Understanding and improving statistical models of protein sequences

Pierre Barrat-Charlaix

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Informatique

École doctorale n°130: Informatique, Télécommunications et électronique (Paris)

réalisée

au Laboratoire de biologie computationnelle et quantitative

sous la direction de Martin WEIGT

présentée par

Pierre BARRAT-CHARLAIX

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Comprendre et améliorer les modèles statistiques de séquences de protéines

soutenue le 9 novembre 2018

devant le jury composé de :

M.	Alessandro LAIO	Rapporteur
M.	Clément NIZAK	Rapporteur
M.	Guillaume ACHAZ	Examinateur
M^{me}	Aleksandra WALCZAK	Examinatrice
M.	Martin WEIGT	Directeur de thèse

Pierre Barrat-Charlaix: *Understanding and improving statistical models of protein sequences,* November 2018

Tchouang-tseu et Houei-tsu se promenaient sur le pont enjambant la rivière Hao. Tchouang-tseu dit :

– Regarde les vairons, qui nagent et bondissent tout leur soûl. C'est ça qui rend les poissons heureux.

Houei-tseu dit :

–Tu n'es pas un poisson, alors comment sais-tu ce qui rend les poissons heureux ?

Tchouang-tseu dit :

– Tu n'es pas moi, alors comment sais-tu que je ne sais pas ce qui rend les poissons heureux ?

Houei tseu dit :

– C'est vrai, je ne suis pas toi, je n'ai donc, c'est certain, aucune idée de ce que tu sais. D'un autre côté, tu n'es pas un poisson, c'est certain, et cela prouve simplement que tu ne peux pas savoir ce qui rend les poissons heureux.

Tchouang-tseu dit :

– Revenons à ta question initiale. Tu as dit : "Comment sais-tu ce qui rend les poissons heureux ?" Donc lorsque tu as posé cette question, tu savais que je le savais. Je le sais parce que je suis ici, au-desuss de la rivière Hao.

Tchouang-Tseu, *17*

ABSTRACT

In the last decades, progress in experimental techniques have given rise to a vast increase in the number of known DNA and protein sequences. This has prompted the development of various statistical methods in order to make sense of this massive amount of data. Among those are pairwise co-evolutionary methods, using ideas coming from statistical physics to construct a global model for protein sequence variability. These methods have proven to be very effective at extracting relevant information from sequences only, such as structural contacts or effects of mutations. While co-evolutionary models are for the moment used as predictive tools, their success calls for a better understanding of they functioning. In this thesis, we propose developments on existing methods while also asking the question of how and why they work. We first focus on the ability of the so-called Direct Coupling Analysis (DCA) to reproduce statistical patterns found in sequences in a protein family. We then discuss the possibility to include other types of information such as mutational effects in this method, followed by potential corrections for the phylogenetic biases present in available data. Finally, considerations about limitations of current co-evolutionary models are presented, along with suggestions on how to overcome them.

RÉSUMÉ

Dans les dernières décennies, les progrès des techniques expérimentales ont permis une augmentation considérable du nombre de séquences d'ADN et de protéines connues. Cela a incité au développement de méthodes statistiques variées visant à tirer parti de cette quantité massive de données. Les méthodes dites co-évolutives en font partie, utilisant des idées de physique statistique pour construire un modèle global de la variabilité des séquences de protéines. Ces méthodes se sont montrées très efficaces pour extraire des informations pertinentes des seules séquences, comme des contacts structurels ou les effets mutationnels. Alors que les modèles co-évolutifs sont pour l'instant utilisés comme outils prédictifs, leur succès plaide pour une meilleure compréhension de leur fonctionnement. Dans cette thèse, nous proposons des élaborations sur les méthodes déjà existantes tout en questionnant leur fonctionnement. Nous étudions premièrement sur la capacité de l'Analyse en Couplages Directs (DCA) à reproduire les motifs statistiques rencontrés dans les séquences des familles de protéines. La possibilité d'inclure d'autres types d'information comme des effets mutationnels dans cette méthode est présentée, suivie

de corrections potentielles des biais phylogénétiques présents dans les données utilisées. Finalement, des considérations sur les limites des modèles co-évolutifs actuels sont développées, de même que des suggestions pour les surmonter.

À vous, troupe légère,
Qui d'aile passagère
Par le monde volez

Joachim du Bellay

REMERCIEMENTS

Il me faut d'abord remercier Martin d'avoir dirigé ma thèse. Je pense que nous nous sommes bien entendus pendant ces trois années, et j'ai le sentiment d'avoir beaucoup appris. Ce manuscrit, j'espère, en est la preuve.
Merci ensuite à Guillaume Achaz, Alessandro Laio, Clément Nizak et Aleksandra Walczak, d'abord pour accepter de faire partie du jury de cette thèse en tant qu'examinateurs ou que rapporteurs, et ensuite pour la patience dont ils devront faire preuve afin de lire ce long manuscrit.
Puis viennent les membres de l'équipe. D'abord ceux qui ne sont plus au laboratoire au moment où j'écris : merci à Eleonora, à Alice, à Guido, à Christoph. Et surtout merci à Matteo, avec qui j'ai eu la chance de travailler. Et puis ceux qui sont toujours là : Giancarlo, Nika, Kai, Edwin, Anna, Edoardo, Maureen, Carlos. D'ailleurs, pourquoi ne pas les remercier pendant la prochaine pause café en plus d'ici ? Enfin, merci aux autres membres du laboratoire. Ce sont toutes ces personnes qui créent l'environnement scientifique stimulant nécessaire à la recherche.
Merci également aux Cubains de m'avoir accueilli quelques mois. Merci à Alejandro, de m'avoir fourni un toit et un lit à la Havane, sinon une chaise. Et puis à Amanda, Edwin, JJ, Carlos, et d'autres, de m'avoir fait découvrir José Marti et son pays.

Une pensée maintenant pour mes amis. Je les ai rencontrés pendant ma thèse, mon master, à l'ENSTA, en classe préparatoire, ou avant. D'autres, je les connais depuis toujours. Chacun à leur manière, ils contribuent au goût qu'a ma vie. Je mentionne évidemment Mourtaza, lui aussi englouti par sa thèse, et le remercie d'être une source constante d'entropie.
Je veux aussi écrire un mot pour ma famille. Aux cousins et cousines de Paris, de Lyon, du Champsaur, et d'ailleurs, que j'aime à revoir le plus souvent possible. À mes grands parents, Papi, Mami, Bernard, Danielle, Blanche – on peut admirer quelqu'un sans l'avoir connu.

Merci à mes parents de m'avoir guidé.
Et merci à Sophie de partager ma vie.

CONTENTS

ACRONYMS

MSA Multiple Sequence Alignment

HMM Hidden Markov Models

PDB Protein DataBank

DCA Direct Coupling Analysis

MaxEnt Maximum-Entropy Principle

KL-distance Kullback-Leibler distance

i.i.d. independent and identically distributed

MF Mean-Field

PLM Pseudo-Likelihood Maximization

BML Boltzmann Machine Learning

PSICOV Pseudo-Sparse Inverse Covariance

MCMC Markov Chain MonteCarlo

APC Average Product Correction

ACE Adaptive Cluster Expansion

MIC Minimum Inhibitory Concentration

RBM Restricted Boltzmann Machine

DI Direct Information

PI Path Information

MI Mutual Information

SCA Statistical Coupling Analysis

1 INTRODUCTION

1.1 A WORD ABOUT PROTEIN SEQUENCES

Protein are molecules essential to almost all cellular processes. Sophisticated experimental techniques developed in the last decades have given rise to a vast increase of amino acid sequence data. Thanks to next generation sequencing, databases have been subject to an exponential growth in their number of entries: as of 2018, the UniProt database [78] now contains more than 100 million protein sequences. However, most of those proteins have not been experimentally studied, and little is known about their biological function properties. Only a small fraction of UniProt sequences are manually annotated, 0.5% in the SwissProt database, meaning that some of their biological features has been studied either experimentally or through curator-evaluated computational analysis.

The sole knowledge of the amino acid sequence does not allow for a clear understanding of the function of the corresponding protein. From a molecular point of view, proteins are characterized by the complex three-dimensional structure resulting from the folding of the amino acid chain. This structure is a fundamental determinant of the molecular function. Through the exposure of certain active sites, the specific and exclusive binding to some molecules, or through reactive conformational changes, it allows proteins to perform a vast array of essential cellular functions, ranging from catalytic activity to gene regulation or signaling. Proteins do not operate alone in those activities, but combine with others to form a cellular pathway, a self-regulated chain of chemical reactions aiming at a precise function for the cell. The knowledge of the interaction network of proteins in the cell is therefore crucial to grasp cellular activity. From the evolutionary point of view, it is of great interest to understand the effect of mutations in DNA and in the proteins. The rapid adaptation of viruses to immune systems, the alarming rise in antibiotic resistant bacterial strains or the development of cancerous cells are all due to mutational events in the involved organisms. Is it possible, from the knowledge of protein sequences, to deduce the effect of possible mutations on the phenotype of the corresponding organism?

The extreme increase in "raw" sequence knowledge could prove very important in answering those questions. However, without proper theoretical methods, the sequences alone are not of much help. The development of such methods is usually called computational biology and has been the subject of intensive research in recent years. In this

respect, one of the most impressive achievement of computational methods has been the classification of protein sequences in families of homologous domains: frequently observed sequence modules which share a common ancestor. An fundamental property of protein families is the conservation of structure and function across its members. Essentially, a family groups different amino acid sequences which encode for biologically similar molecules. This feature calls for statistical methods to model sequence variability inside the family. A central focus of the present work is to study the ability of statistical physics inspired techniques to model variability in protein sequences, and to extract relevant biological informations from it.

1.2 PROTEIN FAMILIES

1.2.1 *Proteins*

Proteins are the main constituent of the cell, accounting for most of its dry mass. Chemically, they are polymers made from a succession of amino acids linked by peptide bonds. Amino acids consist of core atoms with carboxyl (COOH) and amino (NH_2) groups on its ends, and of a side chain. The bond between the carboxyl and amino groups of two amino acid is called the peptide bond, at the basis of the polypeptide backbone forming the protein. There are 20 different side chains, giving rise to the 20 amino acids. Their unique chemical properties – charge, size, hydrophobicity, ... – allow for the complex structures and functions of proteins.
Each protein is chemically defined by its sequence of amino acids. Formally, it can be represented as a string of letters, where each letter stands for one of the 20 amino acids. The protein is usually characterized by four levels of organization:

- Primary structure: the linear sequence of amino acids.
- Secondary structure: local arrangement of the amino acids, consisting of the α helices and the β sheets along with less structured loops.
- Tertiary structure: three dimensional shape of an entire chain, where the elements of the secondary structure fold in a compact structure.
- Quaternary structure: Combination of different polypeptide chains, forming complexes consisting of multiple sub-units.

In parallel to those four levels, an important element of organization is the protein domain. The domain is a polypeptide chain, usually of the order of 100 amino acids, that can *independently* fold into a stable structure. Domains can be seen as modular units, the combination

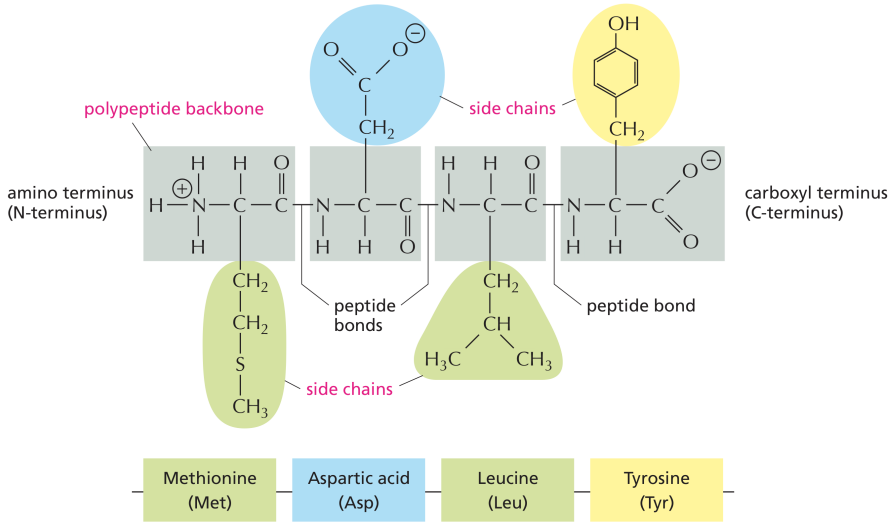

Figure 1.1: The polypeptide chain forming the protein consists of a backbone linking amino and carboxyl groups of amino acids, and of a variety of side chains giving the protein its chemical properties (Source [2]).

of which builds larger proteins. They can usually be associated to particular functions, such as binding to a specific molecule or to a specific DNA fragment. Proteins usually consists of one to dozens of domains, resulting in a broad distribution of their sizes.

1.2.2 *Protein families*

Known sequences can be organized into families. In the course of evolution, protein or domains having similar biological functions but present in different organisms have accumulated mutations, and now present a high variety in amino acid sequences. However, they usually share very similar three dimensional structures and biological activity [2]. These proteins or domains are said to share a common ancestor, and constitute the members of a family. The Pfam database lists 16712 domain families (as of version 31.0 released in March 2017 [63]), built from 26.7 million sequences in the Uniprot reference proteome [78]. Homologous sequences are similar in some aspects, showing portions of very conserved residues. However, accumulated mutations during millions of years of evolution lead to a high diversity, with an average sequence similarity as low as 30% between members of the family. Moreover, because of deletions and insertions, sequences have varied in length and are not directly comparable. For this reason, a Multiple Sequence Alignment (MSA) is built to represent the family: sequences are organized in an array where each line represents one protein domain, and residues conserved across the family are placed in the same columns. Mismatches in length of sequences due to deletions or

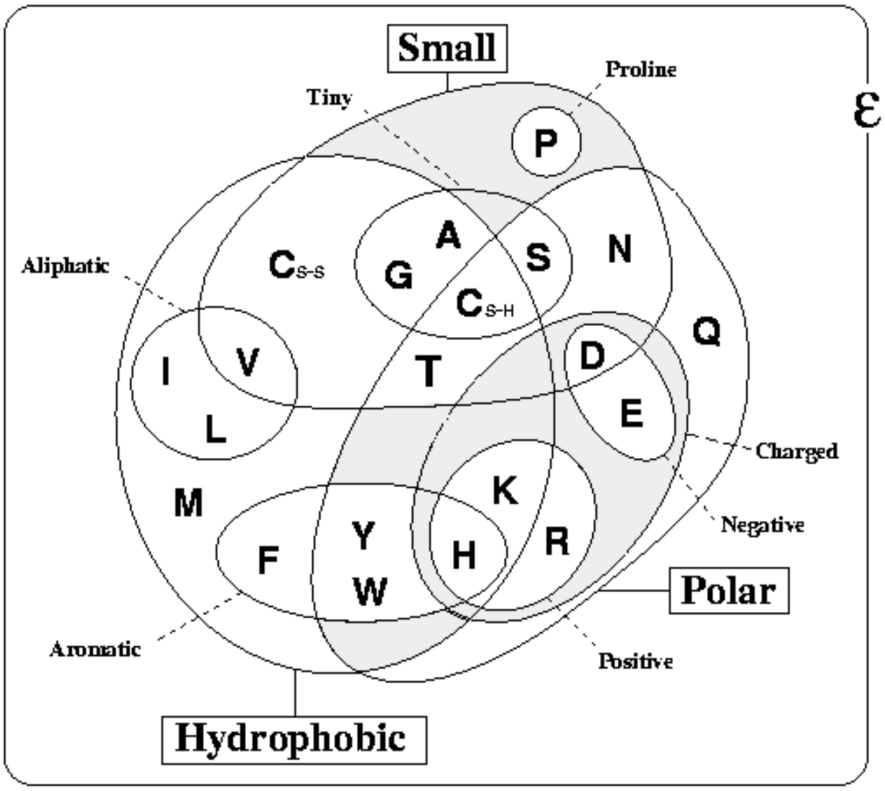

Figure 1.2: Different physico-chemical properties of the 20 amino acids. (Source [51]).

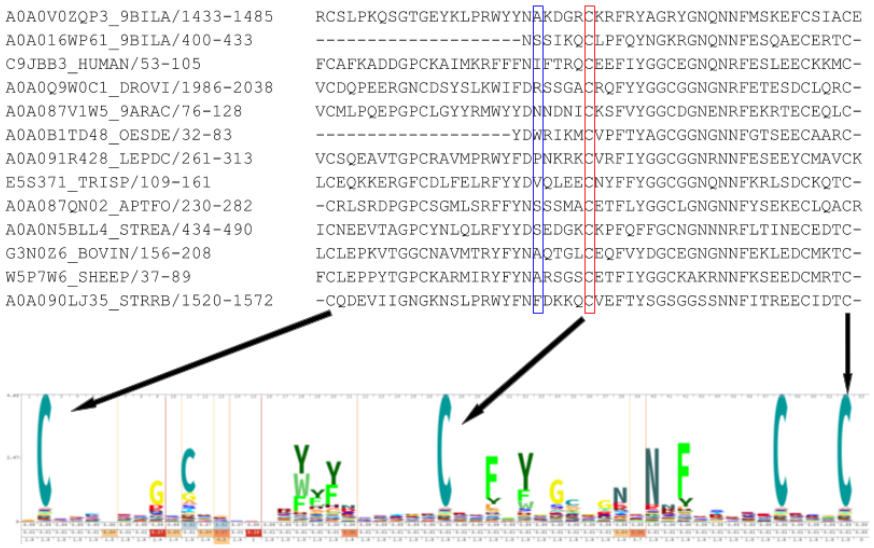

Figure 1.3: Part of the MSA of the PF00014 Pfam family (Trypsin inhibitor), with its corresponding HMM logo [72]. The red-boxed column shows high conservation of a cysteine, while the blue-boxed column seems to be very variable. This is well reflected in the logo, looking at the second "large" C. In this example, columns corresponding to insertions have been removed. PF00014 contains $M = 11819$ sequences of length $L = 53$.

insertions are compensated by the addition of gaps, represented by the symbol "-". An example of such an alignment is shown in figure 1.3.

Formally, an alignment can be represented as an array $\{a_i^m\}$, where $i \in \{1 \dots L\}$ is an index running over the sequences' length, and $m \in \{1 \dots M\}$ is the sequence or line number. Each a_i^m is a number between 1 and $q = 21$, standing for one of the 20 amino acids or for the gap symbol. Typical alignments contain in the range of $M = 10^2 - 10^5$ sequences, with very variable length ranging from 20 residues to several hundreds. Depending on the family, the structures of many or no members are known. For the case of PF00014, 259 structures are present in the Protein DataBank (PDB). However, those are highly redundant, often characterizing the exact same domain. Two of those structures are visible in figure [REF - MAKE FIGURE].
Due to the conserved function and structure of the members of the family, mutations do not happen at random: some positions in the sequences allow for some mutations without disrupting the function, while others cannot mutate and are observed to be completely conserved. As a result, MSAs present strong statistical patterns visible in the example of figure 1.3. These patterns are essential in many aspects: they enable us to find new family members and are an indication of the evolutionary constraints acting on the family. The HMM-logo

represented in figure 1.3 is a simple way to visualize conservation of columns in the MSA [72]. However, quantifying those requires adapted theoretical tools as we will see in the next section.

1.2.3 *Profile hidden Markov models*

Protein families are intricately related to one of the most powerful tools in bioinformatics, the profile Hidden Markov Models (HMM). Profile HMM's (abbreviated as HMM in the following) are a statistical representation of an MSA used to find new members of a family, or to find a family corresponding to a given sequence [25, 26]. In fact, the families in the Pfam database are constructed using the HMMer software [34, 35].
The idea behind HMM's is to model statistical variability in an alignment based on the frequency at which amino acids and gaps are found in each of its columns. Formally, it consists of a Markov chain based on a directed graph (see figure 1.4 for a representation). This graph contains three types of nodes or states:

- Match states: they model the frequency at which amino acids are found in a non-insert column of an alignment.
- Insertion states: they account for potential insertions of amino acids in some sequences of the family, resulting in gaps for the others in the MSA.
- Deletion states: they account for the deletion of amino acids in some proteins, generating a gap in the corresponding aligned sequence.

To each match and insertion state corresponds an amino acid distribution. In the case of insertion states, the background frequency of each amino acid is used. In the case of match states, the distribution of residues in the corresponding column of the MSA is used. When the Markov chain reaches one of these states, it emits a residue in accordance with the corresponding distribution. When it reaches a deletion state, a gap symbol is emitted.
Transition probabilities exist between two "layers" of the graph only in one direction, from left to right. Therefore, the trajectory of the Markov chain in this graph (starting at the "Begin" state and ending at the "End") results in one aligned sequence.
Parameters of the HMM, that is emission probabilities for match states and transition probabilities between different states, must be learned on an initial alignment, called the seed. This learning process mainly depends on the conservation profile of the seed alignment. The emission distribution of each state should match frequencies at which amino acids are found in different columns of the seed alignment. Once the HMM is trained, it can be used to find new members of the

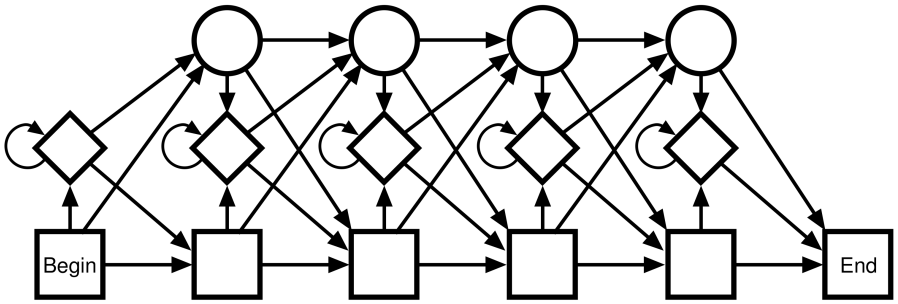

Figure 1.4: Structure of a Hidden Markov Models (HMM). Squares represent match states. A distinct emission probability is associated to each of them, corresponding to the conservation pattern in a column of the MSA. Diamonds represent insert states, with potential transition to the same insert state represented. Circles are deletion states. Source [50].

seed family: sequences that would be obtained with a high probability – compared to a null model – by the Markov chain. The found sequence is aligned to the seed by computing the most likely corresponding path from "Begin" to "End", using the Viterbi algorithm [25]
In the case of the Pfam families, manually curated alignment of about 100 sequences are used as seeds, and profile-HMM are trained. They are then used to scan sequence databases, such as Uniprot, to find homologous sequences. In this way, large alignments of thousands of homologous sequences can be constructed.

1.2.4 *Co-evolution in protein families*

By construction, profile-HMM's are based mostly on the profile of the seed alignment, that is on the frequency at which amino acids are found in each of its columns. The only allowed "interaction" between columns is due to potentially different transition probabilities from match to deletion states, from deletion to deletion states and from match to match states. According to the HMM, the probability of finding a gap or an amino acid at a given position of a sequence only depends on the state of the previous position. This means that this model is able to represent short range "gap-gap" or "gap-amino acid" correlations. Nonetheless, the distribution of amino acids at a position is completely independent from that of other positions.
However, correlation in the usage of amino acids at at different columns of an MSA is observed for all known families. This phenomenon, called co-evolution, is an indicator of epistasis: mutations at different positions in a sequence can not always be considered to have independent or additive effects. On the contrary, the effect of a mutation may depend on the rest of the sequence, thus having

different consequences for different proteins.
One of the possible explanation for epistasis is the presence of structural constraints due to the three-dimensional fold of the protein. If two amino acids are in contact in this fold, a mutation at one of the corresponding positions in the sequence may have to be compensated by a mutation at the other position. In this scenario, two columns of the MSA will show a correlation in their usage of amino acids. This idea has been directly confirmed recently in [58, 67], where it is shown that pairs of sequence positions experiencing strong epistatic effects are close by in the protein fold, and that the structure may potentially be reconstructed from them.
The idea of using co-evolutionary signal to predict structural contacts in a protein has been present for a long time [24, 37, 59]. However, it faces one important limitation. Indeed, apparent correlation between two columns of an MSA can be due to a direct interaction of the two corresponding residues, such as a contact, but also to indirect effects [12]. Imagine residues A and B distant in the structure, but both in contact with C: this may lead to an indirect co-evolution signal between A and B. For this reason, methods based solely on the direct measurement of the correlation between columns of the MSA fail to accurately predict structural contacts.
An important task is thus to disentangle direct and indirect sources of correlations. This cannot be achieved by looking at pairs of columns independently, but calls for a *global* sequence model as will be seen in chapter 2.

Although HMM's can be considered as global sequence models, as they assign a probability to any given sequence of amino acids, they are fundamentally unable to take correlations into account. Even though the profile-HMM identifies members of a family among known natural protein, it cannot be considered as a good statistical model for sequences. Indeed, any artificial sequence that respects the column-wise conservation profile of the family will be considered as a potential member by the HMM, even though it does not respect the correlation patterns between columns and is therefore very unlikely to represent a functional protein. This will be shown in more details in section 2.5.2.
The current state of biology and bioinformatics thus calls for more sophisticated sequence models. The Direct Coupling Analysis (DCA), introduced in 2009 [82] and based on statistical physics ideas, is such a model. Its description is the subject of the following chapter.

2

THE POTTS MODEL FOR PROTEIN SEQUENCES

2.1 MOTIVATION: GLOBAL STATISTICAL MODELS FOR PROTEIN SEQUENCES

If members of a protein family share a common three-dimensional structure and biological function, it is natural to assume that they also share similar evolutionary pressure. The constraints that natural selection imposes on their sequences should be similar. If this is the case, quantities such as the probability for a given sequence to be a functional member of the family or the effect of mutations on members of the family could be described by one single model, representing the evolutionary constraints acting on this family.
The multiple sequence alignment represents essential information for identifying those constraints. Indeed, statistical patterns present in the MSA are a direct indication that mutations are not randomly selected. Almost all families display very conserved columns in their MSA, indicating a residue that cannot be mutated without a major detrimental effect for functionality. Pairs of columns can also display correlation patterns, meaning that pairs of amino-acids appear with a frequency different of what would be expected based on the conservation in their respective columns. This could for instance be an indicator of compensatory mutations or residue-level co-evolution.
The aim of the DCA is to construct a probabilistic model using such statistical features of the MSA in order to have a quantitative description of evolutionary pressure on the sequences. DCA assigns probability score $P(\underline{A}|\boldsymbol{J},\boldsymbol{h})$ to every sequence of amino-acids or gaps $\underline{A}$ of the length of the considered MSA. The specific functional form of P is given by a class of probabilistic distributions named Potts models (see Eq. 2.1), and $\boldsymbol{J}$ and $\boldsymbol{h}$ are sets of parameters defining the model, referred to as couplings and fields.
Models of this form, originally coming from statistical physics, have been successfully used in different biological contexts, ranging from the description of patterns of neuron firing [31, 65, 70], the prediction of contacts in protein structures [57, 62, 82], or the movement of flocks of birds [14, 15].
In the case of protein sequences, one of the main characteristic of DCA is that it relies on a *global* model. Indeed, the score described in Eq. 2.1 depends jointly on the full sequence, and cannot be factorized over columns of the MSA. This is a crucial difference with modeling techniques such as the HMM. This choice is biologically well motivated. Contacts between residues in the protein fold impose constraints on

the corresponding pair of columns in the MSA, making mutations at those positions possibly correlated. Moreover, there is ample evidence of epistatis in proteins, meaning that the effect of a mutation depends not only on the local change in amino-acid, but also on the full background sequence [11, 42, 58, 64]. As a consequence, global models seem to be necessary to correctly represent relevant statistical features found in MSAs.
The following sections will address the choice of Potts models as a functional form for DCA models, the inference of the essential parameters $\boldsymbol{J}$ and $\boldsymbol{h}$, and results obtained on proteins using those methods.

2.2 MAXIMUM-ENTROPY MODELING

2.2.1 *The Potts model*

In this work, we will call Potts models probability distributions that have the functional form given by Eq. 2.1. They are an extension of the Ising model, extensively used in statistical physics, to spins or discrete variables that take q states. In the Ising case we have $q = 2$, whereas in the case of an MSA we have $q = 21$ to account for the 20 amino-acids found in proteins and the gap symbol. Every configuration of the L q-state variables $\underline{A} = (a_1 \dots a_L)$ is assigned a probability

$$P(a_1 \dots a_L | \boldsymbol{J}, \boldsymbol{h}) = \frac{1}{Z(\boldsymbol{J}, \boldsymbol{h})} \exp\big(-\mathcal{H}(\underline{A})\big), \tag{2.1}$$

where the so-called Hamiltonian $\mathcal{H}$ is defined by

$$\mathcal{H}(\underline{A}) = -\sum_{i=1}^{L} h_i(a_i) - \sum_{1 \le i < j \le L} J_{ij}(a_i, a_j). \tag{2.2}$$

Field parameters $\boldsymbol{h}$ are local, acting on one variable at a time. On the other hand, couplings $\boldsymbol{J}$ reflect the symmetric interaction between two distinct variables, with $J_{ij}(a, b) = J_{ji}(b, a)$ and $J_{ii}(a, b) = 0 \quad \forall i, j, a, b$. The partition function Z is a normalization constant, defined by

$$Z = \sum_{\{\underline{A}\}} \exp\big(-\mathcal{H}(\underline{A})\big), \tag{2.3}$$

where the sum runs over all possible configurations of variables in $\underline{A}$. Another convenient representation of the Potts model consists of writing the variables a_i in the form of a q-dimensional vector σ_i indexed by α:

$$\sigma_{i,\alpha} = \begin{cases} 1 & \text{if } \; a_i = \alpha \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

The full configuration $\underline{A}$ is then written as an $L \times q$ vector of 0 or 1's $\underline{\sigma}$. This allows rewriting Eq. 2.2 in a matrix form

$$\mathcal{H}(\underline{A}) = -\boldsymbol{h}\underline{\sigma} - \frac{1}{2}\,\underline{\sigma}\boldsymbol{J}\underline{\sigma}. \tag{2.5}$$

In this formulation, $\boldsymbol{h} = \{h_i(a)\}, i = 1 \dots L, a = 1 \dots q$ is an $L \times q$ vector, where as $\boldsymbol{J}$ is an $Lq \times Lq$ matrix built of $L(L-1)/2$ blocks J_{ij}, each of size $q \times q$.

Two reasons are commonly given in the literature to justify the use of Potts-like models for protein sequences. The first is that the correlation observed between two columns in an MSA may not be a good indicator of a direct functional or structural interaction between the two corresponding residues. In the case of structure for instance, correlation could be caused either by a contact between two residues, making mutations at the two positions inter-dependent, but could also arise between distant residues because of indirect interactions.
The idealized scenario would be the one where residues A and B are distant in the structure, but close to a third residue C. Possible correlation between A and B is then the result of indirect interactions.
The structure of the Potts model is such that the full probability distribution P, and thus the corresponding correlations, is described by means of the direct couplings $\boldsymbol{J}$. In the case of protein sequences, the hope is that those couplings reflect interactions which are biologically interpretable, such as structural proximity of the corresponding residues.
The other motivation for using such a class of models is given by the Maximum-Entropy Principle (MaxEnt).

2.2.2 *The maximum-entropy principle*

The maximum-entropy principle, first introduced by Jaynes [46, 47] can be seen as a principled way to obtain functional forms of probability distributions in inference problems. Given some data $\boldsymbol{X}$ consisting of M configurations of L spins, one identifies a set of *relevant observables* $\mathcal{O}^p(\boldsymbol{X}) \equiv \langle \mathcal{O}^p(x) \rangle_{\boldsymbol{X}}$ of the data, where p stands as a label for different observables. MaxEnt states that in order to describe $\boldsymbol{X}$, one has to find the distribution P that reproduces the chosen observables of the data, but is as general as possible in other regards. Quantitatively, P should have the maximum possible Shanon entropy

$$S = -\sum_x P(x) \log P(x), \tag{2.6}$$

while reproducing the means of the observables over the data.

$$\langle \mathcal{O}^p(x) \rangle_P = \langle \mathcal{O}^p(x) \rangle_{\boldsymbol{X}}. \tag{2.7}$$

This optimization problem in P can be solved using Lagrange multipliers, yielding the following parametrization

$$P(x|\{\lambda^p\}) = \frac{1}{Z(\{\lambda^p\})} \exp\left(\sum_p \lambda^p \mathcal{O}^p(x)\right), \tag{2.8}$$

where $\{\lambda^p\}$ is the set of Lagrange multipliers corresponding to each constraint.
In the case of protein sequences, the chosen observables are the single and two-site frequencies in the MSA:

$$\begin{aligned} f_i(a) &= \frac{1}{M}\sum_{m=1}^{M}\sigma_{ia}^m \\ f_{ij}(a,b) &= \frac{1}{M}\sum_{m=1}^{M}\sigma_{ia}^m\sigma_{jb}^m, \end{aligned} \tag{2.9}$$

where the notation defined in Eq. (2.4) is used. Intuitively, $f_i(a)$ (resp. $f_{ij}(a,b)$) is the frequency at which amino-acid a (resp. a and b) is found at column i (resp. i and j) of the MSA.
The constraints on P can then be written

$$\begin{aligned} \sum_{\{\underline{A}\}} P(\underline{A})\sigma_{i,a} &\equiv P_i(a) = f_i(a), \\ \sum_{\{\underline{A}\}} P(\underline{A})\sigma_{i,a}\sigma_{j,b} &\equiv P_{ij}(a,b) = f_{ij}(a,b). \end{aligned} \tag{2.10}$$

Combining constraints in Eq. (2.10) and Eq. (2.8), it is immediate to see that one recovers the Hamiltonian of Eq. (2.2), with couplings $J_{ij}(a,b)$ and fields $h_i(a)$ acting as Lagrange multipliers.

At this point, two important remarks should be made. The first is that in the MaxEnt setting, the choice of the Potts distribution to model protein sequence data is a consequence of the chosen observables of the data. One does not use pairwise couplings because they should represent a direct interaction in structural contacts, but because they enforce the constraint that P reproduces the pairwise frequencies of amino-acids found in the MSA.
The choice of f_i and f_{ij} as constraints is in a sense arbitrary. It is motivated by the simplicity of the observable, its natural interpretation in terms of co-evolution, and also by the typical size of MSAs [82]: the number of available sequences usually does not allow one to accurately compute higher order moments of the data, such as the three-body distribution. This does not imply that higher order terms in P are *a priori* useless, but rather that one choses to ignore them for practical reasons in the standard DCA approach.
The case of profile models also enters in the MaxEnt setting. If one chooses to ignore the pairwise distribution and to consider only the single site frequencies f_i, a model without coupling is obtained, very similar to an HMM. As section 1.2 showed, these models are at the basis of the construction of MSAs. Yet, they do not model some essential statistical features of the sequences as they do not reproduce correlations between usage of amino-acids. This highlights the fact that the right choice of observables is key in designing a good statistical model of the MSA.

The second is that by construction, the only information about the data present in P is the average value of observables. In a sense, the MaxEnt model never "sees" the full data $\boldsymbol{X}$, but only the quantities $\langle \mathcal{O}^p(x) \rangle_{\boldsymbol{X}}$. This does *not* mean that the inference procedure – finding numerical values for the Lagrange multipliers so that constraints are satisfied – should discard all information about the data, as will be explained for the pseudo-likelihood based inference method. However, two datasets $\boldsymbol{X}$ and $\boldsymbol{Y}$ that give equal values of the observables, $\mathcal{O}^p(\boldsymbol{X}) = \mathcal{O}^p(\boldsymbol{Y})$, will be exactly as likely according to the MaxEnt model, even if they differ for other statistical measures.

As a last note, the MaxEnt principle and its applications to protein sequences are closely related to the inverse Ising problem, and the inverse statistical physics in general. Here, opposed to the classical statistical physics, knowledge about "microscopic" configurations of the system of interest is available, and what misses is a model to describe them. Thus, the aim is to derive a quantitative model using the observables as a starting point.

2.3 INFERENCE METHODS FOR THE INVERSE POTTS PROBLEM

The MaxEnt principle gives a functional form for the distribution P. However, values of the Lagrange multipliers need to be computed so that the constraints are actually satisfied: this is the inference problem. Formally, given data – an MSA $\boldsymbol{A} = (\{\underline{A}^m\}, m \in \{1 \dots M\})$ with $\underline{A}^m = \left(\{a_i^m\}, i \in \{1 \dots L\}\right)$ – the correct values of parameters $\boldsymbol{J}$ and $\boldsymbol{h}$ are the ones that satisfy constraints in Eq. (2.10) for this data. This is equivalent to maximizing the log-likelihood of the data under P:

$$\begin{aligned}\mathcal{L}(\boldsymbol{A}|\boldsymbol{J},\boldsymbol{h}) &= \frac{1}{M}\sum_{m=1}^{M} \log P(\underline{A}^m) \\ &= \sum_{1\le i<j\le L}\sum_{a,b=1}^{q} J_{ij}(a,b) f_{ij}(a,b) + \sum_{i=1}^{L}\sum_{a=1}^{L} h_i(a) f_i(a) - \log Z(\boldsymbol{J},\boldsymbol{h}).\end{aligned} \tag{2.11}$$

An important assumption of Eq. (2.11) is that different sequences of the MSA are independent and identically distributed (*i.i.d.*). We will see in part 5 that because of phylogenetic relations between proteins, this is only an approximation for actual sequence alignments.
As an alternate formulation, it is straightforward to show that the log-likelihood of the data is, up to a sign and a constant, the Kullback-Leibler distance (KL-distance) between P and the empirical probability distribution defined by the data:

$$P^{obs}(\underline{A}) = \frac{1}{M}\sum_{m=1}^{M} \delta_{\underline{A},\underline{A}^m}, \tag{2.12}$$

where δ is the Kronecker symbol.
The log-likelihood is a concave function in the parameters, as can easily be checked from computing its Hessian matrix. As a consequence, the global maximum is attainable by simple optimization methods such as gradient ascent. However, the exact numerical computation of the likelihood function or of its gradient is unfeasible: the number of terms in the sum defining the partition function (Eq. (2.3)) is L^q, with $q = 21$ and L of the order of 100 for typical protein domains.
Many approximation methods are available to tackle this problem, and three of them will be described below. The Mean-Field (MF) approximation is the first "efficient" method that has been proposed for this problem in the context of protein sequences, and though very limited, it can give an estimate of the relative strength of coupling matrices J_{ij} [57]. The Pseudo-Likelihood Maximization (PLM) method is currently the state of the art unsupervised method for predicting contacts in protein structures using DCA [27]. Finally, the Boltzmann Machine Learning (BML) is a computationally expensive method relying on Montecarlo sampling, able to achieve arbitrarily accurate solutions to the inference problem.

2.3.1 *Mean-field approximation*

First introduced in [57] for protein sequences, he MF approximation relies on a high-temperature expansion of the Legendre transform of the free energy of Hamiltonian $\mathcal{H}$ [77]:

$$G = -\log Z + \sum_{i=1}^{L}\sum_{a=1}^{q-1} h_i(a)P_i(a). \tag{2.13}$$

The sum over states a of variable i runs only up to $q-1$ because of the lattice-gas gauge choice, see section 2.4.1. In practice, one places a factor β in front of the couplings in the Hamiltonian, Eq. (2.2). The functional in Eq. (2.13) can be computed through a first order expansion around $\beta = 0$. This is the so called Plefka expansion [38], assuming high temperature or low couplings.
The linear response equations are then used to relate couplings and fields to one and two point statistics:

$$\begin{aligned} h_i(a) &= \frac{\partial G}{\partial P_i(a)}, \\ (C^{-1})_{ij}(a,b) &= \frac{\partial h_i(a)}{\partial P_j(b)}, \end{aligned} \tag{2.14}$$

where we have introduced the connected correlation matrix $\boldsymbol{C}$:

$$C_{ij}(a,b) = P_{ij}(a,b) - P_i(a)P_j(b). \tag{2.15}$$

The resulting MF equations reads

$$P_i(a) = \frac{1}{z_i} \exp\left(h_i(a) + \sum_{j \neq o} \sum_{b=1}^{q-1} J_{ij}(a,b) P_j(b) \right) \tag{2.16}$$

and

$$(C^{-1})_{ij}(a,b) = \begin{cases} -J_{ij}(a,b) & \text{for } i \neq j \\ \frac{\delta_{a,b}}{P_i(a)} + \frac{1}{P_i(q)} & \text{for } i = j \end{cases} \tag{2.17}$$

To solve the inference problem, one just has to replace $P_i(a)$ and $P_{ij}(a,b)$ by their empirical counterparts $f_i(a)$ and $f_{ij}(a,b)$ (see Eq. (2.10)). By inverting the empirical correlation matrix $\boldsymbol{C}$, it is possible to obtain values for the couplings $\boldsymbol{J}$ in a single step. In the same way, the fields are recovered by inverting equation (2.16).
The careful reader will notice that by definition, the connected correlations matrix $\boldsymbol{C}$ has zero modes:

$$\forall i,j,a, \quad \sum_{b=1}^{q} C_{ij}(a,b) = \sum_{b=1}^{q} P_{ij}(a,b) - P_i(a)P_j(b) = 0. \tag{2.18}$$

Therefore, it is not invertible. This is a result of the over-parametrization of the model, and is fixed by operating in the lattice-gas gauge (see section 2.4.1).

The MF approximation was the first efficient method for inferring Potts models in the context of protein sequences. Its main advantage is that it is computationally quite inexpensive, since the only needed operation is to inverse a matrix once. Resulting parameters usually allow for accurate prediction of contacts in the protein structure (see [57] for this biological application, as well as a rigorous derivation of equations (2.16) and (2.17)). Similar schemes such as the gaussian DCA [5] or the Pseudo-Sparse Inverse Covariance (PSICOV) [48] have been used with success.
However, it is important to note that MF suffers from severe drawbacks. First, the $i = j$ case of Eq. (2.17) is inconsistent for the empirical statistics f_i and f_{ij}, showing that the approximation remains very crude. Seconds, even though inferred parameters lead to good biological predictions in some cases, they cannot claim to be a statistical description of the MSA. Values of the MF couplings are usually very high, leading to glassy behavior and a distribution P that is hard to sample by Markov Chain MonteCarlo (MCMC) simulations. This is consistent with theoretical work showing that this approximation typically *overestimates* couplings [7].
As a result, for protein MSAs, a mean-field inferred model does *not* satisfy the constraint in Eq. (2.10), and is only an approximate solution to the MaxEnt problem.

2.3.2 *Pseudo-likelihood maximization*

The PLM method aims at approximating the likelihood in Eq. (2.11) by a tractable expression [4]. We introduce the quantity $P_r(a_r|a_{\setminus r})$

$$P_r(a_r|a_{\setminus r}) = \frac{\exp\left(h_r(a_r) + \sum_{j\neq r} J_{rj}(a_r, a_j)\right)}{\sum_{b=1}^{q} \exp\left(h_r(b) + \sum_{j\neq r} J_{rj}(b, a_j)\right)}, \tag{2.19}$$

which is the probability to find site r in state a_r *given* that the rest of the configuration is $a_{\setminus r} = (a_1 \dots a_{r-1}\ a_{r+1} \dots a_L)$, where $\underline{a}$ is the mth sequence in the MSA.
The *pseudo*-likelihood is then written as

$$\begin{aligned} p\mathcal{L}(\boldsymbol{A}|\boldsymbol{J}, \boldsymbol{h}) &= \sum_{r=1}^{L} \mathcal{L}_r \\ &= \sum_{r=1}^{L} \left(\frac{1}{M} \sum_{m=1}^{M} \log P_r^m(a_r^m|a_{\setminus r}^m) \right). \end{aligned} \tag{2.20}$$

Essentially, the approximation is to factorize P in L single site distributions depending on a_r, with the rest of the configuration being fixed to its data value $a_{\setminus r}^m$.
From there, two strategies can be designed. The first is to directly maximize the pseudo-likelihood 2.20, for instance through gradient ascent. This is the so-called symmetric PLM [28].
The second is to notice that functions $\mathcal{L}_r$ each depend on a different set of parameters h_r and $\boldsymbol{J}_r = \{J_{ri}\}_{i\neq r}$, forgetting for a moment that couplings have to be symmetric, *i.e.* $J_{ij} = J_{ji}$. In the *asymmetric* PLM [27], each $\mathcal{L}_r$ is maximized independently, and couplings are combined by a simple average

$$J_{ij} = \frac{1}{2}\left(J_{ij}^i + J_{ji}^j\right), \tag{2.21}$$

where the superscripts i (resp. j) mean that the coupling estimate comes from maximizing $\mathcal{L}_i$ (resp. $\mathcal{L}_j$). The asymmetric version of PLM has been shown to be faster and more accurate than the symetric one [27], and will be the one used throughout this work.

PLM is the most commonly used co-evolutionary tool for predicting contacts in protein structures from sequence data. It has been shown to give high quality results over a large number of protein families [27]. Statistically, the pseudo-likelihood is a *consistent* estimator: if sequences in the MSA were samples drawn from a Potts distribution, and if the number of available sequences M was infinite, the estimates of PLM would be exact. This contrasts with MF, which will show inconsistencies even if this ideal case if the couplings of the "underlying" Potts model are not small enough.

The distribution P inferred by PLM is typically closer to fulfilling constraints in Eq. (2.10) than MF – see chapter 3 for quantitative information. However, it fundamentally remains an approximation, and large deviation of the two point statistics from the inferred model P_{ij} from the data f_{ij} can be observed.

Last, it is important to notice that in order to compute $\mathcal{L}_r$, knowledge of the full sequences $\underline{A}$ is needed. This contrasts with what was said about the MaxEnt principle in section 2.2.2: the only information the model should have about the data is the average value of the chosen observables, in this case the one and two point statistics. The MaxEnt principle has been subject to critics for this reason [3].
One way to solve what appears as a conflict is to consider the MaxEnt principle and the inference method as fulfilling two different goals. The MaxEnt principle can be seen as a way to parametrize the distribution P that one wants to fit to the data, whereas an inference method such as PLM is a necessary approximation needed to quantify parameters of P.
Moreover, the result of the inference is still in agreement with MaxEnt: the only information needed to know if distribution P was correctly inferred is the pairwise statistics of the model, P_{ij}, and of the data, f_{ij}. In this sense, the pairwise statistics f_{ij} is still sufficient to determine the correct model. If data was indeed sampled from a Potts distribution, PLM guarantees that those two quantities will match. In this case, the knowledge of the full sequence needed for PLM is in a sense irrelevant: if pairwise statistics are sufficient to define the Potts distribution, knowledge of full sequences does not add any information to the resulting model.

