Towards a genome-scale coevolutionary analysis

Giancarlo Croce

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS
DEPUIS 1257

THÈSE DE DOCTORAT
DE SORBONNE UNIVERSITÉ

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

réalisée au
Laboratoire de Biologie Computationnelle et Quantitative

sous la direction de Martin WEIGT
et la co-direction de Olivier TENAILLON

présentée par

Giancarlo CROCE

Sujet de la thèse :

Towards a genome-scale coevolutionary analysis

Présentée et soutenue publiquement le 22 Novembre 2019

devant un jury composé de :

M.	Erik	AURELL	Rapporteur
M.	Paolo	DE LOS RIOS	Rapporteur
Mme	Alessandra	CARBONE	Examinatrice
M.	Philippe	LOPEZ	Examinateur
M.	Martin	WEIGT	Directeur de thèse
M.	Olivier	TENAILLON	Co-directeur de thèse

Giancarlo Croce: *Towards a genome-scale coevolutionary analysis,*
November 2019

ABSTRACT

Advances in sequencing technologies have revolutionized the life sciences. The explosion of genomic sequence data has prompted the development of a wide variety of methods, at the interface between bioinformatics, machine learning, and physics, which aim at gaining a deeper understanding of biological systems from such data.

Pairwise coevolutionary methods, in particular Direct Coupling Analysis (DCA), can extract a multitude of information from sequence data alone, such as structural contacts or phenotypic effects of amino-acid substitutions in proteins. While they have been mainly applied to a number of single exemplary proteins, it is now time for a broader application at the level of the whole genome.

In this thesis, we build upon and extend these models to address biological questions at the genome scale. In a first project, we investigate the protein-protein interaction network by combining coevolutionary signals at multiple but interconnected scales. In a subsequent project, we discuss the possibility of including complementary information to sequences, such as typical patterns of contacts, to improve the inter-protein contact prediction. Finally, through an extensive genome-wide study of *E. coli* strains, we show how the machinery of DCA can be used to investigate the fitness landscape properties at the local and global scales.

RÉSUMÉ

Les progrès des technologies de séquençage ont révolutionné les sciences de la vie. L'explosion de données de séquences génomiques a conduit au développement d'une grande variété de méthodes, à l'interface entre la bioinformatique, l'apprentissage automatique et la physique, qui visent à approfondir la compréhension des systèmes biologiques à partir de telles données.

Les méthodes coévolutives, telles que l'analyse par couplage direct (DCA), peuvent extraire une multitude d'informations à partir de données de séquence uniquement, telles que des contacts structurels ou des effets phénotypiques de substitutions d'acides aminés dans des protéines. Bien qu'elles aient été principalement appliquées à un certain nombre de protéines exemplaires, il est maintenant temps de les appliquer au niveau du génome entier.

Dans cette thèse, nous nous appuions sur ces modèles et les développons pour traiter des questions biologiques à l'échelle du génome. Dans un premier projet, nous avons étudié le réseau d'interactions protéine-protéine en combinant des signaux coévolutifs à des échelles multiples mais interconnectées. Dans un projet ultérieur, nous discutons de la possibilité d'inclure des informations complémentaires aux séquences, telles que des schémas de contacts typiques, afin d'améliorer la prédiction de contacts entre protéines. Enfin, à travers une vaste étude portant sur l'ensemble du génome des souches d'*E. Coli*, nous montrons comment les mécanismes de la DCA peuvent être utilisés pour étudier les propriétés du paysage de la fitness à l'échelle locale et globale.

ACKNOWLEDGMENTS

First and foremost I want to express my special appreciation and sincere gratitude to my advisor Martin Weigt who made this work possible. He introduced me to the world of biology and his guidance and expert advice have been invaluable throughout all the stages of this thesis.

Besides my advisor, I would like to thank my co-supervisor Olivier Tenaillon for extended discussions and valuable suggestions that have contributed greatly to the improvement of the thesis.

I would also wish to express my gratitude to Paolo De Los Rios and to Erik Aurel for accepting to be my thesis reporter and to Alessandra Carbone and Philippe Lopez for their participation to my thesis committee.

I thank my fellow labmates in LCQB and especially to Pierre, Edoardo, Anna, Kai and Maureen for the stimulating discussions, for the dinners we had together, and for all the fun we have had in the last three years. A special thanks go to Edwin and Carlos for hosting me in Cuba. I have very fond memories of my time there. I would also like to thank all the friends I met outside the lab during this years in Paris. In particolar modo Stefania e Luca sia per le interminabilie cene e discussioni sia per essere sempre stati presenti durante gli alti e bassi di questa avventura parigina. Un grazie anche a Denny che, da Pavia, ha seguito le mie orme qui a Parigi per poi partire a Colonia dove, ne sono certo, avrai modo di dimostrare la tua bravura.

Infine un ringraziamento speciale va alla mia famiglia per la loro costante presenza e incoraggiamento. A loro dedico questo mio lavoro.

CONTENTS

I INTRODUCTION

1 PROTEINS, MULTIPLE SEQUENCE ALIGNMENTS AND CO-EVOLUTION 3
- 1.1 Proteins 3
- 1.2 Protein databases 5
- 1.3 Protein family and Multiple Sequence Alignment 5
  - 1.3.1 Profile Hidden Markov Models 7
  - 1.3.2 Protein domains and the Pfam database 9
- 1.4 Epistasis and coevolution 10
  - 1.4.1 Protein-structure prediction using residue co-evolution 11
  - 1.4.2 Coevolution in a more general sense 11

2 STATISTICAL ANALYSIS OF PROTEIN SEQUENCE DATA 13
- 2.1 Maximum entropy modelling 14
  - 2.1.1 Profile models 15
  - 2.1.2 Direct coupling Analysis 15
  - 2.1.3 Criticism to MaxEnt 17
- 2.2 Inference methods for the inverse potts problem 19
  - 2.2.1 Boltzmann Machine learning 20
  - 2.2.2 Mean Field 21
  - 2.2.3 Pseudolikelihood Maximization 22
- 2.3 Technical points 23
  - 2.3.1 Gauge Transformation 23
  - 2.3.2 Regularization 24
  - 2.3.3 Sequence re-weighting 25
  - 2.3.4 Frobenius norm 26
  - 2.3.5 Average Product Correction 27
- 2.4 Application of DCA 27
  - 2.4.1 Contact prediction 28
  - 2.4.2 Protein-protein interaction 29
  - 2.4.3 Mutational landscape 33
  - 2.4.4 Scoring of sequences 34
  - 2.4.5 Genome-wide DCA 36

II PROTEIN-PROTEIN INTERACTIONS

3 PHYLETIC DIRECT COUPLING ANALYSIS 41
- 3.1 Motivation 41
- 3.2 Article 42

4 FILTER DCA 65
- 4.1 Introduction on CNN 66
  - 4.1.1 Deep learning for inter-protein residue-residue contacts 69

4.2 Filter DCA: Methods 69
4.2.1 Dataset 69
4.2.2 DCA performance on data 70
4.2.3 Secondary structure and contact patterns 71
4.2.4 Filter score 75
4.2.5 Learning procedure 76
4.2.6 Filter size 77
4.3 Results 79
4.4 Conclusion and Discussion 79

