically targets protein interfaces located closeby or overlapping small ligand binding pockets, and SC$_{geom}$ is designed to deal with lowly conserved interfaces, *e.g.* antigen binding sites. These SCs are implemented in JET^2 (`http://www.lcqb.upmc.fr/JET2`). The fourth property, NIP, is used exclusively in a fourth strategy, SC$_{dock}$ (**Fig. 2.3c**), implemented in dynJET^2 (`http://www.lcqb.upmc.fr/dynJET2`). It reflects the propensity of each protein residue to bind partners and non-partners in docking calculations. To evaluate docking conformations, we used a coarse-grained empirical energy function comprising a Lennard-Jones potential for van der Waals interactions and a Coulomb potential for electrostatics[155]. JET^2 and dynJET^2 revisit the idea formalized in the Joint Evolutionary Trees (JET) method[84].

JET^2 was initially assessed against 238 protein complexes representing a wide spectrum of functional and structural classes[156,157]. It allowed to detect lowly conserved binding sites that were missed by JET, and to define protein-protein interfaces close to or overlapping ligand-binding pockets with improved sensitivity and precision. It performed similarly to machine learning algorithms employing tens of parameters[158–161]. JET^2 was further applied to more than 20 000 proteins, corresponding to the non-redundant set (at 40% identity) of all chains for which a high-quality 3D structure is available in the PDB. Predictions were evaluated on more than 15 000 experimentally characterised ISs. This is, to our knowledge, the largest evaluation of a protein binding site prediction method. The overall performance of JET^2 on all interfaces are: Sen = 52.52, PPV = 51.24, Spe = 80.05, Acc = 75.89. The generated knowledge base is available to the community at: `http://www.jet2viewer.upmc.fr/jet2_viewer`. JET^2 was also used in recent rounds of CAPRI and in the CASP13-CAPRI experiment. dynJET^2 was assessed against the set of ISs and IRs defined for `P-262`. On average, the predictions cover 60% of the protein surface, a bit more than the experimental estimation, and match IRs with an F1-score of 0.57 ± 0.19. A large amount of predicted patches better matched IRs, compared to ISs. This result is expected from a good protein interface prediction algorithm, as the notion of IR seems more biologically pertinent than that of IS in many cases, especially when the IR synthesises the variability inherent to structure ensembles of the same complex (**Fig 2.2b**). The agreement with experimental IRs depends on the type of interactions. In particular, the detection of the binding sites at the surface of antibodies and G proteins is very sharp, while the interfaces of the receptors and the enzymes regulators are the most difficult to detect. Evolutionary conservation, physicochemical properties and local geometry (SC_{cons}, SC_{notLig}, SC_{geom}) are generally able to better capture protein interface signals than the coarse-grained empirical energy function used in the docking experiment (SC_{dock}). Nevertheless, there are a number of cases where docking-based data provide valuable information to improve predictions by unveiling interfaces that could not be detected otherwise.

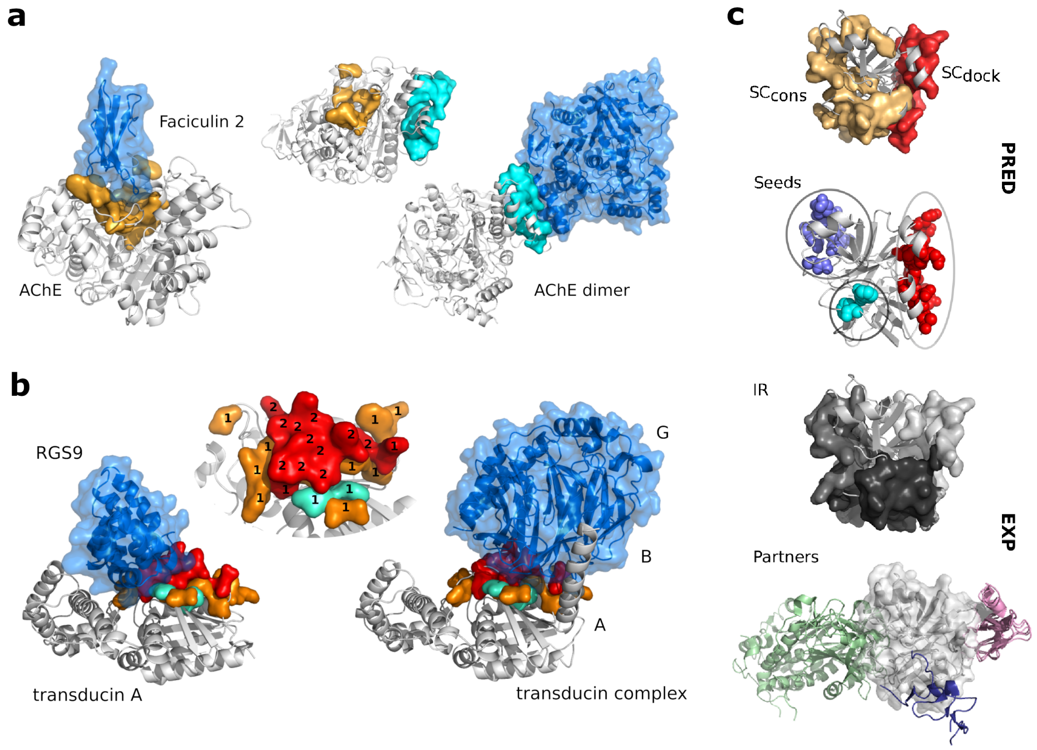


Figure 2.4: **Predictions of protein interfaces with different partners.** The protein of interest is represented as a grey cartoon while its partners are displayed as coloured cartoons (and transparent surfaces on the left). (**a**) Different partners binding to distinct regions (PDB code: 1MAH). The predicted patches are coloured according the scoring scheme used: SC_{cons} in orange and SC_{geom} in cyan. (**b**) Multi-usage of the same protein surface region (PDB codes: 1FQJ, on the left, and 1GOT, on the right). Residues (true positives) predicted by JET^2 (SC_{notLig}) are displayed as a coloured opaque surface: cluster seed, extension and outer layer are in red, orange and cyan. On the predicted patches are indicated, for each residue, the number of interactions it is involved in. (**c**) Heavy chain of the anticoagulation factor X displayed with the patches predicted by SC_{cons} (beige) and SC_{dock} (red), the patches' clustered seeds, the three experimental IRs for this protein and the corresponding partners.

2.2.3 Learning about interface origins and partners

Beyond their predictive power, JET^2 and dynJET^2 permit to dissect interaction surfaces and unravel their complexity. Indeed, the predicted patches are directly interpretable in terms of evolutionary origin and geometrical properties. This is illustrated by the example of acetylcholinesterase for which SC$_{cons}$ predicted a patch matching the conserved substrate binding site and SC$_{geom}$ predicted well the protein homodimeric interface (**Fig 2.4a**). The organisation of the predicted patches in different layers can also inform us about the specificity determinants of molecular recognition. For instance, the binding sites of transducin β-γ and of the regulatory protein RGS9 share 73% of their residues and are detected at 58% and 62% by the same SC$_{not\mathrm{L}ig}$patch (**Fig 2.4b**). Among the true positives, 64% (9/14) of the seed residues are involved in both interactions whereas most if not all residues from the extension and the outer layer are specific of the interaction with one partner. Finally, crossing the information coming from different SCs can help infer the existence of several IRs. An example is given by the heavy chain of the anticoagulation factor X, whose interface with the light chain ((**Fig 2.4c**, in light grey) is well detected by SC$_{dock}$ (in red), while the SC$_{cons}$ prediction (in beige) covers two experimental IRs (in grey and dark grey). In that case, looking at the seeds generated by SC$_{notLig}$, SC$_{geom}$, and SC$_{dock}$(second panel) allows resolving the ambiguity. On the P−262 dataset, we predicted 562 seeds (2.14 per protein), among which one quarter of the seeds are completely inside an IR (100% precision) and almost than half of the seeds detect an IR with very high (≥80%) precision. We also observed that the accumulation of seeds with different properties in a protein region is an indicator that this region will likely interact with many partners.

2.2.4 (Multiple) Protein-DNA interface prediction

JET^2 clustering algorithm was adapted and new scoring strategies were devised to predict DNA-binding sites at the surface of proteins. For this, we compiled a new dataset of 187 protein-DNA complex structures (HR-PDNA187), representative of all types of protein-DNA interactions, along with a subset of associated unbound structures (APO82). HR-PDNA187 comprises the largest body of non-redundant known high-resolution crystallographic protein-DNA complex structures. It covers all major groups of DNA-protein interactions according to Luscombe *et al.* classification[162]. It is freely available at: http://www.lcqb.upmc.fr/JET2DNA and could serve as a reference benchmark for the community. By analysing this dataset, we showed that the support-core-rim model could be useful to describe signals encoded in protein-DNA interfaces. We identified two different spatial organisations where the support and the core alternate their positions (**Fig. 2.5a**). This variety reflects the different ways a protein may bind to DNA. Nevertheless, in both organisations, the core plays the same physical role by stacking into the DNA grooves, while support residues tend to accommodate the DNA backbones. Similarly to what we did for protein-protein interfaces, we devised a scoring scheme relying mainly on conservation (D-SC$_{cons}$), to deal with the general case, and a scoring scheme focusing mainly on geometry (D-SC$_{geom}$), to deal with cases where no evolutionary information is available or where the whole protein surface displays a homogeneous conservation signal (**Fig. 2.5b**). In addition, we devised a scheme designed to specifically detect very concave interfaces, where the protein binds to DNA by "enveloping" it (*e.g.* polymerases). Interface propensities specific to DNA binding are considered in the three D-SC. They are notably different from protein-protein interface propensities. Indeed, protein-DNA interfaces are enriched in positively charged and polar residues (especially in the rim and the core), while protein-protein interfaces are enriched in hydrophobic residues (especially in the support, and the core).