2.3.3 *Boltzmann machine learning*

As stated at the beginning of section 2.3, the likelihood is a concave function of parameters $\boldsymbol{J}$ and $\boldsymbol{h}$. This means that a simple optimization scheme such as gradient ascent is guaranteed to find the maximum value of $\mathcal{L}$ [1]. The gradient of the likelihood with respect to the parameters is

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial h_i(a)} &= f_i(a) - P_i(a), \\ \frac{\partial \mathcal{L}}{\partial J_{ij}(a,b)} &= f_{ij}(a,b) - P_{ij}(a,b). \end{aligned} \tag{2.22}$$

Unfortunately, the exact computation of $P_i(a)$ and $P_{ij}(a,b)$ is intractable. The idea of BML is to draw samples from the current distribution P through MCMC sampling, and use these samples to estimate the gradient. Parameters of P are then updated in the direction of the

gradient [41, 76]. This process is iterated until satisfactory convergence is reached.
BML has one main advantage: it can be arbitrarily accurate. The only limitation to the accuracy of the estimation of the gradient is the size of the MCMC samples. With enough computational power, one can achieve estimations of $\boldsymbol{J}$ and $\boldsymbol{h}$ as accurate as desired. Another benefit is that given the expression of the gradient in Eq. (2.22), it is clear that the only information BML uses is the pairwise statistics of the data f_{ij}.
The major drawback of this method is the computational cost. Each estimation of the gradient requires a sample from P, ideally large enough to avoid fluctuations. Even though the likelihood is a concave function, it does not mean that the number of necessary gradient ascent steps is small. For typical datasets, there seems to exist "flat" directions in parameter space, making the learning process very long [32]. Furthermore, it is impractical to consider computing the Hessian matrix to speed up convergence, as its size is of the order of $L^2q^2 \times L^2q^2$.

All these problems will be addressed in section 3, where we refer to an efficient implementation of the Boltzmann machine. However, even in this case, the inference process is too slow to compete with PLM or MF methods, and impractical for most biological applications. The main advantage of the BML is that the resulting distribution P satisfies constraints (2.10) with very high accuracy, making it a plausible statistical model of the MSA.

2.4 TECHNICAL POINTS

2.4.1 *Gauge invariance*

The constraints in equation (2.10) are not all independent. Indeed, the single site frequencies $f_i(a)$ sum up to 1, and the pairwise frequencies $f_{ij}(a, b)$ have the $f_i(a)$'s as marginals. As a result, the number of truly independent observables is $N \cdot (q-1)$ for the single site frequencies and $N(N-1)/2 \cdot (q-1)^2$ for the pairwise. This means that the Potts model is over-parametrized: with couplings $J_{ij}(a, b)$ and fields $h_i(a)$, there are more free parameters than constraints.
This results in what is called the *gauge invariance*: it is possible to modify the parameters of the Potts model without changing the probability distribution defined in Eq. (2.1). For any arbitrary function $K_{ij}(a)$ with $1 \leq i, j \leq N$ and $a \in \{1 \dots q\}$ and for arbitrary constants c_i and

c_{ij}, $1 \leq i, j \leq N$, the following transformation does not change the probabilities:

$$\begin{aligned} J_{ij}(a,b) &\to J_{ij}(a,b) + K_{ij}(a) + K_{ji}(b) + c_{ij}, \\ h_i(a) &\to h_i(a) - \sum_{j=1\,(j\neq i)}^{N} K_{ij}(a) + c_i. \end{aligned} \tag{2.23}$$

It can be easily verified that this transformation leads to adding a constant to the energies of all configurations $\underline{A} = (a_1 \dots a_L)$. Since this constant will be compensated for in the partition function Z (see Eq (2.3)), probabilities defined by the model are unchanged.
In most applications of Direct Coupling Analysis (DCA), particularly contact prediction, it is usual to work in the so-called *zero-sum gauge*, also known as *Ising gauge*. This gauge is obtained by the transformation

$$\begin{aligned} J_{ij}(a,b) &\to J_{ij}(a,b) - J_{ij}(a,\cdot) - J_{ji}(\cdot,b) + J_{ij}(\cdot,\cdot), \\ h_i(a) &\to h_i(a) - h_i(\cdot) + \sum_{j=1\,(j\neq i)}^{N} \left(J_{ij}(a,\cdot) - J_{ij}(\cdot,\cdot)\right), \end{aligned} \tag{2.24}$$

where the notation $g(\cdot)$ stands for the average $q^{-1}\sum_{a=1}^{q} g(a)$. After this transformation, parameters of the Potts model are such that

$$\sum_{b=1}^{q} J_{ij}(a,b) = \sum_{a=1}^{q} J_{ij}(a,b) = \sum_{a=1}^{q} h_i(a) = 0. \tag{2.25}$$

Importantly, this gauge minimizes the Frobenius norm of couplings $||J_{ij}||$, and is considered optimal for contact prediction for this reason [28, 82] (see sections 2.4.3 and 2.5.1).
In the case of the MF implementation of DCA, the so-called *lattice-gas* gauge is used, in which one state (usually q) is chosen as a reference for the energies, so that

$$\forall a, b, \quad J_{ij}(a,q) = J_{ij}(q,b) = h_i(q) = 0. \tag{2.26}$$

This allows for the inversion of the correlation matrix described in 2.17 by removing trivially zero modes: lines and columns corresponding to the state q are removed from the correlation matrix, making it full-rank in case of sufficient data.

2.4.2 *Regularization*

Although large, the number of sequences in MSAs is far from being infinite. Therefore, the use of a regularization in the DCA inference is essential to avoid overfitting. Indeed, the number of parameters of the Potts distribution for sequences typically of length $L \sim 100$ is of the order of 10^6, making overfitting a risk. In order to reliably measure

the observables f_{ij} that DCA aims to fit, the number of sequences in an MSA needs to be large with respect to the possible number of states of columns i and j. That is, we need the number of sequences M to be large with respect to $q^2 \sim 400$. For most protein families DCA is used on, values of M in the range of 100–1000 barely satisfy this condition. A simple example illustrates the need for regularization: in the typical setting, DCA will be used on an MSA of about $M = 1000$ sequences. Suppose amino acids A and B appear 10 times in columns i and j, leading to measured frequencies $f_i(A) = 10^{-2}$ and $f_j(B) = 10^{-2}$. In the absence of correlation, the expected value of the joint appearance of A and B is $f_{ij} = 10^{-4}$. Therefore, if the joint presence of A and B is observed in one sequence, the frequency $f_{ij} = 1/M = 10^{-3}$ will exceed by a factor 10 the expected value, leading to a high measured correlation and the need for a strong positive coupling. On the other hand, if A and B are never observed together, an infinitely negative coupling is needed for the Potts model to render this. Both of this large coupling values are not justified by the available data, and the use of a regularization term suppresses this problem.

In inference methods relying on the maximization of the likelihood, such as PLM or BML, ℓ_2 regularization is usually used [28, 32]. Instead of directly maximizing the likelihood in Eq. (2.11), one adds a penalty term proportional to the parameters, defining the regularized likelihood:

$$\mathcal{L}(\mathcal{D}_{nat}|\boldsymbol{J},\boldsymbol{h}) = \log P(\mathcal{D}_{nat}|\boldsymbol{J},\boldsymbol{h}) - \lambda_J||\boldsymbol{J}||^2 - \lambda_J||\boldsymbol{h}||^2, \tag{2.27}$$

with parameters λ_J and λ_h defining the strength of regularization. ℓ_2 regularization can also be interpreted as a gaussian prior on the parameters of the Potts model [7]. In this case, the λ parameters should scale as $1/M$, vanishing for infinite number of sequences. However, in typical implementations of DCA, a fixed value of $\lambda = 10^{-2}$ or 10^{-3} is chosen. The ℓ_2 term usually makes the optimization of Eq. (2.27) easier and faster. However, as will be seen in section 3.3, it biases the Potts probability distribution.

In the MF technique, another common regularization scheme is the use of pseudocounts [7, 57]. The empirical correlation matrix obtained from the MSA is typically not invertible due to finite sampling. Pseudocounts are used to "add" random observations to the data in the following way:

$$\begin{aligned} f_i(a) &= \frac{1}{\lambda + M}\left(\sum_{m=1}^{M} \sigma_{i,a}^m + \frac{\lambda}{q}\right), \\ f_{ij}(a,b) &= \frac{1}{\lambda + M}\left(\sum_{m=1}^{M} \sigma_{i,a}\sigma_{j,b}^m + \frac{\lambda}{q^2}\right). \end{aligned} \tag{2.28}$$

It was observed in [57] that a very large pseudocount $\lambda \simeq M$ is needed for optimal contact predictions using this method.

2.4.3 *Phylogenetic biases: sequence re-weighting and* APC *correction*

Sequences found in an MSA are not *i.i.d.* samples from a background distribution. They are the result of an evolutionary process, and are related to each other by phylogenetic relations. Importantly, this means that they cannot be considered as independent observations, as sequences that share a recent common ancestor are likely to share a large part of their amino acid composition. This results in statistical biases described in more details in chapter 5.
To remedy this, a simple re-weighting scheme has been introduced in [82]. Each sequence $\underline{A}^m$ in the MSA is attributed a weight w^m defined as the inverse of the number of sequences that share more than $\delta = 80\%$ identity with $\underline{A}^m$. Formally,

$$w^m = \left(1 + \sum_{n \neq m} \Theta\left(d_{m,n} < \delta\right)\right)^{-1}, \tag{2.29}$$

with $d_{m,n}$ standing for the Hamming distance between sequences $\underline{A}^m$ and $\underline{A}^n$, and Θ being the Heaviside function. As a consequence, an isolated sequence in terms of Hamming distance will have a weight of 1, whereas a sequence with many neighbours will be down-weighted. Frequencies are then computed accordingly:

$$\begin{aligned} f_i(a) &= \frac{1}{M_{eff}} \sum_{m=1}^{M} w^m \sigma_{ia}^m \\ f_{ij}(a,b) &= \frac{1}{M_{eff}} \sum_{m=1}^{M} w^m \sigma_{ia}^m \sigma_{jb}^m, \end{aligned} \tag{2.30}$$

with $M_{eff} = \sum_{m=1}^{M} w^m$. The value of the threshold $\delta = 0.8$ is arbitrary, and the most commonly used in DCA inferences. Contact prediction results are robust with respect to variations around this value, and consistently better than without any re-weighting.

Another proposed method to take phylogenetic biases into account is the so-called Average Product Correction (APC). First introduced in [24], APC is a correction to the the interaction score between two columns of the MSA. In the case of DCA the interaction score used to predict contacts is the Frobenius norm of the coupling matrix $F_{ij} = ||J_{ij}||^2 = \sum_{a,b=1}^{q} J_{ij}(a,b)^2$ (see section 2.5.1). In this case, the APC applies the following transformation to the scores:

$$F_{ij}^{APC} = F_{ij} - \frac{F_{i\cdot} F_{\cdot j}}{F_{\cdot\cdot}}, \tag{2.31}$$

where the $\cdot$ stands for averaging over i or j. This correction is now used in all implementations of DCA that are used for contact predictions.
The rational behind equation (2.31) is that DCA couplings either reflect a direct structural or functional interaction of i with j, or a background

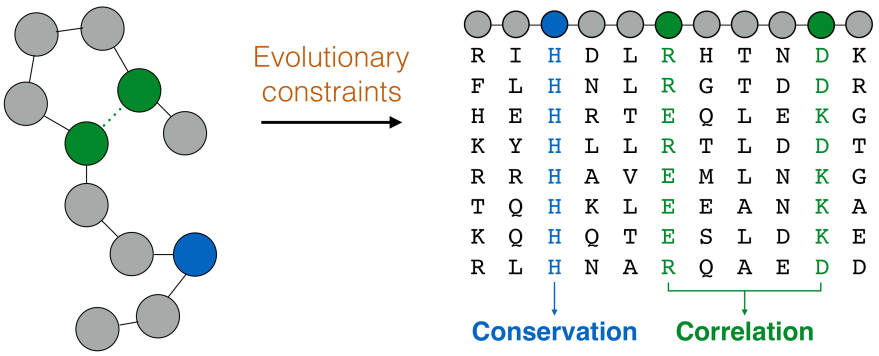

Figure 2.1: Structure imposes constraints on the evolution of a protein sequence, which appear as statistical patterns in an MSA. Here, contact between the two green residues is essential to maintain the fold, leading co-evolution. The consequence is the correlation of the two corresponding column in the alignment. This idea is at the basis of contact prediction methods including DCA.

interaction due to phylogenetic biases. Since this background phylogenetic bias is expected to be roughly the same between any pair of columns, removing the average interaction of i/j to all other columns should suppress it [12]. It has recently been argued that APC is more of a correction to entropic biases than to phylogeny [80]. However, the exact reason of the success of the APC remains unclear.

2.5 STATE OF THE ART: APPLICATIONS OF DCA

DCA was first introduced as a tool to predict pairs of contacting residues in a protein structure from the knowledge of sequences of homologs [82]. For this purpose, a Potts model (see Eq. (2.1)) is inferred based on statistical patterns found in an MSA of homologous sequences. The success of this application has encouraged the use of the DCA model for other purposes. Main achievements in this regard are the prediction of protein-protein interactions, and the scoring of mutations or even full protein sequences. They are described in the next sections. A more detailed review can be found in [19].

2.5.1 DCA *predicts residue-residue contacts*

Predicting a protein's structure from its sequence is one of the oldest bioinformatics or biophysics question. The function of a protein is usually determined by its structure, this knowledge is essential to understanding how proteins operate. However, experimental characterization using X-ray diffraction, NMR, or more recently electron microscopy, remains expensive and time consuming. On the other hand, methods relying on molecular dynamics simulations are computationally very expensive and cannot readily be applied to large

sequences.
The recent accumulation of sequence data and its classification in families lead to the use of statistical methods to help structure prediction. Forgetting the idea of directly determining the full structure from the sequence, those methods focus on predicting pairs of residues which are likely to be in contact in the fold. The underlying idea is to use the observed correlation in amino acid usage in different columns in an MSA, termed co-evolution. If the contact between residues at two positions i and j in the sequence is essential to the structure, they should co-evolve. If one of the two positions mutates, the other may have to mutate in a compensatory way to maintain contact. Thus, a correlation will be measured between the corresponding columns of the alignment. Figure 2.1 illustrates this idea with a sketch protein.
The first applications of this idea directly used correlation between columns as a score indicating the likelihood of contact [24, 37, 59]. A commonly used correlation score is the mutual information of columns i and j:

$$MI_{ij} = \sum_{a,b=1}^{21} f_{ij}(a,b) \log \frac{f_{ij}(a,b)}{f_i(a) f_j(b)}, \tag{2.32}$$

where $f_i(a)$ and $f_{ij}(a,b)$ are respectively single site and pairwise frequencies defined in 2.9.
However, this is limited by the fact that correlations can result from direct or indirect interactions between variables, as is discussed in section 1.2.4. To accurately predict contacts, it is necessary to disentangle direct and indirect effects. The idea behind DCA is to find a model that reconstructs the observed correlations using a network of direct couplings. In practice, this model takes the Potts form of Eq. (2.1), where the direct couplings $\boldsymbol{J}$ are responsible of correlations between variables in the probability distribution P.

The contact prediction proceeds as follows. Pairs of columns (i,j) are ranked according to the frobenius norm of the J_{ij} coupling after the APC is applied, see section 2.4.3 (previous works used the so-called direct information [57] – see article in section 3 for a definition). Pairs with $|i-j| \leq 4$ are discarded from this ranking: strong couplings between close-by positions in the sequence are frequently caused by stretches of gaps in the alignment. Moreover, this constraint leads to predictions for pairs which cannot be said to be structurally close from the sequence only (note this excludes pairs corresponding to one turn in an α-helix). The top pairs of this ranking are then used as contact predictions.
As an illustration, figure 2.2 shows the result of this prediction for the Pfam family PF00014 using the 30 top pairs. The performance of this method has now been evaluated for numerous protein families [28, 57], demonstrating a consistent improvement over purely correlation

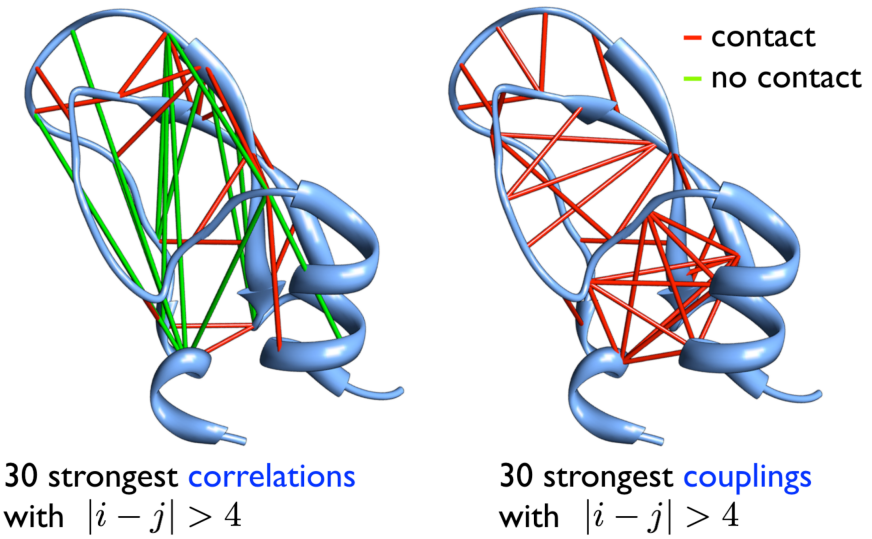

Figure 2.2: Contact prediction based on the 30 top pairs for the mutual information score of Eq. (2.32) (left) and for the DCA score (right). Predictions are mapped onto the protein structure (PDB: 5PTI [83]): green bars link distant residues (distance $> 8\mathring{A}$), and red bars link contacting residues (distance $< 8\mathring{A}$). Source: [19]

based techniques. This is illustrated in figure 2.3. The currently best stand-alone method in this aspect is the PLM implementation of DCA [27].

In terms of structure prediction, determining contacts from the sequence beforehand is of great help when computationally folding the protein. Recently, thousands of new protein structures have been predicted using such co-evolutionary methods [62]. They have also been extensively used in recent CASP competitions. Interestingly, the quality of the inference of the DCA model matters little in the accuracy of contact predictions. Very crude approximations like the MF method still allow for very good predictions. Moreover, as will be seen in the article of chapter 3, a very accurate method in terms of reproducing statistical features of the MSA like the BML does not outperform the much more approximate PLM in contact predictions. This can be explained by the fact that to achieve good contact predictions, it is only necessary to recover the topology of interactions of the graph, and not the exact value of parameters. The procedure used to predict contacts illustrates that only the *ranking* of the strongest inferred couplings matters, and not their precise numerical values. Moreover, the values and ranking of smaller coupling parameters has no influence at all in this matter. For this reason, it is important to make a distinction between inferring an exact DCA model, that is solving the maximum entropy problem from section 2.2.2, and recovering the topology of an interaction network.

Lastly, it is important to note that the current best contact prediction

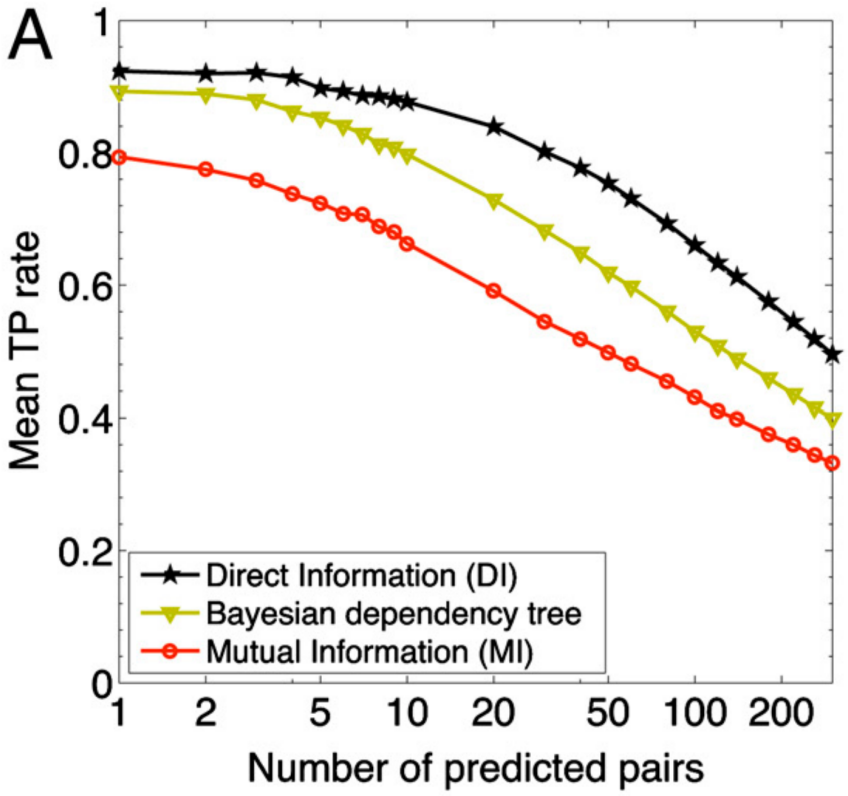

Figure 2.3: Fraction of correctly predicted contacts as a function of the number of pairs (i, j) for which a prediction is made, averaged over 131 protein families. Contact is defined as a distance smaller than $8\mathring{A}$ between the two residues. *DI* is a score based on the couplings of the DCA model, here inferred using the MF method (see Methods of article in chapter 3 for a definition). The Bayesian dependency tree is a model used in [12], also attempting to disentangle direct and indirect sources of correlation. Source: [57].

are obtained using so-called meta methods [49, 74, 81]. These methods use information of many different sources, such as results of DCA but also secondary structure or solvent accessibility predictions. These informations along with known protein structures are used to train deep neural networks. The obtained networks have a significantly higher accuracy in predicting contacts. Differently from DCA techniques, these methods are supervised and need to be trained on already known protein structures.

2.5.2 *Scoring mutations and sequences*

The Potts model defines a probability distribution on all possible sequences, based on their likelihood to belong to the family it has been trained on. It can thus be used to quantitatively score sequences, predicting whether they could be plausible members of the family, *i.e.* similar in structure and function. This opens the way to two interesting applications: scoring of mutations in natural sequences, and generating artificial sequences for a given family by sampling the Potts distribution P.

In the former application, one tries to quantitatively assess whether a mutation – *i.e.* the change in one or a few amino acids – is deleterious for the function of a given protein. Thanks to next-generation sequencing, experimental quantitative characterization of mutational landscapes is becoming increasingly accessible, and has been performed for a number of proteins [45, 54, 55]. The setting is quite simple: a reference wild-type sequence $\underline{A}^{wt}$ is chosen, and a set of mutant sequences is designed. Those can be single mutants – changing every possible amino acid of $\underline{A}^{wt}$ into every other one, one at a time –, or larger modifications such as double or more mutants. A proxy biological function of every mutant is then experimentally determined, through the measurement of, *e.g.* , structural stability, binding affinity to some known target, or global fitness of the organism in which the sequence is expressed.
The DCA model can simply be used as a computational predictor of the result of such experiments. The simplest scoring system is to compare the energies (or log-probabilities) of the wild-type sequence and of the corresponding mutant:

$$\Delta E^{mut} = \mathcal{H}(\underline{A}^{mut}) - \mathcal{H}(\underline{A}^{wt}). \tag{2.33}$$

This computational score can be compared to experimental results to assess the quality of DCA in this task. This has successfully been done for a number of cases, comprising viral, bacterial or human proteins [13, 16, 33, 36, 44, 52]. Such predictions are of high interest both from a medical and evolution point of view, as mutations are responsible of antibiotic resistance in bacterias, of viruses' capacity of adaptation,

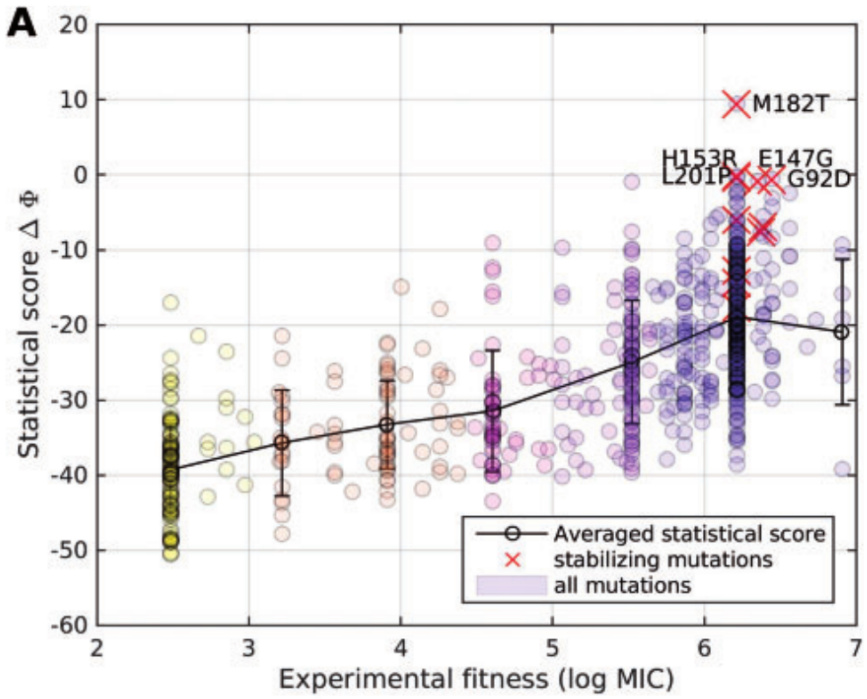

Figure 2.4: Scoring mutations for the beta-lactamase TEM-1 protein, responsible for antibiotic resistance. The Minimum Inhibitory Concentration (MIC) of 990 single mutants of this protein has been experimentally characterized in [45]. The statistical energy score obtained from the DCA is computed for each of those mutants, and compared to the experimental value. Source: [33].

and of some diseases in humans. The ability of DCA-like models to predict the effect of mutations based on the knowledge of homologous sequences is the idea on which the integrative modeling described in chapter 4 is based.

Scoring mutations in sequences can be thought of as a *local* reconstruction of the fitness landscape of a protein. There are encouraging results that DCA could also lead to a *global* reconstruction of the fitness landscape, assessing the functionality of sequences far from any natural one.
First experiments in this regard have been conducted in [68, 75]. The WW domain family (Pfam PF00397) was used, a short 35 residues long domain. Using a small curated alignment of 120 WW sequences, called NAT in the following, artificial sequences were designed in three ways:

- (R - random): Random scrambling of the alignment was performed, killing all existing statistical patterns.
- (IC - independent conservation): Each column of the MSA was shuffled independently, thereby perfectly conserving the intra-column frequencies but removing any correlation between columns.

- (CC - coupled conservation): starting from the IC dataset, a simulated annealing procedure was used to reproduce the pairwise frequencies $f_{ij}(a, b)$ found in the original MSA. This procedure inspired the one described in section 5.1.4.

Importantly, it has been shown that if performed properly, the procedure of the CC dataset is equivalent to sampling from a perfectly inferred DCA model, that is a maximum entropy model that exactly reproduced the pairwise statistics $f_{ij}(a, b)$ of the original alignment [9].
Artificial sequences from the 3 datasets were tested for folding stability in [75] and for binding specificity to a target in [68]. None of the R or IC sequences could fold into the proper structure. However, 31% of the CC and 67% of the NAT folded correctly. This result is a very strong indication as to what information is needed to specify a functional protein: the sole knowledge of column-wise statistics is not sufficient (IC), but the pairwise distribution of amino acids in columns is necessary and maybe close to sufficient (CC). These results are summarized in figure 2.5.
This is a good indication that DCA-like models based on pairwise statistics could be a good representation of the fitness landscape of proteins and potentially be generative. The above states procedure of generating artificial sequences results in an alignment, but is unable to predict the functionality of the single sequence. It was first shown in [4] that a PLM implementation of DCA could be used over those artificial sequences to predict which of them could fold and which could not. The energy of each sequence in the model is computed, and those of low energy are considered more likely to be functional.
This was further investigated in [19], where sequences generated by sampling the DCA distribution P were compared with the ones mentioned above. Results are summarized in figure 2.6. DCA energy distribution of sequences generated at random (like R), by an independent profile model and by an Adaptive Cluster Expansion (ACE) [8, 17] inferred DCA were computed. Whereas random sequences lie at very high energies, distribution of those coming from an independent and from the DCA model overlap. Energies of sequences from [75] were then computed. CC and IC sequences show overlapping energies, with CC being lowest on average, while the natural sequences have a small overlapping with IC sequences. The main result is that energy seems to be a very good discriminator between folding and non-folding sequences: almost no sequence that is in the energy range of the IC data is found to be folding, even if it is a natural one. On the contrary, CC sequences with energies similar to the natural sequences are the only functional artificial ones (see figure 2.6. Importantly, the pairwise coupling terms of the Potts model are needed to achieve this. The energy of a profile model or a HMM based score fail to discriminate between folding and non-folding sequences [20].

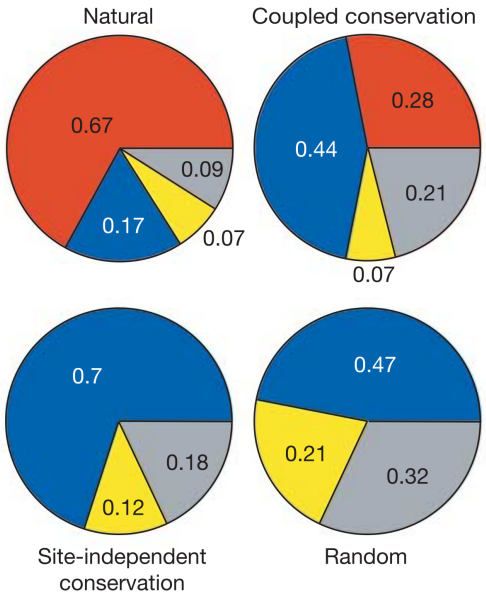

Figure 2.5: Folding experiment for artificial sequences of [75]. Grey: not expressed. Yellow: insoluble. Blue: soluble but not folded. Red: folded. Source: [75].

As in contact prediction, the different inference methods used for obtaining the DCA model do not lead to large differences in scoring sequences, either for local mutational landscape or global generative purposes. Again, the only thing the model needs to do is to have a correct *ranking* of the energies of sequences. The precise numerical value of those energies does not directly matter. However, it is important to notice that *scoring* and *generating* sequences are two different tasks. In order to generate sequences such as the CC ones, it is necessary that the DCA distribution reproduces pairwise frequencies of the natural MSA as closely as possible [9]. For this purpose, accurate inference methods are needed, such as the ACE or the BML.

2.5.3 *Other applications of* DCA

PROTEIN-PROTEIN INTERACTION DCA can be used to determine the existence of a biological interaction between members of two protein families, and also to help determine the structural interface of this interaction if it exists. In practice, alignments of two families A and B are used, with sequences of A and B belonging to the same species being concatenated. In the case of paralogs, that is when several sequences of A and B belong to the same organism, the so-called matching problem has to be solved beforehand. If simple criteria such as genomic co-localization is not available, DCA can actually be used for such a task [10, 39].
A joint Potts model is then inferred for the concatenated alignments

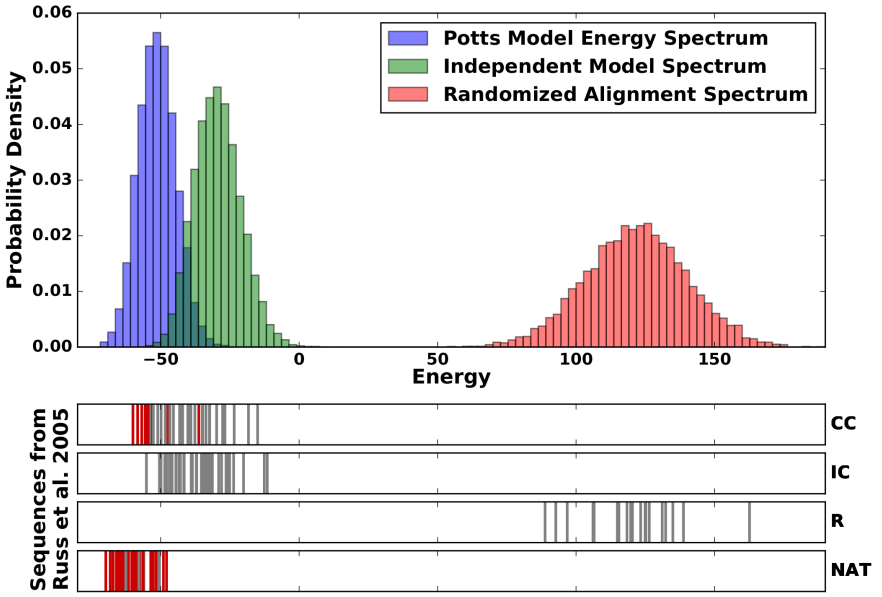

Figure 2.6: **Top**: DCA energy distribution for sequences sampled by the DCA model (blue), the profile model (green) and random sequences (red). **Bottom**: DCA energies of WW sequences from [75]. Each bar indicates a sequence. Red bars indicate a folding sequence. Source: [19].

of families A and B. By construction, this model will include both intra (J^A J^B) and inter (J^{AB}) protein couplings. The latter inter protein couplings represent an interaction between residues of the two proteins according to the model. The strength of these couplings can then be used to estimate whether members of the two families are likely to interact. In [29], the average Frobenius norm F^{APC} of the top 4 inter-protein couplings is used as an interaction score, predicting interacting pairs among the dozens of proteins constituting the small and large ribosomal subunits.
The same idea can be applied to predict the structure of the interaction interface. Similarly as in the contact prediction framework, large inter protein couplings are predicted to be residue-residue contacts in this interface. Using those predictions to guide computational structural prediction methods allows the reconstruction of the interface [43, 61, 71]. The same idea can be applied to predict the structure of the interaction interface. Similarly as in the contact prediction framework, large inter protein couplings are predicted to be residue-residue contacts in this interface. Using those predictions to guide computational structural prediction methods allows the reconstruction of the interface [43, 61, 71].

DCA COUPLINGS REFLECT BIOPHYSICAL INTERACTIONS Strong couplings inferred by DCA correspond to contacts. Is it possible to push

further the interpretation of those parameters? Coupling matrices J_{ij} tend to be very noisy, making interpretation of the detailed interaction between two residues unpractical. However, it has been shown in [21] that averaging over many coupling matrices J_{ij} for different protein families and performing a spectral analysis of the resulting average coupling matrix allows one to recover known bio-chemical interactions. Main eigenmodes of the average coupling matrix are shown to be patterns of electrostatic, hydrophobic/hydrophilic or cysteine-cysteine interactions. Moreover, these statistical couplings are in good agreement with the widely used Miyazawa-Jernigan potentials derived from typical residue-residue contacts found in proteins [56]. Therefore, information contained in the parameters of the DCA model can be interpreted in terms of biochemistry.

3

POTTS MODELS ARE ACCURATE STATISTICAL DESCRIPTION OF PROTEIN SEQUENCE VARIABILITY.

3.1 MOTIVATION

The success of DCA methods in terms of modeling protein sequences is impressive. Summarized in section 2.5, they include the prediction of structural contacts inside and between proteins, the prediction of protein-protein interaction partners, and the scoring of mutations. Less known applications, but still informative about the descriptive capacity of the model, include its ability to score artificial sequences for functionality and the fact that its coupling parameters recover biochemical interactions between amino acids.
The fact that a single class of models is able to quantitatively describe properties of proteins in so many ways calls for explanation. Usually, two hypotheses are given for the cause of this success:

- DCA disentangles indirect from direct correlations. Statistical correlation as measures in an MSA can come from direct interaction of the two corresponding residues (*e.g.* structural contact), or from indirect effects, such as mediation through intermediary residues. The success of Potts models is attributed to their ability to explain correlation using direct coupling parameters. As those direct couplings reflect "real" physical or biological interactions between residues of the protein, the resulting model can be interpreted in a meaningful way.

- The maximum entropy principle is used at the basis of DCA. The Potts model is the most general model that reproduces pairwise statistics measured in MSAs. In a sense, the only information used by the model is this pairwise statistics. If it turns out that this is the correct quantity to reproduce, then Potts model should be good sequence models. On the other hand higher order terms which are not included in Potts model might be important for making a sequence functional.

These two hypotheses raise questions. First, if it is true that direct and indirect effects are explained through couplings $\boldsymbol{J}$, it should be possible to determine what indirect effects "look like": structurally characterizing them, and quantify their effect for the co-evolution of residues. Seconds, the MaxEnt reasoning only holds if pairwise statistics is the correct observable to reproduce. For this reason, it is essential to understand if there exist statistical features of the natural sequences

that are *not* reproduced by the DCA model. In other words, one should understand if higher order terms are needed to accurately describe sequence variability in protein families.
To answer these two questions, it is necessary to be in possession of a very accurately inferred model. Indeed, if the inferred DCA model does not closely fit pairwise statistics, it is insignificant to try to explain indirect correlations using its parameters, or to try finding statistical observables that it cannot reproduce. Popular and efficient inference scheme include the MF approximation and the PLM method. While very useful for some most applications, these methods remain very approximate. Mainly, a sample from the probability distribution P learned using them will be statistically quite different from the natural sequences, even concerning fitted quantities. For this reason, answering the previous questions requires the design of an accurate inference scheme.
In the following article [32], a BML method is used to infer Potts models for the largest Pfam families. Properties of this model are then investigated, the purpose being to understand the limits of DCA for modeling protein sequences.

3.2 ARTICLE

How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins?

Matteo Figliuzzi,[1] Pierre Barrat-Charlaix,[1] and Martin Weigt*,[1]

[1]Sorbonne Université, CNRS, Institut de Biologie Paris Seine, Computational and Quantitative Biology – UMR7238, 75005 Paris, France

***Corresponding author:** E-mail: martin.weigt@upmc.fr.

Associate editor: Claus Wilke

Abstract

Global coevolutionary models of homologous protein families, as constructed by direct coupling analysis (DCA), have recently gained popularity in particular due to their capacity to accurately predict residue–residue contacts from sequence information alone, and thereby to facilitate tertiary and quaternary protein structure prediction. More recently, they have also been used to predict fitness effects of amino-acid substitutions in proteins, and to predict evolutionary conserved protein–protein interactions. These models are based on two currently unjustified hypotheses: 1) correlations in the amino-acid usage of different positions are resulting collectively from networks of direct couplings; and 2) pairwise couplings are sufficient to capture the amino-acid variability. Here, we propose a highly precise inference scheme based on Boltzmann-machine learning, which allows us to systematically address these hypotheses. We show how correlations are built up in a highly collective way by a large number of coupling paths, which are based on the proteins three-dimensional structure. We further find that pairwise coevolutionary models capture the collective residue variability across homologous proteins even for quantities which are not imposed by the inference procedure, like three-residue correlations, the clustered structure of protein families in sequence space or the sequence distances between homologs. These findings strongly suggest that pairwise coevolutionary models are actually sufficient to accurately capture the residue variability in homologous protein families.

Key words: **coevolution, direct coupling analysis, global statistical inference, Boltzmann machine learning.**

Article

Introduction

In the course of evolution, proteins may substitute the vast majority of their amino acids without losing their three-dimensional structure and their biological functionality. Rapidly growing sequence databases provide us with ample examples of such evolutionary related, that is, homologous proteins, frequently already classified into protein families and aligned into large multiple-sequence alignments (MSA). Typical pairwise sequence identities between homologous proteins go down to 20–30%, or even below (Finn et al. 2014). Such low sequence identities are astonishing since even very few random mutations may destabilize a protein or disrupt its functionality.

Assigning a newly sequenced gene or protein to one of these families helps us to infer functional annotations. Structural homology modeling, for example, belongs to the most powerful tools for protein-structure prediction (Arnold et al. 2006; Webb and Sali 2014). However, beyond the transfer of information, the variability of sequences across homologs itself contains information about evolutionary pressures acting in them, and statistical sequence models may unveil that information (Durbin et al. 1998; de Juan et al. 2013).

A first level of information is contained in the variability of individual residues: low variability, that is, conservation, frequently identifies functionally or structurally important sites in a protein. This information is used by so-called profile models (Durbin et al. 1998), which reproduce independently the amino-acid statistics in individual MSA columns. They belong to the most successful tools in bioinformatics; they are at the basis of most techniques for multiple-sequence alignment and homology detection, partially as profile Hidden-Markov models accounting also for amino-acid insertions and deletions (Eddy 1998).

A second level of information is contained in the covariation between pairs of residues, measurable via the correlated amino-acid usage in pairs of MSA columns (de Juan et al. 2013; Cocco et al. 2017). Covariation cannot be captured by profile models, as they treat residues independently. To overcome this limitation, global statistical models with pairwise couplings—exploiting residue conservation and covariation—have recently become popular. Methods like the direct coupling analysis (DCA) (Weigt et al. 2009; Morcos et al. 2011), PsiCov (Jones et al. 2012), or Gremlin (Balakrishnan et al. 2011) allow for the prediction of residue–residue contacts using sequence information alone, and can be used to predict three-dimensional protein structures (Marks et al. 2012; Ovchinnikov et al. 2017) and to assemble protein complexes (Schug et al. 2009; Hopf et al. 2014; Ovchinnikov et al. 2014). Currently, these methods are the central element of various of the best-performing residue-contact predictors in the CASP competition for protein structure prediction (Jones et al. 2015; Wang et al. 2017).

Despite their success in practical applications, not much is known about the reasons for this success and their intrinsic limitations. Typically, two hypotheses are made: 1) The correlated amino-acid usage in two MSA columns may result from a direct residue–residue contact in the protein structure, causing coordinated amino-acid changes to maintain the protein's stability. It may also result indirectly via intermediate residues, making the direct use of covariation for contact prediction impractical. The success of global models is attributed to their capacity to extract direct couplings from indirect correlations. 2) Using the maximum-entropy principle, the simplest models reproducing pairwise residue covariation depend on statistical couplings between residue pairs. Whether or not this model is sufficient to capture also higher-order covariation remains currently unclear.

So far, these two points have not been investigated systematically. The reason is relatively simple: The inference of pairwise models exactly reproducing the empirical conservation and covariation statistics extracted from an MSA requires to sum over all 20^L sequences of aligned length L, an unfeasible task for sequences of typical sizes $L = 50–500$. Approximation schemes like mean-field approximation (Morcos et al. 2011), Gaussian approximation (Jones et al. 2012), or pseudo-likelihood maximization (Balakrishnan et al. 2011; Ekeberg et al. 2013) have been introduced; they perform very well in contact prediction. Their approximate character prohibits, however, the analysis of higher-order correlations and collective effects, since even the pairwise statistics are not well reproduced. More precise approaches have been proposed recently (Sutto et al. 2015; Barton et al. 2016; Haldane et al. 2016), but their high computational cost has limited applications mostly to anecdotal cases so far.

Understanding these basic questions is essential for understanding the success of global coevolutionary models beyond "black box" applications, but also for recognizing their current limitations and thus potentially to open a way towards improved statistical modeling schemes. To this end, we implement a highly precise approach for parameter inference in pairwise statistical models. Applying this approach to a number of very large protein families (containing sufficient sequences to reliably measure higher-order statistical features), we demonstrate that indirectly generated pair correlations are highly collective effects of entire networks of direct couplings, which are based on the structural vicinity between residues.

However, the most interesting finding of the article is the unexpected accuracy of DCA at reproducing higher-order statistical features, which are not fitted by our approach. These nonfitted features include connected three-point correlations, the distance distributions between natural sequences and between artificial sequences sampled from the model, or the clustered organization of sequences in sequence space. Currently we do not find indications, that more involved models (e.g., including three-residue interactions) are needed to reproduce the full sequence statistics: pairwise models are not only necessary as argued above, but seem to be sufficient to describe the sequence variability between homologous proteins.

Results

Direct Coupling Analysis—Methodology and Approximate Solutions

The aim of global coevolutionary sequence models as constructed by DCA is to provide a protein family-specific probability distribution

$$P(\underline{A}) \propto \exp\left(\sum_{j>i} J_{ij}(A_i, A_j) + \sum_{i=1}^{N} h_i(A_i)\right) \qquad (1)$$

for all full-length amino-acid sequences $\underline{A} = (A_1, \ldots, A_L)$. To model sequence variability in an MSA, couplings $J_{ij}(A, B)$ and biases (fields) $h_i(A)$ have to be fitted such that model $P(\underline{A})$ reproduces the empirically observed frequencies $f_i(A)$ of occurrence of amino acid A in the ith MSA column, and cooccurrence $f_{ij}(A, B)$ of amino acids A and B in positions i and j of the same sequence. In other words, the DCA model has to satisfy

$$P_i(A) = f_i(A) \quad \text{and} \quad P_{ij}(A, B) = f_{ij}(A, B) \qquad (2)$$

for all columns i, j and all amino acids A, B, with P_i and P_{ij} being marginal distributions of model $P(\underline{A})$, cf. supplementary methods and section 1, Supplementary Material online.

Equation (2) has two important consequences. First, a precisely inferred DCA model reproduces also pairwise connected correlations (or covariances) $c_{ij}(A, B) = f_{ij}(A, B) - f_i(A) f_j(B)$ found in the MSA. This is a crucial difference with profile models, which show vanishing connected correlations by construction. Second, the inference of $P(\underline{A})$ via equation (2) does *not* use all the information contained in the MSA, but only the pairwise statistics. For this reason, model $P(\underline{A})$ has *a priori* no reason to reproduce any higher-order statistics contained in the alignment. In particular, even though a model of the form of equation (1) will contain higher-order correlations, such as three-residue correlations, these may differ significantly from those found in the original MSA.