III FITNESS LANDSCAPE

5 LOCAL FITNESS LANDSCAPE 85
5.1 Fitness landscape 85
5.2 DCA and Fitness landscape 87
5.3 The dataset 90
5.3.1 Local sampling 90
5.3.2 Global sampling 91
5.3.3 Core genome 92
5.4 Epistasis 92
5.5 Quantifying context dependence 95
5.6 Summary and Outlook on Future Work 100
5.6.1 Towards DCA as evolutionary model? 100

IV CONCLUDING REMARKS

V APPENDIX

A APPENDIX 109
A.1 PhyDCA: Supplementary information 109
A.1.1 Input data 109
A.1.2 Similarity measures 109
A.1.3 Average product correction 110
A.1.4 Results 111
A.1.5 Paralog matching 111
A.1.6 Network analysis 112
A.1.7 All bacteria 117
A.2 FilterDCA: Supplementary information 119
A.2.1 Handling imbalanced datasets 119
A.2.2 Comparison issues 119

BIBLIOGRAPHY 121

ACRONYMS

MSA Multiple Sequence Alignment

HMM Hidden Markov Model

PDB Protein DataBank

MI Mutual Information

DI Direct information

MaxEnt Maximum-Entropy Principle

DCA Direct Coupling Analysis

MF Mean-Field

PLM Pseudo-Likelihood Maximization

BML Boltzmann Machine Learning

APC Average Product Correction

PPV Positive Predictive Value

ML Machine Learning

DL Deep Learning

CNN Convolutional Neural Network

PPI Protein-protein interaction

Part I

INTRODUCTION

The following chapters serve as an introduction to the research field. Chapter 1 contains the fundamental biological concepts that are needed to understand the methods, results, and discussions in the thesis. We focus on proteins and protein databases which play an ever-increasing important role in modern biology. In Chapter 2, we present the theoretical basis of the statistical-mechanics inspired methods which are used in protein sequence analysis. We then introduce Direct Coupling Analysis (DCA) together with some of the most important applications, including the prediction of intra- or inter- protein residue-residue contacts or the prediction of mutational effects.

1

PROTEINS, MULTIPLE SEQUENCE ALIGNMENTS AND COEVOLUTION

1.1 PROTEINS

Virtually every property that characterizes a living organism is affected by proteins. Proteins store and transport a variety of particles; as hormones, they transmit information between cells and organs in complex organism; as antibodies they recognize and latch onto antigens in order to remove them from the body; they control gene expression, thereby turning genes on and off, and many proteins are simply used as structural elements.

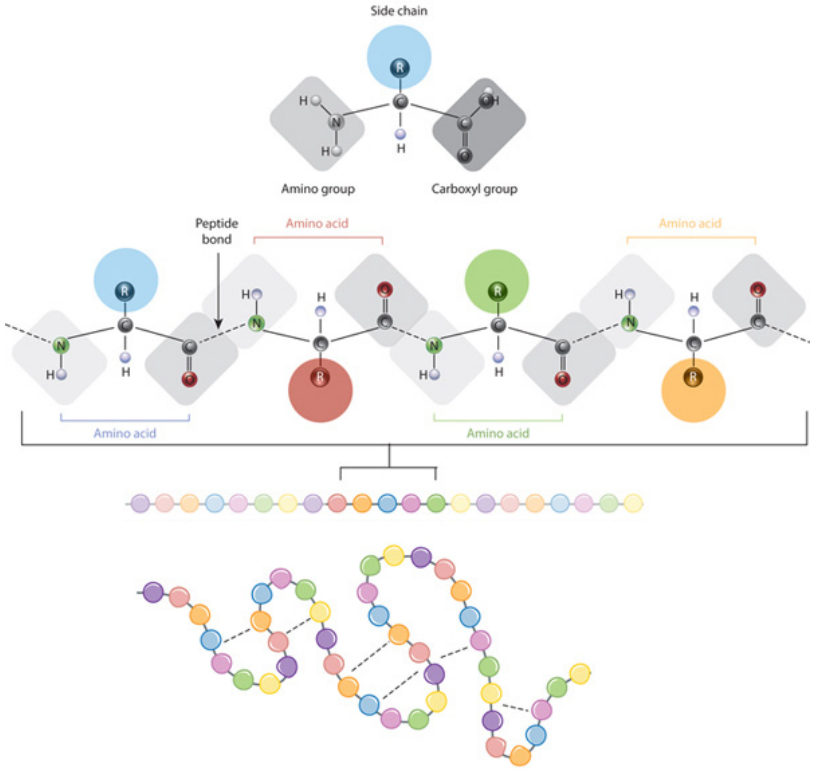

Figure 1.1: When connected together by a series of peptide bonds, amino acids form a polypeptide which then folds into a specific conformation depending on the interactions between its amino acid side chains. Source: [1].

Despite these diverse biological functions, all proteins consist of one or more long polymers chains built from series of up to 20 different

amino acid residues, linked to each other by a peptide bond. All the 20 amino acids ordinarily found in proteins contain amine (-NH2) and carboxyl (-COOH) along with a side chain (R group) specific to each amino acid (see Figure 1.1).

The secret of protein functional diversity lies partly in the physico-chemical diversity of the amino acids - charge, size, hydrophobicity (see Figure 1.2) - but primarily in the diversity of the three-dimensional structures that they can form after folding. Because of side-chain interactions, polymer chains bend, twist and flex into a very large variety of three dimensional stable structures.

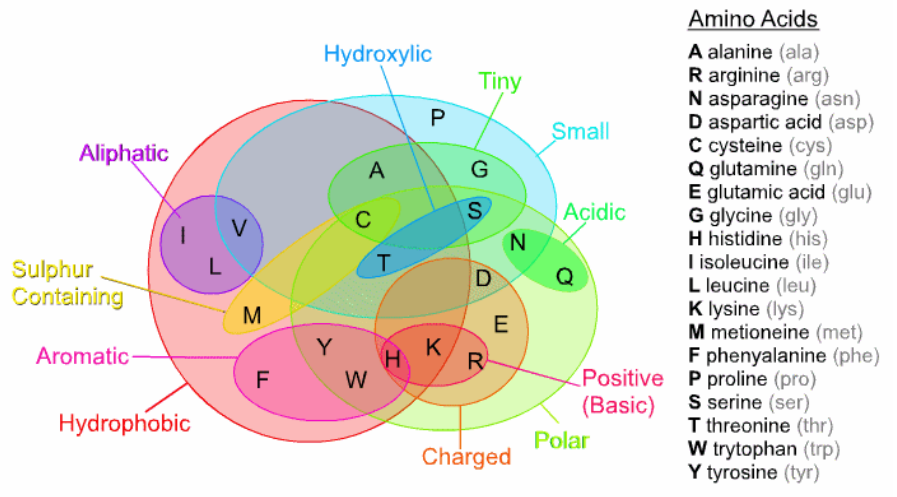

Figure 1.2: The 20 naturally-occurring amino acids clustered by their physico-chemical properties. Source [2].