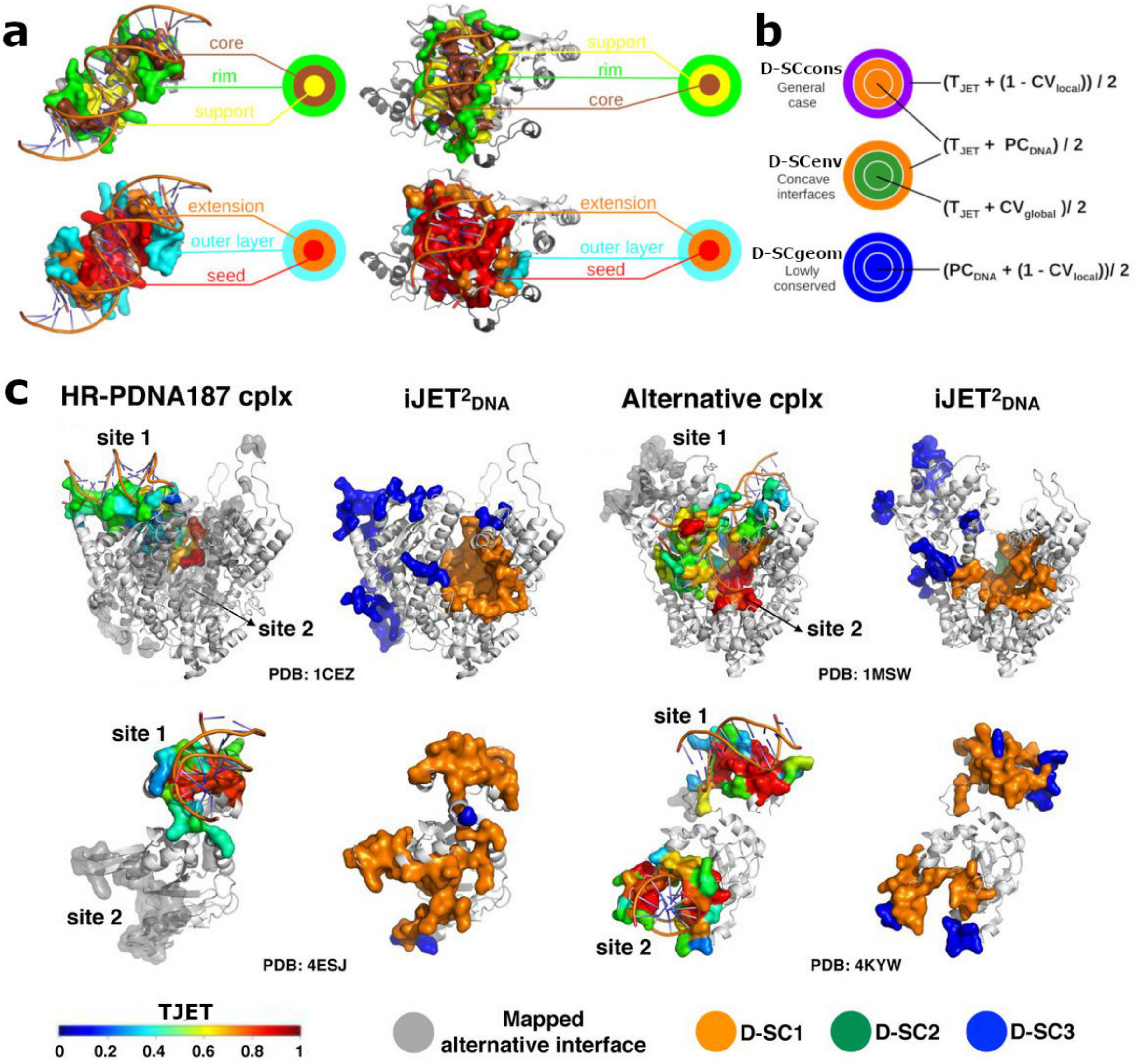


Figure 2.5: **Protein-DNA interface prediction.** (**a**) Sections of two experimental interfaces (on top, PDB codes: 1JE8 and 1D02) and the corresponding JET^2_{DNA} predictions using D-SC$_{env}$ (at the bottom). The colour code is the same as in Fig. 2.3b. **b**) Schematic icons picturing the three scoring schemes in JET^2_{DNA}. CV$_{local}$ and CV$_{global}$ are computed with a radius of 12 Å and 100 Å , respectively. Different colours correspond to different formulas used to detect the layers. (**c**) Examples of multiple binding site predictions. For each protein (each row), the experimental complex present in HR-PDNA187 and another experimental complex displaying a distinct DNA-binding site are shown on the first and third columns, respectively. For each structure, the 'main' DNA-binding site is coloured according to T_{JET} values and the site coming from the other structure is mapped and displayed in transparent grey. The predictions computed on each experimental structure are displayed on the second and fourth columns, respectively, and coloured according to the scoring scheme. On top: RNA polymerase from bacteriophage T7 (PDB: 1CEZ and 1MSW). At the bottom: R.DpnI modification-dependent restriction endonuclease (4ESJ and 4KYW).

JET$^2_{\text{DNA}}$ reached an average F1-score of 61% on HR-PDNA187 and of 58-59% on the subset of 82 proteins for which unbound forms are available. The performance is similar on bound and unbound forms, and for the vast majority of the proteins, the sensitivity on the unbound form is almost as high as (>80%) of even greater than that computed on the bound form. Hence, JET$^2_{\text{DNA}}$ is able to detect interacting residues even when they are 'hidden' by conformational changes. Assessed on a completely independent dataset (taken from[163]) comprising 57 DNA- and RNA-protein complexes, JET$^2_{\text{DNA}}$ achieved an average F1-score of 45. The sensitivity is roughly the same as for HR-PDNA187 but the precision (PPV) is lower. This can be explained by the fact that the experimental reference interfaces are defined using a much more stringent distance cutoff[163] than what we used for HR-PDNA187. The predictive performances are equivalent for both DNA- and RNA-binding sites, showing that these sites share similar properties. Compared to several machine-learning methods[163–165], JET$^2_{\text{DNA}}$ achieves a significantly higher sensitivity (by 20-40%) with similar accuracy. The main advantages of JET$^2_{\text{DNA}}$ compared to the other methods is that it can inform on the type of interactions the predicted patches may be engaged in and it can unravel alternative binding sites. For instance, it detects the highly conserved active site of the RNA polymerase from bacteriophage T7 with D-SC$_{cons}$ (**Fig. 2.5c**, on top, in orange) and the protruding and lowly conserved recognition site of the same protein with D-SC$_{geom}$ (in blue). Another example is given by the modification-dependent restriction endonuclease, where a combination of D-SC$_{cons}$ and D-SC$_{geom}$ is identifies very well the two binding sites of the protein (**Fig. 2.5c**, at the bottom). Both sites play the same role and are highly conserved. D-SC$_{geom}$ enables rescuing lowly conserved subregions that D-SC$_{cons}$ is not able to detect. Notice that the second site is well detected even when displaying a "distorted" conformation in the absence of DNA (on the left).

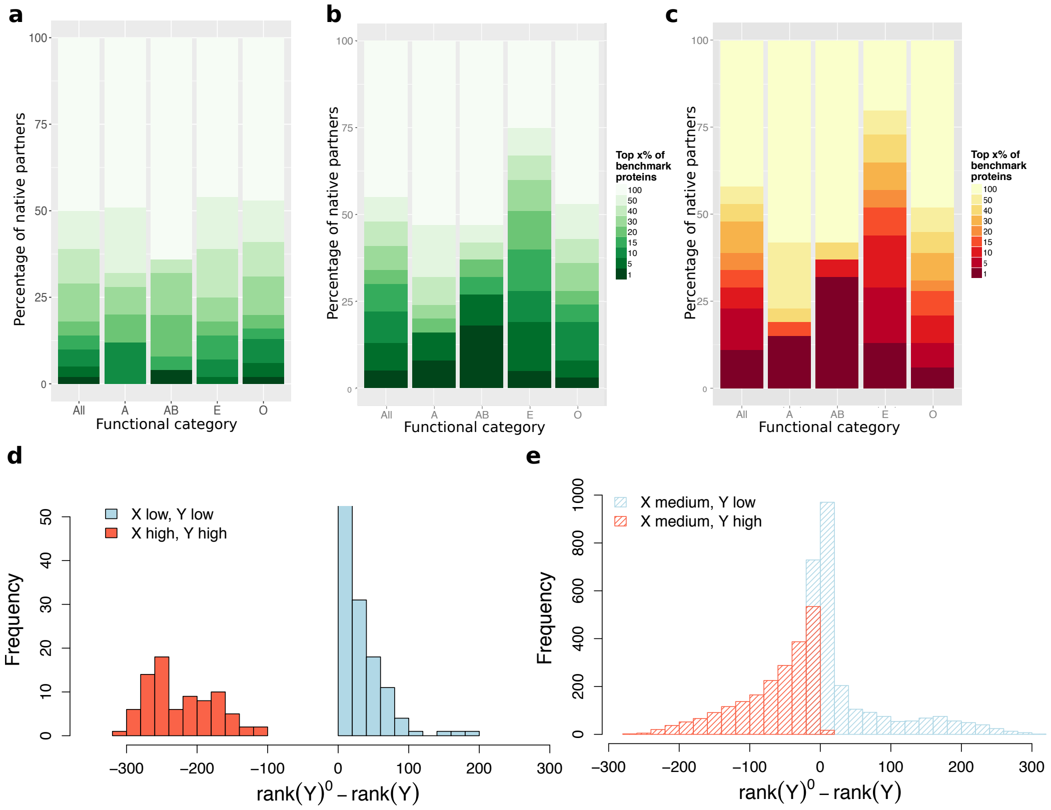


Figure 2.6: **Protein partner discrimination and global social behaviour.** **(a-c)** In y-axis are reported the percentages of cognate partners identified within the top x% of non-interactors, where x varies between 1 and 100 (colour scale). Each protein was docked against its native partner and 371 non-interactors (including itself). The proteins were ranked using: **(a)** shape complementarity score distributions, **(b)** interaction indexes II, **(c)** normalised interaction indexes NII. **(d-e)** Effect of the normalisation on proteins depending on their sociability. Given a protein X, the NII values enable to rank all the proteins from the dataset, from 1^{st} to 352^{nd}. We report the distributions of the number of ranks lost (negative values) or gained (positive values) by any protein Y. **(d)** both X and Y and either highly (in red, S ≥ 0.75) or poorly (in blue, S ≤ 0.19) sociable. **(e)** X has medium sociability while Y is highly (in red) or poorly (in blue) sociable.

2.2.5 Protein sociability

The knowledge of protein interfaces can be used to rank putative complex conformations generated by docking and to help the identification of the "true" partners versus non- or poorly-interacting pairs. Specifically, we performed a high-throughput complete cross-docking (CC-D) experiment, where 352 proteins[157] were docked against each other, using the docking program HEX[166]. To estimate the strength of the interaction between two proteins P_1 and P_2, we considered either the docking score, which is based on shape complementarity, or a measure reflecting the match between docked and known interfaces, which we call II:

$$\mathrm{II}_{P_1,P_2} = max(\mathrm{FIR}_{P_1,P_2}, \mathrm{FIR}_{P_2,P_1}), \tag{2.1}$$

where FIR_{P_1,P_2} and FIR_{P_2,P_1} (Fraction of Interface Residues) are the fractions of the docked interfaces, obtained when docking P_1 against P_2 and reciprocally, composed of residues belonging to the experimental interfaces for the two proteins (see *Materials and Methods*). The maximum is determined over the 2 x 2 000 best-scored poses from the 2 docking calculations involving P_1 and P_2. Experimental interfaces can be viewed as perfect predictions and allow estimating the maximum discriminative power one can expect from the interaction index II. Using this index, we could identify three times more cognate partners in the top 5% than when using the shape complementarity docking score (**Fig. 2.6**, compared **a** and **b**).