To infer DCA parameters, we need to estimate marginal probabilities for single positions and position pairs from model $P(\underline{A})$. Exact calculations of these marginals require to perform exponential sums over q^L terms, with L being the sequence length, and $q = 21$ enumerating amino acids and the alignment gap. These sums are infeasible even for short protein sequences, and have been replaced by approximate expressions, for example, via mean-field (Morcos et al. 2011), Gaussian (Jones et al. 2012), or pseudo-likelihood approximations (Balakrishnan et al. 2011; Ekeberg et al. 2013). These approximations are sufficiently accurate for residue-contact prediction, which is topological in nature: only the existence of a strong direct statistical coupling has to be detected, not necessarily its precise numerical value. As a consequence, these methods do not reproduce the empirical frequencies and thus do not satisfy equation (2), cf. figure 1 for pseudo-likelihood maximization (plmDCA [Ekeberg et al. 2013]). More precise methods based on an adaptive cluster expansion (Barton et al. 2016) or Boltzmann machine learning using Markov-chain Monte Carlo sampling (Ackley et al. 1985) for estimating marginal distributions have

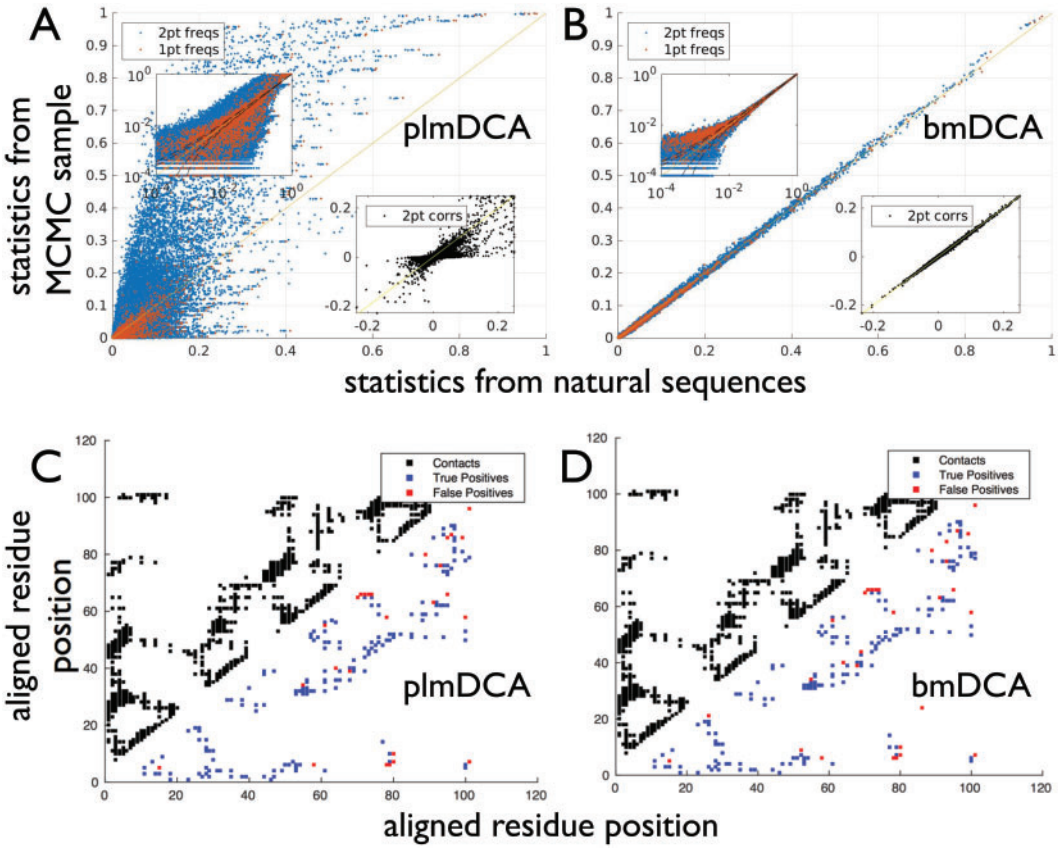

Fig. 1. Fitting accuracy and contact prediction for DCA models inferred using pseudo-likelihood maximization (plmDCA) and Boltzmann-machine learning (bmDCA): While the Potts model inferred by plmDCA (*A*) fails to reproduce the one- and two-residue frequencies (main panel) and the connected two-point correlations (lower left insert) in the PF00072 protein family, the model inferred using our bmDCA algorithm (*B*) is very accurate. Slight deviations visible for very small frequencies in log-scale (upper right insert) are results of the ℓ_2-regularization penalizing strongly negative couplings. Despite these differences, the contact predictions (*C* for plmDCA and *D* for bmDCA) relying on the strongest $2L = 224$ DCA couplings (with $|i - j| > 4$) are close to identical: native contacts (all-atom distance below 8 Å) are shown above the diagonal, predicted contacts (below the diagonal). Very similar results are observed across all studied protein families, cf. supplementary sections 5.1 and 5.5, Supplementary Material online (color online).

been proposed recently (Sutto et al. 2015; Haldane et al. 2016). While decreasing deviations from equation (2) substantially (i.e., fitting quality), they are typically much more computationally expensive and not suitable for large-scale application to hundreds or thousands of protein families.

Accurate Fitting Is Needed to Reproduce the Empirical Residue Covariation in Homologous Protein Families

Since the aim of the current article is to unveil the way DCA disentangles direct couplings and indirect correlations, and to investigate if it captures higher-order statistical observables estimated from the MSA, we have implemented an efficient version of Boltzmann machine (BM) learning described in "Materials and Methods" and, in full detail, in supplementary section 2, Supplementary Material online. In short, BM learning estimates the pairwise marginal distributions of $P(\underline{A})$ by Monte-Carlo sampling, and iteratively updates model parameters until equation (2) is satisfied. In contrast to approximations such as applied in plmDCA, the inference of parameters using BM learning can be made arbitrarily accurate, provided that Monte-Carlo samples are large enough and sufficient iterations are performed. In analogy to earlier notation, we will use bmDCA for the resulting implementation of DCA. As is shown in figure 1 and in table 1, bmDCA reaches very accurate fitting, approaching the statistical uncertainties related to the finite sample size (i.e., the sequence number in each MSA), even for the very large protein families studied here. Obviously bmDCA has a higher computational cost than plmDCA: While plmDCA achieves inference typically in few minutes, bmDCA needs few hours to several days for one family, in dependence of the sequence length and the required fitting accuracy.

Interestingly, the increased fitting accuracy does not improve the contact prediction beyond the one of plmDCA, the currently best unsupervised DCA contact predictor, cf. figure 1C and *D*. Couplings $J_{ij}(A, B)$ are highly correlated between PLM and BM (Pearson correlations of 90–98% across all studied protein families), in particular large couplings are robust and lead to very similar contact predictions. However, the model statistics depends collectively on all $\mathcal{O}(q^2L^2)$ parameters and can thus differ substantially even for small differences in the individual parameters. This sensitivity (so-called criticality) has also been observed in other models inferred from large-scale biological data, cf. (Mora and Bialek 2011).

Table 1. Results for the Ten Selected Protein Families.

Protein Family				Fitting Quality		Contact Prediction		Three-Point Correlations		Collectivity of Correlations		
Pfam	*L*	*M*	PDB	PLM	BM	PLM	BM	PLM	BM	corr(DI, MI)	corr(L2I, MI)	ν
PF00004	132	39277	4D81	0.630	0.954	0.672	0.672	0.333	0.980	0.33	0.42	1.2
PF00005	137	68891	1L7V	0.546	0.948	0.599	0.586	0.718	0.978	0.51	0.65	1.4
PF00041	85	42721	3UP1	0.897	0.973	0.715	0.671	0.893	0.991	0.61	0.77	1.7
PF00072	112	73063	3ILH	0.670	0.978	0.836	0.842	0.803	0.988	0.52	0.69	1.4
PF00076	69	51964	2CQD	0.868	0.977	0.877	0.833	0.963	0.993	0.53	0.72	1.5
PF00096	23	38996	2LVH	0.954	0.987	0.657	0.711	ND	ND	0.95	0.99	2.3
PF00153	97	54582	2LCK	0.800	0.967	0.601	0.563	0.517	0.986	0.45	0.57	1.1
PF01535	31	60101	4G23	0.902	0.994	0.630	0.739	0.120	0.996	0.70	0.91	1.5
PF02518	111	80714	3G7E	0.624	0.970	0.423	0.396	−0.228	0.986	0.47	0.60	1.6
PF07679	90	36141	1FHG	0.823	0.955	0.826	0.826	0.797	0.993	0.48	0.58	1.8

NOTE.—The first four columns give the ID of the selected protein families together with the sequence length L, alignment depth M and a representative protein structure. The fitting quality measures the Pearson correlation between connected two-point correlations in the natural data, and in a sample drawn from the Potts models inferred by plmDCA and bmDCA (better quality emphasized in boldface). The contact prediction gives the fraction of true positives (all-atom distance < 8 Å) within the first 2 L predictions. Columns 9 and 10 provide the Pearson correlation between connected three-point correlations observed in natural and in sampled sequences (due to the dominance of insignificantly small terms, only those with $c^{MSA}_{ijk}(A,B,C) > 0.01$ are considered). PF00096, with only 23 aligned positions is the shortest considered protein family, has no significant three-point correlations, neither in the data nor in the Potts model. The last three columns quantify the collective nature of correlations: the Pearson correlation of direct information/mutual information as compared to the length-two information/mutual information, and the exponent of the approximate power-law decay of the strongest paths (in terms of their path information) with their ranking.

Indirect Correlations Result Collectively from Networks of Direct Couplings

bmDCA provides a highly accurate approach to describe the sequence variability of homologous proteins via a pairwise coevolutionary model. This implementation allows us to ask fundamental questions about how DCA works, its capacities and its possible limitations, without being biased by the specificities of approximate DCA implementations.

The success of global models as inferred by DCA is typically attributed to the idea that they disentangle statistical correlations, which are empirically observed in an MSA and measured via the mutual information (MI), into a network of direct couplings between residues. The strongest direct couplings are biologically interpretable as residue–residue contacts in the three-dimensional protein structure. However, this idea, even if stated in many papers on the subject, has never been examined in detail, and important questions remain unanswered: can indirect effects be explained by a few strong coupling chains, or are they distributed over networks of numerous small couplings? Are these networks structurally interpretable, that is, in relation to a proteins contact map?

Correlations Are Mediated Collectively by Distributed Networks of Coevolutionary Couplings

To answer the first question, we need to quantify the correlation induced by a coupling chain of arbitrary length, connecting any two residues. To this aim, we take inspiration from the concept of direct information (DI) introduced in (Weigt et al. 2009). DI is a proxy of the strength of the direct interaction J_{ij} between two residue positions *i* and *j*; it measures the correlation that *i* and *j* would have if they were only connected by J_{ij}, cf. figure 2A. To measure the indirect correlation between *i* and *j* induced via a chain of intermediate residues, we introduce the concept of path information (PI), as illustrated again in figure 2A and defined in "Materials and Methods". Now, for each protein family, we extracted the 100 most correlated residue pairs (highest MI). Using a modification of Dijkstra's shortest-path algorithm (Dijkstra 1959)—which becomes approximate due to the nonadditivity of PI but delivers highly reliable results as shown in supplementary section 3, Supplementary Material online—we extracted for each residues pairs the 15 strongest coupling paths (highest PI) connecting the two residues.

In figure 2B, we show that the decrease of the average strength of the *k*th strongest path is compatible with a slowly decreasing power law, $\langle PI(k)\rangle \propto k^{-\nu}$, with exponents ν between 1.1 and 2.3. While this fit is only approximate, as visible by the strong deviations for the strongest path at $k = 1$, its slow decay clearly shows that the correlation between two residues typically is not mediated by one or few coupling chains. On the contrary, indirect effects emerge collectively, in the sense that a large number of partially overlapping coupling chains have to be taken into account, each one contributing only a small fraction to the total correlation. It is important to note that the strongest path (rank $k = 1$) is on average much stronger than the others and clearly does not fall onto a power law. For the overwhelming majority of the pairs, this strongest path is the direct one containing only one coupling. Its contribution to the total correlation is, on average, about 12.5% of the total MI. This average is dominated by the shortest protein families, PF00096 and PF01535, who are expected to show less collectivity due to their small number *L* of aligned residues.

On the Structural Basis of Coevolutionary Coupling Networks

As a consequence of the last section, we need to consider the collective effect of multiple paths rather than trying to biologically interpret individual paths beyond the direct one. While this is technically very hard in general, the collective effect of all paths of length two is efficiently computable, cf. "Materials and Methods." The corresponding correlation measure, named here length-two information (L2I) and illustrated in figure 2A, adds the $L - 2$ possible indirect paths of

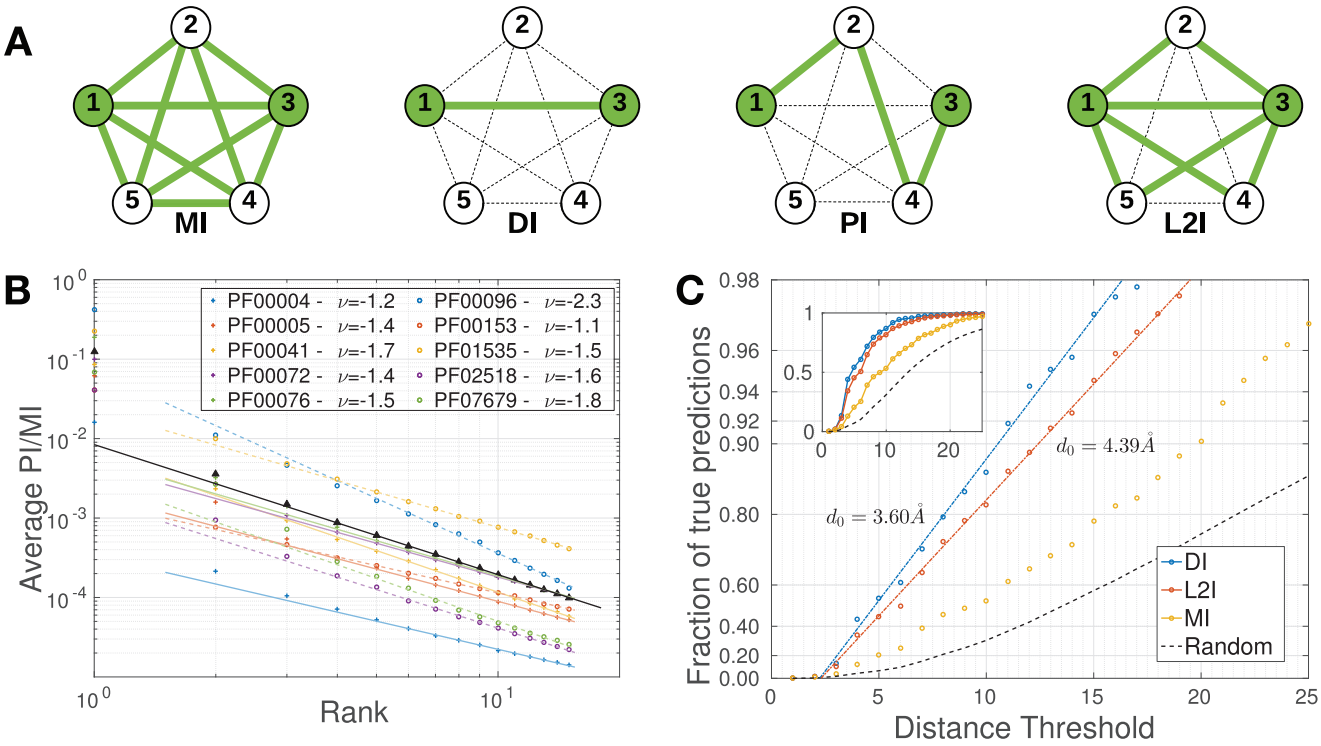

FIG. 2. Collective nature of the correlation between two residue positions: (*A*) Illustration of the correlation measures used in this study. While the mutual information MI depends collectively on the entire network of coevolutionary couplings, the direct information DI is obtained by taking into account only the single direct coupling between the sites of interest (e.g., 1 and 3 in the figure). All other couplings are formally set to zero. The path information PI is the direct generalization of DI to the correlation mediated by a single path (e.g., [1, 2, 4, 3] in the figure). The length-two information L2I measures the collective effect of the direct coupling and all length-two paths (e.g., [1, *k*, 3] with $k = 2, 4, 5$). (*B*) A log–log plot of the average ratio of path information to mutual information (triangular symbols and black fat line: average over all families) as a function of the rank of the corresponding path, showing a very slow (approximately power-law) decay. This illustrates the fact that indirect correlations do not depend on a single (or very few) coupling chains, but are distributed over coupling networks. (*C*) For the 25 highest ranking residue pairs according to DI, L2I, and MI, the fraction of pairs of distance below *d*, as a function of *d*. The scale on the *y*-axis is logarithmic, and chosen in a way that functions of the form $1 - e^{-d/d_0}$ will appear as straight lines, the insert shows a standard linear scale. For DI and PI, these curves show a clear exponential convergence to 1, with characteristic distance scales of 3.6 resp. 4.4 Å. MI does not show any exponential behavior, and thus no characteristic distance scale (color online).

length 2 (one intermediate residue) to the direct path between two residue positions. As expected, L2I captures a much higher fraction of the full mutual information than DI, cf. table 1. However, a large fraction of the mutual information is not yet covered. It is contributed by longer coupling chains: L2I depends only on $2L - 3$ out of the $L(L-1)/2$ couplings between residue pairs. Consistent with this observation, the correlation of L2I with MI is much larger in small proteins, and decreases when going to larger proteins.

L2I allows for an interesting structural interpretation. It is well established that large DI are good predictors for native residue contacts. Is large L2I a good predictor of second neighbors in the protein structure, that is, of residue pairs which are two contacts away? To investigate this question, the blue line in figure 2C displays the fraction of true positive predictions (positive predictive value, averaged over the protein ensemble) within the highest 25 DI as a function of a distance cutoff *d*, which varies between 1 and 25 Å. It starts at 0 for small *d*, and approaches 1 exponentially with a scaling $1 - \exp(-d/d_0)$ of characteristic length $d_0 = 3.6$ Å. At 8 Å distance (typically used as contact definition in DCA studies), an accurate prediction of about 85% true positives (TP) and only 15% false positives (FP) is reached. Measuring the cut-off dependent positive predictive value for the length-two information L2I, we find again an exponential behavior but with characteristic length $d_0 = 4.4$ Å. The fraction of TP therefore reaches 85% only between 11 and 12 Å, a distance compatible with second structural neighbors. The finding that the top DI are dominated by direct contacts, and large L2I by residue pairs which are up to second neighbors in the structure, further underlines the structural basis of coevolutionary constraints as captured by DCA. We also note that the full correlation MI—depending on coupling chains of all possible lengths—does not imply an exponential behavior in figure 2, and no characteristic length scale can be identified.

Pairwise Coevolutionary Models Accurately Reproduce the Residue Variability beyond the Fitted Two-Residue Statistics

Profile models assuming independent residues are not able to extract the full information contained in the MSA of a protein family. In particular, the inclusion of pairwise coevolutionary

couplings is required for the prediction of intra or interprotein residue–residue contacts, which has become the most important application of coevolutionary modeling. Furthermore, studies about protein mutational effects (Levy et al. 2017) and the prediction of protein–protein interactions (Szurmant and Weigt 2018) have underlined the importance of pairwise couplings.

Is there information hidden in large MSA, which cannot be captured by pairwise models? Does one need to include higher-order couplings into the modeling? The highly accurate inference of pairwise models obtained by bmDCA, reproducing faithfully the empirical first- and second-order statistics, allows to address these questions systematically. To this aim, we use MCMC samples from the inferred models to *compare statistical observables, which are not a direct consequence of the fitted covariances*. These comparisons unveil the astonishing capacity of bmDCA to capture local and global statistical features, which are not explicitly fitted by the model: pairwise couplings are not only necessary for characterizing sequence variability between homologs, but they also seem to be sufficient.

First, we observe that the *three-residue statistics* is accurately reproduced by our model including only pairwise couplings: figure 3 (cf. supplementary section 5.2, Supplementary Material online, for other families) shows a density-colored scatter plot of the connected three-point correlations of the natural sequences versus the MCMC sample drawn from the model. Correlations are high across all protein families for the pairwise model, with close to perfect Pearson correlations ranging from 0.978 to 0.997, cf. table 1. Profile models, which by definition do not have any connected three-point correlation, can be seen as null model testing the strength of three-point correlations emerging due to finite sampling. As is shown in figure 3D, they are at least one order of magnitude smaller than those found empirically, underlining the significance of our findings. The only exception is family PF00096, where no significant connected three-point correlations are detectable in the MSA or in the sample. Note that we use connected correlations $c_{ijk}(A,B,C) = f_{ijk}(A,B,C) - f_{ij}(A,B) f_k(C) - f_{ik}(A,C) f_j(B) - f_{jk}(B,C) f_i(A) + 2 f_i(A) f_j(B) f_k(C)$, which are intrinsically harder to reproduce than three-point frequencies $f_{ijk}(A,B,C)$. Note also that our result is far from being obvious: a Gaussian model with the same covariances would have vanishing three-point correlations, while the sequence data and the sample from our DCA model do not. Further more, it is easy to construct models with discrete variables, whose three-point correlations are not reproduced by a pairwise DCA model. This is shown in supplementary section 4, Supplementary Material online, via analytical calculations and numerical simulations.

To complement the three-point statistics, we investigated more global quantities. The first one is the clustered organization of protein families in sequence space. Figure 3A shows all sequences mapped onto their first two principal components for PF00072 (cf. Supplementary Material online for other families). We observe a clear clustering into at least three distinct subfamilies, which identify different functional subclasses of the PF00072 protein family (single domain vs. multi-domain architectures with distinct DNA-binding domains). A sample drawn from a profile model does not reproduce this clustered structure (B), while the MCMC sample of the bmDCA model does, including the fine structure of the clusters (C). Again, this structure is not a simple consequence of the empirical covariance matrix as a sample from a Gaussian model with the same covariances would not show any clustering.

As a last measure, we compared the pairwise Hamming distances between sequences in the natural MSA and in the model-generated sequences. Again the pairwise bmDCA model is needed to reproduce the bulk of the empirical distribution of pair distances. Interestingly, a difference between the two becomes visible in the small-distance tail of the histograms in figure 3G: while natural sequences may be close to identical due to a close phylogenetic relation, small sequence distances are never observed in an equilibrium sample of the bmDCA model, that is, a part of the phylogenetic bias present in the MSA is avoided by the bmDCA model.

Discussion

This article unveils a number of reasons behind the success of global pairwise models in extracting information from the sequence variability of homologous protein sequences. First, we show that residue–residue correlations actually result from the collective variability of many residues, and are not the result of a few strong coupling chains. Therefore, local statistical measures taking into account only a small numbers of residues at a time (like correlation measures) are necessarily limited in their capacity to represent the data, and global modeling approaches are needed.

One of the most astonishing findings is that many features of the data, which are not explicitly fitted by a pairwise modeling, are nevertheless well reproduced by the inferred models. This includes higher-order correlations, like the connected three-point correlations considered here, and more general aspects of the distribution of amino-acid sequences like the histogram of pairwise Hamming distances between pairs of sequences or the clustered organization of the sample in sequence space. Interestingly, only the small distances between phylogenetically closely related sequences are not reproduced in a sample drawn from the inferred DCA model. This capacity to reproduce the sequence variability beyond the fitted empirical observables distinguishes the DCA model (fitting one- and two-residue frequencies) from profile models of independent residues (fitting only one-residue frequencies). While the restriction to pairwise models was initially motivated by the limited availability of sequence data—three-point correlations require to estimate frequencies for $21^3 = 9261$ combinations of amino acids or gaps—we find that even for large MSA pairwise models seem to be sufficient to capture collective effects beyond residue pairs.

Note that this argument does not rule out the existence of higher-order residue effects in the underlying evolutionary processes shaping the sequence variability in homologous protein families (cf. Merchan and Nemenman 2016; Schmidt and Hamacher 2017). However, their statistical

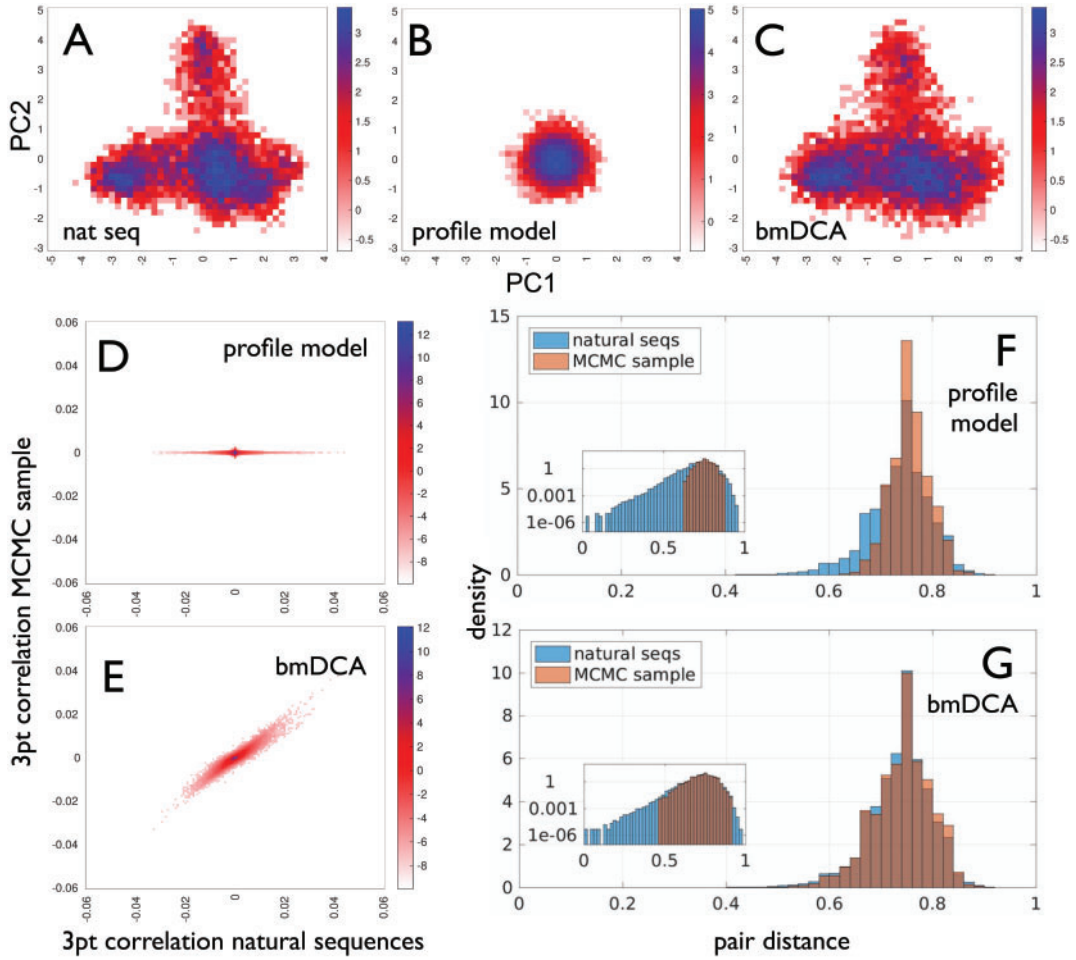

FIG. 3. Nonfitted statistical observables are captured by DCA: (*A*–*C*) Natural sequences (PF00072—*A*) and MCMC samples from inferred profile (*B*) and bmDCA (*C*) models are projected on the first two principal components of the natural MSA. (*D* and *E*) Three-point correlations of samples of the profile (*D*) and bmDCA (*E*) models, as compared to the three-point correlations in the natural sequences. (*F* and *G*) Histograms of all pairwise Hamming distances between natural or MCMC sampled sequences, for profile (*F*) and bmDCA (*G*) models. Surprisingly bmDCA is able to reproduce all three nonfitted statistical properties of the natural MSA, with the difference of the small distances between close homologs, while the profile model not taking into account residue–residue couplings does not. This suggests that accurately inferred pairwise models are necessary and sufficient to capture the residue variability in families of homologous proteins. Similar results are observed across all studied protein families, as is documented in supplementary sections 5.2–5.4, Supplementary Material online (color online).

signature is not strong enough to be detectable via deviations from the behavior of a pairwise model, even in the large families considered here. Random samples drawn from a DCA model based exclusively on the knowledge of the empirical one- and two-residue statistics appear to be statistically indistinguishable from natural sequences.

This finding is particularly interesting in the context of work made few years ago by the Ranganathan lab (Russ et al. 2005; Socolich et al. 2005). Using the small WW domain, they applied a number of diverse procedures to scramble MSA of natural sequences to produce artificial sequences. Scrambling MSA columns to maintain residue conservation while destroying residue correlations, lead in all tested cases to nonfolding amino-acid sequences. A procedure maintaining also pairwise correlations lead to a substantial fraction of folding and functional proteins. Later on it has been observed that the functional artificial sequences actually have the highest probabilities within pairwise coevolutionary models (Balakrishnan et al. 2011). These findings open interesting roads to evolution-guided protein design (Reynolds et al. 2013).

Note, however, that the finite size of the input MSA requires to use regularized inference, which penalizes large absolute parameter values. It leads to a small bias visible in figure 1B: small pair frequencies are slightly but systematically overestimated by DCA. This may smoothen the inferred statistical model, cf. (Otwinowski and Plotkin 2014) for the related case of inferring epistatic fitness landscapes. As a consequence "bad" sequences may be given high probabilities in our model. Based on the findings presented in figures 1B and 3, we expect these effects to be minor. When increasing the regularization strength beyond parameters used in this study, the clustered structure of sampled sequences (fig. 3C) disappears gradually. Data in large MSA allow to use small regularization, thereby simultaneously limiting overfitting of statistical noise and reducing biases in parameter inference.

This may be impossible for small MSA, so the ongoing growth of sequence databases is key for the wide applicability of global statistical sequence models.

One potentially important limitation remains: the distribution of sequences in sequence space is not only determined by functional constraints acting on amino-acid sequences, but also by phylogenetic relations between sequences. Natural sequences are, even beyond the very closely related sequences not reproduced by the DCA model, far from being an independent sample of all possible amino-acid sequences. They are correlated due to finite divergence times between homologs, and due to the human selection bias in sequenced species. Any model reproducing the full empirical statistics of the MSA describes therefore a mixture of functional and phylogenetic correlations, while an ideal model would contain the functional ones and discard the phylogenetic ones. How these can be disentangled remains an important open question.

Materials and Methods

Protein Families

We have selected ten protein families of known three-dimensional structure which belong to the largest 20 Pfam families (Finn et al. 2014), which are not repeat proteins (i.e., they are not just frequent because repeated many times on the same protein), and have an aligned sequence length below 200 amino acids (for computational reasons), cf. table 1. Sequences with more than 50 alignment gaps are removed. The resulting sequence numbers are reported in table 1. The main reason to include only large Pfam families is the possibility to accurately estimate three-point correlations. For each triplet of residue positions, there are $21^3 = 9{,}261$ combinations of amino acids or gaps. Nonsystematic tests in smaller protein families show that our main findings of the paper translate directly to these families.

Boltzmann Machine Learning

DCA infers a Potts model

$$P(A_1,\ldots,A_L) = \frac{1}{Z}\exp\left\{\sum_{i<j} J_{ij}(A_i,A_j) + \sum_i h_i(A_i)\right\} \quad (3)$$

reproducing the single- and two-residue frequencies found in the input MSA:

$$\begin{aligned}\textstyle\sum_{\{A_k|k\neq i\}} P(A_1,\ldots,A_L) &= f_i(A_i)\\ \textstyle\sum_{\{A_k|k\neq i,j\}} P(A_1,\ldots,A_L) &= f_{ij}(A_i,A_j)\end{aligned} \quad (4)$$

with empirical frequencies $f_i(A_i)$ and $f_{ij}(A_i,A_j)$ defined, respectively, as the fraction of sequences in the MSA having amino acid A_i (resp. A_i and A_j) in column i (resp. in columns i and j) (cf. supplementary section 1, Supplementary Material online for a precise definition of these frequency counts, including a sequence weighting to reduce phylogenetic biases). For the sake of contact prediction, this inference can be done with efficient approximation schemes like mean-field of pseudo-likelihood maximization. The objectives of this study—to understand the collective variability of the residues—require a more precise inference based on the classical ideas of Boltzmann-machine learning (Ackley et al. 1985). It consists of an iterative procedure where

(i) for a given set of model parameters $\{J_{ij}, h_i\}$, Markov-chain Monte Carlo (MCMC) sampling is used to estimate the one- and two-point frequencies of the model;

(ii) parameters are adjusted when the estimated model frequencies deviate from the empirical ones.

To reduce finite-sample effects, the model parameters are subject to an ℓ_2-regularization. The likelihood function is convex, guaranteeing convergence to a single globally optimal solution, which reproduces the empirical one- and two-point frequencies with arbitrary accuracy. The direct implementation of Boltzmann-machine learning is computationally very slow. We have therefore introduced a reparameterization of the model, which allows to replace the gradient ascent of the likelihood by a faster pseudo-Newtonian method. Technical details of the implementation are described in supplementary section 2, Supplementary Material online.

From Direct Couplings to Indirect Correlations

Quantifying the Strength of a Coupling Chain

To quantify the strength of a coupling chain, we generalize the direct information introduced in (Weigt et al. 2009). There, the direct probability

$$\begin{aligned}&P_{ij}^{dir}(A_i,A_j) =\\ &\exp\left\{J_{ij}(A_i,A_j) + \tilde{h}_i(A_i) + \tilde{h}_j(A_j)\right\}/Z_{ij}\,.\end{aligned} \quad (5)$$

was defined as the hypothetical distribution of two residues i and j connected only by the inferred direct coupling J_{ij} and having the empirical single-residue frequencies $f_i(A_i)$ and $f_j(A_j)$, thereby removing all indirect effects from model P. Parameters $\tilde{h}_i$ and $\tilde{h}_j$ are to be adjusted to ensure correct marginals. The *path probability* between positions i_1 and i_{L+1} through the length-L path $[i_1, i_2 \ldots i_{L+1}]$ is a direct generalization:

$$P^{path}_{[i_1\ldots i_{L+1}]}(A_{i_1},A_{i_{L+1}}) = \textstyle\sum_{\{A_{i_2}\ldots A_{i_L}\}} f_{i_1}(A_{i_1}) \prod_{l=1}^{L} P^{dir}_{i_{l+1}i_l}(A_{i_{l+1}}|A_{i_l})\,, \quad (6)$$

with $P^{dir}_{ij}(A_i|A_j) = P^{dir}_{ij}(A_i,A_j)/f_j(A_j)$. Equation (6) contains the product of direct probabilities for all links in the path, in analogy to a Markov chain. The sum over all configurations taken by intermediate sites $[i_2 \ldots i_L]$ is performed efficiently by dynamic programming; the definition guarantees the empirical marginals in all sites on the path.

To measure the correlation mediated by direct links or indirect paths, we use variants of the mutual information based on the direct and path probabilities. To this aim, we define the *direct information* (DI) as

$$DI_{ij} = \sum_{A_i,A_j=1}^{q} P_{ij}^{dir}(A_i,A_j)\log\frac{P_{ij}^{dir}(A_i,A_j)}{f_i(A_i)f_j(A_j)}, \quad (7)$$

and the *path information* (PI) as

$$PI_{[i\ldots j]} = \sum_{A_i,A_j=1}^{q} P_{[i\ldots j]}^{path}(A_i,A_j)\log\frac{P_{[i\ldots j]}^{path}(A_i,A_j)}{f_i(A_i)f_j(A_j)}. \quad (8)$$

The full *mutual information* (MI) is defined by replacing P^{dir} or P^{path} by f_{ij}.

The Joint Effect of Paths of Length 2

Quantifying the strength of a group of indirect effects between two sites i and j is in general non trivial. However, it is possible if one only considers all chains of couplings that go through at most one intermediary site k. In other words, one can combine the direct path $[ij]$ and all the chains of the form $[ikj]$ $(k \neq i,j)$ into a single probability distribution:

$$P_{ij}^{L2}(A_i,A_j) \propto \frac{P^{dir}(A_i,A_j)}{z_i(A_i)z_j(A_j)} \cdot \prod_{k\neq i,j} P_{[ikj]}^{path}(A_i,A_j), \quad (9)$$

where z_i and z_j ensure P_{ij}^{L2} to have marginals f_i and f_j. The path probabilities can be simply multiplied since each intermediate residue k appears only once, and they become conditionally independent for given (A_i, A_j). The correlation resulting from this combination of paths is the mutual information of P_{ij}^{L2}, called *L2I*.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online. Code and raw data can be accessed via Github (https://github.com/matteofigliuzzi/bmDCA; last accessed January 2018).

Acknowledgments

M.W. acknowledges funding by the ANR project COEVSTAT (ANR-13-BS04-0012-01), and by the European Union's H2020 research and innovation programme MSCA-RISE-2016 under Grant Agreement No. 734439 INFERNET. This study was undertaken partially in the framework of CalSimLab, supported by the grant ANR-11-LABX-0037-01 as part of the "Investissements d'Avenir" program (ANR-11-IDEX-0004-02).

References

Ackley DH, Hinton GE, Sejnowski TJ. 1985. A learning algorithm for Boltzmann machines. *Cogn Sci*. 9(1): 147–169.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The swiss-model workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22(2): 195–201.

Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. 2011. Learning generative models for protein fold families. *Proteins* 79(4): 1061–1078.

Barton JP, De Leonardis E, Coucke A, Cocco S. 2016. Ace: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics* 32(20): 3089–3097.

Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. 2017. Inverse statistical physics of protein sequences: a key issues review. *Rep Prog Phys*. 81:3.

de Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. *Nat Rev Genet*. 14(4): 249–261.

Dijkstra EW. 1959. A note on two problems in connexion with graphs. *Numer Math*. 1(1): 269–271.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press.

Eddy SR. 1998. Profile Hidden-Markov models. *Bioinformatics* 14(9): 755–763.

Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys Rev E* 87(1): 012707.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res*. 42(D1): D222–D230.

Haldane A, Flynn WF, He P, Vijayan R, Levy RM. 2016. Structural propensities of kinase family proteins from a potts model of residue co-variation. *Protein Sci*. 25(8): 1378–1384.

Hopf TA, Schärfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS. 2014. Sequence co-evolution gives 3d contacts and structures of protein complexes. *Elife* 3:e03430.

Jones DT, Buchan DW, Cozzetto D, Pontil M. 2012. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2): 184–190.

Jones DT, Singh T, Kosciolek T, Tetchner S. 2015. Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* 31(7): 999–1006.

Levy RM, Haldane A, Flynn WF. 2017. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol*. 43:55–62.

Marks DS, Hopf TA, Sander C. 2012. Protein structure prediction from sequence variation. *Nat Biotechnol*. 30(11): 1072–1080.

Merchan L, Nemenman I. 2016. On the sufficiency of pairwise interactions in maximum entropy models of networks. *J Stat Phys*. 162(5): 1294–1308.

Mora T, Bialek W. 2011. Are biological systems poised at criticality?. *J Stat Phys*. 144(2): 268–302.

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA*. 108(49): E1293–E1301.

Otwinowski J, Plotkin JB. 2014. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc Natl Acad Sci USA*. 111(22): E2301–E2309.

Ovchinnikov S, Kamisetty H, Baker D. 2014. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife* 3:e02030.

Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. 2017. Protein structure determination using metagenome sequence data. *Science* 355(6322): 294–298.

Reynolds KA, Russ WP, Socolich M, Ranganathan R. 2013. Evolution-based design of proteins. *Methods Enzymol*. 523:213–235.

Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. 2005. Natural-like function in artificial ww domains. *Nature* 437(7058): 579–583.

Schmidt M, Hamacher K. 2017. Three-body interactions improve contact prediction within direct-coupling analysis. *Phys Rev E* 96(5): 052405.

Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. 2009. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc Natl Acad Sci USA*. 106(52): 22124–22129.

Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. 2005. Evolutionary information for specifying a protein fold. *Nature* 437(7058): 512–518.

Sutto L, Marsili S, Valencia A, Gervasio FL. 2015. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci USA*. 112(44): 13567–13572.

Szurmant H, Weigt M. 2018. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr Opin Struct Biol*. 50:26–32.

Wang S, Sun S, Li Z, Zhang R, Xu J. 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol*. 13(1): e1005324.

Webb B, Sali A. 2014. Protein structure modeling with MODELLER. *Methods Mol Biol*. 1137:1–15.

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci USA*. 106(1): 67–72.

3.3 ARTIFICIAL SEQUENCES: THE EFFECT OF REGULARIZATION

The above article showed that an accurately inferred DCA model is a surprisingly good statistical model of the Multiple Sequence Alignment. The Potts model by construction reproduces conservation and two-point correlation patterns found in the MSA, but also higher order quantities such as three-body correlations, distribution of sequences in a principal component space, and distribution of hamming distances between sequences.

It is interesting to put this result in the context of figure 2.6 of section 2.5.2. There, a PLM-inferred DCA model was able to predict functionality of artificially designed sequences of the WW domain. Natural WW sequences were all located in the lower tail of the energy-distribution of the inferred Hamiltonian. Artificial sequences designed by conserving correlation patterns (CC dataset) are not all folding, and have a higher average energy according to the Hamiltonian. However, the ones that do fold are seen to be the ones of lowest energy. Lastly, sequences designed using the column-wise conservation profile never fold correctly and have higher energies. As a result, all folding sequences are in the lower tail of the DCA energy distribution, while high energy sequences are never folding.

It seems that in addition to be a good statistical model of sequences, DCA provides an energy function able to discriminate between folding and non-folding sequences. This makes the idea of designing artificial protein sequences using a Potts model very promising. A sample from the DCA distribution would have similar statistical properties than natural sequences, and could also contain functional sequences recognizable by statistical energies similar than the natural ones. However, this simple idea cannot be straightforwardly implemented due to the way the model is inferred. Figure 3.1 shows the energy distributions of different sets of sequences using the Hamiltonian of the BML-DCA model inferred on PF00072 (family used as an example in the article above). Shown sequences include a randomized alignment, samples from a profile model and from the inferred DCA model, and the homologs found in the Pfam MSA. As expected, sequences from the Potts distribution have lower energies than those coming from the profile model. However, it is also visible that the natural sequences have lower energies than the DCA ones. This is surprising, as energies are measured using the Hamiltonian of the DCA model which was trained on the natural sequences.

This observation can also be made, though in a less clear manner, in figure 2.6. There, it can be seen that the folding sequences – red bars on the bottom plot – are almost all located in the lowest part of the DCA energy spectrum, instead of being spread over all its width. It seems that in order to be folding, sequences need to have statistical energies in the range of the *natural* sequences. However, the pairwise model

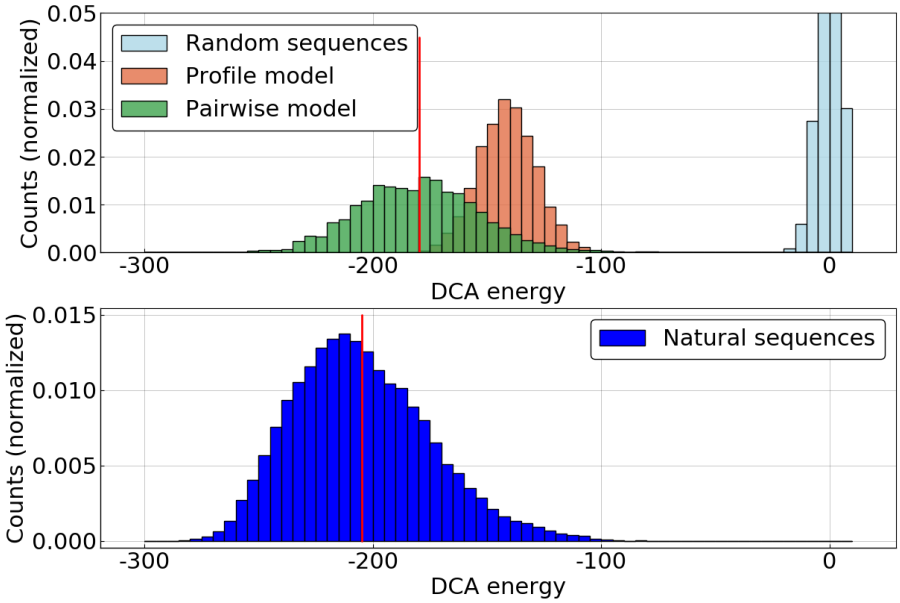

Figure 3.1: DCA energy distributions of different set of sequences. Top panel: Random sequences, and samples of a profile model and the DCA model. Bottom panel: natural sequences used to train the DCA model. Red lines indicate energy averages of the DCA sample and the natural sequences. The model was inferred on the PF00072 family.

generates sequences which have a higher energy than the natural ones. This raises a number of question. Since the model is inferred using the natural sequences, and since it precisely fits their statistical properties such as conservation and correlation, why does it generate sequences which on average have higher energies? One application of DCA could be to design artificial sequences. In this regard, would it be possible to generate sequences in a given energy range by changing the temperature of the model? If so, would these "low temperature" sequences still be similar to natural ones?

3.3.1 *Energy shift due to regularization*

In order to investigate these questions, the inference problem is put in a formal and well defined setting. The MSA of M homologous sequences of length L will be noted $\mathcal{D}_{nat} = \{a_i^m\}$, with $m = 1 \dots M$ and $i = 1 \dots L$. Let us assume that those sequences are an *i.i.d.* sample of a "true" distribution:

$$P^0(a_1 \dots a_L) = \frac{1}{Z^0} \exp\left(-\mathcal{H}^0(a_1 \dots a_L)\right). \tag{3.1}$$

The function $\mathcal{H}^0$ can be imagined as the "true" fitness of sequence $(a_1 \dots a_L)$. The existence of such a function is not necessary to explain the energy shift. However, it is useful to assume that it exists when

investigating the sampling at lower temperature of the DCA model. $\mathcal{H}^0$ could in principle have a very different mathematical form than the Potts Hamiltonian. However, we will show in the next sections that the observed energy shift can be explained simply due to regularization effects, whether $\mathcal{H}^0$ exists or not.
$\mathcal{D}_{nat}$ is used to infer the DCA model, which in practice is another Hamiltonian $\mathcal{H}^{inf}$ parametrized by coupling and field parameters $\boldsymbol{J}$ and $\boldsymbol{h}$. This Hamiltonian is numerically obtained by maximizing the regularized likelihood (see section 2.4.2 for the necessity of regularization):

$$\mathcal{L}(\mathcal{D}_{nat}|\boldsymbol{J},\boldsymbol{h}) = \log P(\mathcal{D}_{nat}|\boldsymbol{J},\boldsymbol{h}) - \lambda_J||\boldsymbol{J}||^2 - \lambda_J||\boldsymbol{h}||^2, \tag{3.2}$$

where $||\boldsymbol{J}|| = \sum_{1\leq i<j\leq L}\sum_{a,b} J_{ij}(a,b)^2$ and $||\boldsymbol{h}|| = \sum_{i=1}^{L}\sum_a h_i(a)^2$ Since Boltzmann Machine Learning is used, after a large enough number of iteration of the learning algorithm, the gradient of the likelihood is close to zero and the following relations stand:

$$\begin{aligned} f_i^0(a) - P_i^{inf}(a) &= \lambda_h h_i(a), \\ f_{ij}^0(a,b) - P_{ij}^{inf}(a,b) &= \lambda_J J_{ij}(a,b), \end{aligned} \tag{3.3}$$

where f_i^0 and f_{ij}^0 are the single site and pairwise frequencies measured in $\mathcal{D}_{nat}$, and P_i^{inf} and P_{ij}^{inf} the corresponding quantities for the inferred model. Equation (3.3) makes clear that the frequencies as measured in the data are only reproduced up to some precision which depends on the regularization. The resulting deviation in frequencies is quite small on the individual term – values used for regularization are typically $\lambda = 0.01$ –, and only has a visible effect for very small frequencies as can be seen in the figure 1 of the article above. However, it becomes a bias when computing energies. Indeed, when computing the average energy of natural sequences in the inferred Hamiltonian taking Eq. (3.3), one obtains

$$\begin{aligned} \langle \mathcal{H}^{inf} \rangle_{\mathcal{D}_{nat}} &= -\frac{1}{2}\sum_{i,j=1}^{L}\sum_{a,b=1}^{q} J_{ij}(a,b) f_{ij}^0(a,b) - \sum_{i=1}^{L}\sum_{a=1}^{q} h_i(a) f_i^0(a) \\ &= \langle \mathcal{H}^{inf} \rangle_{\mathcal{H}^{inf}} - \lambda_J||\boldsymbol{J}||^2 - \lambda_h||\boldsymbol{h}||^2 \end{aligned} \tag{3.4}$$

.