The linear amino acid sequence identifies a protein unambiguously. It determines all its chemical and biological properties and indirectly specifies the higher levels of protein structure (see Figure 1.3). The structure can be described on four distinct levels:

- *primary structure* - the linear sequence of amino acids in the polypeptide backbone[1];
- *secondary structure* - folding patterns within a polypeptide resulting from hydrogen bonds between atoms of the backbone, mainly α-helices and β-sheets along with less structured loops [2];
- *tertiary structure* - the overall three dimensional shape of a polypeptide chain determined by the interactions between the side chains of the various amino acids;

1 The backbone just refers to the polypeptide chain apart from the R groups.
2 Only certain types of secondary structures are possible due to the planar nature of the peptide bonds.

- *quaternary structure* - combination of multiple polypeptide chains assembled via intermolecular interactions.

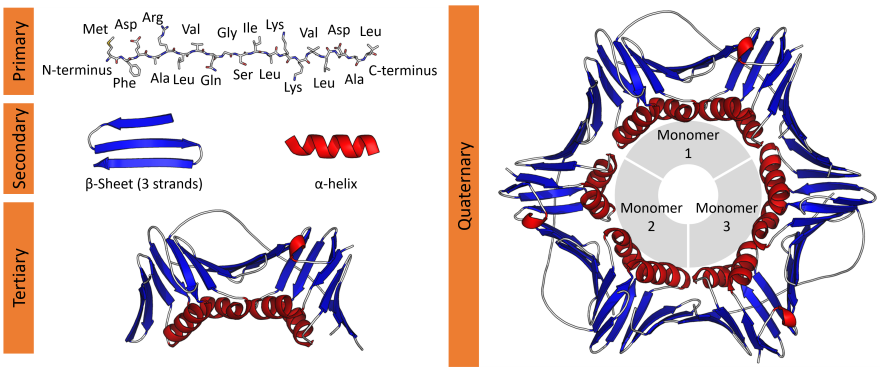

Figure 1.3: The four distinct levels of protein structure. Source: [3].

1.2 PROTEIN DATABASES

Anfinsen's seminal studies [4] established that the sequence of a protein determines its structure and function uniquely. The genotype to the phenotype map (i.e. sequence to structure or to function) is an experimentally formidable task. Experiments designed to provide a manual annotation of a protein - such as the description of its function, its structure or its interactions - are time-consuming and expensive, thereby limiting the number of sequences that can be studied. Experimental structures, obtained with methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy are freely accessible in the Protein DataBank (PDB) [5][6].

In the last decades, the technological breakthroughs in high-throughput sequencing followed by the rise of massive protein sequence databases, opened the door to a plethora of new computational methods [7] which aim at characterizing the phenotype of a protein from sequence data alone. Today, one of the largest public repository of protein sequences is UniProt Knowledgebase (UniProtKB) [8]. It consists of two sections: the UniProtKB/SwissProt database with high-quality manually-annotated records and the UniProtKB/TrEMBL containing computationally analyzed records. As of June 2019, UniProtKB/TrEMBL contains more than 150 million protein sequences compared with the 500.000 of the UniProtKB/SwissProt database. As shown in Figure 1.4, the gap widens across the years.

1.3 PROTEIN FAMILY AND MULTIPLE SEQUENCE ALIGNMENT

It is natural to cluster sequences of the Uniprot database into groups of evolutionarily-related proteins. Homologous proteins, that share a common ancestor, are usually classified into *protein families*. These

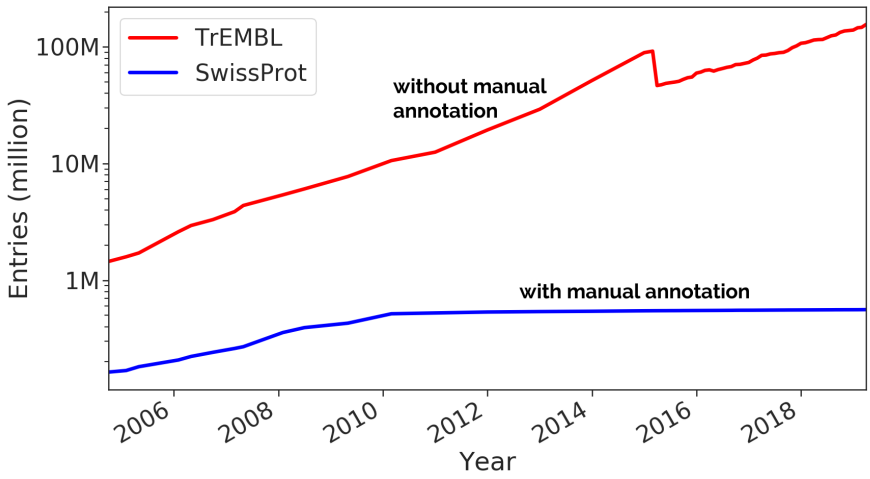

Figure 1.4: Number of entries in the UniProt database across the years. The gap widens between TrEMBL (unreviewed automatically annotated sequences) and SwissProt (manually annotated entries). While the TrEMBL dataset is still exponentially growing, relatively few new entries were added to the SwissProt database since 2010. The research effort is indeed more directed at improving the quality of the annotations (structural or functional) rather than their quantity. Source: [8].

homologs can be orthologs, that were separated by a speciation event, or paralogs, that were separated by a duplication event inside one species. Within a protein family, all members are subjected to comparable evolutionary pressure and, as a result of that, they share similar three-dimensional structure and function. Nevertheless, due to amino acid substitutions, insertions and deletions across millions of years of evolution, they can have a high variety in amino acid sequences: the average sequence identity between two homologous proteins is $20 - 30\%$.

To make data more amenable to statistical analysis, it is useful to arrange homologous sequences into a data matrix: the Multiple Sequence Alignment (MSA) (see Figure 1.5). Formally, an MSA is a rectangular matrix $A = \{a_i^m | i = 1 \dots N, m = 1 \dots M\}$ containing M sequences belonging to the same family, which are aligned to be as similar as possible. Each entry a_i^m of the matrix is either one of the 20 natural amino acids, or the alignment gap $'-'$ which is employed to encode the insertion or deletion of amino acids. Hereafter we consider the gap as a 21st amino acid and we represent amino acids by numbers, i.e. $a_i^m = \{1, \dots, 21\}$. The accuracy of MSA is of critical importance to perform statistical analysis. A large number of tools have been developed by bioinformaticians [9], to align thousands of sequences and produce high-quality alignments and in a reasonable time.