We further introduced a sociability index, or S-index, and used it to normalize the II indices before comparing them. The S-index of a proteins P_i is expressed as

$$\mathrm{S}_{P_i} = \frac{1}{\mathscr{P}} \sum_{P_j \in \mathscr{P}} \mathrm{II}_{P_i,P_j}, \tag{2.2}$$

and represents the degree of "sociability" of a protein: the higher the value of S, the more sociable the protein in the CC-D. Using the S-index to weight II values yielded strikingly improved results (**Fig. 2.6c**). The known partners of 80 proteins (23% of the benchmark set) were identified in the top 5%. Given a protein pair P_1P_2, the normalisation accounts for the sociability of P_1 and P_2 in the following way: if the proteins are highly (resp. poorly) sociable, *i.e.* their S values are high (resp. low), the interaction index II_{P_1,P_2} will be lowered (resp. raised). This procedure has a direct impact on the ranks of the potential partners (**Fig. 2.6e**). Poorly sociable proteins (in blue) generally gain ranks upon normalisation while highly sociable proteins (in red) are systematically penalised by the normalisation. Given a protein P with medium sociability, the down-shifting of highly sociable proteins may greatly help singling out its cognate partner. These results showed that to decide whether P_1 and P_2 interact together, the way P_1 and P_2 behave with all the other proteins in the dataset should be accounted for. Using the same computational experiments, we also showed that proteins from the same functional class have evolved to avoid interactions between them (low NII values), and that the ability to discriminate cognate partners from non-interactors is very much linked to the ability of the docking program to generate near-native interfaces. We found similar results with two other docking programs, namely ZDOCK[167] and MAXDo[140].

A question one may wonder is to what extent our notion of sociability is related to that of stickiness. The notion of stickiness is usually defined based on the content of hydrophobic residues at the surface of the protein[168]. Important efforts have been dedicated to characterising sticky proteins and their interactions[159,168–170]. It was shown that sticky proteins have stronger than average non-functional interactions and that avoiding such non-functional PPIs is an important constraint in protein evolution[159,169]. Sociability, as we define it here, reflect a tendency to glue to anyone in the docking calculations, without explicitly accounting for the physico-chemical properties of the surface. The most sociable proteins in the dataset are inhibitors and proteins with other function, displaying rather small interacting surfaces, without any particular compositional bias. Hence, the notion of sociability goes beyond that of stickiness: while a sticky protein has no preferential partner, we show that a protein might be sociable with all other proteins but display different degrees of sociability, with proteins playing different functional roles in the cell.

2.3 Conclusions and Perspectives

We have proposed to revisit the definition of protein interfaces to account for the multiplicity of protein surface usage and molecular flexibility. We have developed several computational methods to detect protein surface patches likely involved in interactions with other proteins and with nucleic acids. Contrary to machine learning methods, our approaches rely on a small number of biologically and physically relevant descriptors and combine them in a straightforward way. They provide a unique way to understand the origins and properties of the predicted sites and interpret them in light of their functions. For example, enzyme binding sites for substrates or inhibitors typically display a very conserved support and core, while antibody binding sites for antigens are highly protruding and variable. Transcription factors typically display one or two highly conserved binding sites, while polymerases form large concave interfaces. Another advantage of our approaches is their ability to discover alternative hidden binding sites. We have illustrated this with enzymes displaying both a protruding or flat and poorly conserved recognition site, and a highly conserved active site. The examples we showed let us envision a much larger complexity in the usage of protein surfaces by DNA than expected. In addition, we have compiled a new non-redundant dataset of protein-DNA complexes spanning a wide range of functional classes and have made it freely available for the community. We have performed CC-D calculations of several hundreds of proteins and have inferred systemic properties about the way proteins live together. We have proposed a formal definition of protein sociability and have demonstrated that the global social behaviour of a protein, with respect to many others, should be taken into account to identify its cognate partners and discriminate them from non-specific encounters.

A perspective for the JET^2/dynJET^2/JET^{2}DNA suite is to integrate its predictions in a docking engine. The residue-based scores provided by these methods could be mapped onto grids representing the proteins and exploited by fast Fourier transform (FFT) algorithms, to guide the sampling of candidate conformations. This may allow enriching the conformational ensembles with near-native solutions, especially for cases where large conformational changes occur. I have started a collaboration with S. Grudinin (Nano-D team, INRIA) on this subject. He has extensive experience with FFT docking algorithms. Another direction of improvement and extension for this work concerns the reconstruction of the cellular interactome. We have been developing a unified computational framework integrating information coming from CC-D calculations, interface prediction and binding affinity estimation toward a better identification of cellular partners. We have already achieved very high accuracy in discriminating cognate partners from non-interactors for certain functional classes of proteins. We are now moving toward a multi-conformation and multi-partner paradigm, using deep learning architectures. We are also investigating whether the effect of mutations can be predicted using machine learning with good generalisation performance. Being able to reconstruct interaction networks and predict how they are altered, possibly rewired, by mutations, is of paramount importance for personalised medicine. I am co-supervising a PhD student (2019-2022) with A. Carbone (LCQB, SU) on these issues.

Chapter 3

The evolution of protein isoforms

3.1 Motivation

Eukaryotes have evolved a transcription machinery that can augment the protein repertoire without increasing the genome size. It produces several mRNA transcripts from the same gene, by choosing different initiation/termination sites and/or by splicing different exons[171]. Alternative splicing (AS) concerns almost all multi-exon genes in vertebrates[172] and can result in protein coding transcripts with very diverse lengths, domains and amino acid compositions. It has been suggested that two protein isoforms may have completely different cellular partners[173] and may adopt different 3D folds[174]. AS has also gained interest for medicinal purpose, as the ratio of alternatively spliced isoforms is imbalanced in several cancers[175,176].

At the genomic level, the mechanisms responsible for the emergence of a new isoform have been well documented[177–179]. In recent years, deep surveys of the splicing complexity generated by such mechanisms across species have become possible, thanks to the advent of high-throughput sequencing (HTS) technologies like RNA-Seq[172,180]. Nevertheless, we still know relatively little about the actual impact of AS on the functioning of the concerned proteins[181]. This is in part due to the fact that most studies have focused on regulatory aspects. The second reason is technical, as it is very challenging to evaluate how many of the transcripts detected by HTS are translated and functional. Different experimental techniques give very different estimates[182,183]. Finally, AS tends to affect protein regions that are difficult to see experimentally, because they are intrinsically disordered and/or involved in transient interactions[184,185]. Hence, the extent to which and how AS modulates protein functions and interactions remains an open question.

The elusiveness of the significance of AS for protein function has stimulated the development of knowledge bases[186–189] providing sequence- and/or structure-based information and functional annotations at the level of the transcript or the exon. In particular, a pertinent proxy for function is evolutionary conservation. In this respect, a few methods have been proposed to reconstruct transcripts' phylogenies[190,191] or assess the presence of ASEs in several species[192]. One of the problems one has to face when studying transcript diversity across species is to determine a mapping of transcripts and/or of exonic regions between the considered species. Such a mapping is not trivial as genomic exons may vary in terms of sequence and length from one species to another. Moreover, the task may be complicated by the presence of duplicated sequences displaying high sequence similarity. Disentangling the orthology (coming from speciation) from paralogy (coming from duplication) relationships is challenging in such cases. Existing tools for orthology detection at the gene level rely on graph-based clustering of sequences, where sequence similarities are computed using pairwise alignments, or on tree-based methods reconciling the gene family trees with the species tree[193]. At the exon level, previous efforts have used pairwise alignments of genomic sequences[194,195] or multiple sequence alignments (MSAs) of concatenated translated exons[190]. However, there exists no automated and/or general method to detect orthologous exons while accounting for transcript diversity.

Structural characterization can also inform us about the potential function of a protein isoform and rationalize about the mechanisms underlying its (mal)-functioning. However, so far, very few structures of alternatively spliced isoforms have been described and are available in the PDB[196]. It was shown that the boundaries of single constitutive exons or of co-occurring exon pairs tend to overlap those of compact structural units, called protein units[197]. Moreover, some particular features (PFAM domain content, structural content, presence of binding motifs...) could be related to the existence of alternative isoforms. For instance, brain- or other tissue-regulated exons frequently overlap intrinsically disordered regions embedding conserved linear binding motifs[184]. Exciting results were obtained by Birzele and co-authors[198], who collected experimental data for 488 isoforms representing 367 proteins and provided a database of structural templates. They highlighted examples where ASEs affecting highly conserved protein regions resulted in major fold-changing events, suggesting new functional properties of the isoforms. A few cases of isoforms displaying domain atrophy while retaining some activity have also been reported[199].

In the following, I present contributions toward decomposing transcripts into evolutionary-wise building blocks[200], reconstructing plausible evolutionary scenarios explaining an ensemble of transcripts observed in a set of species and systematically assessing the impact of AS on protein structures[201]. They led to the publication of one research article as co-corresponding author and another one is in preparation.

3.2 Contributions

3.2.1 What is an (s-)exon?

Along evolution, genomic exons may diverge, undergo truncation or elongation, get lost or duplicated once, or several times. These events taking place at the level of the gene may be accompanied by a modulation of the usage of the exons in the transcriptome. In this context, decomposing transcripts into entities that are meaningful to date the appearance and fixation of AS events along evolution is not obvious. We propose to introduce the notion of *s(pliced)-exon* defined as the minimal building block necessary to describe the whole transcript variability observed in a set of species. Formally, an s-exon is a group of orthologous exonic regions, represented by a multiple sequence alignment (MSA). To identify s-exons, we combine pairwise alignment-based exon clustering with MSAs of concatenated exonic regions (**Fig. 3.1a**, on the left). The first step consists in comparing all genomic exons versus all by pairwise alignment, constructing a similarity graph and identifying clusters in the graph. Then, genomic exons are split into smaller pieces, called sub-exons, that account for the transcript variability within each species. Finally, for each cluster, an MSA is generated that satisfies the constraints given by the genomic coordinates of the sub-exons. S-exons are defined as column blocks in the MSA, delimited by the sub-exon boundaries. The algorithm is implemented in ThorAxe (`https://github.com/PhyloSofS-Team/thoraxe`), an efficient and fully automated tool that retrieves transcript data from Ensembl[202] and clean them, decomposes the transcripts into s-exons and builds a splice graph describing their variability (**Fig. 3.1a**, on the top right). Each node of the graph is an s-exon and each edge is a junction between two s-exons The graph is annotated with evolutionary conservation levels (the darker the colour the more conserved the s-exon/junction). Such a graph permits a straightforward identification of evolutionary conserved ASEs. Importantly, the transcripts come from different species, and the splice graph represents both the intra- and inter-species variability. While most methods work with genomic exons and perform species comparison after the construction of the splice graph, the original contribution of our work is that both the s-exons and the splice graph are transcript- and species-aware. Moreover, ThorAxe is able to deal with very small exonic regions and to disentangle paralogy from orthology relationships (**Fig. 3.1a**, on the bottom right).