Therefore, in the inferred model, energies of natural sequences are systematically lower than energies of a sample of $\mathcal{H}^{inf}$. This systematic bias is due to regularization and can be quantified using equation 3.4: in the case of figure 3.1, the quantity $\lambda_J||\boldsymbol{J}||^2 + \lambda_h||\boldsymbol{h}||^2$ has a numerical value of 24.7, for a measured energy shift of ~ 25 (red lines in the figure).

3.3.2 *Sampling at lower temperature*

If one trusts the DCA Hamiltonian as a good proxy for the "functionality" of a sequence, then it would be interesting to generate sequences with energies matching the ones of members of the MSA. Since regularization shifts energies of a DCA sample upwards, one way to achieve this is to lower the temperature at which the sampling is performed. Instead of taking samples directly from the Potts distribution in Eq. (2.1), an inverse temperature parameters $\beta = 1/T$ is introduced, and the following distribution is sampled:

$$P_\beta^{inf}(a_1 \dots a_L) = \frac{1}{Z(\beta)} \exp\left(-\beta \mathcal{H}^{inf}(a_1 \dots a_L)\right). \tag{3.5}$$

When lowering the temperature, that is increasing β, average energies in $\mathcal{H}^{inf}$ of sampled sequences are decreased compared to the $\beta = 1$ case. In this way, it is possible to reach a temperature value such that $\langle \mathcal{H}^{inf} \rangle_{P_\beta^{inf}} \simeq \langle \mathcal{H}^{inf} \rangle_{\mathcal{D}_{nat}}$. However, designing artificial sequences requires that the "true" energies $\mathcal{H}^0$ of sampled sequences match those of the natural ones. Of course, in practice, true energies – if they exist – are unknown. However, if the inferred Hamiltonian $\mathcal{H}^{inf}$ is a good proxy to the real one, then sampling from a low temperature $\mathcal{H}^{inf}$ should result in an enrichment in in low energy sequences in $\mathcal{H}^0$.

Naturally, this cannot be directly tested in the case of biological data, since $\mathcal{H}^0$ is unknown. Here, we use simulated data to test this idea on an ideal case where the true Hamiltonian is of the Potts form. The BML-DCA model learned on the PF00072 family is taken as the true underlying model $\mathcal{H}^0$. A sample drawn at $\beta = 1$ plays the role of the MSA $\mathcal{D}_{nat}$. This sample is in turn used to infer a new Potts model $\mathcal{H}^{inf}$. This allows for the direct comparison of the true and inferred models in terms of energies, but also of other statistical measure such as KL-distance.
Figure 3.2 compares the energy behavior of $\mathcal{H}^{inf}$ when varying temperature. As expected, the average energy $\langle \mathcal{H}^{inf} \rangle_{P_\beta^{inf}}$ decreases as β increases, and there exists a temperature for which it reaches the average energy of the "natural" sequences $\langle \mathcal{H}^{inf} \rangle_{\mathcal{D}_{nat}}$. The second and third panels of figure 3.2 directly compare the two averages $\langle \mathcal{H}^{inf} \rangle_{P_\beta^{inf}}$ and $\langle \mathcal{H}^0 \rangle_{P_\beta^{inf}}$. At $\beta = 1$, the energy in $\mathcal{H}^{inf}$ of sampled sequences is higher than that of training ones. Interestingly, this directly translates in a higher *true* energy $\mathcal{H}^0$. As temperature decreases, energies in the two Hamiltonians stay perfectly correlated. Therefore, when temperature is such that sampled and training sequences have similar energies in $\mathcal{H}^{inf}$, they also have similar energies in the true model $\mathcal{H}^0$. This demonstrates that in this setting, the inferred energies can serve as a good enough proxy of the true ones.

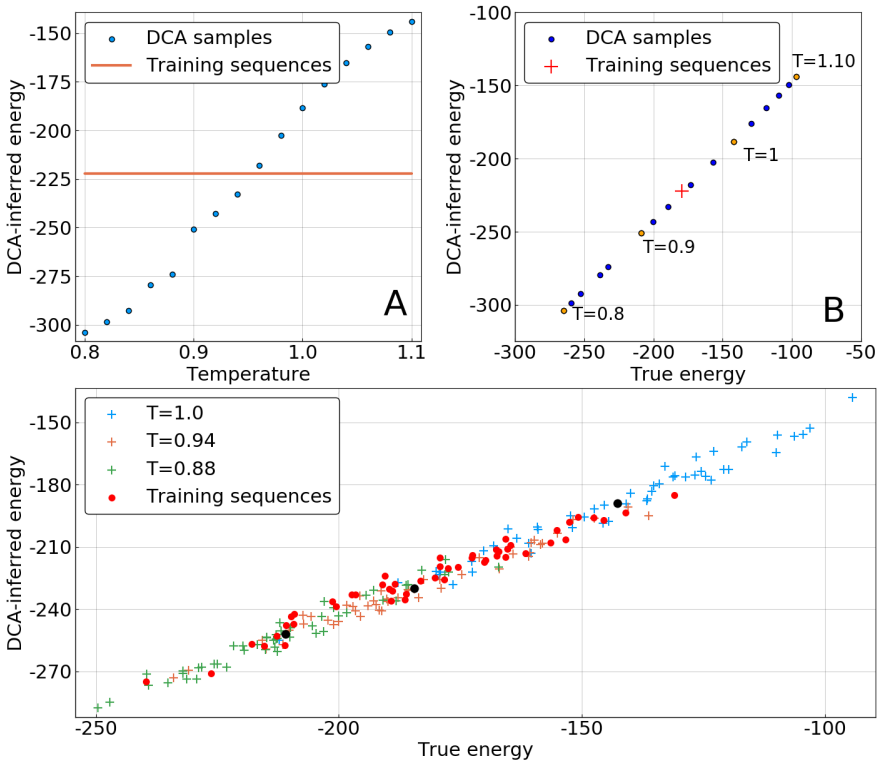

Figure 3.2: ***A.*** Average value of $\mathcal{H}^{inf}$ over samples of P_β^{inf}, as a function of the temperature β^{-1}. Energies of the training sequences are displayed by the horizontal line. ***B.*** Average energies of $\mathcal{H}^{inf}$ and $\mathcal{H}^0$ over samples of P_β^{inf}, for values of β^{-1} between 1.1 and 0.8. ***C.*** Same as ***B.***, but displaying energies of individual sequences instead of averages. Circles are the training sequences, and crosses are samples from the inferred DCA at different temperatures. Black dots are the centers of mass of the crosses, corresponding to what is shown in ***B.***. For visibility, energies of only 50 sequences are displayed for each temperature.

The model is inferred at $T = 1$, meaning that it should reproduce the pairwise frequencies f_{ij}^0 of the training data at this temperature. Figure 3.3 shows the fitting quality of $\beta\mathcal{H}^{inf}$ as a function of β^{-1} – as a reminder, the fitting quality for pairwise statistics is defined as the Pearson correlation between the connected correlations measured in the MSA and in a sample from the inferred model. The statistics are best fitted at $\beta = 1$, as expected. However, the fitting quality quickly deteriorates as the temperature is decreased, reaching values as low as 0.2 for temperatures around $T = 0.8$. The picture changes when looking at the symmetric Kullback-Leibler distance (KL-distance) between $\beta\mathcal{H}^{inf}$ and $\mathcal{H}^0$. The symmetric version of the KL-distance is defined by

$$D_{KL}(P_\beta^{inf}||P^0) + D_{KL}(P^0||P_\beta^{inf}) = \langle \beta\mathcal{H}^{inf} - \mathcal{H}^0 \rangle_{P^0} + \langle \mathcal{H}^0 - \beta\mathcal{H}^{inf} \rangle_{P_\beta^{inf}}. \tag{3.6}$$

As it depends only on energy differences, and not on the partition function of the Hamiltonians as the standard unsymmetric KL-distance, it is easily estimated by sampling from the two models. Figure 3.4 shows this quantity as a function of temperature, as well as the energy differences $\langle \mathcal{H}^0 \rangle_{P_\beta^{inf}} - \langle \mathcal{H}^0 \rangle_{P^0}$ and $\langle \mathcal{H}^{inf} \rangle_{P_\beta^{inf}} - \langle \mathcal{H}^{inf} \rangle_{P^0}$, which correspond to x- and y-axis differences between sampled and training sequences in panels **B** and **C** of figure 3.2. A clear minimum of the KL-distance can be seen around $T \simeq 0.95$, corresponding to the point where the mentioned energy differences vanish. This means that the temperature for which energies of training and sampled sequences are similar in the inferred model is also the one for which the two distributions are the closest in the sense of the KL-distance.

As a comparison, things are quite different when using a profile model instead of a Potts Hamiltonian. We now infer a profile model for $\mathcal{H}^{inf}$, meaning that all the inferred couplings vanish. Since single site frequencies are estimated with a higher accuracy than pairwise ones, it is possible to use a very low regularization in this case. As a result, inferred and true single site frequencies are almost exactly matched, and the bias of Eq. (3.4) is negligible. Figure 3.5 is the equivalent of figure 3.2 for the profile model. Here, according to the inferred model, average energies of sampled and training sequences are the same at $\beta = 1$. However, this does not give any indication about their similarity in the true model. Indeed, sampled sequences have a much higher energy than training ones for $\mathcal{H}^0$.
Decreasing the temperature lowers the energies of sampled sequences both in the true and inferred models. In this manner, it is possible to reach a point for which $\langle \mathcal{H}^0 \rangle_{P_\beta^{inf}} = \langle \mathcal{H}^0 \rangle_{D_{nat}}$. Sequences obtained through the profile model at this temperature would then be expected to be "functional", or at least typical of $\mathcal{H}^0$. However, as this optimal temperature is quite low, the entropy of the corresponding sample

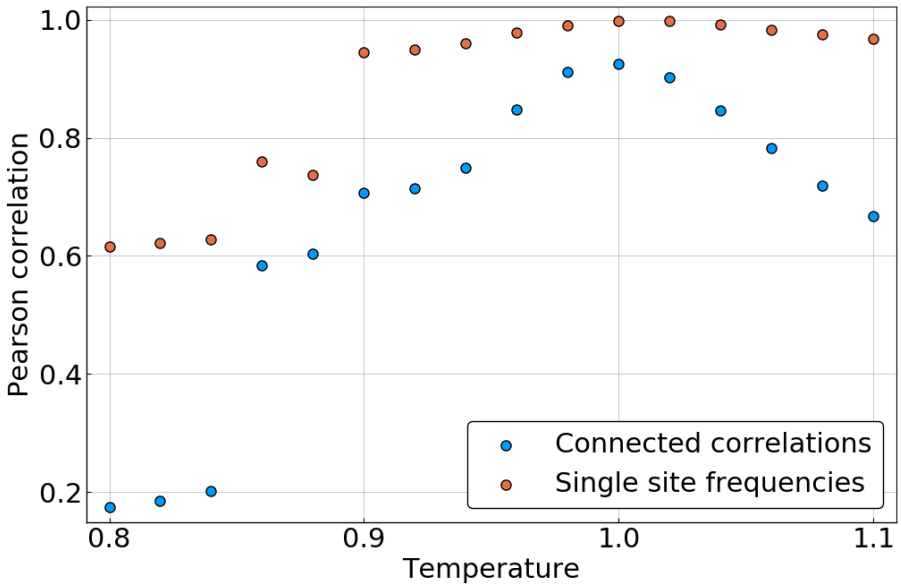

Figure 3.3: Fitting quality of the inferred model. On the y-axis is the Pearson correlation between both connected correlations c_{ij} and single site frequencies f_i of the inferred and true models. Single site frequencies are close to perfectly fitted at $T = 1$ with Pearson correlation of 1, and connected correlations are fitted with a high accuracy at this temperature. However, the fitting quality for connected correlation quickly drops as T is varied.

also decreases and many of the positions in the sequences only have a small variability, as shown in figure 3.6.
Furthermore, in contrast to the case of the Potts model though, energies in $\mathcal{H}^{inf}$ are never a good proxy for energies in $\mathcal{H}^0$. If $\mathcal{H}^0$ were unknown, it would not be possible to use the training data to self-consistently find a sampling temperature such that $\langle \mathcal{H}^0 \rangle_{P_\beta^{inf}} = \langle \mathcal{H}^0 \rangle_{D_{nat}}$.

If DCA is to be used to design artificial sequences, it is necessary to correct for the bias in energies due to regularization. Even in an ideal setting, true energies sequences sampled from an inferred Potts model are higher than those of the natural/training ones. We have shown here that this may be corrected by decreasing the temperature at which sequences are sampled from the DCA model. If the inferred Hamiltonian is a good enough proxy for the "real" energy function, this allows to close the energy gap. It is not possible to apply this method with profile model, as they are likely to not be a proxy for the underlying energy function.
However, sequences sampled in this manner do not reproduce statistical patterns found in the original alignment. Even if the DCA model fits pairwise frequencies with good accuracy, decreasing the temperature

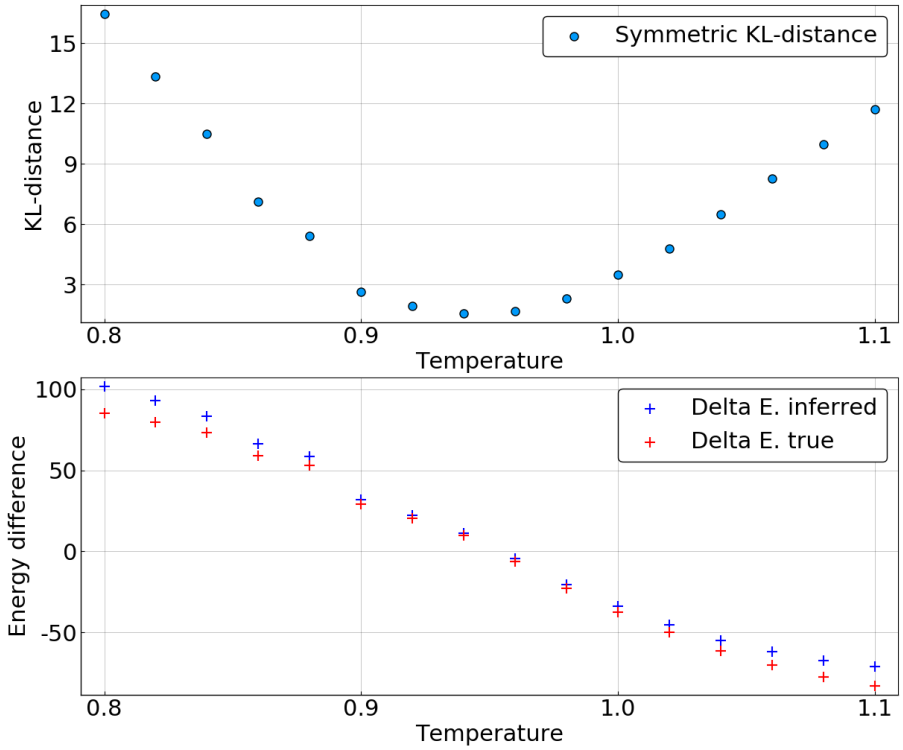

Figure 3.4: **Top.** Symmetric KL-distance between the true model $\mathcal{H}^0$ and the inferred one at temperature β^{-1}, $\beta\mathcal{H}^{inf}$, as a function of the temperature. **Bottom.** Average energy differences between samples from $\beta\mathcal{H}^{inf}$ and the training sequences, measured with the two Hamiltonians $\mathcal{H}^{inf}$ and $\mathcal{H}^0$. When this quantity is 0 for $\mathcal{H}^0$ (unknown in practice), sequences generated by the inferred model have exactly the same average true energies as training sequences.

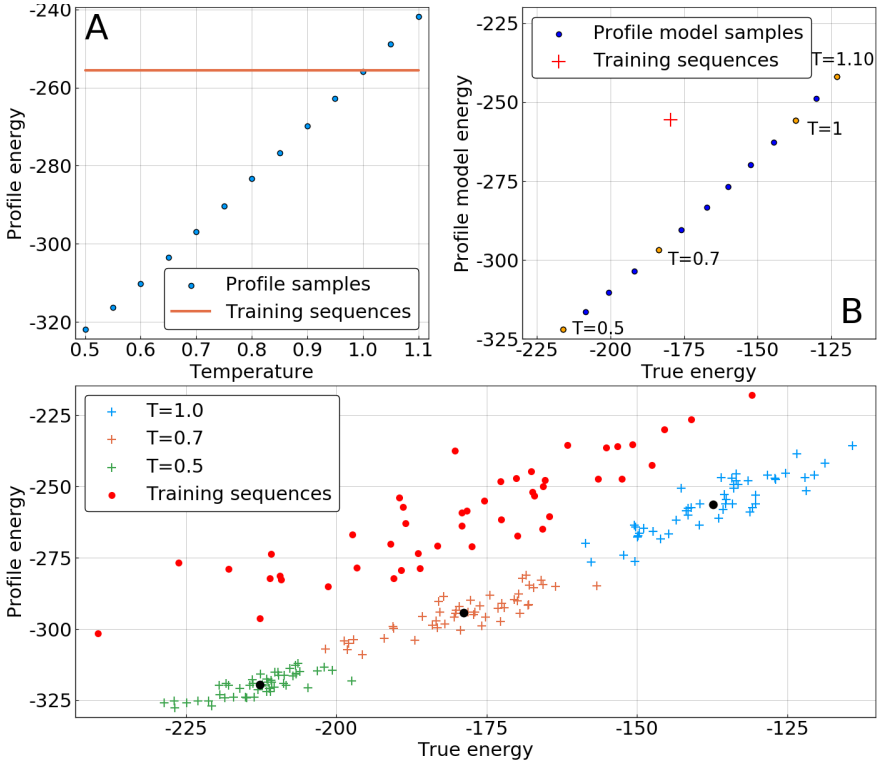

Figure 3.5: Equivalent to figure 3.2, but $\mathcal{H}^{inf}$ is now a profile model without couplings.

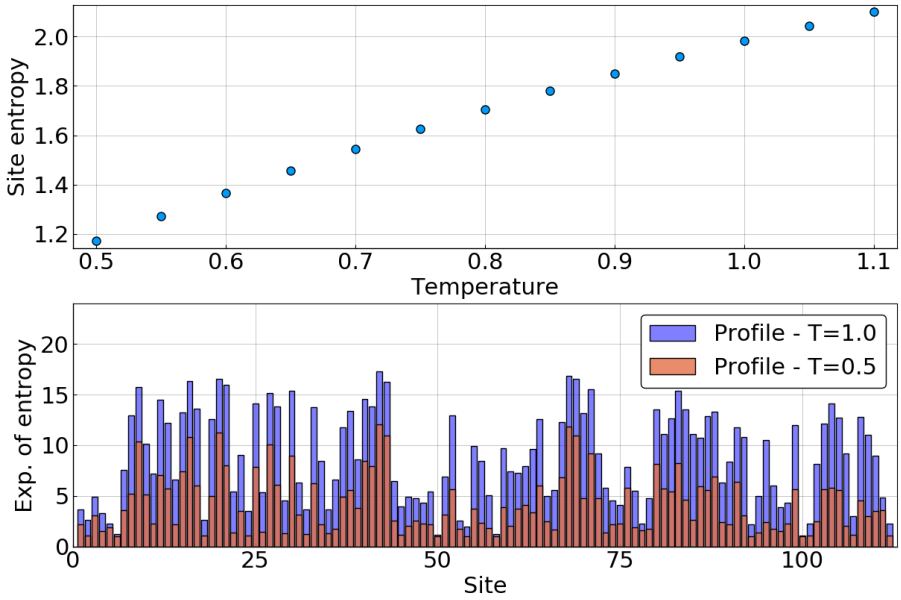

Figure 3.6: **Top.** Per site entropy of the profile model $\mathcal{H}^{inf}$ as a function of temperature. **Bottom.** Exponential of the site-entropy for the profile model for two temperatures.

leads to a biased way of sampling, favoring low energies, and thus modifying frequencies.

4

INTEGRATING HETEROGENEOUS DATA IN THE INVERSE POTTS PROBLEM

4.1 MOTIVATION

DCA was initially introduced as a tool to predict structural contacts from sequences only [82]. As the number of sequences in MSAs is limited, only pairwise statistics are used to construct the model, and other statistical signal is ignored for fear of it being too noisy. This is also coherent with the view that co-evolution at the amino-acid level is due to structural contacts between pairs of residues, and thus influences pairwise correlation in the MSA. The Maximum-Entropy Principle was used in this context to find a functional form to the DCA model based on the chosen statistical observables.
In the article of chapter 3, it is shown that the Potts model inferred using only pairwise statistics measured in the MSA is able to reproduce higher order signal that was not fitted. This indicates that DCA may not only be able to predict contact, but may also be an overall good model of sequence variability. If this is true, one line of development might be to improve its ability to model sequences. While it is hard to use statistical patterns in MSAs beyond the pairwise level due to limited amount of data, it is interesting to try including other sources of information in the modeling process.

DCA is often linked to the problem of inverse statistical physics, in which one attempts to deduce microscopic properties of a system from the observation of its macroscopic behavior. In this analogy, the protein family is the system of interest, single protein sequences represent one of its "microscopic" configuration, while the MSA summarizes its average macroscopic behavior. Thus, frequencies measured in the MSA correspond to macroscopic observables: quantities that do not depend on a precise microscopic configuration (*i.e.* the sequence), but on the average properties of the studied system. In other words, they are an emerging property of the unobserved microscopic constraints which in DCA are represented by the Hamiltonian $\mathcal{H}$. In this sense, they represent what we will call *global* information.
Experimental development in biology now allows to quantify the phenotype (or a proxy for it) of individual protein sequences. As stated in section 2.5.2, quantitative characterization of mutational landscapes has been conducted for a number of proteins [45, 54, 55]. In these experiments, the fitness or phenotype of mutants of existing protein sequences is measured. In our statistical physics analogy, this amounts

to measuring properties of *individual microscopic configurations*. In this sense, these measurements represent information of a different nature than frequencies measured in the MSA, which we will call *local* information.

Statistical physics deals with the question of going from microscopic behavior to macroscopic one, and inverse statistical physics with the one of going in the opposite direction. Here we can here attempt to tackle the problem from both ends by combining statistical patterns of the MSA (*i.e.* global information) with quantitative experiments on mutational effects (*i.e.* local information). The article that follows – [6] – introduces a mathematical framework to properly integrate these two different types of information. This effectively extends the usual setting of DCA in an attempt to increase its accuracy in modeling constraints acting on a protein family.

4.2 ARTICLE

SCIENTIFIC REPORTS

OPEN

Improving landscape inference by integrating heterogeneous data in the inverse Ising problem

Received: 26 August 2016
Accepted: 01 November 2016
Published: 25 November 2016

Pierre Barrat-Charlaix[1,*], Matteo Figliuzzi[1,2,*] & Martin Weigt[1]

The inverse Ising problem and its generalizations to Potts and continuous spin models have recently attracted much attention thanks to their successful applications in the statistical modeling of biological data. In the standard setting, the parameters of an Ising model (couplings and fields) are inferred using a sample of equilibrium configurations drawn from the Boltzmann distribution. However, in the context of biological applications, quantitative information for a limited number of microscopic spins configurations has recently become available. In this paper, we extend the usual setting of the inverse Ising model by developing an integrative approach combining the equilibrium sample with (possibly noisy) measurements of the energy performed for a number of arbitrary configurations. Using simulated data, we show that our integrative approach outperforms standard inference based only on the equilibrium sample or the energy measurements, including error correction of noisy energy measurements. As a biological proof-of-concept application, we show that mutational fitness landscapes in proteins can be better described when combining evolutionary sequence data with complementary structural information about mutant sequences.

High-dimensional data characterizing the collective behavior of complex systems are increasingly available across disciplines. A global statistical description is needed to unveil the organizing principles ruling such systems and to extract information from raw data. Statistical physics provides a powerful framework to do so. A paradigmatic example is represented by the Ising model and its generalizations to Potts and continuous spin variables, which have recently become popular for extracting information from large-scale biological datasets. Successful examples are as different as multiple-sequence alignments of evolutionary related proteins[1–3], gene-expression profiles[4], spiking patterns of neural networks[5,6], or the collective behavior of bird flocks[7]. This widespread use is motivated by the observation that the least constrained (i.e. maximum-entropy[8]) statistical model reproducing empirical single-variable and pairwise frequencies observed in a list of equilibrium configurations is given by a Boltzmann distribution:

$$P(s) = \frac{1}{\mathcal{Z}} \exp\{-\mathcal{H}(\mathbf{s})\}, \quad \mathcal{H} = -\sum_{i<j}^{N} J_{ij} s_i s_j - \sum_{i=1}^{N} h_i s_i, \tag{1}$$

with $s = (s_1,..., s_N)$ being a configuration of N binary variables or 'spins'. Inferring the couplings $\mathbf{J} = \{J_{ij}\}_{1\le i<j\le N}$ and fields $\mathbf{h} = \{h_i\}_{1\le i\le N}$ in the Hamiltonian $\mathcal{H}$ from data, known as the *inverse Ising problem*, is computationally hard for large systems ($N \gg 1$). It involves the calculation of the partition function $\mathcal{Z} = \sum_{\mathbf{s}} e^{-\mathcal{H}(s)}$ as a sum over an exponential number of configurations. The need to develop efficient approximate approaches has recently triggered important work within the statistical-physics community, cf. e.g. refs 9–17.

Despite the broad interest in inverse problems, the methodological setting has remained rather limited: all of this literature, including the biological cases mentioned in the beginning, seeks to estimate model parameters starting from a set of configurations s, which are considered to be at equilibrium and independently drawn from $P(s)$. Real data, however, may be quite different. In biological systems, "microscopic spins configurations" (e.g. amino-acid sequences) are increasingly accessible to experimental techniques, and *quantitative information* for a limited number of *particular* configurations (e.g. three-dimensional structures, measured activities or

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, Biologie computationnelle et quantitative - Institut de Biologie Paris Seine, 75005 Paris, France. [2]Sorbonne Universités, UPMC Univ Paris 06, Institut de Calcul et de la Simulation, 75005 Paris, France. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to M.W. (email: martin.weigt@upmc.fr)

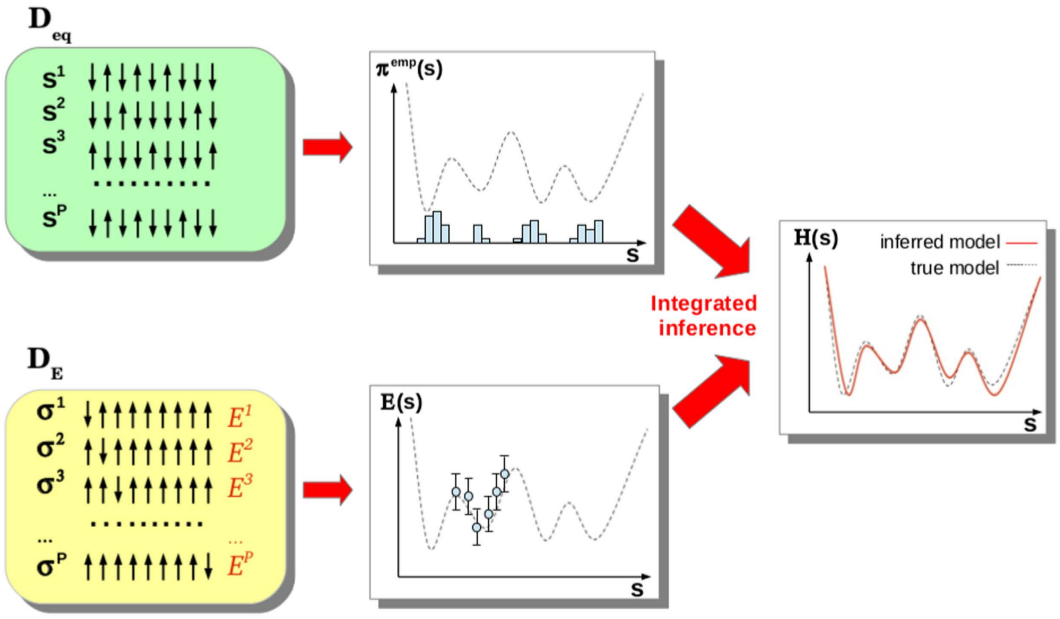

Figure 1. Schematic representation of the inference framework: a sample of equilibrium configurations (dataset D_{eq}) and noisy energy measurements for another set of configurations (dataset D_E) are integrated within a Bayesian approach to infer the model $\mathcal{H}$. The dashed lines represent the underlying true landscape, which has to be inferred, the red line the inferred landscape.

thermodynamic stabilities for selected proteins) is frequently available. It seems reasonable to actually integrate such information into the inverse Ising problem instead of ignoring it. In this work, we use two different types of data (cf. Fig. 1):

- As in the standard inverse Ising problem, part of the data comes as a sample of equilibrium configurations assumed to be drawn from the Boltzmann distribution to be inferred.
- The second data source is a collection of arbitrary configurations together with *noisy* measurements of their energy.

These data sets are limited in size and accuracy. Therefore an optimized integration of both data types is expected to improve the overall performance as compared to the individual use of one single data set.

The inspiration to develop this new integrative framework for the inverse Ising problem is taken from *protein fitness landscapes* in biology, which provide a quantitative mapping from any amino-acid sequence $s = (s_1,..., s_N)$ to a fitness $\phi(s)$ measuring the ability of the corresponding protein to perform its biological function. Fitness landscapes are of outstanding importance in evolutionary and medical biology, but it appears impossible to deduce a protein's fitness from its sequence only. Experimental or computational approaches exploiting other data are urgently needed.

Information about fitness landscapes can be found in the amino-acid statistics observed in natural protein sequences, which are related to the protein of interest. In fact they represent diverse but functional configurations sampled by evolution. It has been recently proposed that their statistical variability can be captured by Potts models (generalization of the Ising model to 21-state amino-acid variables). Indeed, statistical models inferred from large collections of natural sequences have recently led to good predictions of experimentally measured effects[18–21]: in a number of systems, the fitness cost $\Delta\phi(s) \equiv \phi(s) - \phi(s^{ref})$ of mutating any amino acid in a reference protein s^{ref} strongly correlates with the corresponding energy changes in the inferred statistical model,

$$\Delta\phi(s) \sim \log\left[\frac{P(s)}{P(s^{ref})}\right] = -(\mathcal{H}(s) - \mathcal{H}(s^{ref})), \tag{2}$$

suggesting that the Hamiltonian of the inferred models is strictly related to the underlying mutational landscapes.

While evolutionary diverged sequences can be regarded as a global sample of the fitness landscape, further information can be obtained from direct measurements on particular '*microstates*' of the system, *i.e.* individual protein sequences. Recent advances in experimental technology allow for conducting large-scale *mutagenesis* studies: in a typical experiment, a reference protein of interest is chosen, and a large number (10^3–10^5) of mutant proteins (having sequences differing by one or few amino acids from the reference) are synthesized and then characterized in terms of fitness. This provides a systematic *local* measurement of the fitness landscape[22–24]. Regression analysis may be used to globally model mutational landscapes[25]. A second-order parameterization of ϕ arises naturally in this context, when considering an expansion of effects in terms of independent additive effects and pairwise '*epistatic*' interactions between sites[26],

$$\phi(s) = \sum_{i=1}^{N} \varphi_i(s_i) + \sum_{1 \le i \le j \le N} \varphi_{i,j}(s_i, s_j). \tag{3}$$

However, the number of accessible mutant sequences remains small compared to the number of terms in this sum, and mutagenesis data alone are not sufficient to faithfully model fitness landscapes[27].

In situations where no single dataset is sufficient for accurate inference, integrative methods accounting for complementary data sources will improve the accuracy of computational predictions. In this paper, (i) we define a generalized inference framework based on the availability of an equilibrium sample *and* of complementary quantitative information; (ii) we propose a Bayesian integrative approach to improve over the limited accuracy obtainable using standard inverse problems; (iii) we demonstrate the practical applicability of our method in the context of predicting mutational effects in proteins, a problem of outstanding bio-medical importance for questions related to genetic disease and antibiotic drug resistance.

Results

An integrated modeling. *The inference setting.* Inspired by this discussion, we consider two different datasets originating from a *true* model $\mathcal{H}^0$. The first one, $\mathbf{D}_{\rm eq} = \{s^1,..., s^M\}$, is a collection of M equilibrium configurations independently drawn from the Boltzmann distribution $P^0(s)$. For simplicity, we consider binary variables $s_i \in \{0, 1\}$, corresponding to a "lattice-gas" representation of the Ising model in Eq. (1). This implies that energies are measured with respect to the reference configuration $s^{\rm ref} = (0,..., 0)$. The standard approach to the inverse Ising problem uses only this type of data to infer parameters of $\mathcal{H}$: couplings and fields in Eq. (1) are fitted so that the inferred model reproduces the empirical single and pairwise frequencies,

$$\pi_i^{\rm emp} = \frac{1}{M}\sum_{\mu=1}^{M} s_i^\mu, \quad \pi_{ij}^{\rm emp} = \frac{1}{M}\sum_{\mu=1}^{M} s_i^\mu s_j^\mu. \tag{4}$$

The second dataset provides a complementary source of information, which shall be modeled as noisy measurements of the *energies* of a set of P *arbitrary* (i.e. not necessarily equilibrium) configurations σ^a. These data are collected in the dataset $\mathbf{D}_E = \{(E^a, \sigma^a)_{a=1,...,P}\}$, with

$$E^a = \mathcal{H}^0(\sigma^a) + \xi^a \quad a = 1,...,P. \tag{5}$$

The noise ξ^a models measurement errors or uncertainties in mapping measured quantities to energies of the Ising model. For simplicity, we consider ξ^a to be white Gaussian noise with zero mean and variance Δ^2: $\langle \xi^a \xi^b \rangle = \delta_{a,b}\Delta^2$.

As schematically represented in Fig. 1, datasets $\mathbf{D}_{\rm eq}$ and $\mathbf{D}_E$ constitute different sources of information about the energy landscape defined by Hamiltonian $\mathcal{H}^0$. Observables in Eq. (4) are empirical averages computed from equilibrium configurations in $\mathbf{D}_{\rm eq}$, providing *global* information about the energy landscape. On the contrary, configurations in $\mathbf{D}_E$ are arbitrarily given, and a (noisy) measurement of their energies provides *local* information on particular points in the landscape.

A maximum-likelihood approach. To infer the integrated model, we consider a *joint description of the probabilities of the two data types* for given parameters $\mathbf{J}$ and $\mathbf{h}$ of tentative Hamiltonian $\mathcal{H}$. The probability of observing the sampled configurations in $\mathbf{D}_{\rm eq}$ equals the product of the Boltzmann probability of each configuration,

$$P(\mathbf{D}_{\rm eq}|\mathbf{J}, \mathbf{h}) = \exp\left\{-\sum_{\mu=1}^{P} \mathcal{H}(\mathbf{s}^\mu) - M \log \mathcal{Z}(\mathbf{J}, \mathbf{h})\right\}. \tag{6}$$

To derive an analogous expression for the second dataset, we integrate over the Gaussian distribution of the noise $\xi^a = E^a - \mathcal{H}(\sigma^a)$ obtaining a Gaussian probability of the energies (remember configurations in $\mathbf{D}_E$ are arbitrarily given):

$$\begin{aligned} P(\mathbf{D}_E|\mathbf{J}, \mathbf{h}) &= \prod_{a=1}^{P} \int d\xi_a P(\xi_a)\, \delta(E^a - \mathcal{H}(\sigma^a) - \xi^a) \\ &= \frac{1}{(2\pi\Delta^2)^{\frac{P}{2}}} \exp\left\{-\sum_{a=1}^{P} \frac{[E^a - \mathcal{H}(\sigma^a)]^2}{2\Delta^2}\right\} \end{aligned} \tag{7}$$

The combination of these expressions provides the joint log-likelihood for the model parameters given the data:

$$\mathcal{L}(\mathbf{J}, \mathbf{h}\big|\mathbf{D}_{\rm eq}, \mathbf{D}_E) = \log P(\mathbf{D}_{\rm eq}\big|\mathbf{J}, \mathbf{h}) + \log P(\mathbf{D}_E\big|\mathbf{J}, \mathbf{h}) \tag{8}$$

Maximizing the above likelihood with respect to parameters $\{h_i\}_{1 \le i \le N}$ and $\{J_{ij}\}_{1 \le i < j \le N}$ leads to the following self-consistency equations:

$$\begin{aligned} p_i(\mathbf{J}, \mathbf{h}) &= \pi_i^{\rm emp} + \frac{\lambda}{1-\lambda}\frac{1}{M}\sum_{a=1}^{P} \sigma_i^a [E^a - \mathcal{H}(\sigma^a)] \\ p_{ij}(\mathbf{J}, \mathbf{h}) &= \pi_{ij}^{\rm emp} + \frac{\lambda}{1-\lambda}\frac{1}{M}\sum_{a=1}^{P} \sigma_i^a \sigma_j^a [E^a - \mathcal{H}(\sigma^a)] \end{aligned} \tag{9}$$

with $p_i(\mathbf{J}, \mathbf{h}) = \langle \sigma_i \rangle_{\mathcal{H}}$, $p_{ij}(\mathbf{J}, \mathbf{h}) = \langle \sigma_i \sigma_j \rangle_{\mathcal{H}}$ being single and pairwise averages in the model Eq. (1). We have introduced the parameter $\lambda = \frac{1}{1+\Delta^2}$: in practical applications, the error Δ may not be known, and the parameter $0 \leq \lambda < 1$ allows to weigh data sources differently. For $\lambda = 0$ (*i.e.* large noise), the standard inverse Ising problem is recovered: optimal parameters are such that the model exactly reproduces magnetizations and correlations of the sample. For $\lambda > 0$, the second dataset containing quantitative data is taken into account: whenever energies computed from the Hamiltonian $\mathcal{H}$ do not match the measured ones, the model statistics deviates from the sample statistics. Both log-likelihood terms in (8) are concave, and thus their sum: Eq. (9) has a unique solution.

Noiseless measurements. The case of noiseless energy measurements in Eq. (5) (i.e. $\lambda \to 1$) has to be treated separately. First, energies have to be perfectly fitted by the model, by solving the following linear problem:

$$\mathbf{X}\vec{\mathcal{H}} = \vec{E}, \tag{10}$$

where we have introduced a $N(N+1)/2$ dimensional vectorial representation $\vec{\mathcal{H}} = (h_1,\ldots,h_N, J_{12},\ldots,J_{N,N-1})^T$ of the model parameters, and $\vec{E} = (E_1,\ldots,E_P)^T$ contains the exactly measured energies. The matrix

$$\mathbf{X} = \begin{bmatrix} \sigma_1^1 & \cdots & \sigma_N^1 & (\sigma_1^1\sigma_2^1) & \cdots & (\sigma_{N-1}^1\sigma_N^1) \\ \vdots & & \vdots & \vdots & & \vdots \\ \sigma_1^P & \cdots & \sigma_N^P & (\sigma_1^P\sigma_2^P) & \cdots & (\sigma_{N-1}^P\sigma_N^P) \end{bmatrix} \tag{11}$$

specifies which parameters contribute to the energies of configurations in the second dataset. If $K = N(N+1)/2 - rank(X) > 0$, the parameters cannot be uniquely determined from the measurements: The sample $\mathbf{D}_{\text{eq}}$ can be used to remove the resulting degeneracy. To do so, we parametrize the set of solutions of Eq. (10) as follows:

$$\vec{\mathcal{H}} = \vec{\mathcal{H}}_{nh} + \sum_{k=1}^{K} \alpha_k \vec{\mathcal{O}}_k \tag{12}$$

where $\vec{\mathcal{H}}_{nh}$ is any particular solution of the non homogeneous Eq. (10), and $\{\vec{\mathcal{O}}_k\}$ a basis of observables spanning the null space of the associated homogeneous problem $\mathbf{X}\vec{\mathcal{H}} = 0$. The free parameters $\alpha_k \in \mathbb{R}$ can be fixed by maximizing their likelihood given sample $\mathbf{D}_{\text{eq}}$,

$$\mathcal{L}(\alpha|\mathbf{D}_{\text{eq}}) \propto \exp\left\{ -\sum_{k,\mu} \alpha_k \mathcal{O}_k(\mathbf{s}^\mu) - M \log \mathcal{Z}(\alpha) \right\} \tag{13}$$

with $\mathcal{O}_k(\mathbf{s}) = (s_1, \cdots, s_N, (s_1 s_2), \cdots, (s_{N-1}s_N)) \cdot \vec{\mathcal{O}}_k$. The maximization provides conditions for the observables ($k = 1,\ldots, K$),

$$\frac{\partial \mathcal{L}(\alpha)}{\partial \alpha^k} \propto \frac{1}{M}\sum_{\mu=1}^{M} \mathcal{O}_k(\mathbf{s}^\mu) - \langle \mathcal{O}_k \rangle_{\mathcal{H}} = 0. \tag{14}$$

Equation (14) shows that the α_k have to be fixed such that empirical averages $\frac{1}{M}\sum_{\mu=1}^{M} \mathcal{O}_k(\mathbf{s}^\mu)$ equal model averages $\langle \mathcal{O}_k \rangle_{\mathcal{H}}$. Any possible sparsity of the matrix of measured configurations $\mathbf{X}$ (entries are 0 or 1 by definition) can be exploited to find a sparse representation of the $\{\vec{\mathcal{O}}_k\}$. In the protein example discussed above, mutagenesis experiments typically quantify all possible single-residue mutations of a reference sequence (denoted (0,..., 0) without loss of generality). In this case, the pairwise quantities $s_i s_j$ with $1 < i < j < N$ can be chosen as the basis $\{\mathcal{O}_k\}$ of the null space. A particular solution of the non-homogeneous system (10) is given by the paramagnetic Hamiltonian $\mathcal{H}_{nh} = \sum_i E^i s_i$, with E^i being the energy shift due to spin flip $s_i = 0 \mapsto 1$.

Artificial data. We first evaluate our method on artificial data (*Materials and Methods*). Random couplings $\mathbf{J}^0$ and fields $\mathbf{h}^0$ are chosen for a system of $N = 32$ spins. Dataset $\mathbf{D}_{\text{eq}}$ is created by Markov chain Monte Carlo (MCMC) sampling, resulting in $M = 100$ equilibrium configurations. To mimic a protein 'mutagenesis' experiment, one of these configurations is chosen at random as the reference sequence, and the energies of all N configurations differing by a single spin flip from the reference (thereafter referred to as single mutants) are calculated, resulting in dataset $\mathbf{D}_E$ (after adding noise of standard deviation Δ_0). Datasets $\mathbf{D}_E$ and $\mathbf{D}_{\text{eq}}$ will subsequently called "local" and "global" data respectively.

Equations (9) are solved using steepest ascent, updating parameters $\mathbf{J}$ and $\mathbf{h}$ in direction of the gradient of the joint log-likelihood (8). Since the noise Δ_0 may not be known in practical applications, we solve the equations for several values of $\lambda \in [0, 1]$, weighing data sources differently. We expect the optimal inference to take place at a value λ that maximizes the likelihood in Eq. (8), *i.e.* $\lambda_0 = (1 + \Delta_0^2)^{-1}$. For $\lambda = 0$, this procedure is equivalent to the classical Boltzmann machine[28], but for $\lambda > 0$, the term corresponding to the quantitative essay constrains energies of sequences in $\mathbf{D}_E$ to stay close to the measurements. As explained above, the case $\lambda = 1$ has to be treated separately; a similar gradient ascent method is used. Since exact calculations of gradients are computationally hard, the mean-field approximation is used (*Materials and Methods*).

To evaluate the accuracy of the inference, most of the existing literature on the inverse Ising modeling simply compares the inferred parameters with the true ones. However, a low error in the estimation of each inferred

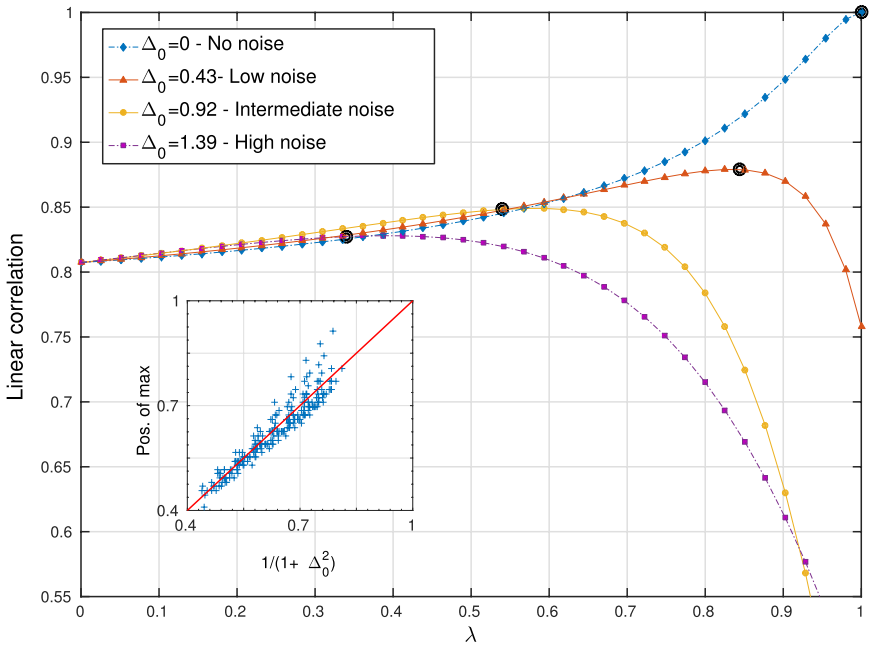

Figure 2. Integration of noisy measurements of energies of single mutants with *M* equilibrium configurations. The linear correlation between predicted and true single mutants energies is shown in dependence of integration parameter λ, for four different values of the noise $\Delta_0 = 0$, 0.43, 0.92, 1.39 added to the energies in the second data set. The integration strength $\lambda_0 = 1/(1 + \Delta_0^2)$, which would be naturally used in case of an *a priori* known noise level, is located close to the optimal inference, cf. the black circles for $\lambda_0 = 1$, 0.84, 0.54, 0.34. The insert shows the value of the integration strength λ reaching maximal correlation, as a function of the theoretical value $1/(1 + \Delta_0^2)$, for 200 independent realisations of the input data at different noise levels. Points are found to be closely distributed around the diagonal (red line).

parameter does not guarantee that the inferred distribution matches the true one. On the contrary, in the case of a high susceptibility of the statistics with respect to parameter variations, or if the estimation of parameters is biased, the distributions of the inferred and true models could be very different even for small errors on individual parameters. For this reason, we introduce two novel evaluation procedures. First, to estimate the accuracy of the model on a local region of the configuration space, we test its ability to reproduce *energies* of configurations in $\mathbf{D}_E$. Then, we estimate the global similarity of true and inferred distributions using a measure from information theory.