The bottom part of Figure 1.5 shows the so-called *sequence logo*. It provides a graphical representation of the conservation of amino

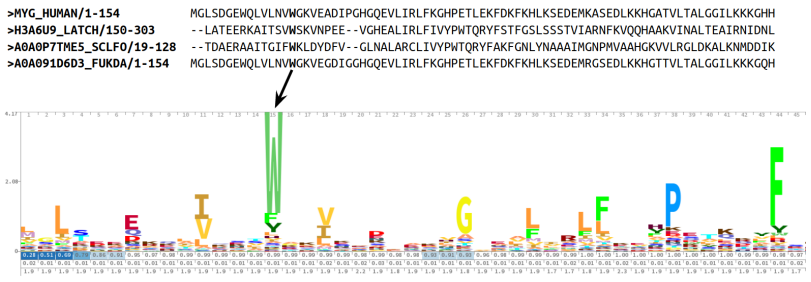

Figure 1.5: Part of the MSA of homologous proteins of the human myoglobin protein, an iron-and oxygen-binding protein. The sequence logo provides a rich description of the MSA. For example position 15 shows high conservation of the amino acid tryptophan (W).

acids. In general, functional or structural important residues are highly conserved within a protein family. At each position the relative sizes of the letters indicate their frequency in the MSA sequences, Eq. (1.3). The total height of the letters depicts the information content of the position, in bits:

$$I_i = \log_2(21) + \sum_{a=1}^{21} f_i(a) \log_2 f_i(a) \tag{1.1}$$

where $f_i(a)$ is the relative frequency of the amino acid a at position i. It is maximal ($I_i = \log_2(21)$) for a totally conserved site, i.e. $f_i(a) \in \{0, 1\} \quad \forall a \in \{1, \ldots, 21\}$.

In the next subsections, we briefly describe Profile Hidden Markov Model (HMM) which is one the most successful tool allowing the detection of homologous proteins and the generation of high-quality MSA.

1.3.1 *Profile Hidden Markov Models*

Profile Hidden Markov Models are probabilistic models underlying bioinformatical programs such as HMMER [10] and HHBlits [11]. *mainly adapted from [9]*

HMMs are typically trained on a small, high-quality and usually manually curated alignment; the so-called *seed* alignment. It consists of $\lesssim 200$ sequences which are with high confidence members of the protein family one aims to model (*cf.* Figure 1.6). The idea behind the HMM is that the visible "symbols" composing a sequence (amino acids or a gaps) are conditioned by internal factors - the hidden "states" - which are not directly observable. The hidden states form a Markov chain, i.e. the transition probability from a state s_i to the next state s_{i+1} is conditionally independent of the states $s_1, \ldots, s_{i-2}$ given s_{i-1}. Once a hidden state is reached, a symbol can be produced with an emission probability. The transitions probabilities between hidden states, as well as the emission probabilities are estimated from the frequencies in the seed alignment.

Three different hidden states are possible: the "match" state (yellow boxes in Figure 1.6), emitting an amino acid symbol with position-dependent probabilities; "delete" state (red circles in Figure 1.6) which represents skipping the position (a gap symbol is emitted instead) and the "insert" states (blue diamonds in Figure 1.6) allowing for addition of excess residues.

Note that profile HMMs assume that the residue in a particular position is independent of the residues in all other positions thus neglecting any higher-order correlations. It is common to use regularization to avoid overfitting due to the limited size of the seed alignments, and to give a high cost for opening a gap and a smaller one to extend it [9].

To align a new sequence to a profile HMM corresponds to finding the most likely sequence of hidden states. A dynamic programming algorithm, the Viterbi algorithm [9], allows to get the most probable path efficiently.

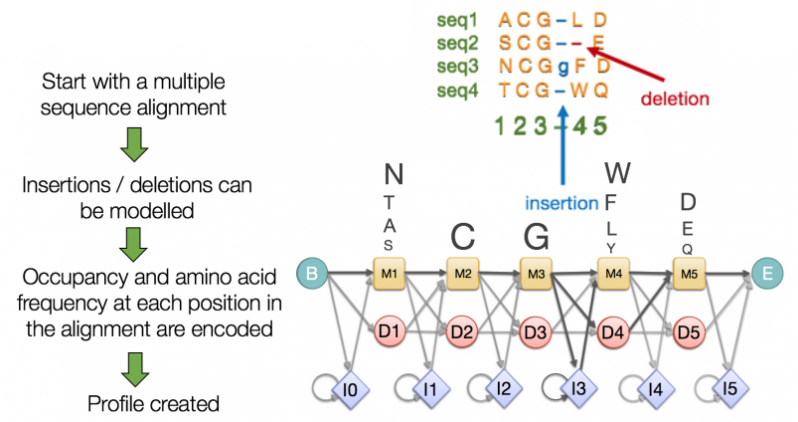

Figure 1.6: A profile HMM modelling a multiple sequence alignment. The boxes in yellow are the match states (M). The diamonds are insert states (I) which are used to model highly variable regions in the alignment. The circular states are delete states (D). They make it possible to jump over one or more columns in the alignment. The emission and transitions probabilities are derived from the observed occupancy of each position in the seed alignment. Source: [12].

HMMER [10] [13] is a software package for sequence analysis. It Includes:

- hmmbuild - build profile HMM from input multiple alignment
- hmmalign - align sequences to HMM and output the resulting MSA
- hmmsearch - search profile against sequence database
- hmmscan - search sequence against profile database

It defines various significance thresholds on the computed scores to decide whether a sequence should be added to the MSA or not: the "log-odds score" - the log ratio between the query sequence path probability and the probability of the sequence in a null model (obtained using only the background amino acid frequencies) - and the "E-value" - the expected number of random sequences achieving an equal or higher log-odds ratio than the query sequence.

HMMER can also build a profile from a single query sequence:

- phmmer - search sequence against a protein database. The profile HMM model is build internally from the single query sequence, using a simple position-independent scoring system (BLOSUM62 [14] scores converted to probabilities, plus a gap-open and gapextend probability).
- jackhmmer - iteratively search sequence against a protein database. The first iteration is identical to a phmmer search. For the next iterations, a multiple alignment of the query together with all target sequences satisfying inclusion thresholds is assembled, a profile HMM is constructed from this alignment.

Iteratively searching through large sequence databases such as UniProt can be time demanding. It requires to compare each sequence to the profile HMM. An alternative is HHblits [11] which implements a profile-profile comparison. It starts from the clustering of UniProtKB/TrEMBL into groups of similar sequences alignable over at least 80% of their length and down to 30% pairwise sequence identity. For each cluster, an HMM is created. Given a query sequence, HHblits first converts it to an HMM. Then it searches the query profile against the HMM database, i.e. against sequence clusters instead of individual sequences. New sequences from the clusters, below a defined E-value threshold, are added to the query MSA. Due to the pre-clustering of the UniProt database, HHblists is much faster than Jackhmmer while leading to virtually identical MSAs.

1.3.2 *Protein domains and the Pfam database*

HMMER is the core utility of the protein family databases Pfam [15] which contains annotations and MSA of *proteins domains. A domain* is a part of the protein that can evolve, function, and exist almost independently of the rest of the polypeptide chain. Many proteins consist of several domains, and evolution uses domains as 'modules' that may be recombined in different arrangements to create multi-domains proteins which differ in function and structure. The Pfam 32.0 version (September 2018) contains 17929 protein domain families. Each family is defined by a high-quality seed alignment from which is built the profile HMM. In Pfam, the profile HMM is then searched

against an extensive sequence collection, based on reference proteomes extracted from UniProt, to produce the full alignment.