3.2.2 Evolutionary history and structural characterization of transcript isoforms

To get insight into the functional implications of AS events, we have developed a unified computational framework combining sequence- and structure-based information. We infer plausible evolutionary scenarios explaining a set of protein coding transcripts in a set of species and we predict the 3D structures of the corresponding protein isoforms. The evolutionary inference is based on the maximum parsimony principle. Specifically, given a a gene tree and the observed transcripts at the leaves, represented as a collection of s-exons, we reconstruct a phylogenetic forest embedded in the gene tree that minimises the number of evolutionary events (creation, loss and mutation of a transcript, see **Fig. 3.1b**). The underlying evolutionary model is comprised of two levels, following[190]. At the level of the gene, exons can be absent, constitutive, or alternative (*i.e.* involved in at least one ASE), whereas at the level of the transcript, exons are either present or absent. The cost associated to the mutation of an exon naturally depends on its impact on the status of the exon at the gene level. For instance when the gain of an exon at the gene level shifts its status from absent to constitutive, the mutation will not be penalised. Our main contribution was to develop heuristics in order to treat complex cases in a computationally tractable way. Specifically, we have implemented a multi-start iterative strategy combined with a systematic local exploration around the best current solution to efficiently search the space of phylogenetic forests. Moreover, we have designed a branch-and-bound algorithm adapted to the problem of assigning transcripts between parent and child nodes. The reconstructed forests are provided with a user-friendly visualisation (**Fig. 3.1c**). In addition to phylogenetic reconstruction, we predict the 3D structures of the protein isoforms based on comparative modelling. To retrieve template structures of (possibly distant) homologs to the studied transcripts, we use the HH-suite[203]. The search is performed in an iterative way, so as to reach a maximum coverage of the query, and in an exon-centered manner, such that we provide an overview of the structural information available for each s-exon. The generated models (see examples on **Fig. 3.1c**) are annotated with sequence (exon boundaries) and structure (secondary structure, solvent accessibility, model quality) information. It becomes very easy to visualise the location of each exon on the modelled structure. Our approach is implemented in a fully automated tool, Phylogenies of Splicing Isoforms Structures or PhyloSofS (`https://github.com/PhyloSofS-Team/PhyloSofS`).

3.2.3 The c-Jun N-terminal kinase family as a case study

As a proof of concept, we used our methods to date the appearance of ASEs inducing functional diversity in the c-Jun N-terminal kinase family (JNK1, JNK2 and JNK3) and to provide a mechanistic explanation for their outcome. JNKs play essential regulatory roles by targeting specific transcription factors. Several alternative JNK transcripts performing different functional tasks have been experimentally identified and characterised[181,204,205]. We considered 7 species, namely *H. sapiens, M. musculus, X. tropicalis, T. rubripes, D. rerio, D. melanogaster* and *C. elegans*. With 60 observed transcripts assembled from a total of 19 different s-exons, this case represents a high degree of complexity for the phylogenetic reconstruction. Most transcripts are comprised of more than 10 s-exons, and the number of transcripts per gene per species varies from 1 to 8 (**Fig. 3.1c**). We inferred a phylogenetic forest containing 7 transcript trees (**Fig. 3.1c**), relating 12 transcripts observed in human across the three genes. Based on it, we dated the appearance of an ASE involving a pair of mutually exclusive s-exons, namely *6* and *7*, in the ancestor common to mammals, amphibians and fishes. We found that the most ancient of these two s-exons is s-exon *7*. By characterising in detail the structural dynamics of two human isoforms, JNK1α and JNK1β, bearing one or the other s-exon, we could detect changes in the side-chain flexibilities of a few residues differing between the two s-exons (**Fig. 3.1c**, in the top right corner, in orange and purple). These residues are predicted as involved in interactions by JET[2] and thus the subtle differences we observed may be responsible for the selectivity of the JNK isoforms toward their substrates[206,207]. Our transcripts phylogeny also highlighted an isoform that was not previously described in the literature, namely JNK1δ (**Fig. 3.1c**). Despite displaying a large deletion (∼ 80 residues), it is conserved across several species and MD simulations suggested that it is stable in solution (**Fig. 3.1c**, in the top right corner, in gray-pink). In total, we could reconstruct a phylogeny for 46 out of the 60 observed transcripts. The 14 orphan transcripts (**Fig. 3.1c**, leaves in grey) are not conserved across the studied species, and thus likely result in non-functional protein products. As a support for this hypothesis, our structural analysis showed that they displayed properties likely reflecting

