Statistical modeling of protein sequences beyond structural prediction : high dimensional inference with correlated data

Alice Coucke

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à
l'École Normale Supérieure

Statistical modeling of protein sequences beyond structural prediction

High-dimensional inference with correlated data

École doctorale nº 564 *Physique en Île-de-France*
Spécialité *Physique théorique*

Soutenue par **Alice Coucke**
le 10 octobre 2016

Dirigée par
Rémi Monasson
et **Martin Weigt**

Composition du jury

M. Andrea De Martino
Université La Sapienza de Rome
Rapporteur

M. Olivier Rivoire
Collège de France
Rapporteur

M. Olivier Martin
INRA
Membre du jury

Mme Aleksandra Walczak
ENS - CNRS
Membre du jury

M. Rémi Monasson
ENS - CNRS
Directeur de thèse

M. Martin Weigt
UPMC
Directeur de thèse

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS

The literature of science is filled with answers found when the question propounded had an entirely different direction and end.

— John Steinbeck

There was a lot more to magic, as Harry quickly found out, than waving your wand and saying a few funny words.

— J.K. Rowling

ABSTRACT

Over the last decades, genomic databases have grown exponentially in size thanks to the constant progress of modern DNA sequencing. A large variety of statistical tools have been developed, at the interface between bioinformatics, machine learning, and statistical physics, to extract information from these ever increasing datasets. In the specific context of protein sequence data, several approaches have been recently introduced by statistical physicists, such as direct-coupling analysis, a global statistical inference method based on the maximum-entropy principle, that has proven to be extremely effective in predicting the three-dimensional structure of proteins from purely statistical considerations.

In this dissertation, we review the relevant inference methods and, encouraged by their success, discuss their extension to other challenging fields, such as sequence folding prediction and homology detection. Contrary to residue-residue contact prediction, which relies on an intrinsically topological information about the network of interactions, these fields require global energetic considerations and therefore a more quantitative and detailed model. Through an extensive study on both artificial and biological data, we provide a better interpretation of the central inferred parameters, up to now poorly understood, especially in the limited sampling regime. Finally, we present a new and more precise procedure for the inference of generative models, which leads to further improvements on real, finitely sampled data.

KEYWORDS: inference, statistical learning, regularization, maximum entropy, protein coevolution, statistical modeling of protein sequences, maximum likelihood, mean field, pseudolikelihood, cluster expansion

RÉSUMÉ

Grâce aux progrès des techniques de séquençage, les bases de données génomiques ont connu une croissance exponentielle depuis la fin des années 1990. Un grand nombre d'outils statistiques ont été développés à l'interface entre bioinformatique, apprentissage automatique et physique statistique, dans le but d'extraire de l'information de ce déluge de données. Plusieurs approches de physique statistique ont été récemment introduites dans le contexte précis de la modélisation de séquences de protéines, dont l'analyse en couplages directs. Cette méthode d'inférence statistique globale fondée sur le principe d'entropie maximale, s'est récemment montrée d'une efficacité redoutable pour prédire la structure tridimensionnelle de protéines, à partir de considérations purement statistiques.

Dans cette thèse, nous présentons les méthodes d'inférence en question, et encouragés par leur succès, explorons d'autres domaines complexes dans lesquels elles pourraient être appliquées, comme la prédiction de repliement de protéines ou la détection d'homologies. Contrairement à la prédiction des contacts entre résidus qui se limite à une information topologique sur le réseau d'interactions, ces nouveaux champs d'application exigent des considérations énergétiques globales et donc un modèle plus quantitatif et détaillé. À travers une étude approfondie sur des données artificielles et biologiques, nous proposons une meilleure interpretation des paramètres centraux de ces méthodes d'inférence, jusqu'ici mal compris, notamment dans le cas d'un échantillonnage limité. Enfin, nous présentons une nouvelle procédure plus précise d'inférence de modèles génératifs, qui mène à des avancées importantes pour des données réelles en quantité limitée.

MOTS-CLEFS : inférence, apprentissage statistique, régularisation, entropie maximale, coévolution des protéines, modélisation statistique des séquences de protéines, vraisemblance maximale, champ moyen, pseudo vraisemblance, développement en grappe

REMERCIEMENTS

Je tiens à remercier Rémi Monasson et Martin Weigt d'avoir accepté de diriger ma thèse entre le laboratoire de physique théorique de l'École Normale Supérieure et le laboratoire de biologie computationnelle et quantitative de l'Université Pierre et Marie Curie. Leur exigence et l'autonomie qu'ils m'ont accordée auront fait de ces trois dernières années une expérience unique dont les enseignements me suivront encore longtemps.

Je voudrais exprimer la plus grande gratitude envers mes collaborateurs. Tout d'abord à Guido Uguzzoni, pour sa disponibilité et sa pédagogie, sans oublier les heures passées à regarder des taux de vrais positifs ou des figures colorées. À John Barton pour des discussions toujours enrichissantes. Merci à Eleonora De Leonardis, pour sa clarté et son enthousiasme, ainsi que pour son amitié qui m'est très chère. Enfin, je suis tout particulièrement reconnaissante envers Simona Cocco pour sa disponibilité et nos fructueuses et agréables interactions tout au long de ma dernière année de thèse.

Je remercie Olivier Martin et Aleksandra Walczack d'avoir accepté de participer au jury de ma thèse, et Andrea De Martino et Olivier Rivoire d'en être les rapporteurs.

Je remercie également Jean-Marc Berroir de m'avoir écoutée et soutenue, Aleksandra Walczack (à nouveau) d'avoir cru en moi, Sylvie Hénon pour ces deux merveilleuses années d'enseignement et Sébastien Balibar pour les nombreuses discussions passionnantes. Merci à Giulio Biroli pour m'avoir appris à associer physique et voile et whisky et chocolat. Merci aussi à Jean-François Allemand de m'avoir convaincue de venir à l'ENS un jour d'août 2010. Enfin, je voudrais remercier Yann Brunel pour son enthousiasme communicatif pour la physique et sa bienveillance, sans qui je ne serais sûrement pas là où je suis.

Viviane Sébille, Sandrine Patacchini et Claire Bourliaud m'ont fourni une aide très précieuse dans un contexte administratif compliqué, qu'elles en soient ici remerciées.

J'ai eu la chance de rencontrer et côtoyer des personnes exceptionnelles entre les rues Lhomond et de l'école de médecine. Par ordre d'apparition Tom, Jonathan, Thibaud, Thimothée, Suzanne, Antoine, Jean, Sophie, Dario, Matteo, Juliana, Alberto, Ulisse, Ralph, Quentin, Andreas, Christoph, Pierre, Lorenzo. Merci pour, pêle-mêle, l'élevage de dumb cane, le taboulé, les jeux de piste du CPER, les débats sur il caffè, la pasta, e la 'nduja calabrese, les jeux de pôt, le benchmarking du flan du $5^{ème}$ arrondissement, l'élucidation du mystère de la machine à glaçons, l'eau municipale et son menu de luxe ... En somme,

d'avoir rendu supportables les coups durs et les désillusions passagères. Je remercie aussi mes amis de plus longue date – je pense à Faustine, Diane, Marie, Elisa, Tancrède, Margot, Raphaëlle, Minh-Tu, Gabi – qui ont pardonné mes errances théoriques ; à Sophia, qui m'a rendu visite partout, ou presque.

Je voudrais remercier ici Declan McCavana et les coaches de la French Debating Association qui m'ont accueillie dans leur *big family*, donnant à mes années de thèse ce je-ne-sais-quoi qui leur manquait.

Je n'aurais probablement jamais pu achever cette thèse sans le soutien indéfectible de ma famille, et je voudrais leur témoigner ici toute mon affection. Merci aux Levené/Gaudron/Benech pour tous ces rires et ces discussions sans fin, sans oublier les embuscades et la sacro-sainte pizza-houmous du dimanche soir. Aux grumeaux, aka Jeanine et Dédé, pardon pour ce que j'ai dit quand j'avais faim ; à Rosa dont je suis si fière et qui ne cesse de m'étonner, notamment pour sa connaissance encyclopédique des champignons, des émulsions et de la musculation ; à Léon avec un "n" le grand lettré de la famille, de loin le plus passionné, aussi bien des discontinuités en géographie que des tartines de beurre sans beurre. Merci à ma grande soeur Caroline, qui m'a appris tant de choses, ainsi qu'à Khéo et Evan, promis maintenant c'est bien fini ! À mes parents, enfin, pour leur anticonformisme que j'ai mis trop longtemps à accepter et l'éducation déplorable qui m'a menée jusqu'ici.

Alaa, me supporter pendant ces trois années a du être au moins aussi frustrant qu'essayer d'améliorer le score APC. Aussi, je laisserai ce bon vieux Charles Bukowski s'exprimer à ma place : "The free soul is rare, but you know it when you see it – basically because you feel good, very good, when you are near or with them".

CONTENTS

ACRONYMS

ACE Adaptive Cluster Expansion

APC Average Product Correction

AUC Area Under the Curve

BML Boltzmann Machine Learning

cplmDCA Compressed Pseudolikelihood Maximization

DCA Direct-Coupling Analysis

DI Direct Information

HMM Hidden Markov Model

i.i.d independent and identically distributed

KL Kullback-Leibler

LP Lattice Proteins

MaxEnt Maximum-Entropy Principle

MC Monte Carlo

MCMC Monte Carlo Markov Chain

mfDCA Mean-Field Direct-Coupling Analysis

MI Mutual Information

MJ Miyazawa-Jernigan

MSA Multiple Sequence Alignment

PCA Principal Component Analysis

PDB Protein Data Bank

plmDCA Pseudolikelihood Maximization

PPV Positive Predictive Value

PSICOV Protein Sparse Inverse Covariance Estimation

ROC Receiver Operating Characteristic

RSA Relative Solvent Accessibility

SCOP Structural Classification of Proteins

FOREWORD

The exponential growth of genomic databases over the recent years has prompted a surge of interest among researchers in the fields of bioinformatics, machine learning, and statistical physics. A large variety of statistical tools have been developed to extract information from these ever increasing datasets. In statistical physics, estimating the probability distribution from which a given dataset may have been generated is referred to as the inverse problem. Direct-coupling analysis is a generic name for several approximate methods to solve the inverse problem based on maximum-entropy modeling, in the specific context of protein sequence data. Such approaches, introduced by statistical physicists, have proven to be extremely effective in predicting the three-dimensional structure of proteins from sequence information alone, reaching a level of accuracy previously thought to be beyond reach.

In this dissertation, we will review these existing methods and, encouraged by their success, explore other challenging fields in which they may be applied. However, we will quickly realize that many things about these models are not fully controlled. Some effort first needs to be put in understanding them better, especially if we want to go beyond protein structure prediction. Indeed, while it has proven possible to make predictions about the tertiary structure of a protein based solely on a map of the interactions between its residues, a more detailed description of these interactions is needed to tackle more complex questions. Fields such as fitness landscape modeling or homology detection can only be addressed through global considerations concerning the energy of whole sequence, and require more quantitative statistical models.

Part I will be therefore dedicated, after a brief word about coevolution in proteins, to the introduction of inverse Potts problems in statistical physics in the context of maximum-entropy modeling. We will present a review of the most popular approximation methods to tackle this problem on biological sequence data. We then explore the specificities of protein domain families in more details, as well as the success of statistical physics approaches in protein structure prediction.

We will discuss the ability of inverse Potts methods to go beyond structural prediction in Part II. We will start by analyzing the data from a recent publication where the authors designed artificial proteins and experimentally tested their ability to fold. Then, we will show that the application of direct-coupling analysis approaches in the context of remote homology detection gives promising but also

unexpected results. It incidentally raises several questions about gaps modeling that we will subsequently address through a more principled approach, which is the object of a publication currently in preparation.

The main focus of Part III will be the inferred Potts couplings. When used for residue-residue contact prediction, these couplings are usually mapped onto simple scalar parameters and subsequently ranked, so that the full information they potentially contain gets lost. A detailed understanding of these crucial parameters is lacking. By analyzing 70 protein families, we will provide a quantitative interpretation of the inferred couplings and describe their properties in great details. We will also assess the crucial role of sampling and regularization by studying artificial lattice proteins. This work has led to the following paper "Direct coevolutionary couplings reflect biophysical residue interactions in proteins", A Coucke, G Uguzzoni, F Oteri, S Cocco, R Monasson, and M Weigt, *Journal of Chemical Physics* (2016) [31], currently under review.

Finally, we will describe in Part IV the adaptive cluster expansion, a new approach to inverse problems. Introduced by our group in the Ising case in the context of neural recordings, it has been recently generalized to the Potts case and we will discuss its application to artificial and protein sequence data. We will compare this new approach to standard direct-coupling analysis models on various datasets and study its ability to accurately recover the model parameters, reconstruct the statistics of the input data, and predict protein structure. This work has been recently published in "ACE: adaptive cluster expansion for maximum entropy graphical model inference", JP Barton, E De Leonardis, A Coucke, and S Cocco, *Bioinformatics* (2016) [13]. We will also go into more details about a new compressed representation of the data which aims at reducing overfitting and improving the quality of the inference; a paper on this topic is currently in preparation.

Part I

FROM COEVOLUTION TO INVERSE STATISTICAL PHYSICS

In this introduction, we first motivate the statistical physics approach to tackle genomic data, and in particular present the general principles of modeling coevolution in proteins (Chapter 1). We then provide a review of the inverse Potts model in the context of maximum-entropy modeling and introduce the direct-coupling analysis approach (Chapter 2). These models are vastly used for the computationally challenging task of estimating the probability distribution from which the given data may have been drawn. Finally we go into more details about the specificity of protein sequence data and show that direct-coupling analysis is very successful in exploiting coevolution to predict residue-residue contacts in the protein structure (Chapter 3).

1

A WORD ABOUT COEVOLUTION IN PROTEINS

More is different.

— P. W. Anderson [4]

Over the last decades, the development of new experimental techniques in biology has given rise to a rapid increase in data availability. Consequently, a large variety of statistical tools have been recently developed to extract information from these growing datasets. In particular, genomic databases have known a spectacular exponential growth thanks to the constant progress of modern DNA sequencing technologies, resulting in about 100 million known protein sequences. However, only a small fraction of these sequences have been manually annotated – from 5% in 2010 to 0.5% in 2015 – meaning that some of their biological features have been experimentally identified by a human being. Experimentally extracting the three-dimensional structure of a protein is indeed still hard and costly. On the other hand, while these annotations are continuously updated, the research effort is directed at improving the quality of the annotations (not only structural, but also functional) rather than the quantity. The vast majority of the protein sequences are therefore unreviewed, automatically annotated entries. In the UniProt database [30], a freely accessible database of protein sequences, the gap between the manually annotated Swiss-Prot [22] and the automatically annotated TrEMBL databases dramatically widens, as shown on Fig. 1.1.

The function of a protein mainly depends on its three-dimensional structure [5]. Knowing the structure of a protein is therefore very informative about its potentiality. The ultimate goal would be go back and forth from genotype to phenotype, or in other words from a single protein sequence (amino-acid chain) to its folded structure and function. It would for instance allow to design new drugs targeting specific agents. Given the very large number of degrees of freedom (torsion angles) in an unfolded amino-acid chain, the number of possible spatial configurations is astronomical and the folding energy landscape extremely complex. For a protein of 100 residues, finding the global energy minimum by exploring each possible configuration at the speed of light would take 10^{75} years (known as Levinthal's paradox [69]). Mapping a single sequence to its three-dimensional structure is therefore extremely difficult, even with the most advanced computational techniques – computational protein folding [1] is actu-

1. Computational protein folding uses far more than sequence information alone. It usually includes the physico-chemical properties of amino acids and molecular dynamics simulations.

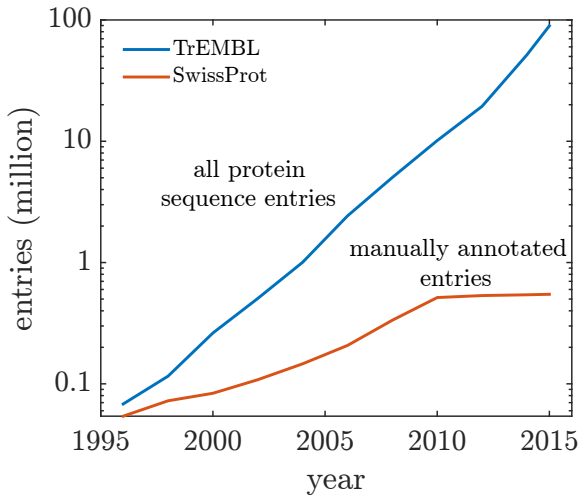


Figure 1.1 – Evolution of the number of entries in the UniProt database. The gap between TrEMBL (unreviewed automatically annotated sequences) and SwissProt (manually annotated entries) dramatically widens.
Source: uniprot.org/statistics

ally one of the most active fields of biophysics and bioinformatics [23]. On the other hand, starting from thousands of sequences coding for the same kind of protein (across different species or different pathways in the same species), thus sharing the same structure, would be much easier. We could indeed exploit the variability and statistical properties of the ensemble. As challenging as it may be, it is therefore very tempting to apply statistical physics tools to sequence data alone and try to infer information about the proteins.

Fortunately, the 100 million protein sequences are classified into 16306 protein domain families in the publicly available Pfam database [49]. Many families contain about 10^3-10^5 evolutionary related homologous proteins, taking the form of a multiple sequence alignment (MSA), where all amino-acid sequences of the family are aligned to be as similar as possible (*cf.* Fig. 1.2 for a schematic view). A key point is that the structure and functionality of the proteins belonging to the same family are very conserved, whereas the amino-acid sequences are quite diverged, with only 20-30% sequence identity on average [48]. Across evolution, mutations indeed occurred, leading to a lot of variability in protein MSAs. Besides, this variability is not homogeneous: although some positions – or MSA columns – may contain a large variety of amino acids, others will be remarkably conserved. The question is whether the statistical properties of the ensemble of sequences may be used to unveil global informations about the protein family. The Pfam database is updated roughly once a year: the number of protein families is growing slowly, whereas the number of sequences per family is continuously increasing. Creating

MSAs of ever improving quality is indeed a major topic of bioinformatics – the main ideas about sequences alignment will be presented in Chapter 3.

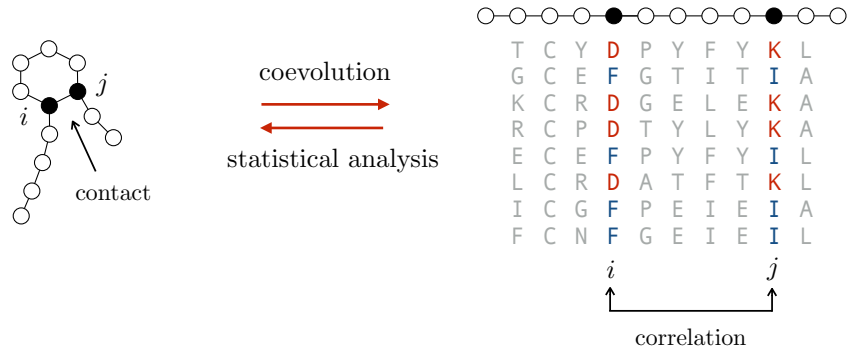


Figure 1.2 – Schematic view of the basic connection between correlation patterns in the MSA and residue-residue contacts. Inspired by [82].

The conservation of structure and function across protein families induces important constraints on the sequence variability. Single-column variability is a first step toward identifying these constraints: a very conserved position indicates residues whose mutations have deleterious effects and disrupt the integrity of the protein. However, compensatory mutations can happen to preserve the protein function, even if single-site mutations are deleterious [54, 72]: if two residues of a protein form a contact, a destabilizing mutation at one position is expected to be compensated by a mutation of the other position over the evolutionary timescale, to maintain the protein structure. A natural idea is therefore to take the reverse path and analyze the statistical correlations induced by coevolution between residues across protein families to infer structural information about proteins (*cf.* Fig. 1.2). Over the last few years, it has prompted a surge of interest among researchers [68, 99, 116], especially in the context of the inverse problem in statistical physics (*cf.* Chapter 2): what is the simplest statistical model for protein sequences capable of reproducing the empirically observed correlations in the MSA?

The difficulty of such an approach mainly lies in disentangling direct (resulting from native contacts in the 3D structure) and indirect (mediated through chains of native contacts) correlations, while dealing with a limited and biased sampling (few and phylogenetically related sequences). A strong correlation between two residues (columns in the MSA) may indeed result from two situations: either residue i is in contact with residue j, or residues i and j are both in contact with residue k (*cf.* Fig. 1.3). The direct measure of statistical correlations has therefore remained of very limited accuracy for unveiling structural constraints [38, 50, 87, 88]. First introduced in 2009 [116], direct-coupling analysis (DCA) is a global statistical inference method based on the maximum-entropy principle (MaxEnt) [61, 62],

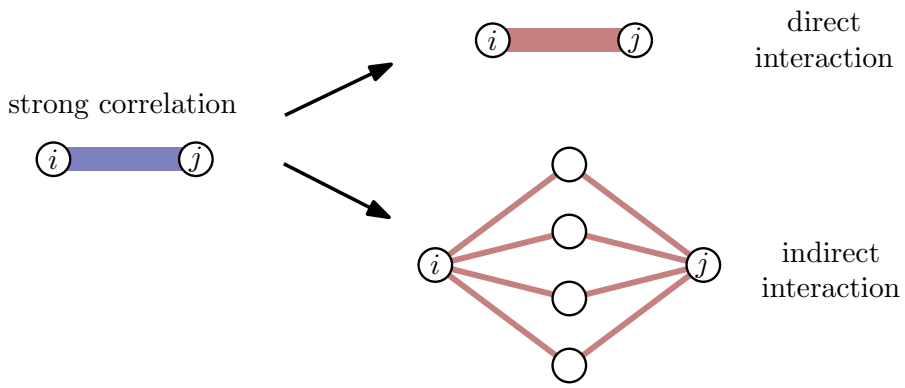


Figure 1.3 – A strong correlation between two residues i and j may indicate either a direct interaction (native contact) between them or an indirect interaction mediated through chains of contacts.

that uses pairwise correlations in amino-acid occurrence from large multiple sequence alignments and has been very successful in the field of protein structural prediction [33, 34, 42, 43, 77, 82] (*cf.* Chapter 3).

Encouraged by this success, the exploitation of coevolution in proteins goes beyond structural prediction, with recent applications to describing fitness effects of mutations [46, 73, 83], or designing artificial proteins with native properties [65]. Experiments actually show that a significant fraction of artificial sequences generated to respect the two-point correlation patterns from the natural MSA acquire the native fold, whereas none of them do fold when reproducing only the single-site frequencies (*cf.* Part II). Very recently, DCA related approaches were found to be able to locate drug resistance regions of the HIV virus [25], and to be very promising in detecting homology in artificial data [60].

2

INVERSE POTTS MODEL

Interpreting patterns of statistical correlations in data is a fundamental problem across scientific disciplines. The goal is to estimate a global probability distribution describing the system from samples of a large number of variables. This model should explain some statistical properties through a network of effective interactions between the variables and may be used to make predictions. The main challenge is to disentangle direct from indirect interactions, through the analysis of correlations between the variables. Usually, the datasets come from experimental measurements and provide a reduced and often biased sample of the possible configurations of the system, increasing the difficulty of this approach, known as *inverse problem* in statistical physics.

Here, we focus on a specific family of statistical models referred to as Potts models [119] – a generalization of the Ising model – which assign a probability distribution $P(\underline{a}|J,h)$ to a configuration (or sequence) $\underline{a} = (a_1, ..., a_N)$ of N variables (or protein residues) taking any value from an alphabet of size q ($q = 21$ for proteins), given the parameters J, h (see Eq. (2.1) for a definition). Inverse Potts models have been applied to various fields, such as patterns of neuron firing activity based on multi-electrode recordings [14, 26, 98, 102], prediction protein 3D structure [57, 78, 82, 110], fitness effect of mutations [25, 45, 46, 76], and gene expression networks [8], all based on the analysis of statistical correlations in experimental data.

However, solving the Potts inverse problem is challenging as the required computational time scales exponentially with the system size, becoming rapidly infeasible for realistic systems. Many approximations have been developed over the recent years to tackle this problem, the most celebrated of which – in the context of protein sequence data – will be presented in this chapter.

2.1 MAXIMUM-ENTROPY MODELING

2.1.1 *Potts model*

Potts models – or pairwise Markov random fields – are a particular family of undirected graphical models. Considering a system of N variables, this model assigns to every configuration $\underline{a} = (a_1, ..., a_N)$ a probability

$$P(\underline{a}|J,h) = \frac{1}{\mathcal{Z}(J,h)} \exp\left(-E(\underline{a}|J,h)\right) = \frac{1}{\mathcal{Z}(J,h)} \exp\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} J_{ij}(a_i,a_j) + \sum_{i=1}^{N} h_i(a_i)\right) . \tag{2.1}$$

Each variable can be found in one of the q possible states, or in other words a_i may take any value from an alphabet of size q; typically $q = 21$ for protein sequences (20 amino acids and 1 alignment gap), or $q = 2$ (Ising case) for neurons (spiking, not spiking). Potts parameters $\{h_i(a)\}_{a=1,\dots,q}$ and $\{J_{ij}(a,b)\}_{a,b=1,\dots,q}$ are respectively local fields on a single variable and direct couplings between pairs of variables. The latter take the form of $q \times q$ matrices, with positive and negative entries. Each entry of these matrices is the coupling between a pair of Potts states a, b; the higher the value of the entry, the more probable it is to find the pair a, b at positions i and j. The coupling matrices will be extensively described in Part III in the context of protein residue interactions.

The energy, or Hamiltonian, of the system

$$E(\underline{a}) = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} J_{ij}(a_i,a_j) - \sum_{i=1}^{N} h_i(a_i) , \tag{2.2}$$

is naturally anti-correlated with the probability P of observing configuration $\underline{a}$, low energies indicating a favorable configuration. The normalization constant $\mathcal{Z}$, or partition function, writes

$$\mathcal{Z}(J,h) = \sum_{\underline{a}} \exp\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} J_{ij}(a_i,a_j) + \sum_{i=1}^{N} h_i(a_i)\right) . \tag{2.3}$$

It will be sometimes more convenient to adopt a slightly different notation using spins $\widehat{\sigma_i} = \{\sigma_{ia}\}_{a=1,\dots,q}$ which are binary q-dimensional vectors given by

$$\sigma_{ia} = \begin{cases} 1 & \text{if} \quad a_i = a \\ 0 & \text{else} \end{cases} . \tag{2.4}$$

Any configuration can therefore be described by a binary vector $\underline{\sigma} = (\widehat{\sigma_1}, \dots, \widehat{\sigma_N})$ composed of N blocks of size q. With these notations, Eq. (2.1) now writes

$$P(\underline{\sigma}|J,h) = \frac{1}{\mathcal{Z}(J,h)} \exp\left(\sum_{\substack{i,j \\ i<j}}^{N}\sum_{a,b=1}^{q} \sigma_{ia} J_{ij}(a,b) \sigma_{jb} + \sum_{i=1}^{N}\sum_{a=1}^{q} h_i(a)\sigma_{ia}\right) . \tag{2.5}$$

Interestingly, Potts models naturally arise in the context of the maximum-entropy principle (MaxEnt).

2.1.2 *Maximum-entropy principle*

As explained in the introduction of this chapter, we wish to learn the joint probability distribution P of N random variables, given some realizations of these variables. Supposing that the B samples $\underline{\sigma}^{(1)}, ..., \underline{\sigma}^{(B)}$ we have access to are independent and identically distributed (i.i.d), the probability distribution should be *coherent with the data*. In other words, the empirical frequencies $f_i(a)$ and correlations $f_{ij}(a,b)$ from the data

$$\begin{aligned} f_i(a) &:= \frac{1}{B}\sum_{\tau=1}^{B} \sigma_{ia}^{(\tau)} \,, \\ f_{ij}(a,b) &:= \frac{1}{B}\sum_{\tau=1}^{B} \sigma_{ia}^{(\tau)}\sigma_{jb}^{(\tau)} \,, \end{aligned} \tag{2.6}$$

should be matched by its one- and two-point marginals

$$\sum_{\underline{\sigma}} \sigma_{ia} P(\underline{\sigma}) = f_i(a) \qquad \text{and} \qquad \sum_{\underline{\sigma}} \sigma_{ia}\sigma_{jb} P(\underline{\sigma}) = f_{ij}(a,b) \,. \tag{2.7}$$

However, these conditions can be satisfied by an infinite number of probability distributions. We are looking for the *least constrained* of them, or namely, the one with the maximum (Shannon) entropy [61, 62]:

$$S[P] := -\sum_{\underline{\sigma}} P(\underline{\sigma}) \log P(\underline{\sigma}) \,. \tag{2.8}$$

As in any optimization problem, constraints (2.7) are enforced by Lagrange multipliers $h_i(a)$, $J_{ij}(a,b)$:

$$\begin{aligned} S = &-\sum_{\underline{\sigma}} P(\underline{\sigma}) \log P(\underline{\sigma}) + \sum_{i<j}\sum_{a,b} J_{ij}(a,b)\left(\sum_{\underline{\sigma}} \sigma_{ia}\sigma_{jb}P(\underline{\sigma}) - f_{ij}(a,b)\right) \\ &+ \sum_{i}\sum_{a} h_i(a)\left(\sum_{\underline{\sigma}} \sigma_{ia}P(\underline{\sigma}) - f_i(a)\right) + \lambda\left(\sum_{\underline{\sigma}} P(\underline{\sigma}) - 1\right) \,, \end{aligned} \tag{2.9}$$

the last term guaranteeing the normalization of P. By differentiating Eq. (2.9) for an arbitrary $\underline{\sigma}$, we get:

$$\begin{aligned} \frac{\partial S}{\partial P(\underline{\sigma})} = 0 = &-1 - \log P(\underline{\sigma}) + \sum_{i<j}\sum_{a,b} \sigma_{ia} J_{ij}(a,b)\sigma_{jb} \\ &+ \sum_{i}\sum_{a} h_i(a)\sigma_{ia} + \lambda \,. \end{aligned} \tag{2.10}$$

with $\mathcal{Z} = \exp(1-\lambda)$

P therefore takes the form of a Boltzmann distribution with the Hamiltonian of a Potts model defined at Eq. (2.5).

If the variety of biological effects involved in protein evolution could certainly not be reduced to pairwise interactions, the importance of higher-order terms is not clear, but frequently requires even

more samples [21]. Note that MaxEnt only gives the general form of the probability distribution. The Potts parameters still have to be inferred from the data, as explained in the next section. Some of the existing methods – such as pseudolikelihood [7, 94] – actually require the whole knowledge of the samples, not only frequencies and correlations. Other methods which only need frequencies and correlations rely on approximations – mean-field [67] or variational approximations [115] – leading to statistically inconsistent estimators [1]. For these reasons, MaxEnt has been sometimes criticized [6].

2.2 APPROXIMATIONS TO THE INVERSE PROBLEM

Formally, the inverse Potts problem is solved by the set of fields and couplings that maximize the average log-likelihood over the data D, in the so-called *maximum-likelihood* approach. Denoting the Potts parameters $\mathbf{J} = \{J_{ij}(a,b), h_i(a)\}$, the average log-likelihood $\mathcal{L}$ writes

$$\begin{aligned}\mathcal{L}(\mathbf{J}|D) =& \frac{1}{B}\sum_{\tau=1}^{B} \log P(\underline{\sigma}^{(\tau)}) \\ =& \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{a,b=1}^{q} J_{ij}(a,b) f_{ij}(a,b) \\ &+ \sum_{i=1}^{N}\sum_{a=1}^{q} h_i(a) f_i(a) - \log \mathcal{Z}(\mathbf{J}) \ ,\end{aligned} \tag{2.11}$$

where B is the number of samples in the data (*e.g.* the number of sequences in a MSA).

Alternatively, the cross-entropy between the data and the model $S \equiv -\mathcal{L}(\mathbf{J}|D)$ can be written as the sum of the entropy of the data and the Kullback-Leibler (KL) divergence [2] of the model with respect to the data [105]. Defining the empirical measure over the observed configurations through

$$P_{obs}(\underline{\sigma}) = \frac{1}{B}\sum_{\tau=1}^{B} \delta_{\underline{\sigma},\underline{\sigma}^{(\tau)}} \ , \tag{2.12}$$

with δ the Kronecker delta function, the cross-entropy indeed rewrites

$$S = -\sum_{\underline{\sigma}} P_{obs}(\underline{\sigma}) \log P_{obs}(\underline{\sigma}) + D(P_{obs}||P) \ . \tag{2.13}$$

1. The exact parameters cannot be recovered even in the limit of an infinitely large number of samples drawn from the Potts model.

2. The Kullback-Leibler divergence is a measure of the difference between probability distributions. It can be thought of a distance between probability distributions P and Q, except that it is not symmetric under the exchange of P and Q. It is always non-negative and equals 0 if and only if $P = Q$.

Hence, maximizing the log-likelihood – or minimizing the cross-entropy – over the parameters $\mathbf{J} = \{J, h\}$ ensures that the "best" (in the sense of the KL divergence) Potts measure is found.

The log-likelihood is indeed concave, as can be easily shown considering its Hessian, that is the Fisher information matrix up to a sign (see Part IV of this dissertation for more details). The log-likelihood (resp. cross-entropy) therefore has a maximum (resp. minimum), guaranteeing that the maximum-likelihood approach has a solution.

Note however that the computation of $\mathcal{Z}$ required in Eq. (2.11) involves a summation over the q^N possible configurations. In the case of protein sequences with $q = 21$ and $N = 50 - 500$ residues typically, there are $10^{65} - 10^{650}$ possible configurations, making any exact computation of $\mathcal{Z}$ impossible. Many approximated methods have been proposed, and four of them will be presented below.

Direct-coupling analysis (DCA) is a generic name for approximating the inverse problem in the maximum-entropy approach, in the specific case of biological sequence data. As mentioned at the beginning of this chapter, it has been successfully applied for the inference of protein and RNA residue contacts, protein-protein interaction networks, and fitness landscape. The most used methods in this context are the mean-field and the pseudolikelihood approximations.

2.2.1 *Boltzmann machine learning*

The inverse problem can be tackled with the Boltzmann machine learning (BML) approach developed in the 1980's [1], avoiding the computation of $\mathcal{Z}$. Given an input set of fields and couplings, the model frequencies and correlations f_i^{MC} and f_{ij}^{MC} are computed through Monte Carlo (MC) simulations. The Potts parameters are then updated according to the gradient of the log-likelihood [96], until the model correlations match the imposed values (2.7):

$$\begin{aligned} h_i(a) &\to h_i(a) + \left(f_i^{MC}(a) - f_i(a)\right) \eta_i(a) \,, \\ J_{ij}(a,b) &\to J_{ij}(a,b) + \left(f_{ij}^{MC}(a,b) - f_{ij}(a,b)\right) \eta_{ij}(a,b) \,, \end{aligned} \tag{2.14}$$

where $\{\eta_i, \eta_{ij}\}$ are parameter-specific weight factors, also updated at each iteration.

Eq. (2.14) can be seen as the minimization of the KL divergence between the MC equilibrium distribution and the empirical measure over the observed configurations. Given the convexity of the optimization problem, this gradient ascent is supposed to converge to the exact solution. However, thermalization is needed to estimate the model correlations. Each MC step requires huge computational efforts to estimate the change in energy due to a change in the configurations, which may be prohibitive for large system sizes. More

over, this data-driven approach leads to overfitting in the case of poor sampling.

Besides, the number of updates can be extremely large without a good initial guess for the Potts parameters, rendering the algorithm very slow to converge. This starting point can however be provided by other faster inference methods, as will be illustrated in Part IV of this dissertation.

2.2.2 *Mean-field approximation*

The mean-field approximation allows for a computation of $\mathcal{Z}$ in polynomial time and the equations in the Ising case ($\sigma_i = \pm$) read [89, 95]

$$\tanh^{-1} m_i = h_i + \sum_j J_{ij} m_j \ , \tag{2.15}$$

with m_i the magnetization at site i. Connected correlations $C_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$ can be obtained from the linear response [67, 117]:

$$C_{ij} = \frac{\partial m_i}{\partial h_j} \ , \qquad (C^{-1})_{ij} = \frac{\partial h_i}{\partial m_j} \ . \tag{2.16}$$

An equation involving the connected-correlation matrix and the couplings can therefore be derived:

$$J_{ij} = -(C^{-1})_{ij} \ . \tag{2.17}$$

In the context of protein sequences, mean-field direct-coupling analysis (mfDCA) – based on the naive mean-field inversion – was the first efficient method to infer the Potts parameters given a MSA. It is based on the high temperature (small couplings) expansion of the Legendre transform of the free energy [53, 92], generalized to the Potts case. The full derivation can be found in the supplementary material of [82].

The proper generalization of Eq. (2.15) to the Potts case reads

$$P_i(a) = \frac{1}{z_i} \exp\left(h_i(a) + \sum_{jb} J_{ij}(a,b) P_j(b) \right) \ , \tag{2.18}$$

where $P_i(a) = \sum_{\underline{\sigma}} \sigma_{ia} P(\underline{\sigma})$ is the marginal of the probability distribution P, and $z_i = \sum_a \exp\left(h_i(a) + \sum_{jb} J_{ij}(a,b) P_j(b) \right)$, a normalization constant. Using the linear response, the inferred Potts couplings therefore read

this formula is exact in the "q-gauge" (cf. Section 2.3.1)

$$J_{ij}(a,b) = -(C^{-1})_{ij}(a,b) \ , \tag{2.19}$$

with the connected-correlation matrix $C_{ij}(a,b) = f_{ij}(a,b) - f_i(a) f_j(b)$ depending on the empirical values $f_i(a)$ and $f_{ij}(a,b)$ computed from the dataset. Practically, the inverse problem can therefore be solved in

one single step, by calculating and inverting the connected-correlation matrix directly from the data. The complexity of the procedure is thus of $\mathcal{O}(q^3N^3)$. The idea of inverting the connected-correlation matrix also arises in the Gaussian analogy of mfDCA, the Protein Sparse Inverse Covariance Estimation (PSICOV) approach [64].

Notice that the connected-correlation matrix always displays N zero modes due to the identity

$$\sum_b C_{ij}(a,b) = \sum_b f_{ij}(a,b) - f_i(a) = 0\,, \tag{2.20}$$

and is therefore not invertible. This problem is related to the over-parametrization of the system and the zero modes can be removed by fixing the Potts gauge (*cf.* Section 2.3.1). However, in case of insufficient data availability, the connected-correlation matrix may still not be invertible (even after fixing the gauge). Some Potts states (amino acids) indeed may never be observed in the data and the matrix may not be of full rank. The empirical frequencies and correlations need to be adjusted with a regularization variable, as will be discussed in Section 2.3.2.

2.2.3 *Pseudolikelihood maximization*

First implemented on the inverse Ising case [94], the pseudolikelihood method is now the most used tool in the field of protein structure prediction. The pseudolikelihood approximation of the direct-coupling analysis (plmDCA) [41, 42] (or equivalently GREMLIN [9, 66]) has been shown to outperform any other existing method in this specific context. Avoiding the complete computation of $\mathcal{Z}$, like BML, the pseudolikelihood related methods however require the complete knowledge of the configurations – not only the frequencies and correlations. The runtime complexity of the pseudolikelihood approaches is of $\mathcal{O}(Bq^2N^2)$. Note the linear dependence in B, the number of samples (or homologous sequences in the MSA).

In the following, we will use the notations of Eqs. (2.1 - 2.3), with the variables $a_i \in \{1,...,q\}$. plmDCA substitute the probability P in the log-likelihood (Eq. (2.11)) by the conditional probability of observing one variable a_r in the configuration $\underline{a}^{(\tau)} = (a_1^\tau,...,a_N^\tau)$ given observation of all other variables $a_{\setminus r}^\tau = (a_1^\tau,...,a_{r-1}^\tau,a_{r+1}^\tau,...,a_N^\tau)$:

$$P(a_r = a_r^\tau | a_{\setminus r}^\tau) = \frac{\exp\left(h_r(a_r^\tau) + \sum_{i\neq r} J_{ri}(a_r^\tau, a_i^\tau)\right)}{\sum_{l=1}^q \exp\left(h_r(l) + \sum_{i\neq r} J_{ri}(l, a_i^\tau)\right)}\,. \tag{2.21}$$

The parameters $\mathbf{h}_r$ and $\mathbf{J}_r = \{J_{ri}\}_{i\neq r}$ can be computed via the maximization of the pseudo log-likelihood at site r:

$$\mathcal{PL}_r(\mathbf{h}_r, \mathbf{J}_r) = \frac{1}{B}\sum_{\tau=1}^{B} \log P_{\{\mathbf{h}_r,\mathbf{J}_r\}}(a_r = a_r^\tau | a_{\setminus r}^\tau)\,. \tag{2.22}$$

The total pseudo log-likelihood then writes

$$\mathcal{L}_{\text{pseudo}}(\mathbf{J}|D) = \sum_{r=1}^{N} \mathcal{PL}_r(\mathbf{h}_r, \mathbf{J}_r) \,. \tag{2.23}$$

Contrary to mfDCA, this procedure is statistically consistent, *i.e.* it guarantees to extract the exact parameter values in the limit of an infinitely large sample drawn from the Potts model. However, this consideration might not be relevant in the case of real biological data, which are of course not extracted from Potts models.

Besides, for a finite sample, this method returns two different values for the couplings J_{ri}: $J^{\star,i}_{ri}$ and $J^{\star,r}_{ir}$ obtained from the maximization of $\mathcal{PL}_i$ and $\mathcal{PL}_r$ respectively. One simple way to reconcile these values is to replace them by the average: $J_{ri} = \frac{1}{2}\left(J^{\star,i}_{ri} + J^{\star,r}_{ir}\right)$. This approach is referred to as *asymmetric pseudolikelihood maximization* [41], and will be used in this dissertation.

A prior probability distribution (typically Gaussian) can be considered for the model parameters, which discounts large values resulting from insufficient statistics in the original data. It takes the form of a regularization term added to the objective function, described in Section 2.3.2.

2.2.4 *Adaptive cluster expansion*

Another method to accurately estimate the partition function lies in cluster expansions. Widely used in statistical mechanics [55, 90], such expansions are limited by the system size or only consider fixed cluster sizes, and do not tackle overfitting issues. The adaptive cluster expansion (ACE), first developed in the Ising case [27, 28], proposes a method adapted to the specificity of the data, fully accounting for the complex patterns of statistical correlations present in experimental samples. It has been successfully applied in the Ising case ($q = 2$) to real data with as many as several hundred variables, including studies of neural activity [14, 113], or human immunodeficiency virus (HIV) fitness based on protein MSA data [11, 76].

Practically, ACE builds global solutions from local ones. Let us consider the susceptibility matrix and its inverse:

$$\chi = \left.\frac{\partial \mathbf{p}}{\partial \mathbf{J}}\right|_{\mathbf{J}} , \qquad \chi^{-1} = \left.\frac{\partial \mathbf{J}}{\partial \mathbf{p}}\right|_{\mathbf{p}} , \tag{2.24}$$

with the Potts parameters $\mathbf{J} = \{J_{ij}(a,b), h_i(a)\}$ and the one- and two-site correlations from the MaxEnt model $\mathbf{p} = \{f_{ij}(a,b), f_i(a)\}$. χ describes the *direct problem* and how the correlations respond to a small variation in the Potts parameters. On the other hand, χ^{-1} is a naturally associated to the *inverse problem*, and measures the response of the inferred parameters to a small change in the correlations.

These two matrices are crucially different, in the sense that χ^{-1} is much sparser and shorter range than χ (for more details, see [27]). It means that even if the system is described by strong long-range correlations, the Potts parameters may only depend on a small (compared to the system size) number of correlations. This property is essential as it ensures that the inverse problem is actually solvable and meaningful.

Given these considerations, ACE proposes to accurately estimate the parameters by building a sparse network of interactions. The regularized cross-entropy or negative log-likelihood

$$\begin{aligned} S = \log \mathcal{Z} - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{a,b=1}^{q} J_{ij}(a,b) f_{ij}(a,b) \\ + \sum_{i=1}^{N} \sum_{a=1}^{q} h_i(a) f_i(a) - \frac{1}{B} \log P_0(\mathbf{J}) \ , \end{aligned} \tag{2.25}$$

where P_0 is a prior distribution for the parameters typically Gaussian (*cf.* Section 2.3.2), is decomposed into sum of contributions from clusters of variables $\Gamma = \{i_1, ..., i_k\}$, $k \leqslant N$:

$$S = \sum_{\Gamma} \Delta S_\Gamma \ , \qquad \Delta S_\Gamma = S_\Gamma - \sum_{\Gamma' \subset \Gamma} \Delta S_{\Gamma'} \ , \tag{2.26}$$

where the summation is over all possible subsets of the N variables. The cluster entropy ΔS_Γ is recursively defined as the remaining contribution once all contributions from smaller clusters have been removed. S_Γ is the minimum of Eq. (2.25) restricted to the variables in Γ, thus depending only on $f_{ij}(a,b)$, $f_i(a)$, for $i,j \in \Gamma$. S_Γ is tractable if the clusters are small. ΔS_Γ is the contribution to the cross-entropy from the cluster Γ which is not captured by any subset of Γ.

The key of this approach is to approximate the cross-entropy – and therefore the Potts parameters which minimize it – by truncating the sum in Eq. (2.26) to a restricted set of clusters Γ contributing most to the cross-entropy. The convergence of Eq. (2.26) is therefore made faster, especially since contributions for overlapping clusters sharing the same interaction subgraph partially compensate, as shown in [27, 28]. Moreover, by neglecting clusters which contribute less to the cross-entropy (poorly sampled Potts states), overfitting can be reduced. Note that, by construction, the summation over all possible clusters would give the exact value of the cross-entropy.

The algorithm is quickly described below. For more details, see the pseudocode in [28]. A threshold t is defined on the cross-entropy to separate the significant clusters from negligible ones. Initially large, this threshold is progressively lowered (outer loop) until enough clusters are included to yield an inferred model fitting the imposed correlations in Eq. (2.7), within the statistical error due to finite sampling (see Part IV of this dissertation for more details). On the other hand,

an inner loop iteratively constructs the set of clusters Γ with contributions to the cross entropy $|\Delta S_\Gamma| > t$, giving rise to an approximation of the cross-entropy and the Potts parameters at threshold t.

We give here a description of the inner loop, based on [13] co-authored by the author of this dissertation. Given a list L_k of clusters of size k, beginning with $k = 2$,

1. For each cluster $\Gamma \in L_k$
   a) Compute S_Γ by numerical minimization of Eq. (2.25) restricted to Γ.
   b) Record the parameters minimizing Eq. (2.25), called $\mathbf{J}_\Gamma$.
   c) Compute ΔS_Γ using Eq. (2.26).
2. Add all clusters $\Gamma \in L_k$ with $|\Delta S_\Gamma| > t$ to a new list $L'_k(t)$.
3. Construct a list L_{k+1} of clusters of size $k+1$ from overlapping clusters in $L'_k(t)$.

After the summation of clusters terminates, the approximate value of the Potts parameters – minimizing the cross-entropy given the current value of t – is computed by

$$\mathbf{J}(t) = \sum_k \sum_{\Gamma \in L'_k(t)} \Delta \mathbf{J}_\Gamma \,, \qquad \Delta \mathbf{J}_\Gamma = \mathbf{J}_\Gamma - \sum_{\Gamma' \subset \Gamma} \Delta \mathbf{J}_{\Gamma'} \,. \tag{2.27}$$

Note that this formula generally yields sparse solutions because nonzero couplings are only included if some clusters containing them have been selected. In this algorithm the dominant contribution to the computational complexity often comes from the evaluation of the partition function $\mathcal{Z}$ for large cluster sizes k, which requires $\mathcal{O}(q^k)$ operations to compute. Note that this is much smaller than for the exact computation which would be of $\mathcal{O}(q^N)$, ensuring reasonable execution time of the algorithm.

ACE adapts the complexity of the inferred Potts model to the level of the sampling in the data, reducing overfitting. This is achieved first by a sparse inference procedure that omits interactions that are unnecessary for reproducing the statistics of the data to within the error bounds due to finite sampling. On the other hand, less frequently observed Potts states are regrouped into a unique state according to a threshold on entropy or frequency, as will be extensively discussed in Part IV. Initially developed in the Ising case, this procedure has been adapted to the Potts case and will be illustrated to both real and artificial data sets, also in Part IV.

2.3 MODEL PARAMETERS

Besides the chosen approximate method to solve the inverse problem, several parameters need yet to be fixed before any application to experimental data. Among them are the gauge invariance due to

the over-parametrization of the problem, regularization accounting for finite-sample effects, and reweighting dealing specifically with biased samples.

2.3.1 *Gauge invariance*

The Nq frequencies $f_i(a)$ and $\frac{1}{2}N(N-1)q^2$ correlations $f_{ij}(a,b)$ $(i<j)$, estimated from the data are not independent. The former sum up to 1, and the latter have the frequencies as marginals. Therefore not all constrains in Eq. (2.7) are independent: the total number of non redundant parameters is actually $\frac{1}{2}N(N-1)(q-1)^2+N(q-1)$. This number is smaller than the total number $Nq+\frac{1}{2}N(N-1)q^2$ of Potts parameters $h_i(a)$ and $J_{ij}(a,b)$. The model is therefore over-parametrized, a fact referred to as *gauge invariance* in physics language. We can reparametrize the model without changing probabilities[3] using an arbitrary $K_{ij}(a), 1\leqslant i,j\leqslant N, a\in\{1,...,q\}$:

$$\begin{aligned} J_{ij}(a,b) &\to J_{ij}(a,b)+K_{ij}(a)+K_{ji}(b)\ ,\\ h_i(a) &\to h_i(a)+\sum_{j(j\neq i)} K_{ij}(a)\ . \end{aligned} \tag{2.28}$$

The inferred fields and couplings can be expressed in the so-called "zero-sum gauge", in which

$$\sum_{c=1}^{q} J_{ij}(a,c)=\sum_{c=1}^{q} J_{ij}(c,a)=\sum_{c=1}^{q} h_i(c)=0\ , \tag{2.29}$$

for all states a and all variables i,j. In practice, the couplings $J_{ij}(a,b)$ can be simply put in the zero-sum gauge through

$$\begin{aligned} J_{ij}(a,b) &\to J_{ij}(a,b)-J_{ij}(\cdot,b)-J_{ij}(a,\cdot)+J_{ij}(\cdot,\cdot)\ ,\\ h_i(a) &\to h_i(a)-\sum_j J_{ij}(a,\cdot)\ , \end{aligned} \tag{2.30}$$

where $g(\cdot)$ denotes the uniform average of $g(a)$ over all states a at fixed position. The zero-sum gauge minimizes the Frobenius norm of the coupling matrices, which is used as a scalar measure of the coupling strength. It allows for the ranking of residue pairs (i,j) in order to predict residue-residue contacts [29, 42, 116].

cf. Chapter 3

Alternatively, a gauge state c_i per variable can be chosen such that

$$J_{ij}(a,c_j)=J_{ij}(c_i,b)=h_i(c_i)=0\ , \tag{2.31}$$

for all states a,b and variables i,j. The couplings and fields are transformed as follows:

$$\begin{aligned} J_{ij}(a,b) &\to J_{ij}(a,b)-J_{ij}(c_i,b)-J_{ij}(a,c_j)+J_{ij}(c_i,c_j)\ ,\\ h_i(a) &\to h_i(c_i)-\sum_{j\neq i}\left(J_{ij}(a,c_j)-J_{ij}(c_i,c_j)\right)\ . \end{aligned} \tag{2.32}$$

3. Although the gauge transformation conserves the probability, it modifies entirely the Potts parameters (*i.e.* the couplings and fields)

Usually, c_i is chosen as the most frequent observed state for the variable i, this specific gauge being referred to as "consensus gauge". Besides, mfDCA typically uses the "q-gauge", where $c_i = q$ for all sites i.

2.3.2 *Data preprocessing for finite-sample effects*

2.3.2.1 *Regularization*

Experimental data are often not i.i.d; they form a finite and usually small-size sample. For instance, a Potts model describing a protein family with sequences of $50-500$ amino acids requires *ca.* 10^6 to 10^8 parameters. Few protein families are large enough to directly determine these parameters, and regularization is essential to avoid overfitting. Moreover, adding a regularization term helps the hill-climbing optimization (in plmDCA or ACE) to rapidly find the maximum of the (pseudo) likelihood, or alternatively guarantees the inversion of the connected-correlation matrix (in mfDCA). Different regularization schemes and their effects have been extensively addressed in [12].

A prior probability distribution (typically Gaussian) is considered for the model parameters, yielding to a penalty term in the objective function. The following l_2-penalty is therefore added to the log-likelihood of the data:

$$\gamma \sum_{i=1}^{N} \sum_{a=1}^{q} h_i(a)^2 + \gamma \sum_{i<j}^{N} \sum_{a,b=1}^{q} J_{ij}(a,b)^2 \,. \tag{2.33}$$

For $\gamma \sim 1/B$, this factor can be thought of as a weakly informative prior [52], whose main purpose is to ensure that solutions of the inverse problem are not infinite due to issues of undersampling (*e.g.* parameters corresponding to a state that is never observed). For plmDCA, the standard value of the regularization parameter is $\gamma = 10^{-2}$ as it gives optimal results for contact prediction [42].

Other forms of regularization are also possible, such as l_1 introduced in [94] for the inverse Ising problem, which forces a fraction of the Potts parameters to be set to 0 and effectively reduces the number of parameters. Since we are interested, in context of contact prediction, in the accuracy of the strongest couplings (which will be ranked, as explained in Chapter 3), l_1 penalty might be less appropriate than l_2, as it makes no difference for the ranking whether the weakest couplings are small or precisely set to 0 [42]. Note that these forms of regularizations are not invariant under gauge transformations. Thus, the results of the inference including the regularization do have some dependence on the gauge choice.