1.4 EPISTASIS AND COEVOLUTION

Profile-HMMs are some of the most successful tools for searching and aligning homologous sequences. However, they treat each protein residue independently: being based on single-site conservation patterns, they assume that a residue in a particular position is independent of the rest of the chain. Models of this type are intrinsically unable to take into account the fact that when two or more simultaneous mutations are present, the phenotypic effect can be different from a function of the simple sum of individual amino acid changes. The non-additivity of single mutation effects is called *epistasis.*

Epistasis between residues causes residues to constrain each other's evolution or, in other terms, to *coevolve.* For instance, the deleterious effect of a mutation in one site can be reverted by a second-site interacting residue. As a result of that, correlated mutations will be observed between the corresponding columns of the MSA.

While a comprehensive explanation for epistasis is still lacking, recent studies [16, 17] have found that epistatic pairs of residues tend to be close in 3D structure. Studies such as [16, 17], made possible thanks to recent advances in high-throughput sequencing technologies[3], can be seen as a proof-of-concept of the long-standing idea [19, 20] of using correlated amino acid substitutions in the MSA to predict structural contacts in a protein. We will describe it in the next section.

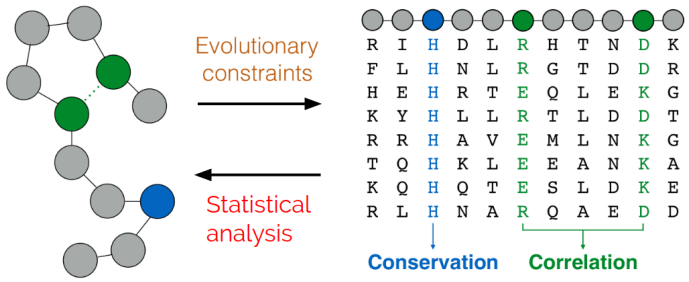

Figure 1.7: The conservation of the same structure or function within members of a proteins family imposes constraints on the evolution of a protein sequence, which, in turn, lead to statistical patterns in an MSA. Functionally or structurally important residues are usually higly conserved (blue sites). Correlated mutations between columns of the MSA (green sites) are signatures of epistasis and, often, of residue-residue contacts. Therefore, one can use the correlation patterns of the MSA to identify contacts in the 3d structure.

3 Notably *deep mutational scanning* [18] which can assess the phenotypic effects of thousands of mutations simultaneously.

1.4.1 *Protein-structure prediction using residue coevolution*

The protein structure prediction problem asks the following question: given a protein sequence - or the corresponding MSA of homologous sequences - can we predict structural contacts in the 3d structure? More than 20 years ago [19, 20], it was suggested that correlated mutations between columns of the MSA could be used to predict contacts, see Figure 1.7. The first applications of this idea used the Mutual Information (MI), or related pairwise correlation measures, between columns i and j of the MSA to identify co-evolving pairs of residues [21–23]:

$$MI_{ij} = \sum_{q_1,q_2=1}^{21} f_{ij}(q_1,q_2) \log \frac{f_{ij}(q_1,q_2)}{f_i(q_1)f_j(q_2)} \tag{1.2}$$

where $f_{ij}(q_1,q_2)$ and $f_i(q_1)$ are respectively the empirical one- and two-point frequencies in the MSA:

$$f_i(a) = \frac{1}{M}\sum_{m=1}^{M} \delta_{a,a_i^m} \qquad f_{ij}(a,b) = \frac{1}{M}\sum_{m=1}^{M} \delta_{a,a_i^m}\delta_{b,a_j^m} \tag{1.3}$$

with M being the number of sequences in the alignment, and $\delta_{a,b}$ the Kronecker symbol, which equals one if amino acids a and b are equal.

But this approach has had only limited success. Indeed, the true covariation signal in the MSA is often masked by ancillary indirect-correlations: residue i might be correlated to a residue j without being in contact with it, because both are in contact with a third residue k. As MI looks at pairs of columns independently - or, in other words, is intrinsically *local* - it can not disentangle direct from indirect correlations.

To overcome this problem, the *global* statistical model DCA [24, 25] was introduced in 2009. We describe it in the following chapter.

1.4.2 *Coevolution in a more general sense*

Finally, note that coevolution does not happen only for residues within proteins, but it is *ubiquitous* in nature. It occurs when two or more related biological systems do not evolve independently. This means that one exerts selective pressures on the other, thereby affecting each other's evolution. The initial ideas on the mutual influence between species can be traced back to Darwin's 1862 [26] work on orchids and pollinators. At the molecular level, coevolutionary signatures between genes/proteins is a consequence of physical interactions and/or functional relationships. In Chapter 3 we will combine different coevolutionary scales, from correlated presence-absence patterns of proteins across species, up to correlations in the amino-acid usage, to predict currently unknown, but biologically sensible interactions.

2

STATISTICAL ANALYSIS OF PROTEIN SEQUENCE DATA

Recent technological advances in high-throughput sequencing have generated vast amounts of genetic data. Consequently, a large variety of statistical tools have been recently developed to extract information on biological systems from these growing datasets. It became rapidly manifest that so-called *inverse problems* from statistical physics offered a powerful tool. In inverse problems, the usual procedure of statistical physics is reversed: instead of calculating a set of observables from an underlying model, one aims to infer the parameters of a model from a set of observations. Models of this type have found applications in different biological contexts, including the reconstruction of neural networks [27–29], the prediction of contacts in protein structures [24, 25, 30], or the movement of flocks of birds [31, 32]. For a complete review of inverse problems in statistical physics and applications to biological systems we refer to [33].

In the case of protein sequences, the starting point for the statistical modeling is a multiple sequence alignment $A = \{a_i^m | i = 1 \dots N, m = 1 \dots M\}$ containing M aligned sequences of length N belonging to the same protein family, *cf.* Section 1.3. Statistical patterns in the MSA are signatures that mutations are not randomly selected. The conservation of structure and function across protein families induces constraints on the sequence variability. The aim of the Direct Coupling Analysis (DCA) is to construct a probabilistic model to seize the sequence variability in the MSA and to relate it to the biological structure and function of the protein family.

To this end, one has first to determine the functional form of the probability distribution, and second to infer the parameters of the distribution.

This chapter is organized as follows. Section 2.1 presents the maximum entropy framework, which is typically used to justify the use of Boltzmann distributions, together with some significant shortcomings of this approach. Next, we focus on a specific class of probabilistic distributions known as Potts models $P(\mathbf{a}|\mathbf{J},\mathbf{h}) \propto \exp\Big(-\sum h_i(a_i) - \sum J_{ij}(a_i,a_j)\Big)$, where $\mathbf{a} = (a_1,\dots,a_N)$ is full-length protein sequence, which is used by DCA to model the sequence variability of an MSA . Section 2.2 contains a concise overview of the inference methods of the DCA parameters $\mathbf{J},\mathbf{h}$. Section 2.4 presents some results obtained on proteins using DCA methods.

2.1 MAXIMUM ENTROPY MODELLING

The Maximum-Entropy Principle (MaxEnt), introduced by Jaynes in [34], can be seen as a principled way to obtain functional forms of probability distributions $P(\mathbf{a})$ in inference problems.