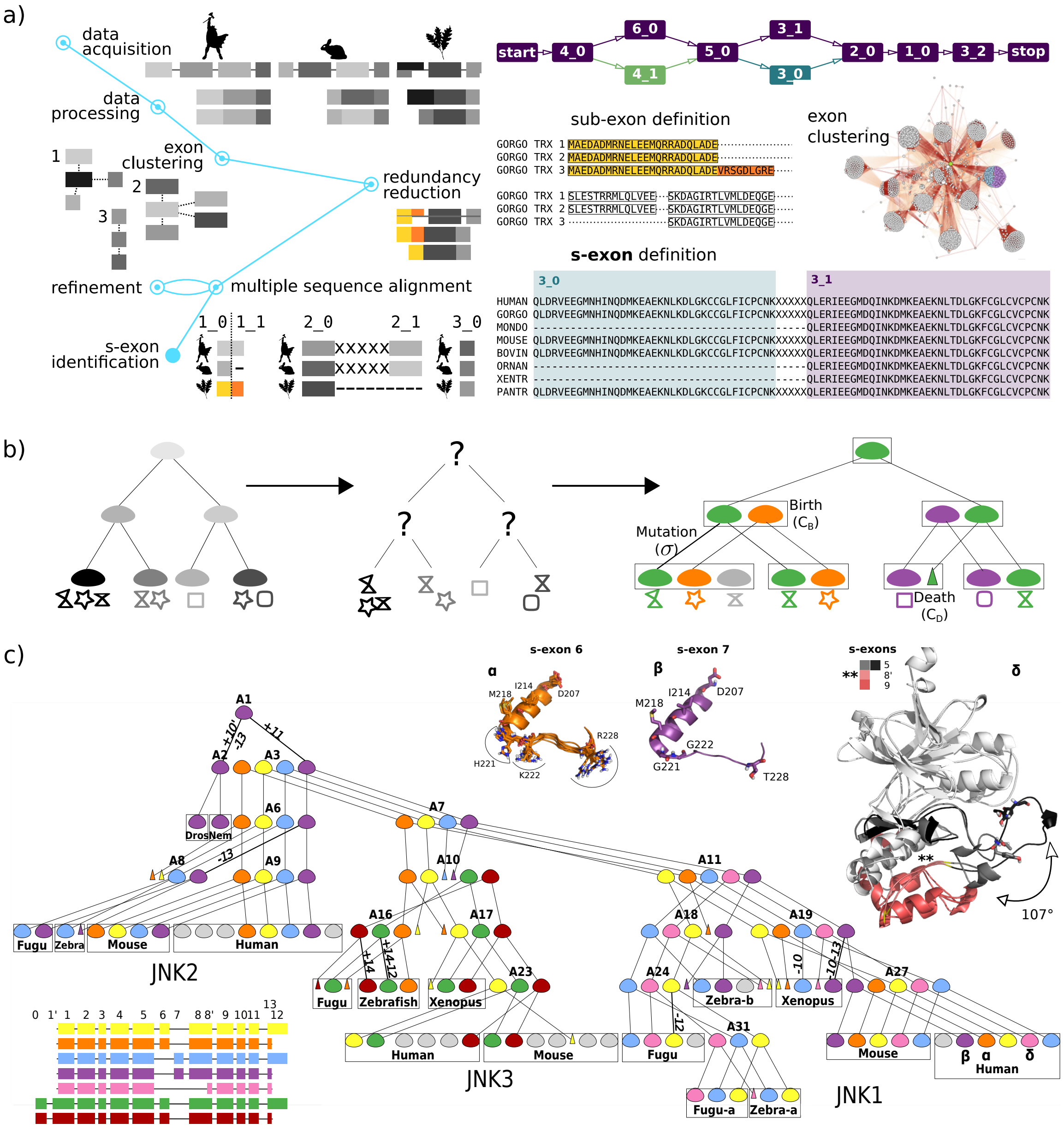

a)
data acquisition
data processing
exon clustering
redundancy reduction
refinement
multiple sequence alignment
s-exon identification
1
2
3
1_0 1_1 2_0 2_1 3_0
XXXXX
XXXXX
start
4_0
6_0
4_1
5_0
3_1
3_0
2_0
1_0
3_2
stop
sub-exon definition
GORGO TRX 1 MAEDADMRNELEEMQRRADQLADE
GORGO TRX 2 MAEDADMRNELEEMQRRADQLADE
GORGO TRX 3 MAEDADMRNELEEMQRRADQLADEVRSGDLGRE
GORGO TRX 1 SLESTRRMLQLVEE SKDAGIRTLVMLDEQGE
GORGO TRX 2 SLESTRRMLQLVEE SKDAGIRTLVMLDEQGE
GORGO TRX 3 SKDAGIRTLVMLDEQGE
exon clustering
s-exon definition
3_0
3_1
HUMAN QLDRVEEGMNHINQDMKEAEKNLKDLGKCCGLFICPCNKXXXXXQLERIEEGMDQINKDMKEAEKNLTDLGKFCGLCVCPCNK
GORGO QLDRVEEGMNHINQDMKEAEKNLKDLGKCCGLFICPCNKXXXXXQLERIEEGMDQINKDMKEAEKNLTDLGKFCGLCVCPCNK
MONDO ---QLERIEEGMDQINKDMKEAEKNLTDLGKFCGLCVCPCNK
MOUSE QLDRVEEGMNHINQDMKEAEKNLKDLGKCCGLFICPCNKXXXXXQLERIEEGMDQINKDMKEAEKNLTDLGKFCGLCVCPCNK
BOVIN QLDRVEEGMNHINQDMKEAEKNLKDLGKCCGLFICPCNKXXXXXQLERIEEGMDQINKDMKEAEKNLTDLGKFCGLCVCPCNK
ORNAN ---QLERIEEGMDQINKDMKEAEKNLTDLGKFCGLCVCPCNK
XENTR ---QLERIEEGMEQINKDMKEAEKNLTDLGKFCGLCVCPCNK
PANTR QLDRVEEGMNHINQDMKEAEKNLKDLGKCCGLFICPCNKXXXXXQLERIEEGMDQINKDMKEAEKNLTDLGKFCGLCVCPCNK
b)
?
?
?
Birth (C_B)
Mutation (σ)
Death (C_D)
c)
s-exon 6
s-exon 7
s-exons
α
β
δ
**
5
8'
9
M218
I214
D207
R228
H221
K222
G222
G221
T228
107°
A1
A2
A3
A6
A7
A8
A9
A10
A11
A16
A17
A18
A19
A23
A24
A27
A31
+10'
-13
+11
-13
+14
+14-12
-10
-10-13
-12
DrosNem
Fugu
Zebra
Mouse
Human
JNK2
Fugu
Zebrafish
Xenopus
Human
Mouse
JNK3
Zebra-b
Xenopus
Fugu
Fugu-a
Zebra-a
Mouse
β α δ
Human
JNK1
0 1' 1 2 3 4 5 6 7 8 8' 9 10 11 12 13

Figure 3.1: **Evolution and structures of transcript isoforms.** **(a)** Schematic workflow of ThorAxe pipeline. The different steps of the method are explained on the left and illustrated on the right. The s-exons are displayed as columns of gray boxes on the bottom left, as nodes in the splice graph on the top right and as coloured bocks in MSAs on the bottom right. An exemple of a pair of highly similar mutually exclusive homologous exons (3_0 and 3_1) is given. **(b)** Principle of PhyloSofS' reconstruction of transcripts phylogenies. Given a gene tree and transcripts observed at the leaves, PhyloSofS infers a forest of transcript trees embedded in the gene tree. The problem addressed here is that of a partial assignment: how to pair transcripts so as to maximize their similarity? **(c)** Transcripts' phylogeny reconstructed by PhyloSofS for the JNK family. The forest comprises 7 trees, 19 deaths (triangles) and 14 orphan transcripts (in grey). Mutation events are indicated on branches by the symbol + or - followed by the number of the exon being included or excluded (*e.g.* *+11*). The cost of the phylogeny is 69 (with $C_B = 3$, $C_D = 0$ and $\sigma = 2$). On the bottom left corner are displayed the exon compositions of the human isoforms for which a phylogeny could be reconstructed. On the top right corner, are displayed some representative MD conformations of the human isoforms JNK1α (orange), JNK1β (purple) and JNK1δ (gray-pink). For JNK1α (resp. JNK1β), we focus on the s-exon *6* (resp. *7*). The clustering was performed based on position 228 (RMSD cutoff of 1.5 Å) and yielded 8 conformations for JNK1α (resp. 1 for JNK1β). For JNK1δ, we show a superimposed pair of conformations illustrating the amplitude of the A-loop motion. Exons *5*, *8'* and *9* are indicated by colours and labels. For clarity, *8'* is also indicated by two stars on the structure.

structural instability (large truncations, poorer quality, larger and more hydrophobic surfaces). We found alternative phylogenies with equivalent cost, but they differ only slightly from the one reported here. Moreover, we showed that our phylogeny was robust to small parameter changes.

3.2.4 Tracking functional ASEs in protein evolution and structures

We then scaled up the analysis to an original set of 50 human genes (16 gene families) that we compiled. The rationale for choosing these genes was that several splice variants had been experimentally shown to perform different functional tasks. We considered one-to-one orthologs of these genes in 12 species, namely (*H. sapiens*, *G. gorilla*, *M. mulatta*, *M. domestica*, *R. norvegicus*, *M. musculus*, *B. taurus*, *S. scrofa*, *O. anatinus*, *X. tropicalis*, *D. rerio*, *D. melanogaster* and *C. elegans*). We collected about 900 transcripts annotated in Ensembl, comprising between 8 and 91 s-exons. We found that the vast majority of the documented functional splice variants (corresponding to 48 ASEs) were conserved in several (more than 3) species and that the variations between them affected interactions with cellular partners. Moreover, we identified 11 conserved new ASEs. About half of the detected ASEs are insertions/deletions and the other half is mutually exclusive tuples or alternative starts/ends. Structural information is available for a bit less than 50% of the 105 associated s-exons. As illustrative examples, let us mention TPM1 and CAMK2B. For TPM1, the splice graph produced by ThorAxe allowed the identification of a couple of complex ASEs involving some substitutions and insertions of homologous exons. All exons were predicted as folding into an α-helix and there exist cryo-EM structures showing they interact with actin. Hence, the ASEs likely modulate the speed at which actin filaments form and their length. For CAMK2B, the splice graph clearly showed that most of the AS concerns the linker between the two domains (kinase and hub) of the protein. Most of the events consist in insertion of disordered segments. We could identify a triplet of highly similar s-exons being part of the same insertion event. They are enriched in prolines, suggesting that they may be involved in interactions with cellular partners.

3.3 Conclusions and perspectives

We have developed a couple of computational methods to investigate the evolution of protein transcript isoforms and the impact of AS events on their structural dynamics. Our work allows to put together, for the first time, two types of information, one coming from sequence analysis and phylogenetic inference, and the other from molecular modelling. We should stress that the problem of pairing transcripts across homologous and paralogous genes between different species, addressed here, is much more complex than that of inferring the transcripts' phylogeny of each gene separately. Indeed, in the former case, the problem size is bigger, one needs to reconcile the gene tree with the species tree, and the sequences are more divergent. By applying our method to the JNK family, we could date functional ASEs, identify a new conserved isoform that seems stable in solution and provide mechanistic explanation for AS-induced phenotypes. By scaling up the analysis to a few tens of genes, we assessed the evolutionary origins of functional ASEs documented in the literature and we discovered new ASEs. We should stress that our iterative exon-centred strategy for modelling the 3D structures if the isoforms is the first of its kind.

The reliability of the transcript expression data clearly constitutes a present limitation of our methods. However, as experimental evidence accumulates and precise quantitative data become available, they will become instrumental in assessing the contribution of AS in protein evolution. Although PhyloSofS was applied here to study the evolution of transcripts in different species, it has broad applicability and can be used to study transcript diversity and conservation among diverse biological entities. The entities could be at the scale of one individual/species (tissue/cell differentiation), or different species (matching cell types), or a population of individuals affected (or not) by a multifactorial disorder.

Chapter 4

Future work

In the next four years, I will pursue my efforts to elucidate the relationship between genotype and phenotype, particularly in the context of protein interactions and alternative splicing (AS). I will leverage the expertise I have acquired and the work we have done for identifying functional sites and predicting the impact of sequence variations at these sites. The three themes developed in this dissertation are intertwined and my research plan will emphasise the cross-talks and overlaps between them. I intend to target most of my efforts to describe, understand and exploit the way AS generates protein functional diversity in eukaryotes. As stated in the third chapter, our knowledge on this matter is still relatively limited. Yet, this topic is of paramount importance, both for our fundamental understanding of living organisms and for the development of new treatments.

Consistent with the literature, our preliminary results on a benchmark set of 50 genes have shown that evolutionary conserved ASEs tend to impact regions directly or indirectly involved in interactions with other proteins, small molecules or nucleic acids (**Fig. 4.1**). We found cases where a disordered region or loop is inserted nearby an interaction site, or an interacting segment is replaced by a very similar one, or several similar binding motifs are inserted. These examples illustrate the way AS modulates the strength, the speed and the selectivity of functional cellular interactions. Predicting precisely the impact of the AS-induced variations is a very challenging task because we still do not know much about the determinants of molecular specificity and binding affinity. State-of-the-art methods predicting binding affinities and how they are altered by mutations crucially lack generalisation performance[208,209]. Moreover, the task is complicated by the fact that a significant proportion of the alternatively spliced regions are intrinsically disordered. I propose a new view on the problem, by inverting the question: what can evolutionary conserved ASEs tell us about the molecular determinants of protein function(s) and interactions?

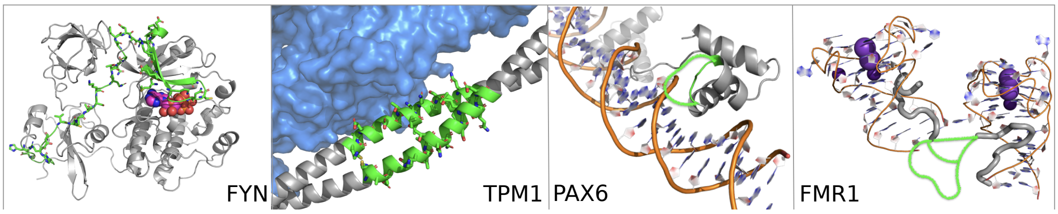


Figure 4.1: **Examples of conserved ASEs affecting interactions.** The protein is displayed as a grey cartoon, with the region affected by the ASE coloured in green. When the structure of the region is not known, it is simply represented by a hand-drawn line. When the ASE involves two mutually exclusive segments, the residues differing between the two segments are highlighted in sticks. FYN interacts with a small molecule (in spheres), TPM1 with a protein partner (in surface), PAX6 with DNA (in cartoon) and FMR1 with RNA (in cartoon).