Error correction of local data. We first test the ability of our approach to predict the true single-mutant energies, when noisy measurements are presented in $\mathbf{D}_E$, *i.e.* to correct the measurement noise using the equilibrium sample $\mathbf{D}_{\rm eq}$. For every λ, $\mathbf{J}$ and $\mathbf{h}$ are inferred and used to compute predicted energies of the N configurations in $\mathbf{D}_E$. The linear correlation between such predicted energies (measured with the inferred Hamiltonian) and the true energies (measured with the true Hamiltonian) is plotted as a function of λ in Fig. 2.

In the very low noise regime, $\Delta_0 \simeq 0$, the top curve in Fig. 2 reaches its peak at $\lambda \simeq 1$, which is expected as local data is then sufficient to accurately "predict" energies from single mutants. On the contrary, in the high noise regime, the maximum is located close to $\lambda = 0$, pointing to the fact that local data is of little use in this case. Between those two extremes, an optimal integration strength can be found, yielding a better prediction of energies in $\mathbf{D}_{\rm E}$ as for any of the datasets taken individually. It is interesting to notice that even for highly noisy data, integrating the two sources of information with the right weight λ results in an improved modeling.

The insert of Fig. 2 shows the integration strength λ at which the best correlation is reached, against the corresponding theoretical value $\lambda_0 = 1/(1 + \Delta_0^2)$ for different realizations. On average, optimal integration is reached close to the theoretical case of equation (8). This result highlights the possibility of using this integrative approach to *correct measurement errors* in the energies of single mutants. If a dataset such as $\mathbf{D}_{\rm eq}$ provides global information about the energy landscape, and the measurement noise Δ_0 can be estimated, an appropriate integration can then be used to infer more accurately the energies of the single mutants.

Global evaluation of the inferred Ising model. To assess the ability of our integrative procedure to provide a *globally* accurate model, we use the Kullback-Leibler divergence $D_{KL}(P^0||P)$ between the true model $P^0 \propto e^{-\mathcal{H}^0}$ and the inferred $P \propto e^{-\mathcal{H}}$ (*Materials and Methods*). The symmetric expression

$$\begin{aligned} \Sigma(P^0, P) &= D_{KL}(P^0||P) + D_{KL}(P||P^0) \\ &= \langle \mathcal{H}^0 - \mathcal{H} \rangle_P + \langle \mathcal{H} - \mathcal{H}^0 \rangle_{P^0} \end{aligned} \tag{15}$$

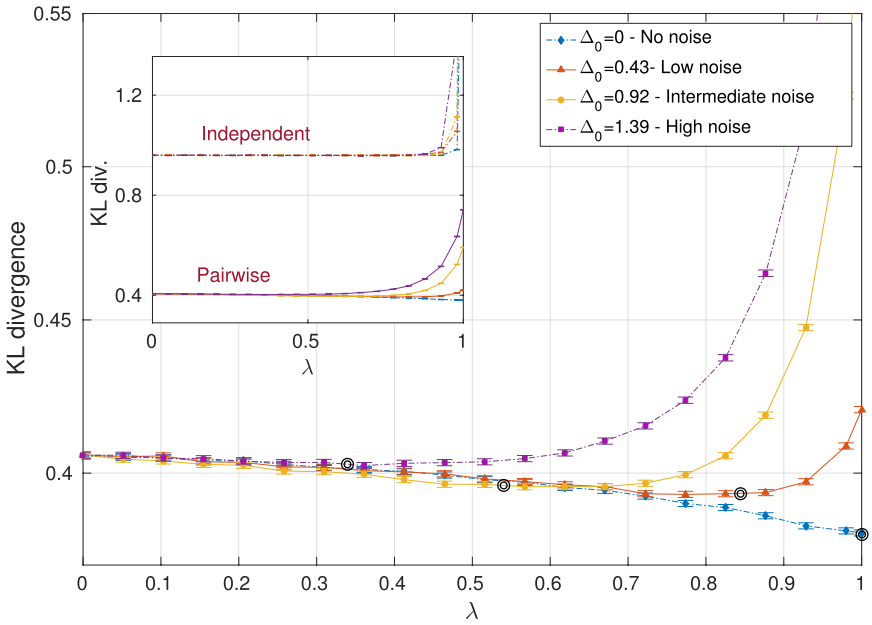

Figure 3. Symmetric Kullback-Leibler divergence $\sum$ between true and integrated pairwise models versus strength of integration λ, estimated from $M_K = 3 \cdot 10^6$ MCMC samples. Different curves correspond to different noise levels added to single mutants energies used for integration, so that datasets $\mathbf{D}_E$ are the same as in Fig. 2. **Insert** - Comparison with an independent model using only fields $\mathbf{h}$, with the same methodology.

simplifies to the average difference between true and inferred energies. It can be consistently estimated using MCMC samples $\mathbf{D}$ (resp. $\mathbf{D}^0$) drawn from P (resp. P^0), without the need to calculate the partition functions. $\Sigma(P^0, P)$ has an intuitive interpretation in terms of distinguishability of models: It represents the log-odds ratio between the probability to observe samples $\mathbf{D}$ and $\mathbf{D}^0$ in their respective generating models, and the corresponding probability with models $\mathcal{H}$ and $\mathcal{H}^0$ swapped:

$$\Sigma(P^0, P) = \frac{1}{M_K} \log\left[\frac{P(\mathbf{D}|\mathcal{H})P(\mathbf{D}^0|\mathcal{H}^0)}{P(\mathbf{D}|\mathcal{H}^0)P(\mathbf{D}^0|\mathcal{H})}\right]. \tag{16}$$

where M_K is the number of sampled configurations in $\mathbf{D}^{\mathcal{H}}$ and $\mathbf{D}^0$.

The inferred model undoubtedly benefits from the integration, as a minimal divergence between the generating and the inferred probability distributions is found for an intermediate value of λ, outperforming both datasets taken individually (Fig. 3). It has to be noted that even in the noiseless case $\Delta_0 = 0$, the minimum in $\Sigma(P^0, P)$ obtained at $\lambda = 1$ depends crucially on the availability of the equilibrium sample D_{eq}. The local data D_E are not sufficient to fix uniquely all model parameters, and the degeneracy in parametrization is resolved using D_{eq} as explained at the end of Sec. 0.

As a comparison, the same analysis is done using an independent modeling that uses only fields $\mathbf{h}$, and no couplings. The inset of Fig. 3 clearly shows that the pairwise modeling outperforms the independent one. Even the limit $\lambda \to 1$, where D_{eq} becomes irrelevant in the independent model, the performance of the integrative pairwise scheme is not attained.

Biological data. To demonstrate the practical utility of our integrative framework, we apply it to the challenging problem of predicting the effect of amino-acid mutations in proteins. To do so, we use three types of data: (*i*) Multiple-sequence alignments (MSA) of homologous proteins containing large collections of sequences with shared evolutionary ancestry and conserved structure and function; they are obtained using HMMer[29] using profile models from the Pfam database[30]. Due to their considerable sequence divergence (typical Hamming distance ~0.8*N*), they provide a *global* sampling of the underlying fitness landscape. (*ii*) Computational predictions of the impact of all single amino-acid mutations on a protein's structural stability[31] are used to *locally* characterize the fitness landscape around a given protein. The noise term ξ^a represents the limited accuracy of this predictor, and the uncertainty in using structural stability as a proxy of protein fitness. (*iii*) Mutagenesis experiments have been used before to simultaneously quantify the fitness effects of thousands of mutants[22,23]. While datasets (*i*) and (*ii*) play the role of $\mathbf{D}_{eq}$ and $\mathbf{D}_E$ in inference, dataset (*iii*) is used to assess the quality of our predictions (ideally one would use the most informative datasets (i) and (iii) to have maximally accurate predictions, but no complementary dataset to test predictions would be available in that case).

To apply the inference scheme to such protein data, three modifications with respect to simulated data are needed. First, the relevant description in this case is a 21-state Potts Model (Supporting Information), since each variable s_i, $i = 1,..., N$, can now assume 21 states (20 amino acids, one alignment gap)[32]. Second, since measured

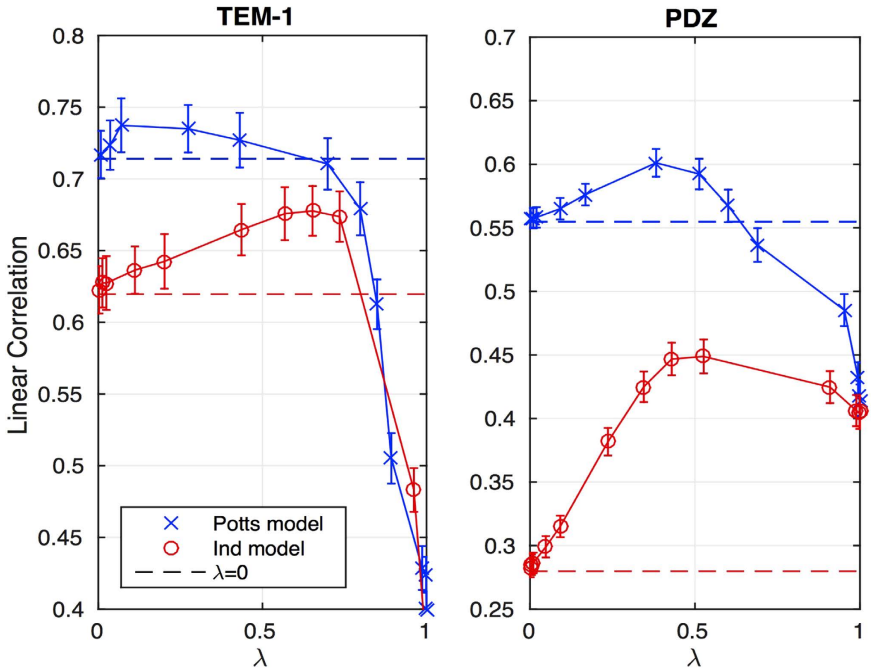

Figure 4. Linear correlation between experimental fitness effects and predictions from integrated models, at different values of λ, for 742 single mutations in the beta-lactamase TEM-1 (left panel) and for 1426 single mutations in the PSD-95 PDZ domain (right panel). Error bars represent statistical errors estimated via jack-knife estimation.

fitnesses and model energies are found in a monotonous non-linear relation, we have used the robust mapping introduced in ref. 18 (reviewed in the Supporting Information). Third, since correlations observed in MSA are typically too strong for the MF approximation to accurately estimate marginals, we relied on Markov Chain Monte Carlo (*Materials and Methods*), which has recently been shown to outperform other methods in accuracy of inference for protein-sequence data[33,34].

We have tested our approach for predicting the effect of single amino-acid mutations in two different proteins: the β-lactamase TEM-1, a bacterial enzyme providing antibiotic resistance, and the PSD-95 signaling domain belonging to the PDZ family. In both systems computational predictions can be tested against recent high-throughput experiments quantifying the *in-vivo* functionality of thousands of protein variants[22,23]. Figure 4 shows the Pearson correlations between inferred energies and measured fitnesses as a function of the weight λ: Maximal accuracy is achieved at finite values of λ when both sources of information are combined, significantly increasing the predictive power of the models inferred considering the statistics of homologs only ($\lambda = 0$). When repeating the integrated modeling with a paramagnetic model where all sites are treated independently, $\mathcal{H}^{\mathrm{ind}}(\mathbf{a}) = -\sum_{i=1}^{N} h_i(a_i)$ (only single-site frequencies are fitted in this case) the predictive power drops as compared to the Potts model, cf. the red lines in Fig. 4.

Conclusion

In this paper, we have introduced an integrative Bayesian framework for the inverse Ising problem. In difference to the standard setting, which uses only a global sample of independent equilibrium configurations to reconstruct the Hamiltonian of an Ising model, we also consider a local quantification of the energy function around a reference configuration. Using simulated data, we show that the integrated approach outperforms inference based on each single dataset alone. The gain over the standard setting of the inverse Ising problem is particularly large when the equilibrium sample is too small to allow for accurate inference.

This undersampled situation is particularly important in the context of biological data. The prediction of mutational effects in proteins is of enormous importance in various bio-medical applications, as it could help understanding complex and multifactorial genetic diseases, the onset and the proliferation of cancer, or the evolution of antibiotic drug resistance. However, the sequence samples provided by genomic databases, like the multiple-sequence alignments of homologous proteins considered here, are typically of limited size, including even in the most favorable situations rarely more than 10^3–10^5 alignable sequences. Fortunately, such sequence data are increasingly complemented by quantitative mutagenesis experiments, which use experimental high-throughput approaches to quantify the effect of thousands of mutants. While it might be tempting to use these data directly to measure mutational landscapes from experiments, it has to be noted that current experimental techniques miss at least 2–3 orders of magnitude in the number of measurable mutants to actually reconstruct the mutational landscape.

In such situations, where no single dataset is sufficient for accurate inference, integrative methods like the one proposed here will be of major benefit.

Methods

Data. *Artificial data.* For a system of $N = 32$ binary spins, couplings $\mathbf{J}^0$ and fields $\mathbf{h}^0$ are chosen from a Gaussian distribution with zero mean, and standard deviation $\sim 0.8/\sqrt{N}$ for J and 0.2 for h (analogous results are obtained for other parameter choices, as long as these correspond to a paramagnetic phase). Dataset $\mathbf{D}_{\text{eq}}$ is created by Markov chain Monte Carlo (MCMC) simulation, resulting in $M = 100$ equilibrium configurations. A large number ($\sim 10^5$) of MCMC steps are done between each of those configurations to ensure that they are independent. One of these configurations is chosen at random as the reference sequence ("wild-type"), and the energies of all N configurations differing by a single spin flip from the reference are computed ("single mutants"). Gaussian noise of variance Δ_0^2 can be added to these energies, resulting in dataset $\mathbf{D}_E$.

Biological data. Detailed information about the analysis of biological data is provided in the Supporting Information.

Details of the inference. For artificial data, Eq. (9) are solved using steepest ascent, updating parameters $\mathbf{J}$ and $\mathbf{h}$ in direction of the gradient. To ensure convergence, we have added an additional ℓ_2-regularization to the joint likelihood: $\gamma_h(\|\mathbf{h}\|_2)^2 + \gamma_J(\|\mathbf{J}\|_2)^2$. A gradient ascent method has been analogously used for the case $\lambda = 1$. To estimate the gradient, it is necessary to compute single and pair-wise probabilities $p_i(\mathbf{J}, \mathbf{h})$ and $p_{ij}(\mathbf{J}, \mathbf{h})$. Their exact calculation requires summation over all possible configurations of N spins, which is intractable even for systems of moderate size N, so we relied on the following approximation schemes.

Mean-field inference. In the analysis of artificial data we relied on the mean-field approximation (MF) leading to closed equations for p_i and p_{ij}:

$$\begin{aligned} p_i &= e^{h_i + \sum_{j \neq i} J_{ij} p_j} / (1 + e^{h_i + \sum_{j \neq i} J_{ij} p_j}) \\ p_{ij} - p_i p_j &= -(\mathbf{J}^{-1})_{ij}. \end{aligned} \tag{17}$$

The main advantage of the MF approximation is its computational efficiency: The first term is solved by an iterative procedure, the second requires the inversion of the couplings matrix $\mathbf{J}$. However, the approximation is only valid and accurate at "high temperatures", *i.e.* small couplings[35]. This condition is verified in the case of the artificial data described above.

MCMC inference. Correlations observed in MSA of protein sequences are typically too strong for the MF approximation to accurately estimate marginals of the model. Therefore we use MCMC sampling of $M^{MC} = 10^4$ independent equilibrium configurations to estimate marginals at each iteration of the previously described learning protocol.

Global evaluation of the inferred Ising model. The Kullback-Leibler divergence $D_{KL}(P||Q) = \sum_s P(s) \log \{P(s)/Q(s)\}$ is a measure of the difference between probability distributions P and Q. It is zero for $P \equiv Q$, and otherwise positive. In the case of Boltzmann distributions $P \propto e^{-\mathcal{H}_P}$ and $Q \propto e^{-\mathcal{H}_Q}$, its expression simplifies to

$$D_{KL}(P||Q) = \langle \mathcal{H}_Q - \mathcal{H}_P \rangle_P + \log \mathcal{Z}_Q - \log \mathcal{Z}_P. \tag{18}$$

Evaluating this expression requires the exponential computation of the partition function of both models $\mathcal{H}_P$ and $\mathcal{H}_Q$. To overcome this difficulty, we use the symmetrized expression in Eq. (15), which only involves the average of macroscopic observables.

The symmetrized Kullback-Leibler divergence is computed by obtaining $M_K = 128000$ equilibrium configurations from both P and Q, using them to estimate the averages in Eq. (15).

References

1. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106,** 67–72 (2009).
2. Mora, T., Walczak, A. M., Bialek, W. & Callan, C. G. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences* **107,** 5405–5410 (2010).
3. Ferguson, A. L. *et al.* Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38,** 606–617 (2013).
4. Lezon, T. R., Banavar, J. R., Cieplak, M., Maritan, A. & Fedoroff, N. V. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences* **103,** 19033–19038 (2006).
5. Schneidman, E., Berry, M. J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440,** 1007–1012 (2006).
6. Cocco, S., Leibler, S. & Monasson, R. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences* **106,** 14058–14062 (2009).
7. Bialek, W. *et al.* Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences* **109,** 4786–4791 (2012).
8. Jaynes, E. T. Information theory and statistical mechanics. *Physical Review* **106,** 620 (1957).
9. Roudi, Y., Tyrcha, J. & Hertz, J. Ising model for neural data: model quality and approximate methods for extracting functional connectivity. *Physical Review E* **79,** 051915 (2009).
10. Sessak, V. & Monasson, R. Small-correlation expansions for the inverse ising problem. *Journal of Physics A: Mathematical and Theoretical* **42,** 055001 (2009).

11. Mézard, M. & Mora, T. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris* **103,** 107–113 (2009).
12. Cocco, S., Monasson, R. & Sessak, V. High-dimensional inference with the generalized hopfield model: Principal component analysis and corrections. *Physical Review E* **83,** 051123 (2011).
13. Cocco, S. & Monasson, R. Adaptive cluster expansion for inferring boltzmann machines with noisy data. *Physical Review Letters* **106,** 090601 (2011).
14. Nguyen, H. C. & Berg, J. Mean-field theory for the inverse ising problem at low temperatures. *Physical Review Letters* **109,** 050602 (2012).
15. Aurell, E. & Ekeberg, M. Inverse ising inference using all the data. *Physical Review Letters* **108,** 090201 (2012).
16. Nguyen, H. C. & Berg, J. Bethe–peierls approximation and the inverse ising problem. *Journal of Statistical Mechanics: Theory and Experiment* **2012,** P03004 (2012).
17. Decelle, A. & Ricci-Tersenghi, F. Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of ising models. *Physical Review Letters* **112,** 070603 (2014).
18. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary inference of mutational landscape and the context dependence of mutations in beta-lactamase tem-1. *Molecular Biology and Evolution* (2016).
19. Asti, L., Uguzzoni, G., Marcatili, P. & Pagnani, A. Maximum-entropy models of sequenced immune repertoires predict antigen-antibody affinity. *PLoS Comput Biol* **12,** e1004870 (2016).
20. Morcos, F., Schafer, N. P., Cheng, R. R., Onuchic, J. N. & Wolynes, P. G. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences* **111,** 12408–12413 (2014).
21. Mann, J. K. *et al.* The fitness landscape of hiv-1 gag: Advanced modeling approaches and validation of model predictions by *in vitro* testing. *PLoS Comput Biol* **10,** e1003776 (2014).
22. McLaughlin, R. N. Jr., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491,** 138–142 (2012).
23. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase tem-1. *Proceedings of the National Academy of Sciences* **110,** 13067–13072 (2013).
24. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein. *RNA* **19,** 1537–1551 (2013).
25. Hinkley, T. *et al.* A systems analysis of mutational effects in hiv-1 protease and reverse transcriptase. *Nature genetics* **43,** 487–489 (2011).
26. de Visser, J. A. G. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics* **15,** 480–490 (2014).
27. Otwinowski, J. & Plotkin, J. B. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proceedings of the National Academy of Sciences* **111,** E2301–E2309 (2014).
28. Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. A learning algorithm for boltzmann machines. *Cognitive Science* **9,** 147–169 (1985).
29. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research* **41,** e121 (2013).
30. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Research* **42,** D222–D230 (2014).
31. Dehouck, Y., Kwasigroch, J. M., Gilis, D. & Rooman, M. Popmusic 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* **12,** 151 (2011).
32. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108,** E1293–E1301 (2011).
33. Sutto, L., Marsili, S., Valencia, A. & Gervasio, F. L. From residue coevolution to protein conformational ensembles and functional dynamics. *Proceedings of the National Academy of Sciences* **112,** 13567–13572 (2015).
34. Haldane, A., Flynn, W. F., He, P., Vijayan, R. S. K. & Levy, R. M. Structural Propensities of Kinase Family Proteins from a Potts Model of Residue Co-Variation. *Protein Science* **25,** 1378–1384 (2016).
35. Plefka, T. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general* **15,** 1971 (1982).

Acknowledgements

MW acknowledges funding by the ANR project COEVSTAT (ANR-13-BS04- 0012-01). This work undertaken partially in the framework of CALSIMLAB, supported by the grant ANR-11-LABX-0037-01 as part of the "Investissements d' Avenir" program (ANR-11-IDEX-0004-02).

Author Contributions

M.F. and M.W. designed research; P.B.C., M.F. and M.W. performed research; P.B.C. and M.F. analyzed data; and P.B.C., M.F. and M.W. wrote the paper.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Barrat-Charlaix, P. *et al.* Improving landscape inference by integrating heterogeneous data in the inverse Ising problem. *Sci. Rep.* **6**, 37812; doi: 10.1038/srep37812 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5

DIRECT COUPLING ANALYSIS FOR PHYLOGENETICALLY CORRELATED DATA

Statistical models of proteins sequences such as DCA are built on two assumptions: sequences $\underline{A}$ are distributed according to some fixed equilibrium distribution $P^0(\underline{A})$, and two homologous sequences found in an alignment are independent samples from P^0, *i.e.* $P(\underline{A}^1, \underline{A}^2) = P^0(\underline{A}^1)P^0(\underline{A}^2)$.
However, the evolutionary history of proteins is in evident contradiction with that second assumption. The very notion of protein family implies that present sequences derive from a common ancestor. If the divergence time between members of a family is usually long enough to result in large sequence diversity, it can also be very short for subsets of sequences. This is commonly seen in MSAs, where sequences differing only by a few amino acids are frequent.
If the branching event separating sequences $\underline{A}^1$ and $\underline{A}^2$ took place at time Δt in the past, the joint probability should be written as $P(\underline{A}^1, \underline{A}^2|\Delta t)$, *a priori* different from the product of the two equilibrium probabilities. This is made evident in the case $\Delta t = 0$, where $\underline{A}^1 = \underline{A}^2$ and $P(\underline{A}^1, \underline{A}^2|\Delta t = 0) = P^0(\underline{A}^1)\delta_{\underline{A}^1,\underline{A}^2}$. This extreme situation can be observed in protein families, where protein sequences of closely related organisms are distinct only by a few mutations.

This poses an important problem to the inference of a statistical model, as the expression of the likelihood of the data in Eq. (2.11) becomes approximate. Such an approximation leads to *biased* statistics, such as those represented in figure 5.1. For instance, closely related organisms over-represented in the family may bias statistics toward certain sequences. Likewise, figure 5.2 shows the possible appearance of spurious correlations, not due to any functional or structural interaction between residues but to the way data has been collected. Direct

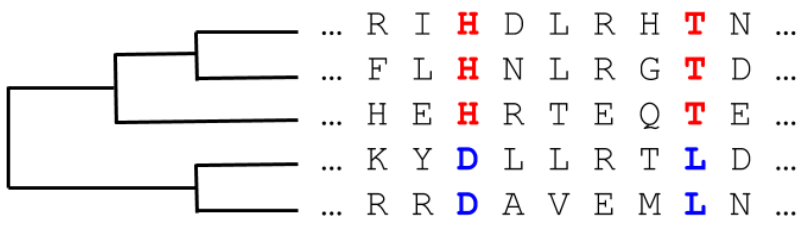

Figure 5.1: Homologous proteins constituting an MSA are related by common ancestors through a phylogenetic tree.

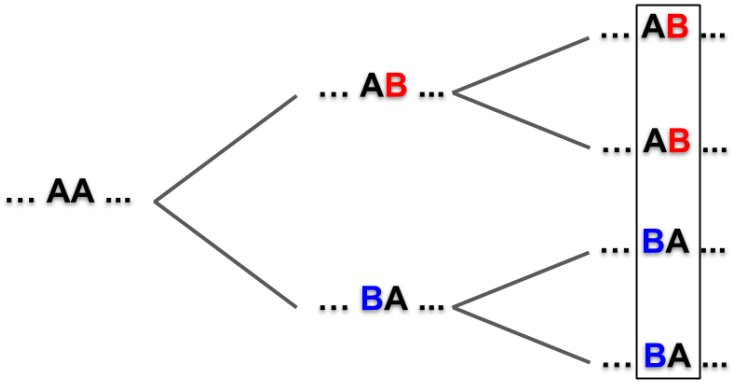

Figure 5.2: Toy example of a possible spurious correlation due to phylogenetic bias. Here, the sequence consists of two letter drawn at random. A recent common ancestor for the two top sequences and the two bottom ones biases distribution at the leaves of the tree.

inference of a DCA model thus leads to the existence of couplings parameters that attempt to model those biased statistics. As a result, the parameters of the DCA model cannot be expected to accurately represent functional constraints acting on the protein, *even if* the single sequences where distributed according to P^0.
Usual implementations of DCA use the so-called re-weighting (section 2.4.3)scheme to account for phylogeny: sequences with more than 80% identity are down-weighted, counting for one observation in total. In the $\Delta t = 0$ case, this has the correct effect of considering $\underline{A}^1$ and $\underline{A}^2$ as a single observation. In the general setting however, this is only a crude correction for the biases. Here, we aim at designing a more principled method of taking phylogenetic effects into account.

5.1 METHODS

Quantitatively, the evolutionary process can be defined by its propagator $P^0(\underline{A}^2|\underline{A}^1, \Delta t)$: the probability of observing sequence $\underline{A}^2$ knowing that it has sequence $\underline{A}^1$ as an ancestor at a time Δt in the past. For the evolutionary process to be stationary, the propagator should satisfy the condition

$$\sum_{\underline{A}^1} P^0(\underline{A}^2|\underline{A}^1, \Delta t) P^0(\underline{A}^1) = P^0(\underline{A}^2). \tag{5.1}$$

The equilibrium distribution of sequences can be recovered from this expression by taking $\Delta t \to \infty$, making sequence $\underline{A}^2$ independent from $\underline{A}^1$. If this propagator is known, the topology of the evolutionary tree allows one to have an analytical expression for the likelihood

of observing the existing sequences of the MSA using an algorithm designed by Felsenstein [30].
Let $\mathcal{T}$ stand for the evolutionary tree, with nodes of $\mathcal{T}$ indexed by n. Following [86], let $\mathcal{L}^n(\underline{A})$ be the probability of observing existing sequences that share n as an ancestor, given that the sequence of this ancestor is $\underline{A}$, and without any information on the sequences at potential intermediary nodes. If n represents a leaf node, *i.e.* an existing sequence $\underline{A}^n$, we trivially have $\mathcal{L}^n(\underline{A}) = \delta_{\underline{A},\underline{A}^n}$. For an internal node of the tree, the following recursion stands:

$$\mathcal{L}^n(\underline{A}) = \prod_{m \in \mathcal{C}(n)} \left(\sum_{\underline{B}} P(\underline{B}|\underline{A}, \Delta t^m) \mathcal{L}^m(\underline{B}) \right), \tag{5.2}$$

where $\mathcal{C}$ stands for the indexes of the children of node n, and Δt^m the time separating node m from its direct ancestor n. Figure 5.3 illustrates this idea. This recursion can be conducted from the leaves to the root r of the tree, with $\mathcal{L}^r(\underline{A})$ as a result. Since the sequence of the root of the tree is unknown, it is necessary to sum over all sequences one more time. The probability of observing existing sequences given the tree and the propagator P^0 is then

$$\mathcal{L}\left(\{A^m\},\, m \in \mathit{leaves} | P^0\right) = \sum_{\underline{A}} P^0(A) \mathcal{L}^r(\underline{A}). \tag{5.3}$$

If the propagator would depend on parameters $\boldsymbol{J}$ and $\boldsymbol{h}$ of the DCA model, it could be possible to optimize $\mathcal{L}$ over those parameters, finding the most likely Potts distribution accounting for observed sequences and their phylogenetic tree.
However, this approach suffers from two major problems. The first is that the propagator $P^0(\underline{A}^2|\underline{A}^1, \Delta t)$ associated to the Potts model is not known *a priori*. A possibility for estimating it would be to sum over all possible evolutionary trajectories from $\underline{A}^1$ to $\underline{A}^2$, but it is intractable in practice. The second is that each use of the recursion relation (5.2) involves the summation over all possible sequences of the children of node n. This amounts to summing over 20^L terms, L being the sequence length, and so for every node in the tree. Thus, a direct application of this scheme impossible for systems of realistic sizes.
The following sections propose two approximations based on the previously described idea, intending to make the computation of the likelihood tractable.

5.1.1 *Approximating dynamics: independent sites evolution*

In order to reduce the complexity of the problem, we choose to use an approximation commonly used in evolutionary biology and phylogeny. The independent sites approximation – also referred to as "single site" approximation in the following– considers each column of the MSA as evolving independently from the others. In this setting, instead of

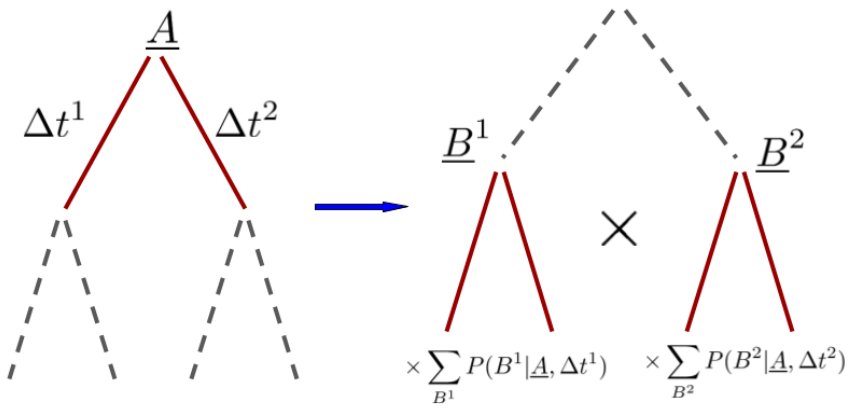

Figure 5.3: Illustration of equation 5.2. $\mathcal{L}^n(\underline{A})$, represented by the left tree, is the probability of observing configurations of the leaves given that the common ancestor is $\underline{A}$. This probability can be decomposed into a product over $\underline{A}$'s children, necessitating summation over all possible configurations of the children.

considering probabilities of observing full sequences, as in $\mathcal{L}^n(\underline{A})$, we focus on the distribution of amino acids at one MSA column only. The equivalent of equation (5.2) becomes

$$\mathcal{L}_i^n(\alpha) = \prod_{m \in \mathcal{C}(n)} \left(\sum_{\beta=1}^{q} P(\beta|\alpha, \Delta t^m) \mathcal{L}_i^m(\beta) \right), \tag{5.4}$$

where $\mathcal{L}_i^n(\alpha)$ is the probability of observing the state of column i in existing sequences that share n as an ancestor, given that the sequence of this ancestor contains α at this position. Summations over all possible configurations of internal nodes are replaced by summations over one symbol β, resulting in a complexity of $\mathcal{O}(L \times N \times q)$ for computing the L site-wise likelihood, where N is the number of internal nodes of the tree and L the length of the sequences.
In order to apply this idea, a propagator is designed using the Felsenstein model for evolution [30], using a constant mutation rate μ: in time Δt, one or more mutations happen with probability $(1 - e^{-\mu \Delta t})$. In this case, the new residue at position i is choosen according to its stationary distribution $P_i^0(\alpha) = \omega_i(\alpha)$. In the case of no mutational event, residue at i stays equal to that of its ancestor. The following propagator summarizes this process:

$$P_i(\beta|\alpha, \Delta t) = e^{-\mu \Delta t} \delta_{\alpha,\beta} + (1 - e^{-\mu \Delta t}) \omega_i(\beta). \tag{5.5}$$

Using this simple dynamical model and applying the recursion of Eq. (5.4), it is possible to compute the likelihood of the observed data in a reasonable time.

5.1.2 *Approximating dynamics: independent pairs evolution*

Using the independent sites approximation, one recovers the most likely single site stationary distribution $\omega_{i,}(\alpha)$, given the existing alignment and the topology of the evolutionary tree. However, this method is intrinsically unable to correct for spurious correlations such as that displayed in figure 5.2. A way to take two point statistics into account is therefore needed. On the other hand, performing the phylogenetic inference with a model of the full sequence is intractable, as is explained at the beginning of this section.
To deal with this dilemma, we choose to use an independent pairs approximation: each pair of sites i and j is thought of as evolving independently from the others, with a propagator similar to that of Eq. (5.5). The probability in time δt that i changes from α to γ, and j from β to δ is defined as

$$\begin{aligned} P_{ij}(\gamma,\delta|\alpha,\beta,\Delta t) =& e^{-2\mu\Delta t}\delta_{\alpha,\gamma}\delta_{\beta,\delta} \\ &+e^{-\mu\Delta t}(1-e^{-\mu\Delta t})\left(\delta_{\alpha,\gamma}\omega_{ij}(\delta_j|\gamma_i)+\delta_{\beta,\delta}\omega_{ij}(\gamma_i|\delta_j)\right) \\ &+(1-e^{-\mu\Delta t})^2\omega_{ij}(\gamma,\delta)\ , \end{aligned} \tag{5.6}$$

where $P^0_{ij}(\gamma,\delta)=\omega_{ij}(\gamma,\delta)$ is the stationary pairwise distribution at sites i and j, and $\omega_{ij}(\gamma|\delta)=P^0_{ij}(\gamma,\delta)/P^0_i(\delta)$ is the conditional probability of observing γ in i knowing δ in j. In turn, the Felsenstein's recursion relation becomes

$$\mathcal{L}^n_{ij}(\alpha,\beta)=\prod_{m\in\mathcal{C}(n)}\left(\sum_{\gamma,\delta=1}^{q}P(\gamma,\delta|\alpha,\beta,\Delta t^m)\mathcal{L}^m_i(\gamma,\delta)\right). \tag{5.7}$$

The summation over all possible configurations of two sites and the computation of the likelihood for all pairs now results in a still feasible complexity of $\mathcal{O}(L^2\times N\times q^2)$.
Of course, a naive application of this method poses a major consistency problem: two pairs sharing one residue *cannot* evolve independently. As a result, the inference of the most likely pairwise statistic $\omega_{ij}(\alpha,\beta)$ for each pair will give inconsistent results. For three residues i,j and k, one might have

$$\sum_{\beta=1}^{q}\omega_{ij}(\alpha,\beta)\neq\sum_{\gamma=1}^{q}\omega_{ik}(\alpha,\gamma). \tag{5.8}$$

To settle this inconsistency, we propose to optimize for the most likely pairwise distribution under the constraint that its partial sums correspond to the single site distribution obtained using the independent

site approximation scheme (superscript *is*). In other words, for all i and j, the following will stand:

$$\sum_{\beta=1}^{q} \omega_{ij}(\alpha,\beta) = \omega_i^{is}(\alpha) \quad \text{and} \quad \sum_{\alpha=1}^{q} \omega_{ij}(\alpha,\beta) = \omega_j^{is}(\beta), \tag{5.9}$$

where $\omega_i^{is}(\alpha)$ stands for the result of the scheme described in 5.1.1.
The hope is that by extending the phylogenetic inference beyond a site-wise description, the background pairwise statistics of the evolutionary process might be recovered, therefore improving the inference of the DCA coupling parameters.

5.1.3 *Optimization: maximizing the likelihood*

The independent sites or independent pairs approximations allow for a computationally efficient estimation of the likelihood. In order to correct empirical frequencies f for phylogenetic bias, it is now needed to find stationary frequencies ω that maximize the approximated likelihood: equation (5.4) (resp. (5.7)) has to be optimized over $\omega_i(\alpha)$ (resp. $\omega_{ij}(\alpha,\beta)$). Since each site i or each pair (i,j) is independent from the others (depending on the approximation used), the optimization is conducted over either q or q^2 parameters. However, the gradient of the likelihood in both approximations is intractable, and its concavity is unknown, making the use of standard gradient ascent techniques impractical.
Here, we rely on a stochastic optimization scheme which was empirically found to be efficient in this scenario, inspired from [22]. Parameter space – *i.e.* the $\omega_i(a)$ or the $\omega_{ij}(a,b)$ – is randomly sampled by making global or local random moves: in global moves, all parameters to be optimized are simultaneously changed, while in local moves only one is changed. The moves are only accepted if they lead to an increase if the likelihood. Their magnitude is decreased throughout the optimization, starting with large displacement in parameter space and ending with small adjustments. After a pre-defined number of moves are made, and the best parameters found are returned.
This scheme is rather empirical and does not guarantee convergence. However, in testing scenarios where the stationary frequencies ω are known, it was found to always lead to the correct solution.

In the case of the independent pairs approximation, $\omega_{ij}(\alpha,\beta)$ needs to be optimized under the constraints defined in equation (5.9). For this reason, moves proposed by the stochastic exploration of parameter space need to satisfy the constraints at all times. Here, we use a re-parametrization trick inspired by the definition of Direct Information (see appendix a.1): temptative pair frequencies are written as

$$\omega_{ij}(\alpha,\beta) = \frac{1}{Z(J,h_i,h_j)} \exp\left(J(\alpha,\beta) + h_i(\alpha) + h_j(\beta)\right). \tag{5.10}$$

The optimization is then conducted over the coupling parameter J. Whenever J is changed, compensatory fields h_i and h_j are re-estimated in order to satisfy the marginalization constraints. In this way, optimization is conducted in the space of frequencies that do satisfy equation (5.9).

5.1.4 *Inferring* DCA *models based on corrected statistics*

To infer Potts models based on frequencies corrected through the method describe above, we used the PLM method (see section 2.3.2). Its main advantage is its statistical consistency, meaning that with enough *i.i.d.* samples we are guaranteed to recover the original model (if it is of the Potts form). However, the PLM inference is not directly based on frequencies f_{ij} but on the samples themselves. The methods described above, however, do not yield corrected samples but corrected frequencies.

In order to use the PLM inference, we designed a way to construct a sequence alignment which has a given target pairwise statistic f_{ij}^{target}, using a simulated annealing strategy based on the work in [75]. The idea is to start with an alignment having the correct target profile f_i^{target}. Single variables a_i^m are then permuted in the following way: at each move t, a column i and two lines m and n are chosen at random, and an attempt to exchange a_i^m and a_i^n is made. The probability of the exchange to take place is

$$P(exchange) = min\left(1, \exp\left(\beta||\boldsymbol{C}^{t+1} - \boldsymbol{C}^{target}|| - \beta||\boldsymbol{C}^{t} - \boldsymbol{C}^{target}||\right)\right), \tag{5.11}$$

where $\boldsymbol{C}^t$ and $\boldsymbol{C}^{t+1}$ are the connected correlation matrices of the current alignment before and after the exchange, $\boldsymbol{C}^{target}$ the correlation matrix corresponding to the target frequencies, $||\cdot||$ stands for the Frobenius norm of matrices, and β is an inverse temperature parameter. Thus, a move is more likely to be accepted if it makes the connected correlation matrix of the alignment closer to that of the target. Parameter β is initialized at a low value and slowly increased as more moves are made. In this way, when β goes to infinity, we hope to have $\boldsymbol{C} \longrightarrow \boldsymbol{C}^{target}$.

Importantly, this procedure never changes the single point marginals of the alignment, since exchanges are made inside one column. Because the starting point has the correct single point marginals f_i^{target}, the designed sample will always keep these marginals. In this way, since we expect $\boldsymbol{C} \longrightarrow \boldsymbol{C}^{target}$ as the temperature goes to zero, we will also have $f_{ij} \longrightarrow f_{ij}^{target}$. In practice, the result of this procedure gives a Frobenius norm $||\boldsymbol{C} - \boldsymbol{C}^{target}||$ typically smaller than noise due to finite size samples on each $f_{ij}(a, b)$.

This procedure allows us to construct a sample based on the corrected pairwise frequencies ω_{ij}, using the independent pairs approximation described above: the target frequencies are simply set to the ones resulting from the optimization of the likelihood: $f_{ij}^{target} = \omega_{ij}$. However, this is not possible when using the independent site correction, since only the single site frequencies ω_i are corrected. In this case, we build an artificial pairwise frequencies matrix defined by

$$\omega_{ij}(a,b) = f_{ij}(a,b) - f_i(a)f_j(b) + \omega_i(a)\omega_j(b), \tag{5.12}$$

The pairwise statistics defined in this way will have the corrected single site frequencies as marginals, but uncorrected connected correlations.
However, a major drawback of this method is that this manner of combining different frequencies gives rise to inconsistencies, with some terms $\omega_{ij}(a,b)$ being larger than 1 or smaller than 0. It is therefore impossible for our simulated annealing procedure to construct an alignment exactly reproducing these frequencies.
Once the corrected pairwise statistics are computed and a corresponding alignment built, the PLM method is used to infer the DCA model.

5.2 RESULTS: TOY MODEL

5.2.1 *Design of the toy model*

In order to test the methodology, we first try our methods on a toy model. As the aim of correcting data for phylogenetic bias is ultimately to have a better DCA inference, we choose our toy model to be of the Potts form. In this manner we know that without any phylogeny and with enough samples from the toy model, the parameters $\boldsymbol{J}$ and $\boldsymbol{h}$ should be recovered with high accuracy.
For computational efficiency, the length of the model is set as $L = 25$, with $q = 4$ states for its variables. Couplings and fields are drawn from a normal distribution, with couplings taking a ferromagnetic form:

$$J_{ij}^0(a,b) = s_{ij}x_{ij}^J \cdot \delta_{a,b} \quad \text{and} \quad h_i^0(a) = x_i^h(a), \tag{5.13}$$

where $\{x_{ij}^J\}, i,j \in \{1\dots L\}$ and $\{x_i^h(a)\}, i \in \{1\dots L\}, a \in \{1\dots q\}$ are gaussian variables:

$$x_{ij}^J \sim \mathcal{N}(\mu_J, \sigma_J) \;\text{ and }\; x_i^h \sim \mathcal{N}(\mu_h, \sigma_h), \tag{5.14}$$

and s_{ij} are discrete variables taking values in $\{0,1\}$:

$$s_{ij} = \begin{cases} 1 & \text{with probability } c/L, \\ 0 & \text{with probability } 1 - c/L. \end{cases} \tag{5.15}$$

To mimic the effect of structural contacts, we dilute the couplings by taking a value of $c = 3$, making the $\boldsymbol{J}^0$ matrix sparse. Therefore, each site i shares a direct coupling J_{ij} with on average 3 other sites j.

5.2.2 *Artificial data*

To simulate the effect of phylogeny, we sample the toy model P^0 using MCMC on a binary tree. Two MCMC chains are initialized from a root configuration, itself drawn from a fair sample of P^0, resulting in two new configurations. This process is then iterated, taking the two resulting configurations as new roots, thus growing the tree. For K iterations – "duplications" –, the resulting tree will have 2^K leaves. For each MCMC run, a number of "mutations" is drawn from a Poisson distribution with parameters $\mu\tau$. For each of those mutations, a site i is chosen at random and its new state is drawn from the local conditional probability $P^0(a_i|\underline{A}_{\setminus i})$ in a Gibbs sampling manner.
This scheme guarantees that the number of mutational events will correspond to dynamical models in Eqs. (5.5) and (5.6). However, the way residues are re-drawn after a mutation depends on the full current sequence through distribution P^0, unlike the simplifying assumptions of the propagators.
For simplicity reasons, $\mu\tau$ is set to be identical for all branches of the tree, taking values 3, 5 or ∞ (*i.e.* $\mu\tau \gg L$), resulting in respectively strong, weak and absent phylogenetic effects. In the following, the samples corresponding to finite values of τ will be referred to as *biased* samples, while the one corresponding to $tau \to \infty$ will be referred to as a "fair" or *i.i.d.* sample. 12 duplication events are performed, resulting in a tree of $2^{12} = 4096$ leaves and $2^{12} - 1$ internal nodes. Finally, in order not to depend on the particular choice of the root configuration, 30 repetitions of the sampling process are performed for each τ.
For concision of the main text, only results concerning the $\mu\tau = 3$ are shown. This represents the hardest case, as phylogeny effects are more pronounced for short branch lengths. Results for the $\mu\tau = 5$ case are shown in appendix b in the form of figures.

5.2.3 *Phylogenetic inference corrects one and two points statistics*

To assess the quality of the phylogenetic correction, we first compare single site and pairwise statistics before and after our inference to the same observables measured in an *i.i.d.* sample drawn from P^0.
In the case of the independent sites approximation, the single site statistics are corrected. This observable as measured in the biased sample – *i.e.* sample coming from the leaves of the tree, without correction, referred to in the figures as the "tree" sample – will be referred to as f_i^t. After correction, we refer to it as f_i^{inf}, and as f_i^0 in the case of the

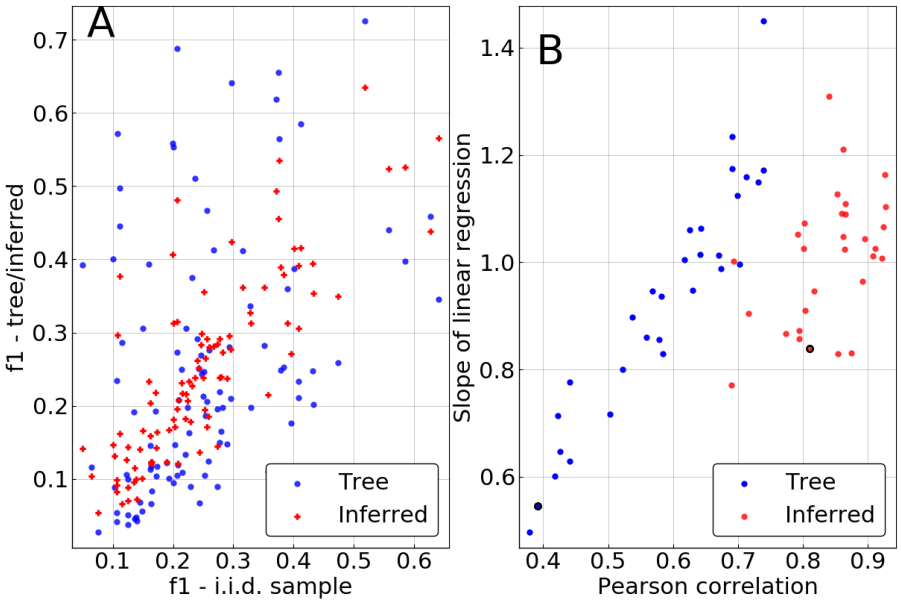

Figure 5.4: Result of the single site phylogenetic inference. **A.** Single site statistics of a sample of P^0 coming from a tree, before ("Tree") and after ("Inferred") the phylogenetic inference, against "true" single site statistics coming from the fair sample. **B.** Slope of the linear regression and pearson correlation corresponding to the plot in **A**, for the 30 repetitions of the experiment. The black-circled points correspond to the repetition displayed in **A**.

i.i.d. sample.
As demonstrated in figure 5.4, the inference clearly improves the estimation of single site frequencies over naive counting in the biased sample. Pearson correlations between f_i^{inf} and f_i^0 are significantly higher than between f_i^t and f_i^0, being larger than 0.8 in 29 out of 30 repetitions. This contrasts with the remarkably low correlations of 0.4 that can be achieved for some realizations of the tree if no correction is performed. Similarly, the slope of a linear regression of f_i^{inf} against f_i^0 tends to be much closer to 1 in most cases, also showing lower variation from repetition to repetition.