It starts from a set of arbitrarily chosen observables $\{\mathcal{O}^\mu\}_{\mu=1\dots p} : \mathcal{A}^N \to \mathbb{R}$, with $\mathcal{A}^N$ being the space of all possible sequences of length N, which assign a real value to any amino-acid sequence. Next, we require the MaxEnt model $P(\mathbf{a})$ to reproduce the empirical mean of each observable in the MSA:

$$\forall \mu = 1 \dots p : \qquad \sum_{\mathbf{a}\in\mathcal{A}^N} P(\mathbf{a})\mathcal{O}^\mu(\mathbf{a}) = \frac{1}{M}\sum_{\mathbf{a}\in A} O^\mu(\mathbf{a}), \tag{2.1}$$

with the last sum running over all sequences of the MSA A. Besides this consistency requirement, the MaxEnt principle states that, in absence of any good reason to do otherwise, the model should be the least constrained one. Its Shannon entropy has therefore to be maximized:

$$S = -\sum_{\mathbf{a}\in\mathcal{A}^N} P(\mathbf{a}) \log P(\mathbf{a}) \to \max. \tag{2.2}$$

To solve this problem, we introduce a set of Lagrange multipliers $\{\lambda_\mu\}, \mu = 1 \dots p$, one for each observable plus another one $\{\omega\}$ to assure the normalization of probability distribution, and we maximize the following Lagrange functional with respect to $P(\mathbf{a})$:

$$\begin{aligned}\mathcal{F} = &-\sum_{\mathbf{a}\in\mathcal{A}^N} P(\mathbf{a}) \log P(\mathbf{a}) + \omega\Big(\sum_{\mathbf{a}\in\mathcal{A}^N} P(\mathbf{a}) - 1\Big) + \\ &+ \sum_{\mu=1}^{p} \lambda_\mu \Big(\sum_{\mathbf{a}\in\mathcal{A}^N} P(\mathbf{a})\mathcal{O}^\mu(\mathbf{a}) - \frac{1}{M}\sum_{\mathbf{a}\in A} O^\mu(\mathbf{a})\Big).\end{aligned} \tag{2.3}$$

The solution is a Boltzmann-like distribution:

$$P(\mathbf{a}|\{\lambda_\mu\}) = \arg\max_P(F) = \frac{1}{Z(\{\lambda_\mu\})} \exp\Big(\sum_{\mu=1}^{p} \lambda^\mu \mathcal{O}^\mu(\mathbf{a})\Big) \tag{2.4}$$

where the partition function $Z(\{\lambda_\mu\}) = \exp(1-\omega)$ guarantees the normalization of $P(\mathbf{a})$.

Note that the MaxEnt principle provides a way to determine the functional form of the probability distribution, but it does not give any rule how to choose the set of relevant observables Eq. (2.1) or even if such a set exists for a specific problem under consideration. This choice is, in a sense, arbitrary and different choices lead to different models. Ideally one would like to select the *minimal number of observables which make the probability distribution $P(\mathbf{a})$ generative*: samples drawn from the $P(\mathbf{a})$ should be statistically indistinguishable from the data even for observables whose consistency with the data was not imposed by the inference procedure.

2.1.1 *Profile models*

Amino acids that are highly conserved in a MSA frequently identify functionally or structurally important sites in a protein. Profile models encode this infomations into the modelling by considering as set of observables $\mathcal{O}^{ia}(\mathbf{a}) = \delta_{a_i,a}$ for all positions $i = 1,...,N$ and all amino-acid letters $a \in \{1,\dots,21\}$. Their statistics in the MSA is thus characterized by the fraction of sequences having amino acid a in position i:

$$f_i(a) = \frac{1}{M}\sum_{m=1}^{M}\delta_{a,a_i^m} \qquad \delta_{a,a_i^m} = \begin{cases} 1, & \text{if } a = a_i^m \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

The constrains on $P(\mathbf{a})$ becomes:

$$\forall i = 1\dots N \quad P_i(a_i) = \sum_{\{A_k|k\neq i\}} P(a_1,\dots,a_N) = f_i(a_i) \tag{2.6}$$

so that the marginal single-site distributions of $P(\mathbf{a})$ coincide with the empirical frequencies of the data $f_i(a_i)$. The result of MaxEnt maximization is:

$$P(\mathbf{a}|\mathbf{h}) = \frac{1}{Z(\mathbf{h})}\exp\Big(\sum_{i=1}^{N} h_i(a_i)\Big) = \prod_{i=1}^{N}\frac{\exp(h_i(a_i))}{\sum_{a_i=1}^{q}\exp(h_i(a_i))} \tag{2.7}$$

where the local fields $h_i(a_i)$ can be easily computed by inverting Eq. (2.6), $h_i(a_i) = \log(f_i(a_i)) + const$.

Profile models, and their generalization with hidden nodes profile HMM, are among the most successful models in bioinformatics, and they represent the state-of-the-art technique for homology detection and construction of MSA, *cf.* Section 1.3.1. From the consistency Eq. (2.6) it is clear that they are able to reproduce the *conservation* statistics of an MSA, meaning to reproduce the heterogeneous usage of amino acids in the different positions of the sequence. However they are not generative models since they assume all positions to be statistically independent. They neglect that residues do not evolve independently in a protein sequence. The coevolution of residues, discussed in Section 1.4, which is visible via residue correlation, can not be captured by profile models.

2.1.2 *Direct coupling Analysis*

DCA, introduced in 2009 [24, 25], overcomes this limitation by including into the set of observables $\mathcal{O}^{ia,jb}(\mathbf{a}) = \delta_{a_i,a}\delta_{a_j,b}$ for all positions $i,j = 1,...,N$ and all amino-acid letters $a,b \in \{1,\dots,21\}$. The statistical model $P(\mathbf{a})$ is required to reproduce not only the amino-acid usage of single MSA columns, but also the fraction $f_{ij}(a,b)$ of sequences having

simultaneously amino acid a in position i, and amino acid b in position j:

$$f_i(a) = \frac{1}{M}\sum_{m=1}^{M}\delta_{a,a_i^m} \qquad f_{ij}(a,b) = \frac{1}{M}\sum_{m=1}^{M}\delta_{a,a_i^m}\delta_{b,a_j^m}. \tag{2.8}$$

The consistency with the data implies:

$$P_i(a_i) = \sum_{\{A_k|k\neq i\}} P(a_1,\ldots,a_N) = f_i(a_i) \tag{2.9}$$
$$P_{ij}(a_i,a_j) = \sum_{\{A_k|k\neq i,j\}} P(a_1,\ldots,a_N) = f_{ij}(a_i,a_j). \tag{2.10}$$

The solution of the optimization problem Eq. (2.3) is the so-called *generalized Potts model* :

$$P(a_1,\ldots,a_L|\mathbf{J},\mathbf{h}) = \frac{1}{Z(\mathbf{J},\mathbf{h})}\exp(-H(\mathbf{a})) \tag{2.11}$$

where the "energy function" or Hamiltonian H is:

$$H(a_1,...,a_N) = -\sum_{i=1}^{N} h_i(a_i) - \sum_{1\leq i<j\leq N} J_{ij}(a_i,a_j). \tag{2.12}$$

Note that in this formulation each coupling J_{ij} is a $q\times q$ matrix whose entries are the couplings between a pair of Potts states a,b. They are symmetric $J_{ij}(a,b) = J_{ji}(b,a)$ since correlations are symmetric $f_{ij}(a,b) = f_{ji}(b,a)$.