Alternatively, in the context of mfDCA, the connected-correlation matrix may not be of full rank, as some states may never be observed in finite-size samples. To ensure its invertibility, empirical frequencies

and correlations (defined at Eq. (2.6)) are adjusted with a regularization variable λ, referred to as "pseudocount" and introduced in [82]:

$$\begin{aligned} f_i(a) &= \frac{1}{\lambda + B}\left(\frac{\lambda}{q} + \sum_{\tau=1}^{B} \sigma_{ia}^{\tau}\right) , \\ f_{ij}(a,b) &= \frac{1}{\lambda + B}\left(\frac{\lambda}{q^2} + \sum_{\tau=1}^{B} \sigma_{ia}^{\tau}\sigma_{jb}^{\tau}\right) . \end{aligned} \tag{2.34}$$

This is equivalent to adding λ random samples to the data. It was observed in [82], that optimal results for contact prediction in mfDCA are obtained with a fairly large pseudocount parameter $\lambda \sim B$. .

2.3.2.2 *Reweighting*

In the context of protein sequences, phylogenetic relations between proteins and human selection of the sequenced species yield strong sampling biases. This issue has been the object of previous studies [24, 38, 114, 118], but a simple sampling correction can be implemented by counting sequences with more than 80% identity and reweighting them in the frequency counts [82]. A weight w_τ is associated to each sequence $\underline{a}^\tau$, reflecting their importance in the sampling. Sequences too similar to other sequences are attributed a lower weight, whereas isolated sequences contribute with a higher weight to the sampling.

The weight is defined as the inverse number of sequences within Hamming distance $d_H < xN$, with $x \in [0,1]$:

$$w_\tau = \frac{1}{|\{b | 1 \leqslant b \leqslant B; d_H(\underline{a}^{(b)}, \underline{a}^{(\tau)}) \leqslant xN\}|} , \tag{2.35}$$

with $\tau = 1, ..., B$. The value $x \sim 0.2$ was found to be optimal across many protein families [82]. The number of non-redundant sequences is measured as the effective sequence number after reweighting:

$$B_{eff} = \sum_{\tau=1}^{B} w_\tau . \tag{2.36}$$

As a rule of thumb, B_{eff} should be at least 300 to ensure good results in the context of inference on protein sequences.

3

APPLICATION TO BIOLOGICAL DATA

3.1 PROTEIN FAMILIES

3.1.1 *Basic notions*

Some very basic concepts about proteins will be presented in this section, mainly based on [2]

Proteins are long polymer chains consisting of monomeric blocks that are the same for all living cells: the amino acids. Many different protein molecules are present in each cell – forming most of its mass, excluding the water – and they are involved in most of the biological functions within living organisms, such as acting as enzymes to catalyze chemical reactions, maintaining structure, generating movements, responding to stimuli, etc. [2]. There are 20 amino acids, each with a distinctive chemical character given by a specific side group attached to a common core structure. Amino acids therefore display a large variety of physico-chemical properties [19], such as charge, size, acidity, polarity, hydrophobicity (*cf.* Fig. 3.1), which play a central role in determining the shape of the protein.

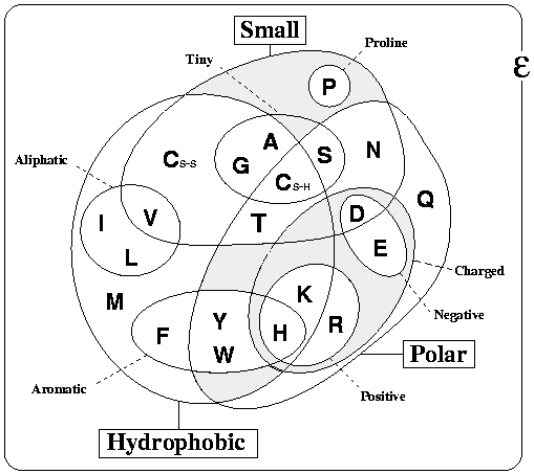


Figure 3.1 – The 20 amino acids display a large variety of physico-chemical properties.
Source: [71]

Each protein molecule defined by its sequence of amino acids – covalently linked by peptide bonds – folds into a specific three-dimensional structure, unique[1] to each type of protein (within some flexibility). The structure can be described on four distinct levels:

1. allosteric proteins may have more than one structural conformation.

— primary structure: linear sequence of amino acid along the polypeptide backbone;
— secondary structure: folding patterns resulting from hydrogen bonds in the backbone, mainly α-helices and β-sheets (shorthand symbols helices and arrows in ribbon drawings of proteins, *cf.* Fig. 3.2);
— tertiary structure: three-dimensional conformation of the protein – α-helices and β-sheets folded into a compact structure – characterized by long range non-convalent interactions, such as hydrogen bonds or disulfide bonds;
— quaternary structure: several polypeptide chains bounded together, forming a multi-subunit protein (called dimer with two subunits, *cf.* Fig. 3.2).

Beyond the four levels of organizations, the protein domain is a unit of major importance. This substructure designates any part of the polypeptide chain folding independently into a stable structure. Typical domain lengths range from 40 to 350 amino acids, also called residues [2].

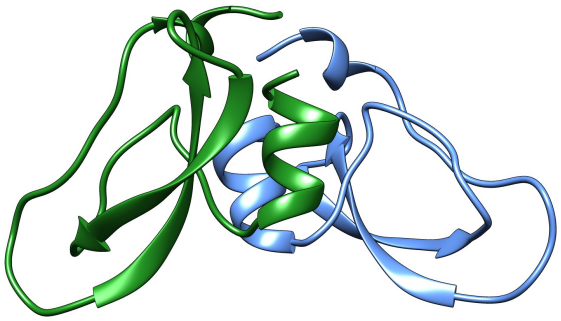


Figure 3.2 – Dimeric assembly of PDB entry 1bhc, corresponding to protein BPT1_BOVIN (residues 39-91) from the Kunitz/Bovine pancreatic trypsin inhibitor domain (Pfam id: PF00014).
Source: [81], ebi.ac.uk/pdbe, rbvi.ucsf.edu/chimera [91].

3.1.2 *Multiple sequence alignments*

3.1.2.1 *Families*

The 100 million known protein sequences are grouped into domain families composed of homologous proteins – *i.e.* of common evolutionary origin – displaying similar 3D structures and functions. The Pfam database [48] lists 16306 different domain families, mainly composed of $10^2 - 10^5$ sequences forming a MSA, where all amino-acid sequences of the family are aligned to be as similar as possible. In the course of evolution, mutations occurred – substitutions, insertions or

deletions – leading to a lot of diversity in the MSA (displaying only about 20%-30% sequence identity). Therefore, it often happens that a residue of a domain cannot be aligned perfectly with the other members of the family and a "gap" symbol is introduced. On the other hand, a mutation with a deleterious effect – that alters 3D structure of the protein – would have caused the denatured protein to be eliminated by natural selection. In the MSA, not every mutation is therefore possible and the conservation of the structure imposes strong constraints on the sequence variability.

Practically, a MSA is a matrix a_i^τ, which lines $\tau = 1, ..., B$ are amino-acid sequences considered to be different versions of the same protein, and which columns $i = 1, ..., N$ are the aligned protein residues. Therefore a_i^τ is either one of the 20 amino acids, or the gap symbol. These symbols are then converted to numbers from 1 to 21 for statistical analysis purposes. As mentioned in Chapter 1, some MSA columns will be highly conserved whereas other will be more variable, depending on the constraints imposed by the conservation of the 3D structure. A sub-alignment of the MSA for the Kunitz/Bovine pancreatic trypsin inhibitor domain (Pfam id: PF00014) is shown on Fig. 3.3. The alignment indeed displays non-trivial statistical patterns, which are analyzed by DCA related methods. In particular, gaps come in long stretches contrary to amino acids, resulting from affine penalties in the alignment procedure (see next section).

```
B3N358_DROER/24-81      --------GACYAYFPLFSYYPESNSCELFIYGGCWGNANRFHSKESCEEKCL
Q86QT1_BOMMO/32-85      -CEQAFDAGLCFGYMKLYSYNQETKNCEEFIYGGCQGNDNRFSTLAECEQKCI
ISC3_BOMMO/9-62         -CEQAFDAGLCFGYMKLYSYNQETKNCEEFIYGGCQGNDNRFSTLAECEQKCI
ISC2_BOMMO/8-61         -CEQAFNSGPCFAYIKLYSYNQKTKKCEEFIYGGCKGNDNRFDTLAECEQKCI
Q967V8_BOMMO/31-84      -CEQAFNSGPCFAYIKLYSYNQKTKKCEEFIYGGCKGNDNRFDTLAECEQKCI
ISC1_BOMMO/8-61         -CEQAFNSGPCFAYIKLYSYNQKTKKCEEFIYGGCQGNDNRFITLAECEQKCI
Q5MBP2_BOMMO/31-84      -CEQAFDAGPRDAYIKLYSYNQETKKCEEFIYGGCLGNDNRFNTLAECEQKCI
Q8WPI5_BOMMO/31-84      -CEQAFDVGPCGAYFKLYSYNQETKKCEEFIYGGCQGNDNRFNTLAECEQKCI
B4G9D5_DROPE/24-79      --------GACLAYIPSWSYN--GRTCEEFIYGGCGGNDNRFNSQAECEAKCL
B5DIF0_DROPS/24-79      --------GACLAYIPSWSYN--GRACEEFIYGGCGGNDNRFNSQAECEAKCL
```

Figure 3.3 – Sub-alignment of 10 sequences from the multiple sequence alignment of PF00014, on the left are protein names and domain coordinates in the full length sequences. Non trivial statistical features arise, in particular for gaps. Red highlights a totally conserved residue, whereas green shows a much more variable residue.

As an example, Pfam domain family PF00014 contains 7005 sequences of length $N = 53$ residues from 176 different species. 253 different structures are available, but with a lot of redundancy – sometimes crystallized several times – and overall very similar. All structures are experimentally determined (X-ray crystallography, NMR spectroscopy) by biologists around the world and classified in the freely accessible Protein Data Bank (PDB) [17, 18]. Besides, protein domains are (mainly) manually classified in the structural classification of proteins (SCOP) database [84], based on the similarity of their struc-

tures. The types of folds are grouped into "classes", the top level of the hierarchical classification, such as "all α-proteins", " membrane and cell surface proteins and peptides", or "small proteins".

3.1.2.2 *Profile hidden Markov models*

mainly adapted from [39] & [32]

Forming multiple sequence alignments and assigning a given sequence to a domain family is closely related to predicting its structure and function. In this section, we will shortly present of one of the most powerful tool in bioinformatics: the profile hidden Markov model (HMM) [39, 40]. It is widely used in the fields of MSA building, homology detection, and structural modeling. The HMMer software [47] is used to build the alignments in the Pfam database.

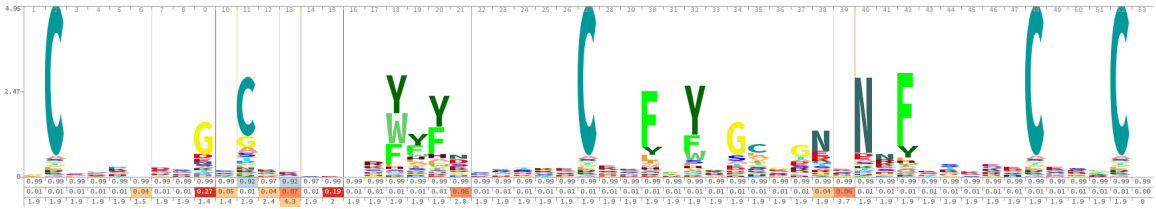


Figure 3.4 – HMM logo of PF00014, providing a graphical representation of the conservation in the MSA as well as the profile-HMM. Each amino acid letter scales according to its frequency at the given position (N = 53).
Source: pfam.xfam.org

Profile models on MSA are based on single-residue conservation and are similar to non-interacting Potts models with local fields only. The corresponding probability distribution is factorized on MSA columns:

$$P(a_1, ..., a_N) = \prod_{i=1}^{N} f_i(a_i) \ , \tag{3.1}$$

with $f_i(a)$ the single-site frequency. The "information" contained in column i taking into account the amino-acid conservation score is therefore

$$I_i = \log_2(21) + \sum_{a=1}^{21} f_i(a) \log_2 f_i(a) \ , \tag{3.2}$$

where the first term is the maximum information, obtained for a totally conserved site, and the second term is the entropy of position i (up to a sign). This defines the "alignment logo", a graphical representation of the sequence conservation, where the amino acid letters are scaled according to their frequency. Fig. 3.4 shows the HMM logo [103] for PF00014, providing additional information about the profile-HMM of this domain family.

Profile-HMM are a generalization of profile models which include the possibility of amino-acid deletions or insertion. They take the

form of directed graphical models summarizing the statistical properties of a MSA, based on single-site conservation. The (visible) symbols composing the sequence (amino acids or gap) are conditioned by hidden states. The underlying probabilistic model is a Markov chain, which jumps from one hidden state to the other depending on a transition probability T; after the transition to a new hidden state, a symbol can be produced with an emission probability E.

Such models display three hidden states for each column i of the MSA:

— match states M_i, emitting the visible outputs with position-dependent probabilities;
— insertion states I_i, allowing for addition of excess residues;
— deletion states D_i, representing the lack of correspondence to the residue i, allowing for its removal.

Match and insertion states lead to the emission of an amino acid, whereas deletion states do not and often lead to gaps. The model also includes a gap penalty for matching an amino acid with a gap. The most used schemes assign a large cost for opening a gap and a smaller to extend it [39], taking the form of an affine penalty $p(l) = a + bl$ depending on the length l of the gap stretch (with $|a| > |b|$). Gaps and amino acids are therefore intrinsically different, and stretches of gaps are much more likely to occur than subsequences of repeated amino acids. We will come back to this asymmetry in Part II.

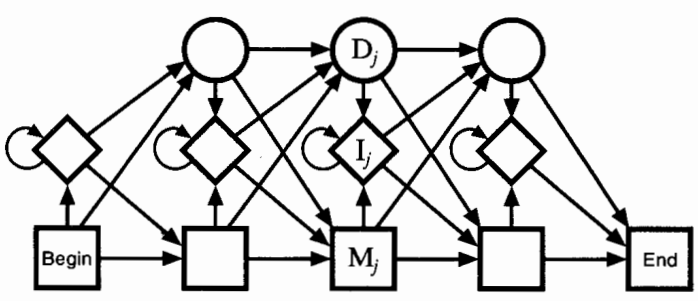


Figure 3.5 – Schematic structure of a profile-HMM. The alignment of a sequence to a profile-HMM corresponds to the most likely path from Begin to End. Squares indicate match states, diamond insertion states, and circles deletion states.
Source: [39]

The parameters T and E are estimated from seed alignments of previously aligned sequences. The quality of the seed alignments is therefore crucial in the procedure: in Pfam, they consist in manually curated alignments of 100-200 sequences, which are constantly improved.

The homology search goes as follows: given a seed alignment, the model parameters E and T are learned and the Uniprot database [30] can be searched for sequences that have a high probability under the model. The alignment of a sequence to a profile-HMM corresponds to

the most likely path from "Begin" to "End" on Fig. 3.5. The probability of a sequence $\underline{s}$ in the database reads [32]

$$P(\underline{s}) = \sum_{\underline{h} \in \mathcal{H}^n} P(\underline{s}, \underline{h}) = \sum_{\underline{h} \in \mathcal{H}^n} P(\underline{s}|\underline{h}) P(\underline{h}) \, , \tag{3.3}$$

where $\mathcal{H}^n$ denotes all possible hidden chains of length n. This calculation requires a summation over an exponentially large space and can be solved by dynamic programming methods. The Viterbi algorithm [32, 39] enables to efficiently find the maximum-likelihood hidden sequence:

$$\underline{h}^* = \arg \max_{\underline{h} \in \mathcal{H}^n} P(\underline{s}, \underline{h}) \, . \tag{3.4}$$

The whole procedure can be done with the "hmmsearch" command of HMMer software [47], with the profile-HMM (built on a seed alignment) as input. It defines various significance thresholds on the computed scores to decide whether a sequence should be added to the MSA or not. Such scores include:

— the log-odds score, *i.e.* the log of the ratio of the probability of the sequence $\underline{s}$ in the model to the probability of the sequence in a random model [39];
— the E-value, *i.e.* the expected number of hits among random sequences with equal or higher log-odds score.

The log-odds score will be referred to as the "HMMer score" in the following, and will be widely used in Part II to compare with DCA energies.

Profile-HMM is the most used method to search for homologous sequences, but it treats the protein residues independently and is only based on single-site conservation patterns. If the poor availability and quality of the sequences at the time it was introduced justified to neglect higher order statistics, the spectacular growth of the genomic databases allows for more engaged methods based on coevolution, such as DCA. Models assuming independent evolution of distinct residues are indeed unable to provide structural information about proteins (see next section), detect protein-protein interactions [44, 116], or describe epistasis [2] [46].

All the MSAs used in this dissertation were downloaded from Pfam and therefore built with a profile-HMM. We then develop approaches based on coevolution, which consists in pairwise models exploiting covariation patterns in MSAs. To be fully consistent, one should first build alignments with a pairwise method and then exploit the statistical correlations. This point has not been addressed in this dissertation, but would surely need to be tackled in the context of DCA approaches.

2. The effect of a mutation depends on the background, through interactions between genes.

3.1.3 *Protein structure prediction*

Prediction of residue-residue contacts in the tertiary structure of a protein, given sequence information alone, is the major application of DCA approaches [78, 116]. Its success makes it a reference in this major field of bioinformatics, and contact maps from plmDCA are at the source of the most advanced computational techniques in contact prediction using deep learning [106].

In this section, the results from DCA approaches will be compared with the direct measure of statistical correlations, mainly based on [82] and [42].

3.1.3.1 *Mutual information*

The mutual information (MI) is computed directly from the one- and two-point correlations in the MSA, after pseudocount regularization and sequence reweighting (*cf.* Section 2.3.2):

$$\mathrm{MI}_{ij} = \sum_{a,b=1}^{21} f_{ij}(a,b) \log \frac{f_{ij}(a,b)}{f_i(a) f_j(b)} \,. \tag{3.5}$$

MI is the KL divergence of the joint distribution $f_{ij}(a,b)$ from its factorized form $f_i(a)f_j(b)$ [116]. It equals 0 if and only if i and j are uncorrelated and it is positive else.

As mentioned in Chapter 1 (*cf.* Fig. 1.3), high mutual information may result from either a strong direct coupling between i and j (which is interpreted as a contact between the residues), or from an indirect interaction mediated through a chain of couplings (the residues are not in contact). As MI is intrinsically local, it cannot disentangle direct from indirect interactions. As a result, the accuracy of the methods directly measuring the correlations remains very limited to unveil structural information from protein MSAs [38, 50, 87, 88], *cf.* Fig. 3.7.

3.1.3.2 *Direct-coupling analysis*

Contrary to the Ising case – where each interaction is described by one scalar coupling J_{ij} – each residue pair (i,j) in the Potts model is characterized by a $q \times q$ matrix $\{J_{ij}(a,b)\}_{a,b=1\ldots q}$. To measure the coupling strength between two sites, the inferred coupling matrix needs to be mapped onto a scalar parameter, which will be subsequently ranked: the larger they are, the higher is the probability that residues i and j are in contact in the tertiary structure.

Previous work have mainly used the so-called direct information (DI) [82, 116], the mutual information of a restricted two-site probability model only including the direct coupling between the two positions to be scored:

$$\mathrm{DI}_{ij} = \sum_{a,b=1}^{21} P_{ij}^{(\mathrm{dir})}(a,b) \log \frac{P_{ij}^{(\mathrm{dir})}(a,b)}{f_i(a) f_j(b)} \,, \tag{3.6}$$

where the isolated two-site model in question displays the direct couplings $J_{ij}(a,b)$ and modified fields $\tilde{h}_i(a)$ to match the empirical frequencies:

$$P_{ij}^{(dir)}(a,b) = \frac{1}{z_{ij}} \exp\left(J_{ij}(a,b) + \tilde{h}_i(a) + \tilde{h}_j(b)\right) . \tag{3.7}$$

More recently, it has been observed that a different score F_{ij}^{APC} - the Frobenius norm F_{ij} of the coupling matrix adjusted by an average product correction (APC) term - improves the contact prediction in the case of pseudolikelihood [42] or mean-field related methods [10, 29]:

$$F_{ij} = \sqrt{\sum_{a,b=1}^{21} J_{ij}(a,b)^2} , \qquad F_{ij}^{APC} = F_{ij} - \frac{\langle F_{ij}\rangle_i \langle F_{ij}\rangle_j}{\langle F_{ij}\rangle_{ij}} , \tag{3.8}$$

where $\langle\cdot\rangle_i$ denotes the position average. The APC score[3] is not gauge invariant, contrary to the DI score. Before computing the norm, the couplings are shifted to the zero-sum gauge (defined at Eq. (2.29)), as it is the gauge that minimizes the Frobenius norm.

Introduced in [38], the APC correction is presented as an entropy correction to suppress effects from phylogenetic biases and insufficient sampling. The origin of this efficiency is unclear, but it does improve the accuracy of contact prediction compared to DI [42].

3.1.3.3 *Comparison*

A residue pair is considered to be a contact in the tertiary structure if its minimal heavy-atom distance is below 8 Å in the crystallized protein structure. This threshold is quite large and is often criticized by structural biologists, claiming that 6 Å is a more reasonable value for residue-residue contact distances. It is usually chosen in DCA related methods as the distance distribution among residue pairs is bimodal with two peaks around 3–5 and 7–8 Å [82]. To avoid trivial contacts – local contacts inside the secondary structure – a minimum separation between the residues along the protein backbone is imposed (usually $|j-i|>4$).

one helix turn is 3.6 Å

Fig. 3.6 is taken from [82] and displays the top 20 predictions of the ranked MI score – directly computed from the correlations in the MSA – and DI score – obtained after inferring the coupling matrices in the mfDCA approximation – for the protein domain family Region 2 of the bacterial Sigma factor (Pfam id: PF04542). Red links indicate true positive predictions and green links indicate false positive. 19 out of 20 predictions with DCA appear to be truly native contacts, leading to a precision of 95% (panel A). On the other hand, only 13 out of 20 predictions with mutual information are truly residue-residue

3. For the sake of simplicity, F^{APC} – the Frobenius norm with the average product correction – will be referred as "APC score" in the following.

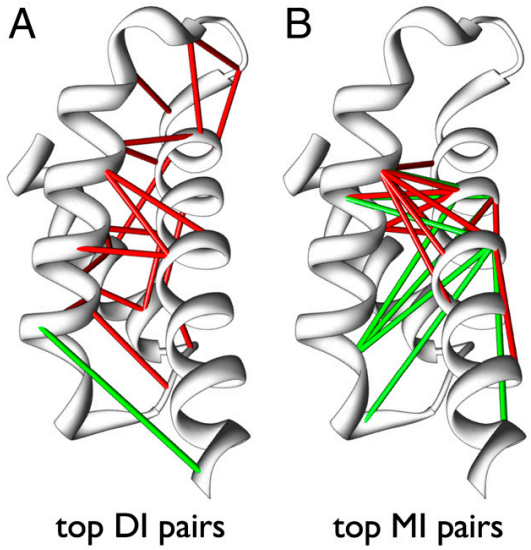


Figure 3.6 – Contact predictions for the family of domains homologous to Region 2 of the bacterial Sigma factor (Pfam id PF04542) mapped to the sequence of the SigmaE factor of E. coli (encoded by rpoE) (PDB id: 1OR7). Panel A shows the top 20 DI predictions, and panel B shows the top 20 MI predictions for residue–residue contacts, both with a minimum separation of five positions along the backbone. Each pair with distance < 8Å is connected by a red link, and the more distant pairs are connected by the green links.
Source: *figure and caption from* [82].

contacts, reaching a precision of only 65% (panel B). MI predictions are also concentrated on only a few residues, whereas DI predictions are more evenly distributed in the protein.

In practice, we compute the positive predictive value (PPV) to compare the different scores, which is the fraction of true predictions among the total number of predictions. The PPV in the top n predictions is usually plotted against n. Fig. 3.7, also taken from [82], displays the average PPV for 131 Pfam families[4] with DI scores obtained with mfDCA (black curve) and MI scores (red curve). The third method – Bayesian dependency tree [24] – is out of the scope of this thesis. Scores obtained from DCA couplings (DI, or even better APC [42]) largely outperform simple covariance analysis.

First introduced in 2009, DCA approaches have helped unveiling the structural information contained in protein MSAs and proved that although residue-residue contacts are the result of complex physicochemical interactions (*e.g.* hydrophobicity or amino-acid charge), they can actually be inferred from purely statistical considerations. Protein contact prediction has today reached a level of precision that was be-

4. Although the y-axis label indicates TP (True Positive) rate, it is actually the PPV which is plotted ...

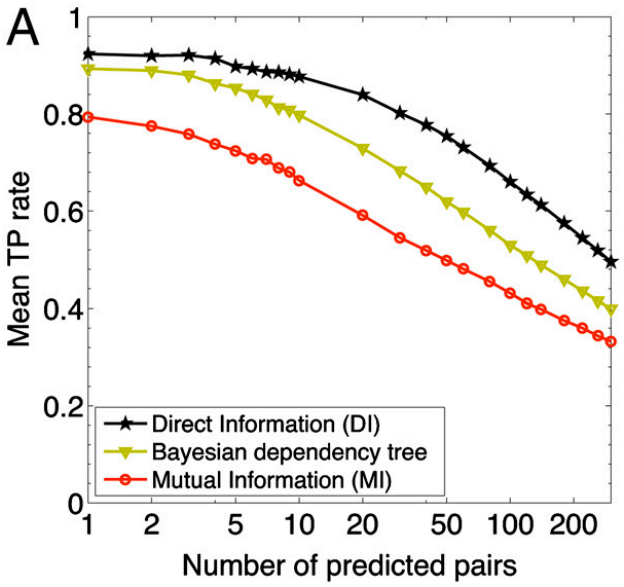


Figure 3.7 – Mean PPV for 131 domain families, as a function of the number of top-ranked predictions. DI scores obtained with DCA clearly outperform MI.
Source: [82].

fore thought to be unattainable. More recently, such approaches have been applied to other challenging fields such as protein-protein interaction networks [44], or mutation fitness landscape [46, 76]. Structural prediction consists in revealing the topology of the network of interaction by detecting strongly and directly interacting pairs of sites. These new topics, however, require a more detailed description of the system and aim at constructing a global energetic model. Most of the work presented in this dissertation aims at better understanding DCA approaches and trying to apply them beyond protein structure prediction.

3.2 LATTICE PROTEINS

LP are exactly solvable models of proteins folding on a 3D lattice. As *in silico* systems, LP allow for precise numerical control, and large samples of sequences corresponding to a single fold can be generated without phylogenetic bias. The many common properties they share with real proteins (efficient folding, non trivial statistical features, existence of families in the profile-HMM sense with conserved folds, etc.), make them an ideal benchmark for better understanding inference methods developed in the context or real protein data. Most of the results presented in this section have been recently published in [60] by members of our team at ENS.

see Chapter 3 of Part III for more details

3.2.1 *Background*

3.2.1.1 *Polymers on a cubic lattice*

A lattice protein is a chain of $N = 27$ residues occupying the sites of a $3 \times 3 \times 3$ simple cubic lattice; each residue position in the chain can be occupied by one of the 20 different amino acids. $\mathcal{N} = 103,346$ self-avoiding conformations unrelated through symmetry have been enumerated [104]. Each conformation defines a possible structure, or fold of a the protein sequence. The geometry of the cube imposes exactly 28 contacts (neighbors on the lattice but not on the backbone) between the protein sites, *cf.* Fig. 3.8.

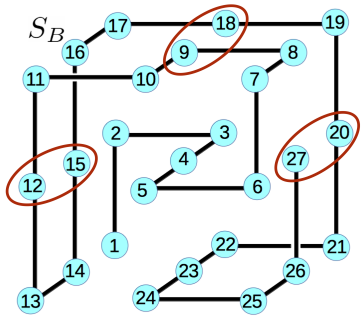


Figure 3.8 – Representative fold of a lattice protein (structure S_B); 3 out of the 28 contacts of this structure have been circled in red.

Given a fold S, an energy is assigned to each amino-acid sequence $\underline{a} = (a_1, ..., a_{27})$:

$$\mathcal{E}(\underline{a}|S) = \sum_{i<j} c_{ij}^{(S)} \, E^{MJ}(a_i, a_j) \, , \tag{3.9}$$

where $c_{ij}^{(S)}$ is the contact map of structure S, *i.e.* the 27×27 adjacency matrix ($c_{ij}^{(S)} = 1$ if i and j are in tertiary contact, but not along the chain, and 0 otherwise). Amino acids in contact interact through the Miyazawa-Jernigan (MJ) statistical potential $E^{MJ}(a, b)$ [79], which will be extensively described in Part III Chapter 1. The probability that a given sequence $\underline{a}$ folds in structure S is defined by

$$\begin{aligned} P_{nat}(S|\underline{a}) &= \frac{e^{-\mathcal{E}(\underline{a}|S)}}{\sum_{S'=1}^{\mathcal{N}} e^{-\mathcal{E}(\underline{a}|S')}} \\ &= \frac{1}{1 + \sum_{S \neq S'} e^{-[\mathcal{E}(\underline{a}|S') - \mathcal{E}(\underline{a}|S)]}} \, , \end{aligned} \tag{3.10}$$

and depends on its energies in all folds S'. A good folder $S^\star$ is a sequence $\underline{a}$ with a minimal energy $\mathcal{E}(\underline{a}|S^\star)$ and the largest energy gap $\mathcal{E}(\underline{a}|S') - \mathcal{E}(\underline{a}|S^\star)$ with competing structures S'. These conditions are satisfied by many sequences which define a protein-like family.

3.2.1.2 *Lattice protein families*

The method to create alignments of sequences folding in the same structure is given in [60], and will be briefly described here. A MSA corresponding to a native fold S is generated through Monte Carlo Markov Chain (MCMC) sampling of $P_{nat}(S, \underline{a})$, with the Metropolis rule. The sequence $\underline{a}$ is mutated into $\underline{a}'$: if $P_{nat}(S|\underline{a}') \geqslant P_{nat}(S|\underline{a})$ the mutation is accepted, otherwise it is accepted with the probability $[P_{nat}(S|\underline{a}')/P_{nat}(S|\underline{a})]^{\beta}$ (< 1). The corresponding effective Hamiltonian also includes contributions coming from all other folds S' with multiple body interactions at any orders > 2:

$$\begin{aligned} \mathcal{H}(\underline{a}) &= -\beta \log P_{nat}(S|\underline{a}) \\ &= \beta \log \left(1 + \sum_{S' \neq S} \exp \sum_{i<j} \left[c_{ij}^{(S)} - c_{ij}^{(S')} \right] E^{MJ}(a_i, a_j) \right) . \end{aligned} \tag{3.11}$$

Supplementary materials of [60] give access to four MSAs of $B = 50000$ sequences generated at regular intervals and folding in their native structure with $P_{nat} > 0.995$ (fine-tuning of the inverse temperature β).

3.2.1.3 *Competitor folds*

The closest competitor to the native fold S is defined in [60] as the structure S' minimizing the gap $\Delta(S'|S)$ defined through

$$e^{-\Delta(S'|S)} = \left\langle e^{-[\mathcal{E}(\underline{a}|S') - \mathcal{E}(\underline{a}|S)]} \right\rangle_{\underline{a}} , \tag{3.12}$$

where the mean is over sequences $\underline{a}$ in the MSA defined above. The number N_S of competitors and their typical gap Δ with the native structure S are approximated through

$$N_S e^{-\Delta} = \sum_{S'(\neq S)} e^{-\Delta(S'|S)} . \tag{3.13}$$

The average contact map of these competitors reads

$$\bar{c}_{ij} = \frac{1}{N_S e^{-\Delta}} \sum_{S'(\neq S)} e^{-\Delta(S'|S)} c_{ij}^{(S)} . \tag{3.14}$$

3.2.2 *Covariation in lattice proteins*

Covariation properties of lattice proteins (LP) have been studied only recently in [60], and the main results are summarized in this section. The same inverse methods used for real proteins have been applied to the generated MSAs, and, as in real data, inferred couplings (with mfDCA, plmDCA or ACE) are very accurate in predicting contacts in the native structure, even if MI predictions are also quite good. Very

interestingly, a linear dependency is observed between the inferred couplings $J_{ij}(a,b)$ and the MJ energetic parameters $E^{MJ}(a,b)$, with a coefficient λ_{ij} depending on the residue pair:

$$J_{ij}(a,b) \approx \lambda_{ij} E_0^{MJ}(a,b) \,, \tag{3.15}$$

where $E_0^{MJ}(a,b)$ is the MJ statistical potential placed in the gauge of the couplings (see Chapter 1 of Part III for more details). This coefficient is interpreted in [60] as a measure of the coevolutionnary pressure on residues i,j due to the design of the native structure. Fig. 3.9 taken from [60] displays – for fold S_B – the empirically measured λ_{ij}

$$\lambda_{ij} = \frac{\sum_{ab} J_{ij}(a,b) E_0^{MJ}(a,b)}{\sum_{ab} E_0^{MJ}(a,b)} \,, \tag{3.16}$$

as a function of $\delta c_{ij} = c_{ij} - \bar{c}_{ij}$, the difference between the native contact map of S_B and the average contact map over the structures in competition with S_B (*cf.* Eq. (3.14)). The dependence of the pressure λ_{ij} is monotonic in δc_{ij}, and both have the same sign.

Pairs of sites that can never be in contact due to geometrical constraints (magenta pluses), corresponding to $\delta c_{ij} = 0$, display weak pressures (at the level of noise in the data). Pairs that are in contact in the native fold, but not in the competitor structures (filled triangles), corresponding to a large positive δc_{ij}, are subject to strong covariation pressures, as it is essential that they stabilize the native fold and not the competitors. On the contrary, pairs in contact both in the native structure and its competitors (small and positive δc_{ij}) show weak pressures (empty triangles), as they are less specific to the native fold. Respectively, pairs not in contact in the native structure display negative pressures. Either they are also not in contact in the competitors – corresponding to small negative δc_{ij} – and the pressure is weak (empty squares), or they are in contact in the competitor folds (large negative δc_{ij}) and are therefore subject to negative design: such conformations should be avoided and the resulting pressure in large negative.

The evolutionary pressure λ_{ij} can therefore be used for contact prediction and gives better results than classical estimators such as the APC score [60] (see also Chapter 3 of Part III). The idea is that large APC scores not corresponding to contacts can result from large couplings anti-correlated with $E^{MJ}(a,b)$. As the APC score is based on the squared couplings (*cf.* Eq. (3.8)), such pairs are given high scores and give rise to false positive. On the contrary, they are associated with large negative evolutionary pressures and are low-ranked with the predictor λ_{ij}.

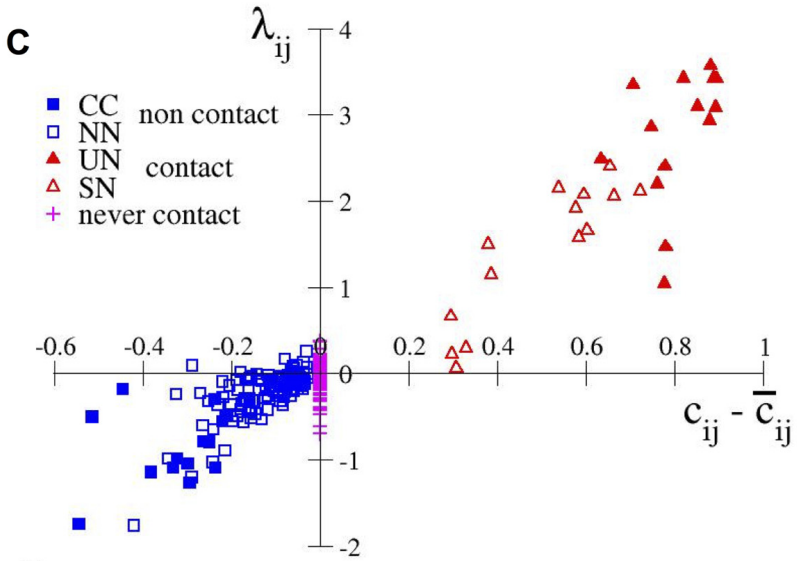


Figure 3.9 – Empirical pressure λ_{ij} for each pair of sites (i,j), *vs.* $\delta c_{ij} = c_{ij} - \overline{c}_{ij}$ for structure S_B. The 195 pairs of sites which can never be in contact on any fold due to the lattice geometry are shown with magenta pluses. The 28 contacts on S_B (red symbols) are partitioned into the Unique-Native (UN, 14 full triangles) and Shared-Native (SN, 14 empty triangles) classes, according to, respectively, their absence or presence in the closest competitor structure. The remaining 128 pairs of sites (blue symbols) are not in contact on S_B, and are partitioned into the Closest-Competitor (CC, 14 full squares) and the Non-Native (NN, 114 empty squares) classes, according to, respectively, whether they are in contact or not in the closest competitor structure.
Source: *figure and caption from* [60]

.

Part II

SCORING OF SEQUENCES

This section is dedicated to illustrating the ability of the direct-coupling analysis (DCA) inference methods to go beyond protein structure prediction. We start by shortly analyzing the data of [107], where the authors designed new proteins based on an alignment of the WW domain, and experimentally tested their ability to fold (Chapter 1). Then, applying DCA in the context of remote homology detection to a dozen of protein domain families, we observe that alignment gaps give rise to several complications and we define a null model to suppress this dominating signal, leading to interesting but unexpected results (Chapter 2). Finally, inspired by this problem of gaps, we develop a more principled approach modeling gaps as missing information and thus gap-rich sequences as partial observations, in the theoretical framework of mean-field inverse Potts models (Chapter 3).

1

A FIRST EXAMPLE: WW DOMAIN

Originally developed in the context of prediction of residue-residue contacts in the tertiary and quaternary structures of proteins, DCA related approaches have proven to be very accurate and are today widely used in the field. Encouraged by this success, DCA has been successfully applied by members of our group to predict the antibiotic drug resistance properties of beta lactamase TEM-1 [46], and – as it has already been mentioned several times in this dissertation – in other challenging fields, such as drug resistance detection in the HIV virus [25], fitness effects of mutations [45, 76], or folding properties of lattice proteins [60].

In the specific context of [107], where properties of artificial sequences have been tested experimentally, we will shortly asses the ability of DCA related approaches to be good predictors of protein folding properties.

1.1 BACKGROUND

The WW domain is a small protein domain involved in specific interactions with protein ligands. Socolich and collaborators [107] designed new artificial sequences using the statistical information from the multiple sequence alignment (MSA) of the WW domain (Pfam id PF00397, $N = 33$ residues), and experimentally tested their ability to fold into the native WW structure. Four groups of new sequences have been generated based on Monte Carlo Markov Chain (MCMC) simulations:

— Natural (NAT): natural WW sequences drawn from the original MSA,

— Coupled conservation (CC): artificial sequences with the same single-site frequencies $f_i(a)$, and the same connected correlations $C_{ij}(a, b)$ than the original MSA,

— Independent-site conservation (IC): artificial sequences with the same single-site frequencies $f_i(a)$,

— Random (R): random sequences, only with the overall same frequencies than the original MSA.

The main result of the paper is that the knowledge of statistical correlations is sufficient but also necessary to create sequences that fold into the native WW structure. Indeed, as displayed on Fig. 1.1, a significant fraction of CC sequences correctly fold, whereas none of IC or R sequences do. This emphasizes the role of the interactions between residues in the folding process - which are not described by

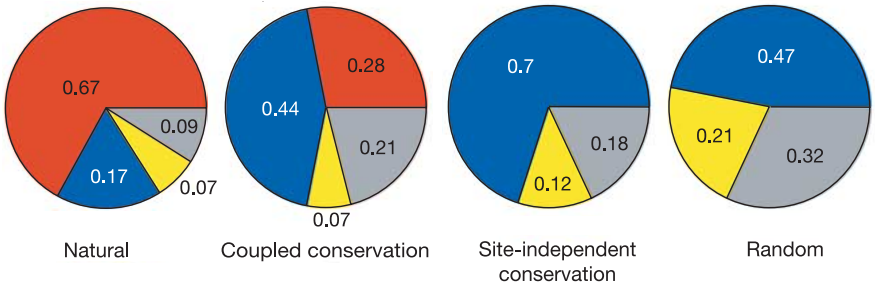


Figure 1.1 – Outcome of folding studies for natural, CC, IC and random WW sequences. Red: natively folded, blue: soluble but unfolded, yellow: insoluble, grey: poor expressing. A significant part of CC sequences are natively folded, whereas IC sequences are not.
Source: [107]

a independent or random model - and justifies the use of a pairwise Potts model, beyond maximum entropy (MaxEnt). Note that some natural sequences do not fold, simply reflecting the imperfections of the experimental procedure to test folding.

1.2 FOLDING PREDICTION WITH DIRECT-COUPLING ANALYSIS

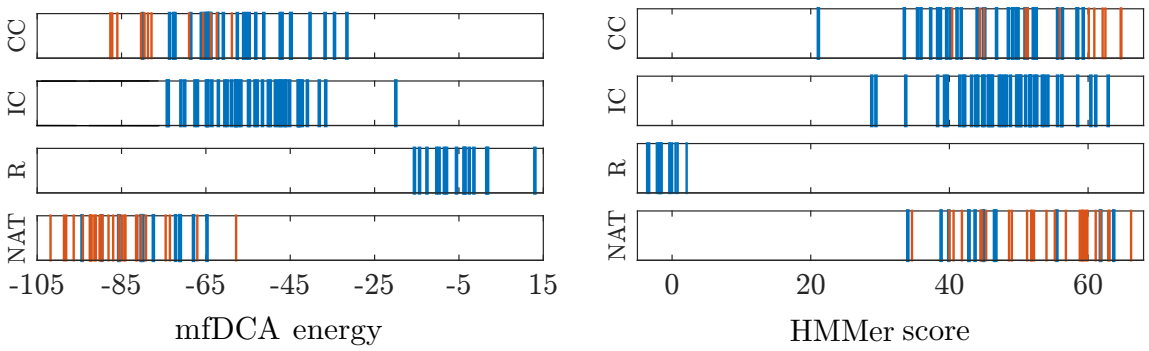


Figure 1.2 – mfDCA energies (left panel) and HMMer scores (right panel) of sequences from [107], experimentally folding (red) and not folding (blue) in the native WW structure. mfDCA seems to be more able to discriminate between natively folded and not folded sequences.

initiated with C. Feinauer from Politecnico di Torino

In this short study, we will show that the energy in the DCA Potts model – inferred on the original WW alignment – of a given sequence is a good predictor of its ability to fold (even among CC and NAT ones), whereas the procedure of [107] scores only the full MSA or artificial protein sequences, and does not provide any information about the potential value of each single sequence.

Each sequence $\underline{a} = (a_1, ..., a_N)$ in the *test* alignment – composed of the four groups of new sequences described above – is assigned the energy (*cf.* Eq. (3.9) in Part I):

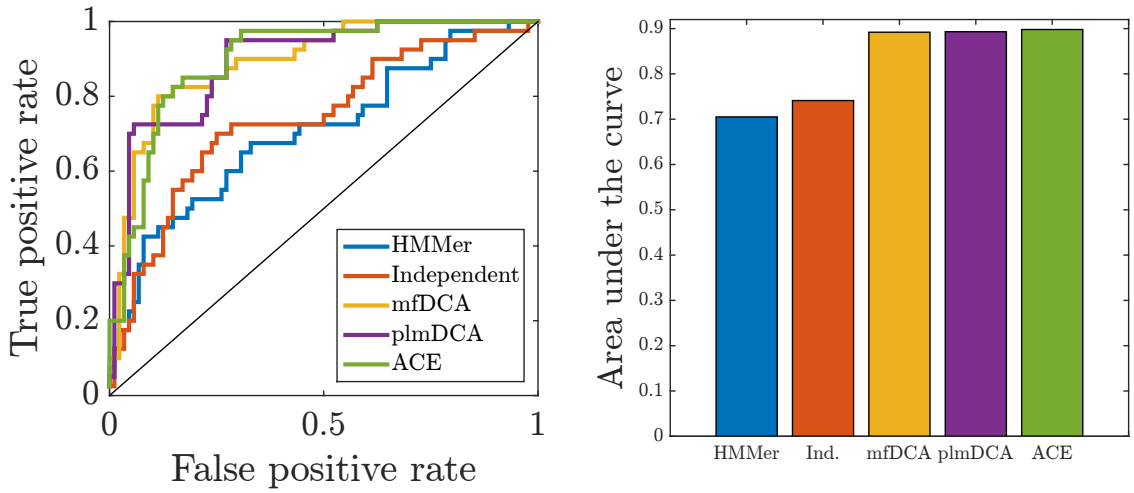
$$E(a_1, ..., a_N) = -\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(a_i, a_j) - \sum_{i=1}^{N} h_i(a_i) \,, \qquad (1.1)$$

"good" sequences are given low energies and high HMMer scores

where $\{h_i(a), J_{ij}(a, b)\}$ are the couplings and fields inferred with DCA in the mean-field approximation (mfDCA) from the original MSA of the WW domain (PF00397), the *training* alignment. Fig. 1.2 shows the energies and HMMer [47] scores (see Chapter 3 of Part I for the definition) of the different groups of sequences. Red (resp. blue) bars correspond to sequences that do (resp. do not) fold experimentally, according to [107]. As expected, natural sequences globally have a lower mfDCA energy than the IC sequences and the folded CC sequences lie in a low energy region mostly free from IC sequences. On the contrary, HMMer gives similar scores to Natural, CC and IC sequences as they share the same single-site frequency patterns. Random sequences are discarded by both models. Similar results are obtained with other pairwise inference methods, such as pseudolikelihood maximization (plmDCA) or adaptive cluster expansion (ACE), and by replacing HMMer by a site-independent (factorized or fields) model (*cf.* Eq. (3.1) in Chapter 3 of Part I).

A graphical representation of the performance of the different models as binary classifiers, *i.e.* discriminating between folded and unfolded sequences is given by the receiver operating characteristic (ROC) and the area under the ROC curve (AUC), displayed on Fig. 1.3. The ROC curve is the true positive rate – proportion of positives (folded sequences) detected (given a top-ranked score) – as a function of the false positive rate – fraction of negative (unfolded sequences) identified as such (given a low score) – for various thresholds (number of predictions). A random guess would go along the diagonal in the ROC space and get a AUC of 0.5; a perfect classifier would have a ROC constant equal to 1 and a AUC of 1. Any method in between recovers non-random information.

In this case, mfDCA, plmDCA, and ACE are much better classifiers than independent-site models such as HMMer or the independent model. It is also quite astonishing that the pairwise models are so similar, given that they cover very different energy ranges, with mfDCA seemingly working at much lower temperature. DCA related approaches therefore seem to be very efficient in predicting whether a given sequence will fold in a native structure or not, in the specific context of [107]. This success is very encouraging, but needs to be confirmed on more data. The next chapter will focus on a dozen of protein families in the context of homology detection.



(a) (b)

Figure 1.3 – ROC (panel (a)) and AUC (panel (b)) curves illustrating the performance of the different models in discriminating between folded and unfolded sequences. Pairwise models are better classifiers.

2

SEQUENCE SCORING AND GAP TREATMENT

In the previous section, we showed that DCA related approaches seem to be able to assess the folding ability of a given sequence, in the specific context of [107]. More data will be available in the near future, leading to further work by our team. In the meantime, we would like to test our method in the context of homology detection. As mentioned in Chapter 3 of Part I, homology detection is a major field of bioinformatics as it allows to assign a sequence to a protein domain family, therefore building sequence alignments and predicting the structure and function of proteins. A vast literature addresses this topic in bioinformatics and machine learning [40, 47, 59, 101, 108, 109].

Methods treating residues independently – such as profile-HMM – perform quite well on phylogenetically close enough proteins, because their sequences are still similar. However, there is no approach working well in all cases for remote homology detection, a much harder problem focusing on proteins which are "far" in a phylogenetic point of view and with low sequence similarity. Interestingly, a few methods tackling this specific problem tend to use non-local information from MSAs [16, 35, 74]. Moreover, for the specific case of RNA [1], coevolution of the secondary structure is taken into account by approaches introduced as covariance models (such as Infernal [86]). The naive idea that covariation patterns – currently not taken into account – may be better preserved than single-site conservation patterns in remote but homologous sequences motivates the use of DCA approaches in this context.

The three domains of the tree of life offer a natural clustering into phylogenetically different groups: archaea, bacteria and eukaryota, to which we can add viruses. In the following, we will divide a dozen MSAs of protein families into sub-alignments, according to these domains of life. The task here is more complex than in the last chapter, as we have no such thing as a binary information whether the sequence is folded/unfolded for instance. We will however study the energy distribution of DCA models across these domains of life and compare it to the HMMer scoring, the currently most used tool in the context of homology detection.

1. Ribonucleic acids are polymeric molecules consisting in chains of nucleotides (A, C, G, U). Their secondary structure is composed of Watson-Crick base pairings.

2.1 SCORING PROCEDURE

Before applying it to broad domain families with both bacterial and eukaryotic sub-families, the scoring procedure will be first illustrated on the Kunitz/Bovine pancreatic trypsin inhibitor domain (Pfam id PF00014, N = 53 residues). The Potts parameters $\{J_{ij}(a,b), h_i(a)\}$ are inferred with mfDCA from the one- and two-point statistics in the MSA – which has been downloaded from Pfam, as well as the profile-HMM (needed to compute the HMMer scores). The original MSA is referred to as the *training* alignment. A *test* alignment is created by searching (using the *hmmsearch* tool of the HMMer software [47]) the whole Uniprot database with the PF00014 profile-HMM [2]. We force HMMer to align also with negative log-odds scores, hoping that among all these aligned sequences there are not only totally unrelated elements, but also distant homologs. This artificial alignment is of course extremely rich in gaps (with some sequences gaped at more than 50%). The idea initially was to use DCA to distinguish between homologs and unrelated sequences.

disabling all significance thresholds

2.1.1 *Gaps are not modeled well by direct-coupling analysis*

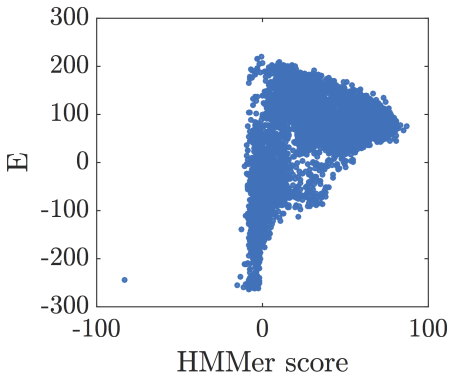


Figure 2.1 – PF00014 - Energies of the sequences in the artificial *test* alignment, with the Potts parameters inferred on the original MSA, as a function of their HMMer scores. Sequences with more than 30% of gaps are given low HMMer scores, but surprisingly extremely low energies. The "sequence" with the lower HMMer score has one of the lowest energies and consists in 52 gaps and 1 amino acid.

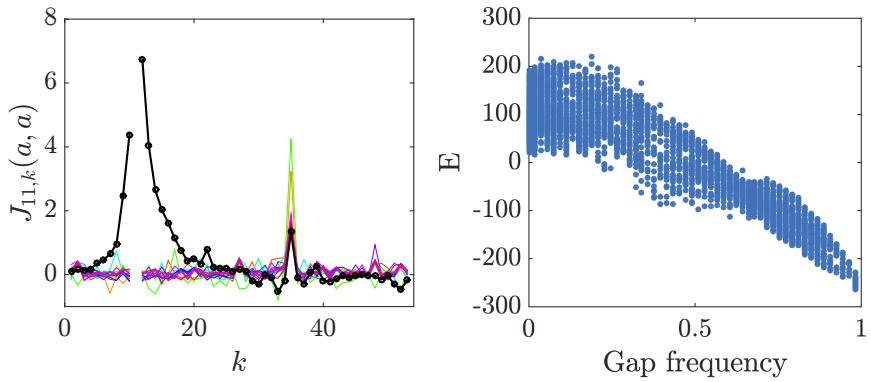
Similarly to the previous section, an energy (*cf.* Eq. (1.1)) is assigned to each sequence of the artificial *test* alignment, with the Potts parameters inferred on the original MSA. The energy distribution is compared with the HMMer score, as displayed on Fig. 2.1. Surpris-

2. The procedure to build alignments from profile-HMM is briefly explained in Part I Section 3.1.2.2

"good" sequences are given low energies and high HMMer scores

ingly, some sequences are given very low energies in the DCA model, whereas they have bad HMMer scores. A quick verification shows that they are very rich in gaps, from 30% to 98% – the sequence with one of the lowest energies (-245) also has the lowest HMMer score[3] (-83) and consists in 52 gaps and 1 amino acid. This "sequence" has of course absolutely no biological meaning and should never have been aligned, but for the sake of this study we disabled HMMer significance thresholds. It anyway seems that DCA dangerously underestimates the energy of gap-rich sequences.

When sequences are distant in evolution – such as proteins present in both bacterial and eukaryotic domains of life – not necessarily all parts of the sequences are well conserved and alignable. As a consequence, broad families – containing a lot of sequences $B_{eff} \gtrsim 500$ in both domains of life – frequently have many gaps. Gaps in alignments are reflecting deletion or insertion mutations in sequences, and therefore a mismatch between two residues to be aligned in the same MSA column. Consequently, gaps are intrinsically different from amino acids. Their distribution along the positions of the MSA is also very different as they tend to come in repeated stretches - especially at the beginning or the end of sequences - which is not the case for amino acids (*cf.* Fig. 3.3 in Part I). However, they are treated as an extra symbol by DCA related approaches.



(a) (b)

Figure 2.2 – PF00014 - Panel (a): couplings $\{J_{i,k}(a,a)\}$ with $i = 11$ as an example and $k \in \{1,...,N = 53\}$, inferred with mfDCA, for a gap (black) and amino acid (colors). Panel (b): energies of the *test* alignment, as a function of the frequency of gaps in the sequences. Gap-Gap couplings are dominant in a range of about 10 sites and the energy of a sequence is highly correlated to the frequency of gaps.

These gap-induced artifacts give rise to strong DCA couplings between gaps. Fig. 2.2a shows the range of the couplings $J_{ij}(a,a)$, with

3. Gap-rich sequences are given low HMMer scores because of gap penalties in the profile-HMM procedure, see Part I Chapter 3.