Our working hypothesis is that ASEs encode precious information about where and how to target a protein in order to modulate its interactions. We will extract this information and use it to expand protein diversity way beyond what is observed today. We will exploit the growing body of RNA-Seq data across species to discover new conserved ASEs not yet annotated in public databases. Indeed, we have seen in our preliminary work that the publicly available annotated transcript data are incomplete. By directly looking at raw RNA-Seq evidence and relying on the percent spliced in (PSI) measure[210], we could further increase by about 35% the number of conserved ASEs in our benchmark set. We will develop algorithms to integrate RNA-Seq data (splice junctions) in ThorAxe splice graphs. On this matter, we have initiated a collaboration with P. De La Grange, co-founder of the Genosplice company. The company has extensive expertise and know-how in ASE detection from RNA-Seq data. They are willing to share their expertise and also their in-house database FastDB[211], which implements a series of filtering and rescuing protocols to get reliable sets of transcripts. The developed computational framework will be applied to the ~20,500 Human protein coding genes available in Ensembl. We will identify the set of ASEs observed in Human and conserved in several species spanning the tree of life and for which RNA-Seq data are available in the Bgee database[212]. We will further create an atlas of ASEs for all known protein domain families. We will use Pfam classification[213], which contains ~ 18 000 domain families. The rationale will be that the functional ASEs detected on one domain of one particular protein can inform us about the functional sites of the whole Pfam family to which the domain belongs. We will gather all available structural data, to describe the conformational flexibility and the interactions of the s-exons. Based on our preliminary work, we estimate that we will be unable to find structural information for a significant portion of the s-exons (see examples of loop insertions on Fig. 2, for PAK6 and FMR1). That is why I have started a collaboration with J. Cortés, who recently developed a method predicting probabilities of folding into the main types of secondary structures[214]. It exploits structural information encoded in tripeptide fragments from coil regions. Our multi-strategy approach will make the construction of a 'structural profile' feasible for virtually all s-exons involved in the detected functional ASEs. This knowledge base shall provide very useful information to biologists and medicinal chemists about the regions of a protein that are important for its function and how they can be modified or targeted to modulate this function. We will also develop a probabilistic model that will learn from the functional ASEs observed today in nature to generate new protein functional diversity. The rationale will be that the means AS has produced to generate protein diversity along evolution can be reused and generalized to expand the protein repertoire way beyond what we observe today. The challenge here will be to design an artificial system able to learn the underlying rules of (functional) AS and to generalize to any protein sequence. Solutions found by living organisms have often proven very complex and diverse, such that generalizing explicative and predictive models is a difficult task. At the same time, the conditions favorable to life are strongly constrained, and hence one can expect strong regularities in the way natural diversity is generated. Deep neural nets are especially suited to deal with complex data and spot their regularities. We will particularly focus on variational auto-encoders (VAE) and deep neural network-powered autoregressive models (DNNAM), which have proven very powerful to predict the outcomes of mutations and insertions/deletions and design protein sequences with desired properties[3,215,216]. DNNAM have the advantage of being reference-free (no need, for example, to align training sequences) and some recent implementations[217] deal with 3D information (protein structures represented as graphs) as input. We will use the knowledge acquired with the atlas to guide the design of our architecture and determine the representation of the input data. We will use the ASEs identified in Human as training set. We expect our initial set to be of the order of a few thousands and we will augment it with information coming from the other species. For both the design of the architecture and the representation of the input data, we will benefit from the know-how and expertise of our collaborator S. Grudinin, who has been developing pioneering deep learning-based methods for protein structure quality assessment[218,219].

This project embeds original concepts concerning the relationship between genotype and phenotype. The idea that the way living organisms generate protein diversity through AS can inform us about the fundamental determinants of protein functioning and can be reused to guide protein design is new. The question we address could not have been answered before, due to conceptual and technological reasons. The rapid growth of data generated by HTS and of studies applying deep learning to biological issues (including AS, but with different goals and views compared to our proposal) makes this project timely. Our preliminary work has been accomplished within the framework of the MASSIV project (ANR-17-CE12-0009, 2018-2021), which I am co-leading with H. Richard, and which will be running for one more year. This will enable starting the proposed plan with a post-doctoral fellow (D. Zea) and two Master students (B. Moindrot and P. Charpentier). I am currently looking for funding to ensure the continuity of the work on the subject (Projet Emergence de la Ville de Paris, Doctoral grant from the SCAI...).

Bibliography

[1] Fowler, D. M.; Fields, S. *Nat. Methods* **2014,** *11,* 801–807. 7

[2] Gasperini, M.; Starita, L.; Shendure, J. *Nature Protocols* **2016,** *11,* 1782–7. 7

[3] Riesselman, A. J.; Ingraham, J. B.; Marks, D. S. *Nat. Methods* **2018,** *15,* 816–822. 7, 10, 34

[4] McLaughlin, R. N.; Poelwijk, F. J.; Raman, A.; Gosal, W. S.; Ranganathan, R. *Nature* **2012,** *491,* 138–142. 7, 14

[5] Firnberg, E.; Labonte, J. W.; Gray, J. J.; Ostermeier, M. *Mol. Biol. Evol.* **2014,** *31,* 1581–1592. 7

[6] Boucher, J. I.; Bolon, D. N.; Tawfik, D. S. *Protein Sci.* **2016,** *25,* 1219–1226. 7

[7] Louie, R. H. Y.; Kaczorowski, K. J.; Barton, J. P.; Chakraborty, A. K.; McKay, M. R. *Proc. Natl. Acad. Sci. U.S.A.* **2018,** *115,* E564-E573. 7, 8, 11

[8] Hopf, T. A. *et al. Nature Biotechnology* **2017,** *35,* 12–135. 10

[9] Flynn, W. F.; Haldane, A.; Torbett, B. E.; Levy, R. M. *Mol. Biol. Evol.* **2017,** *34,* 1291–1306. 8, 11

[10] Figliuzzi, M.; Jacquier, H.; Schug, A.; Tenaillon, O.; Weigt, M. *Mol. Biol. Evol.* **2016,** *33,* 268–280. 7

[11] Barton, J. P. *et al. Nat Commun* **2016,** *7,* 11660. 8

[12] Hart, G. R.; Ferguson, A. L. *Phys Biol* **2015,** *12,* 066006.

[13] Mann, J. K. *et al. PLoS Comput. Biol.* **2014,** *10,* e1003776. 7

[14] Ferguson, A. L. *et al. Immunity* **2013,** *38,* 606–617. 7, 8

[15] Sim, N. L. *et al. Nucleic Acids Res.* **2012,** *40,* W452–457.

[16] Dehouck, Y.; Kwasigroch, J. M.; Gilis, D.; Rooman, M. *BMC Bioinformatics* **2011,** *12,* 151.

[17] Adzhubei, I. A. *et al. Nat. Methods* **2010,** *7,* 248–249.

[18] Cheng, J.; Randall, A.; Baldi, P. *Proteins* **2006,** *62,* 1125–1132.

[19] Capriotti, E.; Fariselli, P.; Casadio, R. *Nucleic Acids Res.* **2005,** *33,* W306–310.

[20] Ng, P. C.; Henikoff, S. *Nucleic Acids Res.* **2003,** *31,* 3812–3814. 7

[21] Levasseur, A.; Pontarotti, P.; Poch, O.; Thompson, J. D. *Evolutionary Bioinformatics* **2008,** *4,* EBO–S597. 7

[22] Lecompte, O.; Thompson, J. D.; Plewniak, F.; Thierry, J.-C.; Poch, O. *Gene* **2001,** *270,* 17–30. 7

[23] Breen, M. S.; Kemena, C.; Vlasov, P. K.; Notredame, C.; Kondrashov, F. A. *Nature* **2012,** *490,* 535–538. 7

[24] McCandlish, D. M.; Shah, P.; Plotkin, J. B. *Genetics* **2016,** *203,* 1335–1351. 7

[25] Stein, R. R.; Marks, D. S.; Sander, C. *PLoS Comput. Biol.* **2015,** *11,* e1004182. 7

[26] Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. *Proc. Natl. Acad. Sci. U.S.A.* **2009,** *106,* 67–72. 7, 15

[27] Haldane, A.; Levy, R. M. *Phys Rev E* **2019,** *99,* 032405. 7

[28] Barton, J. P.; Cocco, S.; De Leonardis, E.; Monasson, R. *Phys Rev E Stat Nonlin Soft Matter Phys* **2014,** *90,* 012132. 7

[29] Joerger, A. C.; Ang, H. C.; Veprintsev, D. B.; Blair, C. M.; Fersht, A. R. *J. Biol. Chem.* **2005,** *280,* 16030–16037. 8

[30] Wong, K. B. *et al. Proc. Natl. Acad. Sci. U.S.A.* **1999,** *96,* 8438–8442. 8

[31] Kumar, S.; Clarke, D.; Gerstein, M. *Nucleic Acids Res.* **2016,** *44,* 10062–10073. 8

[32] Raman, A. S.; White, K. I.; Ranganathan, R. *Cell* **2016,** *166,* 468–480. 8

[33] Monod, J.; Wyman, J.; Changeux, J. P. *J Mol Biol* **1965,** *12,* 88–118. 8, 12

[34] Goguet, M.; Narwani, T. J.; Petermann, R.; Jallu, V.; de Brevern, A. G. *Scientific reports* **2017,** *7,* 1–13. 8

[35] Saladino, G.; Gervasio, F. L. *Curr. Opin. Struct. Biol.* **2016,** *37,* 108–114.

[36] Lu, S.; Jang, H.; Nussinov, R.; Zhang, J. *Sci Rep* **2016,** *6,* 21949.

[37] Kamaraj, B.; Bogaerts, A. *PLoS ONE* **2015,** *10,* e0134638.

[38] Couve, S. *et al. Cancer Res.* **2014,** *74,* 6554–6564.

[39] Chauvot de Beauchene, I. *et al. PLoS Comput. Biol.* **2014,** *10,* e1003749.

[40] Da Silva Figueiredo Celestino Gomes, P. *et al. PLoS ONE* **2014,** *9,* e97519.

[41] Doss, C. G.; Nagasundaram, N. *PLoS ONE* **2012,** *7,* e31677.

[42] Laine, E.; Chauvot de Beauchene, I.; Perahia, D.; Auclair, C.; Tchertanov, L. *PLoS Comput. Biol.* **2011,** *7,* e1002068. 8

[43] Calhoun, S.; Daggett, V. *Biochemistry* **2011,** *50,* 5345–5353.

[44] Dixit, A.; Verkhivker, G. M. *PLoS Comput. Biol.* **2009,** *5,* e1000487.

[45] Liu, J.; Nussinov, R. *Proc. Natl. Acad. Sci. U.S.A.* **2008,** *105,* 901–906. 8

[46] Shan, Y. *et al. Cell* **2012,** *149,* 860–870. 8

[47] Weber, G. *Biochemistry* **1972,** *11,* 864-878. 8

[48] Karplus, M.; McCammon, J. A. *Annu. Rev. Biochem.* **1983,** *52,* 263–300.

[49] Ichiye, T.; Karplus, M. *Proteins* **1991,** *11,* 205–217.

[50] Tai, K.; Shen, T.; Borjesson, U.; Philippopoulos, M.; McCammon, J. A. *Biophys. J.* **2001,** *81,* 715–724.

[51] Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. *FASEB J.* **2012,** *26,* 868–881.

[52] McClendon, C. L.; Hua, L.; Barreiro, A.; Jacobson, M. P. *J Chem Theory Comput* **2012,** *8,* 2115–2126.