A similar comparison is made for pairwise frequencies in the case of the independent pairs approximation. We now compare f_{ij}^t and f_{ij}^{inf} to their counterpart from the *i.i.d.* sample f_{ij}^0. The two top panels of figure 5.5 once again show an improvement resulting from the phylogenetic inference, as pairwise statistics are closer to match f_{ij}^0 after it is performed.
However, one has to keep in mind that some of this improvement is due to the single site correction. Indeed, in the independent pairs approximation, marginals of the pairwise frequencies are constrained to match the corrected single site frequencies f_i^{inf}. In order to evaluate the intrinsic quality of the pairwise method, we focus on the connected

correlations $c_{ij} = f_{ij} - f_i f_j$, thus removing the influence of the single site correction. Bottom panels of figure 5.5 demonstrate that even this intrinsically pairwise quantity is recovered with higher accuracy after inference. Even our very crude approximation – considering every pair as evolving independently – can correct some of the statistical bias due to phylogeny, improving over naive counting in the MSA.

5.2.4 DCA *parameters are recovered with increased accuracy*

We infer DCA models based both on the uncorrected and the corrected frequencies f_{ij}^t and f_{ij}^{inf} using the methodology described in section 5.1.4. To evaluate both of our approximations, we infer the DCA model in the case of the single site correction and the independent pair correction.
In the top panel of figure 5.6, inferred parameters are then compared to the true ones $\boldsymbol{J}^0$ and $\boldsymbol{h}^0$ using Pearson correlation as a measure. Both methods – single site and independent pairs, labeled as pairwise in the figures – lead to a significant improvement in the inference of fields. However, the inference of couplings is deteriorated when using only the single site correction, whereas it is improved in the pairwise case. This may be due to the inconsistencies appearing when combining correlations from the biased sample with corrected single site frequencies, as is explained in section 5.1.4. Indeed, such inconsistencies (frequencies larger than 1 or smaller than 0) were observed for all of the 30 repetitions.
To understand if the inferred DCA models are a better fit to the true distribution, we compute their symmetric Kullback-Leibler distance to P^0:

$$D_{KL}(Q||P^0) + D_{KL}(P^0||Q) = \langle \mathcal{H}_Q - \mathcal{H}_{P^0} \rangle_{\mathcal{H}_{P^0}} + \langle \mathcal{H}_{P^0} - \mathcal{H}_Q \rangle_{\mathcal{H}_Q}. \quad (5.16)$$

An explanation for the use of this quantity is given in the Methods of the article of chapter 4. Shortly, while the standard KL-distance depends on the intractable calculation of the partition function of one of the distributions, its symmetrized version can be easily estimated by MCMC sampling. Figure 5.6 shows a histogram of this quantity for the 30 repetitions of each sampling process. A clear ranking between methods appears, with the inference based on the biased sample being the worse. Both phylogenetic corrections result in a model that is close to P^0, with an advantage to the pairwise method. Surprisingly, the decrease in inference quality of the couplings when using the single site correction does not appear to have a strong influence on Kullback-Leibler distance, as there is a very large drop of this quantity between a biased sample based or a single site correction based DCA. However, this does not stand in the case of "contact prediction", where the single site correction based DCA performs significantly worse than others. Since our artificial model is sparse in the sense that most of

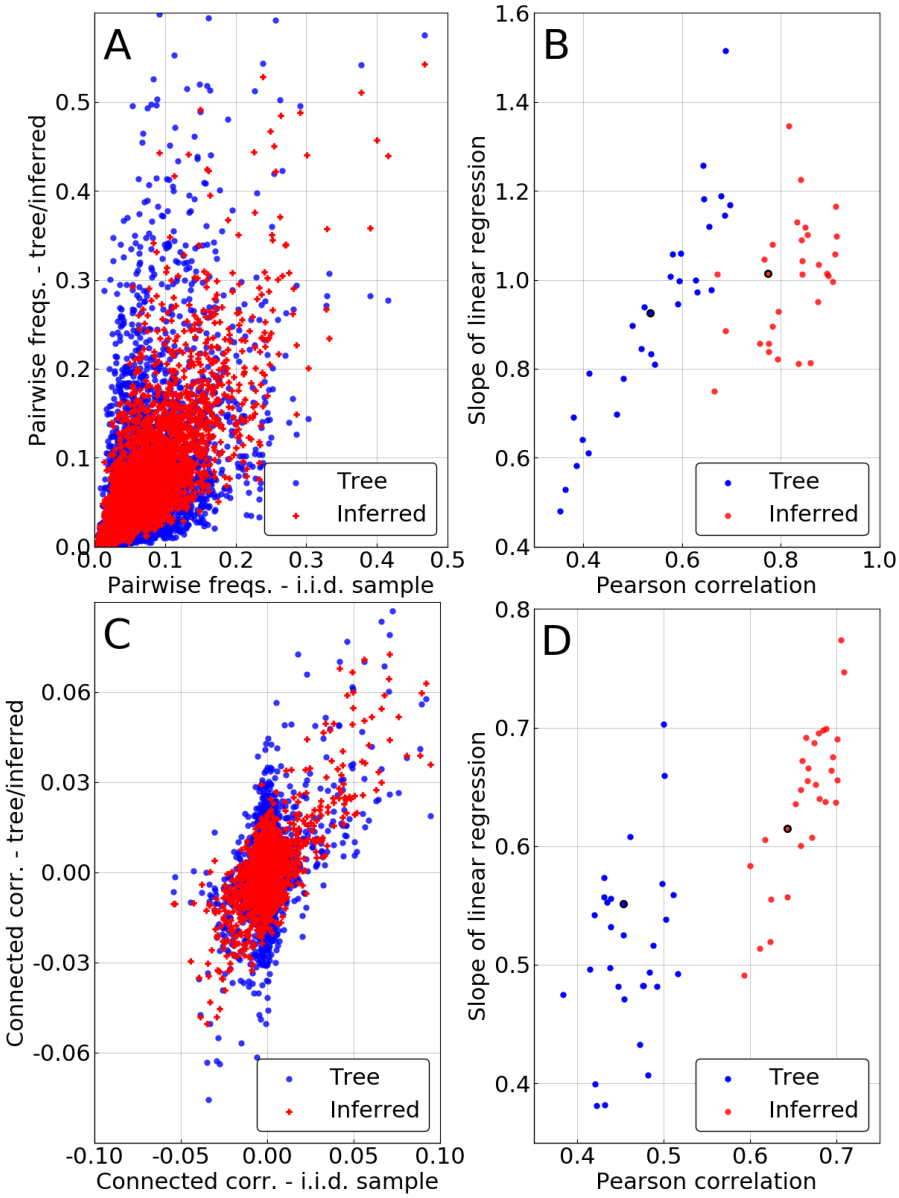

Figure 5.5: Result of the pairwise phylogenetic inference. **A.** Pairwise frequencies $f_{ij}(a,b)$ of a sample of P^0 coming from a tree, before ("Tree") and after ("Inferred") the phylogenetic inference, against "true" pairwise frequencies coming from the fair sample. **B.** Slope of the linear regression and pearson correlation corresponding to the plot in **A**, for the 30 repetitions of the experiment. The black-circled points correspond to the repetition displayed in **A**. **C.** Same as **A** for connected correlations $c_{ij} = f_{ij} - f_i f_j$. **D.** Same as **B** for connected correlations.

the J^0_{ij} matrices are 0, we can use the inferred couplings as "contact" predictors, where a contact is defined as a pair (i, j) for which the coupling J^0_{ij} is non zero. In this case, there is only a slight improvement in using the pairwise phylogenetic correction and even a drop in prediction quality when using the single site method.

5.2.5 *Improvement in the prediction of single mutant's energies*

One of the most promising application of DCA-like methods is the ability to infer the effect of mutations in proteins from the MSA of homologs. Here, we want to investigate the potential of our phylogenetic correction to enhance accuracy of these predictions. To recreate this setting in our toy model, we consider single mutants of "wild-type" artificial sequences. Wild-types can be taken either in the the phylogenetically biased sample, as would be the case in standard DCA, either in the *i.i.d.* sample. For each "wild-type" sequence $\underline{A}^m$, all of its $L \times (q-1)$ single mutants are denoted by $\{\underline{A}^m_\alpha\}$, $\alpha \in \{1 \dots L \times (q-1)\}$. For each of those, the effect of the mutation is defined to be the difference of energy between $\underline{A}^m$ and the mutant:

$$\Delta\mathcal{H}_{m,\alpha} = \mathcal{H}(\underline{A}^m_\alpha) - \mathcal{H}(\underline{A}^m). \tag{5.17}$$

$\mathcal{H}$ can be either the true Hamiltonian $\mathcal{H}^0$, then defining the true mutational effect, or an inferred one, corresponding to the inferred mutational effect. In order to evaluate the influence of both the phylogenetic correction and the DCA methodology on the quality of predictions, we choose to also infer a profile model as a comparison point. As described in the article of chapter 3, profile models have vanishing couplings and reproduce the single site statistics f_i using fields only. They have been used with success for predicting mutational effects in proteins based on the conservation profile of the MSA.

We first focus on the single site phylogenetic correction. For each model, profile $\mathcal{H}^{prof}$ and DCA $\mathcal{H}^{dca}$, and for each statistic, uncorrected f^t_i and corrected f^{inf}_i, we compute the Pearson correlation between $\{\Delta H_{m,\alpha}\}_\alpha$ and the correct energies $\{\Delta H^0_{m,\alpha}\}_\alpha$. This is repeated for each sequence $\underline{A}^m$ in either the biased or the *i.i.d.* sample, and all resulting Pearson correlation are averaged into one score representing the quality of predictions of the energies of single mutants with wild-types in a given sample.
As is shown in figure 5.7, when the reference sequence is taken in the biased sample, all methods seem to perform equally well, apart from the profile model inferred on the biased frequencies. In particular, applying the DCA methodology and thus attempting to fit correlations or using a simple profile model on corrected data seems to result in the same improvement.
The picture changes when the reference sequence is taken in a fair

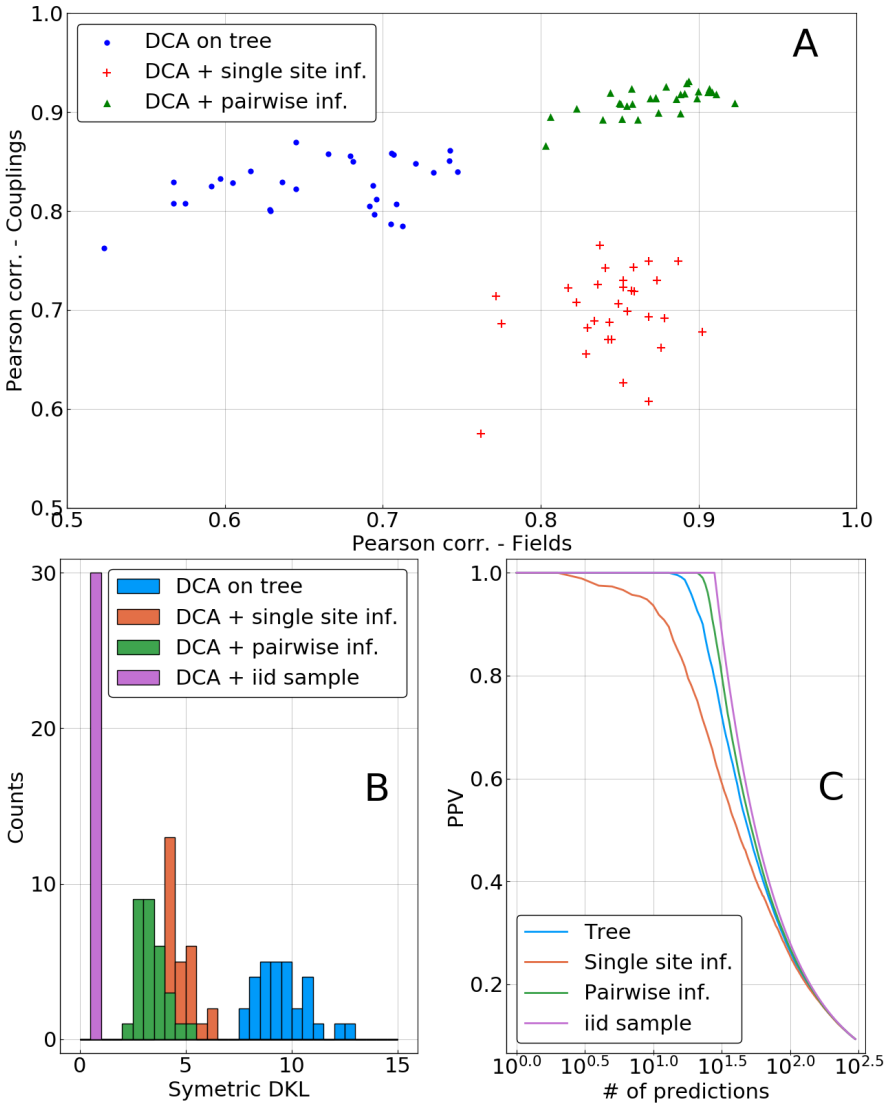

Figure 5.6: DCA model inferred after single site or pairwise phylogenetic correction **A.** Pearson correlation between parameters of inferred and of true DCA models. y-axis: couplings J_{ij} ; x-axis: fields h_i. One point corresponds to one repetition of the MCMC process on the tree. **B.** Histogram of the symmetric Kullback-Leibler distances between inferred and true models for all repetition. **C.** Positive predictive value for predicting non zero couplings (*i.e.* "contacts") using inferred DCA models. DCA inferred on the *i.i.d.* sample performs perfectly in this case.

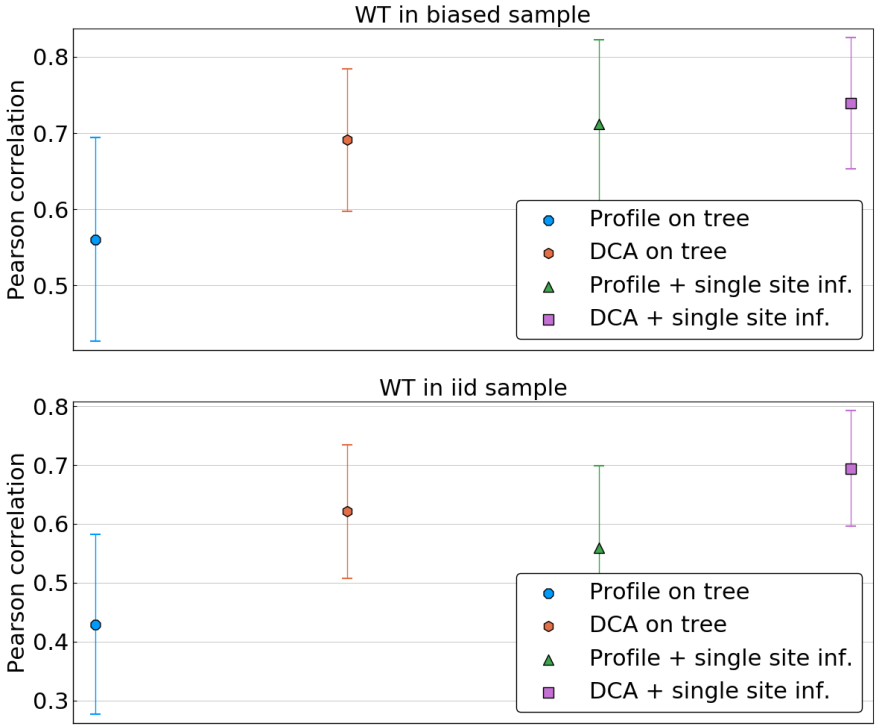

Figure 5.7: Pearson correlation in predicting energies of single mutants averaged over sets of reference sequence. In the top panel, reference sequences are taken in the biased sample, *i.e.* among the leaves of the phylogenetic tree. In the bottom panel, reference sequences are taken in a fair sample of P^0. Predictions are made using four models: respectively a profile model and a Potts model trained on the uncorrected biased sample (resp. "Profile on tree" and "DCA on tree"), and using the corrected single site frequencies (reps. "Profile + single site inf." and "DCA + single site inf."). Error bars indicate the standard deviation across the 30 repetitions of the tree sampling process.

sample. In this case, the performance of both DCA on uncorrected data and of the profile models drop significantly, whereas DCA inferred on corrected frequencies remains as accurate. To investigate this further, we compute the average Pearson correlation as a function of the Hamming distance of the wild-type to the closest sequence in the biased sample. Figure 5.8 shows that while the performance of the uncorrected DCA and the profile models declines rapidly when using a reference sequence far away from the biased sample, the corrected DCA has a more stable performance before large hamming distances are reached.

Since the combination of DCA and of the single site phylogenetic correction outperforms profile models or a naive DCA approach, we now consider inferring the Potts model based on the corrected pairwise frequencies. The same scoring as above is used, using all single

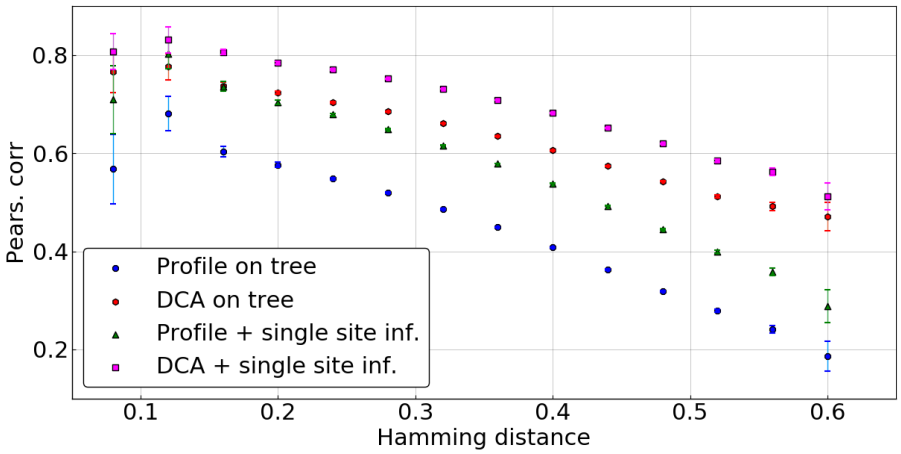

Figure 5.8: Pearson correlation in predicting energies of single mutants averaged over reference sequence at a given hamming distance to the closest sequence in the biased sample, as a function of this hamming distance. Error bars are inversely proportional to the square root of the number of sequences in each hamming distance bin. Profile and Potts models are inferred either directly using biased data, or using corrected single site frequencies.

mutants for wild-type sequences in both samples and computing the average Pearson correlation across wild-types. Figure 5.9 compares the predictions of the DCA models using the tree levels of phylogenetic correction: none, site-wise and pairwise. The latter leads to a significant improvement in accuracy of predictions, outperforming the two other methods. This stands both in the case of a wild-type belonging to the biased sample or to the fair sample.
Again, we investigate the dependence of those predictions on the distance of the wild-type to the closest sequence in the biased sample. The largest increase in Pearson correlation resulting from the pairwise phylogenetic inference once again happens for sequences that are far from the biased sample (figure 5.10). Removing part of the phylogenetic bias seems to have a stronger influence when considering the energy landscape around sequences that are far away from the leaves of the phylogenetic tree. When using those leaves as a sample without accounting for their non-independence, the resulting model seems not to learn much about the energy landscape far away from those points. However, correcting for non-independence even in a rather crude way leads to a much better inference in this regard.

This phenomenon is also apparent in figure 5.11, where the average energy of sequences of the *i.i.d.* sample is shown as a function of the hamming distance to the biased sample. Sequences of the leaves of the tree being "typical" of P^0, it is unsurprising to see that the average energy computed in $\mathcal{H}^0$ increases as the sequence gets further away from them. However, when inferring a Potts model using uncorrected

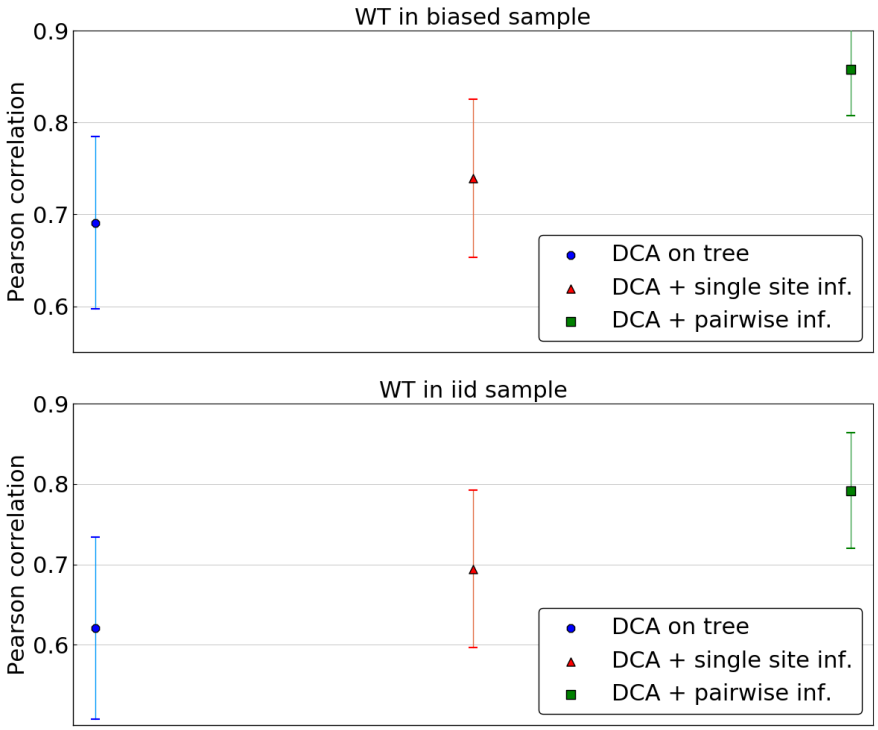

Figure 5.9: Pearson correlation in predicting energies of single mutants averaged over sets of reference sequence. In the top panel, reference sequences are taken in the biased sample, *i.e.* among the leaves of the phylogenetic tree. In the bottom panel, reference sequences are taken in a fair sample of P^0. Predictions are made using a DCA model inferred either directly on biased data, either using corrected single site frequencies, either using corrected pairwise frequencies. Error bars indicate the standard deviation across the 30 repetitions of the tree sampling process.

statistics from the leaves sequences, this increase in average energy is much more abrupt. Energy differences between "typical" sequences and distant ones in $\mathcal{H}^0$ are or the order of 10, but go up to 20 for a model inferred on biased frequencies. In this scenario, the result of the phylogenetic correction is to correct this overshooting, resulting in a more accurate inference of the energies of distant sequences.

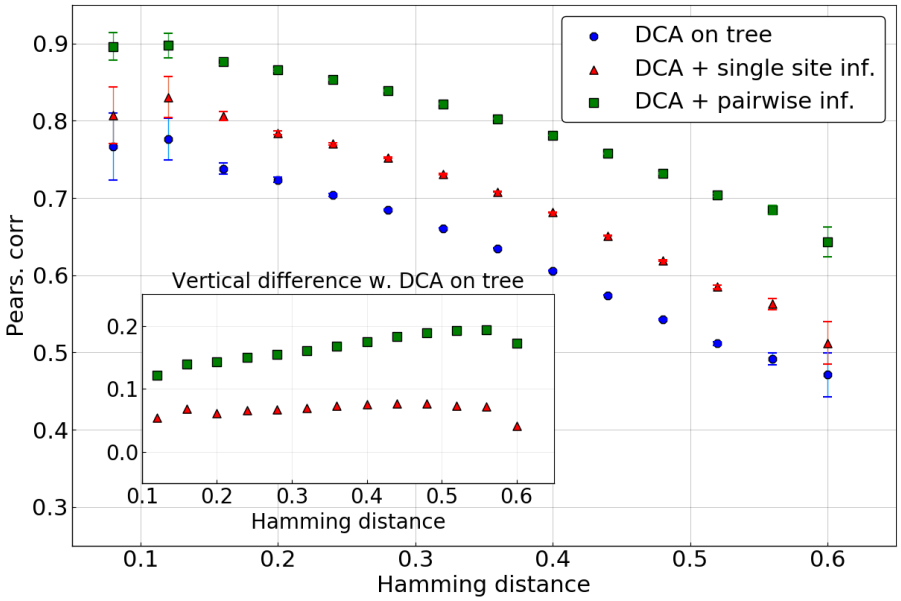

Figure 5.10: Pearson correlation in predicting energies of single mutants averaged over reference sequence at a given hamming distance to the closest sequence in the biased sample, as a function of this hamming distance. Error bars are inversely proportional to the square root of the number of sequences in each hamming distance bin. The Potts model is inferred either directly on biased data, either using corrected single site frequencies, either using corrected pairwise frequencies.

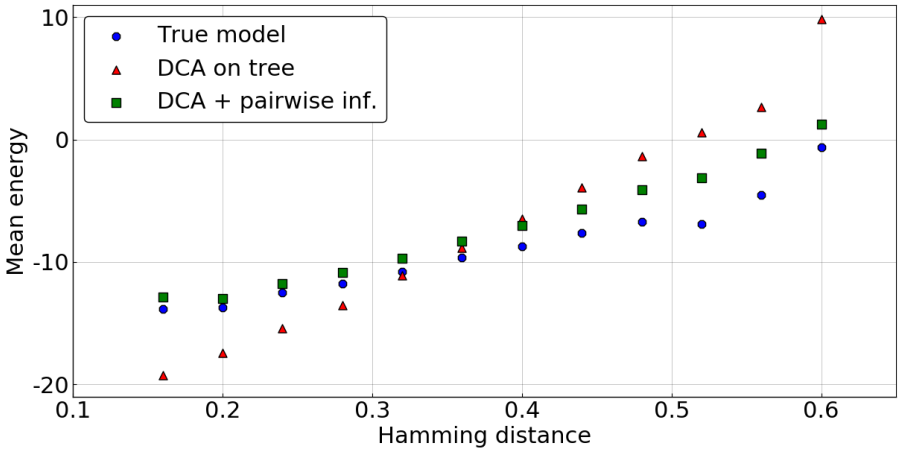

Figure 5.11: Average energy of sequences as a function of the hamming distance of the sequence to the closest point in the biased sample.

6 SOME RESULTS AND OPEN QUESTIONS

According to the results of chapter 3, DCA seems like a good model to capture sequence variability in a protein family. Rather than just a tool to make predictions about the structure of the protein, it appears to capture some of the evolutionary constraints acting on the sequences, explaining its accuracy in reproducing statistical patterns of an Multiple Sequence Alignment, both fitted and not fitted by the method. If these constraints are indeed grasped by the model, it is not surprising that DCA succeeds in predicting diverse features of proteins such as structural contacts, effects of mutations or potential functionality of sequences (*c.f.* section 2.5).
This calls for a deeper understanding of the method and the way it operates. In particular, if constraints acting on sequences are well captured by DCA, we should be able to interpret the inferred parameters in a biologically meaningful way. However, DCA is far from being perfect. It infers a very large number of parameters – $\mathcal{O}(L^2 \cdot q^2)$ coupling parameters for a family with sequences of length L – using a relatively small number of sequences, resulting in severe undersampling. Moreover, typical MSAs may suffer from biases such as those induced by phylogeny, making the disentangling of truly evolutionary constraints and other statistical signal non trivial.
In this section, we question the DCA modeling by asking the question of its robustness and of the interpretation of its parameters.

All results in the following will be shown on the PF00072 family, also used as an example in the article of 3. Most of them – *e.g.* the non-sparsity of coupling matrices or the difficulty to structurally interpret indirect effects on correlation – are robust across protein families studied in chapter 3. The sparse modeling of section 6.2 or the construction of cliques in section 6.1.2 were tested on relatively few families for computational reasons, still giving consistent results across those.

6.1 INTERPRETING "DIRECT" COUPLINGS

6.1.1 *Coupling matrices are not sparse*

As the acronym suggests, couplings obtained with Direct Coupling Analysis (DCA) should represent a *direct interaction*, with biological meaning. The most straightforward reason for such interactions to exist is a structural contact between two residues in the protein fold. The

ability of DCA to predict some of these contacts is a good demonstration that the strongest couplings are interpretable in biological terms. There could be other ways through which two residues may interact. The dimerization of some protein domains creates interfaces, resulting in new contacts that are unseen in a single structure. The folding process of the protein could also be the source of co-evolution between residues. However, all these potential sources of interaction between residues are expected to be sparse to some extent: most residue pairs in the protein should not interacting *directly*.

Yet, as the DCA inference allows for a coupling J_{ij} between any pairs of columns in the MSA, most of the resulting parameters are non-zero. Figure 6.1 shows the distribution of the Frobenius norm of couplings $||J_{ij}||^2 = \sum_{a,b=1}^{q} J_{ij}(a,b)^2$ for the PF00072 family. The parameters have been inferred using the BML implementation of DCA, thus reproducing very accurately statistical features of the alignment (cf. article of chapter 3). It is clearly apparent that DCA introduces a finite coupling for all pairs of residues. The majority of Frobenius norms are around 0.4, only three to four times weaker than the strongest couplings. Whereas the strong couplings almost always represent contacts, the majority of the parameters do not. Indeed, the fraction of couplings corresponding to structural contacts for Frobenius norms around 0.4, where the majority of parameters lie, is roughly equivalent to the overall fraction of contacts in the protein fold. In other words, the vast majority of coupling parameters do not have any interpretation in terms of structural contact.

However, figure 6.2 shows that those parameters seem to be essential for the model to accurately fit the data. In this figure, couplings are sparsified by setting to zero those with a Frobenius norm smaller than some threshold F^{th}. Larger couplings remain unchanged. Properties of the resulting model such as mean energies of natural sequences or fitting quality are evaluated as a function of the Frobenius norm threshold. When F^{th} reaches a value of 0.5, in practice removing weak couplings that do not contain any structural signal according to figure 6.1, the energies of natural sequences measured in the model are heavily changed, and the fitting quality of correlations drops to zero. When structurally meaningful couplings start to be decimated – approximatively $F^{th} = 0.8 - 0.9$, the energies of natural sequences have already been completely modified, rising by about $\Delta E = 150$, and the fitting quality is close to zero for correlations.

This leads to a paradoxical situation. On the one hand, DCA is able to extract structural information from sequence data by disentangling direct and indirect effects through the use of direct couplings. On the other hand, the vast majority of inferred parameters, though necessary for the model to reproduce statistical patterns found in the MSA, cannot be interpreted in terms of structural contacts.

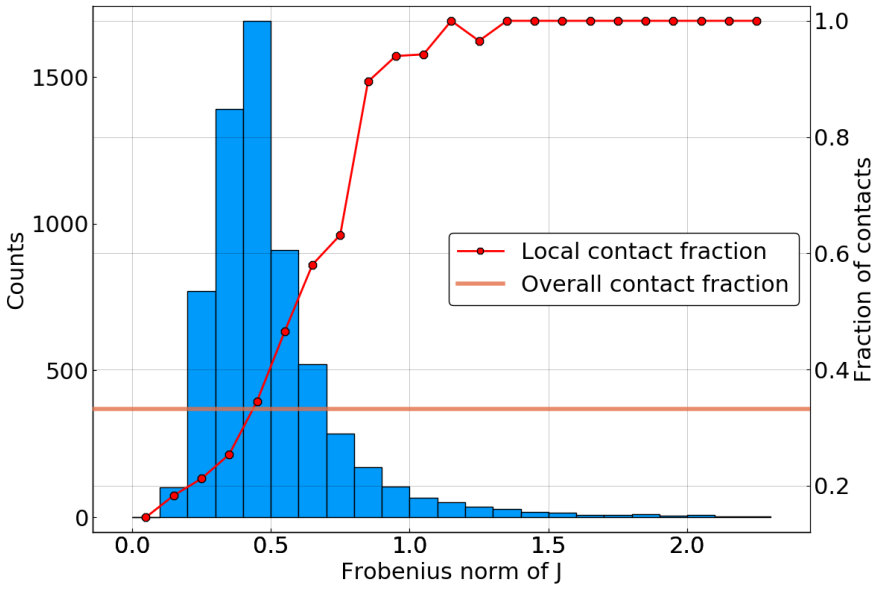

Figure 6.1: **Left axis.** Distribution of the Frobenius norm of couplings $||J_{ij}||$ for a BML inference on the PF00072 family. Gauge used is the zero-sum. Most couplings have finite non-zero values. **Right axis.** Fraction of couplings which correspond to a structural contact for a given value of the Frobenius norm. Strong couplings are clearly structurally interpretable, but the bulk of the distribution is not.

6.1.2 *Chains and networks of couplings*

The ability of DCA to predict contact is usually attributed to its capacity to disentangle two different sources of correlations between columns of the MSA, namely a direct interaction between residues and indirect effects mediated through intermediary residues. As it appears that DCA provides an impressively good model for sequence variability in a protein family – in terms of structural information, but also in terms of mutational landscapes and maybe for designing artificial sequences – it becomes important to understand if parameters of the model can be interpreted. In particular, networks of couplings that mediate correlation between structurally distant residues should themselves be structurally interpretable.

The article of chapter 3 introduced the use of chains of couplings as a measure of indirect correlation effects. It was shown there (figure 2. of the article) that it is possible to extract some structural signal from those chains: pairs of residues strongly linked by all chains of length 2 are typically closer than random residues, and yet further away than pairs with a strong direct coupling. However, it does not seem possible to explain the existence of pairs of strongly correlated but distant residues through individual chains of couplings. Indeed, when considering the strongest chains, it was found that any individual indirect effect is very weak compared to the direct coupling, and that

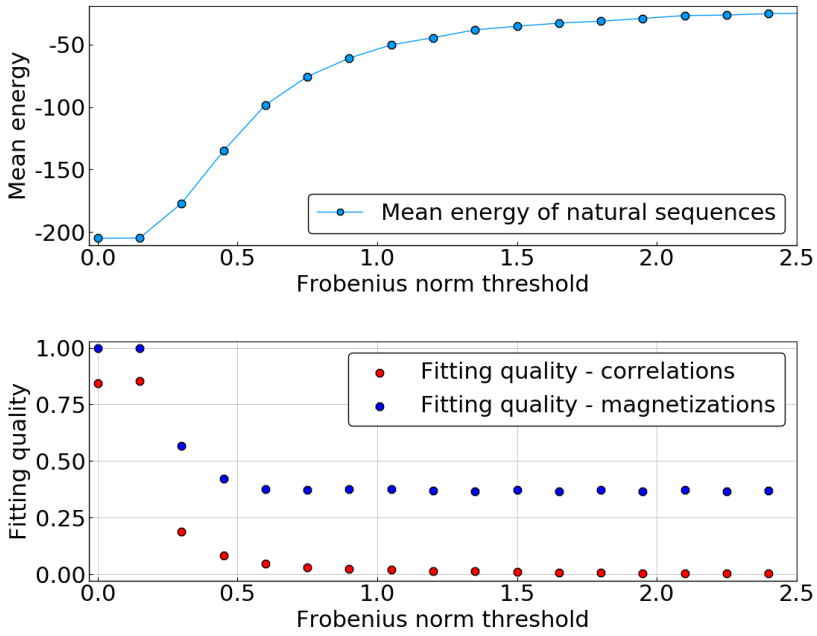

Figure 6.2: Decimating the DCA model for PF00072 by removing couplings with Frobenius norms smaller than a threshold. **Top.** Mean energy of natural sequences in the model as a function of the Frobenius norm threshold. **Bottom.** Fitting quality – Pearson correlation between observables $c_{ij} = f_{ij} - f_i f_j$ and f_i as measured in the alignment or in the model – as a function of the Frobenius norm threshold.

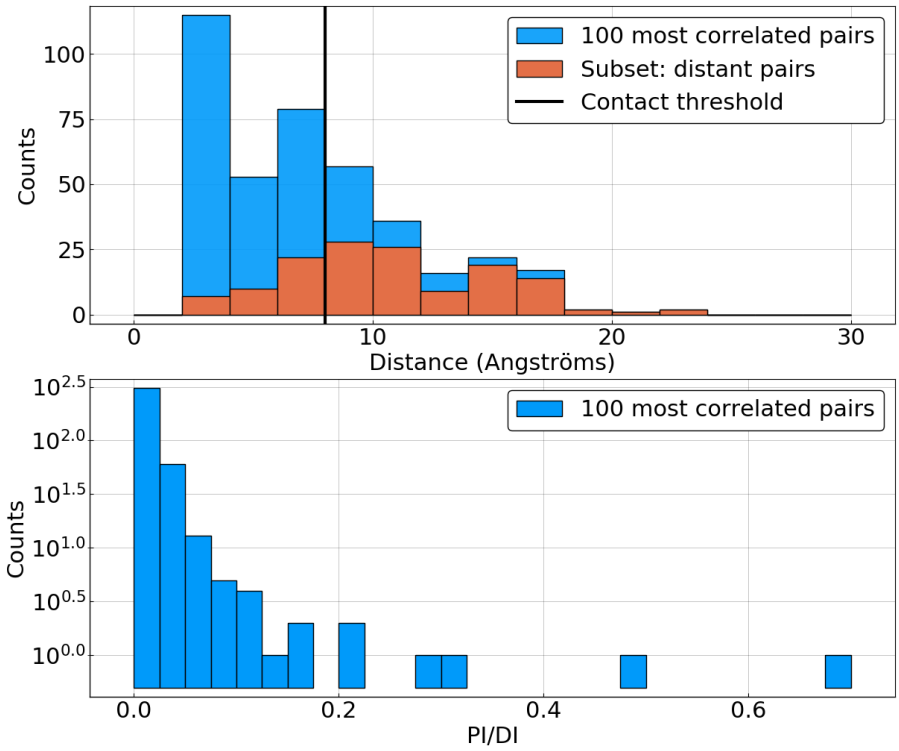

Figure 6.3: **Top.** Histogram of the largest distance found in each coupling chains. The chains considered are the 4 strongest (excluding the direct coupling itself) between each of the 100 most correlated residue pairs in PF00072. The subset of pairs corresponding to non-contacting residues (distant of more than 8$\AA$) is shown as a separate histogram.**Bottom.** Histogram of the strength of chains, the path information PI, scaled to the direct information DI, for each of the chains described above.

correlation seems to be a network effect.

To investigate whether the way correlations are mediated by the network can be structurally explained, we look at the strongest coupling chains for the 100 most correlated pairs of columns in the MSA of PF00072. Figure 6.3 shows the distribution of the largest distance found in each of the 4 strongest chains for each pair, excluding the chain of length 1 corresponding to the direct coupling. As an example, if a chain goes through residues i, j, k, l, the distance we consider here is $max(d_{ij}, d_{jk}, d_{kl})$. If this distance is larger than the threshold defining contacting residues, here 8$\AA$, then the corresponding chain is not structurally explainable. As can be seen in the figure, a large fraction of the maximum distances are larger than the threshold for contact. This gets worse for distant residues, for which almost none of the strongest chain can be structurally interpreted. The same figure also shows the distribution of the strengths of strongest chains (path information PI) scaled by the strength of the corresponding direct coupling for the considered pair (direct information DI). For most

chains, this ratio is smaller than 10%. Importantly, this histogram includes pairs of distant residues, where the direct coupling is likely not to be very strong as hinted by figure 6.1. Thus, even the strongest coupling chains between the most correlated pairs have quite weak effects compared to the direct coupling. Taken individually, coupling chains seem to be either dominated by noise or heavily influenced by the existence of the many non-structural weaker couplings shown in figure 6.1. As such, it remains hard to interpret them.

Since the single one-dimensional coupling chain does not provide much information, we investigate the effect of sub-networks on the correlation. For a pair of correlated columns in the MSA, there should exist a sub-network of sites, which we will call a *clique*, through which the DCA model explains this correlation. If such a clique can be found, one could try to see if it has relations to the structure of the protein.
The problem of finding cliques of sites mediating correlation between two residues is not trivial. As in the definitions of direct information or path information, we would like to assess the effect of a sub-network of sites on the correlation between two residues, keeping *constant* the conservation profile of all members of the sub-network. Yet, as figure 6.2 shows, setting couplings to zero – a corollary of removing sites from the full network – strongly impacts single-site statistics $f_i(a)$ at all sites. One way to achieve a constant conservation while removing sites would be to introduce compensatory fields $\tilde{h}$ to the Hamiltonian, tuned to compensate the vanishing couplings. This is the method used for direct information and path information measures. However, when a whole subnetwork is considered, tuning those compensatory fields amounts to training a new model for each sub-network to be evaluated, making things computationally intractable. Here, we designed an alternative methodology explained in details in the appendix a. Quickly summarized, a sub-alignment including only the sites of the desired sub-network is extracted from the full MSA. Intra-column shuffling moves are then proposed and accepted with a probability that depends only on the couplings between sites of the sub-network, in an MCMC-like fashion. This guarantees that the correct conservation profile is maintained for each site, while the correlations between columns corresponds to the couplings between clique members (see appendix a for more details).
To find cliques that mediate correlation between a given pair of columns, we decimate the model by iteratively removing individual or small groups of sites that contribute least to the mutual information of the considered pair. At each iteration, every column is tried for removal, and the ones which least modify the *MI* are effectively removed. Iterations are repeated until only the direct coupling is left, corresponding to the direct information. For efficiency reasons, residues are removed by groups at the beginning of the decimation,

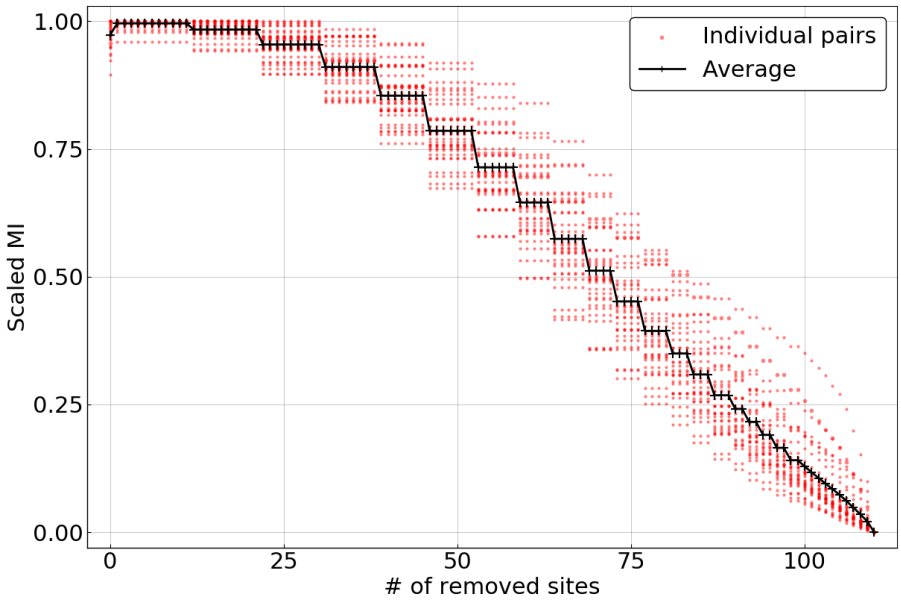

Figure 6.4: For 50 well correlated pairs of PF00072, mutual information *MI* remaining as a function of the number of nodes that are removed from the network. For each pair, *MI* is scaled so that it equals 1 when whole network is present (corresponding to the value measured in the MSA), and zero when all other nodes have been removed and only the direct coupling remains. The flat portions of the curves are explained by the fact that several nodes are removed at once at the beginning of the decimation.

but one by one at the end. Figure 6.4 shows this process of decimation for the 50 most correlated pairs in PF00072: the mutual information mediated by the clique is shown as a function of the number of removed residues. To allow for comparison between pairs, *MI* is scaled so that it is one when all the network is present, and zero when only the direct coupling remains (points higher than 1 in figure 6.4 are due to a statistical bias when estimating *MI*, leading to overestimations for finite size data).

Figure 6.4 indicates that it is possible to remove a significant number of sites without changing much to the correlation of a given pair. For some pairs, removing up to $\sim$ 40 residues (out of 112 in total) leaves the mutual information roughly to its original level. However, it is also visible that there is no sign of well-defined subnetworks or cliques. Once the first sites are removed, *MI* decreases steadily as more decimation iterations are performed, and it does not seem possible to define a cutoff value or a characteristic size for subnetworks. It is therefore unpractical to identify a clique responsible for the observed long-range correlations in MSA. This suggests once again that correlation is mediated by a large portion of the network of couplings, making interpretation of indirect effects very difficult.

6.1.3 *Distinct sets of* DCA *parameters with equally good fitting quality*

The BML method described in chapter 3 is based on iteratively modifying the parameters $\boldsymbol{J}$ and $\boldsymbol{h}$ in the direction of the gradient of the likelihood function. As the optimization problem is convex, this scheme should converge to a unique solution independently of how it is initialized. However, BML is only run for a finite time, and estimation of the gradient through MCMC is always approximative. As a result, exact convergence is never reached. Therefore, the parameters obtained through BML could depend on the way they were initialized.