$i = 11$ as an example, for all symbols a (amino acids and gap) and as a function of $j \in \{1...N\}$. The highest interactions are gap-gap (black curve, other colors are amino acids), in a range of about 10 sites. The second peak reflects a very strong interaction between residues 11 and 35 due to a disulfide bridge in the 3D structure (a contact between two Cysteine, in green on the Figure).

The couplings have been inferred on the original MSA (the *training* alignment) and are absolutely unrelated with the artificial *test* alignment. 89% of its sequences actually contain less than 10% of gaps. These spurious interaction therefore arise even if few gaps are present in the *training* alignment. It however does not impact residue-residue contact prediction as severely, given that residues too close on the backbone which are the most involved in the gap-gap interactions are usually discarded from the ranking (*cf.* Part I Chapter 3).

The high gap-gap couplings lead to an artificially low energy of the gap-rich sequences in the *test* alignment. Fig. 2.2b shows how correlated the mfDCA energy is to the frequency of gaps, from about 30% of gaps in the sequence[4]. Previous work [43] proposed to suppress strong couplings induced by gaps by introducing additional gap parameters learned from the MSA besides usual Potts couplings and fields. This new asymmetry between gaps and amino acids improves contact prediction, especially in the region at the end of the proteins richer in gaps. Here, we present another solution taking the form a null model on the gap distribution.

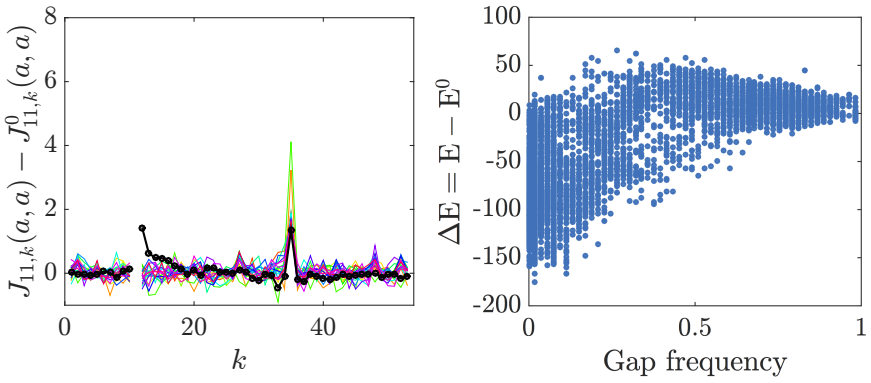
2.1.2 *Null model*

The null model we have implemented requires a manipulation on the *training* MSA. It consists in keeping the gaps at the same positions, but reemitting the amino acids with their overall frequency in Uniprot (totally site-independent). Any signal stemming from amino acids is therefore suppressed and only the spurious signals coming from gaps remain. The training set can be replicated a certain number of times (typically 10) to avoid finite size effects, the randomization of amino acids being done independently in the replicas. The inference is then done as usual with DCA approaches using two different sets of one- and two-sites frequency counts $\{f_i(a), f_{ij}(a,b)\}$ and $\{f^0_i(a), f^0_{ij}(a,b)\}$, coming respectively from the actual training MSA, and from the new randomized (except for gaps) training MSA. It leads to two sets of Potts couplings and fields: $\{h_i(a), J_{ij}(a,b)\}$ and $\{h^0_i(a), J^0_{ij}(a,b)\}$.

The energy difference between the standard and the null model should capture the signal due to amino acids alone. For a given sequence $\underline{a} = (a_1, ..., a_N)$, it reads

$$\Delta E(a_1, ..., a_N) = E(a_1, ..., a_N) - E^0(a_1, ..., a_N)\ , \tag{2.1}$$

4. The scoring of artificial WW domain sequences in the last chapter was not affected by gaps, as there are very few of them in this specific data.



(a) (b)

Figure 2.3 – PF00014 - Panel (a): effective couplings $\{J_{i,k}(a,a) - J^0_{i,k}(a,a)\}$ with $i = 11$ as an example, $k \in \{1, ..., N = 53\}$ and a amino acids (colors) and gap (black), inferred with mfDCA, corrected by the null model. Panel (b): corresponding energies of a test alignment as a function of the frequency of gaps in the sequence. The effect of gaps is suppressed and energies globally anti-correlated to the frequency of gaps.

where $E^0(a_1, ..., a_N) = -\sum_{i<j} J^0_{ij}(a_i, a_j) - \sum_i h^0_i(a_i)$. The range of the new effective couplings $J - J^0$ are displayed on Fig. 2.3a: the strong long range gap-gap couplings have been suppressed, and the energies are slightly anti-correlated with the gap frequency from about 30%, as it should be (Fig. 2.3b). This model will be applied in the following to a dozen of broad protein families, containing sequences in various domains of life.

2.1.3 *Scoring method*

It is necessary at this point to understand the difference between the *training* and the *test* alignments. The *training* alignment is the one on which the parameters of the different models are learned: the frequency counts $\{f_i(a), f_{ij}(a, b)\}$ for DCA related approaches and the profile-HMM for HMMer scoring. The *test* alignment contains the sequences to which a score will be assigned: an energy for DCA and a log-odds score for HMMer. It can of course be very different from the *training* alignment, but a constraint is that their sequences should have the same length N. For example, in the following section, the training set is typically the bacterial sub-alignment of the studied protein family, and the test sets are the eukaryotic, archaea, and viruses sub-alignments.

the detailed procedure is in Appendix B

The computation of the HMMer score should be explained in more details, as it is not straightforward. The inputs of the HMMer software are a profile-HMM and a target database of full length (not aligned) sequences. It finds the most relevant (hit) domains from

the sequences in the database corresponding to the profile. For consistency with the "training/test" point of view, a profile-HMM is built, specific only to the training alignment. This profile is therefore different from the Pfam profile used to align the sequences of the entire family (training+test). Moreover, HMMer requires the full length version of the sequences (available in Uniprot [30]) and cannot use as a target database the aligned MSA from Pfam. We therefore need to make sure that the relevant hit domains correspond to the test alignment, if we want to compare the HMMer scores and the DCA energies.

2.2 RESULTS

PFAM ID	DESCRIPTION	N	B_{eukar}	B_{bact}
PF00004	AAA	132	18844	31242
PF00006	ATP synthase α/β family	215	11041	21821
PF00011	HSP20/α-crystallin domain	102	3660	5812
PF00013	KH domain	60	12576	6502
PF00023	Ankyrin repeat	33	7256	1006
PF00027	Cyclic nucleotide-binding	91	8811	17078
PF00033	Cytochrome b/b6/petB	188	1577	6821
PF00089	Trypsin	220	18275	3897
PF00091	Tubulin/FtsZ/GTPase domain	216	14988	876
PF00664	ABC transporter	275	13285	37386
PF03547	Membrane transport protein	385	1148	7790

Table 1 – List of the selected Pfam families.

Eleven protein families (*cf.* Table 1) have been selected because of their natural division into two bacterial and eukaryotic sub-families, with the condition that they contain enough sequences so that DCA related approaches can be applied on each of the sub-alignments ($B_{eff} \gtrsim 500$). Only six of them, displaying the most significant results will be studied in the following: PF00011, PF00013, PF00027, PF00033, PF00091, and PF00664. Fig. 2.4 displays a graphical representation of the distribution of the Pfam family PF00011 across species, with large bacterial and eukaryotic sub-families.

The main result of this preliminary study is that DCA related approaches (mfDCA and plmDCA) have a stronger tendency to discriminate between sequences from the same family, where this discrimination is, *e.g.* consistent with the phylogenetic distribution. Besides, DCA is sometimes able to detect errors in the labeling of sequences, or at least gives interesting insights about unclassified sequences - automatically annotated unreviewed sequences, metagenomic sequences

Figure 2.4 – Graphical representation of the distribution of PF00011 across species, with bacteria (green) eukaryota (purple), archaea (red) and unclassified sequences (blue).

from environmental samples, etc. - indicating good candidates for further studies.

We will go into details for the first case of PF00091, before reviewing the other protein families. Energies from both mfDCA and plmDCA have been compared to HMMer scores, giving similar results, but only mfDCA energies have been displayed in the following, for the sake of simplicity. In each case, both eukaryotes and bacteria have been treated as training alignments, yielding a similar outcome, but we will present here only the most significant results for each Pfam family.

2.2.1 *PF00091 - Tubulin/FtsZ family GTPase domain*

This family (Tubulin/FtsZ family, GTPase domain) includes the tubulin α, β and γ chains, as well as the bacterial FtsZ family of proteins, all involved in polymer formation. FtsZ is the polymer-forming protein of bacterial cell division, part of the ring formed in the middle of the dividing cell that is required for the constriction of the membrane. The discovery of bacterial tubulins (principal component of microtubules) was a surprise, and little is known about the structure of these proteins [75].

2.2.1.1 *Scoring*

The training set is the eukaryotic sub-family (mainly tubulin). The bacterial sub-family (test alignment) is composed of four groups of sequences differing from each other by their Uniprot annotations: "FtsZ cell division" proteins (in majority), bacterial "tubulins", "puta-

tive uncharacterized" proteins (mainly inferred from homology) and "deleted" proteins (mainly fragments or preliminary data deleted from the latest Uniprot release[5]). Fig. 2.5 displays the comparison between the mfDCA energies and HMMer scores of the bacterial sequences. The different bacterial groups have been colored in blue (FtsZ), red (tubulin), grey (putative) and black (deleted). The squares are two types of eukaryotic sequences: the dark green ones are directly taken from the training set and the light green ones are a random subset of eukaryotic sequences that were not included in the training set in the first place.

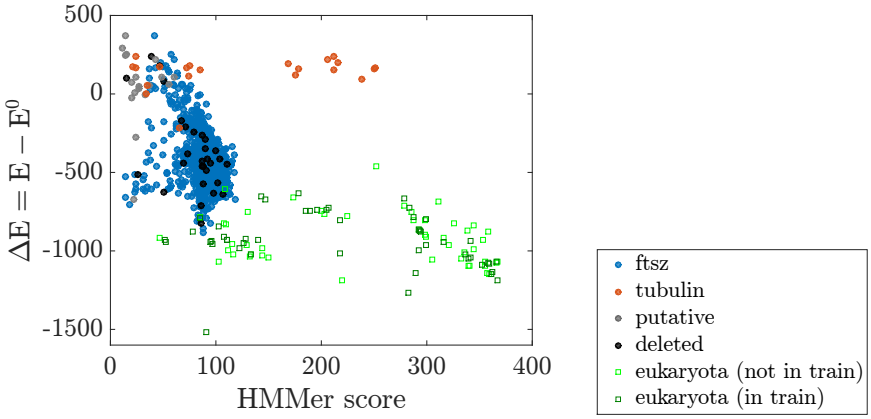


Figure 2.5 – PF00091 - Comparison between mfDCA energies (corrected by the null model) and HMMer scores. Training on the eukaryotic sub-family and testing on the bacterial sub-family. Eukaryotes (light and dark green squares) have lowest energies than bacteria (other colors).

What is particularly striking (and it will be the case for all the studied families), is the range of HMMer scores for the eukaryotic sequences (from the training set). Although the profile-HMM has been built on the training set, their HMMer scores are widely spread from about 20 (which usually is the minimum score for homology detection) to several hundreds. On the contrary, the DCA pairwise models give a very high scores to eukaryotic sequences, much more efficient than HMMer in pointing out sequences similar to the training sub-family. The two groups of eukaryotic sequences having approximately the same mfDCA energies, we see that the overfitting is limited here.

We also denote a very interesting feature: the bacterial tubulins (red) are divided into two groups of low and high HMMer scores. However, they are given similar scores by DCA. Too little is known about these bacterial tubulins to be able to investigate further and

5. The fact that Uniprot is updated more frequently than Pfam can be of interest for us, as the label of some sequences may change from one Uniprot version to the other, with some sequences being misclassified; this information can be used to see whether DCA models could have predicted these changes

verify these predictions in a structural point of view. Besides, low scores are assigned to the putative uncharacterized sequences (grey) by both models. The deleted proteins (black) are mainly drafts of FtsZ proteins, to which they have equivalent scores in both models.

2.2.1.2 *Structural information*

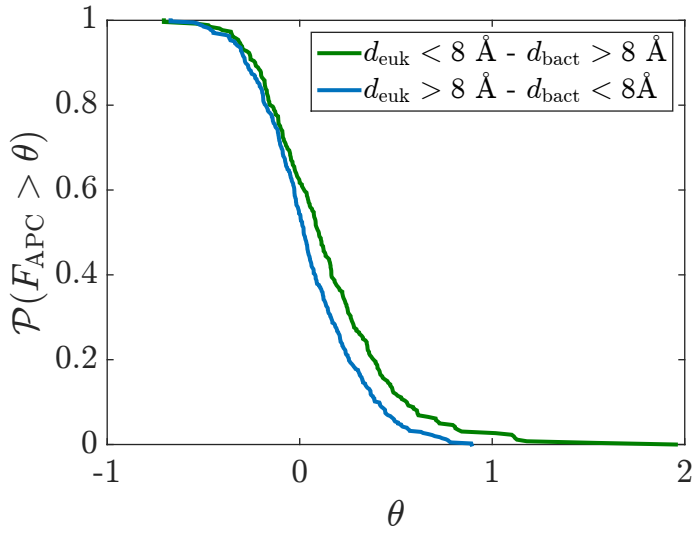


Figure 2.6 – Cumulative distribution function of APC scores for pairs in contact only in the eukaryotic structures (green), and only in the bacterial structures (blue).

The possibilities in linking scoring and structure informations are limited by the scarcity of sequences for which there are available structures. To unveil the structural differences between eukaryotes and bacteria (pointed out by DCA), we compare the average product correction (APC) scores of two group of residue pairs (i, j):

Standard Frobenius-based score used for contact prediction, cf. Eq. (3.8) in Part I

— the pairs that are contacts ONLY in the eukaryotic structure, but not in the bacterial one ($d_{ij}^{euk} < 8$ Å and $d_{ij}^{bact} > 8$ Å),
— the pairs that are contact ONLY in the bacterial structure, but not in the eukaryotic one ($d_{ij}^{euk} > 8$ Å and $d_{ij}^{bact} < 8$ Å).

The chosen structures have PDB[6] id 3CB2-A for the eukaryotic sequence TBG1_HUMAN (energy: -1002) and 2R75 for FtsZ bacterial sequence FTSZ_AQUAE (energy: -237). We would expect the APC scores to be higher for the first group, as the training has been done on the eukaryotic sub-alignment and eukaryotic sequences have been assigned lower DCA energies than bacterial sequences.

Fig. 2.6 displays the cumulative distribution function of the APC scores of the pairs that are in contact only in the eukaryotic structures (green curve) and the APC scores of the pairs that are in contact only

6. The Protein Data Bank (PDB) is a database of all known experimentally-determined structures of proteins.

in the bacterial structures (blue curve). We see that the cumulative distribution is systematically shifted in favor of the training set, which is consistent with the fact the the eukaryotic sequences have lower energies than the bacterial ones.

This preliminary result on the link between energy (scoring) and structure is to consider carefully as it was done on only two structures, and as it was not possible to generalize it further to all studied protein families. Besides, both distributions do remain strongly overlapping on Fig. 2.6. It means that residue pairs in contact only in the bacterial structure also display some coevolution in the eukaryotic structure, and may even be in contact in other eukaryotic structures (which is not easy to verify).

2.2.2 *More protein families*

We will present the results for Pfam families PF00011, PF00013, PF00027, PF00033, PF00664. In each case, the training has been done on both eukaryotic and bacterial sub-alignments, yielding similar results, but only the outcomes for a training on the bacterial sub-families[7] are displayed on Fig. 2.8b, 2.9a, 2.10, 2.11, 2.12, and 2.13. The same color code has been used in these figures for test sequences – eukaryotes (blue dots), archaea (red dots), viruses (purple dots) – and training sequences – bacteria present in training (dark green squares) and not present in training (light green squares). Some sequences are unclassified (black dots) for several reasons: they have been deleted in the current version of Uniprot as Uniprot is updated more frequently than Pfam (usually drafts or duplicates), or no information is known about their biological domain (usually metagenomes).

2.2.2.1 *Specificity to the training sub-family*

Confirming the interesting results obtained for PF00091, DCA pairwise models (corrected by the null model) always give lower energies to sequences in the same domain of life as the training sub-family, whereas HMMer scores are usually widely spread and not able to distinguish between training-like sequences and test sequences. This is independent of whether the sequences in the same domain of life as the training sub-family were actually in the training set or not, showing that the observed discrimination is independent from overfitting effects. On the contrary, without the null model, extremely low energies are given to sequences containing a lot of gaps, equivalent to the energies of the training sequences (*cf.* Fig. 2.8b & Fig. 2.9a for PF00011). These gap-rich sequences are usually given low scores by HMMer, due to gap penalties in the alignment procedure.

7. Some results for a training on the eukaryotic sub-families can be found in Appendix C.

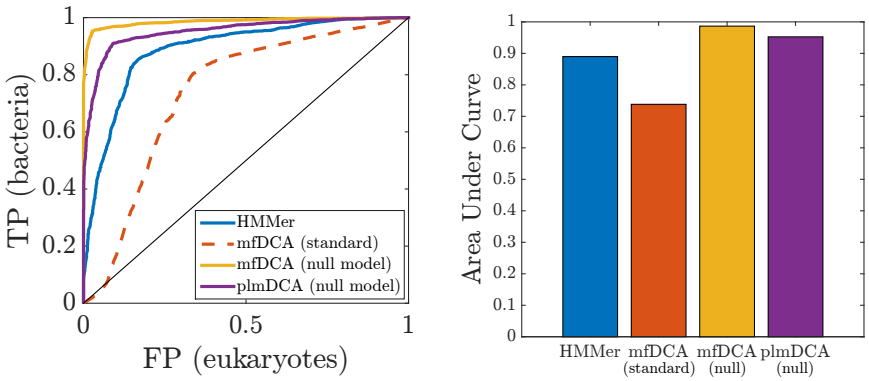


Figure 2.7 – Mean ROC (panel (a)) and AUC (panel (b)) curves – over the five studied Pfam families – illustrating the performance of the different models in discriminating between sequences, depending on whether they belong to the same domain of life than the training sub-family or not. Pairwise models are better classifiers if corrected by the null model on gaps.

To illustrate the properties of the different models to discriminate between the training and test sub-families, we consider the ROC curve (and the corresponding AUC), assessing their performance as binary classifiers. The two classes are in this case the bacteria (training sub-family) considered as true positives and the eukaryotes (test sub-family) considered as false positives [8]. A perfect classifier would give higher scores to all bacteria and therefore would be the constant function 1 in the ROC space; a random guess would go along the diagonal. Figure 2.7 displays the mean ROC and AUC over the five Pfam families studied in this section (PF00011, PF00013, PF00027, PF00033, PF00664) in four models (HMMer score, standard mfDCA, mfDCA corrected by the null model, plmDCA corrected by the null model).

As mentioned above, the energy in the standard mfDCA model is deeply affected by the presence of gaps in the sequence, performing no better than a random classifier for the first top-ranked sequences, and much worse than HMMer (see also Fig. 2.8b). The latter is outperformed by mfDCA and plmDCA when corrected by the null model on gaps. Interestingly, the mean-field approximation leads to a better discrimination between sequences than pseudolikelihood, probably because mfDCA typically tends to overestimate the inferred couplings [12]. The observed discrimination is consistent with the phylogenetic distribution and typically depends on whether they belong to the same domain of life than the training sub-family or not.

8. The ROC and AUC curves for a training on the eukaryotic sub-families can be found in Appendix C, displaying similar results.

2.2.2.2 *Detailed information about sequences*

More over, some eukaryotic or unclassified sequences have low DCA energies equivalent to the bacterial sub-family on which the training is done (see the figures below). Usually these eukaryotic sequences similar to bacteria according to DCA are putative or automatically labeled by homology softwares, meaning that there is no experimental proof and few indications that they really are eukaryotes. They often have been deleted in the last Uniprot version. The unclassified sequences similar to bacteria according to DCA usually are metagneomic sequences from environmental samples, or even misclassified as bacteria the last Uniprot version. Fortunately, the sequences that are pointed out by DCA are the same in the mean-field and the pseudolikelihood approximations, emphasizing the robustness of the procedure.

— **PF00011** (Figs. 2.9a & 2.8b): 4 eukaryotes (blue) and 8 unclassified (black) sequences have DCA energies equivalent to bacteria. All the eurkaryotes are automatically annotated and unreviewed (inferred from homology), which means that very little is known about these sequences. 3 of them have been deleted in the last Uniprot release. All of the 8 unknown sequences are unclassified metagenomic sequences from environmental samples. DCA seems to suggest that they are bacteria.

— **PF00013** (Fig. 2.10): 4 eukaryotes (blue) and 9 unclassified (black) sequences have DCA energies equivalent to bacteria. All the eukaryotes are unreviewed and inferred from homology, 2 of them are "putative uncharacterized", meaning that very little is known about these sequences. More interestingly, 2 of the unclassified sequences actually are labeled as bacteria in the last Uniprot release and the remaining 7 are unclassified metagenomic sequences from environmental samples, which could very well be bacteria.

— **PF00027** (Fig. 2.11): 7 eukaryotes (blue) and 7 unclassified sequences (black) have similar DCA energies than bacterial sequences. All of these eukaryotic sequences are unreviewed and inferred from homology, 1 has been deleted in the latest Uniprot release, 1 is putative uncharacterized. Besides, 5 of the unclassified sequences actually are bacteria and the remaining 2 are unclassified metagenomic sequences from environmental samples, possibly bacteria. HMMer is unable to detect these sequences.

— **PF00033** (Fig. 2.12): a group of high HMMer score eukaryotic sequences have a relatively high DCA energy. Unfortunately, their structure is unknown and we could not investigate further. We notice 15 unclassified sequences which have a low DCA ener-

gies: 1 is actually a bacterial protein and the 14 others are again metagenomic sequences from environmental samples.

— **PF00664** (Fig. 2.13): 72 unclassified sequences have a DCA energy as low as bacterial sequences: 15 are ecological metagenomes and 57 actually are labeled as bacteria in the last Uniprot release. HMMer is unable to detect these sequences.

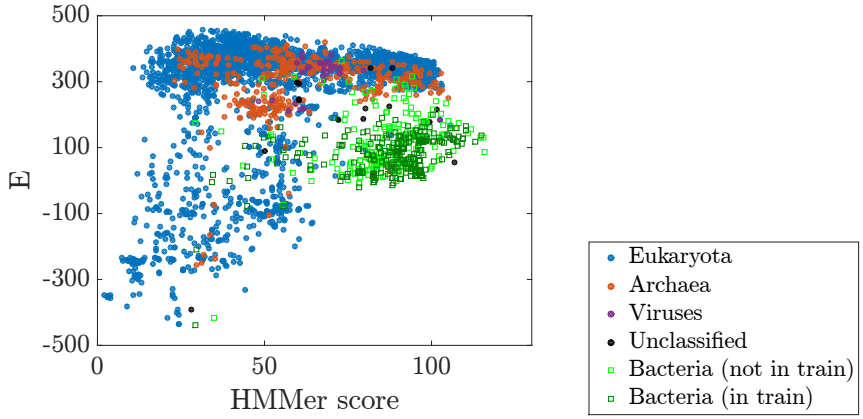


Figure 2.8 – PF00011 - Comparison between standard mfDCA energies and HMMer scores. Training on the bacterial sub-family and testing on the eukaryotic and archaea sub-families. Some eukaryotic (blue) and archaea (red) sequences are given extremely low energies, equivalent to energies of bacterial sequences (light and dark green squares): these sequences have low HMMer score and contain a lot of gaps, their energy is overestimated by mfDCA.

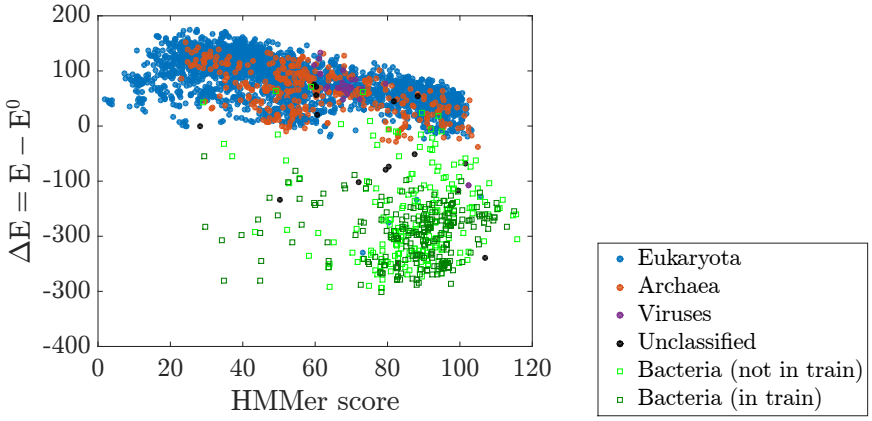


Figure 2.9 – PF00011 - Comparison between mfDCA energies corrected by the null model and HMMer scores. Training on the bacterial sub-family and testing on the eukaryotic and archaea sub-families. Bacteria (light and dark green squares) have lowest energies than eukaryota, archaea and viruses (other colors).

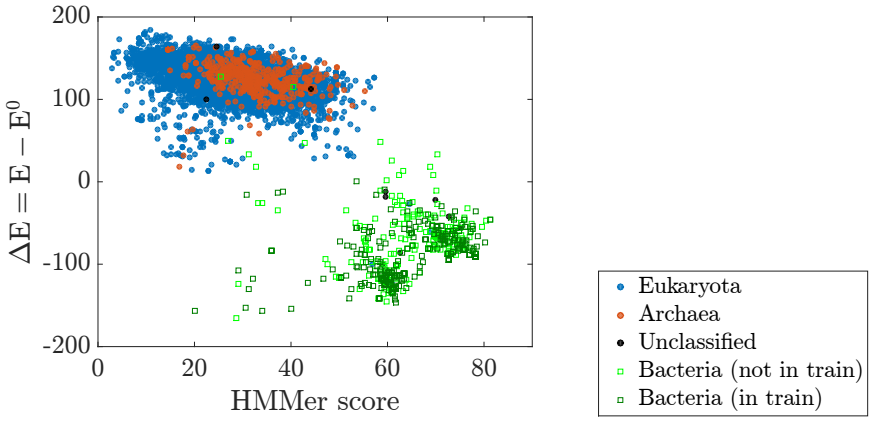


Figure 2.10 – PF00013 - Comparison between mfDCA energies (corrected by the null model) and HMMer scores. Training on the bacterial sub-family and testing on the eukaryotic and archaea sub-families. Bacteria (light and dark green squares) have lowest energies than eukaryota or archaea (other colors).

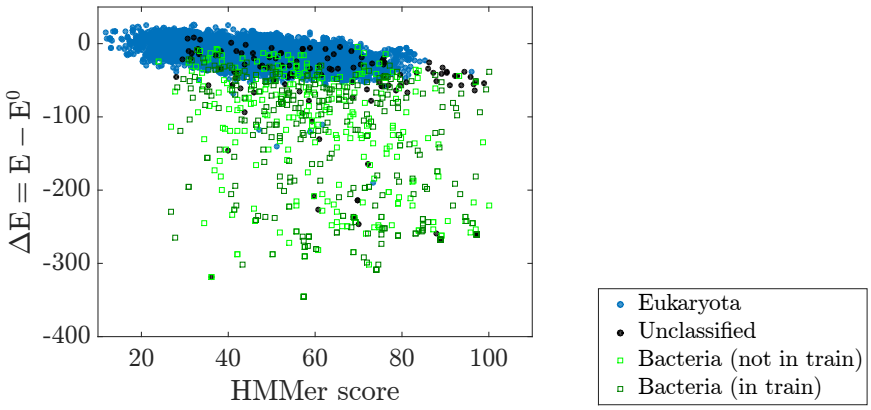


Figure 2.11 – PF00027 - Comparison between mfDCA energies (corrected by the null model) and HMMer scores. Training on the bacterial sub-family and testing on the eukaryotic and archaea sub-families. Bacteria (light and dark green squares) have lowest energies than eukaryotes (other colors).

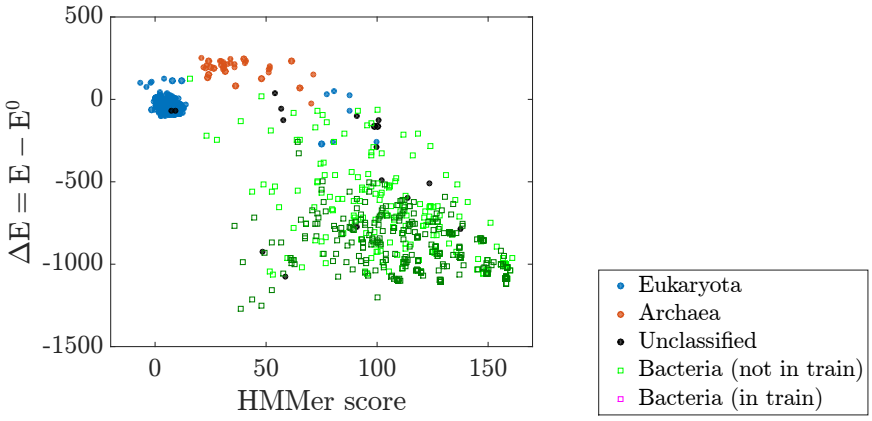


Figure 2.12 – PF00033 - Comparison between mfDCA energies (corrected by the null model) and HMMer scores. Training on the bacterial sub-family and testing on the eukaryotic and archaea sub-families. Bacteria (light and dark green squares) have lowest energies than eukaryota and archaea (other colors).

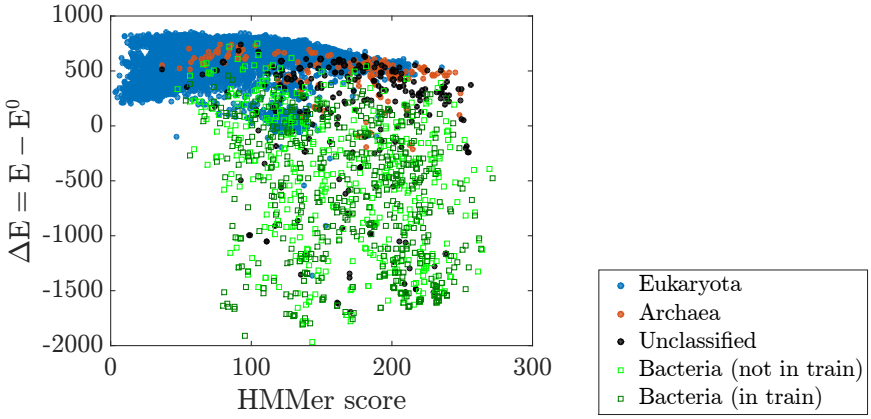


Figure 2.13 – PF00664 - Comparison between mfDCA energies (corrected by the null model) and HMMer scores. Training on the bacterial sub-family and testing on the eukaryotic and archaea sub-families. Bacteria (light and dark green squares) have lowest energies than eukaryotes (other colors).

2.3 OUTLOOK

DCA was originally developed to unveil structural information about proteins from sequence data alone, reaching a level of accuracy that was at the time thought to be beyond reach. It is only natural to use this successful framework in other ambitious fields such as fitness landscape modeling, folding prediction, or remote homology detection. In this chapter, we have made encouraging but preliminary steps in some of these challenging fields and showed for instance that DCA approaches were promising in predicting the folding ability of a given sequence, in a case where it was experimentally verified. More data of this kind will be available in the near future, leading to further exciting work by our team.

Remote homology detection is a hard problem, where no totally satisfactory approach has been developed yet in bioinformatics. Remote homologs indeed display lower sequence similarity and current tools such as profile-HMM only take into account conservation patterns. The idea that covariation patterns may be better preserved in remote homologs is what motivated the use of DCA approaches in the first place. Very surprisingly, it seems to answer a different question: DCA has a stronger tendency than HMMer to discriminate between sequences from the same family, where this discrimination is, *e.g.* consistent with the phylogenetic distribution. However, this tendency does not make it currently a tool for distant homology detection, but it seems to allow for a more detailed description of protein sequences, such as recognizing biological domains or pointing out interesting sequences for further studies. Having a more detailed picture of the different domains of life could also be very useful for phylogeny, although non trivial technical problems may arise. These results, although preliminary, are promising and encourage the use of DCA approaches beyond protein structural prediction.

However, the latter remains intrinsically topological - based on a ranking of couplings parameters, whereas fitness prediction or homology detection require global energetic considerations and therefore a much more detailed and quantitative statistical model. We have already encountered the problem of gaps, treated by DCA related approaches as an extra amino acid, despite strong evidence that they are intrinsically different. Another question raised in this context is the modeling of gaps as a missing information about the sequence. Although less appropriate for remote homology searches, treating missing data in the specific context of statistical inference on sequence samples is an interesting problem in itself. Besides, a quantitative understanding of the inferred Potts parameters is also lacking. We will address these two points in the next two parts of the present dissertation.

3

MODELING OF GAPS AS MISSING INFORMATION

As explained in the previous section, introducing an asymmetry between gaps on one hand and amino acids on the other hand allows for more accuracy in the DCA related approaches in the context of contact prediction [43] or sequence scoring (Chapter 2). The null model we have previously presented is a simple way of discarding strong contribution from gaps to the DCA energy of a sequence. Another solution is to consider gaps as missing information in the data. This approach is particularly relevant in amino acid sequences such as metagenomes [1], where long stretches of gaps reflect a high level of uncertainty on a specific region of the alignment. Related problems have already been addressed in the literature, but in the specific context of hidden nodes [15, 37], or phylogenetic trees reconstruction [70]. In this chapter, we will develop a general theoretical framework for dealing with random missing information in the observed samples, in the general context of Potts inverse problem within the mean-field approximation.

3.1 METHOD

The full derivation is in Appendix D

The probability of observing a sequence $\underline{a} = (a_1, ..., a_N)$ of length N with $a_i \in \{1, ..., q\}$ is (*cf.* Eq. (2.1) in Part I)

$$P(a_1, ..., a_N) = \frac{1}{\mathcal{Z}} \exp\left(\sum_{i=1}^{N} h_i(a_i) + \sum_{i=1}^{N-1}\sum_{j=1}^{N} J_{ij}(a_i, a_j)\right) . \tag{3.1}$$

Given a subset $\mathcal{K}$ of missing entries in sequence $\underline{a}$, the variables that are not observed are set to 0:

$$\forall k \in \mathcal{K} \qquad a_k = 0 . \tag{3.2}$$

The probability $\widetilde{P}(a_1, ..., a_N)$ of the sequence including missing entries (thus with $a_i \in \{0, ..., q\}$) is the marginal probability of the observed symbols:

$$\begin{aligned} \widetilde{P}(a_1, ..., a_N) &= \sum_{\{b_k = 1, ..., q \,|\, k \in \mathcal{K}\}} P(a_1, ..., b_k, ..., a_N) \\ &= \sum_{b_1, ..., b_N} \left(\prod_{\{j \,|\, a_j \neq 0\}} \delta_{a_j, b_j}\right) P(b_1, ..., b_N) . \end{aligned} \tag{3.3}$$

1. Genetic material from environmental samples, as opposed to traditional microbial genome sequencing from *in vitro* cultivated samples.

Given a multiple sequence alignment $A = \{a_i^\tau \mid i = 1...N, \tau = 1...B\}$ of B configurations of length N, where some of the entries are missing [2], the log-likelihood writes

$$\widetilde{\mathcal{L}}(\mathbf{J}, \mathbf{h} \mid A) = \frac{1}{B} \sum_{\tau=1}^{B} \log \widetilde{P}(a_1^\tau, ..., a_N^\tau) \,, \tag{3.4}$$

where $\mathbf{J} = \{J_{ij}(a, b)\}$, and $\mathbf{h} = \{h_i(a)\}$ denote the Potts parameters.

3.1.1 *Maximum-likelihood equations*

The Potts parameters are found by maximizing the log-likelihood:

$$\begin{cases} \dfrac{\partial \widetilde{\mathcal{L}}}{\partial h_k(c)} = 0 \,, \\ \dfrac{\partial \widetilde{\mathcal{L}}}{\partial J_{kl}(c,d)} = 0 \,. \end{cases} \tag{3.5}$$

After computation (*cf.* Appendix D for the full derivation), we get the following maximum-likelihood equations:

$$\begin{cases} BP_k(c) = \displaystyle\sum_{\tau=1}^{B} \delta_{a_k^\tau, c} + \sum_{\{\tau \mid a_k^\tau = 0\}} P(a_k^\tau = c \mid \{a_i^\tau \mid a_i^\tau \neq 0\}) \,, \\ \begin{aligned} BP_{kl}(c,d) = &\sum_{\tau=1}^{B} \delta_{a_k^\tau, c} \delta_{a_l^\tau, d} \\ &+ \sum_{\{\tau \mid a_k^\tau = 0\}} \delta_{a_l^\tau, d} P(a_k^\tau = c \mid \{a_i^\tau \mid a_i^\tau \neq 0\}) \\ &+ \sum_{\{\tau \mid a_l^\tau = 0\}} \delta_{a_k^\tau, c} P(a_l^\tau = d \mid \{a_i^\tau \mid a_i^\tau \neq 0\}) \\ &+ \sum_{\substack{\{\tau \mid a_k^\tau = 0, \\ a_l^\tau = 0\}}} P(a_k^\tau = c, a_l^\tau = d \mid \{a_i^\tau \mid a_i^\tau \neq 0\}) \,, \end{aligned} \end{cases} \tag{3.6}$$

where $c, d \neq 0$ ($c, d \in \{1, ..., q\}$) are observed states at sites k, l, and $P_k(c)$ and $P_{kl}(c, d)$ are the marginals of the probability distribution P.

Up to the first term on the right-hand side, both equations are similar the usual maximum-likelihood equations where marginals should match the frequency counts measured on the data. They contribute only if a_k^τ is observed in the first equation, and both a_k^τ and a_l^τ are observed in the second equation. The following terms account for the missing data and take the form of a combination of the observed part with averages over the observed background, taking the form of conditional probabilities on the observed symbols. In the second equations, three possible cases are taken into account: either a_l^τ is

2. Notice that each sequence has its own subset of missing entries. In other words, a variable can be observed in one sequence, but missing in another.

observed and a_k^τ is not (second term), or a_k^τ is observed and a_l^τ is not (third term), or both a_k^τ and a_l^τ are not observed (last term).

3.1.2 *Mean-field approximation*

Eqs. (3.6) are exact. Within the mean-field approximation at the first order in the couplings, the conditional probabilities are approximated through

$$\begin{aligned} P(a_k^\tau = c \,|\, \{a_i^\tau \mid a_i^\tau \neq 0\}) = & \frac{1}{\mathcal{Z}_k} \exp\left(h_k(c) + \sum_{\{i|a_i^\tau \neq 0\}} J_{ki}(c, a_i^\tau) \right. \\ & \left. + \sum_{\{i|a_i^\tau = 0\}} \sum_{\alpha=1}^{q} J_{ki}(c,\alpha) P_i(\alpha) \right), \end{aligned} \tag{3.7}$$

$$\begin{aligned} P(a_k^\tau = c, a_l^\tau = d \,|\{a_i^\tau \mid a_i^\tau \neq 0\}) = & \frac{1}{\mathcal{Z}_{kl}} \exp\left(h_k(c) + h_l(d) + J_{kl}(c,d) \right. \\ & + \sum_{\substack{\{i|a_i^\tau \neq 0\} \\ i \neq l}} J_{ki}(c, a_i^\tau) + \sum_{\substack{\{i|a_i^\tau = 0\} \\ i \neq l}} \sum_{\alpha=1}^{q} J_{ki}(c,\alpha) P_i(\alpha) \\ & \left. + \sum_{\substack{\{i|a_i^\tau \neq 0\} \\ i \neq k}} J_{il}(a_i^\tau, d) + \sum_{\substack{\{i|a_i^\tau = 0\} \\ i \neq k}} \sum_{\alpha=1}^{q} J_{il}(\alpha, d) P_i(\alpha) \right), \end{aligned} \tag{3.8}$$

where $\mathcal{Z}_k$ and $\mathcal{Z}_{kl}$ ensure that the conditional probabilities are normalized. One could have expected to find a conditional probability instead of $P_i(\alpha)$ in the right-hand side of Eqs. (3.7) & (3.8). This would have however included higher-order terms in the couplings, which have been neglected in the mean-field approximation (but could be included within the higher-order TAP approximation).

In extreme cases of high amount of missing data or low sampling (*cf.* Section 3.2), the convergence of the procedure may be compromised. In addition to the mean-field approximation, the two-point conditional probability can be approximated by the product of the one-point conditional probabilities:

$$\begin{aligned} P(a_k^\tau = c, a_l^\tau = d \,|\, \{a_i^\tau \mid a_i^\tau \neq 0\}) = & P(a_k^\tau = c \,|\, \{a_i^\tau \mid a_i^\tau \neq 0\}) \\ & \times P(a_l^\tau = d \,|\, \{a_i^\tau \mid a_i^\tau \neq 0\}) \,. \end{aligned} \tag{3.9}$$

Mean-field approaches indeed overestimate the coupling parameters, while this approximation tends to underestimate them. Couplings between missing states are actually not taken into account in Eq. (3.9), considering only couplings between missing and observed residues in a given sequence. One can therefore expect that approximation (3.9) makes the inference less constrained, improving the convergence properties of the whole procedure when put in difficulty by small sample sizes or large amount of missing data.

3.1.3 *Iterative Procedure*

In the following, we refer to the the mfDCA equations (*cf.* Eqs. (2.18) & (2.20)) in the q-gauge introduced in Part I Section 2.3.2.

3.1.3.1 *Initialization*

We initialize with $\mathbf{J}^{(0)} = 0$. The mfDCA equations in the q-gauge with zero couplings give $h_k^{(0)}(c) = \log \frac{f_k(c)}{f_k(q)}$, with $f_k(c)$ the single site frequency count computed from the MSA and normalized only on observed states: $\sum_{c=1}^{q} f_k(c) = 1$. In this case, Eqs. (3.6) write

$$\begin{cases} BP_k^{(0)}(c) = \sum_{\tau=1}^{B} \delta_{a_k^\tau,c} + \sum_{\{\tau|a_k^\tau=0\}} f_k(c) \ , \\ BP_{kl}^{(0)}(c,d) = \sum_{\tau=1}^{B} \delta_{a_k^\tau,c}\delta_{a_l^\tau,d} + \sum_{\substack{\{\tau|a_k^\tau=0,\\ a_l^\tau=0\}}} f_k(c)f_l(d) \\ \qquad\qquad + \sum_{\{\tau|a_l^\tau=0\}} \delta_{a_k^\tau,c} f_l(d) + \sum_{\{\tau|a_k^\tau=0\}} \delta_{a_l^\tau,d} f_k(c) \ . \end{cases} \tag{3.10}$$

Eq. (3.10) is equivalent to replacing the missing symbols by new ones with the frequencies of the considered sites, independently of the observed sequence background.

The couplings and fields $\left\{\mathbf{J}^{(1)}, \mathbf{h}^{(1)}\right\}$ are inferred with mfDCA from $\left\{P_k^{(0)}(c), P_{kl}^{(0)}(c,d)\right\}$:

$$\begin{cases} \dfrac{h_k^{(1)}(c)}{h_k^{(1)}(q)} = \log\left(\dfrac{P_k^{(0)}(c)}{P_k^{(0)}(q)}\right) - \sum_{b=1}^{q}\sum_{\substack{j=1\\ j\neq k}}^{L} J_{kj}^{(1)}(c,b)P_j^{(0)}(b) \\ J_{kl}^{(1)}(c,d) = -\left((\mathcal{C}^{(0)})^{-1}\right)_{kl}(c,d) \end{cases} \tag{3.11}$$

with $\mathcal{C}_{kl}^{(0)}(c,d) = P_{kl}^{(0)}(c,d) - P_k^{(0)}(c)P_l^{(0)}(d)$. As explained in Part I Section 2.2.2, the invertibility of the connected-correlation matrix is insured by fixing the Potts gauge (here, we chose the q-gauge).

3.1.3.2 *Iteration*

We use Eqs. (3.6), (3.7) & (3.8) to iterate the procedure. Given $\left\{\mathbf{J}^{(t)}, \mathbf{h}^{(t)}\right\}$ and $P_k^{(t-1)}(c)$, we compute $\left\{P_k^{(t)}(c), P_{k,l}^{(t)}(c,d)\right\}$:

$$\begin{cases} BP_k^{(t)}(c) = \sum_{\tau=1}^{B} \delta_{a_k^\tau,c} + \sum_{\{\tau|a_k^\tau=0\}} P^{(t-1)}(a_k^\tau = c \,|\, \{a_i^\tau \,|\, a_i^\tau \neq 0\}) \,, \\ BP_{kl}^{(t)}(c,d) = \sum_{\tau=1}^{B} \delta_{a_k^\tau,c}\delta_{a_l^\tau,d} \\ \qquad + \sum_{\substack{\{\tau|a_k^\tau=0,\\ a_l^\tau=0\}}} P^{(t-1)}(a_k^\tau = c, a_l^\tau = d \,|\, \{a_i^\tau \,|\, a_i^\tau \neq 0\}) \\ \qquad + \sum_{\{\tau|a_l^\tau=0\}} \delta_{a_k^\tau,c} P^{(t-1)}(a_l^\tau = d \,|\, \{a_i^\tau \,|\, a_i^\tau \neq 0\}) \\ \qquad + \sum_{\{\tau|a_k^\tau=0\}} \delta_{a_l^\tau,d} P^{(t-1)}(a_k^\tau = c \,|\, \{a_i^\tau \,|\, a_i^\tau \neq 0\}) \,. \end{cases} \tag{3.12}$$

A damping parameter ϵ is added to help the convergence of the algorithm:

$$\begin{cases} \widetilde{P}_k^{(t)} = (1-\epsilon)P_k^{(t-1)} + \epsilon P_k^{(t)} \,, \\ \widetilde{P}_{kl}^{(t)} = (1-\epsilon)P_{kl}^{(t-1)} + \epsilon P_{kl}^{(t)} \,. \end{cases} \tag{3.13}$$

The couplings and fields $\left\{J^{(t+1)}, h^{(t+1)}\right\}$ are inferred through mfDCA equations from $\left\{\widetilde{P}_k^{(t)}(c), \widetilde{P}_{kl}^{(t)}(c,d)\right\}$:

$$\begin{cases} \dfrac{h_k^{(t+1)}(c)}{h_k^{(t+1)}(q)} = \log\left(\dfrac{\widetilde{P}_k^{(t)}(c)}{\widetilde{P}_k^{(t)}(q)}\right) - \sum_{b=1}^{q}\sum_{\substack{j=1\\ j\neq k}}^{L} J_{kj}^{(t+1)}(c,b)\widetilde{P}_j^{(t)}(b) \,, \\ J_{kl}^{(t+1)}(c,d) = -\left((\widetilde{C}^{(t)})^{-1}\right)_{kl}(c,d) \,, \end{cases} \tag{3.14}$$

with $\widetilde{C}_{kl}^{(t)}(c,d) = \widetilde{P}_{kl}^{(t)}(c,d) - \widetilde{P}_k^{(t)}(c)\widetilde{P}_l^{(t)}(d)$.

3.2 CONVERGENCE AND RECOVERY OF THE POTTS PARAMETERS

We consider a Potts model with $q = 2$ states ($a_i \in \{1, 2\}$), where the network of interactions is described by an Erdős-Rényi random graph with $N = 30$ variables. Each edge in the interaction graph is included with probability 0.8. Field and coupling values for interacting pairs of sites $\left\{J^{true}, h^{true}\right\}$ are selected from a Gaussian distributions with mean $\mu = 0$ and standard deviation $\sigma_J = 0.05$. A MSA of $B = 10^6$ configurations $A = \{a_i^\tau \mid i = 1...N, \tau = 1...B\}$ is generated through Monte Carlo (MC) sampling. A Potts model with $q = 4$ states ($a_i \in \{1, ..., 4\}$) has also been studied and gives similar results, emphasizing the robustness of the procedure.

The missing data in the MSA is simulated given two types of random distributions:

— *uniform*: x =10%, 30% or 50% of the $B \times N$ entries of the MSA are set to 0 with a uniform distribution.
— *stretch*: fragmented sequences are simulated by stretches of length $l < 0.5N$ of 0 entries, generated at the beginning or the end of a given sequence, with l drawn from a Poisson distribution of mean $0.1N$, $0.3N$ or $0.5N$. $1/4$ of the sequences are left untouched, otherwise the stretch of length l is randomly set in 3 different ways: at the beginning of the sequence (from site 1 to site l), at the end (from site $N - l + 1$ to N), at both (from site 1 to $\lfloor l/2 \rfloor$ AND from site $N - \lfloor l/2 \rfloor + 1$ to N).

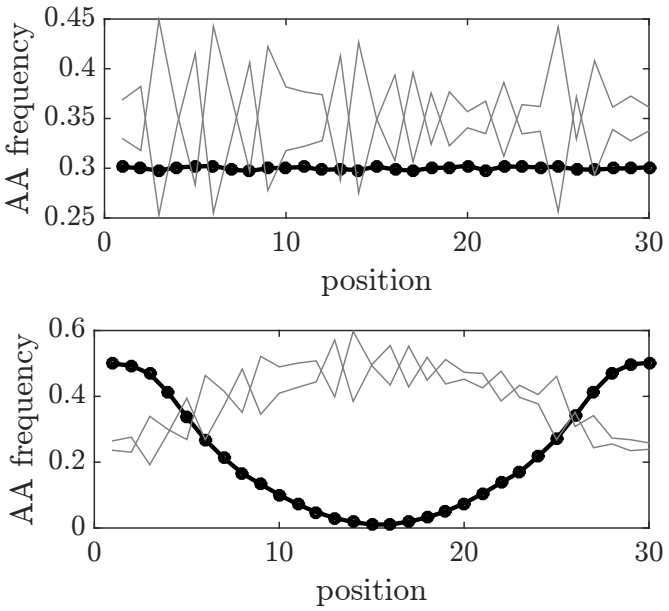


Figure 3.1 – Missing entries (black) and other symbols (grey) frequency in the *uniform* (top panel) and *stretch* (bottom panel) distributions of missing data along the $N = 30$ positions of the MSA, for x =30% and $l = 0.3N$ respectively. The *stretch* distribution roughly simulates gaps in protein MSAs.

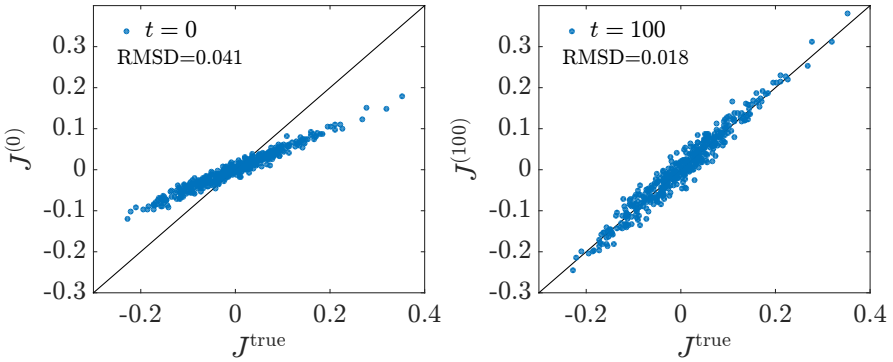
Fig. 3.1 displays the frequency of the 3 types of entries ($a_i^\tau \in \{0, 1, 2\}$) in the generated MSA for a realization of both distributions of missing data. The *stretch* distribution roughly simulates the effect of gaps in protein MSA, which tendency to come in long stretches at the beginning or the end of a sequence introduces dangerous artifacts in DCA related approaches (*cf.* Chapter 2).

3.2.1 *Effect of the amount of missing data*

We test the convergence and efficiency of the algorithm to recover the underlying parameters of the random graph on a MSA of $B = 10^5$ sequences, depending on the amount of missing data: $x \in \{10, 30, 50\}$ and $l \in \{0.1N, 0.3N, 0.5N\}$. Convergence is achieved when a plateau

is reached in the correlation (or RMSD) between the Potts parameters inferred through the procedure and the true parameters. Reducing the damping parameter ϵ (*cf.* Eq. (3.13)) prevents from oscillations in the iterative procedure.

Interestingly, the initial step of the algorithm - replacing the missing data by amino acids with their frequency on the considered site - gives rise to an underestimation of the inferred couplings. This replacement of missing data indeed suppresses correlations between interacting pairs of sites, leading to smaller couplings. It corresponds to the slope (regression coefficient) < 1 at $t = 0$ on Fig. 3.2a.



(a) (b)

Figure 3.2 – *Uniform* distribution of missing data with x =30% - Panel (a): inferred *vs.* true couplings at $t = 0$. Panel (b): inferred *vs.* true couplings at $t = 100$. Convergence is achieved in about 100 iterations and the Potts parameters are accurately recovered through the procedure.

In most cases, convergence is achieved in less than 150 iterations with the Potts parameters being accurately recovered by the procedure (*cf.* Fig. 3.2b). Tables 2 & 3 display the mean Pearson correlation and slope between true and inferred couplings over 10 realizations of both distributions of missing data. The more data is missing, the lower the initial slope ($t = 0$ in Tables 2 & 3), with for instance a mean slope going from 0.814 to 0.251 for 10% to 50% of missing entries in the MSA.

Convergence could not be achieved for x =50% of missing data in the *uniform* distribution, even for small damping parameters ϵ. Better convergence properties are obtained with the *stretch* distribution of missing data because its nature is different: the effective amount of missing entries is lower due to the Poissonian distribution of the stretch length, and one fourth of the configurations are not altered, leading to a better quality of inference.