The DCA model Eq. (2.11) has been initially proposed in the context of structural biology [24, 25]. It has been long recognized that coevolutionary information contained in a protein family allows extracting structural information [19, 20]. However, the true covariation signal in the MSA is often masked by ancillary indirect-couplings: residue i might be correlated to a residue j without being in contact with it because both are in contact with a third residue k. Therefore, any *local* correlation measure (like the MI, cf. Section 1.4.1) which looks at pairs of columns independently from the other columns can not disentangle direct from indirect interactions. On the contray, DCA relies on a *global* model: the probability distribution $P(\mathbf{a}|\mathbf{J},\mathbf{h})$, Eq. (2.11), depends on the full sequence $\mathbf{a}$ and cannot be factorized over columns or column pairs of the MSA. The pairwise couplings $J_{ij}(a,b)$, introduced to capture epistasis between residues, have been observed to predict residue-residue contacts accurately [24, 25].

Note that we still need to solve the inference problem, i.e. to tune the values of the Lagrange multipliers, the local fields $\{h_i(a)\}_{a=1,\ldots,q}$ with $i = 1,...,N$ and the pairwise couplings $\{J_{ij}(a,b)\}_{a,b=1,\ldots,q}$ with $i,j = 1,...,N$, to satisfy the respective consistency Eqs. (2.9,2.10). Many

approximation methods are available to tackle this problem, and we present three of them in the next section.

The consistency Eqs. (2.9,2.10) have two important consequences. First, a precisely inferred DCA model reproduces *conservation* and *covariation* statistics of an MSA. Second, only the single and pairwise statistics of the MSA are used for the inference. There is a priori no reason for which the model should be able to reproduce any higher-order statistics. Astonishingly, Figliuzzi and collaborators [35] demonstrated that a precisely inferred model captures even statistical measures which were not imposed by the inference procedure, like three-residue correlations, the clustered structure of protein families in sequence space or the Hamming distance between sequences, as shown in Figure 2.1. These findings suggest that the pairwise statistics are sufficient to accurately capture the residue variability in the MSA, or, in other terms, that the DCA model $P(\mathbf{a}|\mathbf{J},\mathbf{h})$ is a generative model.

2.1.3 *Criticism to MaxEnt*

Criticism, however, has been leveled against the MaxEnt from several angles [36, 37]. First, MaxEnt modeling requires to select a set of relevant observable which compress the information contained in the whole multiple sequence alignment. Therefore, Eq. (2.4) is not the most unbiased representation of the protein family, but it is the most unbiased for the set of observable that one has arbitrarily chosen. In the case of DCA, the choice of f_i and f_{ij} is motivated by the fact that they are easy to interpret in terms of conservation and coevolution. Furthermore, the typical size of MSAs, usually hundreds or thousands of sequences, often does not allow one to accurately compute higher order moments of the data like the third-order frequencies f_{ijk}.

Second, MaxEnt models assume that the process sampled is at equilibrium, which is certainly questionable for biological systems. Third, as explained in the next section, the best current method for prediction of residue contacts is pseudolikelihood, which requires as input the full sequences of the MSA, not only frequencies and correlations, in contrast with the MaxEnt principle.

In [38] Gao and collaborators suggested that the reason behind the success of DCA is to be searched in population genetics. They applied DCA to whole-genome population-wide sequencing data and showed that DCA yields meaningful results only when a population evolves with a sufficient amount of exchange of genetic material. In this case, the population reaches a dynamic equilibrium, the so-called quasi-linkage equilibrium (QLE), where the distribution of genotypes assumes the form of a Potts model Eq. (2.11). On the contrary, for population evolving with low recombination rate DCA can not be expected to work. Note that in [38] they analysed recently diverged sequences (about 3,000 genomes of the human pathogen *Streptococcus*

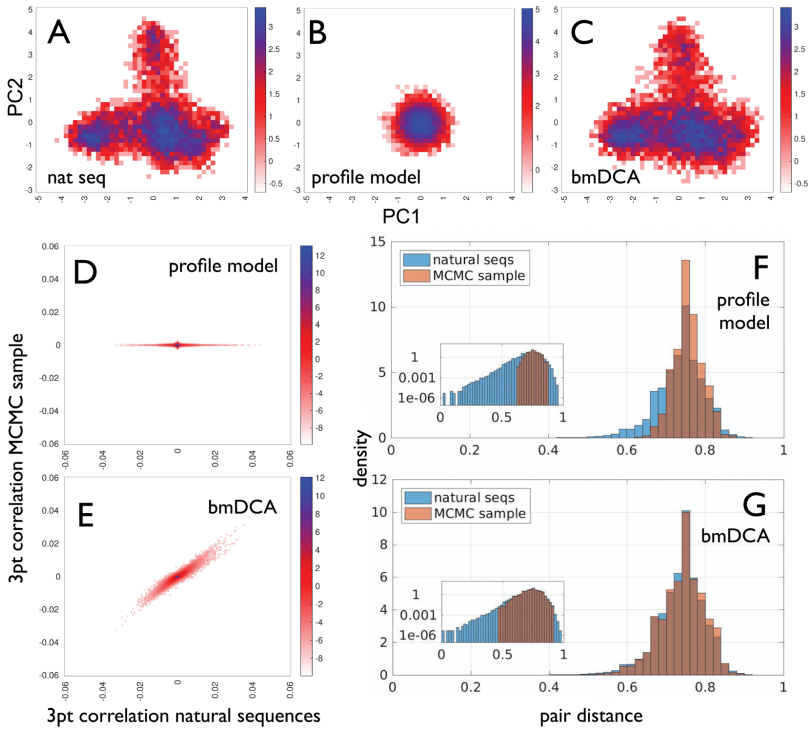

Figure 2.1: *Non-fitted statistical observables are captured by* DCA: The projection on the first two principal components of sequences belonging to the MSA PF00072 (Panel A) or generated by Monte Carlo sampling from the profile model (Panel B) and the DCA (Panel C). Three-point correlations of Monte Carlo samples of the profile (D) and DCA (E) models, as compared to the three-points correlations in the natural sequences. Histograms of all pairwise Hamming distances between natural or Monte Carlo sampled sequences, for profile (F) and DCA (G) models. All model DCA were inferred with the highly precise BML method, described in Section 2.2.1. These findings suggest that accurately inferred DCA model can capture the residue variability of the MSA. Source: [35].

pneumoniae, *cf.* Section 2.4.5) therefore subject to evolutionary pressure different from that of Pfam sequences. However, these findings suggest that population genetics can provide a more rational basis for DCA, at least for genome-scale analysis.

Finally, note that MaxEnt is not needed if we assume priori that $P(\mathbf{a}|\mathbf{J},\mathbf{h})$ takes the form of a Potts model Eq. (2.11). As we will discuss in the next section, the consistency Eqs. (2.9,2.10) directly follow from the vanishing first derivatives of the log-likelihood Eq. (2.15) (see next section).