[53] Rod, T. H.; Radkiewicz, J. L.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2003,** *100,* 6980–6985.

[54] Kern, D.; Zuiderweg, E. R. *Current Opinion in Structural Biology* **2003,** *13,* 748 - 757.

[55] del Sol, A.; Fujihashi, H.; Amoros, D.; Nussinov, R. *Mol. Syst. Biol.* **2006,** *2,* 2006.0019.

[56] Gorfe, A. A.; Grant, B. J.; McCammon, J. A. *Structure* **2008,** *16,* 885–896.

[57] Tsai, C.-J.; del Sol, A.; Nussinov, R. *Journal of Molecular Biology* **2008,** *378,* 1 - 11.

[58] Liu, J.; Nussinov, R. *PLoS Comput. Biol.* **2016,** *12,* e1004966. 8

[59] Ota, N.; Agard, D. A. *J. Mol. Biol.* **2005,** *351,* 345–354. 8

[60] Ho, B. K.; Agard, D. A. *Protein Sci.* **2010,** *19,* 398–411. 15

[61] Seeber, M. *et al. J Comput Chem* **2011,** *32,* 1183–1194.

[62] Gerek, Z. N.; Ozkan, S. B. *PLoS Comput. Biol.* **2011,** *7,* e1002154. 15

[63] Bhattacharyya, M.; Bhat, C. R.; Vishveshwara, S. *Protein Sci.* **2013,** *22,* 1399–1416.

[64] Mariani, S.; Dell'Orco, D.; Felline, A.; Raimondi, F.; Fanelli, F. *PLoS Comput. Biol.* **2013,** *9,* e1003207.

[65] Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. *Bioinformatics* **2013,** *29,* 2053-2055.

[66] LeVine, M. V.; Weinstein, H. *PLoS Comput. Biol.* **2014,** *10,* e1003603.

[67] Tiberti, M. *et al. J Chem Inf Model* **2014,** *54,* 1537–1551.

[68] Skjaerven, L.; Yao, X. Q.; Scarabelli, G.; Grant, B. J. *BMC Bioinformatics* **2014,** *15,* 399.

[69] Chakrabarty, B.; Parekh, N. *Nucleic Acids Res.* **2016,** *44,* W375–382. 8

[70] Clarke, D. *et al. Structure* **2016,** *24,* 826–837. 8, 15

[71] Zhang, Z.; Wriggers, W. *J Phys Chem B* **2008,** *112,* 14026–14035.

[72] Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. *Proc. Natl. Acad. Sci. U.S.A.* **2009,** *106,* 6620–6625.

[73] Romanowska, J.; Nowiski, K. S.; Trylska, J. *J Chem Theory Comput* **2012,** *8,* 2588–2599.

[74] McClendon, C.; Kornev, A.; Gilson, M.; Taylor, S. *Proceedings of the National Academy of Sciences* **2014,** *111,* E4623-E4631.

[75] James, K. A.; Verkhivker, G. M. *PLoS ONE* **2014,** *9,* e113488.

[76] Chopra, N. *et al. PLoS Comput. Biol.* **2016,** *12,* e1004826. 8

[77] Singh, S.; Bowman, G. R. *J Chem Theory Comput* **2017,** *13,* 1509–1517. 8, 15

[78] Schueler-Furman, O.; Wodak, S. J. *Curr. Opin. Struct. Biol.* **2016,** *41,* 159–171. 8

[79] Laine, E.; Karami, Y.; Carbone, A. *Mol. Biol. Evol.* **2019,** . 8

[80] Laine, E.; Auclair, C.; Tchertanov, L. *PLoS Computational Biology* **2012,** *8,*. 8

[81] Allain, A. *et al. Faraday Discuss.* **2014,** *169,* 303–321.

[82] Karami, Y.; Laine, E.; Carbone, A. *BMC Bioinformatics* **2016,** *17 Suppl 2,* 13. 8

[83] Karami, Y.; Bitard-Feildel, T.; Laine, E.; Carbone, A. *Scientific Reports* **2018,** *8,* 16126. 8

[84] Engelen, S.; Trojan, L. A.; Sacquin-Mora, S.; Lavery, R.; Carbone, A. *PLoS Comput. Biol.* **2009,** *5,* e1000267. 9, 15, 17, 20

[85] Mihalek, I.; Res, I.; Lichtarge, O. *J. Mol. Biol.* **2004,** *336,* 1265–1282. 9

[86] Lichtarge, O.; Bourne, H. R.; Cohen, F. E. *J. Mol. Biol.* **1996,** *257,* 342–358. 9, 17

[87] Ferreon, A. C.; Ferreon, J. C.; Wright, P. E.; Deniz, A. A. *Nature* **2013,** *498,* 390–394. 12

[88] Choi, J. H.; Laurent, A. H.; Hilser, V. J.; Ostermeier, M. *Nat Commun* **2015,** *6,* 6968. 12

[89] Frappier, V.; Najmanovich, R. J. *PLoS Comput. Biol.* **2014,** *10,* e1003569. 15

[90] Lockless, S.; Ranganathan, R. *Science* **1999,** *286,* 295–299. 15

[91] Baussand, J.; Carbone, A. *PLoS Comput. Biol.* **2009,** *5,* e1000488. 15

[92] Bonetta, L. *Nature* **2010,** *468,* 851–854. 17

[93] Baker, M. *Nature* **2012,** *484,* 271. 17

[94] Huttlin, E. L. *et al. Cell* **2015,** *162,* 425–440. 17

[95] Rolland, T. *et al. Cell* **2014,** *159,* 1212–1226. 17

[96] Robinson, C. V.; Sali, A.; Baumeister, W. *Nature* **2007,** *450,* 973–982. 17

[97] Schweppe, D. K. *et al. Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114,* 1732–1737.

[98] Kalisman, N.; Adams, C. M.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **2012,** *109,* 2884–2889.

[99] Lasker, K. *et al. Molecular & Cellular Proteomics : MCP* **2010,** *9,* 1689–1702.

[100] Karaca, E.; Melquiond, A. S. J.; de Vries, S. J.; Kastritis, P. L.; Bonvin, A. M. J. J. *Molecular & Cellular Proteomics : MCP* **2010,** *9,* 1784–1794. 17

[101] Meyer, M. J. *et al. Nat. Methods* **2018,** *15,* 107–114. 17

[102] Hayashi, T.; Matsuzaki, Y.; Yanagisawa, K.; Ohue, M.; Akiyama, Y. *BMC Bioinformatics* **2018,** *19,* 62.

[103] Szklarczyk, D. *et al. Nucleic Acids Res.* **2017,** *45,* D362-D368.

[104] Garzon, J. I. *et al. Elife* **2016,** *5,*.

[105] Tsuji, T.; Yoda, T.; Shirai, T. *Sci Rep* **2015,** *5,* 16341.

[106] Mosca, R.; Ceol, A.; Aloy, P. *Nat. Methods* **2013,** *10,* 47–53.

[107] Lopes, A. *et al. PLoS Comput. Biol.* **2013,** *9,* e1003369. 17, 18

[108] Tonddast-Navaei, S.; Skolnick, J. *J Chem Phys* **2015,** *143,* 243149. 17, 19

[109] Yang, L. *et al. Cell reports* **2018,** *24,* 585–593. 17

[110] Coesel, S. *et al. EMBO reports* **2009,** *10,* 655–661. 17

[111] Hudson, W. H.; Ortlund, E. A. *Nat. Rev. Mol. Cell Biol.* **2014,** *15,* 749–760. 17

[112] Janin, J. *et al. Proteins* **2003,** *52,* 2–9. 17

[113] Quignot, C. *et al. Nucleic acids research* **2018,** *46,* W408–W416. 17

[114] Van Zundert, G. *et al. Journal of molecular biology* **2016,** *428,* 720–725.

[115] Hopf, T. A. *et al. Elife* **2014,** *3,* e03430. 17

[116] Lensink, M. F. *et al. Proteins: Structure, Function, and Bioinformatics* **2018,** *86,* 257–273. 17

[117] Fleishman, S. J. *et al. J. Mol. Biol.* **2011,** *414,* 289–302. 17

[118] Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. *Protein Sci.* **1997,** *6,* 53–64. 17

[119] Larsen, T. A.; Olson, A. J.; Goodsell, D. S. *Structure* **1998,** *6,* 421–427.

[120] Bogan, A. A.; Thorn, K. S. *J. Mol. Biol.* **1998,** *280,* 1–9.

[121] Jones, S.; Marin, A.; Thornton, J. M. *Protein Eng.* **2000,** *13,* 77–82.

[122] Glaser, F.; Steinberg, D. M.; Vakser, I. A.; Ben-Tal, N. *Proteins* **2001,** *43,* 89–102.

[123] Chakrabarti, P.; Janin, J. *Proteins* **2002,** *47,* 334–343.

[124] Neuvirth, H.; Raz, R.; Schreiber, G. *J. Mol. Biol.* **2004,** *338,* 181–199.

[125] Ofran, Y.; Rost, B. *Bioinformatics* **2007,** *23,* e13–16. 17

[126] Pupko, T.; Bell, R. E.; Mayrose, I.; Glaser, F.; Ben-Tal, N. *Bioinformatics* **2002,** *18 Suppl 1,* S71–77. 17

[127] Glaser, F. *et al. Bioinformatics* **2003,** *19,* 163–164.

[128] Cheng, G.; Qian, B.; Samudrala, R.; Baker, D. *Nucleic Acids Res.* **2005,** *33,* 5861–5867.

[129] Innis, C. A. *Nucleic Acids Res.* **2007,** *35,* W489–494.

[130] Tyagi, M. *et al. PLoS One* **2012,** *7,* e28896.

[131] Xue, L. C.; Dobbs, D.; Honavar, V. *BMC Bioinf.* **2011,** *12,* 244.

[132] Esmaielbeiki, R.; Nebel, J.-C. *BMC Bioinf.* **2014,** *15,* 171. 17

[133] Esmaielbeiki, R.; Krawczyk, K.; Knapp, B.; Nebel, J.-C.; Deane, C. M. *Briefings Bioinf.* **2016,** *17,* 117–131. 17

[134] Aumentado-Armstrong, T. T.; Istrate, B.; Murgita, R. A. *Algorithms Mol Biol* **2015,** *10,* 7. 17

[135] Fernandez-Recio, J.; Totrov, M.; Abagyan, R. *J. Mol. Biol.* **2004,** *335,* 843–865. 18

[136] Russell, R. B. *et al. Curr. Opin. Struct. Biol.* **2004,** *14,* 313–324. 18

[137] Aloy, P.; Russell, R. B. *Nat. Rev. Mol. Cell Biol.* **2006,** *7,* 188–197.