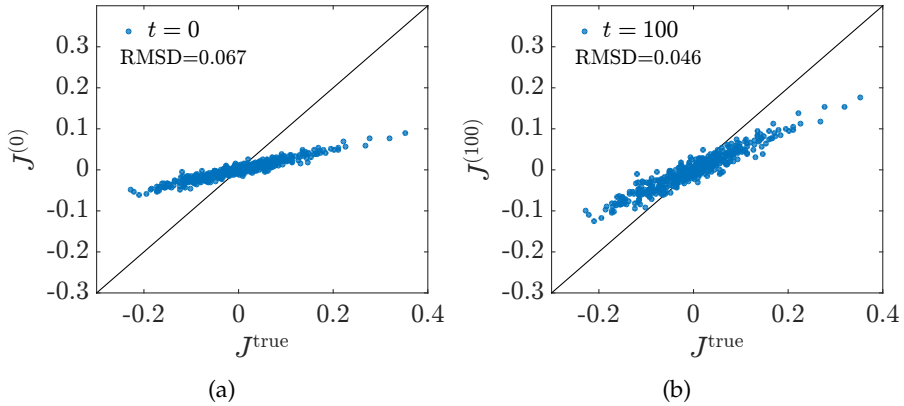
In the case of x =50% of missing data in the *uniform* distribution, approximating the two-point conditional probability by the product of the one-point conditional probabilities (Eq. (3.9)) allows for better

		x =10%	x =30%	x =50%
$t = 0$	slope	0.814	0.490	0.251
$t = t_f$	slope	1.01	1.02	0.505
	Pearson	0.988	0.977	0.961

Table 2 – Mean Pearson correlation coefficients and slope between true and inferred couplings over 10 realizations of the *uniform* distribution of missing data. Coefficient are given at initialization ($t = 0$) and convergence ($t = t_f$), for $B = 10^5$ configurations, depending on the percentage x of 0 entries in the MSA. Red numbers indicate that convergence could not be achieved, and that approximation (3.9) was used.

		$l = 0.1N$	$l = 0.3N$	$l = 0.5N$
$t = 0$	slope	0.871	0.621	0.487
$t = t_f$	slope	1.01	1.02	1.02
	Pearson	0.988	0.979	0.964

Table 3 – Mean Pearson correlation coefficients and slopes between true and inferred couplings over 10 realizations of the *stretch* distribution of missing data. Coefficient are given at initialization ($t = 0$) and convergence ($t = t_f$), for $B = 10^5$ configurations, depending on the mean length of stretches l in the MSA.



(a) (b)

Figure 3.3 – *Uniform* distribution of missing data with x =50% and approximation (3.9) - Panel (a): inferred *vs.* true couplings at $t = 0$. Panel (b): inferred *vs.* true couplings at $t = 100$. Convergence is achieved, but there is still an underestimation of the inferred couplings.

convergence properties (see red numbers in Table 2). As can be expected, this new approximation leads to an underestimation of the inferred couplings (slope < 1) even if convergence is achieved – the Pearson correlation between true and inferred couplings has reached a plateau – as displayed on Fig. 3.3.

3.2.2 *Effect of the sampling*

		$l = 0.1N$	$l = 0.3N$	$l = 0.5N$
$B = 50000$	slope	1.02	1.03	1.06
	Pearson	0.974	0.961	0.934
$B = 10000$	slope	1.02	1.04	0.910
	Pearson	0.898	0.855	0.805
$B = 1000$	slope	1.06	0.8433	0.8360
	Pearson	0.529	0.415	0.395

Table 4 – Mean Pearson correlation coefficients and slopes over 10 realizations of the *stretch* distribution of missing data, depending on the mean length of stretches l in the MSA and the number of configurations B. Red numbers indicate that the algorithm failed to converge, and that approximation (3.9) was used.

We consider the influence of the number of configurations B in the MSA on the convergence of the algorithm and its efficiency in recovering the true Potts parameters of the random graph. As displayed on Table 4 (for the *stretch* distribution of missing data), the effect is twofold: the accuracy in recovering the underlying Potts parameters of the graph strongly decreases as the sampling drops, and convergence cannot be achieved - no matter how small is the damping parameter - for large values of l and small sample sizes. However, approximation (3.9) allows for better convergence properties in these extreme cases. As the sample size drops, the slope between real and inferred couplings slightly increases, reaching values superior to 1 (except when approximation (3.9) is used, see the previous section). mfDCA indeed tends to overestimate the couplings, especially in poor sampling cases [12].

3.3 SEQUENCE ENERGIES ARE ACCURATELY REPRODUCED

Standard DCA approaches consider gaps as an extra symbol. Therefore, the standard model displays $q + 1$ states, contrary to the present method with only q states. The size of the inferred coupling matrices

being different, it is not possible to directly compare both procedures on their accuracy to recover the model parameters. However, some comparison can be obtained through the energy function which assigns a scalar score to a configuration.

We therefore consider a *test* alignment, in which some data may also be missing (or, equivalently, gaps may be present). In the present framework, the energy of a sequence in which some sites are not observed is the average energy over all sequences compatible with the observation, and weighted by the Boltzmann distribution. It is therefore the best-educated guess possible, considering only a partial observation.

The energy of sequence $\underline{a}$ with missing entries ($\exists \mathcal{K}, \forall k \in \mathcal{K}, a_k = 0$) therefore reads

$$\begin{aligned}\widetilde{E}(a_1, ..., a_L) = &\sum_{\{i|a_i\neq 0\}} h_i(a_i) + \sum_{\substack{\{i,j|a_i\neq 0,\\ a_j\neq 0\}}} J_{ij}(a_i, a_j) \\ &+ \sum_{\{i|a_i=0\}} \sum_{\alpha=1}^{q} h_i(\alpha) P(a_i = \alpha \,|\, \{a_l \mid a_l \neq 0\}) \\ &+ \sum_{\substack{\{i,j|a_i=0,\\ a_j\neq 0\}}} \sum_{\alpha=1}^{q} J_{ij}(\alpha, a_j) P(a_i = \alpha \,|\, \{a_l \mid a_l \neq 0\}) \\ &+ \sum_{\substack{\{i,j|a_i\neq 0,\\ a_j=0\}}} \sum_{\beta=1}^{q} J_{ij}(a_i, \beta) P(a_j = \beta \,|\, \{a_l \mid a_l \neq 0\}) \\ &+ \sum_{\substack{\{i,j|a_i=0,\\ a_j=0\}}} \sum_{\alpha,\beta=1}^{q} J_{ij}(\alpha, \beta) P(a_i = \alpha, a_j = \beta \,|\, \{a_l \mid a_l \neq 0\}) \,,\end{aligned} \tag{3.15}$$

with the conditional probabilities given by Eqs. (3.7) & (3.8) within the mean-field approximation. Again, the coupling term takes into account the four possible cases: either both a_i and a_j are observed, or a_j is observed and a_i is not, or a_i is observed and a_j is not, or both a_i and a_j are not observed.

In the following, we will call *training* alignment the MSA with $B = 10^5$ sequences introduced in the last section, on which the iterative procedure is applied.

3.3.1 *Real energies*

We consider a *test* alignment of $B = 4000$ sequences including missing data generated by the *stretch* distribution with $l \in \{0.1N, 0.3N, 0.5N\}$. Two energies relative to the true underlying model are compared on Fig. 3.4:

— E^{true}, the true energy (with true couplings and fields) of the full sequences;
— $\widetilde{E}^{true}$, the true energy (with true couplings and fields) of the same sequences, but including missing entries (given by Eq. (3.15)).

Naturally, the more data is missing, the less accuracy there is in recovering the true energies, with Pearson correlation from 0.994 for less than 10% of non observed entries in the sequence, to 0.158 for more than 50%. This gives an indication about the precision that can be expected in the optimal case (true Potts parameters), which is limited by the amount of missing data in the sequence.

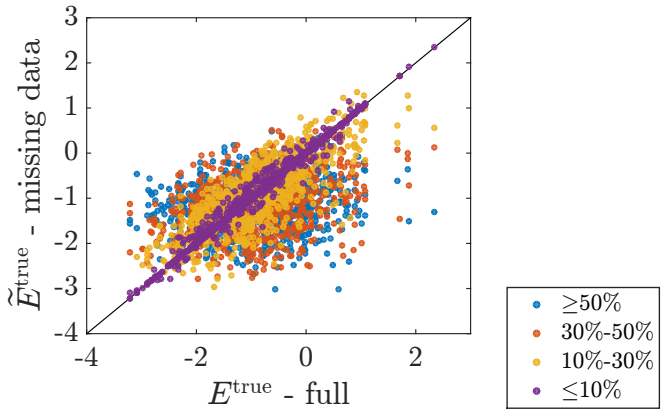


Figure 3.4 – True energy of full sequences *vs.* sequences with missing data. Different colors indicate the amount of missing data in the sequence, generated with the *stretch* distribution. Pearson correlation coefficients are successively 0.994, 0.781, 0.325, 0.158 for less than 10% to more than 50% of gaps in the sequence.

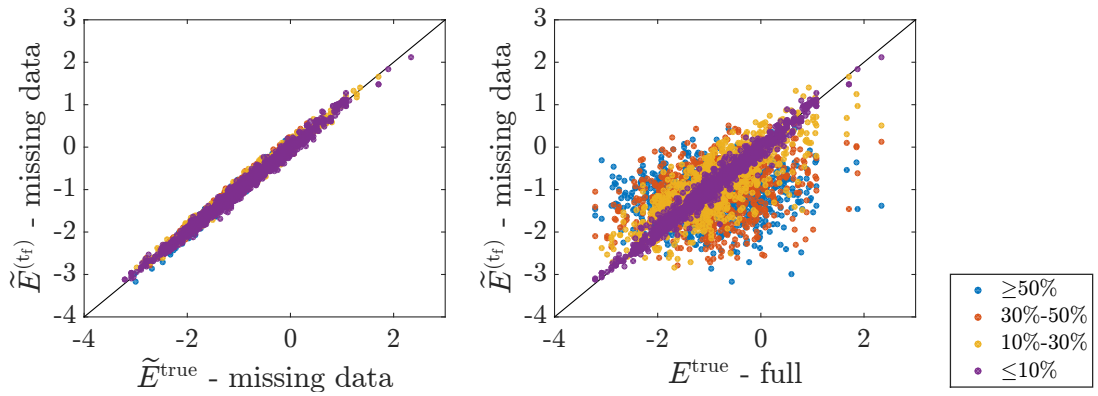
3.3.2 *Inferred energies*

Three energies are compared on Fig. 3.5:
— E^{true} (ideal case),
— $\widetilde{E}^{true}$ (best expected case),
— $\widetilde{E}^{(t_f)}$, the inferred energy of the sequences with missing data.

The energy $\widetilde{E}^{(t_f)}$ is computed with the inferred couplings and fields $\left\{\mathbf{J}^{(t_f)}, \mathbf{h}^{(t_f)}\right\}$ obtained after convergence ($t = t_f$) of the iterative procedure on a *training* alignment including missing data. The *test* alignment has been described in Section 3.3.1.

Fig. 3.5 displays the comparison between true and inferred energies. The Potts couplings have been inferred from a *training* alignment with missing data generated by the *stretch* distribution with $l = 0.3N$. Compared on the same *test* alignment with missing data, $\widetilde{E}^{true}$ and $\widetilde{E}^{(t_f)}$ are very well correlated (panel (a)); whereas the correlation between the true energies of the full sequences E^{true} and the inferred energy on sequences with missing data $\widetilde{E}^{(t_f)}$ naturally

depends on the amount of uncertainty (panel (b), very similarly to Fig. 3.4). Pearson correlation coefficient are of 0.987 for less than 10% of non observed entries in the sequence, to 0.156 for more than 50%.



(a) (b)

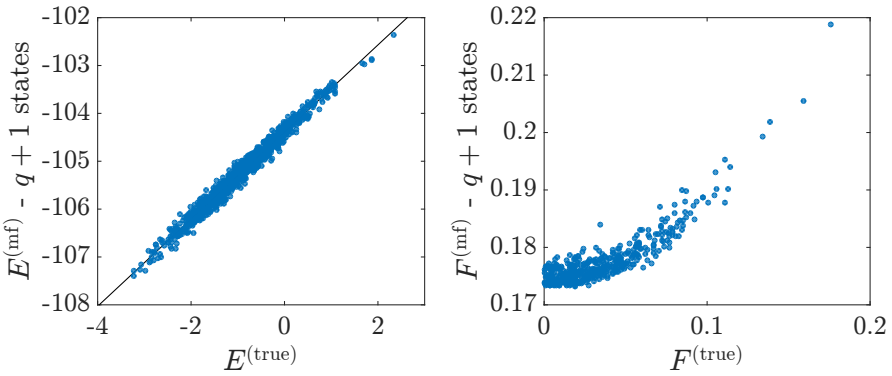
Figure 3.5 – Panel (a): true energies of full sequences *vs.* inferred energies of sequences with missing data (comparison with the ideal case). Panel (b): true *vs.* inferred energies of sequences with missing data (comparison with the best expected case). Different colors indicate the amount of missing data in the sequence – generated by the *Stretch* distribution, both in the *training* ($l = 0.3N$) and *test* ($l \in \{0.1N, 0.3N, 0.5N\}$) alignments. Pearson correlation coefficients are successively 0.987, 0.778, 0.316, 0.156 for less than 10% to more than 50% of gaps in the sequence.

3.4 COMPARISON WITH STANDARD DIRECT-COUPLING ANALYSIS

As stated above, this is not possible to compare directly the true Potts parameters or the inferred Potts parameters obtained after convergence of the iterative procedure on one hand, with the standard mfDCA parameters on the other hand. The latter indeed correspond to $q+1$ Potts states (gap is an extra amino acid) leading to $(q+1)\times(q+1)$-sized coupling matrices and $(q+1)$-sized field vectors, whereas the former describe q Potts states (gap is missing data) with $q \times q$-sized couplings matrices and q-sized field vectors. We can however compare the energies or Frobenius norms in both models.

3.4.1 *Absence of missing data*

Without any missing data (gaps) in the *training* or *test* alignments, the true energies and the energies inferred with $q+1$-states mfDCA (state $q+1$ is therefore *never* observed in both alignments), although very different in absolute value, are very well correlated with a slope of 0.91 and a Pearson coefficient of 0.989 (Fig. 3.6a). On the other hand, very small true couplings give rise to poorly sampled configu-



(a) (b)

Figure 3.6 – Panel (a): true *vs.* standard mfDCA energies of the *test* alignment with missing entries. Panel (b): Frobenius norm of the true *vs.* mfDCA inferred couplings.

rations and are difficult to infer. Strong pseudocounts are therefore used by standard DCA related approaches to ensure that the corresponding inferred parameters are not infinite. This explains why the small Frobenius norms are badly estimated by mfDCA, while there is a good correlation for larger Frobenius norms (*cf.* Fig. 3.6b). In any case, mfDCA gives very satisfactory results without missing entries on both the *training* and *test* alignments. This is of course not the case when gaps are present, as it has been extensively discussed in Chapter 2.

3.4.2 *Presence of missing data*

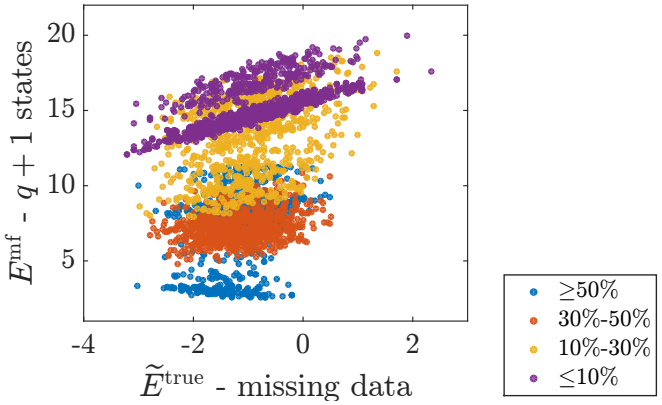


Figure 3.7 – True *vs.* mfDCA energies of sequences. Different colors indicate the amount of missing data (gaps) in the sequence – generated by the *Stretch* distribution, both in the *training* ($l = 0.3N$) and *test* ($l \in \{0.1N, 0.3N, 0.5N\}$).

Fig. 3.7 displays the comparison between true and standard mfDCA energies on a *test* alignment with missing entries (gaps). The mfDCA fields and couplings have been inferred on the same *training* alignment than the iterative procedure in last section, but the missing data is regarded here as an extra amino acid. The correlation between real and mfDCA energies strongly depends on the amount of missing data (gaps), with very low energies for sequences rich in gaps (consistently with Chapter 2).

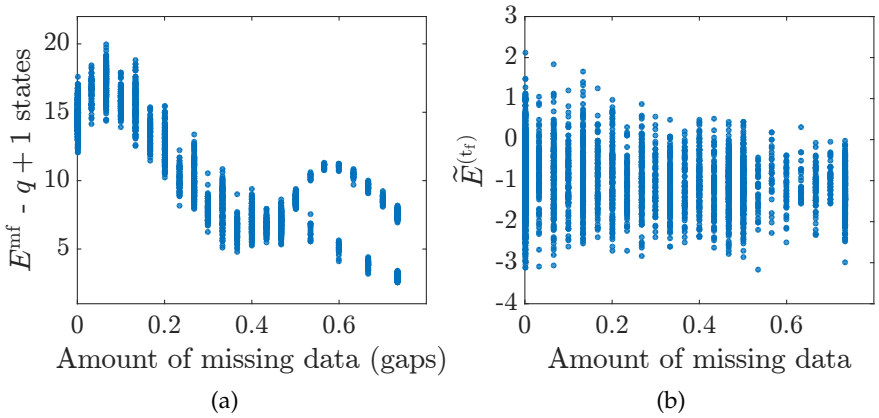


Figure 3.8 – Inferred energy *vs.* amount of missing data (gaps), for mfDCA (panel (a)) and the iterative procedure (panel (b)). The missing data has been generated by the *stretch* distribution, both in the *training* ($l = 0.3N$) and *test* ($l \in \{0.1N, 0.3N, 0.5N\}$).

Similarly to what has been observed in Chapter 2 on real data, mfDCA energies are highly correlated to the amount of gaps in the sequence (*cf.* Fig. 3.8a). This is not the case for energies computed with the iterative procedure, as displayed on Fig. 3.8b.

3.5 OUTLOOK

This general framework for dealing with missing data in the context of the inverse Potts problem within the mean-field approximation proves to be very promising. To the best of our knowledge, this specific problem has never been addressed in the literature. True couplings of an underlying random graph are very accurately recovered in the large sample size situation, even for large amounts of missing data. The recovery of the true parameters of course depends on the sample size, but approximating the two-point conditional probabilities by the product of the one- point conditional probabilities improves the convergence properties of the algorithm for smaller sample sizes and larger amount of missing data. Consistently, the energies of sequences from a *test* MSA also including missing entries are well reproduced, within the expected precision due to the uncertainty, and much better than with the standard mfDCA model.

The next step would be to apply this method to real biological sequences, such as metagenomic sequences which are often fragments with stretches of gaps indicating a high level of uncertainty on a specific region of the MSA. The application to remote homology search is maybe less relevant as the procedure tends to replace gaps according to correlation patterns observed in the data.

Part III

DIRECT COUPLINGS REFLECT BIOPHYSICAL RESIDUE INTERACTIONS

To achieve residue-residue contact prediction, the Potts parameters inferred with direct-coupling analysis (21×21 matrices accounting for direct couplings between the 20 amino acids and the alignment gap) are mapped onto simple scalar parameters and subsequently ranked. The full information they potentially contain gets lost. In this part, we provide a quantitative understanding of the inferred couplings and show they contain detailed and interpretable information about the physico-chemical properties of the amino acids in contact. Our results are based on the analysis of 70 protein families (Chapter 2). We furthermore consider abstract lattice-protein models to better understand the crucial role of sampling on the results (Chapter 3).

1

INTRODUCTORY REMARKS

1.1 MOTIVATIONS

Structural prediction has improved considerably in the last recent years – reaching a precision that was at the time thought to be beyond reach – thanks to direct-coupling analysis (DCA) related approaches today widely used in the field. As explained in Part I Chapter 3, contact prediction is performed by mapping DCA coupling matrices onto simple scalar parameters – average product correction (APC) or direct-information (DI) scores – and subsequently ranking them. Although they are inferred at high computational cost, the full information these 21×21 matrices potentially contain gets lost. Moreover, a better understanding of the Potts parameters is paramount if DCA is to be applied beyond contact prediction to new challenging fields requiring more quantitative considerations and sensitive to the details of the coupling matrices.

For each residue pair (i, j), these inferred matrices have positive and negative entries corresponding to the coupling between a pair (a, b) of amino acids (or Potts states). The interpretation of these entries somewhat depends on the gauge in which the couplings have been inferred; a schematic interpretation in the zero-sum gauge – used in the following, with the sum of the lines and columns of each coupling matrix set to 0 – is that large positive entries indicate that amino acids a, b are favored at positions i, j, and alternatively, large negative entries show that the pair of amino acids is avoided.

See Part I Chapter 2.3.1 Eq. (2.29)

The aim this work is to provide a better quantitative understanding of these inferred couplings. Earlier works have shown that the coevolutionary couplings derived by DCA contain an electrostatic signal [93]. Here, we go considerably further and show that the coevolutionary couplings also contain quantitative and interpretable biological information related to all the physico-chemical properties of amino-acid interactions. These interactions are consistent with knowledge-based amino-acid potentials inferred from known protein structures, such as the statistical potential derived by Miyazawa and Jernigan [79], which will be described below. Other statistical potentials, more recently introduced, may perform better in several tasks, such as describing energetically the native folded states [100], but an extensive comparison between the coupling matrices and scoring potentials is beyond the scope of this dissertation.

Most of the results of this part have been very recently submitted to *"Direct coevolutionary couplings reflect biophysical residue interactions*

in proteins", A Coucke, G Uguzzoni, F Oteri, S Cocco, R Monasson, and M Weigt, Journal of Chemical Physics (2016), [31], currently under review.

1.2 MIYAZAWA-JERNIGAN STATISTICAL POTENTIAL

Developed from the 1980s, the Miyazawa-Jernigan (MJ) knowledge-based potential $E^{MJ}(a,b)$ was derived from the statistics of amino acids in contact in known 3D protein structures. This 20×20 interaction matrix reflects the physico-chemical properties of the amino acids (*cf.* Fig. 3.1 in Part I), torsions angles, solvent exposure and hydrogen bonds geometry [79], finally compressed in a 20×20 interaction matrix between pairs of amino acids.

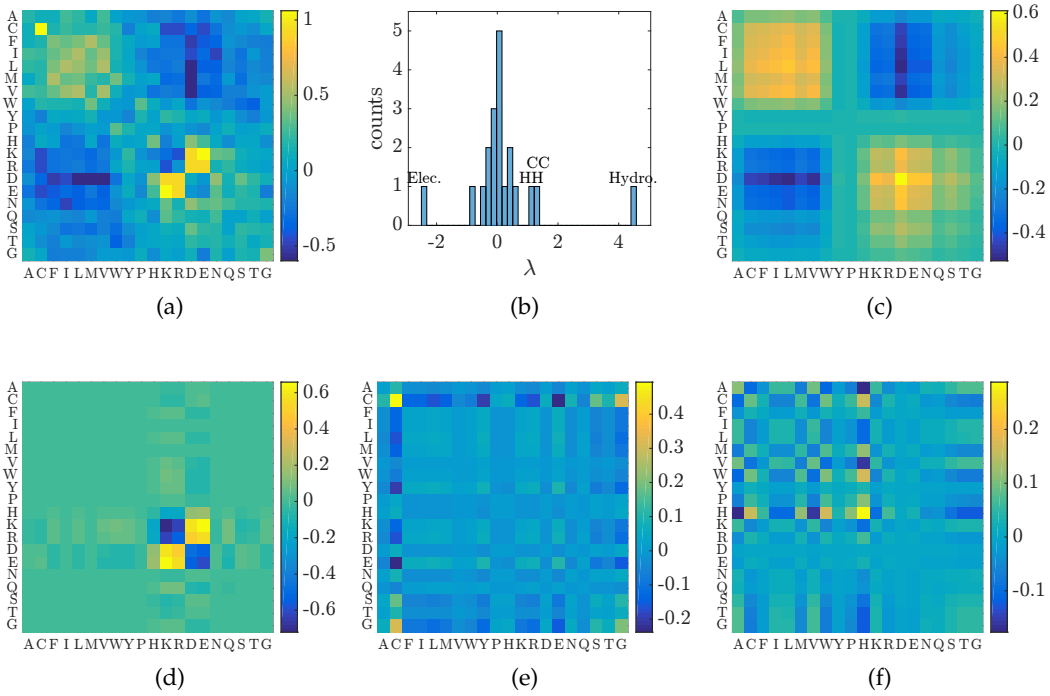


Figure 1.1 – (a) MJ energy matrix $E_0^{MJ}(a,b)$. (b) Spectrum of the MJ matrix dominated by several eigenvalues. MJ's 3 largest spectral modes, displaying physico-chemical interactions: (c) hydrophobicity-hydrophilicity ($\lambda^{(1)} = 4.55$), (d) electrostaticity ($\lambda^{(2)} = -3.51$), (e) Cysteine-Cysteine ($\lambda^{(3)} = 1.28$), and (f) Histidine-Histidine ($\lambda^{(4)} = 1.04$) signals.

In contrast to more detailed potentials including also, *e.g.*, the residue distance, the MJ interaction matrix is a natural starting point for comparison with the DCA-derived coupling matrices. Panel (a) of Fig. 1.1 displays $E_0^{MJ}(a,b)$, the 20×20 matrix provided by Miyazawa and Jernigan in 1996 [80], upon transformation into zero-sum gauge (*cf.* Part I Section 2.3.1), to compare with DCA couplings later on. It has also been multiplied by a factor -1 to comply with the standard

convention that attractive interactions are positive, and repulsive ones are negative:

$$E_0^{MJ}(a,b) = -\left(E^{MJ}(a,b) - E^{MJ}(\cdot,b) - E^{MJ}(a,\cdot) + E^{MJ}(\cdot,\cdot)\right), \quad (1.1)$$

where $g(\cdot)$ denotes the uniform average of $g(a)$ over all 20 amino acids a at fixed position (*cf.* Eq. (2.30) in Part I). In this specific gauge [1], the spectrum of the MJ matrix shows several significant eigenvalues (Fig. 1.1 panel (b)).

Panels (c) to (f) display the first spectral projections of the MJ matrix ($M^{(k)}(a,b) = \lambda^{(k)} v_a^{(k)} v_b^{(k)}$, $k = 1...4$, *cf.* Eq. (2.3) below). They are localized on particular amino acids according to physico-chemical interactions. Panel (c) is related to hydrophobicity/hydrophilicity: amino acids from A to P are hydrophobic, whereas the rest are hydrophilic. Hydrophobic amino acids tend to form contacts with other hydrophobic amino acids but not with hydrophilic ones, according to the signs of the corresponding entries. Panel (d) is related to electrostaticity: amino acids K, R and H are positively charged whereas D and E are negatively charged. A contact between amino acid of the same charge is very unlikely. Panel (e) is localized on the Cysteine-Cysteine entry, as those amino acids tend to form strong chemical disulfide bounds where paired with each other. Finally, panel (f) shows the fourth spectral mode of the MJ matrix, localized on the Histidine-Histidine entry, often forming like-charged contact pairs [56].

The eigenvalues corresponding to hydrophobicity/hydrophilicity ($\lambda^{(1)} = 4.55$), the Cysteine-Cysteine ($\lambda^{(3)} = 1.28$) and Histidine-Histidine interactions ($\lambda^{(4)} = 1.04$) are positive, describing an attractive interaction between like amino acids. On the other hand, the eigenvalue corresponding to electrostaticity ($\lambda^{(2)} = -3.51$) is negative, reflecting the attraction between charges of opposite sign, and repulsion between like charges (antiferromagnetic-like interaction).

1. The original MJ statistical potential – without gauge transformation – is however completely dominated by the hydrophobic eigenmode [121].

2

PROTEIN SEQUENCES DATA

In this chapter, we will consider a set of 70 protein families, from which we infer the coupling matrices with the pseudolikelihood approximation of direct-coupling analysis (plmDCA). After selecting the top ranked residue pairs for each family, we analyze the mean coupling matrix and its spectral modes. Considering structural classifications and solvent exposure helps unveiling the full biological content of the coupling matrices $\{J_{ij}(a,b)\}_{a,b\in\{1,\dots,21\}}$. Our analysis also shows that the distribution of contact distances in the tertiary structure greatly depends on the type of interaction associated to the contact.

2.1 METHOD

2.1.1 *Dataset*

We consider a random set of 70 protein families from the Pfam database [49] satisfying the following criteria: (i) the selected families contain enough sequences ($B_{eff} > 500$) to guarantee a good inference (sufficient sampling), and (ii) possess at least one X-ray crystal structure of resolution below 3 Å in the Protein Data Bank (PDB) [17]. This enables to extract experimental contact maps and to use the first level of SCOP (structural) categorization [84] of PDB structures, *the Class*, that account for the types of folds (*e.g.*, beta sheets). (iii) Every PDB chain that contains a selected domain family has been classified into a unique structural group according to SCOP; (iv) the families are selected to cover a broad range in protein length and to have good sensitivity in the contact prediction. The complete list can be found in Appendix E.

I worked with G. Uguzzoni from our team at LCQB on the selection

Here, a residue pair is considered to be a contact in the tertiary structure if its minimal heavy-atom distance is below 6 Å in the protein structure. A mapping application was developed to map domain family alignments to crystal structures and to extract distances of residue pairs in PDB structures in order to obtain the contact map. The 6 Å threshold is chosen consistently with prior studies [82]. We take into account several crystal structures, when available, to include the structural variability over homologous proteins that are present in the PDB. Therefore, when more structures are at disposal, we take as the distance between residues the minimum distance over the residue pairs in the different PDB structures.

by F. Oteri from our team at LCQB

To avoid both trivial contacts (neighbors on the backbone) and strong but uninformative "gap-gap" signals, we also impose a minimum separation $|j-i| > 10$ along the protein backbone. Indeed, as it has been extensively discussed in Part II, gaps in the multiple sequence alignment (MSA) are not generally modeled well by DCA methods, as they tend to come in long stretches, giving rise to artificially high couplings for closer sites on the backbone.

We use MSAs of protein domains downloaded from the Pfam database version 27.0 [49]. We compute the solvent accessibility of a given residue using the "Naccess" tool [58].

2.1.2 *Mean coupling matrix and its spectral modes*

For each Pfam family n we infer the $\frac{1}{2}N_n(N_n-1)$ (N_n being the aligned length of the proteins in family n) coupling matrices with the plmDCA method [41] at standard regularization ($\gamma = 10^{-2}$), and shift them into the zero-sum gauge. The top ranked residue pairs (i,j) according to the APC score are selected until a rate of 20% of false-positive contact predictions is reached within the selection. Then, only the true-positive predictions (contacts in the tertiary structure) are kept in the selection $\mathcal{S}_n$. The number of selected pairs $|\mathcal{S}_n|$ thus depends on the Pfam family n. We obtain the global selection of residue pairs $\mathcal{S}$ by assembling the selected pairs of each Pfam family together: $\mathcal{S} = \bigcup_{n=1}^{70} \mathcal{S}_n$, with $|\mathcal{S}| = 3790$.

see Eq. (3.8) in Part I

In the following, we consider the mean matrix

$$e(a,b) = \left\langle J_{ij}(a,b) \right\rangle_{ij\in\mathcal{S}} \ , \tag{2.1}$$

where $\langle . \rangle_{ij\in\mathcal{S}}$ denotes the mean over all residue pairs in the above-mentioned selection $\mathcal{S}$, all Pfam families taken together. The matrix e is subsequently symmetrized, as any non-symmetric features of amino-acid interactions originate only from finite-sampling effects in the selection:

$$e(a,b) \to \frac{1}{2}\big(e(a,b) + e(b,a)\big) \ . \tag{2.2}$$

The average coupling matrix e is already in the zero-sum gauge, since the couplings $J_{ij}(a,b)$ are. By considering the mean matrix, we expect site specificities and finite-sampling noise to be averaged out, while the joint global interaction modes should be prominently displayed.

We define the spectral mode k of e by

$$M^{(k)}(a,b) = \lambda^{(k)} v_a^{(k)} v_b^{(k)} \ , \tag{2.3}$$

where $\{\lambda^{(k)}, v^{(k)}\}_{k=1\ldots21}$ are the eigenmodes of e, with the eigenvalues $\lambda^{(k)}$ ranked in decreasing order in absolute value.

2.2 THE COUPLING MATRICES REFLECT BIOLOGICALLY RELEVANT INFORMATION

Strikingly, we find that the mean matrix e and its top three spectral modes display some physico-chemical interactions at the amino-acid scale, consistent with the MJ energy matrix E_0^{MJ}, *cf.* Fig. 2.1. The first spectral mode ($\lambda^{(1)} = -0.0923$) is indeed related to electrostaticity, the second ($\lambda^{(2)} = 0.0363$) and third ($\lambda^{(3)} = -0.0197$) modes are mainly localized on some hydrophobic amino acids (A to P). The third mode illustrates favorable residue pairing between amino acids of opposing size: A on one hand (Van der Waals volume of 67 Å^3) and F, I, L on the other hand (Van der Waals volume of 135 Å^3, 124 Å^3, and 124Å^3 respectively). This coevolutionary effect derives from stericity, and is dominant here because of the abundance of the involved amino acids. The favorable interaction between amino acids of opposite size, and unfavorable between amino acids of the same size can be easily understood: given a contact between two amino acids of opposite size, each single change of a small into a large or a large into a small amino acid induces unfavorable steric effects. A compensatory mutation of the second amino acid would be possible.

I thank R Guerois for useful discussions on stericity

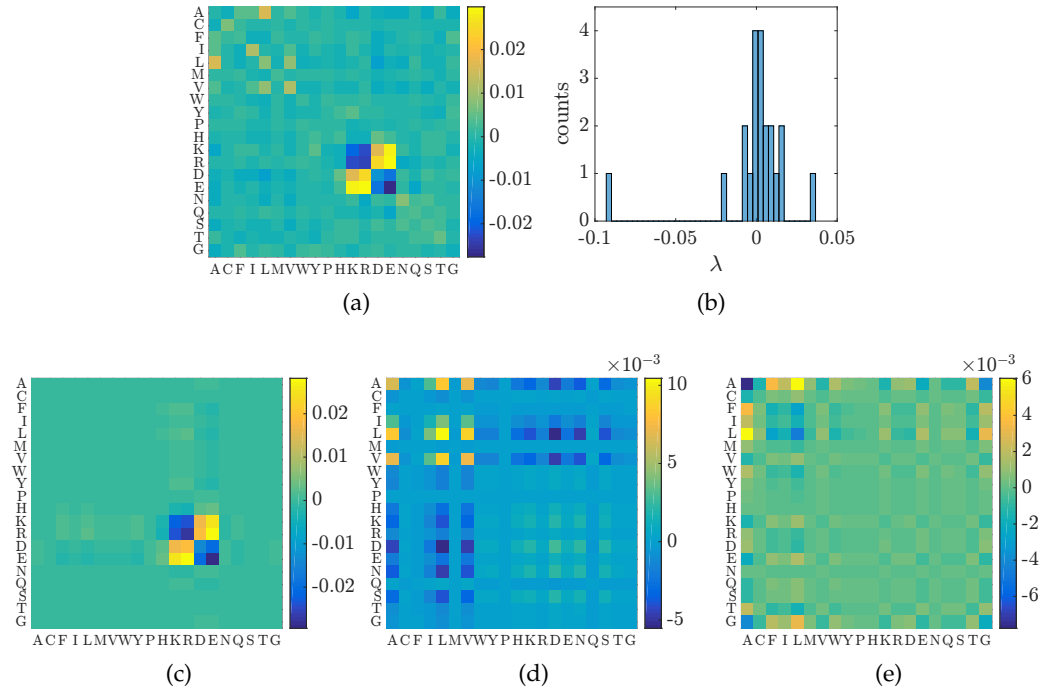


Figure 2.1 – (a) Mean matrix $e(a, b)$ over all residue pairs in the selection, taking all Pfam families together. (b) Spectrum of e, dominated by three eigenvalues. (c) First spectral mode of e ($\lambda^{(1)} = -0.0923$), displaying the electrostatic interaction. (d), (e) Second ($\lambda^{(2)} = 0.0363$) and third ($\lambda^{(3)} = -0.0197$) spectral mode of $e(a, b)$, mainly localized on hydrophobic amino acids (A to P).

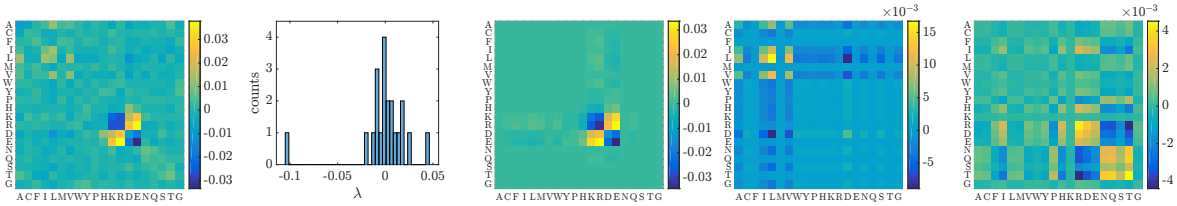
The sign of all eigenvalues is consistent with what has been previously reported for the MJ energy matrix: it is positive for attractive interaction between like amino acids (second mode related to hydrophobicity), negative for attractive interaction between unlike amino acids (first and third modes related to charge and size). Note that the entries and the eigenvalues of $e(a,b)$ are small compared to their counterparts in MJ, a fact we will discuss in Section 2.2.3.

We conclude that the inferred DCA coupling matrices display quantitative and biologically relevant information, beyond their known efficiency to predict tertiary contacts. However, contrary to the MJ statistical potential (Fig. 1.1) which includes the possibility of contacts between hydrophilic amino acids (from H to G) and Cysteine-Cysteine (C-C entry), we do not observe such a signal in the modes of the mean matrix e. The Pearson correlation coefficient between $e(a,b)$ and $E_0^{MJ}(a,b)$ is quite low: 0.58.

2.2.1 *C-C signal and structural classification*

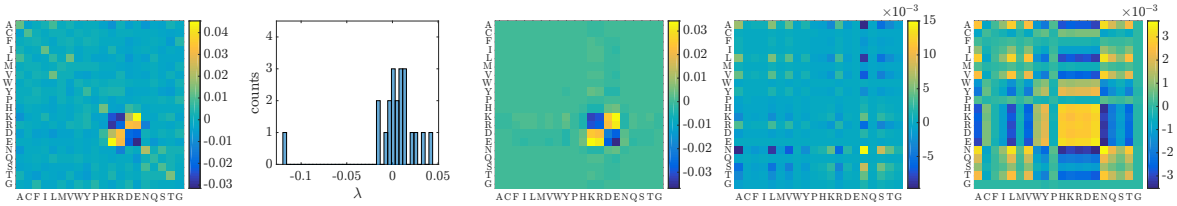
The absence of the Cysteine-Cysteine signal may very well be explained by the scarcity of contacts of this type. In order to gain a more detailed view of the possible contact matrices, we divide up the pool of Pfam families into structural domains based on similarities of their structures using the manual Structural Classification of Proteins (SCOP) database [84] (the repartition is in Appendix E). Five SCOP classes are considered in this analysis : all α-proteins, all β-proteins, α- and β-proteins (mainly antiparallel beta sheets: beta-alpha-beta units and segregated alpha and beta regions), membrane and cell surface proteins and peptides, small proteins. The latter is characterized by the abundance of disulfide bridges between two Cysteines. This gives rise to 5 new selections $\mathcal{S}^{(x)} = \bigcup_{n\in x} \mathcal{S}_n$, where x is the SCOP class ($x \in \{\alpha, \beta, \alpha+\beta, \mathrm{membrane}, \mathrm{small}\}$). We get $|\mathcal{S}^{(\alpha)}| = 300$, $|\mathcal{S}^{(\beta)}| = 493$, $|\mathcal{S}^{(\alpha+\beta)}| = 1814$, $|\mathcal{S}^{(\mathrm{membrane})}| = 879$, and $|\mathcal{S}^{(\mathrm{small})}| = 304$.

Figures 2.2 to 2.6 display, for each of the five SCOP classes, the new mean matrices $e(a,b|x) = \left\langle J_{ij}(a,b)\right\rangle_{ij\in\mathcal{S}^{(x)}}$, their spectra and the top three spectral modes. Electrostatic spectral modes are found in all five SCOP classes (with negative eigenvalues), whereas hydrophobicity-related modes are identified in all but the *small* protein classes. The Cysteine-Cysteine mode is found only in the *small* protein class, as expected (and with a positive eigenvalue). Interestingly, while the hydrophilic signal (amino acids H to G) is still rare in the dominating spectral modes, its presence can be observed in classes α, β and *small*, respectively on the third (Fig. 2.2, panel (e)), second (Fig. 2.3, panel (d)), and third (Fig. 2.6, panel (e)) spectral modes. The third mode of class *small* (Fig. 2.6, panel (e)) even displays both hydrophobic *and*



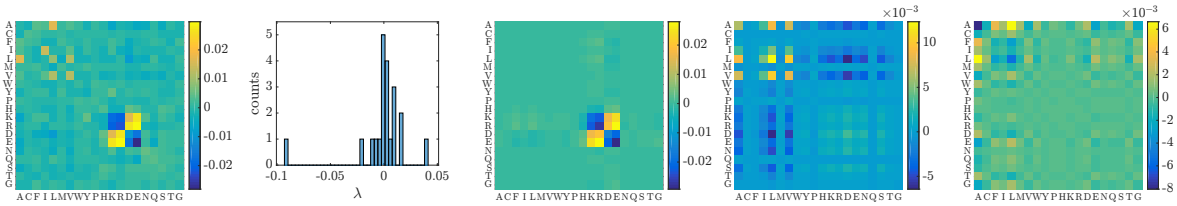
(a) (b) (c) (d) (e)

Figure 2.2 – **α proteins** - (a) $e(a, b|\alpha)$ - (b) Spectrum - (c), (d), (e) Top three spectral modes displaying electrostatic ($\lambda^{(1)} = -0.1043$), hydrophobic ($\lambda^{(2)} = 0.0459$), and hydrophilic ($\lambda^{(3)} = 0.0238$) interactions.



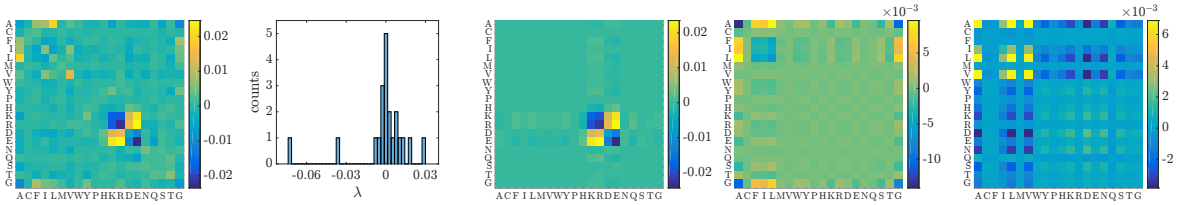
(a) (b) (c) (d) (e)

Figure 2.3 – **β proteins** - (a) $e(a, b|\beta)$ - (b) Spectrum - (c), (d), (e) Top three spectral modes displaying electrostatic ($\lambda^{(1)} = -0.1171$) and hydrophobic/hydrophilic interactions ($\lambda^{(2)} = 0.0405$, $\lambda^{(3)} = 0.0328$).



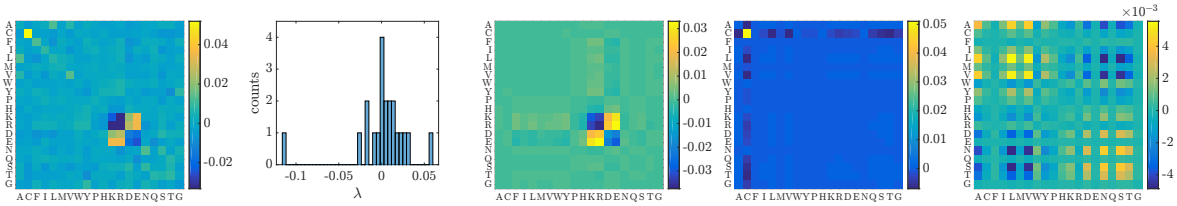
(a) (b) (c) (d) (e)

Figure 2.4 – **$\alpha + \beta$ proteins** - (a) $e(a, b|\alpha + \beta)$ - (b) Spectrum - (c), (d), (e) Top three spectral modes displaying electrostatic ($\lambda^{(1)} = -0.0905$) and hydrophobic ($\lambda^{(2)} = 0.0412$, $\lambda^{(3)} = -0.0198$) interactions.



(a) (b) (c) (d) (e)

Figure 2.5 – **membrane proteins** - (a) $e(a, b|membrane)$ - (b) Spectrum - (c), (d), (e) Top three spectral modes displaying electrostatic ($\lambda^{(1)} = -0.0729$) and hydrophobic ($\lambda^{(2)} = -0.0366$, $\lambda^{(3)} = 0.0299$) interactions.



(a) (b) (c) (d) (e)

Figure 2.6 – **small proteins** - (a) $e(a, b|small)$ - (b) Spectrum - (c), (d), (e) Top three spectral modes displaying electrostatic ($\lambda^{(1)} = -0.1129$, Cysteine-Cysteine ($\lambda^{(2)} = 0.00567$), and hydrophobic/hydrophilic ($\lambda^{(3)} = 0.0306$) interactions.

hydrophilic interactions, similarly to the MJ energy matrix E_0^{MJ} (*cf.* Fig. 1.1, panel (c)).

The spectrum of $e(a, b|\beta)$ is dominated by one eigenvalue ($\lambda^{(1)} = -0.1171$), the second and third eigenvalues being relatively close ($\lambda^{(2)} = 0.0405$, $\lambda^{(3)} = 0.0328$). It causes the separation between the second and third spectral modes (Fig. 2.3, panels (d) and (e)) to be less clear and more sensitive to finite sampling noise than for the other classes, whose spectra are dominated by more than one eigenvalue.

2.2.2 *Hydrophilicity and solvent exposure*

The weakness of a signal involving hydrophilic amino acids (from H to G) may be explained by the scarcity of contacts between two sites localized on the surface of the protein as compared to all other contacts – surface amino acids are indeed most likely to be hydrophilic. We now divide the selected residue pairs in $\mathcal{S}$ into three new classes depending on the solvent exposure – measured by the relative solvent accessibility (RSA) determined using the "Naccess" software [58] – of the involved residues, regardless of the Pfam family they are issued from:

— "surface-surface" contacts: more than half of the surface of both residues is exposed to the solvent,
(selection $\mathcal{S}^{(ss)} = \{ij \in \mathcal{S} \mid RSA(i), RSA(j) > 50\%\}$);

— "core-core" contacts: less than half of the surface is exposed,
(selection $\mathcal{S}^{(cc)} = \{ij \in \mathcal{S} \mid RSA(i), RSA(j) < 50\%\}$);

— "core-surface" contacts: one residue has more than half of its surface exposed, the other has less than half,
(selection $\mathcal{S}^{(cs)} = \{ij \in \mathcal{S} \mid (RSA(i) > 50\%, RSA(j) < 50\%) \vee (RSA(i) < 50\%, RSA(j) > 50\%)\}$).

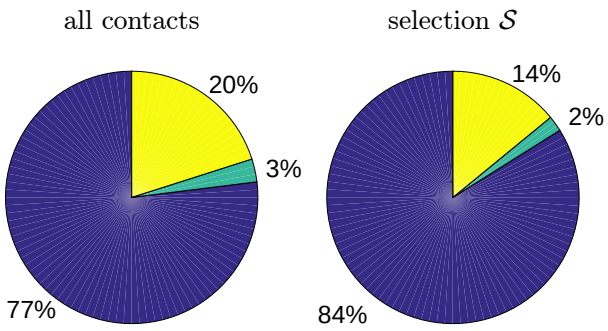
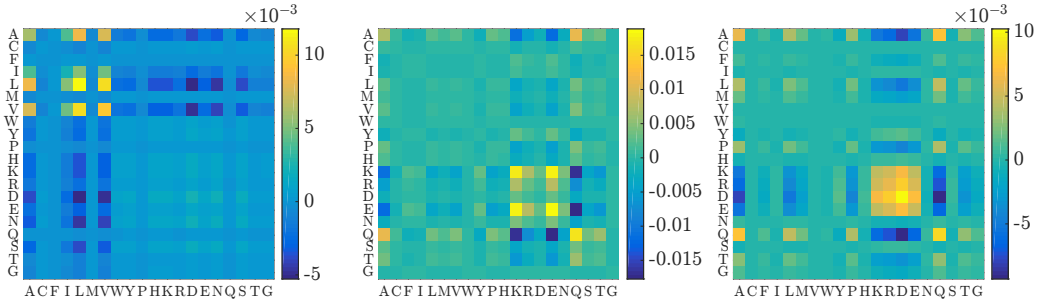


Figure 2.7 – Distribution of core-core (blue), surface-surface (green), and core-surface (yellow) contacts among all contacts (left panel) and contacts in our selection (right panel). Surface-surface contacts are statistically underrepresented in both cases.

Fig. 2.7 displays the repartition of core-core (blue), surface-surface (green), and core-surface (yellow) contacts among all existing tertiary contacts (left panel) and contacts in the selection $\mathcal{S}$ (right panel). As expected, by far the largest part of the tertiary contacts lies in the core of the proteins. Only 2-3% of the (selected) contacts are between surface residues.

Similarly to what has been done before, we consider average coupling matrices for these 3 new classes: $e(a,b|y) = \left\langle J_{ij}(a,b)\right\rangle_{ij\in\mathcal{S}_y}$, with $y \in \{ss, cc, cs\}$ along with their spectral modes. In all classes, the first spectral mode displays the usual electrostatic signal. However, while the second mode of the "core-core" class is localized on hydrophobic amino acids only (from A to P), in agreement with what is observed on Fig. 2.1, the second modes of the "surface-surface" and "core-surface" classes are localized only on hydrophilic (H to G) amino acids, as shown on Fig. 2.8.



(a) (b) (c)

Figure 2.8 – Second spectral modes of the mean matrices (a) $e(a,b|cc)$ over "core-core" contacts, (b) $e(a,b|ss)$ over "surface-surface" contacts, and (c) $e(a,b|cs)$ over "core-surface" contacts. A hydrophilicity-related signal is displayed on the 2 latter.

2.2.3 *Differences with Miyazawa-Jernigan*

2.2.3.1 *Analog of the Miyazawa-Jernigan potential*

The analog of MJ's contact energy (see Eq. (9a) in [79]) in our description would be approximately the quantity $E^{stat}(a,b)$ defined through

$$E^{stat}(a,b) = \log \frac{\left\langle f_{ij}(a,b)\right\rangle_{ij\in\mathcal{S}}}{\left\langle f_i(a)\right\rangle_{i\in\mathcal{S}} \left\langle f_j(b)\right\rangle_{j\in\mathcal{S}}}\,, \tag{2.4}$$

where $\langle . \rangle_{ij\in\mathcal{S}}$ denotes the mean over all residue pairs in the selection $\mathcal{S}$ (all Pfam families taken together), and $\langle . \rangle_{i\in\mathcal{S}}$ and $\langle . \rangle_{j\in\mathcal{S}}$ are the means over all single residues involved in a contact pair in the selection $\mathcal{S}$. E^{stat} is then symmetrized and shifted to the zero-sum gauge. A

straightforward computation gives the analytical expression of E^{stat} in the zero-sum gauge:

$$E^{stat}(a,b) \to \log\left\langle f_{ij}(a,b)\right\rangle_{ij} - \log\left\langle f_{ij}(\cdot,b)\right\rangle_{ij} - \log\left\langle f_{ij}(a,\cdot)\right\rangle_{ij} + \log\left\langle f_{ij}(\cdot,\cdot)\right\rangle_{ij} \ , \tag{2.5}$$

which is by definition the zero-sum gauge transformation of the matrix $\widetilde{E^{stat}}(a,b) = \log\left\langle f_{ij}(a,b)\right\rangle_{ij}$. The denominator of Eq. (2.4) is therefore irrelevant.

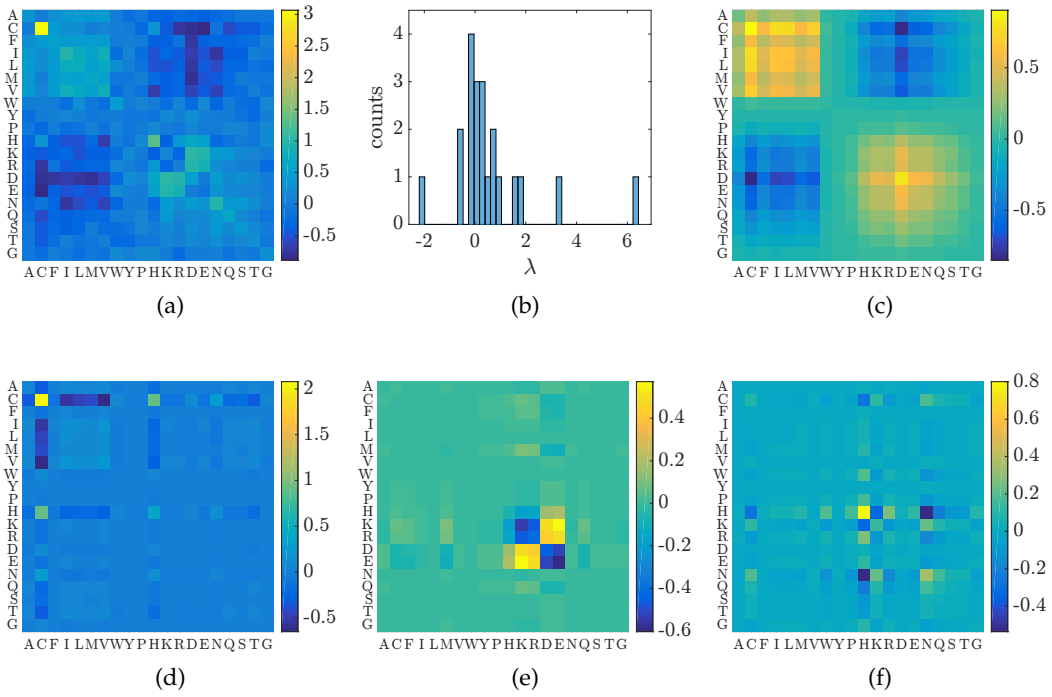


Figure 2.9 – (a) Mean matrix E^{stat} (analog to MJ) over all residue pairs in the selection. (b) Histogram of the spectrum of E^{stat}. (c), (d), (e), (f) First spectral modes of E^{stat} displaying hydrophobic-hydrophilic, Cysteine-Cysteine, electrostatic, and Histidine-Histidine interactions.