Figure 6.5 shows parameters of two BML-inferred models with different initializations. For the first, initial parameters $\boldsymbol{J}^{plm}$ and $\boldsymbol{h}^{plm}$ are inferred using PLM approximation, and the Boltzmann machine is run from there. The resulting PLM-initialized parameters are called $\boldsymbol{J}^{BM/plm}$ and $\boldsymbol{h}^{BM/plm}$. For the second, the BML is initialized to zero, with resulting parameters called $\boldsymbol{J}^{BM/0}$ and $\boldsymbol{h}^{BM/0}$. Both methods are run for a large number of iterations, until the fitting quality has saturated to high values of about ~ 0.9.
The top panel of the figure shows a comparison of $\boldsymbol{J}^{BM/plm}$ (x-axis) with $\boldsymbol{J}^{BM/0}$ and $\boldsymbol{J}^{plm}$ (y-axis). While all these parameters are quite similar in between each other, it is visible that the PLM-initialized couplings are slightly more correlated to their initialization point $\boldsymbol{J}^{plm}$ than the other BML inferred model $\boldsymbol{J}^{BM/0}$. This is made quantitative by looking at Frobenius norms between couplings: we find that $||J^{BM/plm} - J^{plm}|| \simeq 14$, while $||J^{BM/plm} - J^{BM/0}|| \simeq 30$. This implies that even after a large number of iterations and the Boltzmann machines have accurately fitted the pairwise statistics of the data, couplings stay closer to their initialization point than to other couplings inferred from a different initialization.
What is even more striking is that this picture is completely modified when looking at statistical properties of models instead of parameters. The bottom panel of figure 6.5 shows pairwise frequencies $f_{ij}(a, b)$ of the two BML models with PLM and null initialization along with those of the PLM model. With this metric, it is evident that $J^{BM/plm}$ and $\boldsymbol{J}^{BM/0}$ are much closer to each other than to $\boldsymbol{J}^{plm}$. While both BML models have very close pairwise statistics, both fitting the data with quality > 0.9, the PLM inferred model completely fails at fitting statistics of the MSA. Again, this is made quantitative by looking at symmetrized KL-distance (abusively writing a Hamiltonian $\mathcal{H}$ in place of the probability it defines at temperature 1):

$$\begin{aligned} D_{KL}(\mathcal{H}^{BM/plm}||\mathcal{H}^{BM/0}) &= 8.3, \\ D_{KL}(\mathcal{H}^{BM/plm}||\mathcal{H}^{plm}) &= 16.8, \\ D_{KL}(\mathcal{H}^{plm}||\mathcal{H}^{BM/0}) &= 28.4. \end{aligned} \tag{6.1}$$

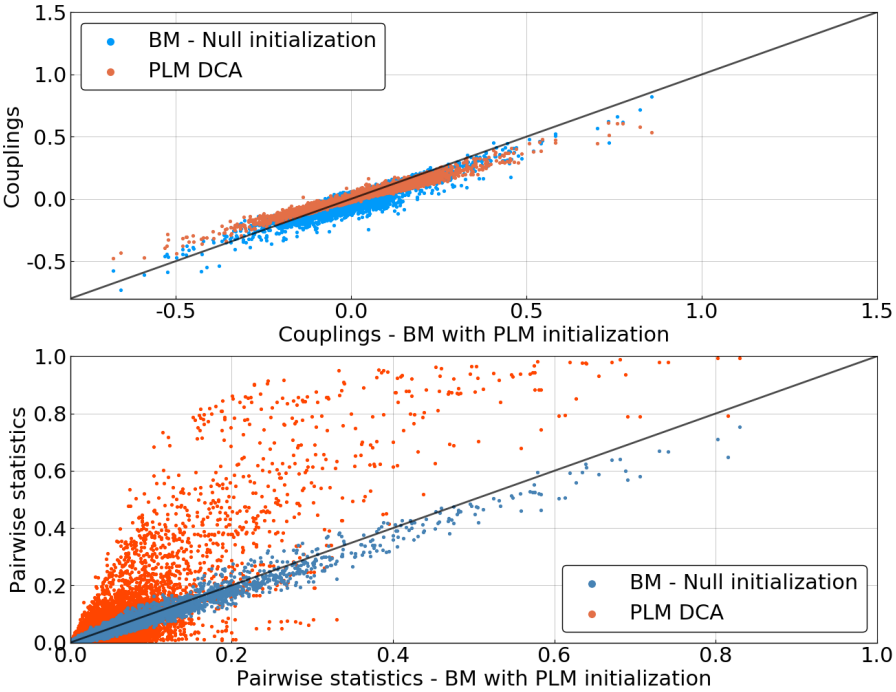

Figure 6.5: Results of the Boltzmann Machine Learning (BML) for two different initialization of the parameters: the PLM-DCA, and a zero (or null) initialization. **Top.** Direct comparison of the $J_{ij}(a,b)$ parameters, with the PLM-initialized BML $J^{BM/plm}$ on the x-axis, and the null-initialized $J^{BM/0}$ on the y-axis. The PLM-DCA couplings J^{plm} themselves are also shown on the y-axis. It is visible that the PLM-initialized model stayed very correlated with its initial parameters. In terms of distance, $||J^{BM/plm} - J^{plm}|| \simeq 14$ while $||J^{BM/plm} - J^{BM/0}|| \simeq 30$. **Bottom.** Comparison of pairwise statistics $f_{ij}(a,b)$ for the same models as above. The PLM-initialized BML is shown on the x-axis, and both the null-initialized BML and the PLM-DCA are on the y-axis. The two BML models are very close in terms of frequencies – and also a good fit to MSA data with a quality of ~ 0.92 –, while the PLM model fails to fit frequencies found in the MSA.

Thus, models $\mathcal{H}^{BM/plm}$ and $\mathcal{H}^{BM/0}$ are distant in parameter space, but closer in the space of probability distributions. Inversely, $\mathcal{H}^{BM/plm}$ stays quite close to its initialization point $\mathcal{H}^{plm}$ in parameter space, but is very far from it in terms of distribution.

The fact that two models with different couplings can be equally good at fitting is made even more striking when initializing the Boltzmann machine with couplings coming from a sparse DCA model. Such models are discussed in more details in section 6.2. Briefly, they include non-zero J_{ij} couplings only for a minority of pairs (i,j). When initializing the BML algorithm from such a well-inferred sparse model, zero couplings naturally relax to non-zero values (since only ℓ_2 regularization is used, as opposed to $l1$, there is no penalty for couplings being

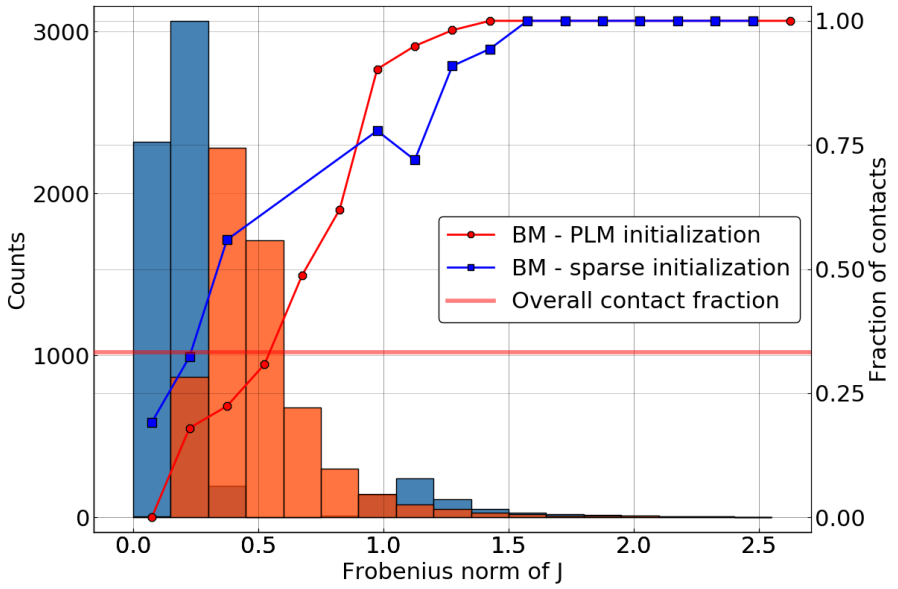

Figure 6.6: Similar to figure 6.1. **Left axis.** Distribution of the Frobenius norm of couplings $||J_{ij}||$ for two BML-inferred models, initialized with PLM-inferred parameters or with sparse parameters (see section 6.2). **Right axis.** Fraction of couplings which correspond to a structural contact for a given value of the Frobenius norm.

different from zero). However, even after many iterations, initially zero couplings remain small compared to initially non-zero couplings. This results in a bimodal distribution of couplings, represented in figure 6.6, and we refer to the resulting model as "pseudo-sparse" for this reason. The difference with the continuous distribution given by a PLM-initialized learning is striking. However, as shown in figure 6.7, the pairwise statistics of the two distributions are highly similar, both again fitting the MSA data with good accuracy. In terms of KL-distance, one measures

$$D_{KL}(\mathcal{H}^{BM/plm}||\mathcal{H}^{p.sparse}) = 3.3. \tag{6.2}$$

The couplings of the pseudo-sparse model have a much nicer interpretation in terms of structure, as seen in figure 6.2. The "strong" couplings, *i.e.* with Frobenius norm larger than ~ 0.9, correspond to structural contacts in 82% of cases, whereas the "weak" couplings, *i.e.* Frobenius norm smaller than ~ 0.5, seem to randomly correspond to contacting or distant residues. The bimodal nature of the distribution of couplings strength is satisfying in terms of interpretation: large parameters detached from the bulk of the distribution are biologically relevant, whereas the majority of small ones are not. However, as far as statistical properties of the DCA model are concerned, the two BML inferred distributions are equivalent.

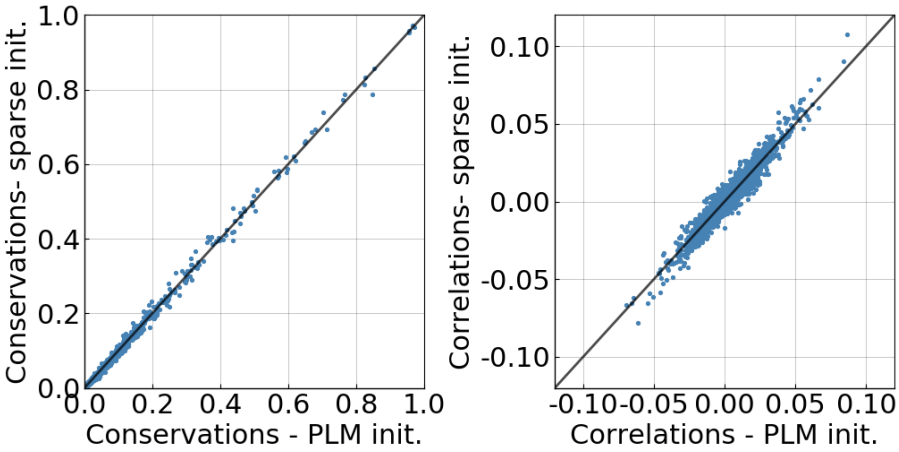

Figure 6.7: **Top.** Frobenius norm of couplings of the PLM-initialized (x-axis) and the sparse-initialized (y-axis) BML models. **Bottom.** Similarity of the single site statistics $f_i(a)$ (left panel) and the connected correlations $c_{ij}(a, b)$ between the PLM-initialized and the sparse-initialized BML models. Both models are a very good fit to MSA statistics, respectively 0.93 and 0.96 in fitting quality of connected correlations.

This poses major understanding problems of the parameters of DCA. On the one hand, the Maximum-Entropy Principle principle states that the "good" model is the one reproducing pairwise statistics found in the MSA. But we show here that down to some accuracy threshold in fitting pairwise statistics, many such models exist with highly varying parameters. This points to the existence of very flat directions in parameter space: large changes of $\boldsymbol{J}$ and $\boldsymbol{h}$ parameters lead to very similar probability distributions. Even if the BML algorithm is run for a large amount of iterations, little movement is made in these directions as the gradient of the likelihood is near to zero. As a result, the inferred models tend to stay close to their initialization point in parameter space, even though they accurately fit the data. In a sense, this means that the problem of finding correct values for all the direct couplings is to some extent ill-defined.

On the other hand, DCA is claimed to divide measured correlation into biologically relevant direct couplings, and indirect effects mediated through the network of direct couplings. The results of this section show that it is actually hard to make sense of how those indirect effects are constructed or to interpret them in relation with the considered protein. This is not so surprising, as these effects are built up from many small parameters which can vary regarding to how the inference is performed.

The combination of those two observations shows the strong limitations on the interpretation of the DCA parameters. While the article of chapter 3 invites one to think that pairwise co-evolutionary models might be the "right" models to describe variability in a Multiple

Sequence Alignment, the lack of understanding of most of the parameters remains frustrating. One way to make DCA models more interpretable would be to make couplings sparse. As stated at the beginning of this section, it is reasonable to assume that all pairs of residues should not interact directly: sparse couplings could represent a more realistic network of interaction. Moreover, reducing the number of parameters with a sparse model would likely solve the degeneracy problem due to the existence of flat directions in parameter space. Finally, since the the small direct couplings do not contain any structural signal, setting them to zero would make the model less prone to fine tuning of the parameters to a specific dataset, and thus overfitting.

6.2 SPARSE DCA MODELS?

6.2.1 *Decimating the couplings*

The most straightforward way to make parameters of the DCA model sparse would be to use $l1$ regularization, effectively imposing a cost for non-zero parameters. However, tests conducted with the BML method of 3 shown convergence problems of the algorithm when $l1$ regularization is used. Here, we try a different approach by slowly decimating the coupling parameters.
The idea is simple: couplings blocks $\boldsymbol{J}_{ij}$ are ranked using their Frobenius norm, and the weakest r are set to zero. r is typically a fraction amounting to 1% or 2% of the total number of coupling blocks $L(L-1)/2$. The model is then re-inferred on all the MSA data, but with some of its parameters now constrained to zero. This procedure is then iterated by setting each time the currently r weakest couplings to zero, and inferring again a new model. After n iterations, a model with only $1 - nr$ non-zero couplings is obtained. Since this method requires $n = 1/r$ inferences to explore all the possible sparsities, it is computationally unpractical to implement it using accurate schemes like BML. In practice, we use the PLM method to infer models, since it offers a good compromise between accuracy and speed and sparsity is easily enforced (as oppose to the Mean-Field approximation for instance).
Figure 6.8 shows the results of this decimation process, again for PF00072. The x-axis of all panels shows the fraction of J_{ij} blocks that are set to zero, with the decimation thus proceeding from left to right. The first panel shows the quality of contact prediction when $L = 112$ predictions are made, *without* using the APC correction (see section 2.4.3). The fraction of correctly predicted contacts slightly raises as couplings are decimated, until reaching the level of the non-decimated PLM using the APC. When too many couplings are removed, quality drops rapidly. A similar behavior is observed for the fitting quality

of the model. Surprisingly, correlations are reproduced with slightly higher accuracy as couplings are decimated, with fitting quality reaching a maximum when around 90% of the parameters are set to zero, and obviously dropping to zero when all are removed (due to the use of the PLM approximation, fitting quality remains lower than that of a BML learned model). The last panels show the pseudo-likelihood of the data according to the decimated model: unlike the likelihood, the pseudo-likelihood (see Eq. (2.20)) can be exactly and efficiently computed. Clearly, the pseudo-likelihood of the data can only be decreased when parameters are removed from the model, since it is now maximized over a subset of parameters. However, it is interesting to see that its decrease is quite slow. To quantify this, a naive null-model is shown where each removed coupling block J_{ij} reduces the pseudo-likelihood by the same amount. The difference between the actual values and this null model are shown in the bottom right panel, and exhibits a clear maximum. Interestingly, the position of this maximum roughly coincides with the points where the correlations are best fitted and where the contact prediction is of highest quality.

The slightly better accuracy in predicting contacts (without the *APC*) or the small improvement in fitting correlations with a PLM approximation are not impressive by themselves. What makes them remarkable is that they are obtained by removing $80 - 90\%$ of the parameters of the DCA model. The remaining parameters are able to better reproduce statistics found in the MSA while making more sense with respect to the structure of the protein: 82% of them correspond to structural contacts.
This contrasts with figure 6.2, where couplings were decimated *without* re-inferring the model. In this case, removing even the smallest couplings has a strong disruptive effect on the resulting probability distribution. This prompts the question of which couplings are set to zero in this new decimation scheme. Figure 6.9 attempts to answer this. On the top panel, Frobenius norms $||J_{ij}||$ of the initial PLM model and of the one decimated at $nr = 90\%$ are compared. Two observations can be made: strong couplings in the initial model remain strong after decimation, well correlated with their initial value even though stronger in norm. However, some couplings which are quite small in the fully connected model end up *not* being decimated, but on the contrary increase in magnitude. This is well visible in the "flat" part of the top panel of 6.9, with Frobenius norms which are relatively high in the decimated model but low in the original one. Furthermore, these "promoted" couplings seem completely uncorrelated to their initial values.
One can ask whether the promoted couplings are reproducible: if the decimation is implemented in a slightly different manner, will the same small couplings emerge from the bulk? To answer this, we

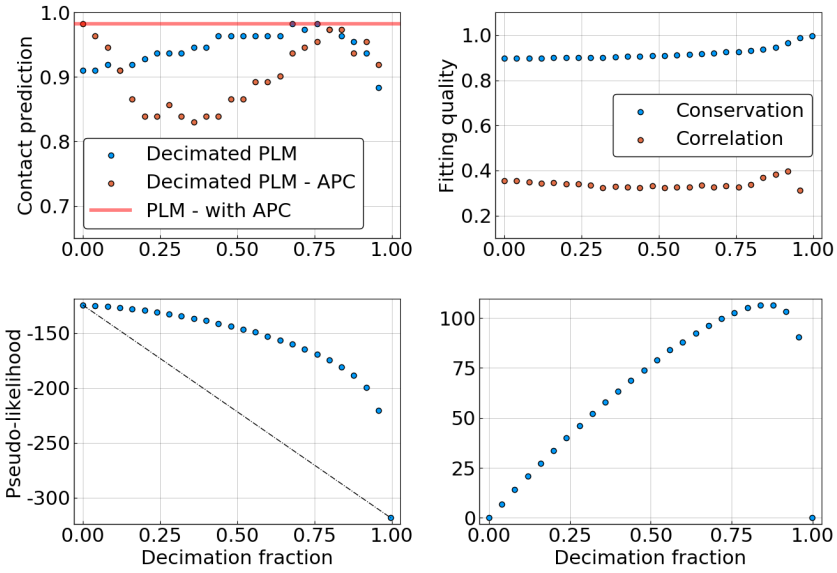

Figure 6.8: Decimating the PLM-DCA model. The x-axis of all plots is the fraction of coupling blocks J_{ij} that have been set to zero. **Top left.** Fraction of correctly predicted contacts when L predictions are made. Prediction is made with and without using the APC correction. The state of the art APC-corrected PLM is shown as a horizontal line. **Top right.** Fitting quality of conservations $f_i(a)$ and connected correlations $c_{ij}(a,b)$. **Bottom left.** Value of the pseudo-likelihood. Straight line shows a "null" model where every J_{ij} block set to zero reduces the pseudo-likelihood by the same amount. **Bottom right.** Vertical difference of the previous panel between the pseudo-likelihood and the null model. The maximum value of this scaled pseudo-likelihood is an estimation of the "best" decimation ratio.

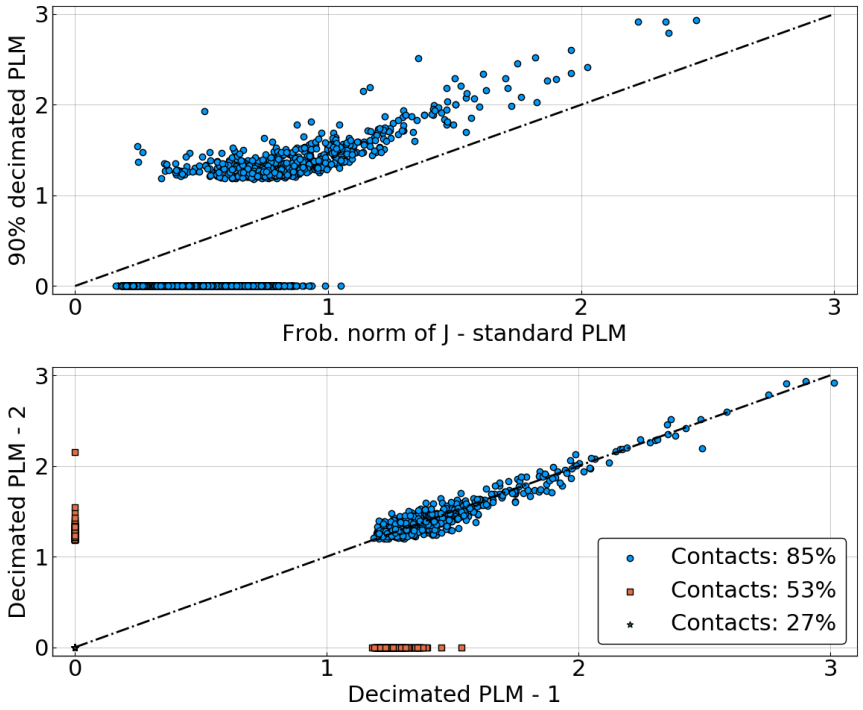

Figure 6.9: All axes represent Frobenius norms of coupling blocks J_{ij}. **Top panel.** 90% decimated PLM model against initial non-decimated one. **Bottom panel.** 90% decimated PLM models trained on two different MSAs. The two MSAs are non-intersecting sub-alignments of the PF00072 family.

perform the same decimation strategy on two non-intersecting sub-alignments of PF00072, each of $M = 5000$ sequences, forming datasets with slightly differing statistics. The Frobenius norms of couplings of the two resulting decimated model (again with a decimation ratio of $nr = 90\%$) are compared in the bottom panel of figure 6.9. If the strongest couplings are almost exactly the same between the two models (*e.g.* , Frobenius norm larger than 1.5), a subset of the weaker ones are not decimated in the same way. Points shown as orange squares in the figure are found to be zero in one model and non-zero in the other. Together, they form about 4% of all parameters, while the ones on which to the two model agree amount to 8%. As the legend of the figure shows, most of the couplings which are non-zero in the two models are contacts, a fraction which sensibly diminishes when considering couplings for which they conflict.
This shows that although the decimation procedure is robust for large coupling values, it is largely variable in the way it keeps initially small couplings.

6.2.2 *Highly accurate sparse models*

The previous section shows that it seems possible to reduce the number of parameters of the DCA model. However, this is achieved using the PLM approximation, thus not accurately fitting statistical features of MSAs. Is it possible to design a model precisely fitting the data with a reduced number of parameters? Achieving accuracy calls for using inference methods such as the BML algorithm. However, it is unpractical to decimate this model in the way that was used previously.
To overcome this, we decide to infer parameters with BML, using the decimated PLM model as a starting point. Two strategies are implemented. The first is simply to let the gradient descent of the BML take place on all the parameters of the initial sparse model, even the ones that were decimated. In this case, initially zero parameters relax to finite values, making the model non-sparse. However, the distribution of couplings remains bimodal, with a clear separation between initially non-zero couplings that remain strong, and initially zero couplings that remain weak. This distribution was shown in figure 6.6, and the corresponding model is referred to as "pseudo-sparse". The second is to constrain the initially decimated parameters to remain zero during the gradient descent. In this way, only the initially non-zero couplings are optimized by the BML.
The corresponding distribution of couplings is shown in figure 6.10, along with the local contact fraction for each Frobenius norm. Naturally, as in the PLM decimated model, couplings are divided into large ones, most of which correspond to contacts, and a majority of zero ones.
However, what is striking is the ability of this sparse model to reproduce statistics found in the MSA. The top panels of figure 6.11 show scatter plots of single site statistics and connected correlations corresponding to a sample of the sparse model and to the MSA. Although not as good of a fit as a fully connected BML trained model (shown in chapter 3 for instance), it is clearly visible that the capacity of this sparse model to reproduce correlations is much higher than , for example, that of a fully connected PLM model (see bottom panel of figure 6.5). Quantitatively, the fitting quality of this sparse model is 0.84, with 0.94 for the full BML model.

Another interesting feature of the sparse modeling lies in its robustness with respect to a global change of the parameters. The bottom panel of figure 6.11 shows the behavior of the heat capacity of the sparse model as function of temperature. The heat capacity C can be related to moments of the Hamiltonian $\mathcal{H}$ using the equation

$$\begin{aligned} C &= \frac{\partial \langle \mathcal{H} \rangle_{\beta\mathcal{H}}}{\partial T} \\ &= \frac{1}{T^2} \left(\langle \mathcal{H}^2 \rangle_{\beta\mathcal{H}} - \langle \mathcal{H} \rangle^2_{\beta\mathcal{H}} \right). \end{aligned} \tag{6.3}$$

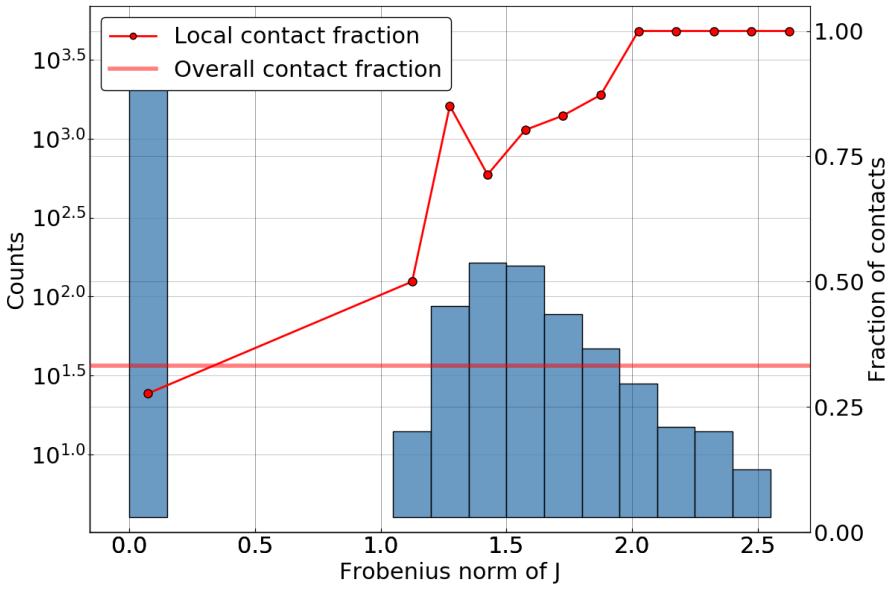

Figure 6.10: Similar to figure 6.1. **Left axis.** Distribution of the Frobenius norm of couplings $||J_{ij}||$ for the accurate BML inferred sparse model. The leftmost part of the histogram corresponds to exactly zero couplings. **Right axis.** Fraction of couplings which correspond to a structural contact for a given value of the Frobenius norm.

Variance and mean of $\mathcal{H}$ can be easily estimated through MCMC sampling, making the computation of C possible. The behavior of the heat capacity as a function of temperature T is quite different for the sparse and the fully-connected models. In the latter, it goes through a large maximum at $T \simeq 1$, indicating that a slight change in temperature leads to large variations of the average energy $\langle \mathcal{H} \rangle$. This can be interpreted as a sort of phase transition of the model, with two different behaviors at $T > 1$ and $T < 1$.
Since a change in temperature can also be written as a global change in effective parameters by the transformations $\boldsymbol{J} \to \boldsymbol{J}/T$ and $\boldsymbol{h} \to \boldsymbol{h}/T$, the large variation of C at the $T \simeq 1$ is a sign of non-robustness or overfitting of the DCA model. This almost completely disappears in the case of the sparse model, where C varies relatively little, showing only a very broad and shallow maximum. This observation is consistent with the idea that a fully connected DCA infers too many parameters, fine tuning the distribution to the available data and lacking robustness.

6.3 GOING BEYOND SPARSE MODELS?

It is striking that very accurate models can be inferred using only a fraction of the coupling parameters. This makes the name Direct Coupling Analysis much more meaningful: statistical patterns found

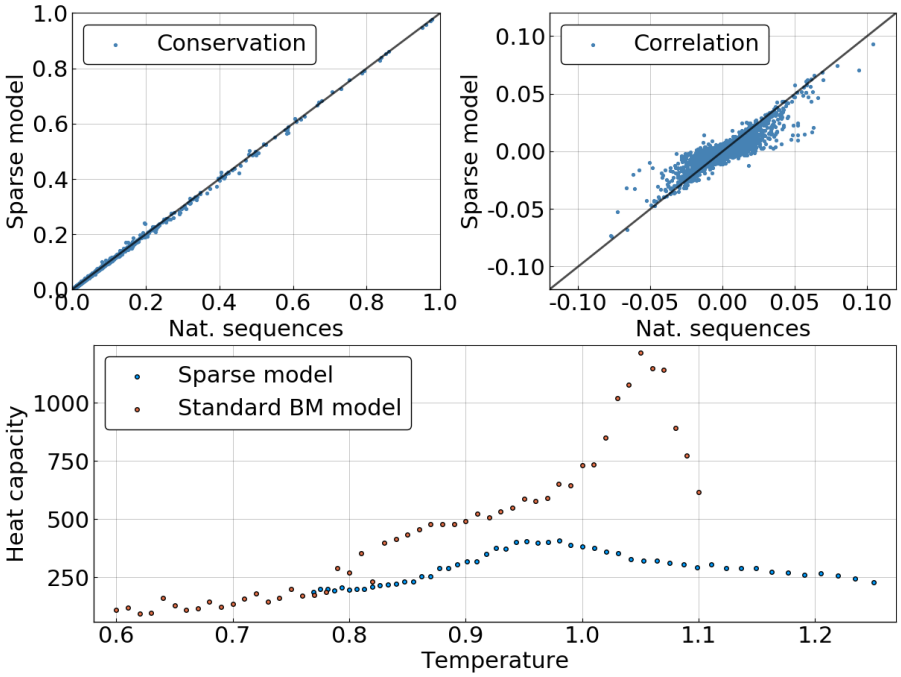

Figure 6.11: **Top panels.** Single site frequencies $f_i(a)$ and connected correlations $c_{ij}(a,b)$ found in the MSA of PF00072 (x-axis), and in a sample from the sparse model (y-axis).**Bottom panel.** Heat capacity $\partial\langle\mathcal{H}\rangle_{\beta\mathcal{H}}/\partial T$ as a function of temperature T, both for the sparse model and the fully connected BML model.

in the MSA are explained by *direct* couplings, and the majority of those correspond to physically interpretable features of the protein: structural contacts. The fact that more than 80% of the couplings that the sparse model of section 6.2.2 uses correspond to structural contacts is highly satisfying in terms of modeling.

However, figure 6.11 clearly shows that not all connected correlations can be fitted using such a reduced number of parameters. Even if the Boltzmann machine is run for a long time, it does not seem possible to achieve a better fitting quality than what figure 6.11 shows. One explanation for this could be that the sparsity derived from decimating the model using the PLM approximation is not optimal, and that with a better choice of non-zero parameters, data could be fitted more accurately.

Here, reasons for the impossibility of sparse coupling matrices of reproducing all statistical variability of the data are discussed. As stated many times, the idea underlying DCA is that correlation patterns found in the MSA can all be described by a sparse network of direct interactions between residues. However, there are hints that some of the measured correlation does not fit this picture. In [18], authors attempt to derive a Hopfield representation of the Potts model used in

DCA. Briefly, this amounts to writing the coupling matrix $\boldsymbol{J}$ as a sum of orthogonal *patterns* $\xi_i^\mu(a)$ to which sequences are "attracted":

$$J_{ij}(a,b) = \sum_{\mu=1}^{K} \xi_i^\mu(a)\xi_j^\mu(b). \tag{6.4}$$

In this representation, the patterns are the eigenmodes of the coupling matrix, its rank being the number of patterns K. In the Mean-Field (MF) approximation, patterns also correspond to the eigenmodes of the Pearson correlation matrix of the alignment. A pattern corresponding to an eigenvalue λ^μ for the Pearson correlation matrix will correspond to an eigenvalue $1/\lambda^\mu$ for the couplings. This is similar to what would be expected from the MF equation $\boldsymbol{J} \sim \boldsymbol{C}^{-1}$ (see Eq. (2.16)). In [18], it was found that both high *and* low eigenvalues of the correlation matrix (respectively $\lambda^\mu \gg 1$ or $\ll 1$)contribute strongly to the likelihood function. Eigenmodes $\xi_i^\mu(a)$ corresponding to low eigenvalues tend to be localized on a few positions i (low inverse participation ratio). In the coupling matrix, they thus naturally form strong localized entries. However, eigenmodes corresponding to *high* eigenvalues are found to be typically spread across all positions. Thus, in the coupling matrix, they result in small but spread out entries.
Interestingly, in [85], it was shown that decomposing the correlation matrix of an MSA in a sparse part and a low-rank part improves the quality of contact prediction. The correlation matrix is written as

$$\boldsymbol{C} = \boldsymbol{S} + \boldsymbol{L}, \tag{6.5}$$

where $\boldsymbol{S}$ is sparse in the sense that most elements $S_{ij}(a,b)$ are zero, and $\boldsymbol{L}$ has most of its eigenvalues equal to zero. The resulting sparse matrix can then be used as a quite efficient contact predictor, almost as good as some DCA implementations.
Lastly, many results have been obtained by the so-called Statistical Coupling Analysis (SCA) method concerning sectors [40, 53]. Sectors are large groups of coherently co-evolving residues, that can usually be linked to structure and have been found to be sensitive to mutations. Mathematically, sectors are based on an independent component analysis of the MSA, roughly corresponding to large eigenmodes of the re-weighted correlation matrix.
This again points to the fact that signal in $\boldsymbol{C}$ may not come only from a sparse network of structurally interpretable couplings, but also contains "spread-out" modes. Removing those drastically improves contact prediction, hinting that structure may not be determinant for their existence. This would explain results in section 6.1.1: energies and statistical properties of the DCA model are dominated by small couplings which do not correspond to any structurally interpretable quantity.

Where could these spread-out eigenmodes of the correlation matrix come from? Several explanations can be given. In [66], it is shown that phylogenetic biases are the source of large eigenvalues in the correlation matrix. If corresponding eigenvectors are removed, contact prediction is improved, at least on simulated data. If this is the case in actual MSA of protein families, being able to accurately correct phylogenetic biases becomes important, as this would disentangle structural or functional sources of correlation from those induced by evolutionary processes themselves.
Another explanation is the potential presence of higher order interactions being responsible for correlation not well explainable by current DCA. It was for instance found in [69] that the introduction of well-chosen three-body interactions can improve the contact prediction ability of an MF inferred DCA model. Another way to introduce higher order interactions is to suppose that the fitness of a sequence is a non-linear function of a simple energy function. As an example, imagine proteins in a folded state have an energy defined by a simple function $E(\underline{A}) = \sum_i h_i(a_i)$, and a zero energy in the unfolded state. Fitness of sequence $\underline{A}$, defined as its probability to be in the folded state, might then be written as

$$f(\underline{A}) = \frac{e^{-E(\underline{A})}}{1 + e^{-E(\underline{A})}}. \tag{6.6}$$

This contrasts strongly with the way DCA is currently used to predict fitness (see section 2.5.2), which assumes the following functional form:

$$f(\underline{A}) = -\mathcal{H}(\underline{A}) = \sum_{1 \leq i < j \leq N} J_{ij}(a_i, a_j) + \sum_{i=1}^{N} h_i(a_i). \tag{6.7}$$

Whereas equation (6.7) is limited to pairwise effects, equation 6.6 has interactions at all orders. At best, the former can be considered as a low order expansion of the latter. It is important to mention that in recent work [60], fitness effects of all double mutants of a protein were accurately fitted with a model similar to Eq. (6.6), using three physical states for the protein instead of two.
One last potential cause for interactions unexplainable by pairwise couplings is the presence of latent variables. It is argued in [73] that if a hidden variable interacts with a system, the variables of this system can appear as coupled at all orders. This phenomena can even lead to critical points in a model inferred without the knowledge of the latent variable. If this is the case for protein sequences, then DCA should be adapted to model for potential hidden sources of interaction, thus disentangling couplings that are proper to the protein from others.

How could current DCA approaches be augmented to take these observations into account? The ideas presented in [69] are one way

to introduce higher order interactions in a feasible way: only those which seem strongly backed by data are used. However, it is hard to deal with orders higher than 3-bodies with this method because of the finite size of available data.
Another way, restricting the model to pairwise interactions, would be to combine the sparse representation of couplings discussed before, and the Hopfield one of [18]:

$$J = J_S + J_L, \tag{6.8}$$

where, for instance, J_S would be similar to the couplings found in section 6.2.2, and J_L would be of the form of equation (6.4). In this way, structurally interpretable parameters are separated from those which are necessary to explain other phenomena.
This idea does not allow for higher order terms in the Hamiltonian. However, it can be extended with the formalism of Restricted Boltzmann Machines (RBMs) [79]. An RBM involves two types of nodes, namely the observable ones – corresponding to observed residues in the DCA setting – and the hidden ones. Interactions only take place between two nodes of different types. This formalization naturally introduces hidden variables while keeping the Potts model of DCA as a particular case, in which the potential acting on hidden variables is Gaussian. However, the choice of other potentials for hidden variables leads to interaction at all orders between residues.

CONCLUDING REMARKS

DCA was proposed in 2009 as a tool to help in protein structure prediction. Taking advantage of the large number of available protein sequences, it was aimed at improving on simple correlation based contact predictors. Further work has since demonstrated that pairwise models can do much more than predicting contacts (section 2.5).

In this thesis, we have tried the limits of such models while improving them in different directions. In chapter 3, we have shown that DCA can serve as a good model for sequence variability inside protein families: statistical features of these sequences are reproduced by the Potts distribution, even though they were not fitted, while the way direct couplings mediate correlation between distant residues has been shown to be interpretable to some extent.
Chapter 4 served as a proof of concept, showing that it is possible to integrate different types of information in the DCA framework, combining global statistical features at the level of the family with local measurement of single sequences in a natural way. Even though this method could not be applied by lack of data, the rapid development of experimental techniques and the increasing number of studies about protein mutational landscapes or protein design could make these ideas relevant in the near future.
In chapter 5, we tried to design corrections for known biases of data DCA is used on. Members of protein families are by definition related by evolution, and thus cannot be considered as completely independent samples of the same distribution. However, current co-evolutionary models account for this in empirical and non-principled ways. Correcting data for phylogenetic effects may lead to important improvements in the ability of DCA to disentangle sources of different statistical signal found in MSAs, namely functional constraints on the sequence and statistical biases due to phylogeny.
Lastly, chapter 6 dealt with limitations of co-evolutionary models. The main concern here is the vast number of parameters inferred by DCA: while potentially millions of couplings are inferred to model a protein family, only a handful can be interpreted and used to predict structural contacts, making current models the opposite of parsimonious. It was shown here that the development of more sparse modeling methods may lead to improvements in interpretability.

To conclude, two lines of future research are proposed. The first follows ideas presented in the last section 6.3. If accurate models of protein families are to be constructed, it is only natural to make them

as parsimonious as possible. As the goal of such methods is ultimately to understand the functional constraints acting on proteins, it is necessary for models to make use of parameters which have biological meaning and can be interpreted.
The second one concerns the exciting field of protein mutational landscapes and protein design. As stated in 3, pairwise models might be used as *generative* models to create new functional protein sequences. At the same time, chapter 4 shows that their capacity to model the fitness landscape proteins evolve in might be increased by incorporating new experimental information. Continuous experimental progress in quantitative characterization of single protein sequences may provide the necessary information to go beyond the use of homologous sequences alone. This provides an exciting opportunity to improve the quality and predictive power of DCA-like methods.

APPENDIX

a

QUANTIFYING INDIRECT EFFECTS: CHAINS AND CLIQUES

A.1 CHAINS OF COUPLINGS

In the first DCA implementations [57, 82], the interaction score used to predict contact was the Direct Information (DI). Once the direct couplings $\boldsymbol{J}$ are inferred, DI is defined for each pair of positions (i, j) as

$$\begin{aligned} DI_{ij} &= \mathcal{I}\left(P_{ij}^{dir}(a,b)\right) \\ &= \sum_{a,b=1}^{q} P_{ij}^{dir}(a,b) \log\left(\frac{P_{ij}^{dir}(a,b)}{f_i(a) f_j(b)}\right), \end{aligned} \tag{a.1}$$

where P_{ij}^{dir} is the so-called *direct probability*, the frequencies f_i and f_j are those measures in the MSA, and the MI is written $\mathcal{I}$. The direct probability is defined as

$$P_{ij}^{dir}(a,b) = \exp\left\{J_{ij}(a,b) + \tilde{h}_i(a) + \tilde{h}_j(b)\right\}. \tag{a.2}$$

The compensatory fields $\tilde{h}_i$ and $\tilde{h}_j$ are computed for each pair (i, j), and ensure that P_{ij}^{dir} has the correct marginals:

$$\sum_{a=1}^{q} P_{ij}^{dir}(a,b) = f_j(b) \quad \text{and} \quad \sum_{b=1}^{q} P_{ij}^{dir}(a,b) = f_i(a). \tag{a.3}$$

The direct probability represents the distribution of amino-acids one would find at columns i and j of the MSA if those columns were only coupled by the direct coupling J_{ij} while keeping the same conservation profile. As such, DI provides a measure of the strength of a direct coupling in a principled way.
In order to evaluate the strength of a *chain* of couplings going through sites $[i_1 \dots i_N]$, we simply extend the definition of the direct probability to include more couplings, defining the *path probability*:

$$P_{i_1 i_N}^{path}(a_{i_1}, a_{i_L} | [i_1 \dots i_N]) = \sum_{a_{i_2} \dots a_{i_{N-1}}=1}^{q} \prod_{l=1}^{N-1} P_{i_l i_{l+1}}^{dir}(a_{i_{l+1}} | a_{i_l}) \cdot f_{i_1}(a_{i_1}), \tag{a.4}$$

with $P_{ij}^{dir}(a_i | b_j) = P_{ij}^{dir}(a_i, b_j) / f_j(b_j)$. In other words, for a chain $[i_1 \dots i_N]$, direct probabilities of every link $(i_l i_{l+1})$ of the chain are multiplied. Since only the distribution of the extremities of the chain is interesting, the resulting expression is summed over all configurations of the intermediary positions. This summation is made possible by the

unidimensionality of the chain: multiplying two direct probabilities P_{ij}^{dir} and P_{jk}^{dir} and summing over configurations of the intermediary variable j amounts to a matrix product. Thus, equation a.4 can be seen as a product of transfer matrices.
Importantly, path probabilities have the correct marginals by construction:

$$\sum_{a=1}^{q} P_{ij}^{path}(a,b|[i\dots j]) = f_j(b) \quad \text{and} \quad \sum_{b=1}^{q} P_{ij}^{path}(a,b|[i\dots j]) = f_i(a), \tag{a.5}$$

for an arbitrary path $[i\dots j]$ linking i and j. This can be verified directly by combining equations (a.4) and (a.3). The path probability, similarly to its direct eponym, quantifies the distributions one would find at positions i_1 and i_N if they were joined only by a one-dimensional chain of coupling going through sites $[i_1, i_2 \dots i_N]$. To quantify the strength of this chain by a scalar measure, we define the Path Information (PI) similarly to Eq. (a.1):

$$PI_{ij}([i\dots j]) = \mathcal{I}\left(P_{ij}^{path}(a,b|[i\dots j])\right). \tag{a.6}$$

A.2 FINDING STRONGEST COUPLING CHAINS: DIJKSTRA'S ALGORITHM AND EXTENSIONS

One can see a Potts model as a graph, with the sites $a_1, \dots, a_L$ being the vertices and the presence of a coupling J_{ij} indicating that there is an edge between vertices i and j. In the case of a Potts model inferred using DCA on a protein's MSA, this graph is *a priori* complete, since every coupling J_{ij} can in principle be non zero. The strength of a coupling path joining two vertices can be defined using path information PI. When looking for the strongest chains of couplings, it is natural to consider Dijkstra's algorithm.
On a graph with positive distances d_{ij} associated to each edge, Dijkstra's algorithm allows one to find the shortest path through edges from vertex i to vertex j [23] . A useful extension to this algorithm is Yen's algorithm [84], which solves the problem of finding the K shortest paths between two arbitrary vertices in a similar graph.
Both these algorithms assume that lenghts of the edges are positive and additive, in the sense that the distance of path $[i\ k\ j]$ is equal to $d_{ik} + d_{kj}$. If one wants to find the strongest chains of couplings in the graph defined by the Potts model, a straightforward procedure is to modify Dijkstra's algorithm to find the path with the strongest information PI, which is similar to using PI as the inverse of a distance. Path information has one of the properties that is essential for Dijkstra-like procedure to function, which is that adding an edge to an existing path reduces the strength of this path, *i.e.* $\mathrm{PI}[i\ k] > \mathrm{PI}[i\ k\ j]$. In terms of distance, this is equivalent to saying that adding an edge

to an existing path increases the length of this path, which is the case for positive lengths on edges. However, path information is clearly not additive: in order to add an edge to an existing path, one has to multiply the probability matrix P^{path} corresponding to the existing path by the direct probability matrix P^{dir} corresponding to the new edge. The definition of path information can actually lead to situations sketched in figure a.1, where Dijkstra's algorithm will obviously fail. Nevertheless, it is still possible to naïvely apply Dijkstra's and Yen's

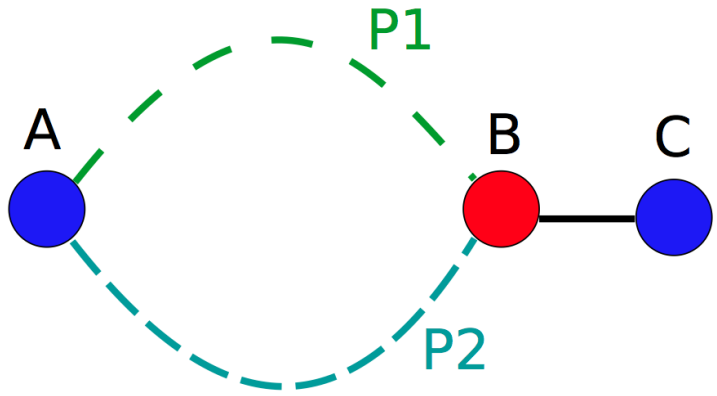

Figure a.1: Using path information as the inverse of a distance can result in non-intuitive situations. In this sketch, imagine a situation where $PI(A, B|P_1) > PI(A, B|P_2)$, that is the strongest path linking A and B is P_1. However, it is possible that $PI(A, C|[P_1 B]) < PI(A, C|[P_2 B])$. In other words, even though the strongest chain from A to B is P_1, the strongest chain from A to C consists in taking path P_2 and going through B. In this kind of scenario, Dijkstra's algorithm will fail to find the correct strongest chain.

algorithm, even if they are not guaranteed to find the strongest paths in practice. As an evaluation of the performance of Yen's algorithm in such a case, it was applied to find the strongest 50 paths for each of the 100 most correlated pairs of sites in the model inferred on PF00004. For each pair, the resulting paths were then sorted by decreasing strength, and the list of sorted paths is compared to the direct output of Yen's algorithm. In an ideal scenario, both should coincide. In the case of using path information as the inverse of a distance, it can however be expected that Yen's algorithm will not output paths in the correct order, and the two lists will thus differ.
In the 100 pairs of arrays constructed in this way, 90 coincide, meaning that Yen's algorithm did not make any obvious mistake in those cases. For the 10 arrays that did not coincide, Spearman's rank correlation between their entries was always higher than 0.98, indicating that they are in practice almost perfectely sorted. Furthermore, differences between the two arrays usually happened after the 10th position, suggesting that top strongest chains were recovered in the correct order. Although this analysis does not prove that the chains found using Dijkstra's algorithm and Yen's algorithm are actually the strongest

chains, it strongly suggests that situations such as that of figure a.1 are of minor importance. To make them even less probable, for the analysis in the main text we have actually extracted the 25 best-scoring paths using our modified algorithm, and then selected the 15 of highest PI for further analysis.