[138] Gray, J. J. *Curr. Opin. Struct. Biol.* **2006,** *16,* 183–193. 18

[139] Sacquin-Mora, S.; Carbone, A.; Lavery, R. *J. Mol. Biol.* **2008,** *382,* 1276–1289. 18

[140] Lopes, A. *et al. PLoS computational biology* **2013,** *9,*. 18, 25

[141] Mintseris, J. *et al. Proteins* **2005,** *60,* 214–216. 18

[142] Lensink, M. F. *et al. Proteins: Structure, Function, and Bioinformatics* **2019,** *87,* 1200–1221. 18

[143] Ripoche, H.; Laine, E.; Ceres, N.; Carbone, A. *Nucleic Acids Res.* **2017,** *45,* 4278.

[144] Laine, E.; Carbone, A. *PLOS Computational Biology* **2015,** *11,* 1-32. 18

[145] Corsi, F.; Lavery, R.; Laine, E.; Carbone, A. *PLOS Computational Biology* **2020,** *16,* e1007624. 18

[146] Dequeker, C.; Laine, E.; Carbone, A. *Proteins: Structure, Function, and Bioinformatics* **2019,** . 18

[147] Dequeker, C.; Laine, E.; Carbone, A. *J Chem Inf Model* **2017,** *57,* 2613–2617. 18

[148] Laine, E.; Carbone, A. *Proteins* **2017,** *85,* 137–154. 18

[149] Laine, E.; Carbone, A. . In *Proceedings of the ICIAP 2013, LNCS Volume 8158*; Springer-Verlag New York, Inc.: 2013. 18

[150] Lensink, M. F.; Wodak, S. J. *Proteins* **2009,** . 18

[151] Cazals, F.; Proust, F.; Bahadur, R. P.; Janin, J. *Protein Science* **2006,** *15,* 2082–2092. 18

[152] Ma, B.; Elkayam, T.; Wolfson, H.; Nussinov, R. *PNAS* **2003,** *100,* 5772–5777. 19

[153] Berman, H. M. *et al. Nucleic Acids Res.* **2000,** *28,* 235–242. 19

[154] Levy, E. D. *J. Mol. Biol.* **2010,** *403,* 660–670. 19

[155] Zacharias, M. *Protein Sci.* **2003,** *12,* 1271–1282. 20

[156] Caffrey, D. R.; Somaroo, S.; Hughes, J. D.; Mintseris, J.; Huang, E. S. *Protein Sci.* **2004,** *13,* 190–202. 20

[157] Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. *Proteins* **2010,** *78,* 3111–3114. 20, 25

[158] Segura, J.; Jones, P. F.; Fernandez-Fuentes, N. *BMC Bioinformatics* **2011,** *12,* 352. 21

[159] Zhang, J.; Maslov, S.; Shakhnovich, E. I. *Mol. Syst. Biol.* **2008,** *4,* 210. 25

[160] Zhang, Q. C. *et al. Nucleic Acids Res.* **2011,** *39,* W283–287.

[161] Maheshwari, S.; Brylinski, M. *Journal of Molecular Recognition* **2015,** *28,* 35–48. 21

[162] Luscombe, N. M.; Austin, S. E.; Berman, H. M.; Thornton, J. M. *Genome Biology* **2000,** *1,* reviews001–1. 22

[163] Yan, J.; Kurgan, L. *Nucleic acids research* **2017,** *45,* e84–e84. 24

[164] Tjong, H.; Zhou, H.-X. *Nucleic Acids Research* **2007,** *35,* 1465–1477.

[165] Segura, J.; Jones, P. F.; Fernandez-Fuentes, N. *Bioinformatics* **2012,** *28,* 1845–1850. 24

[166] Ritchie, D. W.; Kemp, G. J. *Proteins* **2000,** *39,* 178–194. 25

[167] Pierce, B. G.; Hourai, Y.; Weng, Z. *PLoS ONE* **2011,** *6,* e24657. 25

[168] Deeds, E. J.; Ashenberg, O.; Gerardin, J.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2007,** *104,* 14952–14957. 25

[169] Heo, M.; Maslov, S.; Shakhnovich, E. *Proc. Natl. Acad. Sci. U.S.A.* **2011,** *108,* 4258–4263. 25

[170] Deeds, E. J.; Ashenberg, O.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2006,** *103,* 311–316. 25

[171] Graveley, B. R. *Trends Genet* **2001,** *17,* 100–107. 27

[172] Wang, E. T. *et al. Nature* **2008,** *456,* 470–476. 27

[173] Yang, X. *et al. Cell* **2016,** *164,* 805–817. 27

[174] Birzele, F.; Csaba, G.; Zimmer, R. *Nucleic Acids Res.* **2008,** *36,* 550–558. 27

[175] Ward, A. J.; Cooper, T. A. *J. Pathol.* **2010,** *220,* 152–163. 27

[176] Lim, K. H.; Ferraris, L.; Filloux, M. E.; Raphael, B. J.; Fairbrother, W. G. *Proc. Natl. Acad. Sci. U.S.A.* **2011,** *108,* 11093–11098. 27

[177] Akerman, M.; Mandel-Gutfreund, Y. *Nucleic Acids Res.* **2007,** *35,* 5487–5498. 27

[178] Lev-Maor, G. *et al. PLoS Genet.* **2007,** *3,* e203.

[179] Keren, H.; Lev-Maor, G.; Ast, G. *Nat. Rev. Genet.* **2010,** *11,* 345–355. 27

[180] Sultan, M. *et al. Science* **2008,** *321,* 956–960. 27

[181] Kelemen, O. *et al. Gene* **2013,** *514,* 1–30. 27, 29

[182] Gonzalez-Porta, M.; Frankish, A.; Rung, J.; Harrow, J.; Brazma, A. *Genome Biol.* **2013,** *14,* R70. 27

[183] Weatheritt, R. J.; Sterne-Weiler, T.; Blencowe, B. J. *Nat. Struct. Mol. Biol.* **2016,** *23,* 1117–1123. 27

[184] Buljan, M. *et al. Mol. Cell* **2012,** *46,* 871–883. 27, 28

[185] Ellis, J. D. *et al. Mol. Cell* **2012,** *46,* 884–892. 27

[186] Yura, K. *et al. Gene* **2006,** *380,* 63–71. 27

[187] Rodriguez, J. M. *et al. Nucleic Acids Res.* **2013,** *41,* D110–117.

[188] Tranchevent, L. C. *et al. Genome Res.* **2017,** *27,* 1087–1097.

[189] Tapial, J. *et al. Genome research* **2017,** *27,* 1759–1768. 27

[190] Christinat, Y.; Moret, B. M. *BMC Bioinformatics* **2012,** *13 Suppl 9,* S1. 27, 29

[191] Christinat, Y.; Moret, B. M. *IEEE/ACM Trans Comput Biol Bioinform* **2013,** *10,* 1403–1411. 27

[192] Sterne-Weiler, T.; Weatheritt, R. J.; Best, A. J.; Ha, K. C.; Blencowe, B. J. *Molecular cell* **2018,** *72,* 187–200. 27

[193] Trachana, K. *et al. Bioessays* **2011,** *33,* 769–780. 27

[194] Modrek, B.; Lee, C. J. *Nature genetics* **2003,** *34,* 177–180. 27

[195] Xing, Y.; Lee, C. *Bioinformatics* **2005,** *21,* 3701–3703. 27

[196] Hegyi, H.; Kalmar, L.; Horvath, T.; Tompa, P. *Nucleic Acids Res.* **2011,** *39,* 1208–1219. 27

[197] Gelly, J. C.; Lin, H. Y.; de Brevern, A. G.; Chuang, T. J.; Chen, F. C. *Genome Biol Evol* **2012,** *4,* 966–975. 27

[198] Birzele, F. *et al. Nucleic Acids Res.* **2008,** *36,* D63–68. 28

[199] Prakash, A.; Bateman, A. *Genome Biol.* **2015,** *16,* 88. 28

[200] Zea, D. J.; Richard, H.; Laine, E. *in preparation* **2020,** . 28

[201] Ait-hamlat, A. *et al. Journal of Molecular Biology* **2020,** . 28

[202] Yates, A. *et al. Nucleic acids research* **2016,** *44,* D710–D716. 28

[203] Hildebrand, A.; Remmert, M.; Biegert, A.; Soding, J. *Proteins* **2009,** *77 Suppl 9,* 128–132. 29

[204] Bhuiyan, S. A. *et al. BMC Genomics* **2018,** *19,* 637. 29

[205] Stamm, S. *et al. Gene* **2005,** *344,* 1–20. 29

[206] Waetzig, V.; Herdegen, T. *Trends Pharmacol. Sci.* **2005,** *26,* 455–461. 29

[207] Bogoyevitch, M. A.; Kobe, B. *Microbiol. Mol. Biol. Rev.* **2006,** *70,* 1061–1095. 29

[208] Raucci, R.; Laine, E.; Carbone, A. *Structure* **2018,** *26,* 905–915. 33

[209] Geng, C.; Vangone, A.; Folkers, G. E.; Xue, L. C.; Bonvin, A. M. *Proteins: Structure, Function, and Bioinformatics* **2019,** *87,* 110–119. 33

[210] Katz, Y.; Wang, E. T.; Airoldi, E. M.; Burge, C. B. *Nature methods* **2010,** *7,* 1009. 33

[211] De La Grange, P.; Dutertre, M.; Martin, N.; Auboeuf, D. *Nucleic acids research* **2005,** *33,* 4276–4284. 34

[212] Bastian, F. *et al.* . In *International Workshop on Data Integration in the Life Sciences*; 2008. 34

[213] El-Gebali, S. *et al. Nucleic acids research* **2019,** *47,* D427–D432. 34

[214] Estaña, A. *et al. Structure* **2019,** *27,* 381–391. 34

[215] Greener, J. G.; Moffat, L.; Jones, D. T. *Scientific reports* **2018,** *8,* 1–12. 34

[216] Riesselman, A. J. *et al. bioRxiv* **2019,** 757252. 34

[217] Ingraham, J.; Garg, V.; Barzilay, R.; Jaakkola, T. . In *Advances in Neural Information Processing Systems*; 2019. 34

[218] Pagès, G.; Charmettant, B.; Grudinin, S. *Bioinformatics* **2019,** *35,* 3313–3319. 34

[219] Derevyanko, G.; Grudinin, S.; Bengio, Y.; Lamoureux, G. *Bioinformatics* **2018,** *34,* 4046–4053. 34