As shown on Fig. 2.9, the first spectral modes of E^{stat} are very similar to the genuine MJ energy matrix $E_0^{MJ}(a,b)$ – although not in the exact same order – and the Pearson correlation coefficient between the two matrices is 0.81. The order of magnitude of $E^{stat}(a,b)$ and its top eigenvalues are also close to the MJ energy matrix.

2.2.3.2 *Relation with inferred couplings*

The E^{stat} matrix can be related to the inferred couplings in an approximate way as follows. For pairs of site i,j in contact (in the selection $\mathcal{S}$), contrary to sites not in contact, the major contribution to the direct coupling $J_{ij}(a,b)$ comes from the direct correlation $f_{ij}(a,b)/(f_i(a)f_j(b))$

between the sites. Indirect contributions to $f_{ij}(a,b)$, mediated through other sites, are expected to be much smaller. Approximating $J_{ij}(a,b)$ with $\log(f_{ij}(a,b)/(f_i(a)f_j(b)))$ is indeed exact in the case of two interacting sites only. Consequently we introduce the matrix $E^{DIR}(a,b)$ as

$$E^{DIR}(a,b) = \log \frac{\left\langle f_i(a) f_j(b) \exp J_{ij}(a,b) \right\rangle_{ij\in\mathcal{S}}}{\left\langle f_i(a) \right\rangle_{i\in\mathcal{S}} \left\langle f_j(b) \right\rangle_{j\in\mathcal{S}}} \,. \tag{2.6}$$

Again, E^{DIR} is symmetrized and shifted to zero-sum gauge. As displayed on Fig. 2.10, the first spectral modes are very close to the MJ energy matrix (Fig. 1.1), although not in the same order (of decreasing eigenvalue in absolute value). The order of magnitude of $E^{DIR}(a,b)$ and its top eigenvalues are much more similar to the MJ matrix than the other mean matrices $e(a,b|\star)$, with a Pearson correlation coefficient of 0.77.

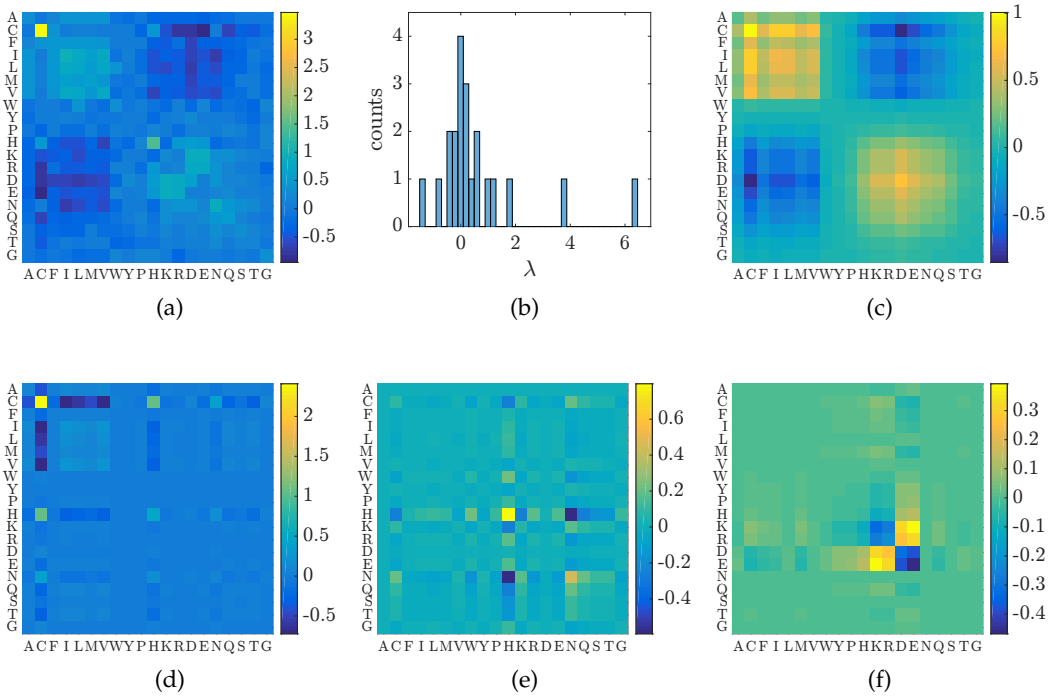


Figure 2.10 – (a) Mean matrix E^{DIR} over all residue pairs in the selection, taking all Pfam families together. (b) Histogram of the spectrum of E^{DIR}. (c), (d), (e), (f) First spectral modes of E^{DIR} displaying hydrophobic-hydrophilic ($\lambda^{(1)} = 6.44$), Cysteine-Cysteine ($\lambda^{(2)} = 3.78$), Histidine-Histidine ($\lambda^{(3)} = 1.80$), and electrostatic ($\lambda^{(4)} = -1.41$) interactions.

This shows that the DCA couplings reflect the full information of the MJ contact energy, provided that the mean is properly weighted by the single-site frequencies. This is consistent with the previous results where the data set of coupling matrices is divided up into structural classes or solvent exposure related classes.

2.3 DISTANCE DISTRIBUTION

2.3.1 *Naive clustering*

Within the SCOP classification defined in Section 2.2.1, we assign each residue pair (i,j) in the selection $\mathcal{S}^{(x)}$ to one spectral mode (k) of $e(a,b|x)$ (with $x \in \{\alpha, \beta, \alpha+\beta, membrane, small\}$), as follows: we first define the score $\pi_{ij}^{(k)}$ *via* the projection of the coupling matrix $J_{ij}(a,b)$ onto the spectral mode (k):

$$\pi_{ij}^{(k)} = \sum_{a,b=1}^{21} J_{ij}(a,b)\, v_a^{(k)}\, v_b^{(k)} \,, \tag{2.7}$$

where $v_a^{(k)}$, $a = 1, ..., b$ are the components of the eigenvector associated to the k^{th} eigenvalue of $e(a,b|x)$. Then, the residue pair (i,j) is assigned to the mode (k) on which the projection $\pi_{ij}^{(k)}$ is maximum.

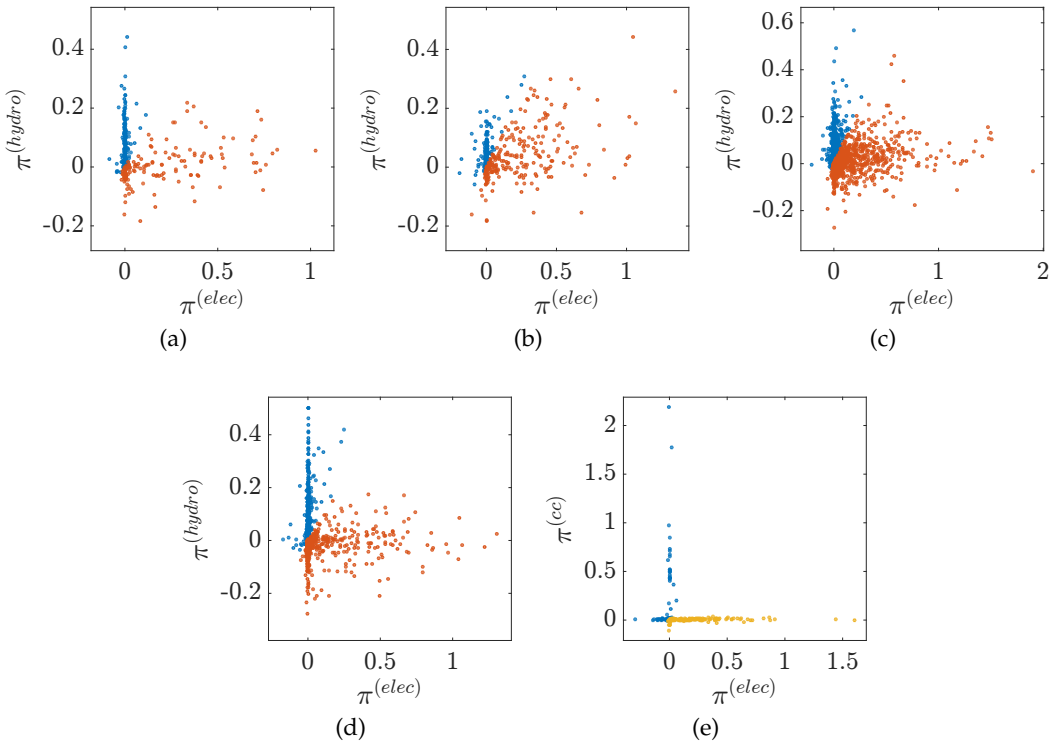


Figure 2.11 – Projection scores $\pi_{ij}^{(k)}$, $k = 1,2$ for all residue pairs (i,j) within SCOP classes (a) α (electrostatic and hydrophobic), (b) β (electrostatic and hydrophobic), (c) $\alpha+\beta$ (electrostatic and hydrophobic), (d) membrane (electrostatic and hydrophobic), and (e) small (electrostatic and Cysteine-Cysteine). Colors indicate the cluster the residue pair has been assigned to (maximum projection score): electrostatic (blue), hydrophobic (red), and Cysteine-Cysteine (yellow).

For each SCOP class, we consider the projection onto the top two spectral modes $k = 1,2$: electrostatic and hydrophobic for the SCOP

classes α, β, $\alpha+\beta$, *membrane*, and electrostatic and Cysteine-Cysteine for the class of *small* proteins (Figs. 2.2 to 2.6). The top two eigenvalues of $e(a, b|x)$ indeed account in each class for about 50% of the sum of all eigenvalues. Figure 2.11 displays the two projection scores $\pi_{ij}^{(k)}$, with $k = 1, 2$, for all residue pairs (i, j) within the five SCOP classes. Each color corresponds to the cluster the residue pairs are assigned to, *i.e.* the mode (k) with maximum projection $\pi_{ij}^{(k)}$.

The projection $\pi_{ij}^{(elec)}$ on the electrostatic modes (red dots on Fig. 2.11) is positive for the vast majority of contacts, reflecting the strength and importance of the electrostatic interaction. Residue pairs assigned to hydrophobic modes (blue dots on Fig. 2.11) usually have a projection $\pi^{(elec)}$ close to zero, pointing out that hydrophobic residues are uncharged. While the assignment procedure seems to be well justified for the SCOP classes α, *membrane*, and *small* (panels (a), (d), (e)), no clear separation is observed for classes β and $\alpha+\beta$ (panels (b) and (c)), in which the values of the projection scores for a given residue pair may be both large and comparable in magnitude. This can be explained by the overlapping supports of the electrostatic and hydrophobic spectral modes in theses classes, the latter also having a hydrophilic signal (amino acids K,H,R,D,E are charged *and* hydrophilic), especially for the β class, *cf.* Fig. 2.3 panel (d) and Fig. 2.4 panel (d). Notice that, for the class *small*, the separation between electrostatic and Cysteine-Cysteine modes is very good as the amino acids supporting those interactions are disjoint (K,H,R,D,E for the former, C for the latter).

2.3.2 *Contact distances*

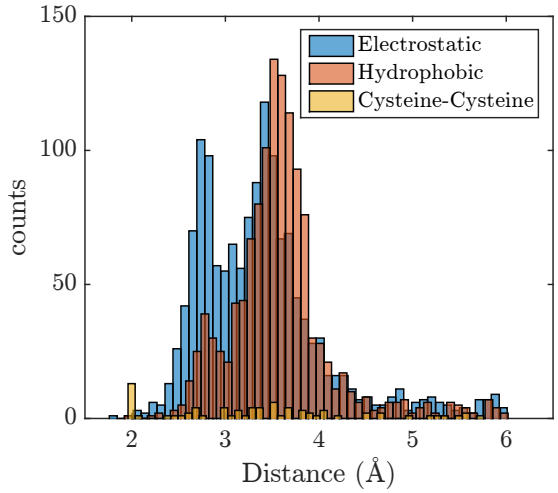


Figure 2.12 – Distribution of distances in the tertiary structure among the selected residue pairs in contact for the different interaction types, pooled across the SCOP classes.

We now study how the native distances in the tertiary structure between the residue pairs vary with the type of interactions they have been assigned to (electrostatic, hydrophobic or Cysteine-Cysteine) as described above. The distance distributions are shown on Fig. 2.12, and vary considerably with the interaction types. The "hydrophobic" type involve residue pairs with a contact distance centered around 3.5 Å, the "electrostatic" type displays a bimodal distance distribution mostly around 2.7 Å and 3.5 Å, and the "Cysteine-Cysteine" type is the only one to have a significant number of pairs in contact at short distance 2 Å. Notice that 3.5 Å is the typical distance between heavy atoms, twice the Van der Waals distance (1.7 Å), on the other hand 2.7 Å corresponds to the distance between atoms linked by a strong to moderate hydrogen bond [63], and 2 Å is the distance between two Cysteine involved in a disulfide bridge.

I thank E. Westhof for his insight on contact distances

2.4 CLUSTERING OF THE COUPLING MATRICES

The naive clustering presented in the previous section unveils important variations in the native distances in the tertiary structure, depending of the type of interaction. In this section, we propose a more refined clustering method, confirming the results obtained in Sections 2.2.1 & 2.3.

2.4.1 *Method*

Within each SCOP class x, we consider the matrix $\mathcal{J}^{(x)}$ of size $|\mathcal{S}^{(x)}| \times 21^2$, where each element $(\mathcal{J}^{(x)})_{n=ij,p=ab}$ is the coupling entry $J_{ij}(a,b)$. The singular value decomposition of this collection of coupling matrices $\mathcal{J}$ writes

$$(\mathcal{J}^{(x)})_{np} = \sum_{\mu=1}^{21^2} U_n^{(\mu)} E_p^{(\mu)} \sigma^{\mu} \,, \tag{2.8}$$

with U the matrix of left-singular vectors, *i.e.* eigenvectors of

$$\left(\mathcal{J}^{(x)} \times (\mathcal{J}^{(x)})^{\top}\right)_{n=ij,m=kl} = \sum_{a,b=1}^{21} J_{ij}(a,b) J_{kl}(a,b) \,, \tag{2.9}$$

E the matrix of right-singular vectors, *i.e.* eigenvectors of

$$\left((\mathcal{J}^{(x)})^{\top} \times \mathcal{J}^{(x)}\right)_{p=ab,r=cd} = \sum_{i,j=1}^{|\mathcal{S}^{(x)}|(|\mathcal{S}^{(x)}|-1)/2} J_{ij}(a,b) J_{ij}(c,d) \,, \tag{2.10}$$

and σ the diagonal matrix of singular values.

The spectrum of $\mathcal{J}^{(x)}$ is usually dominated by only a few singular values, so the summation in Eq. (2.8) can be truncated. In the following, we apply the Principal Component Analysis (PCA) to perform a

dimensional reduction by only considering the few singular vectors related to the largest singular values. Finally, the signals from different types of contacts are disentangled by clustering these singular vectors into several classes, with a simple k-means algorithm.

This algorithm is implemented with the correlation distance (one minus the sample correlation between points), and 100 replicates (number of times the clustering is made using new initial conditions to help find the best local minimum). When the euclidean distance is used instead, the identified contact classes are roughly the same, but with a supplementary class of all the coupling matrices with lowest norms. Indeed, if two coupling matrices are proportional with a large proportionality coefficient, the euclidean distance is large whereas the correlation distance is 0. The clustering is moreover remarkably stable, independent of the initial conditions.

The number of clusters – an input in the k-means algorithm – is chosen to be close to the number of singular values clearly outside of the bulk in the spectrum of $\mathcal{J}^{(x)}$ (from 2 to 7 depending on the considered SCOP class x). To find the optimal number of singular vectors in the PCA dimensional reduction and the optimal number of clusters, we consider the so-called *silhouette* of the clustering. This method is a measure of the consistency within clusters, *i.e.* of how well each coupling matrix lies within its cluster. For each coupling matrix $J_{p=ij}$, let $a(p)$ be the average distance correlation of J_p with all other coupling matrices within the same cluster. Let $b(p)$ be the lowest average distance correlation of J_p to any other cluster, of which it is not a member. The silhouette is defined as

$$s(p) = \frac{b(p) - a(p)}{\max\{a(p), b(p)\}} \,. \tag{2.11}$$

By definition, the silhouette is smaller than one in absolute value: $-1 < s(p) < 1$. Silhouette values close to 1 indicate appropriately clustered data points. Silhouette values close to -1 suggest that the corresponding data points would be better allocated to the neighboring cluster.

Appendix F displays examples of silhouettes for clusterings of normally distributed random numbers.

With this method, all the residue pairs in $\mathcal{S}^{(x)}$ are divided into different clusters. What we call in the following "class of contacts" is defined as the center of mass of the cluster: $(E^{(k)})_{ab} =< J_{ij}(a,b) >_{ij\in(k)}$, where the mean is over the contact pairs attributed to the cluster (k).

2.4.2 *Results*

Similarly to the results of Section 2.2.1, the identified clusters also display physico-chemical interactions at the amino-acid scale: hydrophobicity, electrostaticity and Cysteine-Cysteine, closely related to the spectral modes of the MJ matrix. Figures 2.13 to 2.17 display, for each of the five SCOP classes x, the histogram of the singular values of $\mathcal{J}^{(x)}$, the silhouette of the clustering, and the classes of contacts

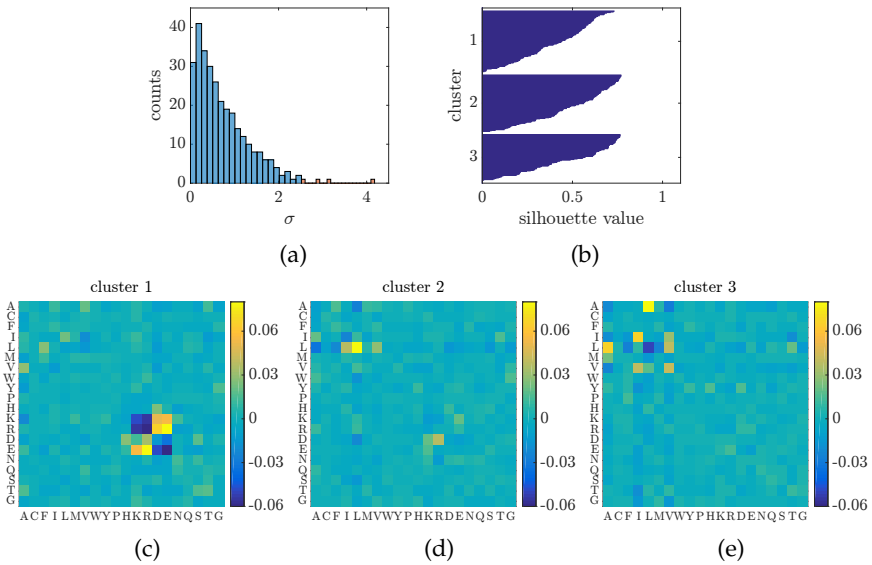


Figure 2.13 – **α proteins** - (a) spectrum of $\mathcal{J}^{(\alpha)}$ - (b) Silhouette values - (c), (d), (e) Classes of contacts displaying electrostatic and hydrophobic interactions.

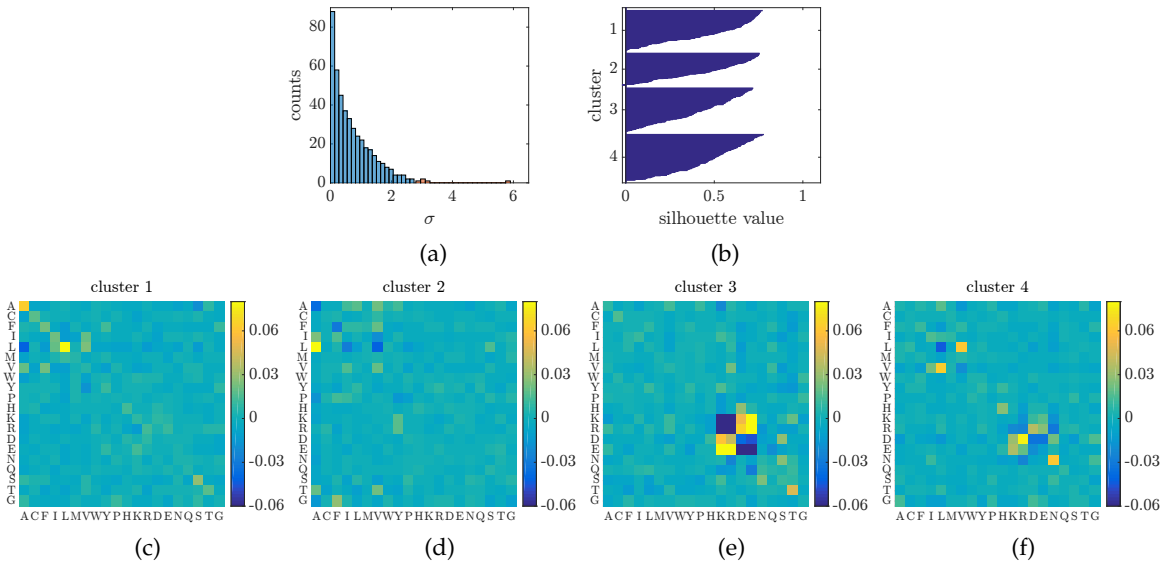


Figure 2.14 – **β proteins** - (a) spectrum of $\mathcal{J}^{(\beta)}$ - (b) Silhouette values - (c), (d), (e), (f) Classes of contacts displaying electrostatic and hydrophobic interactions.

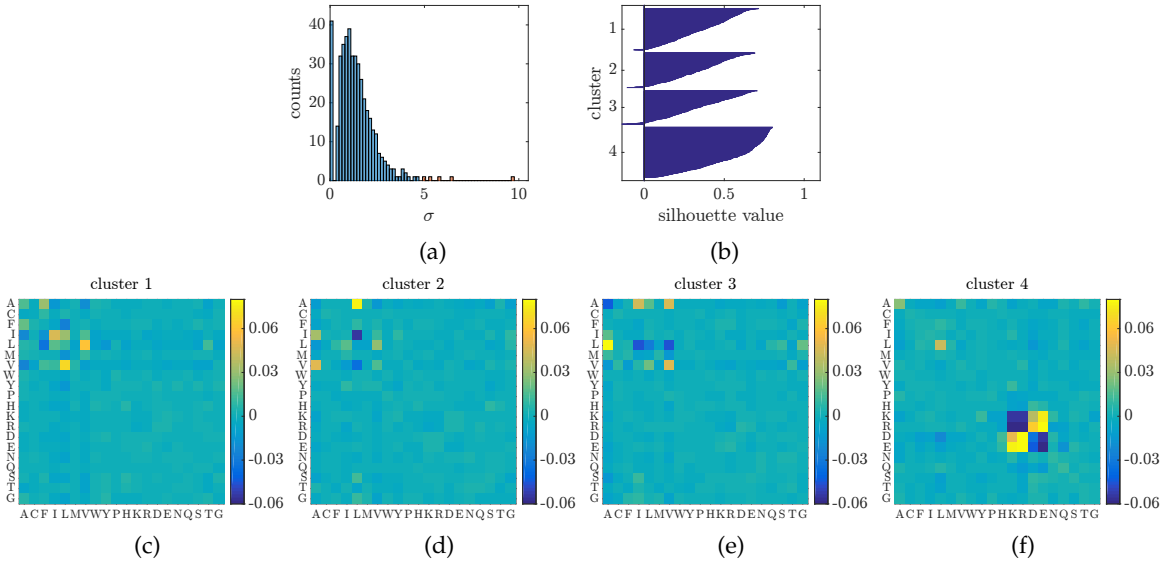


Figure 2.15 – **$\alpha + \beta$ proteins** - (a) spectrum of $\mathcal{J}^{(\beta)}$ - (b) Silhouette values - (c), (d), (e), (f) Classes of contacts displaying electrostatic and hydrophobic interactions.

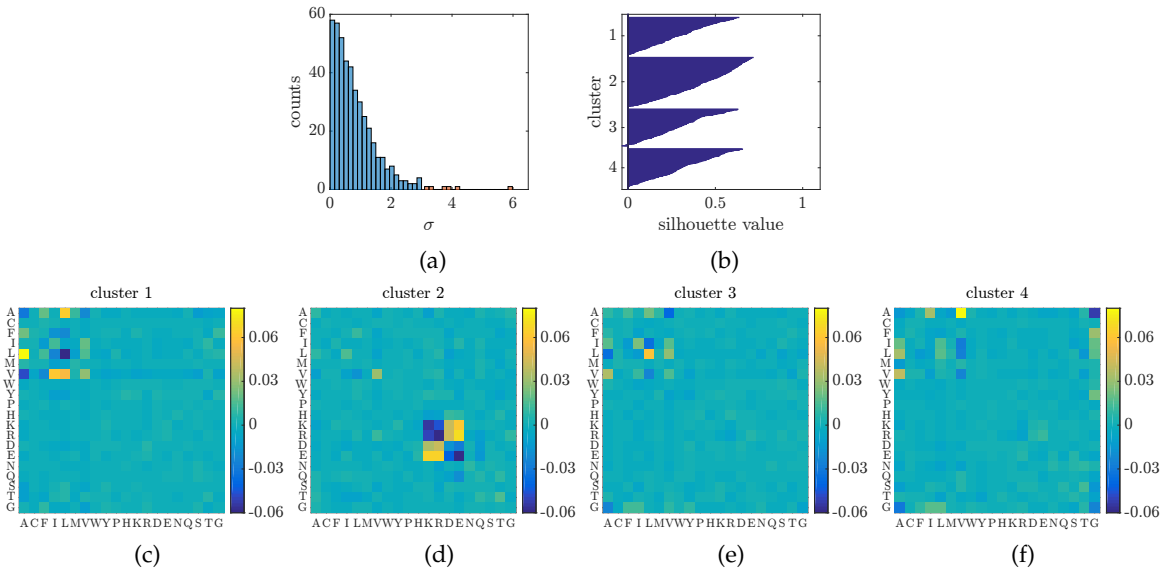


Figure 2.16 – **membrane proteins** - (a) spectrum of $\mathcal{J}^{(\mathrm{membrane})}$ - (b) Silhouette values - (c), (d), (e), (f) Classes of contacts displaying electrostatic and hydrophobic interactions.

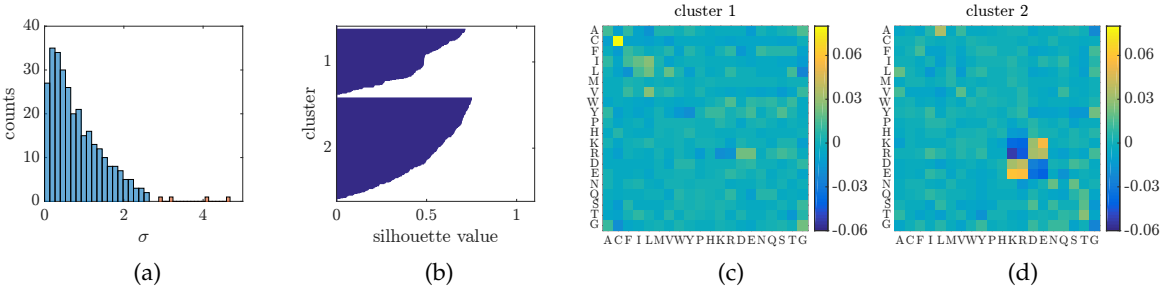


Figure 2.17 – **small proteins** - (a) spectrum of $\mathcal{J}^{(\mathrm{small})}$ - (b) Silhouette values - (c), (d) Classes of contacts displaying electrostatic and Cysteine-Cysteine interactions.

SCOP	PFAM	PCA	CLUSTERS	HYDRO.	ELEC.	CC	MIXED
α	12	4	3	2	1	0	0
β	12	4	4	2	1	0	1
$\alpha+\beta$	18	5	4	4	1	0	0
membrane	16	6	4	3	1	0	0
small	13	4	2	0	1	1	0

Table 5 – Set of chosen parameters for the different SCOP classes: number of Pfam families in the class, number of singular vectors (or dimensions) selected for PCA, and number and types of clusters identified by the k-means algorithm.

(center of mass of the different clusters). Again consistently with Section 2.2.1, electrostatic clusters are found in all five SCOP classes, whereas hydrophobicity-related clusters are identified in all but the *small* protein classes. The Cysteine-Cysteine cluster is found only in the *small* protein class, as expected.

Table 5 contains all the parameters used in each SCOP group: the number of Pfam families (see Appendix E for a complete list), the number of singular vectors considered for the PCA analysis, the number of clusters and their categories. For instance, in the SCOP group of all α-proteins, 4 singular vectors have been used to identify 3 clusters (*cf.* Fig. 2.13). 2 of them are "Hydrophobic" (as they involve only hydrophobic amino acids) and 1 is "Electrostatic". The number of singular vectors and clusters have been chosen to optimize the silhouette distribution (panel (b)). Although the clustering method is more refined, it also depends on more input parameters than the naive approach with the mean matrices, which remains quite simple.

The silhouette values are generally large and positive, underlining the efficiency of the clustering method. The clustering is also remarkably stable: the clusters barely depend on the initial conditions. The silhouette values of the different hydrophobic clusters tend to be more peaked, with smaller values, underlying the fact that hydrophobic clusters have overlapping supports. Negative silhouette values are observed only for the $\alpha+\beta$ class, confirming what has been observed in Section 2.2.1. Panel (c) of Fig. 2.11 indeed showed that there is no clear separation between electrostaticity and hydrophobicity in this SCOP class. This is also the case for class β, where cluster 4 displays both electrostatic and hydrophilic signals (panel (f) of Fig. 2.14), although all silhouette values are positive. The separation between the spectral modes of $e(a,b|\beta)$ was indeed not so clear (*cf.* Section 2.2.1).

This more refined clustering also enables to study the distribution of native distances in the tertiary structure between the residue pairs depending the type of interactions (or cluster) they have been as-

signed to, similarly to what has been done in Section 2.3. The distribution is strictly identical to Fig. 2.12, confirming this result.

2.5 TOWARD AN IMPROVED CONTACT PREDICTION

2.5.1 *Using the unveiled structure of the coupling matrices*

An important question is whether the detailed structure of the inferred couplings revealed in this work could be used to improve structural predictions, based so far on the Frobenius norms of the couplings only (with an APC adjustment, *cf.* Chapter 3 of Part I). It was indeed recently shown [60] that for artificial lattice proteins the projection of the couplings onto the MJ matrix (which has been defined as the evolutionary pressure, see Eq. (3.16) of Part I) is more effective for protein contact prediction than the usual Frobenius-based estimator. However, the sampling situation is extremely favorable in lattice proteins. This is not the case for real proteins, as we will see in the next chapter. Making any improvement to the structural predictions is therefore extremely challenging.

Fig. 2.18 displays the projection of all coupling matrices in the selection $\mathcal{S}$ onto the MJ matrix $\pi^{(\mathrm{MJ})}$ (panel (a)) and onto its first three spectral modes $\pi^{(\mathrm{elec})}$, $\pi^{(\mathrm{hydro})}$, and $\pi^{(\mathrm{cc})}$ (panels (b), (c), (d)) as a function of the APC score. Note that the definition of the projections $\pi^{(k)}$ slightly differs from Eq. (2.3) as it involves here the modes (k) of MJ and not the modes of $e(a, b|x)$. Each of the two methods has its strengths and weaknesses. On one hand, the MJ matrix displays the whole variety of residue-residue contacts but is quite different from the couplings without a proper frequency weighting, as it has been extensively discussed above. On the other hand, $e(a, b)$ has to be learned from the couplings first, leading to more overfitting – contrary to MJ which only results from residue-residue contact statistics in knwon structures. In any case, equivalent results are obtained in both methods, but we will focus here on the projection onto the modes of MJ for the sake of simplicity.

Interestingly, the projections onto the full MJ matrix and onto its electrostatic spectral mode seem quite informative: an important part of the contacts (blue dots) have a high projection and a rather small APC score. Respectively, few of non contacts (red dots) have a high projection. Several attempts have therefore been made by the author of this dissertation to improve the structural prediction by combining both APC and the projections – such as using the union of both predictors, or fitting the bulk of non contacts with gaussian functions (*cf.* Section 2.5.2 below). Unfortunately, none of them is quantitatively improving the accuracy of the structural predictions.

This might be partly explained by the presence of homo-oligomer contacts in the selection, or in other words, residue pairs that are

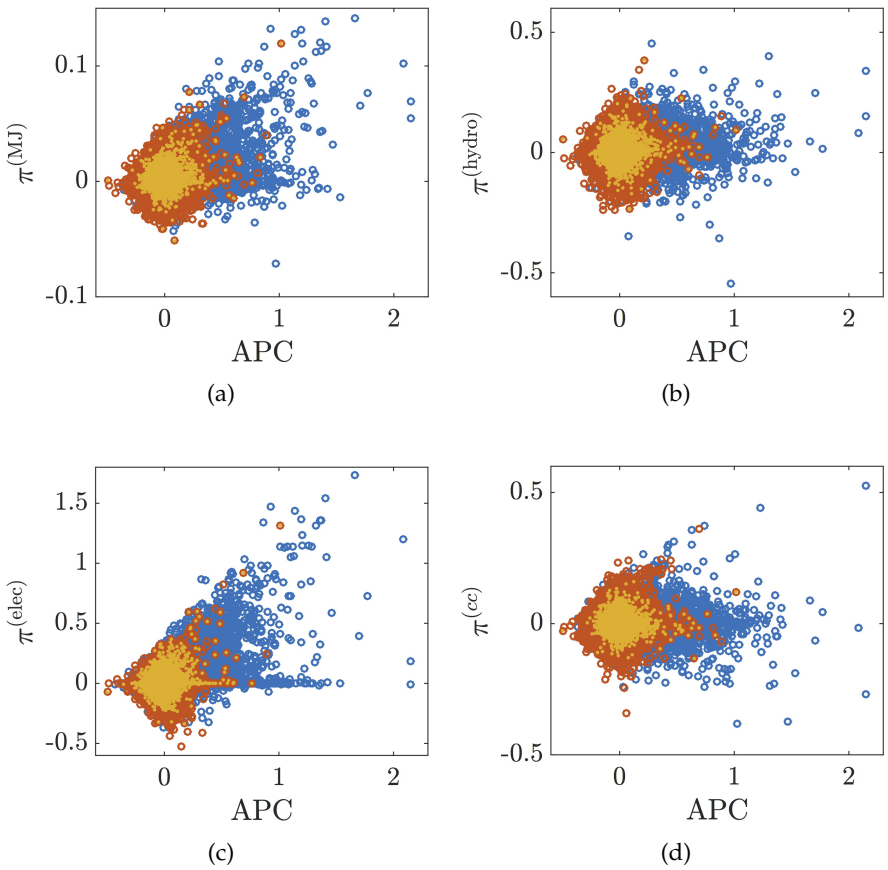


Figure 2.18 – Projections of the coupling matrices in the selection $\mathcal{S}$ onto the MJ matrix and its first three spectral modes as a function of their APC scores, for the tertiary structure contacts (blue), non contacts (red), and for the pairs in contact in the quaternary structure but not in the tertiary structure (yellow).

This data has been made available by G. Uguzzoni, from our team at LCQB

not in contact in the tertiary structure but in contact in the quaternary structure, an arrangement of multiple folded protein subunits. Many protein domains indeed form homo-oligomers with copies of the domain on different chains in the quaternary assembly. It has been shown that this kind of physical constraints constitute a source of coevolution signals detected by DCA approaches [36]. In order to avoid mis-categorization between true- and false- positive predictions, we therefore include in the analysis the homo-oligomers contact map when present and superimpose the contact maps obtained by parsing the monomers and the homo-oligomers structures.

As displayed on Fig. 2.18 (yellow dots), many of the highest APC or projections which are not tertiary contacts – and therefore misclassified as false-positive predictions – are actually coevolving in the quaternary structure. This illustrates how dependent the performance of the structural prediction is in the definition of the contacts, and therefore how complex it is to improve it.

2.5.2 *Attempt: combining the APC and projection scores*

In this paragraph, we will shortly describe an attempt to combine both APC and the projection scores to improve the structural predictions. We focus on the projection onto the electrostatic mode, displayed on panel (c) of Fig. 2.18, showing interesting features. As displayed on Fig. 2.19, the angle θ defines a straight line in the space (APC,$\pi^{(elec)}$); $\theta = 0$ corresponds to the APC score only, whereas $\theta = \pi/2$ is equivalent to the projection alone. Each residue pair (therefore point on the Figure) will be attributed a score corresponding to its orthogonal projection on this straight line. The scores are subsequently ranked and the positive predictive value (PPV) is plotted (*cf.* Chapter 3 of Part I for the definition) and compared to the PPV with APC only. Each value of θ corresponds to a different combination of the APC score and the projection. Unfortunately, it seems that the choice of $\theta = 0$ could not be outperformed and that considering APC alone remains the best option.

Alternatively, since the ranking in the euclidean space was not effective, a more elaborated option is to consider a distance in the metric space defined by the bulk of non contacts. The latter is fitted by a gaussian function $f : t \mapsto \exp\left(-(t-t_0)^2/2\sigma_0^2\right)$. A residue pair of coordinates (x, y) in the space (APC,$\pi^{(elec)}$) will be attributed a score λ such that $y = \lambda f(x/\lambda)$; *cf.* Fig. 2.20. The scores are then ranked and the corresponding PPV is compared to the one obtained with APC. Again, this new score is outperformed by APC alone.

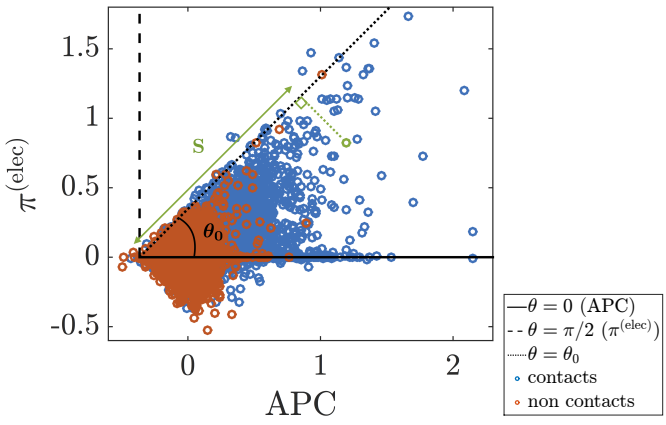


Figure 2.19 – Projections of the coupling matrices in the selection $\mathcal{S}$ onto the electrostatic mode of the MJ matrix. $\theta = 0$ corresponds to APC (full line), $\theta = \pi/2$ is the projection (dashed line), and $\theta \in [0, \pi/2]$ is a combination of both scores. Here $\theta_0 \approx 0.76$ (dotted line) and the score of the green point of coordinates (APC$= 1.198$,$\pi^{(elec)} = 0.8201$) corresponding to its orthogonal projection is of $S = 1.434$.

We propose here two methods based on the idea that the combination of both the already effective APC score and the newly defined

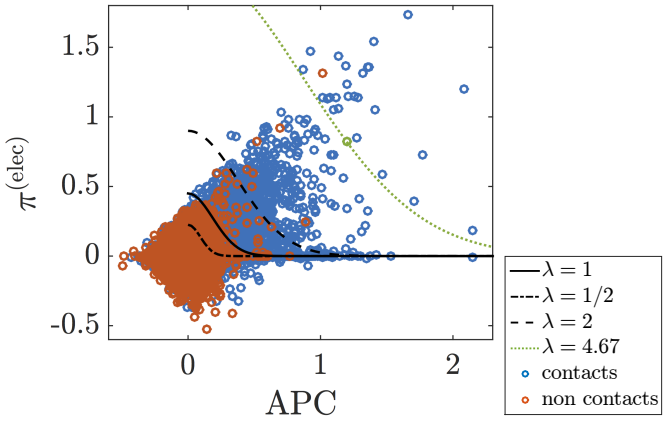


Figure 2.20 – Projections of the coupling matrices in the selection $\mathcal{S}$ onto the electrostatic mode of the MJ matrix. For a point of coordinates (x, y), the score λ is defined such that $y = \lambda f(x/\lambda)$, with $f : t \mapsto \exp\left(-t^2/0.07\right)$ $(\lambda = 1)$. The score of the green point of coordinates $(\text{APC} = 1.198, \pi^{(elec)} = 0.8201)$ is $\lambda = 4.67$.

projection score might improve the structural prediction in proteins. These methods show no improvement compared to the use of APC alone. As mentioned above, the presence of homo-oligomer contacts yields false positives which are actually in contact, but not in the tertiary structure. Considering these as true positive does improve the prediction with the projection, but also with APC, leading to no substantial change in the previous results. On the other hand, although the PPV, or precision – number of selected predictions that are true positive – is the reference to compare different methods in the field of protein structure prediction, it is maybe not the best way to asses the predictive properties of a model.

Notice that the most advanced techniques in protein structure prediction – which are the most effective in separating signal from noise – are meta methods, such as PconsC2 [106], including also the vicinity of a potential contact, secondary-structure predictions etc. Interestingly, they are mostly improving the contact prediction between secondary structures – *i.e.* filling the vicinity of the predicted contact map – but frequently not adding new structurally informative contacts. Such meta methods display much better PPV than usual approaches, but the gain of information is rather limited.

2.6 OUTLOOK

DCA exploits the statistical correlations implied by coevolution in protein multiple sequence alignments to infer residue-residue contacts within the tertiary structure. The probabilistic model takes the form of a $q = 21$-states Potts model, whose parameters are inferred to reproduce the one- and two-residue statistics of the data. Usually, the inferred coupling matrices $\{J_{ij}(a, b)\}$ are mapped onto scalar pa-

rameters to measure the coupling strength between two residues and thereby predict contacts, without exploring the full information they contain.

By studying extensively 70 Pfam protein families, we show that these couplings reflect the physico-chemical properties of amino-acid interactions, such as electrostatic, hydrophobic/hydrophilic, Cysteine-Cysteine and steric interactions. Some of these interaction modes are present in a small fraction of residue pairs only, and are not easily seen in the global analysis over the 70 protein families. We show, however, that Cysteine-Cysteine and hydrophilic signals are unveiled, when we consider the SCOP structural classification (small proteins) and solvent exposure (surface contacts).

Using this detailed information to improve structural predictions is quite challenging. If interesting features are displayed by the projection of the coupling matrices onto the Miyazawa-Jernigan matrix and its spectral modes, we could not quantitatively improve the results obtained with the standard APC score. This might be explained on one hand by the presence of homo-oligomer contacts, which are not in contact in the tertiary structure, but in the quaternary assembly of multiple folded protein subunits. Source of coevolution, these false-positive are detected by DCA. This stresses how important is the definition of contacts and structure. On the other hand, as will be extensively discussed in the next chapter, sampling plays a crucial role in the information contained in the coupling matrices. The limited sampling in real proteins drastically restricts their potentiality.

Nevertheless, even at the current state of sequence sampling, the coupling matrices contain important quantitative information which can directly be implemented into protein-structure prediction: our work indicates that the type of interaction reflected by the inferred couplings is correlated with the distances in the tertiary structure between the residues in contact. Cysteine-Cysteine tend to form very strong chemical bonds such as disulfide bridges and therefore are the only contact type associated to very short distances $\sim$ 2 Å. Electrostatic contacts give rise to distances with a bimodal distribution, centered around 2.7 Å and 3.5 Å. Finally, hydrophobic contacts are mainly located around 3.5 Å.

3

LATTICE PROTEINS

Introduced in Part I, lattice proteins (LP) are exactly solvable models of proteins, folding on a 3D lattice into a compact conformation given by a self-avoiding walk on a cube of dimension $3 \times 3 \times 3$ [104]. Real proteins and LP share many common properties (efficient folding, non trivial statistical features, existence of families in the profile-HMM sense with conserved folds, etc.), but LP as *in silico* systems allow for precise numerical control. It is easy to generate even large samples of sequences (MSA) corresponding to a single fold, defining the equivalent of a protein family, without any phylogenetic sampling bias. LP are therefore an ideal benchmark – in a relatively realistic and fully controllable context – for studying and better understanding inference methods developed in the context of real protein data [60].

After presenting the dataset and a short analysis of the profile-HMM specificity of LP, we will hereafter use the LP framework to study in detail the effect of sampling quality *vs.* regularization strength in the inference of the coevolutionary couplings $\{J_{ij}(a, b)\}$.

3.1 DATASET AND BACKGROUND

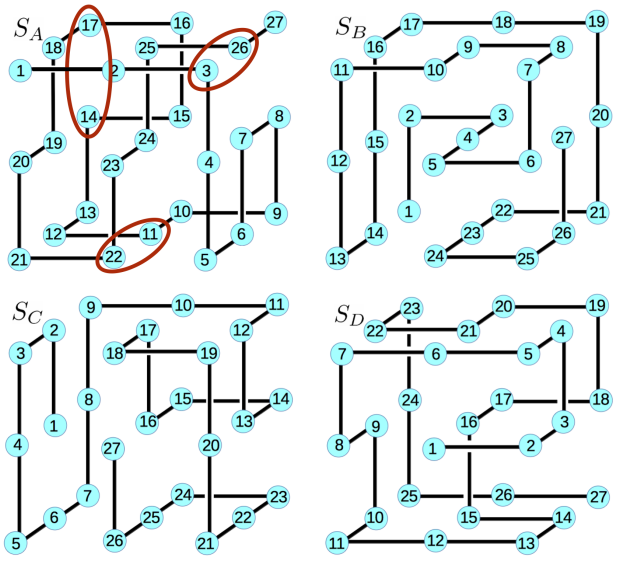


Figure 3.1 – Four representative LP structures used for the analysis. Three among the 28 contacts of structure S_A have been circled in the top left panel.

We consider four LP folds in this chapter, S_A, S_B, S_C, S_D, taken from [60] and shown on Fig. 3.1. MSAs corresponding to these folds have been generated by Monte Carlo Markov Chain (MCMC) sampling, each containing $B = 50000$ sequences folding with probability $P_{nat} > 0.995$. The same inverse methods based on maximum-entropy and Potts modeling used for real proteins (mfDCA, plmDCA and ACE) can applied to infer the pairwise couplings $J_{ij}(a,b)$ from the empirical one- and two-point statistical correlations measured on the MSA of the lattice proteins.

Below are a few reminders of the main results presented in Part I Chapter 3 in more details. As in real data, inferred couplings are excellent predictors of contacts in the structure. Interestingly, a linear dependency is moreover observed between the inferred couplings $J_{ij}(a,b)$ and MJ energetic parameters $E_0^{MJ}(a,b)$ used to compute the energy, both in the zero-sum gauge and for a given residue pair (i,j): $J_{ij}(a,b) \approx \lambda_{ij} E_0^{MJ}(a,b)$. We display this dependency for fold S_B on Fig. 3.2. The prefactor λ_{ij} is interpreted in [60] as a measure of the coevolutionary pressure on the residues (i,j), due to the design of the native structure. Large positive λ_{ij} indicate positive design, and generally correspond to residues (i,j) in contact in the native structure, but not in its competitor folds S'. Conversely, large negative λ_{ij} reflect negative design and generally correspond to residues (i,j) in contact in competitor structures but not in the native structure.

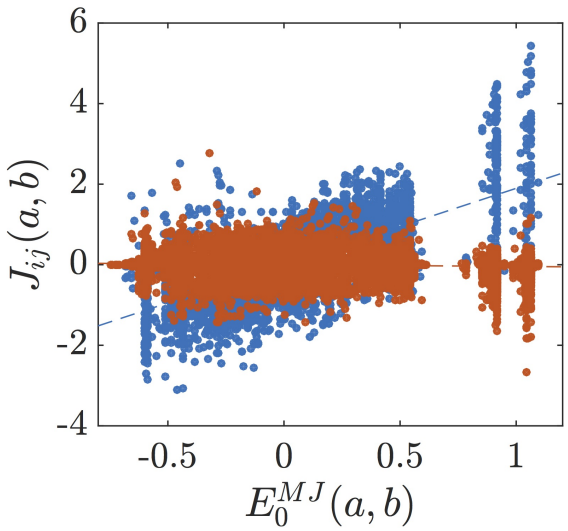


Figure 3.2 – Couplings $J_{ij}(a,b)$ for fold S_B *vs.* the MJ matrix $E_0^{MJ}(a,b)$ across all residue (i,j) and amino-acid (a,b) pairs, inferred with plmDCA. Blue dots correspond to pairs in contact, while red dots correspond to pairs not in contact. Broken lines: linear fits for contacts (blue, slope ~ 1.90) and not in contact (red, slope ~ -0.04). Person correlation for contacts is 0.69 and for non contacts is -0.11.

3.2 PROFILE-HMM SPECIFICITY OF LATTICE PROTEINS

Lattice proteins share many common features with real protein sequences, including specificity of their profile-HMM [40], which will be illustrated in this section. For each of the four studied folds, a profile-HMM is built on a sub-alignment of 1000 sequences. The percentage of hits – or detected homologs – found in each of the four alignments is then recorded, using the *hmmsearch* command of the HMMer software [47]. The tables below display the percentage of hits detected on the target alignments containing the natural sequences of $N = 27$ residues (Tab. 6), and on target alignments of natural sequences with 15-sites random sequences appended at the beginning and the end (Tab. 7). The latter configuration aims at modeling full length sequences, similar to real sequences data available in the Uniprot database.

	S_A	S_B	S_C	S_D
HMM S_A	**84.09**	0.002	0	0
HMM S_B	0	**76.43**	0	0
HMM S_C	0.049	0	**83.52**	0
HMM S_D	0.007	0.017	0	**85.41**

Table 6 – Percentage of hits detected on the target alignments of natural sequences.

	S_A	S_B	S_C	S_D
HMM S_A	**74.45**	0.026	0.002	0
HMM S_B	0.014	**66.50**	0.002	0.004
HMM S_C	0.053	0.010	**67.10**	0.008
HMM S_D	0.011	0.065	0.017	**71.84**

Table 7 – Percentage of hits detected on the target alignments of natural sequences with 15-sites random sequences appended at the beginning and at the end.

Notice that a profile-HMM built on a subpart of a MSA associated to a given fold is very family-specific, and gives high scores to sequences with a high P_{nat} for this fold. Sequences belonging to other families have such lower scores that almost none of them is reported as homologous. The percentage of hits in the alignment corresponding to the profile-HMM decreases when random sub-sequences are appended at the beginning and at the end of the natural sequence. Moreover, the number of hits in the other alignments increases. This is expected, as the LP "domain" is harder to detect from much longer sequences.

3.3 PROPERTIES OF THE INFERRED COUPLINGS

For each fold, the coupling matrices are computed using plmDCA in zero-sum gauge (as in Chapter 2) for four different values of the sampling and regularization parameters:

— large sample size ($B = 50000$ sequences) and strong regularization ($\gamma = 10^{-2}$, standard value for plmDCA),
— large sample size ($B = 50000$ sequences) and weak regularization ($\gamma = 1/B = 2 \times 10^{-5}$),
— small sample size ($B = 500$ sequences extracted from the MSA) and strong regularization ($\gamma = 10^{-2}$),
— small sample size ($B = 500$ sequences extracted from the MSA) and weak regularization ($\gamma = 10^{-4}$).

As mentioned in Section 3.1, the inferred coupling matrices are closely related to the MJ potential, but varying the sampling and regularization strength provides interesting insights. The default regularization parameter is set in plmDCA to the value $\gamma = 10^{-2}$, as it gives the best results for contact prediction [42]. This regularization strength penalizes large couplings and sparsifies the 20×20 matrix. With smaller regularization penalties, $\gamma = 10^{-5} - 10^{-4}$, couplings can acquire larger values.

3.3.1 *Effect of the regularization*

Figure 3.3 displays the coupling matrix $J_{14,17}$ of a representative residue pair (14,17) in contact in structure S_A (Fig. 3.1) at strong ($\gamma = 10^{-2}$, panel (a)) and weak ($\gamma = 1/B = 2 \times 10^{-5}$, panel (b)) regularizations. Left and bottom colorbars are single site frequencies f_{14} and f_{17}, and red squares indicate zero frequency. The characteristics of the mean coupling matrix will be described in Section 3.4.

Strikingly, decreasing the regularization strength enables new interaction signals to emerge, *e.g.* hydrophobic and Cysteine-Cysteine interactions, which are consistent with the MJ matrix, *cf.* panel (a) of Fig. 1.1. The correlation between $J_{ij}(a,b)$ and $E_0^{MJ}(a,b)$ for all (i,j) in contact in the four studied folds therefore increases, with an average Pearson coefficient raising from 0.51 (strong regularization) to 0.70 (weak regularization).

The unveiling of interactions at weak regularization depends, however, on the amino-acid statistics on the involved sites. For example, for the pair $(14,17)$ displayed on Fig. 3.3, electrostatic and hydrophilic amino acids (H to G) have sufficiently large frequencies on sites 14 and 17 to produce enough correlation statistics for the corresponding interaction. On the contrary, no interaction signal is revealed at low regularization for amino acids F, I and L, as they are never found on site 17 (vertical band of zero couplings on panel (b)). Decreasing the

regularization in the latter case merely results in increasing noise, as discussed in the next subsection.