A.3 CLIQUES

As shown in sections 3 and 6.1, individual coupling chains are rarely informative about how the DCA network disentangles direct and indirect correlations. For this reason, it is interesting to attempt to measure how an entire subnetwork of sites correlated a given pair of positions i and j. Such subnetworks will be called *cliques* in the following.
A clique of size $N < L$ (where L is the length of the aligned sequences) can be defined by the nodes composing it, $\underline{A}_C = (a_{i_1}, \dots, a_{i_N})$. Defining a probability distribution for these nodes based on the inferred couplings is not trivial. Such a distribution $P(a_{i_1}, \dots, a_{i_N})$ would have to satisfy marginalisation relations similar to those in equation a.3 for each of its variables. Indeed, what we want to measure is the way direct couplings mediate correlation with a fixed conservation profile, not the way the model reproduces this conservation. If the ideas defining direct probability were to be used here, it would needed to compute compensatory fields $\tilde{h}_{i_l}$ for each of the N variables in the clique such that the correct marginals are exactly recovered:

$$P(a_{i_1}, \dots, a_{i_N}) \propto \exp\left\{ \sum_{1 \leq k < l \leq N} J_{i_k, i_l}(a_{i_k}, a_{i_l}) + \sum_{l=1}^{N} \tilde{h}_{i_l}(a_{i_l}) \right\}. \quad \text{(a.7)}$$

This leads to re-inferring the compensatory fields using a BML algorithm for *each* clique that has to be evaluated, which is computationally untractable.

To overcome this problem, a way to sample from $P(a_{i_1}, \dots, a_{i_N})$ without having to compute fields $\tilde{h}$ was derived. We start with a large sample of an inferred DCA model – or possibly with the studied MSA – noted $\{\underline{A}_i^m\}$, $i = 1 \dots L$, $m = 1 \dots M$. If the model is accurately inferred, $\{\underline{A}_i^m\}$ will have the same conservation profile as the original sequence alignment.
If a clique $(a_{i_1}, \dots, a_{i_N})$ is considered, only columns belonging to this clique are kept, resulting in a reduced sample $\{\underline{A}_i^m\}_C$, $i \in \{i_1 \dots i_N\}$. *Swapping* moves are then attempted on this sample: two lines m and n are chosen at random, along with one column i_l. The move consists in swapping variables $a_{i_l}^m$ and $a_{i_l}^n$. It is accepted with a probability $min(1, e^{-\Delta E_s})$, with

$$\begin{aligned} \Delta E_s = &\mathcal{H}_c(a_{i_1}^m \dots a_{i_l}^n \dots a_{i_N}^m) + \mathcal{H}_c(a_{i_1}^n \dots a_{i_l}^m \dots a_{i_N}^n) \\ &- \mathcal{H}_c(a_{i_1}^m \dots a_{i_l}^m \dots a_{i_N}^m) - \mathcal{H}_c(a_{i_1}^n \dots a_{i_l}^n \dots a_{i_N}^n), \end{aligned} \quad \text{(a.8)}$$

where $\mathcal{H}_c$ is the Hamiltonian of the clique, including *only* couplings:

$$\mathcal{H}_c(a_{i_1}, \ldots, a_{i_N}) = - \sum_{1 \leq k < l \leq N} J_{i_k, i_l}(a_{i_k}, a_{i_l}). \tag{a.9}$$

Since these move only allow intra-column swapping, they conserve the single site frequencies of the sample at all times. In this way, when enough moves are made and equilibrium is reached, the reduced sample $\{\underline{A}_i^m\}_C$ will include sequences distributed according to a Hamiltonian involving all couplings in the clique but with single site frequencies matching exactly those of the original inferred model. Relevant measures can then be computed from this sample, such as the correlation between two of its columns.

In section 6.1.2, a decimation strategy to find relevant cliques connecting two fixed sites i and j is mentioned. Here, we describe this procedure in more detail. One starts with a sample from the full DCA model, $\{\underline{A}_k\}$. For each $k \neq i, j$, position k is removed from the model, resulting in a clique of $L - 1$ nodes. A sample from this clique is obtained using the strategy described above, and the MI $\mathcal{I}_{ij}([k])$ between sites i and j is computed with this sample. The node k^* which maximizes $\mathcal{I}_{ij}([k])$ is then removed permanently from the network. This operation is then repeated, removing one node every time, until only the clique of size 2 that nodes i and j form remains. At this point, the remaining Mutual Information (MI) is by construction the DI.
This method allows to iteratively decimate the network, in a way that removes nodes contributing the least to the MI between i and j at every step. Figure 6.4 shows the remaining MI at each step of the process. Importantly, the scheme described here is a greedy one, making a locally optimal choice at each iteration. It does not guarantee that the cliques it finds at each step are the optimal ones in terms of correlating nodes i and j.
It is important to note that this method is computationally quite expensive. At each step, all remaining nodes of the network have to be evaluated, meaning that the swapping procedure described above has to be conducted about $L(L-1)/2$ times, and this only to conduct the decimation with respect to *one* pair of sites (i, j). In an attempt to speed up the process, several nodes $k_1^*, \ldots, k_K^*$ can be removed at each step, at least at the beginning of the decimation. The hope is that nodes which are removed first matter little in the correlation between i and j, and removing many at the same time does not reduce the accuracy of the algorithm. This explains the "staircase" look of figure 6.4.

b

DIRECT COUPLING ANALYSIS FOR PHYLOGENETICALLY CORRELATED DATA: SUPPLEMENTARY FIGURES

In chapter 5 of the main text, it is mentioned that artificial data was generated for two different times of the branches of the tree, $\tau = 0.3$ and $\tau = 0.5$. Only the figures for the first case were shown in the main text. Here, the same plots for the second case are shown.
$\tau = 0.5$ is an *easier* case for the phylogenetic correction, in the sense that the bias is not as strong as in the $\tau = 0.30$ case. For this reason, many of the improvement are not as strongly visible as those shown in the main text. This can be seen for instance on figure 5.6 (main text) or b.1, which represents the histograms of KL-distance for corrected and uncorrected models. Another difference is visible between figures 5.11 and b.6: in the $\tau = 0.5$ case, the DCA model inferred on uncorrected statistics does not have as strong a bias towards giving higher energies to sequences that are far away from the sample it was trained on.
The last figure b.7 shows the accuracy of our inference of parameter μ using the method described in section 5.1.3.

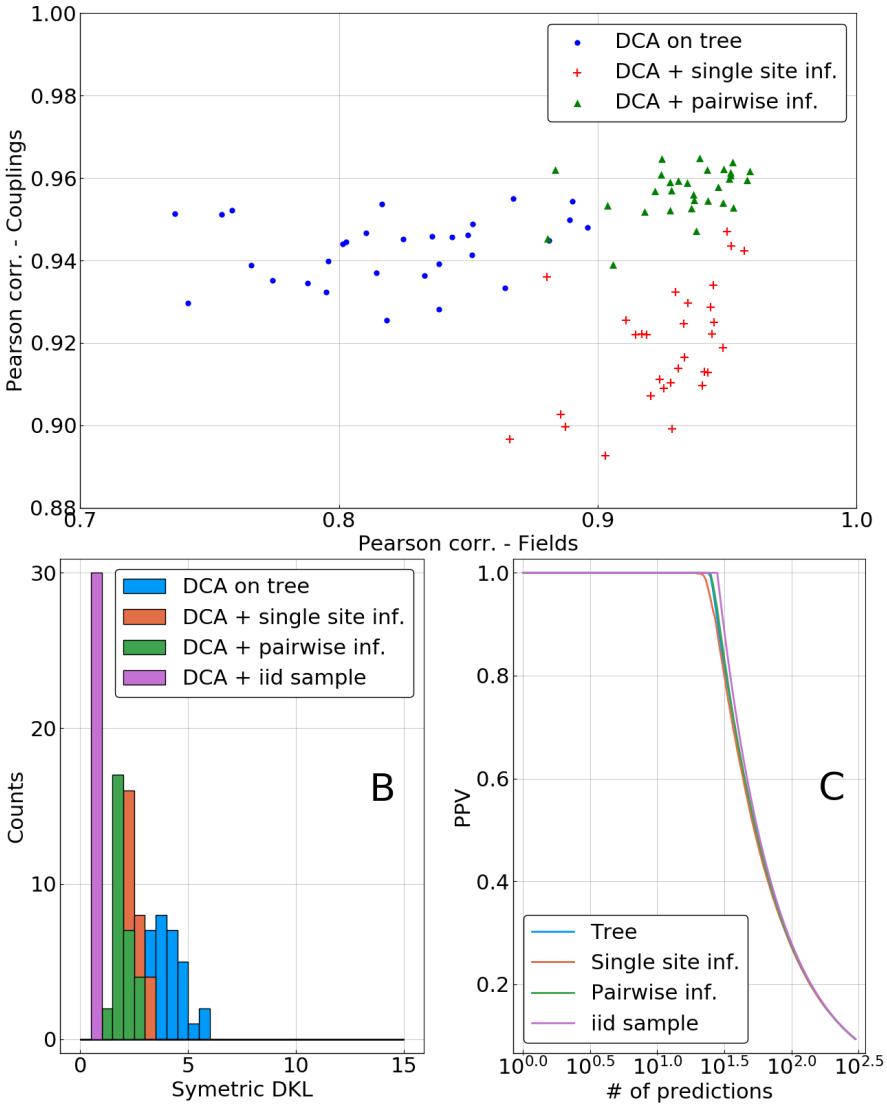

Figure b.1: DCA model inferred after single site or pairwise phylogenetic correction **A.** Pearson correlation between parameters of inferred and of true DCA models. y-axis: couplings J_{ij} ; x-axis: fields h_i. One point corresponds to one repetition of the MCMC process on the tree. **B.** Histogram of the symmetric Kullback-Leibler distances between inferred and true models for all repetition. **C.** Positive predictive value for predicting non zero couplings (*i.e.* "contacts") using inferred DCA models. DCA inferred on the *i.i.d.* sample performs perfectly in this case.

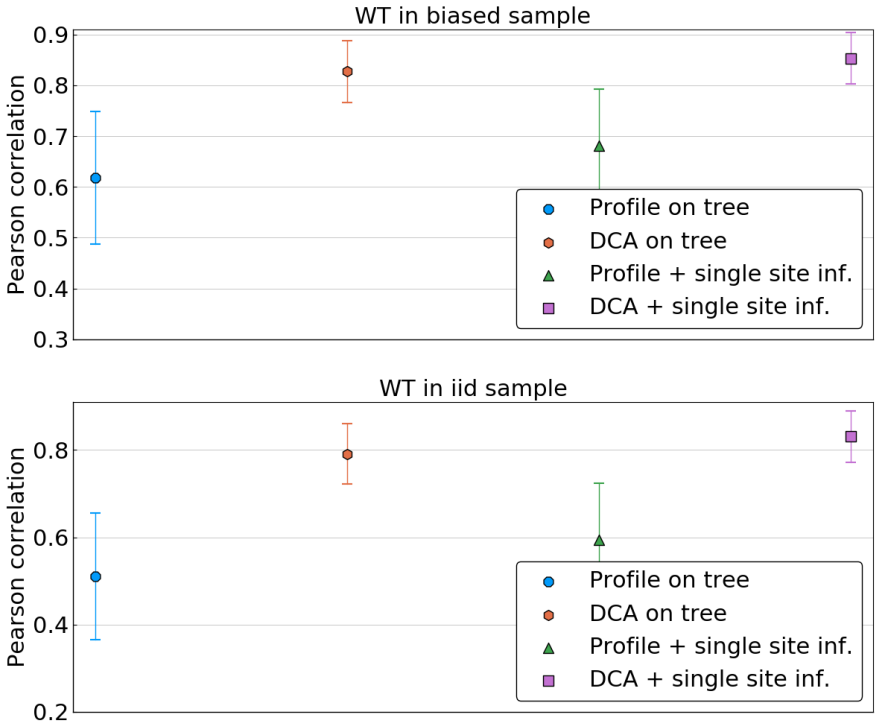

Figure b.2: Pearson correlation in predicting energies of single mutants averaged over sets of reference sequence. In the top panel, reference sequences are taken in the biased sample, *i.e.* among the leaves of the phylogenetic tree. In the bottom panel, reference sequences are taken in a fair sample of P^0. Predictions are made using four models: respectively a profile model and a Potts model trained on the uncorrected biased sample (resp. "Profile on tree" and "DCA on tree"), and using the corrected single site frequencies (reps. "Profile + single site inf." and "DCA + single site inf."). Error bars indicate the standard deviation across the 30 repetitions of the tree sampling process.

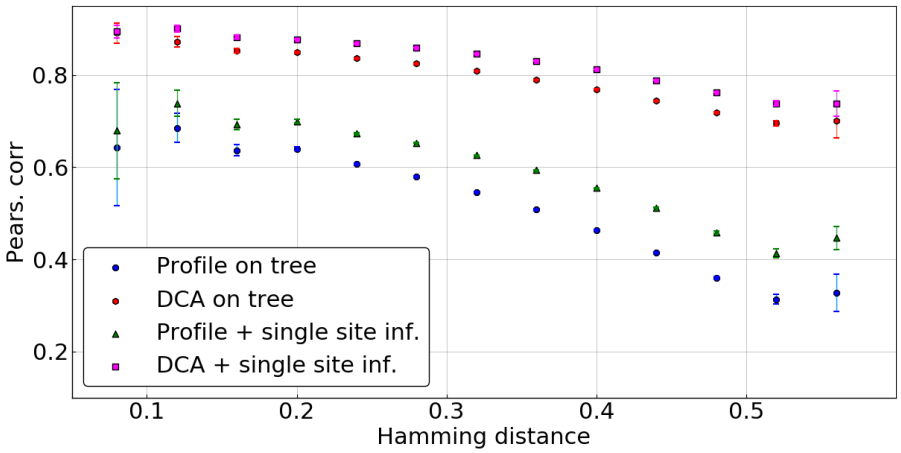

Figure b.3: Pearson correlation in predicting energies of single mutants averaged over reference sequence at a given hamming distance to the closest sequence in the biased sample, as a function of this hamming distance. Error bars are inversely proportional to the square root of the number of sequences in each hamming distance bin. Profile and Potts models are inferred either directly using biased data, or using corrected single site frequencies.

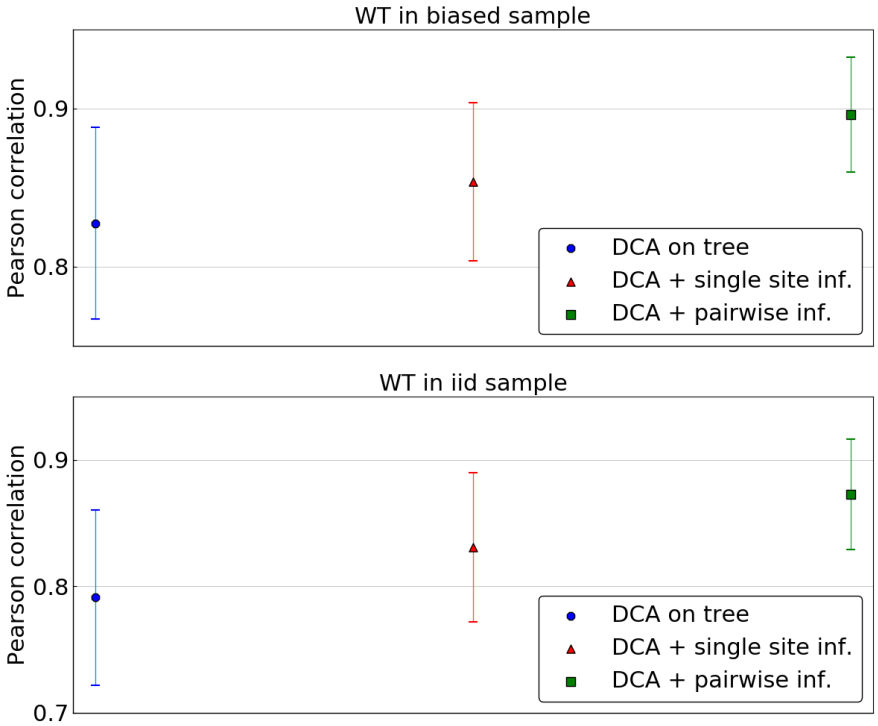

Figure b.4: Pearson correlation in predicting energies of single mutants averaged over sets of reference sequence. In the top panel, reference sequences are taken in the biased sample, *i.e.* among the leaves of the phylogenetic tree. In the bottom panel, reference sequences are taken in a fair sample of P^0. Predictions are made using a DCA model inferred either directly on biased data, either using corrected single site frequencies, either using corrected pairwise frequencies. Error bars indicate the standard deviation across the 30 repetitions of the tree sampling process.

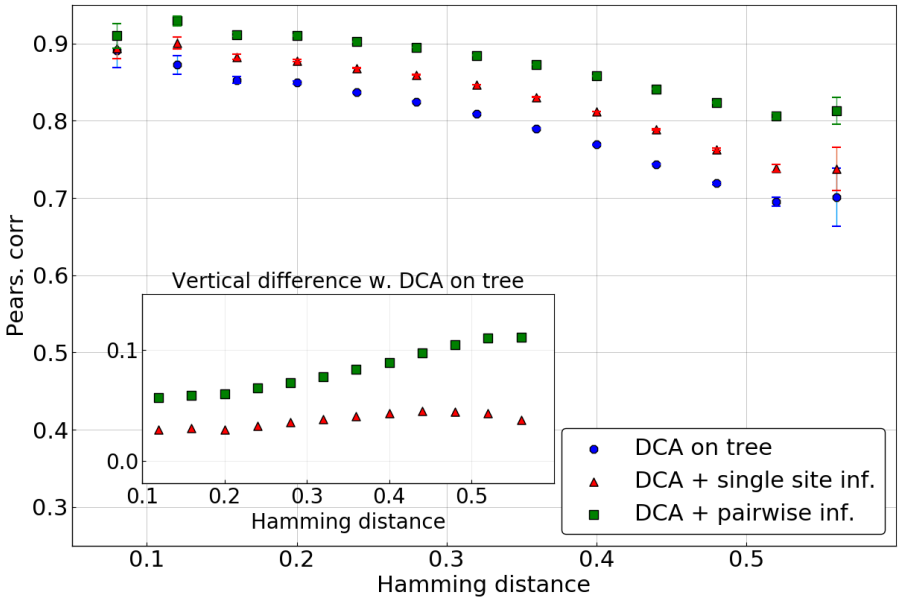

Figure b.5: Pearson correlation in predicting energies of single mutants averaged over reference sequence at a given hamming distance to the closest sequence in the biased sample, as a function of this hamming distance. Error bars are inversely proportional to the square root of the number of sequences in each hamming distance bin. The Potts model is inferred either directly on biased data, either using corrected single site frequencies, either using corrected pairwise frequencies.

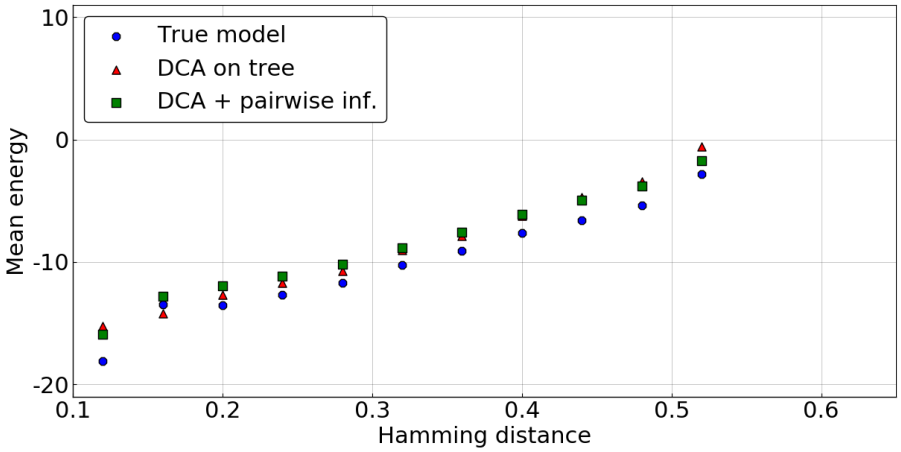

Figure b.6: Average energy of sequences as a function of the hamming distance of the sequence to the closest point in the biased sample.

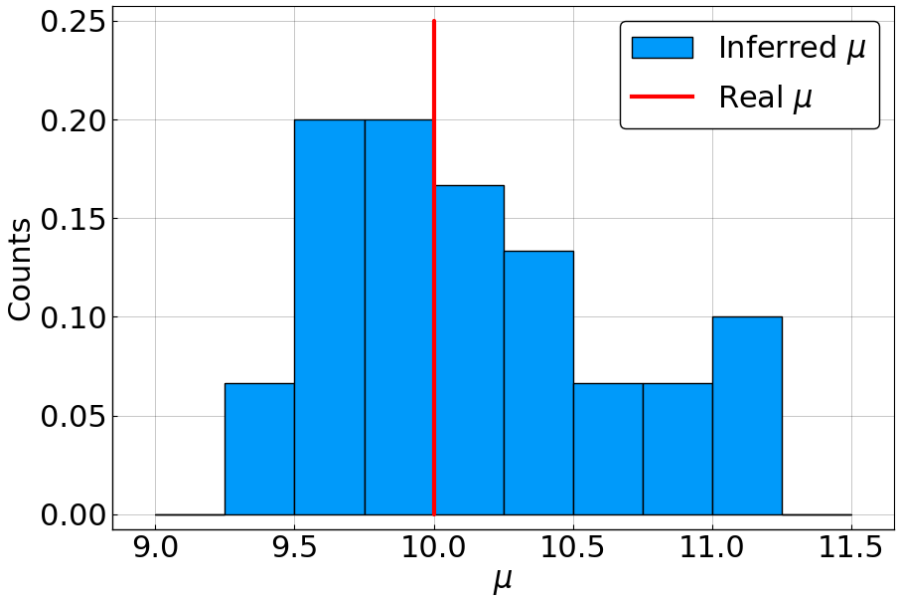

Figure b.7: Histogram of the inferred μ values for the 30 different repetitions of the simulated data. In red is the value $\mu = 10$ used to generate the data.

BIBLIOGRAPHY

[1] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. "A Learning Algorithm for Boltzmann Machines*." en. In: *Cognitive Science* 9.1 (Jan. 1985), pp. 147–169. ISSN: 1551-6709. DOI: 10.1207/s15516709cog0901_7.

[2] Bruce Alberts. *Essential Cell Biology*. en. Fourth edition. New York, NY: Garland Science, 2013. ISBN: 978-0-8153-4454-4 978-0-8153-4455-1.

[3] Erik Aurell. "The Maximum Entropy Fallacy Redux?" en. In: *PLOS Computational Biology* 12.5 (May 2016). Ed. by Andrea Pagnani, e1004777. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004777.

[4] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G. Carbonell, Su-In Lee, and Christopher James Langmead. "Learning Generative Models for Protein Fold Families." en. In: *Proteins: Structure, Function, and Bioinformatics* 79.4 (Apr. 2011), pp. 1061–1078. ISSN: 08873585. DOI: 10.1002/prot.22934.

[5] Carlo Baldassi, Marco Zamparo, Christoph Feinauer, Andrea Procaccini, Riccardo Zecchina, Martin Weigt, and Andrea Pagnani. "Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners." In: *PLoS One* 9.3 (Mar. 2014). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0092721.

[6] Pierre Barrat-Charlaix, Matteo Figliuzzi, and Martin Weigt. "Improving Landscape Inference by Integrating Heterogeneous Data in the Inverse Ising Problem." en. In: *Scientific Reports* 6.1 (Dec. 2016). ISSN: 2045-2322. DOI: 10.1038/srep37812.

[7] J. P. Barton, S. Cocco, E. De Leonardis, and R. Monasson. "Large Pseudocounts and L 2 -Norm Penalties Are Necessary for the Mean-Field Inference of Ising and Potts Models." en. In: *Physical Review E* 90.1 (July 2014). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.90.012132.

[8] J. P. Barton, E. De Leonardis, A. Coucke, and S. Cocco. "ACE: Adaptive Cluster Expansion for Maximum Entropy Graphical Model Inference." en. In: *Bioinformatics* 32.20 (Oct. 2016), pp. 3089–3097. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btw328.

[9] William Bialek and Rama Ranganathan. "Rediscovering the Power of Pairwise Interactions." In: *arXiv:0712.4397 [q-bio]* (Dec. 2007). arXiv: 0712.4397 [q-bio].

[10] Anne-Florence Bitbol, Robert S. Dwyer, Lucy J. Colwell, and Ned S. Wingreen. "Inferring Interaction Partners from Protein Sequences." In: *Proc Natl Acad Sci U S A* 113.43 (Oct. 2016), pp. 12180–12185. ISSN: 0027-8424. DOI: 10.1073/pnas.1606762113.

[11] Michael S. Breen, Carsten Kemena, Peter K. Vlasov, Cedric Notredame, and Fyodor A. Kondrashov. "Epistasis as the Primary Factor in Molecular Evolution." In: *Nature* 490.7421 (Oct. 2012), pp. 535–538. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11510.

[12] Lukas Burger and Erik van Nimwegen. "Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments." en. In: *PLoS Computational Biology* 6.1 (Jan. 2010). Ed. by Philip E. Bourne, e1000633. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000633.

[13] Thomas C. Butler, John P. Barton, Mehran Kardar, and Arup K. Chakraborty. "Identification of Drug Resistance Mutations in HIV from Constraints on Natural Evolution." en. In: *Physical Review E* 93.2 (Feb. 2016). ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.93.022412. arXiv: 1508.01469.

[14] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale. "Scale-Free Correlations in Starling Flocks." en. In: *Proceedings of the National Academy of Sciences* 107.26 (June 2010), pp. 11865–11870. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1005766107.

[15] Andrea Cavagna, Irene Giardina, Alberto Orlandi, Giorgio Parisi, Andrea Procaccini, Massimiliano Viale, and Vladimir Zdravkovic. "The STARFLAG Handbook on Collective Animal Behaviour: 1. Empirical Methods." en. In: *Animal Behaviour* 76.1 (July 2008), pp. 217–236. ISSN: 00033472. DOI: 10.1016/j.anbehav.2008.02.002.

[16] R. R. Cheng, O. Nordesjö, R. L. Hayes, H. Levine, S. C. Flores, J. N. Onuchic, and F. Morcos. "Connecting the Sequence-Space of Bacterial Signaling Proteins to Phenotypes Using Coevolutionary Landscapes." In: *Mol Biol Evol* 33.12 (Dec. 2016), pp. 3054–3064. ISSN: 0737-4038. DOI: 10.1093/molbev/msw188.

[17] S. Cocco and R. Monasson. "Adaptive Cluster Expansion for the Inverse Ising Problem: Convergence, Algorithm and Tests." en. In: *Journal of Statistical Physics* 147.2 (Apr. 2012), pp. 252–314. ISSN: 0022-4715, 1572-9613. DOI: 10.1007/s10955-012-0463-4.

[18] Simona Cocco, Remi Monasson, and Martin Weigt. "From Principal Component to Direct Coupling Analysis of Coevolution in Proteins: Low-Eigenvalue Modes Are Needed for Structure Prediction." en. In: *PLoS Computational Biology* 9.8 (Aug. 2013).

Ed. by Björn Wallner, e1003176. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003176.

[19] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Remi Monasson, and Martin Weigt. "Inverse Statistical Physics of Protein Sequences: A Key Issues Review." In: *Reports on Progress in Physics* 81.3 (Mar. 2018), p. 032601. ISSN: 0034-4885, 1361-6633. DOI: 10.1088/1361-6633/aa9965. arXiv: 1703.01222.

[20] Alice Coucke. "Statistical Modeling of Protein Sequences beyond Structural Prediction : High Dimensional Inference with Correlated Data." PhD thesis. Paris Sciences et Lettres, Oct. 2016.

[21] Alice Coucke, Guido Uguzzoni, Francesco Oteri, Simona Cocco, Remi Monasson, and Martin Weigt. "Direct Coevolutionary Couplings Reflect Biophysical Residue Interactions in Proteins." en. In: (June 2016). DOI: 10.1101/061390.

[22] Rupika Delgoda and James Douglas Pulfer. "A Guided Monte Carlo Search Algorithm for Global Optimization of Multidimensional Functions †." en. In: *Journal of Chemical Information and Computer Sciences* 38.6 (Nov. 1998), pp. 1087–1095. ISSN: 0095-2338. DOI: 10.1021/ci9701042.

[23] E. W. Dijkstra. "A Note on Two Problemsin Connexionwith Graphs." en. In: *Numerische Mathematik* (1959), p. 3.

[24] S.D. Dunn, L.M. Wahl, and G.B. Gloor. "Mutual Information without the Influence of Phylogeny or Entropy Dramatically Improves Residue Contact Prediction." en. In: *Bioinformatics* 24.3 (Feb. 2008), pp. 333–340. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btm604.

[25] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. en. Cambridge: Cambridge University Press, 1998. ISBN: 978-0-511-79049-2. DOI: 10.1017/CBO9780511790492.

[26] S. R. Eddy. "Profile Hidden Markov Models." en. In: *Bioinformatics* 14.9 (Oct. 1998), pp. 755–763. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/14.9.755.

[27] Magnus Ekeberg, Tuomo Hartonen, and Erik Aurell. "Fast Pseudolikelihood Maximization for Direct-Coupling Analysis of Protein Structure from Many Homologous Amino-Acid Sequences." en. In: *Journal of Computational Physics* 276 (Nov. 2014), pp. 341–356. ISSN: 00219991. DOI: 10.1016/j.jcp.2014.07.024.

[28] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. "Improved Contact Prediction in Proteins: Using Pseudolikelihoods to Infer Potts Models." en. In: *Physical Review E* 87.1 (Jan. 2013). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.87.012707. arXiv: 1211.1281.

[29] Christoph Feinauer, Hendrik Szurmant, Martin Weigt, and Andrea Pagnani. "Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon." In: *PLoS One* 11.2 (Feb. 2016). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0149166.

[30] Joseph Felsenstein. *Journal of Molecular Evolution © Springer-Verlag 1981 Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach*.

[31] Ulisse Ferrari, Stephane Deny, Matthew Chalk, Gasper Tkacik, Olivier Marre, and Thierry Mora. "Separating Intrinsic Interactions from Extrinsic Correlations in a Network of Sensory Neurons." en. In: (Feb. 2018). DOI: 10.1101/243816.

[32] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. "How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins?" en. In: *Molecular Biology and Evolution* 35.4 (Apr. 2018), pp. 1018–1027. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msy007.

[33] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. "Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1." en. In: *Molecular Biology and Evolution* 33.1 (Jan. 2016), pp. 268–280. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msv211.

[34] R. D. Finn, J. Clements, and S. R. Eddy. "HMMER Web Server: Interactive Sequence Similarity Searching." en. In: *Nucleic Acids Research* 39.suppl (July 2011), W29–W37. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkr367.

[35] Robert D. Finn et al. "The Pfam Protein Families Database: Towards a More Sustainable Future." en. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D279–D285. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv1344.

[36] William F. Flynn, Allan Haldane, Bruce E. Torbett, and Ronald M. Levy. "Inference of Epistatic Effects Leading to Entrenchment and Drug Resistance in HIV-1 Protease." In: *Mol Biol Evol* 34.6 (June 2017), pp. 1291–1306. ISSN: 0737-4038. DOI: 10.1093/molbev/msx095.

[37] Anthony A. Fodor and Richard W. Aldrich. "Influence of Conservation on Calculations of Amino Acid Covariance in Multiple Sequence Alignments." en. In: *Proteins: Structure, Function, and Bioinformatics* 56.2 (Aug. 2004), pp. 211–221. ISSN: 1097-0134. DOI: 10.1002/prot.20098.

[38] A Georges and J S Yedidia. "How to Expand around Mean-Field Theory Using High-Temperature Expansions." en. In: *Journal of Physics A: Mathematical and General* 24.9 (May 1991), pp. 2173–2192. ISSN: 0305-4470, 1361-6447. DOI: 10.1088/0305-4470/24/9/024.

[39] Thomas Gueudré, Carlo Baldassi, Marco Zamparo, Martin Weigt, and Andrea Pagnani. "Simultaneous Identification of Specifically Interacting Paralogs and Interprotein Contacts by Direct Coupling Analysis." In: *Proc Natl Acad Sci U S A* 113.43 (Oct. 2016), pp. 12186–12191. ISSN: 0027-8424. DOI: 10.1073/pnas.1607570113.

[40] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. "Protein Sectors: Evolutionary Units of Three-Dimensional Structure." en. In: *Cell* 138.4 (Aug. 2009), pp. 774–786. ISSN: 00928674. DOI: 10.1016/j.cell.2009.07.038.

[41] Allan Haldane, William F. Flynn, Peng He, R.S.K. Vijayan, and Ronald M. Levy. "Structural Propensities of Kinase Family Proteins from a Potts Model of Residue Co-Variation: Structural Propensities of Kinase Family Proteins." en. In: *Protein Science* 25.8 (Aug. 2016), pp. 1378–1384. ISSN: 09618368. DOI: 10.1002/pro.2954.

[42] Michael J. Harms and Joseph W. Thornton. "Evolutionary Biochemistry: Revealing the Historical and Physical Causes of Protein Properties." en. In: *Nature Reviews Genetics* 14.8 (Aug. 2013), pp. 559–571. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3540.

[43] Thomas A Hopf, Charlotta P I Schärfe, João P G L M Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre M J J Bonvin, and Debora S Marks. "Sequence Co-Evolution Gives 3D Contacts and Structures of Protein Complexes." In: *eLife* 3 (). ISSN: 2050-084X. DOI: 10.7554/eLife.03430.

[44] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P.I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. "Mutation Effects Predicted from Sequence Co-Variation." In: *Nat Biotechnol* 35.2 (Feb. 2017), pp. 128–135. ISSN: 1087-0156. DOI: 10.1038/nbt.3769.

[45] Hervé Jacquier et al. "Capturing the Mutational Landscape of the Beta-Lactamase TEM-1." In: *Proc Natl Acad Sci U S A* 110.32 (Aug. 2013), pp. 13067–13072. ISSN: 0027-8424. DOI: 10.1073/pnas.1215206110.

[46] E. T. Jaynes. "Information Theory and Statistical Mechanics." In: *Phys. Rev.* 106.4 (May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[47] E. T. Jaynes. *Probability Theory: The Logic of Science*. en. Cambridge University Press, 2003.

[48] David T. Jones, Daniel W. A. Buchan, Domenico Cozzetto, and Massimiliano Pontil. "PSICOV: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments." eng. In: *Bioinformatics* 28.2 (Jan. 2012), pp. 184–190. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr638.

[49] David T. Jones, Tanya Singh, Tomasz Kosciolek, and Stuart Tetchner. "MetaPSICOV: Combining Coevolution Methods for Accurate Prediction of Contacts and Long Range Hydrogen Bonding in Proteins." In: *Bioinformatics* 31.7 (Apr. 2015), pp. 999–1006. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu791.

[50] Anders Krogh. "An Introduction to Hidden Markov Models for Biological Sequences." In: *Computational Methods in Molecular Biology, Elsevier*. 1998.

[51] C. D. Livingstone and G. J. Barton. "Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation." eng. In: *Comput. Appl. Biosci.* 9.6 (Dec. 1993), pp. 745–756. ISSN: 0266-7061.

[52] Jaclyn K. Mann, John P. Barton, Andrew L. Ferguson, Saleha Omarjee, Bruce D. Walker, Arup Chakraborty, and Thumbi Ndung'u. "The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing." In: *PLoS Comput Biol* 10.8 (Aug. 2014). ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1003776.

[53] Richard N. McLaughlin Jr, Frank J. Poelwijk, Arjun Raman, Walraj S. Gosal, and Rama Ranganathan. "The Spatial Architecture of Protein Function and Adaptation." en. In: *Nature* 491.7422 (Nov. 2012), pp. 138–142. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11500.

[54] Richard N. McLaughlin, Frank J. Poelwijk, Arjun Raman, Walraj S. Gosal, and Rama Ranganathan. "The Spatial Architecture of Protein Function and Adaptation." In: *Nature* 491.7422 (Nov. 2012), pp. 138–142. ISSN: 0028-0836. DOI: 10.1038/nature11500.

[55] Daniel Melamed, David L. Young, Caitlin E. Gamble, Christina R. Miller, and Stanley Fields. "Deep Mutational Scanning of an RRM Domain of the Saccharomyces Cerevisiae Poly(A)-Binding Protein." In: *RNA* 19.11 (Nov. 2013), pp. 1537–1551. ISSN: 1355-8382. DOI: 10.1261/rna.040709.113.

[56] Sanzo Miyazawa and Robert L. Jernigan. "Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading." In: *Journal of Molecular Biology* 256.3 (Mar. 1996), pp. 623–644. ISSN: 0022-2836. DOI: 10.1006/jmbi.1996.0114.

[57] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. "Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families." en. In: *Proceedings of the National Academy of Sciences* 108.49 (Dec. 2011), E1293–E1301. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1111471108.

[58] C. Anders Olson, Nicholas C. Wu, and Ren Sun. "A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain." In: *Curr Biol* 24.22 (Nov. 2014), pp. 2643–2651. ISSN: 0960-9822. DOI: 10.1016/j.cub.2014.09.072.

[59] A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick. "Ab Initio Folding of Proteins Using Restraints Derived from Evolutionary Information." eng. In: *Proteins* Suppl 3 (1999), pp. 177–185. ISSN: 0887-3585.

[60] Jakub Otwinowski. "Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function." en. In: *arXiv:1802.08744 [q-bio]* (Feb. 2018). arXiv: 1802.08744 [q-bio].

[61] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. "Robust and Accurate Prediction of Residue–Residue Interactions across Protein Interfaces Using Evolutionary Information." en. In: *eLife* 3 (2014). DOI: 10.7554/eLife.02030.

[62] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. "Protein Structure Determination Using Metagenome Sequence Data." en. In: *Science* 355.6322 (Jan. 2017), pp. 294–298. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aah4043.

[63] *Pfam 31.0 Is Released*. en. Mar. 2017.

[64] A. I. Podgornaia and M. T. Laub. "Pervasive Degeneracy and Epistasis in a Protein-Protein Interface." en. In: *Science* 347.6222 (Feb. 2015), pp. 673–677. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1257360.

[65] Lorenzo Posani, Simona Cocco, Karel Ježek, and Rémi Monasson. "Functional Connectivity Models for Decoding of Spatial Representations from Hippocampal CA1 Recordings." en. In: *Journal of Computational Neuroscience* 43.1 (Aug. 2017), pp. 17–33. ISSN: 0929-5313, 1573-6873. DOI: 10.1007/s10827-017-0645-9.

[66] Chongli Qin and Lucy J. Colwell. "Power Law Tails in Phylogenetic Systems." en. In: *Proceedings of the National Academy of Sciences* 115.4 (Jan. 2018), pp. 690–695. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1711913115.

[67] Nathan J Rollins, Kelly P Brock, Frank J Poelwijk, Michael A Stiffler, Nicholas P Gauthier, Chris Sander, and Debora S Marks. "3D Protein Structure from Genetic Epistasis Experiments." en. In: (May 2018). DOI: 10.1101/320721.

[68] William P. Russ, Drew M. Lowery, Prashant Mishra, Michael B. Yaffe, and Rama Ranganathan. "Natural-like Function in Artificial WW Domains." en. In: *Nature* 437.7058 (Sept. 2005), pp. 579–583. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature03990.

[69] Michael Schmidt and Kay Hamacher. "Three-Body Interactions Improve Contact Prediction within Direct-Coupling Analysis." en. In: *Physical Review E* 96.5 (Nov. 2017). ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.96.052405.

[70] Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. "Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population." en. In: *Nature* 440.7087 (Apr. 2006), pp. 1007–1012. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature04701.

[71] Alexander Schug, Martin Weigt, José N. Onuchic, Terence Hwa, and Hendrik Szurmant. "High-Resolution Protein Complexes from Integrating Genomic Information with Molecular Simulation." en. In: *Proceedings of the National Academy of Sciences* 106.52 (Dec. 2009), pp. 22124–22129. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0912100106.

[72] Benjamin Schuster-Böckler, Jörg Schultz, and Sven Rahmann. "HMM Logos for Visualization of Protein Families." en. In: *BMC Bioinformatics* (2004), p. 8.

[73] David J. Schwab, Ilya Nemenman, and Pankaj Mehta. "Zipf's Law and Criticality in Multivariate Data without Fine-Tuning." en. In: *Physical Review Letters* 113.6 (Aug. 2014). ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.113.068102.

[74] Marcin J. Skwark, Daniele Raimondi, Mirco Michel, and Arne Elofsson. "Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns." en. In: *PLoS Computational Biology* 10.11 (Nov. 2014). Ed. by Guanghong Wei, e1003889. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003889.

[75] Michael Socolich, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner, and Rama Ranganathan. "Evolutionary Information for Specifying a Protein Fold." en. In: *Nature* 437.7058 (Sept. 2005), pp. 512–518. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature03991.

[76] Ludovico Sutto, Simone Marsili, Alfonso Valencia, and Francesco Luigi Gervasio. "From Residue Coevolution to Protein Conformational Ensembles and Functional Dynamics." en. In: *Proceedings of the National Academy of Sciences* 112.44 (Nov. 2015),

pp. 13567–13572. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1508584112.

[77] Toshiyuki Tanaka. "Mean-Field Theory of Boltzmann Machine Learning." en. In: *Physical Review E* 58.2 (Aug. 1998), pp. 2302–2310. ISSN: 1063-651X, 1095-3787. DOI: 10.1103/PhysRevE.58.2302.

[78] The UniProt Consortium. "UniProt: A Hub for Protein Information." en. In: *Nucleic Acids Research* 43.D1 (Jan. 2015), pp. D204–D212. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gku989.

[79] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. "Learning Protein Constitutive Motifs from Sequence Data." en. In: *arXiv:1803.08718 [q-bio]* (Mar. 2018). arXiv: 1803.08718 [q-bio].

[80] Susann Vorberg, Stefan Seemayer, and Johannes Soeding. "Synthetic Protein Alignments by CCMgen Quantify Noise in Residue-Residue Contact Prediction." en. In: (June 2018). DOI: 10.1101/344333.

[81] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model." en. In: *PLOS Computational Biology* (2017), p. 34.

[82] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. "Identification of Direct Residue Contacts in Protein-Protein Interaction by Message Passing." en. In: *Proceedings of the National Academy of Sciences* 106.1 (Jan. 2009), pp. 67–72. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0805923106.

[83] Alexander Wlodawer, Jochen Walter, Robert Huber, and Lennart Sjölin. "Structure of Bovine Pancreatic Trypsin Inhibitor: Results of Joint Neutron and X-Ray Refinement of Crystal Form II." In: *Journal of Molecular Biology* 180.2 (Dec. 1984), pp. 301–329. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(84)80006-6.

[84] Jin Y. Yen. "Finding the K Shortest Loopless Paths in a Network." en. In: *Management Science* (July 1971), p. 6.

[85] Haicang Zhang, Yujuan Gao, Minghua Deng, Chao Wang, Jianwei Zhu, Shuai Cheng Li, Wei-Mou Zheng, and Dongbo Bu. "Improving Residue–Residue Contact Prediction via Low-Rank and Sparse Decomposition of Residue Correlation Matrix." en. In: *Biochemical and Biophysical Research Communications* 472.1 (Mar. 2016), pp. 217–222. ISSN: 0006291X. DOI: 10.1016/j.bbrc.2016.01.188.

[86] Erik van Nimwegen. "Finding Regulatory Elements and Regulatory Motifs: A General Probabilistic Framework." en. In: *BMC Bioinformatics* 8.Suppl 6 (2007), S4. ISSN: 14712105. DOI: 10.1186/1471-2105-8-S6-S4.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

Final Version as of September 14, 2018 (classicthesis v4.6).

Pierre BARRAT-CHARLAIX **9 novembre 2018**

Sujet : Comprendre et améliorer les modèles statistiques de séquences de protéines

Résumé : Dans les dernières décennies, les progrès des techniques expérimentales ont permis une augmentation considérable du nombre de séquences d'ADN et de protéines connues. Cela a incité au développement de méthodes statistiques variées visant à tirer parti de cette quantité massive de données. Les méthodes dites co-évolutives en font partie, utilisant des idées de physique statistique pour construire un modèle global de la variabilité des séquences de protéines. Ces méthodes se sont montrées très efficaces pour extraire des informations pertinentes des seules séquences, comme des contacts structurels ou des effets mutationnels. Alors que les modèles co-évolutifs sont pour l'instant utilisés comme outils prédictifs, leur succès plaide pour une meilleure compréhension de leur fonctionnement. Dans cette thèse, nous proposons des élaborations sur les méthodes déjà existantes tout en questionnant leur fonctionnement. Nous étudions premièrement la capacité de l'Analyse en Couplages Directs (DCA) à reproduire les motifs statistiques rencontrés dans les séquences des familles de protéines. Puis est présentée la possibilité d'inclure d'autres types d'information comme des effets mutationnels dans cette méthode, suivie de corrections potentielles des biais phylogénétiques présents dans les données utilisées. Finalement, des considérations sur les limites des modèles co-évolutifs actuels sont exposées, de même que des suggestions pour les surmonter.

Mots clés : co-évolution, protéines, modèles statistiques, inférence statistique, entropie maximale, physique statistique, phylogénie

Subject : Understanding and improving statistical models of protein sequences

Abstract: In the last decades, progress in experimental techniques have given rise to a vast increase in the number of known DNA and protein sequences. This has prompted the development of various statistical methods in order to make sense of this massive amount of data. Among those are pairwise co-evolutionary methods, using ideas coming from statistical physics to construct a global model for protein sequence variability. These methods have proven to be very effective at extracting relevant information from sequences, such as structural contacts or effects of mutations. While co-evolutionary models are for the moment used as predictive tools, their success calls for a better understanding of they functioning. In this thesis, we propose developments on existing methods while also asking the question of how and why they work. We first focus on the ability of the so-called Direct Coupling Analysis (DCA) to reproduce statistical patterns found in sequences in a protein family. We then discuss the possibility to include other types of information such as mutational effects in this method, and then potential corrections for the phylogenetic biases present in available data. Finally, considerations about limitations of current co-evolutionary models are presented, along with suggestions on how to overcome them.

Keywords : co-evolution, proteins, statistical models, statistical inference, maximum-entropy, statistical physics, phylogeny