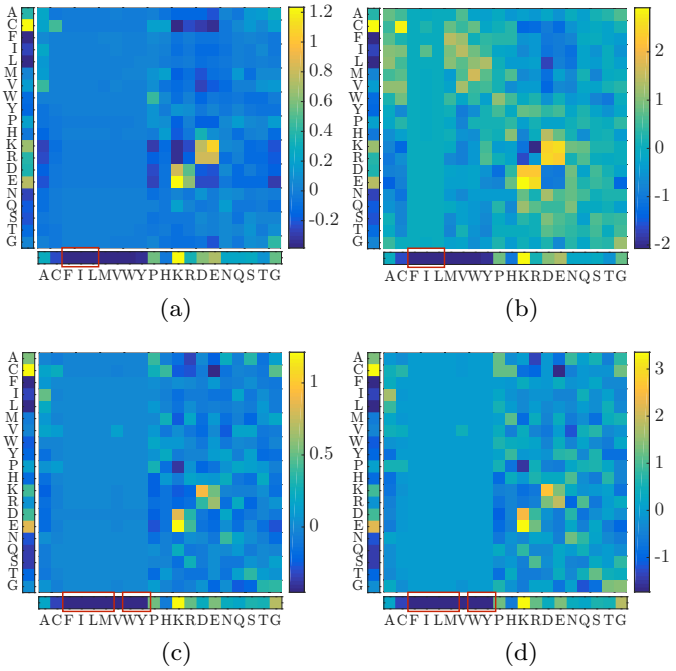


Figure 3.3 – Coupling matrices of pair (14,17), structure S_A. Left and bottom colorbars are single site frequencies f_{14} and f_{17}. Red squares indicate zero frequency. (a) $B = 50000$, $\gamma = 10^{-2}$, (b) $B = 50000$, $\gamma = 2 \times 10^{-5}$, (c) $B = 500$, $\gamma = 10^{-2}$, (d) $B = 500$, $\gamma = 10^{-4}$.

3.3.2 *Effect of the sampling*

The length of LP is $N = 27$ residues, which is small compared to real biological proteins (typically $50 - 500$ amino acids in a single domain). Moreover, the MCMC procedure used to generate MSAs ensures that the sequences are well distributed in sequence space. In consequence, inference based on good sampling ($B = 50000$ sequences) becomes very accurate. The situation for real biological sequences is less optimal, as the effective number of sequences B_{eff} is much smaller (we have chosen $B_{eff} = 500$ as a lower bound for the 70 Pfam families studied in the present work), and only very few proteins reach values close to $B = 50000$.

To test our analysis in a more realistic situation, we therefore select sub-alignments of $B = 500$ sequences for each of the four structures. The bottom panels of Fig. 3.3 display the coupling matrices obtained in this poor sampling situation, at strong (panel (c)) and weak (panel (d)) regularizations. Contrary to the good sampling case, no new interaction signal compatible with MJ is revealed at low regularization. Globally, the coupling matrices of all residue pairs in contact are even less correlated with MJ, as the Pearson correlation goes from

0.42 (small sample size, strong regularization) down to 0.36 (small sample size, weak regularization). The difference between couplings at strong and weak regularization seems to be due to noise for poor sampling.

In the last chapter, the couplings for real protein sequences have been inferred at (plmDCA standard) high regularization ($\gamma = 10^{-2}$). Consistently with what has been described in the last paragraph, and since real biological sequences are not very well sampled ($B_{eff} \simeq 500 - 1000$), decreasing the regularization does not change the mean matrices $e(a, b)$ and their spectral modes previously defined; they contain simply more noise.

SAMPLING	REGULARIZATION	CORRELATION
$B = 50000$	$\gamma = 10^{-2}$	0.51 / −0.15
	$\gamma = 1/B$	0.70 / −0.14
$B' = 500$	$\gamma = 10^{-2}$	0.42 / −0.05
	$\gamma = 10^{-4}$	0.36 / −0.04

Table 8 – Pearson correlation coefficients between $J_{ij}(a, b)$ and the MJ energy matrix $E_0^{MJ}(a, b)$ across all residue pairs (contacts / non contacts) in the four studied folds for different sample sizes and regularization strengths

To sum up the effects of the different parameters (regularization and sampling), Table 8 gathers the Pearson correlation coefficients between $J_{ij}(a, b)$ and $E_0^{MJ}(a, b)$ for all amino-acid and residue pairs in the 4 studied folds ($4 \times 28 = 112$ pairs). As we have discussed above, with a good sampling, the correlation between $J_{ij}(a, b)$ and $E_0^{MJ}(a, b)$ globally increases when the regularization decreases. On the contrary, with poor sampling (as it is the case for real biological data), the correlation slightly decreases when the regularization decreases. However, the inferred signal appears pretty stable at strong regularization, which may be a reason why plmDCA needs this high regularization on real protein data.

3.4 MEAN COUPLING MATRIX

Similarly to what has been done for real sequences data in Chapter 2, we compute the mean matrix

$$e(a, b|LP) = \left\langle J_{ij}(a, b) \right\rangle_{ij} , \tag{3.1}$$

where the mean $\langle . \rangle_{ij}$ is over all residues pairs in contact in the four studied folds ($28 \times 4 = 112$ coupling matrices). The four cases of different sampling and regularization parameters defined in Section 3.3

give rise to four different matrices $e(a,b|LP)$: $(B = 50000, \gamma = 10^{-2})$, $(B = 50000, \gamma = 1/B)$, $(B' = 500, \gamma = 10^{-2})$, and $(B' = 500, \gamma = 10^{-4})$, which are displayed on Fig. 3.4 to 3.7, along with their spectrum and spectral modes.

Consistently to what has been previously stated, the correlation between $e(a,b|LP)$ and the MJ energy matrix E_0^{MJ} is maximum (0.94) in the case of large sample size and weak regularization (Fig. 3.5). Table 9 displays the Pearson correlation coefficients between $e(a,b|LP)$ in the four cases (panels (a) of the figures) and the MJ energy matrix E_0^{MJ}.

SAMPLING	REGULARIZATION	CORRELATION
$B = 50000$	$\gamma = 10^{-2}$	0.76
	$\gamma = 1/M$	0.94
$B' = 500$	$\gamma = 10^{-2}$	0.74
	$\gamma = 10^{-4}$	0.72

Table 9 – Pearson correlation coefficients between the mean matrices $e(a,b|LP)$ and the MJ energy matrix $E_0^{MJ}(a,b)$ for different samplings and regularization strengths.

Interestingly, the regularization strength seems to play an important role in determining the order of magnitude of the entries of the matrix $e(a,b|LP)$ and its dominant eigenvalues. With a fixed sampling $B = 50000$, the top eigenvalues are divided by 5 with the regularization going from $\gamma = 10^{-2}$ to $\gamma = 2 \times 10^{-5}$ (*cf.* panels (b) of Fig. 3.4 and 3.5). On the contrary, decreasing B at fixed regularization does not affect the top eigenvalues (*cf.* panels (b) of Fig. 3.4 and 3.6).

In the optimal case of large sample size and weak regularization, where the correlation with the MJ energy matrix is maximal (*cf.* Table 9), the entries of $e(a,b|LP)$ and its top eigenvalues are larger than the MJ energy matrix (*cf.* Fig. 1.1). The presence of negative and positive designs indeed causes the inferred couplings to be larger. It illustrates the strong influence of the evolutionary pressure and positive/negative design in LP [60].

The situation for real proteins is less stable, as structure is only partially conserved over protein families, and contacts stabilizing a structure may not always be exactly the same across thousands of distant homologs. This probably explains why the entries and top eigenvalues of the mean coupling matrix $e(a,b)$ are much smaller in real proteins than in the MJ energy matrix.

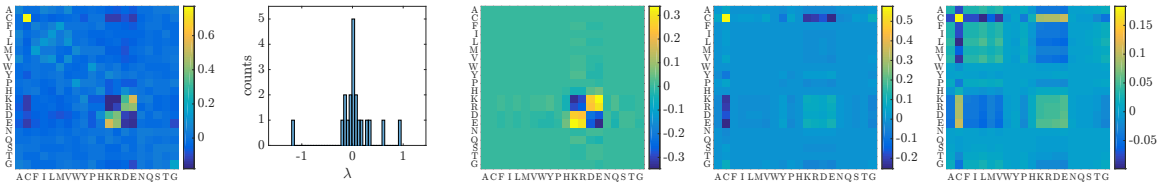


Figure 3.4 – $(B = 50000, \gamma = 10^{-2})$. (a) mean matrix $e(a, b|LP)$ over all residue pairs in contact across the 4 studied fold. (b) Histogram of the spectrum of $e(a, b|LP)$. (c), (d), (e) First spectral modes of $e(a, b|LP)$ displaying electrostatic, Cysteine-Cysteine, and mixed Cysteine-Cysteine/hydrophobic/hydrophilic interactions.

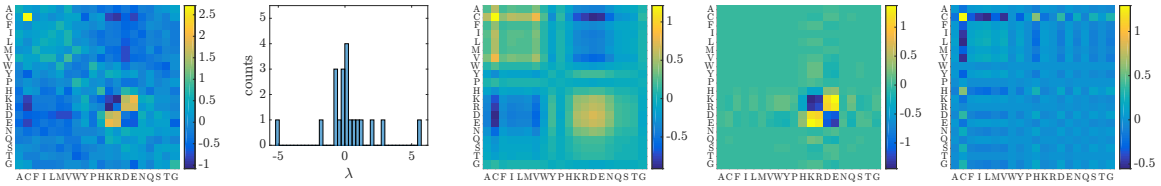


Figure 3.5 – $(B = 50000, \gamma = 1/B = 2 \times 10^{-5})$. (a) mean matrix $e(a, b|LP)$ over all residue pairs in contact across the 4 studied fold. (b) Histogram of the spectrum of $e(a, b|LP)$. (c), (d), (e) First spectral modes of $e(a, b|LP)$ displaying electrostatic, Cysteine-Cysteine, and hydrophobic/hydrophilic interactions.

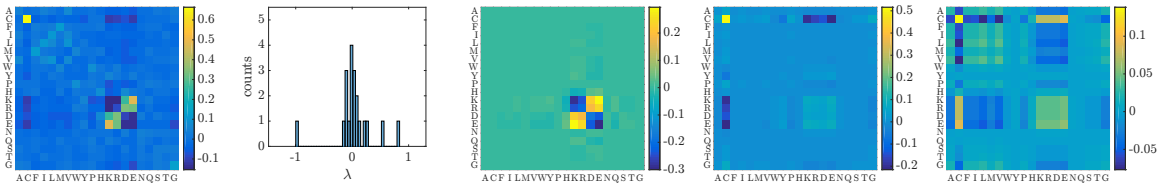


Figure 3.6 – $(B' = 500, \gamma = 10^{-2})$. (a) mean matrix $e(a, b|LP)$ over all residue pairs in contact across the 4 studied fold. (b) Histogram of the spectrum of $e(a, b|LP)$. (c), (d), (e) First spectral modes of $e(a, b|LP)$ displaying electrostatic, Cysteine-Cysteine, and mixed Cysteine-Cysteine/hydrophobic/hydrophilic interactions.

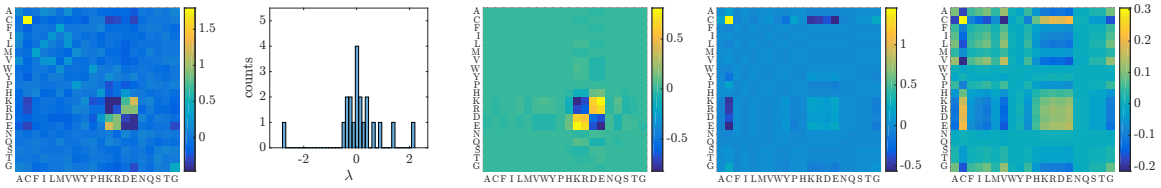


Figure 3.7 – $(B' = 500, \gamma = 10^{-4})$. (a) mean matrix $e(a, b|LP)$ over all residue pairs in contact across the 4 studied fold. (b) Histogram of the spectrum of $e(a, b|LP)$. (c), (d), (e) First spectral modes of $e(a, b|LP)$ displaying electrostatic, Cysteine-Cysteine, and mixed Cysteine-Cysteine/hydrophobic/hydrophilic interactions.

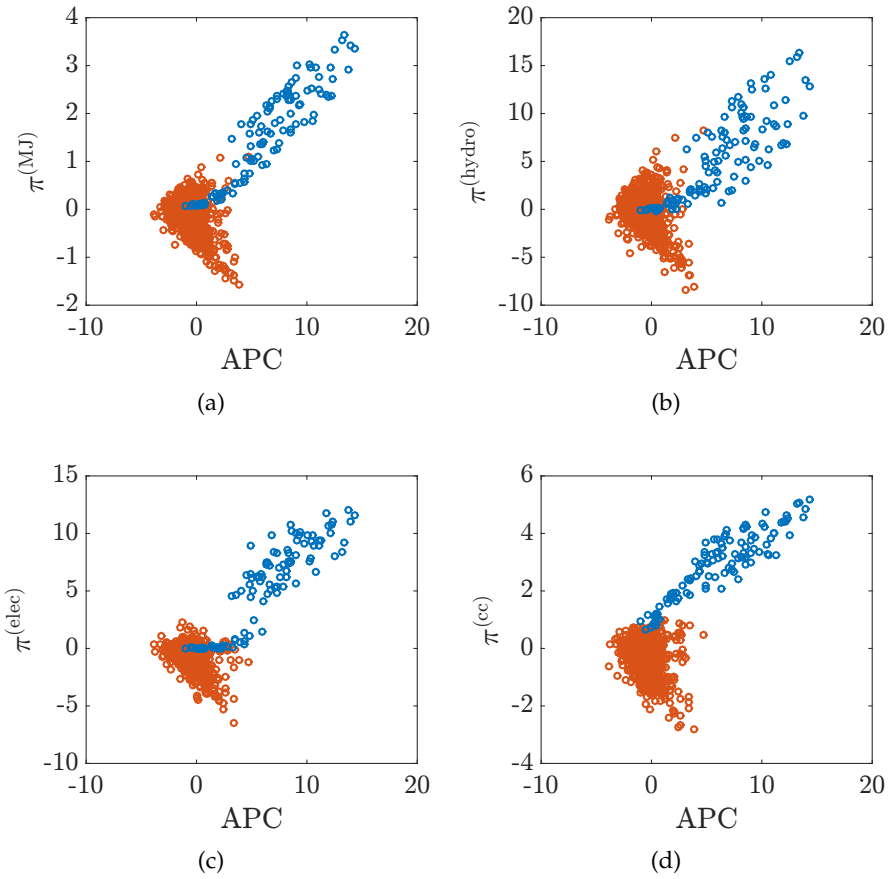
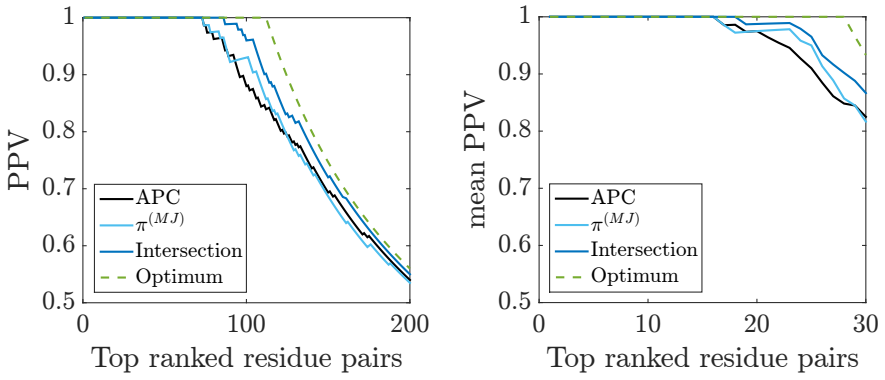


Figure 3.8 – Projections of the coupling matrices $J_{ij}(a, b)$ onto the MJ matrix as a function of the APC scores, across all residue pairs in contact (blue) and not in contact (red) in the four LP folds. Inference is performed on $B = 50000$ sequences and with a weak regularization strength.

3.5 STRUCTURAL PREDICTIONS

Recent work by our team at ENS [60] shows that the projection of the couplings onto the MJ matrix improves the precision of the contact prediction compared with the usual APC score. Fig. 3.8 displays the projection $\pi^{(k)}$ of all coupling matrices from the four folds onto the MJ matrix and its first spectral modes as a function of the APC score (equivalently to Fig. 2.18 for real proteins). The coupling matrices have been inferred with plmDCA at low regularization strength. In this case – and contrary to real proteins – the separation between contacts (blue) and non contacts (red) is clearer in terms of projection than of APC, the former therefore being a better classifier.

The reason is twofold. First the projection, contrary to the Frobenius norm, has a sign, and allows for the distinction of positive design (positive projection, likely to correspond to contact in the native fold) from negative design (negative projection, likely not to correspond to



(a) (b)

Figure 3.9 – Positive predicted values as a function of the number of predictions all folds taken together (panel (a)), or averaged over the folds (panel (b)). Various classifiers are displayed: standard APC score (black), projection $\pi^{(MJ)}$ (light blue), intersection of the APC and $\pi^{(MJ)}$ classifiers (dark blue), and the optimal theoretical classifier (dashed green).

a contact, but with a large APC). Secondly the projection measures the magnitude of the coupling matrix along one direction in the 20×20-dimensional space of amino-acid pairs, and is thus not sensitive to the noise in the 399 remaining orthogonal directions, contrary to APC.

Fig. 3.9 displays the precision (or PPV, number of true predictions divided by the total number of predictions) for structural predictions with different scores. As already shown in [60], the projection $\pi^{(MJ)}$ (light blue) is indeed more efficient than the standard APC (black). Because they usually do not display the same false positives, combining both methods further improves the score. The "intersection" of both classifiers – the first n residue pairs are predicted in contact if they belong to the top ranked n scores of both APC and the projection – is naturally better (dark blue), approaching the perfect classifier (dashed green).

3.6 OUTLOOK

Study of lattice proteins (LP) – synthetic protein models folding on a 3D lattice with energetics ruled by the Miyazawa-Jernigan statistical potential – gives useful insights on the effect of regularization strength and sampling on contact classes. Decreasing the regularization strength (from the default plmDCA value $\gamma = 10^{-2}$ to $\gamma = 1/B$, where B is the sample size) allows for a richer interaction signal to emerge in the coupling matrices, highly correlated with the Miyazawa-Jernigan energy matrix. However, this rich interaction pattern may be inferred only if the sequence sample is sufficiently large.

For sample sizes representative of current real protein databases, decreasing the regularization strength simply makes the correlation with the Miyazawa-Jernigan energy matrix worse, as the inferred couplings merely reproduce the sampling noise in the amino-acid pairwise correlations. With such poor sampling, strong regularization is more reliable: the inferred interaction signal becomes relatively insensitive to the sample size, explaining why plmDCA on real proteins was found to perform consistently with a constant regularization of $\gamma = 10^{-2}$. Note that this picture somewhat depends on the inference method considered: more precise inference procedures could allow for detecting a larger correlation with MJ even with poor sampling [64, 112]. The adaptive cluster expansion (ACE) [13], is on of these methods that will be presented in the next part of this dissertation.

The order of magnitude of the mean coupling matrices $e(a, b|\mathrm{LP})$ seem to be determined by both regularization – increasing the penalty induces strongly damped couplings – and evolutionary pressure. The presence of positive and negative designs indeed causes the inferred couplings to be larger, with the entries of $e(a, b|\mathrm{LP})$ exceeding the MJ matrix. The role of evolutionary pressure is less clear for real proteins, because of partially conserved structures and variety of stabilizing contacts. This could explain why the mean coupling matrix on real proteins $e(a, b)$ has much smaller entries and eigenvalues than the MJ potential.

If the detailed structure of the inferred couplings unveiled in this work can be use to improve structural predictions in the LP context, the applicability to real protein data appears currently limited due to two reasons. First, the projection in is done on the MJ matrix used in the generative model of the lattice proteins, *i.e.* complementary information not coming from the data is used. In real proteins, the reference coupling matrix has to be inferred from data first and is thus expected to be less accurate. Second, the currently limited sampling in real proteins was shown to impose a strong regularization during the inference of the DCA model parameters, which even in lattice proteins reduces the correlation between inferred couplings and the MJ matrix. We however anticipate this situation to improve soon due to the rapid growth of available genomic data, leading to a better and better sampling of protein families.

Part IV

ADAPTIVE CLUSTER EXPANSION

In this chapter, we focus on the Potts version of the adaptive cluster expansion (ACE) algorithm – presented in Part I and initially developed in the Ising case. The inference procedure is now adapted to the level of sampling in the data, both by proposing a compressed representation of this data and by inferring a sparse network omitting unsufficientely well sampled interactions. Chapter 1 shortly introduces two technical points: the analytical computation of the statistical errors on inferred Potts parameters due to finite sampling, and a compressed representation of the data. In Chapter 2, we then illustrate the ACE method – and compare it with standard direct-coupling analysis (DCA) approaches – on three artificial and biological datasets, and assess its ability to recover the true underlying parameters when known, reproduce the statistics of the input data, or predict structural contacts in protein family data. In Chapter 3, we finally explore in more details the compressed representation of the data and the effect of the compression parameter on the inference.

1 BACKGROUND

In this short chapter, we will address two technical points that will prove useful in the following. First, we will go into details describing the Fisher information matrix, that provides an interesting insight on the statistical errors on the inferred parameters due to finite sampling. Second, we will introduce a compressed representation of the data, where the number of explicitly modeled Potts states depends on the variable. It reduces the complexity of the inferred Potts models to the level of the sampling in the data, enabling to both decrease the computational time of ACE and reduce overfitting.

1.1 FISHER INFORMATION MATRIX AND FINITE SAMPLING ERRORS

1.1.1 *Expression of the finite sampling errors*

This section is mainly based on [28]

The Hessian of the cross-entropy – also called the Fisher information matrix – is defined through

$$\chi = \frac{\partial^2 S}{\partial J \partial J} = \begin{pmatrix} \chi_{ia,i'a'} & \chi_{ia,j'b'k'c'} \\ \chi_{jbkc,i'a'} & \chi_{jbkc,j'b'k'c'} \end{pmatrix} , \tag{1.1}$$

where $J = \{J_{ij}(a,b), h_i(a)\}$ denotes the Potts parameters. Its entries can be expressed as averages over the Potts Gibbs measure $\langle\cdot\rangle_J$:

$$\begin{aligned} \chi_{ia,i'a'} &= \langle \sigma_{ia}\sigma_{i'a'}\rangle_J - \langle\sigma_{ia}\rangle_J \langle\sigma_{i'a'}\rangle_J , \\ \chi_{ia,j'b'k'c'} &= \left\langle \sigma_{ia}\sigma_{j'b'}\sigma_{k'c'}\right\rangle_J - \langle\sigma_{ia}\rangle_J \left\langle\sigma_{j'b'}\sigma_{k'c'}\right\rangle_J , \\ \chi_{jbkc,j'b'k'c'} &= \left\langle \sigma_{jb}\sigma_{kc}\sigma_{j'b'}\sigma_{k'c'}\right\rangle_J - \left\langle\sigma_{jb}\sigma_{kc}\right\rangle_J \left\langle\sigma_{j'b'}\sigma_{k'c'}\right\rangle_J . \end{aligned} \tag{1.2}$$

With $x = \{x_{ia}, x_{ia,jb}\}$ an arbitrary $(N(N-1)/2q^2 + Nq)$-dimensional vector, the quadratic form

$$x^\dagger \cdot \chi \cdot x = \left\langle \left(\sum_{ia} x_{ia}(\sigma_{ia} - \langle\sigma_{ia}\rangle_J) + \sum_{\substack{i<j \\ ab}} x_{ia,jb}(\sigma_{ia}\sigma_{jb} - \langle\sigma_{ia}\sigma_{jb}\rangle_J) \right)^2 \right\rangle , \tag{1.3}$$

is semi-definite positive. The cross-entropy S is therefore a convex function, guaranteeing that it has a minimum.

χ can also be used to estimate the statistical deviations due to finite sampling B. If the data were generated by a Potts model with parameters J, the frequencies $f_i(a), f_{ij}(a,b)$ would obey a normal law with the covariance matrix $\frac{1}{B}\chi$. The typical uncertainties of the one- and

two-point frequencies therefore simply derive from the covariance matrix:

$$\begin{aligned} \delta f_i(a) &= \sqrt{\frac{1}{B}\chi_{ia,ia}} = \sqrt{\frac{\langle\sigma_{ia}\rangle_J\,(1-\langle\sigma_{ia}\rangle_J)}{B}}\,, \\ \delta f_{ij}(a,b) &= \sqrt{\frac{1}{B}\chi_{iajb,iajb}} = \sqrt{\frac{\langle\sigma_{ia}\sigma_{jb}\rangle_J\,(1-\langle\sigma_{ia}\sigma_{jb}\rangle_J)}{B}}\,. \end{aligned} \tag{1.4}$$

Practically, estimates of the expected deviations are obtained by replacing the Gibbs averages by the empirical averages $f_i(a)$ and $f_{ij}(a,b)$.

More interestingly, the *inverse* Fisher information matrix χ^{-1} can also be used to estimate the statistical fluctuations of the *inferred* parameters due to a finite number of sampled configurations B. According to the asymptotic theory of inference, in the limit of large B the cross-entropy obeys a normal law centered on the minimum of $S_{\text{Potts}}(J|\mathbf{f})$, and of covariance matrix $\frac{1}{B}\chi^{-1}$, with $\mathbf{f} = \{f_{ij}(a,b), f_i(a)\}$ denoting the data. Consequently the statistical errors on the couplings and fields are given by:

$$\begin{aligned} \delta h_i(a) &= \sqrt{\frac{1}{B}(\chi^{-1})_{ia,ia}}\,, \\ \delta J_{ij}(a,b) &= \sqrt{\frac{1}{B}(\chi^{-1})_{iajb,iajb}}\,. \end{aligned} \tag{1.5}$$

χ^{-1} is not necessarily well defined, and a regularization term $\gamma = 1/B$ needs to be included before inverting the Hessian:

See also Part I Section 2.3.2

$$\begin{aligned} \chi_{ia,ia} &\to \chi_{ia,ia} + \gamma\,, \\ \chi_{iajb,iajb} &\to \chi_{iajb,iajb} + \gamma\,. \end{aligned} \tag{1.6}$$

Removing the zero modes of χ also guarantees the uniqueness of $S_{\text{Potts}}(J|\mathbf{f})$. However, the regularization breaks the gauge invariance and the gauge choice may have an impact of the statistical errors. This problem will be addressed in Chapter 3.

1.1.2 *Approximated errors on the inferred parameters*

The inversion of χ is computationally feasible only for a small system size N and Potts states number q, as it is of size $\left(qN + q^2N(N-1)\right) \times \left(qN + q^2N(N-1)\right)$. Typical values for proteins being $N \approx 100$ and $q = 21$, χ is in this case a $10^6 \times 10^6$ matrix, computationally impossible to invert. We can however approximate the statistical errors in a two-variable system, where the analytical expressions of the couplings and fields are known. In the following we will derive the approximate errors in the consensus gauge, as it will be the chosen gauge for the study in Chapter 3.

1.1.2.1 *Two-variable couplings and fields*

A simple calculation gives the couplings and fields for a system of two variables:

$$\begin{aligned} h_i(a) &= \log f_i(a) \,, \\ J_{ij}(a,b) &= \log \frac{f_{ij}(a,b)}{f_i(a) f_j(b)} \,. \end{aligned} \tag{1.7}$$

Assuming this analytical expression, the couplings and fields write in the consensus gauge:

cf. Eq. (2.31) in Part I

$$\begin{aligned} h_i(a) =& \log f_i(a) - \log f_i(c_i) \\ &+ \sum_{j=1}^{N} \left(\log \frac{f_{ij}(a,c_j)}{f_i(a) f_j(c_j)} - \log \frac{f_{ij}(c_i,c_j)}{f_i(c_i) f_j(c_j)} \right) \,, \\ J_{ij}(a,b) =& \log f_{ij}(a,b) - \log f_{ij}(c_i,b) - \log f_{ij}(a,c_j) + \log f_{ij}(c_i,c_j) \,. \end{aligned} \tag{1.8}$$

1.1.2.2 *Approximated errors*

Following Eq. (1.8), the approximate variances for inferred fields and couplings due to finite sampling in consensus gauge are:

$$\begin{aligned} \sigma_{h_i(a)} =& (N-2) \frac{1 - f_i(a)}{B f_i(a)} \\ &+ (N-2) \frac{1 - f_i(c_i)}{B f_i(c_i)} \sum_{j \neq i} \left(\log \frac{f_{ij}(a,c_j)}{f_i(a) f_j(c_j)} - \log \frac{f_{ij}(c_i,c_j)}{f_i(c_i) f_j(c_j)} \right) \,, \\ \sigma_{J_{ij}(a,b)} =& \frac{1 - f_{ij}(a,b)}{B f_{ij}(a,b)} + \frac{1 - f_{ij}(c_i,b)}{B f_{ij}(c_i,b)} + \frac{1 - f_{ij}(a,c_j)}{B f_{ij}(a,c_j)} + \frac{1 - f_{ij}(c_i,c_j)}{B f_{ij}(c_i,c_j)} \,, \end{aligned} \tag{1.9}$$

and the approximate statistical errors are given by

$$\begin{aligned} \delta h_i(a) &= \sqrt{\sigma_{h_i(a)}} \,, \\ \delta J_{ij}(a,b) &= \sqrt{\sigma_{J_{ij}(a,b)}} \,. \end{aligned} \tag{1.10}$$

Note that of all possible gauge states c_i (such that $h_i(c_i) = J_{ij}(c_i, b) = 0$), the consensus (of maximum frequency $f_i(c_i)$) states gives the lowest statistical errors.

1.1.3 *Absolute and relative errors between true and inferred couplings*

Suppose that the true Potts parameters are known, we define the absolute error between true and inferred parameters as

$$\begin{aligned} \Delta h &= \sqrt{\frac{1}{qN} \sum_i \sum_a \left(h_i^{inf}(a) - h_i^{true}(a) \right)^2} \,, \\ \Delta J &= \sqrt{\frac{1}{q^2 N(N-1)/2} \sum_{i<j} \sum_{ab} \left(J_{ij}^{inf}(a,b) - J_{ij}^{true}(a,b) \right)^2} \,. \end{aligned} \tag{1.11}$$

Given the finite sampling errors (*cf.* Eq. (1.5) and the two-site approximation Eq. (1.10)), the relative errors between true and inferred parameters write

$$\begin{aligned}\epsilon_h &= \sqrt{\frac{1}{qN}\sum_i\sum_a \frac{\left(h_i^{inf}(a)-h_i^{true}(a)\right)^2}{\delta h_i(a)^2}}\,,\\ \epsilon_J &= \sqrt{\frac{1}{q^2N(N-1)/2}\sum_{i<j}\sum_{ab}\frac{\left(J_{ij}^{inf}(a,b)-J_{ij}^{true}(a,b)\right)^2}{\delta J_{ij}(a,b)^2}}\,.\end{aligned} \tag{1.12}$$

1.2 COMPRESSED REPRESENTATION OF THE DATA

The number of Potts states each variable may take on is not necessarily the same for all variables. States with zero or very small frequencies may be very few observed in real, finitely-sampled data. The relative error on the corresponding frequencies and correlations due to finite sampling is large. For instance, several columns of a protein multiple sequence alignment (MSA) may contain much less than 21 symbols because of functional constraints, but also of finite size effects due to the limited number of available sequences.

We describe here a restricted Potts model where the number of states q_i depends on the site i. A total (before the gauge reparametrization) number of $\left(\sum_{i=1}^N q_i + \sum_{i<j} q_i q_j\right)$ parameters (fields and couplings) are therefore inferred, instead of $\left(Nq + q^2N(N-1)/2\right)$, wich can be much larger. Reducing the number of Potts states per site to a minimal number therefore limits the overfitting and reduces the computational time ($\mathcal{O}(q_i^l)$ operations instead of $\mathcal{O}(q^l)$, with $q_i < q$ and l the cluster size in the expansion, see Part I Chapter 2). To do so, infrequently observed states are effectively grouped together according to a given compression parameter.

Two conventions for a compressed representation of the data can be implemented. First, for each variable, only the k states observed with a frequency larger than a cutoff value

$$f_i(a) > f_0\,, \tag{1.13}$$

are explicitly modeled, while all the $q-k$ low frequency states are grouped together into the same state. Alternatively, ordering the states decreasingly by their contribution to the total single-site entropy S_i^q, only the first k states are treated explicitly to capture a cutoff fraction η_0 of S_i^q:

$$\begin{aligned}S_i^k &= -\sum_{a=1}^k f_i(a)\log f_i(a) - \left(1-\sum_{a=1}^k f_i(a)\right)\log\left(1-\sum_{a=1}^k f_i(a)\right)\\ &\geqslant \eta_0 S_i^q\,,\end{aligned} \tag{1.14}$$

with the remaining $q - k$ states regrouped into one last state. In the following, we will focus on the first convention which can be used to explore a wide range of compression regimes. The entropy-based scheme easily gives rise to strong compressions, relevant only in poor sampling situations (*cf.* Chapter 3).

The frequency of the regrouped Potts state is then the sum of the frequencies of the states which have been grouped together: $f_i(k+1) = \sum_{a=k+1}^{q} f_i(a)$. With these notations, the final number of Potts states at site i is $q_i = k+1$. Once the restricted Potts model is inferred, the complete (q states) model can be recovered by modifying the fields of the regrouped state a':

$$h_i(a') = h_i(k+1) + \log\left(\frac{f_i(a')}{f_i(k+1)}\right) , \tag{1.15}$$

but keeping the same value for the corresponding couplings

$$J_{ij}(a', b) = J_{ij}(k+1, b) . \tag{1.16}$$

The fields for the zero frequency states are still fixed from regularization alone.

The inevitable loss of information induced by the state compression – at least when non zero frequency states are regrouped – is balanced by a huge gain in computational time of the ACE algorithm, as well as a reduced overfitting leading to an improved inference quality. The choice of the compression parameters f_0 and η_0 along with the effects of this compressed representation of the data will be extensively discussed in Chapter 3.

2

COMPARISON WITH STANDARD METHODS ON VARIOUS DATASETS

In this chapter, we apply the Potts version of ACE to various datasets and asses its ability to recover the true underlying model parameters when known, reproduce the statistics of the input data, and achieve structural prediction of contacts in protein family data. A comparison with standard DCA approaches (mean-field and pseudolikelihood approximations) will also be performed.
Most of the results of this section have been recently published in *"ACE: adaptive cluster expansion for maximum entropy graphical model inference", JP Barton, E De Leonardis, A Coucke, and S Cocco, Bioinformatics (2016)*, [13].

2.1 DATASETS

We will focus in the following on three datasets. First, we study artificial data (**ER05**) from a Potts model with $q = 21$ states, where the network of interactions is described by an Erdős-Rényi random graph with $N = 50$ variables. Each edge in the interaction graph is included with probability 0.05. Field and coupling values for interacting pairs of sites are selected from a Gaussian distribution, with mean $\mu = 0$ and standard deviation $\sigma_J^2 = 1$ for couplings and $\sigma_h^2 = 5$ for fields. If i and j interact, J_{ij} is a 21×21 matrix whose elements are chosen according to the above distributions. $B = 10^4$ configurations are generated through Monte-Carlo sampling. The infrequently observed Potts states have been compressed with $f_0 = 0.05$ (*cf.* Eq. (1.13)). Chapter 3 provides discussion on the choice of f_0. The inference has been performed with and a regularization of $\gamma = 1/B = 10^{-4}$ in the gauge of the compressed state [1].

Second, we analyze the trypsin inhibitor protein family (**PF00014**), with $N = 53$ sites and $B = 4915$ sequences [29, 42, 82]. After reweighting in the frequency count the sequences with more than 80% identity, there are $B_{eff} = 2051$ sequences left. The rarely observed Potts states have been grouped together with $f_0 = 0.05$ (*cf.* Eq. (1.13)). The inference has been performed with regularizations $\gamma = 2/B_{eff} = 10^{-3}$ and $\gamma = 1$ in the minimum consensus gauge, where the least observed state per site is gauged to zero. Additionally, we noted in Part II that gaps in the MSA tend to be present in long stretches and are not generally modeled well in the Potts model representation with pair-

The reweighting procedure is described in Part I Section 2.3.2

1. In other words, the gauge symbol c_i is the compressed state, for more details about the gauge see Part I Section 2.3.1

wise interactions. Such stretches of highly correlated gaps slow down the inference procedure with ACE because they give rise to large clusters. Here we have processed the data to replace gaps by random amino acids with the same frequency as observed in the non-gapped sequences, which corresponds to the initial step of the procedure described in Part II Chapter 3.

Finally, in the framework of lattice proteins, we consider an alignment of $B = 5 \times 10^{-4}$ with $N = 27$ sites folding on a $3 \times 3 \times 3$ cube with structure $\mathbf{S_B}$ and with a probability $P_{nat} > 0.995$ [60, 104]. The never observed amino acids have been removed ($f_0 = 0$), and the inference has been performed in consensus gauge with regularization $\gamma = 5/B = 10^{-4}$.

See Part I Chapter 3 for definitions

2.2 RECOVERY OF THE ER05 PARAMETERS

In the following figures, the error bars correspond to the approximate statistical errors on the inferred parameters due to finite sampling, introduced in the last chapter at Eq. (1.10).

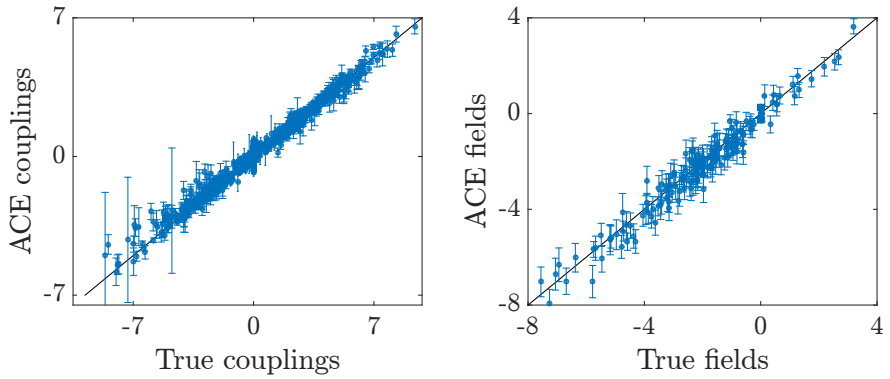


Figure 2.1 – ACE accurately recovers the the true fields (panel (a)) and couplings (panel (b)) corresponding to Potts states with $f_i(a) \geqslant 0.05$ for the ER05 model. Error bars denote approximated standard deviations in estimated parameters due to finite sampling (Eq. (1.10)).

Fig. 2.1 shows that the 2×10^4 underlying parameters of the ER05 model corresponding to the explicitly treated Potts states ($f_i(a) \geqslant 0.05$) are accurately recovered by ACE. The infrequently observed states being discarded by the compression, the remaining states are better sampled and therefore have smaller statistical uncertainties (*cf.* Eq. (1.4) in the last chapter). The error bars on the inferred parameters are thus fairly small. The model inferred by the standard plmDCA method at strong regularization and without compression contains around 10^6 parameters, which are compared to the true ones on panels (a) & (b) of Fig. 2.2. Those corresponding to the explicitly

modeled states in ACE (red dots) are recovered fairly well (with some errors in the fields), but parameters corresponding to compressed states are difficult to infer due to insufficient sampling.

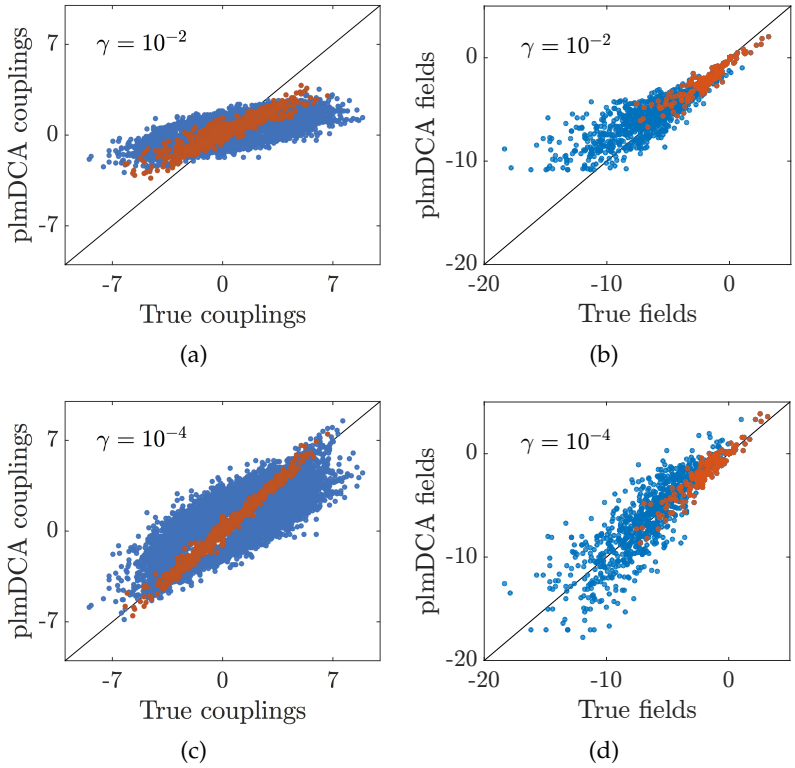


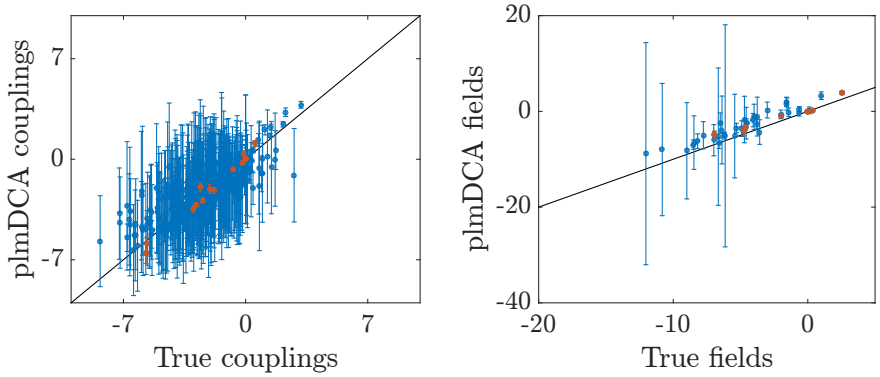
Figure 2.2 – plmDCA fairly well recovers the couplings (left panels) and fields (right panels) on explicitly modeled states (red dots), compressed states (blue dots) are hard to infer. Standard regularization $\gamma = 10^{-2}$ (top panels) gives less accurate results than weak regularization $\gamma = 1/B = 10^{-4}$ (bottom panels).

Panels (c) & (d) of Fig. 2.2 display the plmDCA couplings and fields with a weaker regularization $\gamma = 1/B = 10^{-4}$, similar to the one used in ACE. The recovery of the parameters is improved, but fields are still much less precisely inferred than with ACE. As expected, the statistical errors due to finite sampling are small on explicitly modeled states (red dots) and very large (especially for fields) on rarely observed states (blue dots), as shown on the error bars of Fig. 2.3. These poorly inferred parameters will have a huge impact on the generative properties of plmDCA (*cf.* Section 2.4 below).

2.3 INFERENCE OF STRUCTURAL CONTACTS FOR PF00014

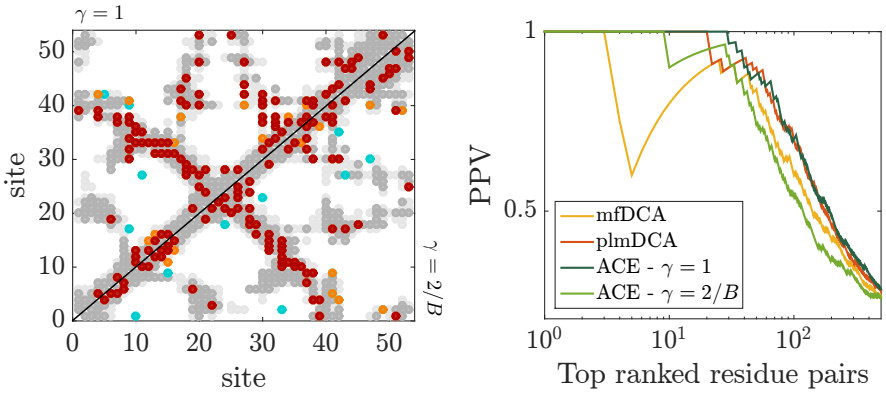
The inferred couplings are used to predict contacts in the tertiary structure of PF00014. We compare results from ACE and from standard contact prediction methods using direct-coupling analysis in the mean-field (mfDCA) and pseudolikelihood (plmDCA) approximations.

See Part I Chapter 3 for the method



(a) (b)

Figure 2.3 – plmDCA couplings and fields at weak regularization $\gamma = 1/B = 10^{-4}$ for sites $i = 2$ and $j = 4$. The statistical errors due to finite sampling (error bars) are small on explicitly modeled states (red dots) and very large on rarely observed states (blue dots).



(a) (b)

Figure 2.4 – Panel (a): Contact map for PF00014 inferred by ACE. The top 100 predicted contacts are shown, with true predictions in red and false predictions in blue. The remaining contact residues in the structure are shown in gray. Close contacts (< 6 Å) are darkly shaded and further contacts (< 8 Å) are lightly shaded. The upper triangular part is for strong regularization ($\gamma = 1$) and the lower triangular part is for weak regularization ($\gamma = 2/B$). Panel (b): Positive predictive values (PPV) as a function of the number of predictions, for various models. ACE is competitive with standard approaches.

We consider residues in contact if within 6 Å of each other in the structure, and we exclude trivial contacts along the protein backbone ($|j - i| \leqslant 4$). The accuracy in recovering the contact map with ACE can be increased by using a large regularization ($\gamma = 1$), consistently with standard approaches using strong regularization or pseudocount like mfDCA or plmDCA. Fig. 2.4 shows that ACE is competitive with DCA related approaches

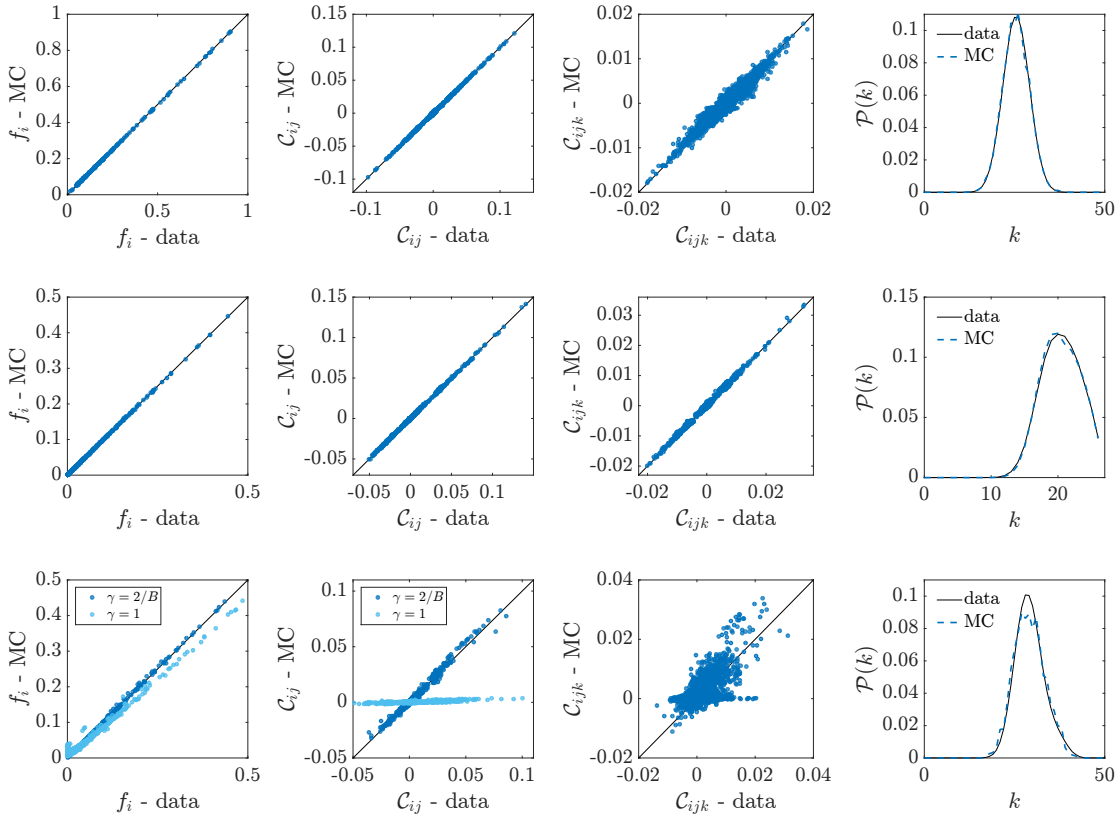


Figure 2.5 – Fit for ACE-inferred models describing ER05 (top panels), S_B (middle panels), and PF00014 (bottom panels). The frequencies, and two-point connected correlations are accurately reproduced. ACE also captures higher order correlations, such as the three-point connected correlations and the probability $\mathcal{P}(k)$ of observing a sequence with k mutations from the consensus configuration. A too strong regularization ($\gamma = 1$ for PF00014) strongly affects the generative properties of the model.

This very large regularization also induces strongly damped couplings and the ACE algorithm converges much faster, with smaller clusters. As the algorithm is typically slower than standard DCA approaches, it is important to notice that if one is interested in recovering protein structures, the speed can be increased by using such regularizations. However, the generative properties of the algorithm are lost (*cf.* Section 2.4 below).

2.4 REPRODUCIBILITY OF THE STATISTICS OF THE DATA

The generative properties of an inference method are assessed by its ability to reproduce the statistics of the input data. We therefore compute the statistical correlations of the model through Monte-Carlo (MC) sampling of a number of configurations from the inferred model and compare them to the true ones (from the training alignment). Four quantities are compared: the frequencies $f_i(a)$ and two-

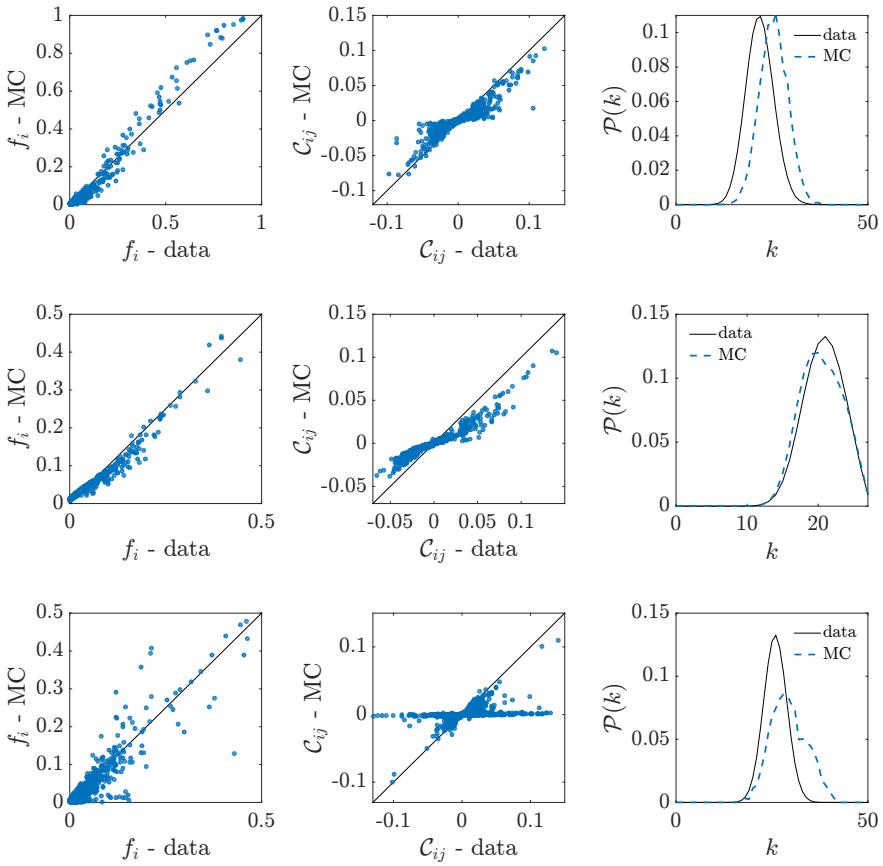


Figure 2.6 – Fit for plmDCA-inferred models describing ER05 (top panels), S_B (middle panels), and PF00014 (bottom panels). plmDCA is outperformed by ACE in reproducing the frequencies and two-point connected correlations, or the probability $\mathcal{P}(k)$ of observing a sequence with k mutations from the consensus configuration.

point connected correlations $\mathcal{C}_{ij}(a,b) = f_{ij}(a,b) - f_i(a)f_j(b)$ directly fitted by the inference method, and higher order statistics not directly taken into account by the model such as the three-point connected correlation[2] $\mathcal{C}_{ijk}(a,b,c)$ and the distribution $\mathcal{P}(k)$ of Hamming distances k between the sampled sequences and the consensus sequence[3].

Fig. 2.5 shows that ACE displays excellent generative properties: not only the one- and two-point statistics of the input data are accurately reproduced, but ACE gives very satisfactory results on the higher order functions. The bottom panels of Fig. 2.5 also display the fits for one- and two-point statistics with $\gamma = 1$. Although a very strong regularization increases the efficiency of ACE to recover the

2. The three-point connected correlation reads $\mathcal{C}_{ijk}(a,b,c) = f_{ijk}(a,b,c) - f_i(a)f_{jk}(b,c) - f_j(b)f_{ki}(c,a) - f_k(c)f_{ij}(a,b) + 2f_i(a)f_j(b)f_k(c)$.

3. $\mathcal{P}(k)$ is equivalently the probability of observing a sequence with k mutations from the consensus sequence, *i.e.* the configuration in which each site takes on the most probable value.

contact map of PF00014 (*cf.* Fig. 2.4), it causes an over damping of the inferred couplings and strongly affects the generative properties, as expected. Furthermore, the generative properties of ACE outperform plmDCA. Poorly inferred couplings and fields (*cf.* Fig. 2.2 & 2.3) indeed affect the generative properties of plmDCA, which cannot accurately reproduce the statistics of the input data. Even the frequencies are poorly recovered (*cf.* Fig. 2.6).

Besides, algorithms based on iterative rounds of MC simulation and Boltzmann machine learning (BML) are capable of inferring models that accurately reproduce the observed correlations, but they are typically slow to converge [1, 76, 112]. If one is interested in the generative properties of the inferred model, running such algorithms from a good initial guess of parameters, such as those obtained by ACE or even plmDCA, could help to accelerate the inference procedure. An illustration will be given in the next chapter.

See Part I Chapter 2 for more details

2.5 REPRODUCIBILITY OF THE ENERGY DISTRIBUTION

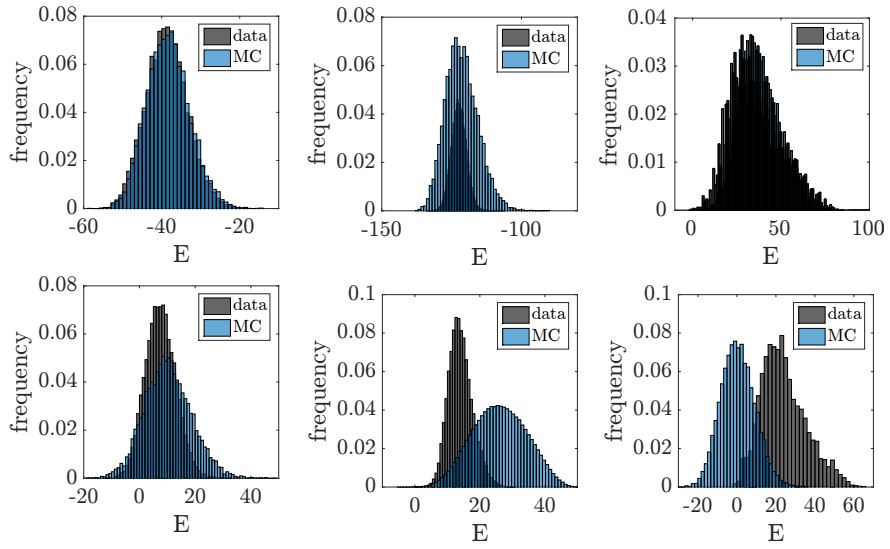


Figure 2.7 – Energies distribution of sequences from the input data (black) and from the sampled model (blue) for ER05 (left panels), S_B (middle panels), and PF00014 (right panels). MC generated configurations with ACE (top panels) have similar energies than the input sequences. This is not the case for plmDCA (bottom panels).

A last aspect of statistical consistency lies in comparing the distribution of energies for configurations sampled from the inferred model to the distribution obtained from the original data. It indicates whether the real data could have been generated from the inferred model. This ability to estimate the energy of a configuration is actu-

ally paramount when comparing the likelihood of a sequence in two different models, as we have done in Part II of this dissertation.

The top panels of Fig. 2.7 show that the distributions of energies of sequences from the input data and sequences generated by the ACE inferred models closely overlap. A small discrepancy is observed in PF00014 (right panel), because of the reweighting procedure – the histogram is normalized by the sequence weights. The energy distribution for the lattice protein model is broader than for the data, although the peak is fit correctly.

Contrary to model inferred with ACE, the distribution of energies is less well reproduced with plmDCA, as shown on the bottom panels of Fig. 2.7. This is again consistent with Fig. 2.2 & 2.3, which show that the true Potts parameters are poorly inferred due to insufficient sampling affecting the generative properties of the model.

2.6 OUTLOOK

In this chapter, we applied the Potts version of ACE to various datasets. The complexity of the inferred Potts models is adapted to the level of the sampling in the data by both regrouping less frequently observed Potts states into a unique state (according to a threshold on entropy or frequency), and then by a sparse inference procedure that omits interactions that are unnecessary for reproducing the statistics of the data to within the error bounds due to finite sampling. We then compared ACE with standard maximum-entropy inference methods based on pseudolikelihood and mean-field approximations. The latter are particularly fast and adapted to find structural contacts and use large regularizations. Inference with ACE is generally slower than mean-field and pseudolikelihood approaches.

However, we showed that ACE was very efficient in recovering the underlying model parameters when known, and in constructing good generative models of the data when using a Bayesian value of the regularization strength ($\gamma \sim 1/B$), outperforming plmDCA. The distribution of energies is also better described by the models inferred with ACE than with plmDCA, a paramount property for comparing sequence scorings (*cf.* Part II). In analogy with standard DCA methods, using ACE with strong regularizations improves the contact prediction while the generative properties of the inferred model are degraded.

Reducing the number of explicitly Potts states according to their sampling allows for a faster and more precise inference of the model parameters while reducing overfitting, and can also be applied to other inference methods. This is precisely the topic of the next chapter.

3

ROLE OF THE COMPRESSED REPRESENTATION OF THE DATA

This work has been initiated in collaboration with E. De Leonardis from LPS-ENS & LCQB

In the very last chapter of this dissertation, we will extensively explore the effect of the compressed representation of the data introduced Chapter 1 – *i.e.* the adaptation of the complexity of the inferred Potts model to the level of sampling – on the quality of the inference with the ACE and plmDCA methods. To carry out this study, we consider artificial data from Erdős-Rényi random graphs. We will start by analyzing the impact of both compression and gauge on the conditioning of the Fisher information matrix (*cf.* Chapter 1 Section 1.1). The variations of the approximate Kullback-Leilbler (KL) divergence between the true and the inferred distributions with the compression will then be discussed, leading to indications regarding the optimal choice for the compression parameter. Finally, the influence of the compression parameter on the recovery of the true underlying model parameters and the statistics of the input data will be investigated.

3.1 METHOD AND DATASETS

We consider artificial data from two types of Potts model where the network of interactions is described by an Erdős-Rényi random graph: a Potts model with $q = 5$ states and $N = 15$ variables for the study of the conditioning of the Fisher information matrix, and a Potts model with $q = 10$ states and $N = 50$ variables for the rest of this chapter. In both cases, each edge in the network is included with probability 0.05 with an maximum connectivity of 7 and the Potts parameters on interacting sites are selected from Gaussian distributions of mean $\mu = 0$ and standard deviations $\sigma_J^2 = 1$ and $\sigma_h^2 = 5$.

In each case ($q = 5$, $N = 15$ and $q = 10$, $N = 50$), 10 realizations (new network of interactions and new set of fields and couplings) are generated. For each realization, $B = 10^2$, $B = 10^3$, $B = 10^4$, and $B = 10^5$ configurations are generated through Monte-Carlo sampling. The observed Potts states are described with the compression scheme on frequency introduced in Chapter 1 Section 1.2, with parameters $f_0 = [10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$, in the limit $f_0 \geqslant 1/B$.

For the purposes of this study, a version of plmDCA including data compression (*i.e.* the number of Potts states depends on the variable) has been implemented and will be referred to as cplmDCA. Two regularization strengths have been used: the standard high regularization $\gamma = 10^{-2}$ and a sample size dependent regularization $\gamma = 1/B$. After

being inferred, the couplings and fields are shifted to the consensus gauge to compare with the true underlying model parameters.

The ACE inference has been performed by gauging the consensus (and also minimum consensus and compressed states for the $q = 5$-state Potts models), with the standard regularization $\gamma = 1/B$. Whichever the chosen gauge, the couplings and fields are then shifted to the consensus gauge to compare with the true underlying model parameters.

Taking into account the 10 realizations of the Erdős-Rényi model, the 4 sample sizes, the 2 to 5 (depending on the sampling) values of the compression parameter in frequency, the ACE and cplmDCA algorithms have been used to infer the Potts parameters of 140 different models.

3.2 CONDITIONING OF THE FISHER INFORMATION MATRIX AND GAUGE CHOICE

The Fisher information matrix χ can be used to estimate the statistical fluctuations of the inferred parameters due to finite sampling, as explained in Chapter 1 Section 1.1. Here, χ is analytically computed on random graphs of $N = 15$ variables and $q = 5$ Potts states, the inversion of χ being computationally infeasible for larger values of these parameters. Moreover, to ensure that χ is positive definite, a sampling-dependent regularization term of $1/B$ is added to its diagonal (*cf.* Eq. (1.6)).

We introduce the condition number κ of the Fisher information matrix:

$$\kappa(\chi) = \|\chi\| \cdot \|\chi^{-1}\| = \frac{\sigma_{max}(\chi)}{\sigma_{min}(\chi)} , \tag{3.1}$$

where σ_{max} and σ_{min} are the largest and smallest singular value of χ respectively. It measures the stability of the inverse Fisher information matrix, or in other words how sensitive it is to small changes in the statistics of the input data. It is particularly of interest here, as the statistical errors on the inferred parameters are related to the inverse of χ. κ ranges from $\kappa = 1$ (identity matrix, "perfectly" conditioned) to $\kappa = \infty$ (singular matrices, not invertible).

χ is computed for various compression parameters f_0 in the three following gauges:

cf. Part I Section 2.3.2 for the gauge definition

— consensus, where the gauged symbol c_i at site i is the most frequent state;
— minimum consensus, where the gauged symbol c_i is the least frequent state within the k explicitly modeled states (*cf.* Section 1.2);
— regrouped, where the gauged symbol is the regrouped state.

Interestingly, the condition number barely depends on the compression parameter as shown on Fig. 3.1, meaning that compressed

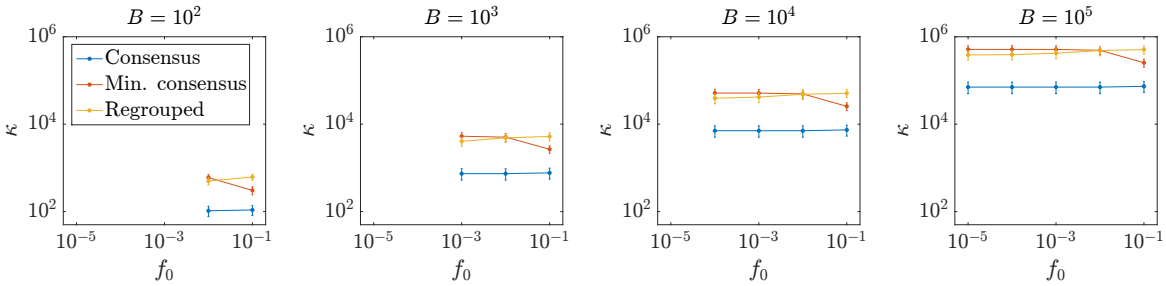


Figure 3.1 – Mean condition numbers $\kappa(\chi)$ of over 10 realizations of the model, as a function of the compression parameter $f_0 \leqslant 1/B$ for gauges consensus (blue), minimum consensus (red), and regrouped (yellow) and sample sizes (a) $B = 10^2$, (b) $B = 10^3$, (d) $B = 10^4$, and (d) $B = 10^5$. Error-bars are standard deviations over the 10 realizations.

representation of the data does not affect the inversion of the Fisher information matrix and hence the finite sampling errors. We also note that the condition number strongly depends on the sampling *via* the diagonal regularization term $1/B$. Indeed, any zero mode will be replaced by $\sigma_{min} = 1/B$ and give rise to a B multiplicative factor to the condition number.

Moreover, it seems that the condition number is smaller for the consensus gauge (blue), meaning that σ_{max} is smaller. The consensus gauge will therefore be the chosen gauge for ACE inference in the following. Although we must stress here that this is only an indication about the gauge choice, it has also been empirically observed that the convergence of the algorithm was often faster in this gauge.

3.3 MINIMIZING THE KULLBACK-LEIBLER DIVERGENCE

3.3.1 *Theoretical framework*

The analytical computation is done in the Ising case for the simplicity of the notations, the generalization to the Potts case being straightforward. We denote $\mathbf{J}^B = \{J^B_{ij}, h^B_i\}$ the parameters inferred on a sample of size B, and $\mathbf{J}^{true} = \{J^{true}_{ij}, h^{true}_i\}$ the true underlying model parameters. The inferred cross-entropy at sampling B writes

$$S_B = -\sum_{\boldsymbol{\sigma}} P_{\mathbf{J}^B}(\boldsymbol{\sigma}) \log P_{\mathbf{J}^B}(\boldsymbol{\sigma}) \,, \tag{3.2}$$

where the sum is over all possible configurations $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_N\}$. The inferred probability distribution at finite sampling B is

$$P_{\mathbf{J}^B}(\boldsymbol{\sigma}) = \frac{1}{\mathcal{Z}_B} \exp\left(\sum_{i=1}^{N} h^B_i \sigma_i + \sum_{\substack{k,l=1 \\ k<l}}^{N} J^B_{kl} \sigma_k \sigma_l \right) \,. \tag{3.3}$$

The Kullback-Leibler (KL) divergence between the true and the inferred distributions writes

$$\begin{aligned} D(P_{J^{true}}||P_{J^B}) &= \sum_{\sigma} P_{J^{true}}(\sigma) \log \frac{P_{J^{true}}(\sigma)}{P_{J^B}(\sigma)} \\ &= -S_{true} - \sum_{\sigma} P_{J^{true}}(\sigma) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l - \log \mathcal{Z}^B \right\} \\ &= -S_{true} + \log \mathcal{Z}^B - \sum_{\sigma} P_{J^{true}}(\sigma) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l \right\} . \end{aligned}$$

However, Eqs. (3.2) & (3.3) give

$$\log \mathcal{Z}^B = S_B + \sum_{\sigma} P_{J^B}(\sigma) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l \right\} .$$

The KL divergence between the true and the inferred distributions then writes

$$\begin{aligned} D(P_{J^{true}}||P_{J^B}) =& (S_B - S_{true}) - \sum_{\sigma} P_{J^{true}}(\sigma) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l \right\} \\ &+ \sum_{\sigma} P_{J^B}(\sigma) \left\{ \sum_i h_i^B \sigma_i + \sum_{k<l} J_{kl}^B \sigma_k \sigma_l \right\} . \end{aligned}$$

Moreover, a reasonable approximation is

$$\begin{aligned} S_{true} &= -\sum_{\sigma} P_{J^{true}}(\sigma) \log P_{J^{true}}(\sigma) \\ &\approx S_{B\to\infty} = -\sum_{\sigma} P_{J^{B\to\infty}}(\sigma) \log P_{J^{B\to\infty}}(\sigma) , \end{aligned} \tag{3.4}$$

because the true underlying parameters are recovered by the inference method in the perfect sampling case: $P_{J^{B\to\infty}}(\sigma) \to P_{J^{true}}(\sigma)$. Therefore,

$$\begin{aligned} D(P_{J^{true}}||P_{J^B}) =& (S_B - S_\infty) + \sum_i h_i^B \left(\langle \sigma_i \rangle^B - \langle \sigma_i \rangle^\infty \right) \\ &+ \sum_{k<l} J_{kl}^B \left(\langle \sigma_k \sigma_l \rangle^B - \langle \sigma_k \sigma_l \rangle^\infty \right) , \end{aligned} \tag{3.5}$$

where $\langle \cdot \rangle^B = \sum_\sigma \cdot P_{J^B}(\sigma)$, and $\langle \cdot \rangle^\infty = \sum_\sigma \cdot P_{J^{B\to\infty}}(\sigma) \approx \sum_\sigma \cdot P_{J^{true}}(\sigma)$.

It naturally generalizes to the q-state Potts case:

$$\begin{aligned} D(P_{J^{true}}||P_{J^B}) =& (S_B - S_\infty) + \sum_{i=1}^{N} \sum_{a=1}^{q} h_i^B(a) \left(\langle \sigma_{ia} \rangle^B - \langle \sigma_{ia} \rangle^\infty \right) \\ &+ \sum_{\substack{k,l=1 \\ k<l}}^{N} \sum_{c,d=1}^{q} J_{kl}^B(c,d) \left(\langle \sigma_{kc} \sigma_{ld} \rangle^B - \langle \sigma_{kc} \sigma_{ld} \rangle^\infty \right) . \end{aligned} \tag{3.6}$$

The dependence of Eq. (3.6) in the compression parameter is not straightforward: the Potts parameters $\mathbf{J}^B = \{J^B_{ij}, h^B_i\}$ are inferred on explicitly modeled states only ($f_i(a) > f_0$) and then completed to the full q-state form with Eqs. (1.15) & (1.16). The number of Potts states depending on f_0, each new value of the compression parameters gives rise to new couplings and fields, as well as a new value of the final entropy S_B, computed by the ACE procedure.

To simulate a perfect sampling and compute the one- and two-point statistics $\langle \sigma_{ia} \rangle^\infty$ and $\langle \sigma_{kc} \sigma_{ld} \rangle^\infty$ in Eq. (3.6), $B = 10^9$ configurations are generated through MC sampling of the true underlying model parameters. The value of entropy S_∞ is then inferred from these correlations with ACE.

3.3.2 *Results*

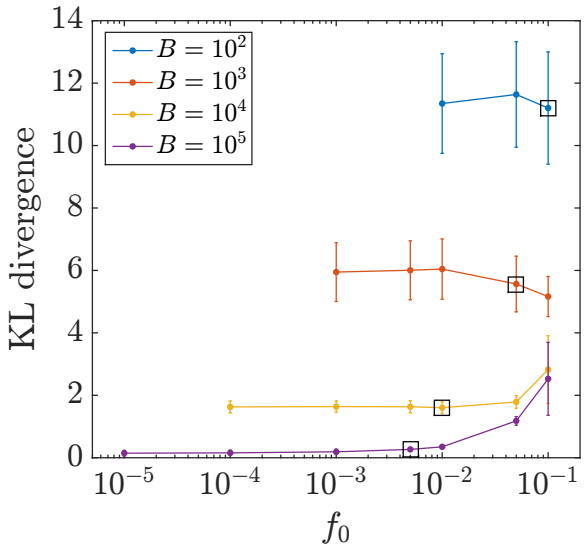


Figure 3.2 – Mean Kullback-Leibler divergence over 10 realizations between true and inferred probability distributions, as a function of the compression parameter f_0 and for sample sizes $B = 10^2$ (blue), $B = 10^3$ (red), $B = 10^4$ (yellow), and $B = 10^5$ (purple). Black squares indicate $f_0^\star = 1/\sqrt{B}$. Error-bars are standard deviations over the 10 realizations.

Fig. 3.2 displays the mean KL divergence between the true and the inferred distributions for various sample sizes and compression parameters. As expected, the KL divergence decreases as the sampling increases, becoming very close to zero for large sample sizes. The variations of the standard deviations on the 10 realizations (error bars on the figure) are large for small sample sizes, but very small as the sampling increases.

Very interestingly, the KL divergence depends fairly little on the reduction parameter. However, its behavior is qualitatively different depending on the sampling. For small sample sizes ($B = 10^2 - 10^3$), it slightly decreases for large compression parameters. It means that reducing the number of explicitly modeled Potts states – *i.e.* increasing f_0 and hence reducing the number of parameters to infer – limits the overfitting and improves the quality of the inference.

related to the bias-variance tradeoff in supervised learning

On the contrary, for larger sample sizes ($B = 10^4 - 10^5$) it slightly increases for large compression parameters. Indeed one can expect that if the sampling is already large enough, reducing too much the number of explicitly modeled Potts states will end up in an important loss of information affecting the quality of the inference.

Moreover, a natural choice for the compression parameter is $f_0^\star = 1/\sqrt{B}$ such that pair correlations between independent states with frequencies of $f_0^\star$ are at the threshold of detection $1/B$ (*i.e.* observed at least once within the B samples). For the considered sample sizes $B = \{10^2, 10^3, 10^4, 10^5\}$, $f_0^\star$ takes on the values $\{0.1, 0.03, 0.01, 0.003\}$ respectively, pointed out by black squares on Fig. 3.2. This choice seems indeed relevant as it corresponds to almost optimal values of the computed KL divergence.

3.4 COMPRESSION AND RECOVERY OF THE ER05 PARAMETERS

3.4.1 *Inference with the adaptive cluster expansion*

Consistently with what has been presented in Chapter 2, the underlying true parameters of the model are accurately recovered by ACE. Fig. 3.3 displays the Pearson correlation coefficients, the absolute errors, and the relative errors (introduced at Eqs. (1.12) & (1.11)) as a function of the compression parameter f_0 and for different sample sizes. Explicitly modeled states only ($f_i(a) > f_0$) are displayed on the left panels (a) & (d), whereas the complete q-state model (*cf.* Eqs. (1.15) & (1.16)) are shown on the right panels (b) & (d).

As expected, the correlations and errors get globally better as the sample size increases. As the compression parameter f_0 increases, the number of explicitly modeled Potts states is reduced and only the sufficiently well sampled states are explicitly treated. The correlation between true and inferred couplings on explicitly modeled states (panel (a)) therefore increases, reaching values close to 1 for large sample sizes. Besides, the absolute error ΔJ (panel (c)) strongly decreases.

On the other hand, the correlations and absolute errors between the true and the complete q-state inferred model (panels (b) & (d)) are globally fairly independent from the reduction parameter. For the largest sample size $B = 10^5$ and large compression parameters, the correlation (resp. absolute error) slightly decreases (resp. increases).

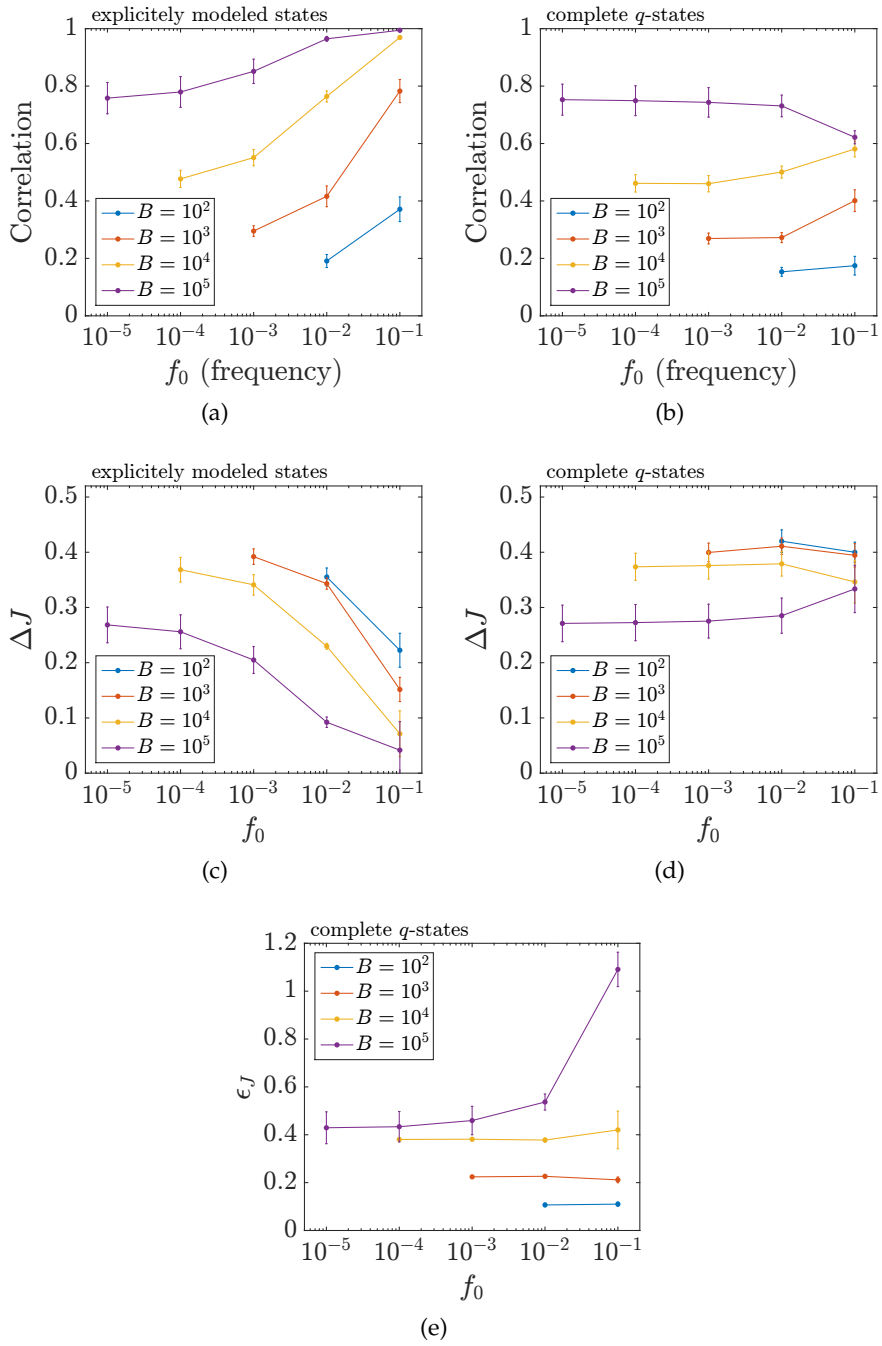


Figure 3.3 – Mean Pearson correlation (top panels), absolute error ΔJ (middle panels), and relative error ϵ_J (bottom panel) over 10 realizations between the true and ACE inferred couplings for explicitely modeled states only (panels (a) & (c)) and complete q-state model (panels (b), (d) & (e)), as a function of the compression parameter f_0 and for sample sizes $B = 10^2$ (blue), $B = 10^3$ (red), $B = 10^4$ (yellow), and $B = 10^5$ (purple). Error-bars are standard deviations over the 10 realizations.

Reducing too much the number of Potts states with a large sampling indeed results in a loss of information about the system that affects the quality of the inference. On the contrary, for smaller sample sizes $B = 10^2 - 10^4$, the correlations and absolute errors get slightly better for large compression parameters, underlying that reducing the number of Potts states for small sampling reduces the overfitting. This confirms what has been previously stated about the KL divergence between the true and the inferred distributions (Section 3.3).

Finally, the same behavior is observed for the relative error ϵ_J (panel (e)), which gets a lot bigger for the largest sampling as the compression parameter increases. Notice that the expected value for the relative error is around 1. Smaller errors are measured because of the 2-site approximation (Eq. (1.10)) used here, which overestimates the finite sampling errors.

3.4.2 *Inference with the compressed pseudolikelihood maximization*

Very similar results are obtained with the compressed version of the pseudolikelihood maximization (*cf.* Fig. 3.4), implemented for the purposes of this study. It therefore validates the methods of compression limiting the overfitting within plmDCA, the most used approximation in the context of protein sequence data. The gain in computational time is less relevant regarding pseudolikelihood, because it also linearly depends on the number of sequences in the input alignment, contrary to ACE which only takes one- and two-point (compressed) frequencies as inputs.

see Part I Section 2.2.3 for more details

The role of regularization, however, is less clear (*cf.* Fig. 3.4). It naturally depends on the sampling, as a weak regularization gives smaller errors for a large sample size ($B = 10^5$). On the contrary, a strong regularization seems to be more efficient in small sampling cases. Moreover, this behavior is qualitatively different for large compression parameters ($B = 10^4 - 10^4$ panel (a)). To better understand the role of regularization, we consider a specific realization of the ER05 model at $B = 10^4$, $f_0 = 0.01$. Fig. 3.5a displays the inferred couplings compared to the true underlying model couplings at strong ($\gamma = 10^{-2}$) and weak ($\gamma = 1/B = 10^{-4}$) regularizations.

Strong regularization clearly overdamps large couplings, and as seen in Chapter 2, the recovery of the true couplings is less accurate. However, very surprisingly, the absolute error decreases:

$$\Delta J(\gamma = 10^{-2}) = 0.2246 < \Delta J(\gamma = 10^{-4}) = 0.2701 \, .$$

This is entirely due to the zero true couplings (the maximum connectivity being fixed to 7, some pairs of sites do not interact and the corresponding true couplings are zero), which are very badly inferred at low regularization (vertical bar at $x = 0$ on Fig. 3.5a). Indeed, these zero true couplings induce low correlations in the MC-generated align-

ment and therefore poor inference, giving rise to abnormally high couplings. High regularization suppresses this effect.

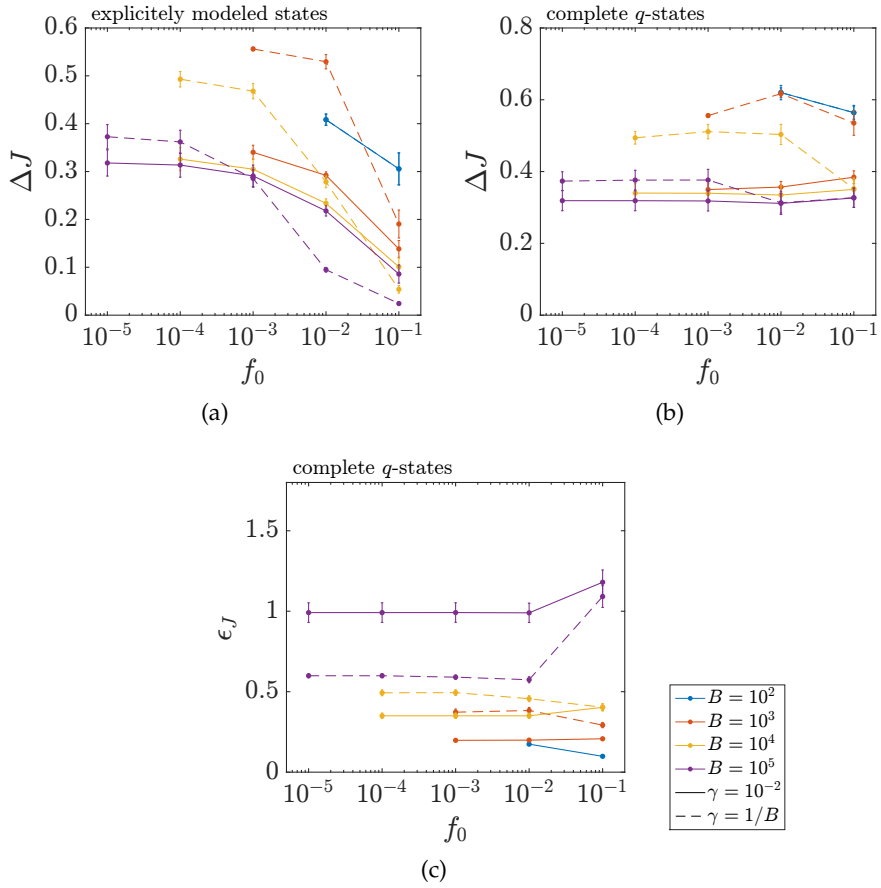


Figure 3.4 – Mean absolute error ΔJ (top panels) and relative error ϵ_J (bottom panel) over 10 realizations between the true and cplmDCA inferred couplings for explicitely modeled states only (panels (a)) and complete q-states model (panels (b) & (c)), as a function of the compression parameter f_0, for standard regularization $\gamma = 10^{-2}$ (stars) and sampling-dependent regularization $\gamma = 1/B$ (diamonds), and for sample sizes $B = 10^2$ (blue), $B = 10^3$ (red), $B = 10^4$ (yellow), and $B = 10^5$ (purple). Errorbars are standard deviations over the 10 realizations.

Actually, if zero real couplings are discarded, the absolute errors becomes:

$$\widetilde{\Delta J}(\gamma = 10^{-4}) = 0.0956\ .$$

Interestingly, increasing the compression parameter and thus reducing the number of explicitly modeled Potts states discards these poorly sampled sites and reduce the abnormally high inferred couplings, as displayed on Fig. 3.5b, on which the above-mentioned vertical bar shrinks.

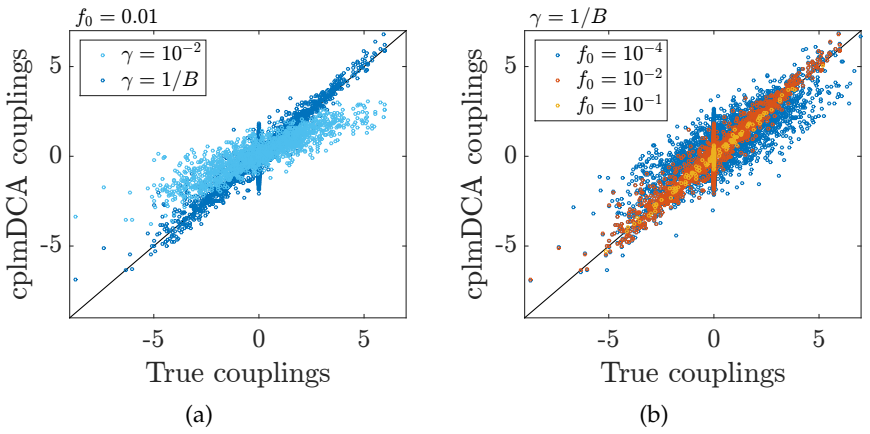


Figure 3.5 – Comparison between cplmDCA-inferred couplings and true underlying model couplings for a specific realization of the ER05 model at $B = 10^4$. Panel (a): comparison between strong (light blue) and weak (dark blue) regularizations at fixed compression $f_0 = 0.01$. Panel (b): comparison between compression parameters $f_0 = 10^{-4}$ (blue), $f_0 = 10^{-2}$ (red), and $f_0 = 10^{-1}$ (yellow) at fixed weak regularization $\gamma = 1/B$.

3.5 COMPRESSION AND REPRODUCIBILITY OF THE STATISTICS

The very last section of this dissertation will address the influence of the compressed representation of the data on the generative properties of the inferred model. For each of the 10 Erdős-Rényi random graph realizations, sample sizes, and values of the compression parameter in frequency f_0 (a total of 140 different models), we compute the one-, two-, and three-point correlation functions through Monte-Carlo sampling of a given number of configurations from the inferred model with ACE and cplmDCA, as in Chapter 2. The Pearson correlation between the MC statistics and to the true ones is then computed for each of the 140 models. Consistently with Chapter 2, ACE gives better results than cplmDCA, as shown on Fig. 3.6, on explicitly modeled states only. This is especially the case for magnetizations (left panels) and three-point connected correlations (right panels). As expected, the Pearson correlation between the true and MC statistics increases with the sample size.

The role of compression is however less clear than for the KL divergence between the true and the inferred distributions or the recovery of the true underlying Potts parameters. The generative properties (on explicitly modeled states) seem rather unaffected by the compressed representation of the data, as the Pearson correlation is globally flat. It has a slight tendency to increase with the compression parameter, as poorly inferred states are removed, consistently with the results of the previous sections. However, for largest values of the compression parameter $f_0 = 0.1$, the correlation between

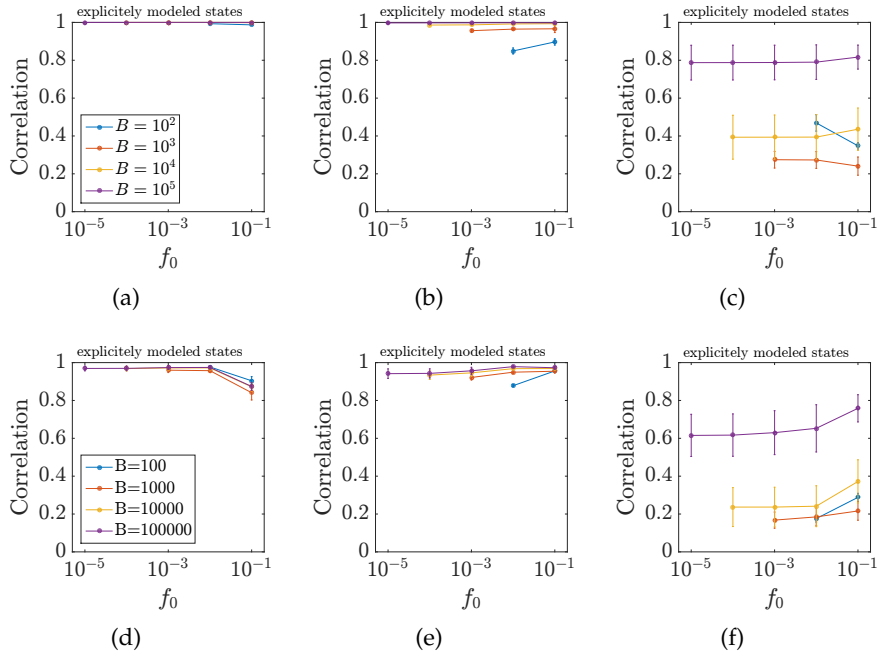


Figure 3.6 – Mean Pearson correlation coefficient over the 10 realizations between the input statistics and the MC statistics from ACE (top panels) and plmDCA, $f_i(a)$ (left panels), $C_{ij}(a,b)$ (middle panels), and $C_{ijk}(a,b,c)$ (right panels), as a function of the compression parameter f_0 and sample sizes $B = 10^2$ (blue), $B = 10^3$ (red), $B = 10^4$ (yellow), and $B = 10^5$ (purple). Error-bars are standard deviations over the 10 realizations.

true and MC statistics sometimes decreases. It is the case for magnetizations inferred with cplmDCA for all sample sizes (panel (d)) and three-point connected correlations with ACE for small sample sizes (panel (c)). In Chapter 2, we noted that magnetizations are particularly badly reproduced by plmDCA (*cf.* Fig. 2.6), regardless of the sample size. Three-point connected correlations are in any case hard to infer for small sample sizes.

More than the compression parameter (when at least the non observed states are removed), what seems to matter most is the strength of the regularization. Fig. 3.7 gives an overview of the generative properties for a given realization of the Erdős-Rényi model with $B = 10^4$ configurations, depending on the inference method: ACE (top panels) with $f_0 = 0.1$, standard plmDCA without any compression and standard regularization $\gamma = 10^{-2}$, compressed version cplmDCA with $f_0 = 0.1$ and standard regularization $\gamma = 10^{-2}$, cplmDCA with $f_0 = 0.1$ weak regularization $\gamma = 1/B = 10^{-4}$, and cplmDCA with $f_0 = 0.1$ standard regularization and refined with BML (bottom panels).

As usual, ACE (top row) outperforms any version of plmDCA, regardless of the presence of compression or the regularization strength. The standard version of plmDCA with no compression scheme (second

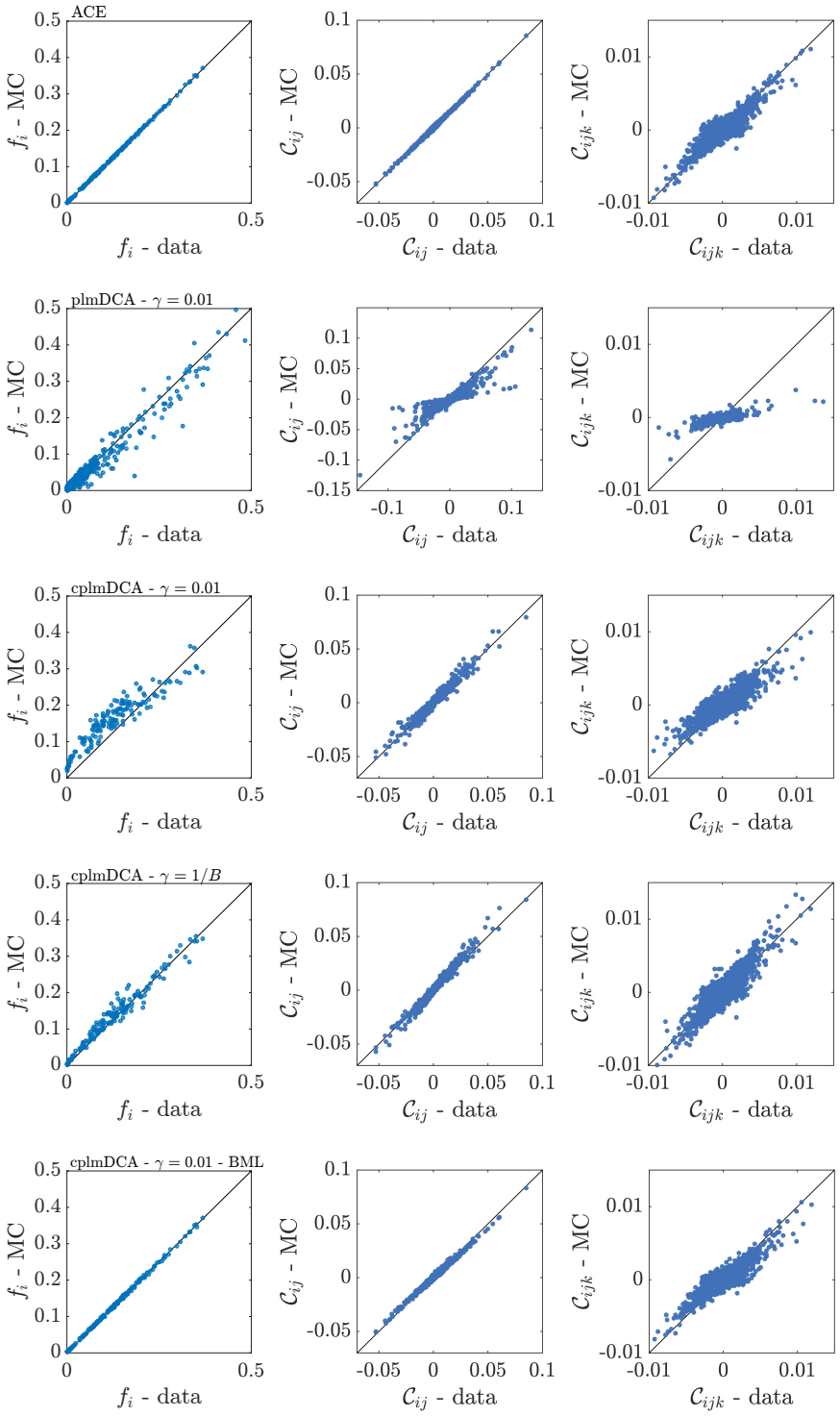


Figure 3.7 – Fit for ACE with $f_0 = 0.1$ (top panels) and various versions of the pseudolikelihood maximization, from top to bottom: plmDCA standard without compression and strong regularization, cplmDCA with compression $f = 0.1$ and standard strong regularization, cplmDCA with compression $f = 0.1$ at weak regularization, plmDCA with compression $f = 0.1$ at standard strong regularization refined with BML.

row) also includes poorly sampled states and the two- and three-point connected correlations cannot be accurately reproduced. Adding a compression scheme regroups these states and improves the generative properties on explicitly modeled states (third to fifth rows). With cplmDCA at fixed compression $f_0 = 0.1$, decreasing the regularization from standard $\gamma = 10^{-2}$ (third row) to weak $\gamma = 1/B$ (fourth row) gives much better fits.

Finally, as mentioned in Chapter 2 Section 2.4, the output set of fields and couplings of cplmDCA can be used as starting values for a BML routine. In this case, the procedure can lead to rapid convergence of the model even when the starting error is large due to strong regularization. Besides, running plmDCA or cplmDCA with weak regularization greatly increases the computational time. The latter can therefore be reduced by using a strong regularization to infer the Potts parameters and refining them with BML. The generative properties of the resulting model (last row) are excellent, comparable to ACE.

3.6 OUTLOOK

In this last chapter, we have explored the role played by the compressed representation of the data on the quality of the inference, by studying the conditioning of the Fisher information matrix, the Kullback-Leibler divergence, the recovery of the true underlying model parameters, and the generative properties of the inferred models. 140 models have been generated corresponding to 10 realizations of Erdős-Rényi random graphs, 4 sample sizes ($B = 10^2$ to 10^5), and various values of the compression parameters f_0. Potts parameters have been inferred for each of these models with ACE and a compressed version of pseudolikelihood maximization called cplmDCA implemented for the purposes of this study.

If the KL divergence for the completed models (the regrouped states are expanded according to Eqs. (1.15) & (1.16)) depends fairly little on the compression parameter, interesting variations are observed at strong compression depending on the sampling. For small sample sizes, the KL divergence between the true and the inferred distributions can be reduced by decreasing the number of explicitly modeled states (increasing f_0), therefore limiting the overfitting. On the contrary, for large sample sizes, all Potts states are expected to be well represented in the samples and the loss of information induced by the compression affects the quality of the inference.

This behavior is confirmed by the analysis of statistical errors on the inferred Potts parameters with ACE and cplmDCA. Naturally, the precision on the inferred parameters increases with the compression parameter, as insufficiently sampled states are removed. Moreover, regularization plays an important role in both the accuracy of the in-

ferred couplings and the generative properties of the model. Decreasing the standard plmDCA regularization from 10^{-2} to $1/B$ improves the generative properties of the model but also increases the required computational time. On the other hand, Boltzmann machine learning routines have been shown to give good generative models but are typically slow to converge. A compromise can be found by using a strong regularization and then refining the inferred Potts parameters with BML.

A possible application for the compressed representation would be protein domain families with few sequences. Some Pfam protein domains indeed contain a limited number of sequences, and although Pfam is frequently updated and grows continuously, the tendency is rather in adding more sequences to already large families than completing small families. DCA related approaches are of course struggling in poor sampling cases. However, we have seen in this chapter that the compression of the Potts states is particularly efficient in these cases, and it would be very interesting to see whether the compression could help improving the inference with only few sequences.

To provide some insight about the feasibility of this application, we display on Fig. 3.8 the distribution of single-site entropies given by $S_i = -\sum_{a=1}^{q} f_i(a) \log f_i(a)$ for two protein domain families: the response regulator receiver domain (Pfam id: PF00072) and the P53 DNA-binding domain (Pfam id: PF00870). The former is one of the largest family in the Pfam database with $B_{eff} \gtrsim 150000$ sequences and $N = 112$ residues and the latter is a small family with $B_{eff} = 117$ sequences and $N = 196$ residues. The single-site entropies seem to vary considerably from one family to the other, and are typically higher in the large family. It indicates that the effective number of Potts states (q_i, $i = 1, ..., N$) may also greatly vary from one family to the other and also, more importantly, from one site to the other.

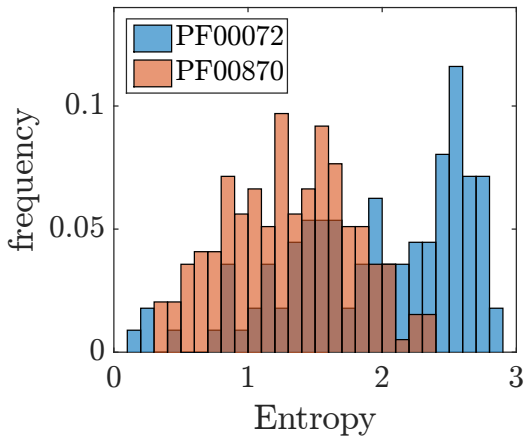


Figure 3.8 – Single-site entropy distribution for PF00072 (blue) and PF00870 (red), with respectively $B_{eff} \gtrsim 150000$ and $B_{eff} = 117$ sequences.

Part V

CONCLUDING REMARKS

1

SUMMARY OF THE RESULTS

Proposed in 2009, direct-coupling analysis (DCA) is a global statistical inference method taking the form of a q-state Potts model, which describes the variability of sequences across homologous protein families. Using pairwise correlations in amino-acid occurrence from large multiple sequence alignments – readily available thanks to rapidly increasing sequence databases – DCA is able to make structural predictions about proteins (*i.e.* contacts on the 3D fold) from purely statistical considerations based on sequence information alone, and is today widely used in the field (Part I). Encouraged by the success of this approach, we explored other challenging fields in which it may be applied, such as protein folding or homology detection (Part II). Contrary to residue-residue contact prediction, which remains an intrinsically topological information about the network of interactions, these fields require global energetic considerations and therefore a more quantitative and detailed model. We indeed realized that a better understanding of DCA models, of their couplings parameters (which were not fully exploited), and of the role of other aspects of the problem (gauge, regularization, sampling, etc.) are paramount to succeed in going beyond protein structure prediction (Parts III & IV).

In Part II, we focused on two possible applications for DCA approaches: sequence folding prediction and homology detection. We proved in Chapter 1 that DCA is a good predictor of whether a given artificial sequence will fold in a native structure or not, outperforming non-pairwise models such as HMMer. This work is based on a recent publication, where artificial proteins sequences for the WW domain were designed based on the original multiple sequence alignment, and their folding properties assessed experimentally. This is a very encouraging result toward the applicability of DCA methods to protein design.

Another possible field of application is remote homology detection, a computationally hard problem where methods treating residues independently, such as HMMer, are often not satisfactory. Pairwise models considering covariation patterns – currently not taken into account – could perform better. While alignment gaps are quite rare in the artificial sequences for the WW domain, they are much more frequent in remote homologs. We showed that they give rise to strong couplings in DCA approaches, which treat them as an extra amino acid, despite strong evidence that they are intrinsically different. The sequence scoring with DCA energies can therefore become completely

irrelevant, because mostly dominated by gap signals. We introduced, in Chapter 2, a correction based on a null model of the gap distribution, so that only the signal stemming from amino acids is captured, getting rid of the spurious gap-induced interactions.

We applied this model to a dozen of protein domain families divided into sub-families according to their natural clustering into domains of life (eukaryotes, bacteria, archaea). Surprisingly, DCA has a stronger tendency than HMMer to discriminate between sequences from the same family, consistently with the phylogenetic distribution. It is sometimes able to detect errors in the labeling of sequences, or at least to give interesting insights about unclassified sequences, pointing out good candidates for further studies. Having a more detailed picture of the different domains of life could also be very useful for phylogeny. Although they do not quite answer the question about remote homologies asked in the first place, these preliminary results are promising for the application of DCA approaches beyond structural prediction in proteins.

In a more principled approach presented in Chapter 3, we proposed a theoretical framework in which alignment gaps are modeled as missing information, and thus gap-rich sequences as partial observations. This specific problem had never been addressed in the literature to the best of our knowledge; related approaches indeed consider hidden nodes (and not random missing entries in the dataset). Our iterative procedure has been tested on random distributions of gaps in Erdős-Rényi graphs. We proved that the true underlying model parameters can be accurately recovered, depending on the sample size and the amount of missing entries in the sequences. The true energies are also well reproduced, within the expected precision due to the uncertainty, and much better than with the standard DCA model.

Then, in order to better understand the specificities of DCA approaches, we studied extensively in Part III its central parameters: the Potts couplings. These $q \times q$ matrices are usually mapped onto scalar parameters which are subsequently ranked, losing a large part of the information they potentially contain. We showed that the couplings contain quantitative and interpretable biological information related to the physico-chemical properties of amino-acid interactions. These interactions are consistent with state-of-the-art knowledge-based amino-acid potentials.

Our results are based on the analysis of 70 protein multiple sequence alignments (MSAs) in Chapter 2, from which we inferred the Potts parameters. The average coupling matrix (over the top-ranked residue pairs for each family) and its spectral modes display interesting features: electrostaticity, hydrophobicity, and stericity. The full biological content of the coupling matrices – Cysteine-Cysteine and hydrophilic interactions – was however unveiled by considering struc-

tural classifications and solvent exposure. We also showed that the distribution of contact distances in the tertiary structure greatly depends on the type of interaction associated to the contact. Despite our several attempts, we could not however use this information to improve the structural prediction in proteins. Notice that only meta methods, such as PconsC2 [106] – using also *e.g.* the vicinity of a potential contact, or secondary-structure predictions – are currently able to better separate signal from noise. Interestingly, such meta methods are mostly improving the contact prediction between secondary structures – *i.e.* filling the predicted contact map – but frequently not adding new structurally informative contacts.

We furthermore considered abstract lattice-protein models in Chapter 3 to better understand the crucial role of sampling and regularization on the inferred Potts parameters. Decreasing the regularization strength allows for a richer signal to emerge in the coupling matrices – consistent with amino-acid interactions and evolutionary pressure – but only if the sample size is sufficiently large. Otherwise, the signal is strongly affected by sampling noise. This is precisely the case for real proteins, where the number of non-redundant sequences (B_{eff}) is still limited. Consistently with a recent publication by our group, we used the detailed structure of the inferred couplings to improve structural predictions for lattice proteins in the sufficient sampling case. Note that this picture somewhat depends on the inference method considered: more precise inference procedures, such as the adaptive cluster expansion, allow to detect a stronger signal.

Finally, in Part IV, we focused on the adaptive cluster expansion (ACE), recently generalized to the Potts case. This method is adapted to the level of noise in the data by inferring a sparse network omitting insufficiently well sampled interactions, while proposing a compressed representation of the data (introduced in Chapter 1). In Chapter 2, we compared ACE to standard DCA approaches on several datasets (Erdős-Rényi graphs, lattice proteins, and real proteins). We showed that ACE outperforms DCA methods based on pseudolikelihood approximations (plmDCA) in recovering the underlying model parameters on artificial data, and in constructing good generative models. More over, ACE is competitive with standard approaches in predicting protein contacts. The distribution of energies is also better described by the models inferred with ACE, a paramount property for comparing sequence scorings.

By reducing the size of the system, the compression of the number of Potts states decreases the computational time of the procedure. We explained in Chapter 3 that it also reduces overfitting in the finite sampling case, improving the quality of the inference. The Kullback-Leibler divergence between the true and the inferred distributions may indeed be lowered by decreasing the number of explicitly mod-

eled states. This behavior is confirmed by the analysis of statistical errors on the inferred Potts parameters with ACE and a compressed version of the pseudolikelihood approximation of direct-coupling analysis (cplmDCA), implemented for the purposes of this study. The role of state compression on the generative properties of the model remains unclear and needs to be further investigated.

2

OUTLOOK AND FUTURE WORK

I review here several aspects of the work I have presented in this dissertation which, in my opinion, call for further investigation.

The first problem that needs to be addressed is related to remote homology detection, where similarities need to be found among evolutionary distant proteins. As explained above – under the naive assumption that single-site conservation patterns may be less preserved than covariation patterns among remote homologs – pairwise models inferred with DCA seems to be a natural method to tackle this problem. I was quite surprised to see that, on the contrary, pairwise models have a stronger tendency than independent-site methods to *discriminate* (rather than finding similarities) between sequences from the same family, consistently with their phylogenetic distribution.

At first, I was quite disappointed that DCA could not straightforwardly be used as a tool for remote homology search, but it does answer a different question related to the detection of dissimilarities among sequences. To my mind, we need to put effort in finding a use case, where this ability could be of interest. Strikingly, members of my group working on neuroscience problems also showed that inferred pairwise model outperform single-site approaches in retrieving a rat's current environment from place cells recordings [97]. The ability of pairwise models to discriminate between neuron sequences was paramount in the success of this study. In a way, we need to find the equivalent in the context of protein sequences.

R. Monasson with S. Rosay, and later L. Posani.

This also raises the question of sequence alignment. All multiple sequence alignments used in this dissertation have been downloaded from the Pfam database and thus built with profile models, treating residues independently and based on single-site frequency patterns. The methods we develop are, however, based on pairwise models and exploit covariation patterns, currently not taken into account in alignment methods. Furthermore, standard DCA approaches are limited to sequences of a fixed length N – corresponding to the size of the alignment from which the Potts parameters have been inferred – contrary to profile-HMM specifically designed (*via* insertion states) to produce sequences longer than the profile length. Developing a similar approach to align sequences, but with pairwise Markov random fields, would address both issues.

I mentioned several times in this dissertation the future availability of artificially designed protein data. R. Ranganathan and his team at

UT Southwestern have developed the techniques introduced for the WW domain (*cf.* Part I) and have reduced the time and cost of the experimental procedure. They have applied similar methods to the chorismate mutase, an enzyme catalyzing the production of amino acids. Our group recently obtained preliminary results on this data, confirming the ability of DCA approaches to predict protein folding. New sequences with specifically low DCA energies have also been submitted for experimental design to the aforementioned collaborators, and the results should be known soon. Protein design really is an exciting domain of research with limitless potentialities, and I look forward to the development of DCA methods in this field.

M. Figliuzzi from our group at LCQB

On a different topic, the theoretical framework we have developed for inverse Potts models with missing information in samples needs to be applied to real data. An interesting application would be metagenomic sequences, which often come in fragments indicating a high level of uncertainty on specific regions of the alignments. This approach is not really relevant for homology search as it replaces the gaps of a given sequence according to the correlation patterns in the observed alignment, thus slightly favoring this sequence energetically.

Besides, it could be really interesting to apply the compressed representation of the data (*cf.* Part IV) to Pfam families with few sequences. DCA methods naturally struggle on poorly sampled data, and we have shown in the last part of this dissertation, that grouping insufficiently observed Potts states together helps reducing overfitting. Such domain families include proteins of major importance, such as complex protein interfaces involved in many cellular pathways [120], HIV p7 nucleocapsid protein essential for viral replication [51] (with a lot of sequence redundancy in a single patient, limiting B_{eff}), or the tumor suppressor protein p53 [111] (with $B_{eff} = 117$).

A last problem I wish I had the time to tackle is related to the detection of intermediate contacts. Protein folding can indeed be a slow process, pausing in many well-define intermediate states [20]. The corresponding contacts are crucial for the folding at an intermediary stage, but not present in the final, native structure. Provided that these conformations are stabilized long enough, they may very well induce coevolution between residues. In the same way that contacts in dimeric assemblies are detected by DCA approaches [36] and lead to false positive (because in contact in the quaternary but not tertiary structure), it would be absolutely fascinating to see if some false predictions could actually be explained by intermediate states in the folding process. Some data may be available in [3, 85].

In my opinion, what is particularly complex in the field of protein sequence data is that we consider real, biased, finitely sampled

data that is of course not generated by a known model. There is no such thing as the "best" method, outperforming any other approach for all use cases. For instance, given the results we obtained in Part II, the mean-field approximation of direct-coupling analysis seems to be better than pseudolikelihood in discriminating between protein sub-families, consistently with their phylogenetic distribution. The pseudolikelihood approach, however, outperforms all approximations in protein structure prediction but its generative properties are surpassed by adaptive cluster expansion.

A theoretically perfect model, which is both consistent – in the sense that it recovers the exact parameter values in the limit of an infinitely large sample drawn from the Potts model – and generative can very well be defeated by simple approximations on real data. Each new field of application requires a benchmark of all existing methods, which is quite unsatisfactory in my humble opinion. But what makes biology a fascinating field for statistical physicists is the limitless number of applications; while we try to understand the biological world with our own tools, we discover new ways to improve these tools and every question raises another.

I think that we also need to keep in mind that the availability and the quality of the data is constantly improving. Therefore, we could rapidly reach a point where databases contain large samples of rather unbiased configurations, and then it will be paramount that our models are statistically consistent and generative.

Part VI

APPENDIX

A

PUBLICATION ABSTRACTS

A.1 JOURNAL OF CHEMICAL PHYSICS [31] (UNDER REVIEW)

Direct coevolutionary couplings reflect biophysical residue interactions in proteins

Alice Coucke,[1,2] Guido Uguzzoni,[2] Francesco Oteri,[2] Simona Cocco†,[3] Remi Monasson†,[1] and Martin Weigt†[2]

[1)] *Laboratoire de Physique Théorique, Ecole Normale Supérieure and CNRS-UMR8549, PSL Research University, 24 Rue Lhomond, 75005 Paris, France*

[2)] *Sorbonne Universités, UPMC, Institut de Biologie Paris-Seine, CNRS, Laboratoire de Biologie Computationnelle et Quantitative UMR 7238, 75006 Paris, France*

[3)] *Laboratoire de Physique Statistique, Ecole Normale Supérieure and CNRS-UMR8550, PSL Research University, Sorbonne Universités UPMC, 24 Rue Lhomond, 75005 Paris, France*

† These authors are joint last authors on this work.

Coevolution of residues in contact imposes strong statistical constraints on the sequence variability between homologous proteins. Direct-Coupling Analysis (DCA), a global statistical inference method, successfully models this variability across homologous protein families to infer structural information about proteins. For each residue pair, DCA infers 21×21 matrices describing the coevolutionary coupling for each pair of amino acids (or gaps). To achieve the residue-residue contact prediction, these matrices are mapped onto simple scalar parameters; the full information they contain gets lost. Here, we perform a detailed spectral analysis of the coupling matrices resulting from 70 protein families, to show that they contain quantitative information about the physico-chemical properties of amino-acid interactions. Results for protein families are corroborated by the analysis of synthetic data from lattice–protein models, which emphasizes the critical effect of sampling quality and regularization on the biochemical features of the statistical coupling matrices.

I. INTRODUCTION

Across evolution, the structure and function of homologous proteins are remarkably conserved. As a consequence, neighboring residues in the three-dimensional structure tend to coevolve, leading to strong constraints on the sequence variability. Direct Coupling Analysis (DCA)[1,2], a global inference method based on the maximum-entropy principle[3,4], successfully exploits pairwise correlations in amino-acid occurrence, which are easily observable in large multiple-sequence alignments, to infer spatial residue-residue contacts within the tertiary protein structure. This approach uses a global statistical model $P(a_1,...,a_L)$ for an amino-acid sequence $(a_1,...,a_L)$ of length L, whose parameters are fields/biases $\{h_i(a)\}$ and statistical couplings $\{J_{ij}(a,b)\}$, where a,b are amino acids or alignment gaps (denoted for simplicity by $\{1,...,21\}$ throughout the paper). These parameters are learnt from site-specific amino-acid frequencies, and from the covariance between amino-acid pairs estimated from multiple-sequence alignments (MSA), which are readily available thanks to rapidly increasing sequence databases[5,6]. Contact prediction is performed by measuring the total coupling strength between two residues. The coupling matrices - inferred at high computational cost - are mapped onto simple scalar parameters, and the full information they potentially contain gets lost.

The aim of our work is to provide a better quantitative understanding of these inferred couplings. Earlier works have shown that the coevolutionary couplings derived by DCA contain an electrostatic signal[7]. In the present study, we go considerably further and show that the coevolutionary couplings also contain quantitative and interpretable biological information related to all the physico-chemical properties of amino-acid interactions, not only electrostaticity, but also hydrophobicity/hydrophilicity, Cysteine-Cysteine bonds, Histidine-Histidine and steric interactions. These interactions are consistent with knowledge-based amino-acid potentials inferred from known protein structures, such as the statistical potential derived by Miyazawa and Jernigan[8].

To carry out our study, we first consider a set of 70 Pfam[6] protein families from which we infer the coupling matrices. After selecting the top ranked residue pairs for each family, we analyze the mean coupling matrix and its spectral modes. Considering structural classifications and solvent exposure helps unveiling the full biological content of the coupling matrices $\{J_{ij}(a,b)\}_{a,b\in\{1,...,21\}}$. Our analysis also shows that the distribution of contact distances in the tertiary structure greatly depends on the type of interaction associated to the contact.

In a second part of the article, to better understand the effect of sampling and regularization on

A.2 BIOINFORMATICS [13]

Bioinformatics Advance Access published July 1, 2016

Bioinformatics, 2016, 1–9
doi: 10.1093/bioinformatics/btw328
Advance Access Publication Date: 21 June 2016
Original Paper

Sequence analysis

ACE: adaptive cluster expansion for maximum entropy graphical model inference

J. P. Barton[1,2,*], E. De Leonardis[3,4], A. Coucke[4,5] and S. Cocco[3,*]

[1]Departments of Chemical Engineering and Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, [2]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard, Cambridge, MA 02139, USA, [3]Laboratoire de Physique Statistique de L'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure & Université P.&M. Curie, Paris, France, [4]Computational and Quantitative Biology, UPMC, UMR 7238, Sorbonne Université, Paris, France and [5]Laboratoire de Physique Théorique de L'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure & Université P.&M. Curie, Paris, France

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on March 25, 2016; revised on May 15, 2016; accepted on May 18, 2016

Abstract

Motivation: Graphical models are often employed to interpret patterns of correlations observed in data through a network of interactions between the variables. Recently, Ising/Potts models, also known as Markov random fields, have been productively applied to diverse problems in biology, including the prediction of structural contacts from protein sequence data and the description of neural activity patterns. However, inference of such models is a challenging computational problem that cannot be solved exactly. Here, we describe the adaptive cluster expansion (ACE) method to quickly and accurately infer Ising or Potts models based on correlation data. ACE avoids overfitting by constructing a sparse network of interactions sufficient to reproduce the observed correlation data within the statistical error expected due to finite sampling. When convergence of the ACE algorithm is slow, we combine it with a Boltzmann Machine Learning algorithm (BML). We illustrate this method on a variety of biological and artificial datasets and compare it to state-of-the-art approximate methods such as Gaussian and pseudo-likelihood inference.

Results: We show that ACE accurately reproduces the true parameters of the underlying model when they are known, and yields accurate statistical descriptions of both biological and artificial data. Models inferred by ACE more accurately describe the statistics of the data, including both the constrained low-order correlations and unconstrained higher-order correlations, compared to those obtained by faster Gaussian and pseudo-likelihood methods. These alternative approaches can recover the structure of the interaction network but typically not the correct strength of interactions, resulting in less accurate generative models.

Availability and implementation: The ACE source code, user manual and tutorials with the example data and filtered correlations described herein are freely available on GitHub at https://github.com/johnbarton/ACE.

Contacts: jpbarton@mit.edu, cocco@lps.ens.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

B

HMMER SCORING PROCEDURE

The several steps of the scoring procedure are detailed below:

1. The first task is to divide a protein domain family downloaded from the Pfam database into sub-families, according to the domains of life (Eukaryota, Bacteria, Archaea). This is done by retrieving the labeling of the aligned sequences in the MSA, given in the tab "species" in Pfam (`pfam.xfam.org/family/PF00011#tabview=tab7`). Some discrepancies may be observed between Pfam and Uniprot, as the latter is updated more regularly. Moreover, some sequences are uncharacterized, or of unknown domain.
2. A profile-HMM is built (*hmmbuild* command of the HMMer software) on the training sub-alignment, which therefore has to be retrieved in full length from the Uniprot database (by searching by sequence name). A problem is that the profile length may differ from the original MSA. A simple way to solve the problem is to "fake" the full length alignment by using the semi-aligned version from Pfam also including insertions.
3. All the sequences from the original MSA are scored with the profile-HMM corresponding to the training sub-family (*hmmsearch* command of the HMMer software). It rarely happens (2% of the sequences on average) that HMMer breaks a sequence into two relevant domains that do not exist in the original MSA. This is always due to the presence of insertions in the middle of the sequence, too penalized by HMMer to be detected as a whole. The log-odds score assigned to such sequences is the sum of the scores of both domains.

C

TRAINING ON THE EUKARYOTIC SUB-FAMILIES

To illustrate the properties of the different models to discriminate between the training and test sub-families, we consider the ROC curve (and the corresponding AUC, area under the ROC curve), assessing the performance of binary classifiers. The two classes are in this case the eukaryotes (training sub-family) considered as true positive and the bacteria (test sub-family) considered as false positive. A perfect classifier would give higher scores to all eukaryotes and therefore would be the constant 1 in the ROC space; a random guess would go along the diagonal. Figure C.1 displays the mean ROC and AUC over the five Pfam families studied in this section (PF00011, PF00013, PF00027, PF00033, PF00664) in three models (HMMer score, mfDCA corrected by the null model, plmDCA corrected by the null model). HMMer is outperformed by mfDCA and plmDCA when corrected by the null model on gaps.

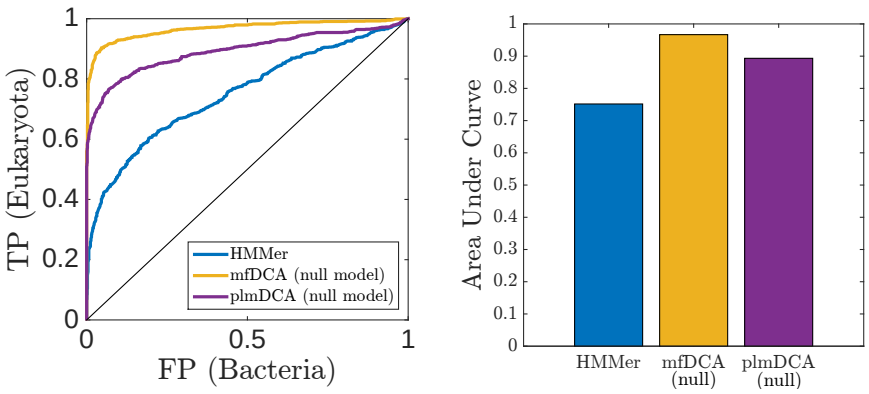


Figure C.1 – Mean ROC (panel (a)) and AUC (panel (b)) curves over the 5 studied Pfam families, illustrating the performance of the different models in discriminating between sequences, depending on whether they belong to the same domain of life than the training sub-family (here eukaryotes) or not. Pairwise models are better classifiers.

D

MAXIMUM-LIKELIHOOD EQUATIONS WITH MISSING DATA

D.1 FIRST MAXIMUM-LIKELIHOOD EQUATION

The derivative of the log-likelihood with respect to the field $h_k(c)$ writes:

$$\begin{aligned}\frac{\partial\widetilde{\mathcal{L}}}{\partial h_k(c)} &= \frac{\partial}{\partial h_k(c)}\sum_{m=1}^{M}\log\sum_{b_1,\dots,b_L}\left(\prod_{\{j|a_j\neq 0\}}\delta_{a_j,b_j}\right)P(b_1,\dots,b_L)\\ &= \sum_{m=1}^{M}\frac{1}{\tilde{P}(a_1^m,\dots,a_L^m)}\sum_{b_1\dots b_L}\left(\prod_{\{j|a_j^m\neq 0\}}\delta_{a_j^m,b_j}\right)\\ &\times\underbrace{\frac{\partial}{\partial h_k(c)}\left(\frac{1}{\mathcal{Z}}\exp\sum_{i=1}^{L}h_i(b_i)+\sum_{i,j=1}^{L}J_{ij}(b_i,b_j)\right)}_{\alpha_k(c)}\ .\end{aligned}$$

But

$$\alpha_k(c) = (\delta_{b_k,c} - P_k(c))P(b_1,\dots,b_L)\ ,$$

with $P_k(c)$ the marginal of the probability distribution P:

$$P_k(c) := \frac{1}{\mathcal{Z}}\sum_{b_1\dots b_L}\delta_{b_k,c}\exp\sum_{i=1}^{L}h_i(b_i)+\sum_{i,j=1}^{L}J_{ij}(b_i,b_j)\ .$$

We then have

$$\begin{aligned}\frac{\partial\widetilde{\mathcal{L}}}{\partial h_k(c)} &= -MP_k(c)+\sum_{m=1}^{M}\frac{1}{\widetilde{P}(a_1^m,\dots,a_L^m)}\\ &\times\underbrace{\sum_{b_1\dots b_L}\left(\prod_{\{j|a_j^m\neq 0\}}\delta_{a_j^m,b_j}\right)\delta_{b_k,c}P(b_1,\dots,b_L)}_{\chi_k^m(c)}\ .\end{aligned}$$

The quantity $\chi_k^m(c)$ depends on the observation of a_k^m:

— If $a_k^m \neq 0$ (observed):

$$\begin{aligned}\chi_k^m(c) &= \sum_{b_1 \dots b_L} \left(\prod_{\substack{\{j|a_j^m \neq 0\} \\ j \neq k}} \delta_{a_j^m, b_j} \delta_{a_k^m, b_k} \right) \delta_{b_k, c} P(b_1, ..., b_L) \\ &= \delta_{a_k^m, c} \sum_{b_1 \dots b_L} \left(\prod_{\{j|a_j^m \neq 0\}} \delta_{a_j^m, b_j} \right) \delta_{b_k, c} P(b_1, ..., b_L) \\ &= \delta_{a_k^m, c} \times \widetilde{P}(a_1^m, ..., a_L^m) \ .\end{aligned}$$

— If $a_k^m = 0$ (not observed):

$$\begin{aligned}\chi_k^m(c) &= \sum_{b_1 \dots b_L} \left(\prod_{\substack{\{j|a_j^m \neq 0\} \\ j \neq k}} \delta_{a_j^m, b_j} \right) \delta_{b_k, c} P(b_1, ..., b_L) \\ &= \widetilde{P}(a_1^m, ..., a_k^m = c, ..., a_L^m) \ .\end{aligned}$$

Therefore

$$\begin{aligned}\frac{\partial \widetilde{\mathcal{L}}}{\partial h_k(c)} &= \sum_{m=1}^{M} \delta_{a_k^m, 0} \frac{\widetilde{P}(a_1^m, ..., a_k^m = c, ..., a_L^m)}{\widetilde{P}(a_1^m, ..., a_L^m)} + \sum_{m=1}^{M} \delta_{a_k^m, c} - M P_k(c) \\ &= \sum_{\{m|a_k^m = 0\}} P(a_k^m = c | \{a_i^m | a_i^m \neq 0\}) + \sum_{m=1}^{M} \delta_{a_k^m, c} - M P_k(c) \ ,\end{aligned}$$

because $\widetilde{P}(a_1^m, ..., a_L^m) = P(\{a_i^m | a_i^m \neq 0\})$.
We finally get the first maximum-likelihood equation:

$$\boxed{M P_k(c) = \sum_{m=1}^{M} \delta_{a_k^m, c} + \sum_{\{m|a_k^m = 0\}} P(a_k^m = c | \{a_i^m | a_i^m \neq 0\})} \qquad \text{(D.1)}$$

D.2 SECOND MAXIMUM-LIKELIHOOD EQUATION

The derivative of the log-likelihood with respect to the couplings $J_{kl}(c, d)$ writes:

$$\begin{aligned}\frac{\partial \widetilde{\mathcal{L}}}{\partial J_{kl}(c, d)} &= \frac{\partial}{\partial J_{kl}(c, d)} \sum_{m=1}^{M} \log \sum_{b_1, ..., b_L} \left(\prod_{\{j|a_j \neq 0\}} \delta_{a_j, b_j} \right) P(b_1, ..., b_L) \\ &= \sum_{m=1}^{M} \frac{1}{\widetilde{P}(a_1^m, ..., a_L^m)} \sum_{b_1 \dots b_L} \left(\prod_{\{j|a_j^m \neq 0\}} \delta_{a_j^m, b_j} \right) \\ &\times \underbrace{\frac{\partial}{\partial J_{kl}(c, d)} \left(\frac{1}{\mathcal{Z}} \exp \sum_{i=1}^{L} h_i(b_i) + \sum_{i,j=1}^{L} J_{ij}(b_i, b_j) \right)}_{\beta_{kl}(c,d)} \ .\end{aligned}$$

But

$$\beta_k(c) = (\delta_{b_k,c}\delta_{b_l,d} - P_{kl}(c,d))P(b_1,...,b_L) \ ,$$

with $P_{kl}(c,d)$ the marginal of the probability distribution P:

$$P_{kl}(c,d) := \frac{1}{\mathcal{Z}} \sum_{b_1 \ldots b_L} \delta_{b_k,c}\delta_{b_l,d} \exp \sum_{i=1}^{L} h_i(b_i) + \sum_{i,j=1}^{L} J_{ij}(b_i,b_j) \ .$$

We then have

$$\frac{\partial \tilde{\mathcal{L}}}{\partial J_{kl}(c,d)} = -MP_{kl}(c,d) \sum_{m=1}^{M} \frac{1}{\tilde{P}(a_1^m,...,a_L^m)} \times \underbrace{\sum_{b_1 \ldots b_L} \left[\prod_{\{j|a_j^m \neq 0\}} \delta_{a_j^m,b_j} \right] \delta_{b_k,c}\delta_{b_l,d} P(b_1,...,b_L)}_{\xi_{kl}^m(c,d)} \ .$$

The quantity $\xi_{kl}^m(c,d)$ depends on the observation of a_k^m and a_l^m:

— If $a_k^m \neq 0$ and $a_l^m \neq 0$ (both observed):

$$\begin{aligned} \xi_{kl}^m(c,d) &= \sum_{b_1 \ldots b_L} \left(\prod_{\substack{\{j|a_j^m \neq 0\} \\ j \neq k, j \neq l}} \delta_{a_j^m,b_j} \right) \delta_{b_k,c}\delta_{b_l,d}\delta_{a_k^m,b_k}\delta_{a_l^m,b_l} P(b_1,...,b_L) \\ &= \delta_{a_k^m,c}\delta_{a_l^m,d} \times \widetilde{P}(a_1^m,...,a_L^m) \ . \end{aligned}$$

— If $a_k^m = 0$ and $a_l^m = 0$ (both not observed):

$$\begin{aligned} \xi_{kl}^m(c,d) &= \sum_{b_1 \ldots b_L} \left(\prod_{\substack{\{j|a_j^m \neq 0\} \\ j \neq k}} \delta_{a_j^m,b_j} \right) \delta_{b_k,c}\delta_{b_l,d} P(b_1,...,b_L) \\ &= \tilde{P}(a_1^m,...,a_k^m = c,...,a_l^m = d,...,a_L^m) \ . \end{aligned}$$

— If $a_k^m \neq 0$ and $a_l^m = 0$:

$$\begin{aligned} \xi_{kl}^m(c,d) &= \sum_{b_1 \ldots b_L} \left(\prod_{\substack{\{j|a_j^m \neq 0\} \\ j \neq k}} \delta_{a_j^m,b_j} \right) \delta_{b_k,c}\delta_{b_l,d}\delta_{a_k^m,b_k} P(b_1,...,b_L) \\ &= \delta_{a_k^m,c}\tilde{P}(a_1^m,...,a_l^m = d,...,a_L^m) \ . \end{aligned}$$

— If $a_k^m = 0$ and $a_l^m \neq 0$:

$$\xi_{kl}^m(c,d) = \delta_{a_l^m,d}\tilde{P}(a_1^m,...,a_k^m = c,...,a_L^m) \ .$$

Therefore

$$\begin{aligned}\frac{\partial\widetilde{\mathcal{L}}}{\partial J_{kl}(c,d)} = &\sum_{m=1}^{M}\delta_{a_k^m,c}\delta_{a_l^m,d} - MP_{kc}(l,d)\\ &+\sum_{\substack{\{m|a_k^m=0,\\ a_l^m=0\}}}\frac{\tilde{P}(a_1^m,...,a_k^m=c,...,a_l^m=d,...,a_L^m)}{\tilde{P}(a_1^m,...,a_L^m)}\\ &+\sum_{\{m|a_l^m=0\}}\delta_{a_k^m,c}\frac{\tilde{P}(a_1^m,...,a_l^m=d,...,a_L^m)}{\tilde{P}(a_1^m,...,a_L^m)}\\ &+\sum_{\{m|a_k^m=0\}}\delta_{a_l^m,d}\frac{\tilde{P}(a_1^m,...,a_k^m=c,...,a_L^m)}{\tilde{P}(a_1^m,...,a_L^m)}\ .\end{aligned}$$

Using the identity $\widetilde{P}(a_1^m,...,a_L^m) = P(\{a_i^m|a_i^m\neq 0\})$, we finally obtain the second maximum-likelihood equation:

$$\begin{aligned}MP_{kl}(c,d) = &\sum_{m=1}^{M}\delta_{a_k^m,c}\delta_{a_l^m,d}\\ &+\sum_{\substack{\{m|a_k^m=0,\\ a_l^m=0\}}}P(a_k^m=c,a_l^m=d|\{a_i^m|a_i^m\neq 0\})\\ &+\sum_{\{m|a_l^m=0\}}\delta_{a_k^m,c}P(a_l^m=d|\{a_i^m|a_i^m\neq 0\})\\ &+\sum_{\{m|a_k^m=0\}}\delta_{a_l^m,d}P(a_k^m=c|\{a_i^m|a_i^m\neq 0\})\end{aligned} \tag{D.2}$$

E

LIST OF PFAM FAMILIES ANALYZED IN PART III

E.1 LIST OF THE 70 PFAM FAMILIES

PF00226, PF00250, PF00618, PF00804, PF00806, PF00808, PF01638, PF02561, PF02909, PF06439, PF07647, PF12840, PF00011, PF00080, PF00169, PF00355, PF00595, PF00805, PF01458, PF05592, PF07559, PF10282, PF13360, PF14602, PF00032, PF00115, PF00208, PF00375, PF00529, PF00654, PF00689, PF00909, PF01035, PF01127, PF01699, PF01715, PF03349, PF07238, PF12700, PF13609, PF00005, PF00013, PF00069, PF00152, PF00290, PF00300, PF00445, PF00814, PF00849, PF01042, PF01244, PF01487, PF01713, PF03460, PF08334, PF08501, PF09360, PF12697, PF00014, PF00057, PF00084, PF00090, PF00105, PF00200, PF00412, PF00593, PF01774, PF01807, PF02953, PF07648.

E.2 STRUCTURAL CLASSIFICATION OF PROTEINS

The five structural groups we used are the following:

— **alpha proteins**, corresponding to the SCOP group *a* (domains consisting of α-helices);
Pfam IDs: PF00226, PF00250, PF00618, PF00804, PF00806, PF00808, PF01638, PF02561, PF02909, PF06439, PF07647, PF12840.

— **beta proteins**, corresponding to the SCOP group *b* (domains consisting of β-helices);
Pfam IDs: PF00011, PF00080, PF00169, PF00355, PF00595, PF00805, PF01458, PF05592, PF07559, PF10282, PF13360, PF14602.

— **alpha and beta proteins**, combining the SCOP groups *c*, *d* and *e* (group SCOP that includes proteins with both beta and alpha folds);
Pfam IDs: PF00005, PF00013, PF00069, PF00152, PF00290, PF00300, PF00445, PF00814, PF00849, PF01042, PF01244, PF01487, PF01713, PF03460, PF08334, PF08501, PF09360, PF12697.

— **membrane proteins**, corresponding to SCOP group *f* (membrane and cell surface proteins and peptides);
Pfam IDs: PF00032, PF00115, PF00208, PF00375, PF00529, PF00654, PF00689, PF00909, PF01035, PF01127, PF01699, PF01715, PF03349, PF07238, PF12700, PF13609.

— **small proteins**, corresponding to SCOP group *g* (usually dominated by metal ligand, heme, and/or disulfide bridges);
Pfam IDs: PF00014, PF00057, PF00084, PF00090, PF00105, PF00131, PF00200, PF00412, PF00593, PF01774, PF01807, PF02953, PF07648.

F

SILHOUETTE OF A CLUSTERING

As stated in the main text, the silhouettes $s(p)$ of a clustering method take the values $-1 < s(p) < 1$. Silhouette values close to 1 indicate appropriately clustered data points. Silhouette values close to -1 suggest that the corresponding data points would be better allocated to the neighboring cluster.

Fig. F.1 displays three very simple cases with the clustering of normally distributed random numbers and the corresponding silhouette values.

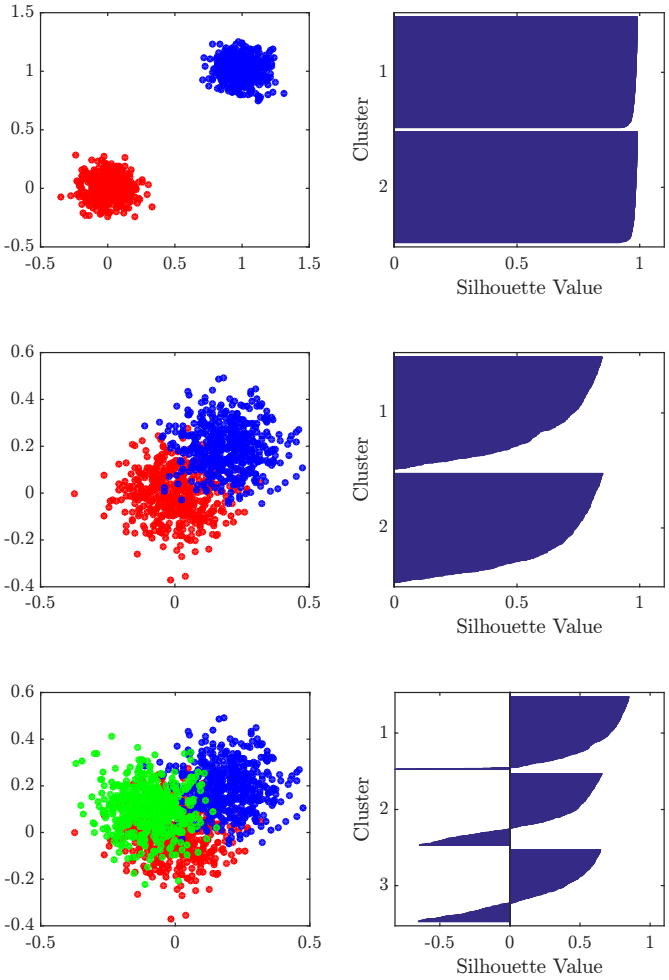


Figure F.1 – k-means clustering of 3 cases of normally distributed random numbers and the corresponding silhouette values.

BIBLIOGRAPHY

[1] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. « A learning algorithm for Boltzmann machines. » In: *Cognitive science* 9.1 (1985), pp. 147–169.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. « Molecular Biology of the Cell (Garland Science, New York, 2002). » In: *There is no corresponding record for this reference* (1997).

[3] Eric Alm, Alexandre V Morozov, Tanja Kortemme, and David Baker. « Simple physical models connect theory and experiment in protein folding kinetics. » In: *Journal of molecular biology* 322.2 (2002), pp. 463–476.

[4] Philip W Anderson et al. « More is different. » In: *Science* 177.4047 (1972), pp. 393–396.

[5] Christian B Anfinsen. *Studies on the principles that govern the folding of protein chains*. 1972.

[6] Erik Aurell. « The maximum entropy fallacy redux? » In: *PLoS Comput Biol* 12.5 (2016), e1004777.

[7] Erik Aurell and Magnus Ekeberg. « Inverse Ising inference using all the data. » In: *Physical review letters* 108.9 (2012), p. 090201.

[8] Marc Bailly-Bechet, Alfredo Braunstein, Andrea Pagnani, Martin Weigt, and Riccardo Zecchina. « Inference of sparse combinatorial-control networks from gene-expression data: a message passing approach. » In: *BMC bioinformatics* 11.1 (2010), p. 1.

[9] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G Carbonell, Su-In Lee, and Christopher James Langmead. « Learning generative models for protein fold families. » In: *Proteins: Structure, Function, and Bioinformatics* 79.4 (2011), pp. 1061–1078.

[10] 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 (2014), e92721.

[11] John P Barton, Mehran Kardar, and Arup K Chakraborty. « Scaling laws describe memories of host–pathogen riposte in the HIV population. » In: *Proceedings of the National Academy of Sciences* 112.7 (2015), pp. 1965–1970.

[12] John 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. » In: *Physical Review E* 90.1 (2014), p. 012132.

[13] John P Barton, Eleonora De Leonardis, Alice Coucke, and Simona Cocco. « Adaptive Cluster Expansion: inference of graphical models describing functional constraints. » In: *Bioinformatics* 32 (2016).

[14] John Barton and Simona Cocco. « Ising models for neural activity inferred via selective cluster expansion: structural and coding properties. » In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.03 (2013), P03002.

[15] Claudia Battistin, John Hertz, Joanna Tyrcha, and Yasser Roudi. « Belief propagation and replicas for inference and learning in a kinetic Ising model with hidden spins. » In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.5 (2015), P05021.

[16] Asa Ben-Hur and Douglas Brutlag. « Remote homology detection: a motif based approach. » In: *Bioinformatics* 19.suppl 1 (2003), pp. i26–i33.

[17] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. « The protein data bank. » In: *Nucleic Acids Research* 28.1 (2000), pp. 235–242.

[18] Helen Berman, Kim Henrick, and Haruki Nakamura. « Announcing the worldwide protein data bank. » In: *Nature Structural & Molecular Biology* 10.12 (2003), pp. 980–980.

[19] Matthew J Betts and Robert B Russell. « Amino acid properties and consequences of substitutions. » In: *Bioinformatics for geneticists* 317 (2003), p. 289.

[20] William Bialek. *Biophysics: searching for principles*. Princeton University Press, 2012.

[21] William Bialek and Rama Ranganathan. « Rediscovering the power of pairwise interactions. » In: *arXiv preprint arXiv:0712.4397* (2007).

[22] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Alan J Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios. « UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. » In: *Plant Bioinformatics: Methods and Protocols* (2016), pp. 23–54.

[23] Joseph D Bryngelson, Jose Nelson Onuchic, Nicholas D Socci, and Peter G Wolynes. « Funnels, pathways, and the energy landscape of protein folding: a synthesis. » In: *Proteins: Structure, Function, and Bioinformatics* 21.3 (1995), pp. 167–195.

[24] Lukas Burger and Erik Van Nimwegen. « Disentangling direct from indirect co-evolution of residues in protein alignments. » In: *PLoS Comput Biol* 6.1 (2010), e1000633.

[25] Thomas C Butler, John P Barton, Mehran Kardar, and Arup K Chakraborty. « Identification of drug resistance mutations in HIV from constraints on natural evolution. » In: *Physical Review E* 93.2 (2016), p. 022412.

[26] Simona Cocco, Stanislas Leibler, and Rémi Monasson. « Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. » In: *Proceedings of the National Academy of Sciences* 106.33 (2009), pp. 14058–14062.

[27] Simona Cocco and Rémi Monasson. « Adaptive cluster expansion for inferring Boltzmann machines with noisy data. » In: *Physical Review Letters* 106.9 (2011), p. 090601.

[28] Simona Cocco and Rémi Monasson. « Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. » In: *Journal of Statistical Physics* 147.2 (2012), pp. 252–314.

[29] 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. » In: *PLoS Comput Biol* 9.8 (2013), e1003176.

[30] UniProt Consortium et al. « UniProt: a hub for protein information. » In: *Nucleic Acids Research* (2014), gku989.

[31] Alice Coucke, Guido Uguzzoni, Francesco Oteri, Simona Cocco, Remi Monasson, and Martin Weigt. « Direct coevolutionary couplings reflect biophysical residue interactions in proteins. » In: *The Journal of Chemical Physics* (2016, submitted).

[32] Nello Cristianini and Matthew W Hahn. *Introduction to computational genomics: a case studies approach*. Cambridge University Press, 2006.

[33] Angel E Dago, Alexander Schug, Andrea Procaccini, James A Hoch, Martin Weigt, and Hendrik Szurmant. « Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. » In: *Proceedings of the National Academy of Sciences* 109.26 (2012), E1733–E1742.

[34] Eleonora De Leonardis, Benjamin Lutz, Simona Cocco, Remi Monasson, Hendrik Szurmant, Martin Weigt, and Alexander Schug. « Protein and RNA Structure Prediction by Integration of Co-Evolutionary Information into Molecular Simulation. » In: *Biophysical Journal* 108.2 (2015), 13a–14a.

[35] Russell J Dickson and Gregory B Gloor. « Protein sequence alignment analysis by local covariation: coevolution statistics detect benchmark alignment errors. » In: *PLoS One* 7.6 (2012), e37645.

[36] Ricardo N Dos Santos, Faruck Morcos, Biman Jana, Adriano D Andricopulo, and José N Onuchic. « Dimeric interactions and complex formation using direct coevolutionary couplings. » In: *Scientific reports* 5 (2015).

[37] Benjamin Dunn and Yasser Roudi. « Learning and inference in a nonequilibrium Ising model with hidden nodes. » In: *Physical Review E* 87.2 (2013), p. 022127.

[38] Stanley D Dunn, Lindi M Wahl, and Gregory B Gloor. « Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. » In: *Bioinformatics* 24.3 (2008), pp. 333–340.

[39] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[40] Sean R. Eddy. « Profile hidden Markov models. » In: *Bioinformatics* 14.9 (1998), pp. 755–763.

[41] Magnus Ekeberg, Tuomo Hartonen, and Erik Aurell. « Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. » In: *Journal of Computational Physics* 276 (2014), pp. 341–356.

[42] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. « Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. » In: *Physical Review E* 87.1 (2013), p. 012707.

[43] Christoph Feinauer, Marcin J Skwark, Andrea Pagnani, and Erik Aurell. « Improving contact prediction along three dimensions. » In: *PLOS Comp Biol* 10.10 (2014), e1003847.

[44] 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 (2016), e0149166.

[45] Andrew L Ferguson, Jaclyn K Mann, Saleha Omarjee, Thumbi Ndung'u, Bruce D Walker, and Arup K Chakraborty. « Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. » In: *Immunity* 38.3 (2013), pp. 606–617.

[46] 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. » In: *Molecular biology and evolution* 33.1 (2016), pp. 268–280.

[47] Robert D Finn, Jody Clements, and Sean R Eddy. « HMMER web server: interactive sequence similarity searching. » In: *Nucleic Acids Research* (2011), gkr367.

[48] Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, et al. « Pfam: the protein families database. » In: *Nucleic Acids Research* (2013), gkt1223.

[49] Robert D Finn, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, et al. « The Pfam protein families database: towards a more sustainable future. » In: *Nucleic Acids Research* (2015), gkv1344.

[50] Anthony A Fodor and Richard W Aldrich. « Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. » In: *Proteins: Structure, Function, and Bioinformatics* 56.2 (2004), pp. 211–221.

[51] Eric O Freed. « HIV-1 assembly, release and maturation. » In: *Nature Reviews Microbiology* 13.8 (2015), pp. 484–496.

[52] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. « A weakly informative default prior distribution for logistic and other regression models. » In: *The Annals of Applied Statistics* (2008), pp. 1360–1383.

[53] Antoine Georges and Jonathan S Yedidia. « How to expand around mean-field theory using high-temperature expansions. » In: *Journal of Physics A: Mathematical and General* 24.9 (1991), p. 2173.

[54] Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. « Correlated mutations and residue contacts in proteins. » In: *Proteins: Structure, Function, and Bioinformatics* 18.4 (1994), pp. 309–317.

[55] Jean-Pierre Hansen and Ian R McDonald. *Theory of simple liquids*. Elsevier, 1990.

[56] Jan Heyda, Philip E Mason, and Pavel Jungwirth. « Attractive interactions between side chains of histidine-histidine and histidine-arginine-based cationic dipeptides in water. » In: *The Journal of Physical Chemistry B* 114.26 (2010), pp. 8744–8749.

[57] Thomas A Hopf, Lucy J Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, and Debora S Marks. « Three-dimensional structures of membrane proteins from genomic sequencing. » In: *Cell* 149.7 (2012), pp. 1607–1621.

[58] Simon J Hubbard and Janet M Thornton. « Naccess. » In: *Computer Program, Department of Biochemistry and Molecular Biology, University College London* 2.1 (1993).

[59] Tommi Jaakkola, Mark Diekhans, and David Haussler. « A discriminative framework for detecting remote protein homologies. » In: *Journal of computational biology* 7.1-2 (2000), pp. 95–114.

[60] Hugo Jacquin, Amy Gilson, Eugene Shakhnovich, Simona Cocco, and Rémi Monasson. « Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. » In: *PLoS Comput Biol* 12.12 (2016), e1004889.

[61] Edwin T Jaynes. « Information theory and statistical mechanics. II. » In: *Physical Review* 108.2 (1957), p. 171.

[62] Edwin T Jaynes. « Information theory and statistical mechanics. » In: *Physical Review* 106.4 (1957), p. 620.

[63] George A Jeffrey and George A Jeffrey. *An introduction to hydrogen bonding*. Vol. 12. Oxford university press New York, 1997.

[64] David T Jones, Daniel WA Buchan, Domenico Cozzetto, and Massimiliano Pontil. « PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. » In: *Bioinformatics* 28.2 (2012), pp. 184–190.

[65] David de Juan, Florencio Pazos, and Alfonso Valencia. « Emerging methods in protein co-evolution. » In: *Nature Reviews Genetics* 14.4 (2013), pp. 249–261.

[66] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. « Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. » In: *Proceedings of the National Academy of Sciences* 110.39 (2013), pp. 15674–15679.

[67] Hilbert J. Kappen and Francisco de Borja Rodríguez. « Efficient learning in Boltzmann machines using linear response theory. » In: *Neural Computation* 10.5 (1998), pp. 1137–1156.

[68] Alan S Lapedes, Bertrand G Giraud, LonChang Liu, and Gary D Stormo. « Correlated mutations in models of protein sequences: phylogenetic and structural effects. » In: *Lecture Notes-Monograph Series* (1999), pp. 236–256.

[69] Cyrus Levinthal. « How to fold graciously. » In: *Mossbauer spectroscopy in biological systems* 67 (1969), pp. 22–24.

[70] Kevin Liu, Sindhu Raghavan, Serita Nelesen, C Randal Linder, and Tandy Warnow. « Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. » In: *Science* 324.5934 (2009), pp. 1561–1564.

[71] Craig D Livingstone and Geoffrey J Barton. « Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. » In: *Computer applications in the biosciences: CABIOS* 9.6 (1993), pp. 745–756.

[72] Steve W Lockless and Rama Ranganathan. « Evolutionarily conserved pathways of energetic connectivity in protein families. » In: *Science* 286.5438 (1999), pp. 295–299.

[73] Sara Lui and Guido Tiana. « The network of stabilizing contacts in proteins studied by coevolutionary data. » In: *The Journal of chemical physics* 139.15 (2013), p. 155103.

[74] Jianzhu Ma, Sheng Wang, Zhiyong Wang, and Jinbo Xu. « MRFalign: protein homology detection through alignment of Markov random fields. » In: *PLoS Comput Biol* 10.3 (2014), e1003500.

[75] Kira S Makarova and Eugene V Koonin. « Two new families of the FtsZ-tubulin protein superfamily implicated in membrane remodeling in diverse bacteria and archaea. » In: *Biology direct* 5.1 (2010), p. 33.

[76] 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 (2014), e1003776.

[77] Debora S Marks, Thomas A Hopf, and Chris Sander. « Protein structure prediction from sequence variation. » In: *Nature biotechnology* 30.11 (2012), pp. 1072–1080.

[78] Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. « Protein 3D structure computed from evolutionary sequence variation. » In: *PloS one* 6.12 (2011), e28766.

[79] Sanzo Miyazawa and Robert L Jernigan. « Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. » In: *Macromolecules* 18.3 (1985), pp. 534–552.

[80] 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 (1996), pp. 623–644.

[81] Salameh Moh'd A, Alexei S Soares, Duraiswamy Navaneetham, Dipali Sinha, Peter N Walsh, and Evette S Radisky. « Determinants of affinity and proteolytic stability in interactions of Kunitz family protease inhibitors with mesotrypsin. » In: *Journal of Biological Chemistry* 285.47 (2010), pp. 36884–36896.

[82] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. « Direct-coupling analysis of residue coevolution captures native contacts across many protein families. » In: *Proceedings of the National Academy of Sciences* 108.49 (2011), E1293–E1301.

[83] Faruck Morcos, Nicholas P Schafer, Ryan R Cheng, José N Onuchic, and Peter G Wolynes. « Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. » In: *Proceedings of the National Academy of Sciences* 111.34 (2014), pp. 12408–12413.

[84] Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia. « SCOP: a structural classification of proteins database for the investigation of sequences and structures. » In: *Journal of molecular biology* 247.4 (1995), pp. 536–540.

[85] Sehat Nauli, Brian Kuhlman, Isolde Le Trong, Ronald E Stenkamp, David Teller, and David Baker. « Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2. » In: *Protein science* 11.12 (2002), pp. 2924–2931.

[86] Eric P Nawrocki, Diana L Kolbe, and Sean R Eddy. « Infernal 1.0: inference of RNA alignments. » In: *Bioinformatics* 25.10 (2009), pp. 1335–1337.

[87] Angel R Ortiz, Wei Ping Hu, Andrzej Kolinski, and Jeffrey Skolnick. « Method for low resolution prediction of small protein tertiary structure. » In: *Pacific Symposium on Biocomputing*. Vol. 97. 1997, pp. 316–327.

[88] Angel R Ortiz, Andrzej Kolinski, Piotr Rotkiewicz, Bartosz Ilkowski, and Jeffrey Skolnick. « Ab initio folding of proteins using restraints derived from evolutionary information. » In: *Proteins: Structure, Function, and Bioinformatics* 37.S3 (1999), pp. 177–185.

[89] Giorgio Parisi. *Statistical field theory*. Addison-Wesley, 1988.

[90] Alessandro Pelizzola. « Cluster variation method in statistical physics and probabilistic graphical models. » In: *Journal of Physics A: Mathematical and General* 38.33 (2005), R309.

[91] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. « UCSF Chimera—a visualization system for exploratory research and analysis. » In: *Journal of computational chemistry* 25.13 (2004), pp. 1605–1612.

[92] T Plefka. « Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. » In: *Journal of Physics A: Mathematical and general* 15.6 (1982), p. 1971.

[93] Andrea Procaccini, Bryan Lunt, Hendrik Szurmant, Terence Hwa, and Martin Weigt. « Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. » In: *PloS one* 6.5 (2011), e19729.

[94] Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. « High-dimensional Ising model selection using l1-regularized logistic regression. » In: *The Annals of Statistics* 38.3 (2010), pp. 1287–1319.

[95] Federico Ricci-Tersenghi. « The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. » In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.08 (2012), P08015.

[96] Martin Riedmiller and Heinrich Braun. « A direct adaptive method for faster backpropagation learning: The RPROP algorithm. » In: *Neural Networks, 1993., IEEE International Conference On*. IEEE. 1993, pp. 586–591.

[97] Sophie Rosay. « A statistical mechanics approach to the modelling and analysis of place-cell activity. » PhD thesis. Ecole normale supérieure-ENS PARIS, 2014.

[98] Yasser Roudi, Joanna Tyrcha, and John Hertz. « Ising model for neural data: model quality and approximate methods for extracting functional connectivity. » In: *Physical Review E* 79.5 (2009), p. 051915.

[99] William P Russ, Drew M Lowery, Prashant Mishra, Michael B Yaffe, and Rama Ranganathan. « Natural-like function in artificial WW domains. » In: *Nature* 437.7058 (2005), pp. 579–583.

[100] Dmitry Rykunov and Andras Fiser. « New statistical potential for quality assessment of protein models and a survey of energy functions. » In: *BMC bioinformatics* 11.1 (2010), p. 1.

[101] Hiroto Saigo, Jean-Philippe Vert, Nobuhisa Ueda, and Tatsuya Akutsu. « Protein homology detection using string alignment kernels. » In: *Bioinformatics* 20.11 (2004), pp. 1682–1689.

[102] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. « Weak pairwise correlations imply strongly correlated network states in a neural population. » In: *Nature* 440.7087 (2006), pp. 1007–1012.

[103] Benjamin Schuster-Böckler, Jörg Schultz, and Sven Rahmann. « HMM Logos for visualization of protein families. » In: *BMC bioinformatics* 5.1 (2004), p. 1.

[104] Eugene Shakhnovich and Alexander Gutin. « Enumeration of all compact conformations of copolymers with random sequence of links. » In: *The Journal of Chemical Physics* 93.8 (1990), pp. 5967–5971.

[105] Claude Elwood Shannon. « A mathematical theory of communication. » In: *The Bell System Technical Journal* 27 (1958), 379—423,623—656.

[106] Marcin J Skwark, Daniele Raimondi, Mirco Michel, and Arne Elofsson. « Improved contact predictions using the recognition of protein like contact patterns. » In: *PLoS Comput Biol* 10.11 (2014), e1003889.

[107] Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, Kevin H Gardner, and Rama Ranganathan. « Evolutionary information for specifying a protein fold. » In: *Nature* 437.7058 (2005), pp. 512–518.

[108] Johannes Söding. « Protein homology detection by HMM–HMM comparison. » In: *Bioinformatics* 21.7 (2005), pp. 951–960.

[109] Johannes Söding, Andreas Biegert, and Andrei N Lupas. « The HHpred interactive server for protein homology detection and structure prediction. » In: *Nucleic acids research* 33.suppl 2 (2005), W244–W248.

[110] Joanna I Sułkowska, Faruck Morcos, Martin Weigt, Terence Hwa, and José N Onuchic. « Genomics-aided structure prediction. » In: *Proceedings of the National Academy of Sciences* 109.26 (2012), pp. 10340–10345.

[111] Sylvanie Surget, Marie P Khoury, and Jean-Christophe Bourdon. « Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. » In: *Onco Targets Ther* 7 (2014), pp. 57–68.

[112] Ludovico Sutto, Simone Marsili, Alfonso Valencia, and Francesco Luigi Gervasio. « From residue coevolution to protein conformational ensembles and functional dynamics. » In: *Proceedings of the National Academy of Sciences* 112.44 (2015), pp. 13567–13572.

[113] Gaia Tavoni, Ulisse Ferrari, Francesco Paolo Battaglia, Simona Cocco, and Rémi Monasson. « Inferred Model of the Prefrontal Cortex Activity Unveils Cell Assemblies and Memory Replay. » In: *bioRxiv* (2015), p. 028316.

[114] Elisabeth RM Tillier and Thomas WH Lui. « Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. » In: *Bioinformatics* 19.6 (2003), pp. 750–755.

[115] Martin J Wainwright and Michael I Jordan. « Graphical models, exponential families, and variational inference. » In: *Foundations and Trends® in Machine Learning* 1.1-2 (2008), pp. 1–305.

[116] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. « Identification of direct residue contacts in protein–protein interaction by message passing. » In: *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 67–72.

[117] Max Welling and Yee Whye Teh. « Approximate inference in Boltzmann machines. » In: *Artificial Intelligence* 143.1 (2003), pp. 19–50.

[118] Kurt R Wollenberg and William R Atchley. « Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. » In: *Proceedings of the National Academy of Sciences* 97.7 (2000), pp. 3288–3291.

[119] Fa-Yueh Wu. « The potts model. » In: *Reviews of modern physics* 54.1 (1982), p. 235.

[120] Jinchao Yu, Marek Vavrusa, Jessica Andreani, Julien Rey, Pierre Tufféry, and Raphaël Guerois. « InterEvDock: a docking server to predict the structure of protein–protein interactions using evolutionary information. » In: *Nucleic acids research* (2016), gkw340.

[121] Hui Zeng, Ke-Song Liu, and Wei-Mou Zheng. « The Miyazawa-Jernigan Contact Energies Revisited. » In: *Open Bioinformatics Journal* 6 (2012), pp. 1–8.

Résumé

Grâce aux progrès des techniques de séquençage, les bases de données génomiques ont connu une croissance exponentielle depuis la fin des années 1990. Un grand nombre d'outils statistiques ont été développés à l'interface entre bioinformatique, apprentissage automatique et physique statistique, dans le but d'extraire de l'information de ce déluge de données. Plusieurs approches de physique statistique ont été récemment introduites dans le contexte précis de la modélisation de séquences de protéines, dont l'analyse en couplages directs. Cette méthode d'inférence statistique globale fondée sur le principe d'entropie maximale, s'est récemment montrée d'une efficacité redoutable pour prédire la structure tridimensionnelle de protéines, à partir de considérations purement statistiques.

Dans cette thèse, nous présentons les méthodes d'inférence en question, et encouragés par leur succès, explorons d'autres domaines complexes dans lesquels elles pourraient être appliquées, comme la prédiction de repliement de protéines ou la détection d'homologies. Contrairement à la prédiction des contacts entre résidus qui se limite à une information topologique sur le réseau d'interactions, ces nouveaux champs d'application exigent des considérations énergétiques globales et donc un modèle plus quantitatif et détaillé. À travers une étude approfondie sur des données artificielles et biologiques, nous proposons une meilleure interpretation des paramètres centraux de ces méthodes d'inférence, jusqu'ici mal compris, notamment dans le cas d'un échantillonnage limité. Enfin, nous présentons une nouvelle procédure plus précise d'inférence de modèles génératifs, qui mène à des avancées importantes pour des données réelles en quantité limitée.

Mots-Clefs

inférence, apprentissage statistique, régularisation, entropie maximale, coévolution des protéines, modélisation statistique des séquences de protéines, vraisemblance maximale, champ moyen, pseudo vraisemblance, développement en grappe

Abstract

Over the last decades, genomic databases have grown exponentially in size thanks to the constant progress of modern DNA sequencing. A large variety of statistical tools have been developed, at the interface between bioinformatics, machine learning, and statistical physics, to extract information from these ever increasing datasets. In the specific context of protein sequence data, several approaches have been recently introduced by statistical physicists, such as direct-coupling analysis, a global statistical inference method based on the maximum-entropy principle, that has proven to be extremely effective in predicting the three-dimensional structure of proteins from purely statistical considerations.

In this dissertation, we review the relevant inference methods and, encouraged by their success, discuss their extension to other challenging fields, such as sequence folding prediction and homology detection. Contrary to residue-residue contact prediction, which relies on an intrinsically topological information about the network of interactions, these fields require global energetic considerations and therefore a more quantitative and detailed model. Through an extensive study on both artificial and biological data, we provide a better interpretation of the central inferred parameters, up to now poorly understood, especially in the limited sampling regime. Finally, we present a new and more precise procedure for the inference of generative models, which leads to further improvements on real, finitely sampled data.

Keywords

inference, statistical learning, regularization, maximum entropy, protein coevolution, statistical modeling of protein sequences, maximum likelihood, mean field, pseudolikelihood, cluster expansion