2.2 INFERENCE METHODS FOR THE INVERSE POTTS PROBLEM

The MaxEnt principle proposes an analytical form of the probability distribution $P(\mathbf{a}|\mathbf{J},\mathbf{h})$; still, the values of the Lagrange multipliers need to be tuned to satisfy the respective consistency Eqs. (2.9,2.10). Given an MSA A, a commonly used strategy, using a Bayesian framework, is to maximize the posterior:

$$P(\mathbf{J},\mathbf{h}|A) \propto P(A|\mathbf{J},\mathbf{h}) \cdot P(\mathbf{J},\mathbf{h}) \tag{2.13}$$

where the $P(A|\mathbf{J},\mathbf{h})$ is the *likelihood* and $P(\mathbf{J},\mathbf{h})$ is the *prior distribution* which allows to encode prior beliefs about the parameters of the model, before the data are observed. We will discuss prior and regularization in Section 2.3.2.

Assuming, for the moment, uniform prior distribution, maximizing the posterior is equivalent to maximize the likelihood $\mathcal{L}$. In practice, it is often more convenient to consider the average log-likelihood:

$$\frac{1}{M}\log\mathcal{L} = \frac{1}{M}\log P(A|\mathbf{J},\mathbf{h}) \tag{2.14}$$

with M being the number of sequences in the MSA. If we assume that the sequences of the MSA are independent and identically distributed, we obtain:

$$\begin{aligned}\frac{1}{M}\log\mathcal{L} = \frac{1}{M}\log\prod_{i=1}^{M} P(\mathbf{a}^m|\mathbf{J},\mathbf{h}) = \frac{1}{M}\sum_{i=1}^{M}\log P(\mathbf{a}^m|\mathbf{J},\mathbf{h}) = \\ \sum_{i=1}^{L}\sum_{a=1}^{q} h_i(a) f_i(a) + \sum_{1\le i<j\le N}\sum_{a,b=1}^{q} J_{ij}(a,b) f_{ij}(a,b) - \log Z(\mathbf{h},\mathbf{J}).\end{aligned} \tag{2.15}$$

One can readily see that the parameters $\mathbf{J},\mathbf{h}$ maximizing Eq. (2.15) are solutions of the consistency Eqs. (2.9,2.10).

By computing the Hessian, it can be easily checked that the log-likelihood Eq. (2.15) is a concave function in the parameters $\mathbf{h}$,$\mathbf{J}$[33]. This means that a simple optimization scheme such as gradient ascent

is guaranteed to find its maximum value. However, the exact computation of the log-likelihood and of its gradient is computationally unfeasible due to the last term containing the partition function:

$$\begin{aligned} Z(\mathbf{J},\mathbf{h}) = & \sum_{\mathbf{a}} \exp\Big(-H(\mathbf{a})\Big) \\ = & \sum_{\mathbf{a}} \exp\Big(\sum_{i=1}^{N} h_i(a_i) + \sum_{1\le i<j\le N} J_{ij}(a_i,a_j)\Big). \end{aligned} \tag{2.16}$$

The sum contains q^N terms, with $q = 21$. This means that the direct calculation of Z is impossible even for small proteins. As a reference, the WW domain (PF00397), one of the shortest in the Pfam database, contains 31 residues, and the partition function contains a sum with $\sim 10^{40}$ terms.

Approximation methods are required and three of them will be described below:

1. Boltzmann Machine Learning (BML). It is able to achieve arbitrarily accurate DCA models, but requires computationally expensive Monte Carlo simulations.

2. The Mean-Field (MF) approximation. It is the computationally most efficient approximative inference scheme but provides only a rough estimate of the Potts parameters.

3. The Pseudo-Likelihood Maximization (PLM). Computationally less demanding than BML while leading to virtually identical perfomance for unsupervised contact prediction, but it fails in accurately reproducing the empirical statistics.

2.2.1 *Boltzmann Machine learning*

Introduced in [39], the Boltzmann Machine Learning is the most straightforward method to tackle the inverse problem. Starting from an initial guess of fields and couplings, the model one- and two-point marginals $P_i(a), P_{ij}(a,b)$ are estimated through Monte Carlo simulations $f_i^{MC}(a), f_{ij}^{MC}(a,b)$, thus avoiding the exact computation of the partition function Z. The parameter values $\mathbf{h}, \mathbf{J}$ are consequently updated following the gradient of the log-likelihood:

$$h_i(a) \to h_i(a) + \eta_i(a)\Big(f_i^{MC}(a) - f_i(a)\Big) \tag{2.17}$$

$$J_{ij}(a,b) \to J_{ij}(a,b) + \eta_{ij}(a,b)\Big(f_{ij}^{MC}(a,b) - f_{ij}(a,b)\Big) \tag{2.18}$$

where $\{\eta_i, \eta_{ij}\}$ are small parameters (typical values for DCA, are $\mathcal{O}(10^{-2})$ or $\mathcal{O}(10^{-3})$), the so-called *learning rates.* Due to the convexity of the log-likelihood, in the case of sufficiently precise Monte Carlo

and small enough learning rate, the procedure is guaranteed to converge with arbitrary precision to the exact solution. One can achieve estimations of $\mathbf{h}$ and $\mathbf{J}$ which satisfy with high accuracy Eqs. (2.9,2.10), as shown in Figure 2.2. However, each step of the learning process requires an accurate Monte Carlo estimation of $f_i^{MC}(a), f_{ij}^{MC}(a,b)$, which can be computationally demanding for large systems. Even with efficient implementation going beyond simple gradient descent [40–42] the BML is applicable only to protein families smaller than about $L = 200$. Large-scale studies for hundreds or thousands of protein families are therefore out of reach.

2.2.2 *Mean Field*

First introduced in [25] for protein sequence, the approximation it is essentially an high-temperature expansion of the Legendre transform $\mathcal{G}$ of the free energy, i.e. the logarithm of the partition function Eq. (2.16):

$$\mathcal{G} := -\log Z + \sum_{i=1}^{N}\sum_{a=1}^{q-1} P_i(a)h_i(a) \tag{2.19}$$

where the sum over a of variable i runs only up to $q-1$ because of the lattice-gas gauge choice, see Section 2.3.1.

The above functional can be computed through a small-coupling expansion, i.e., a Taylor expansion around zero coupling. Then, from linear response theory, the following equations hold:

$$\begin{aligned} h_i(a) &= \frac{\partial \mathcal{G}}{\partial P_i(a)} \\ (C^{-1})_{ij}(a,b) &= \frac{\partial h_i(a)}{\partial P_j(b)} \end{aligned} \tag{2.20}$$

where $\mathbf{C}$ is the connected correlation matrix:

$$C_{ij}(a,b) = P_{ij}(a,b) - P_i(a)P_j(b). \tag{2.21}$$

The resulting MF equations are:

$$P_i(a) = \frac{1}{z_i}\exp\left(h_i(a) + \sum_{j\neq i}\sum_{b=1}^{q-1} J_{ij}(a,b)P_j(b)\right) \tag{2.22}$$

and

$$(C^{-1})_{ij}(a,b) = \begin{cases} -J_{ij}(a_i,a_j) & \text{for } i \neq j \\ \frac{\delta_{a,b}}{P_i(a)} + \frac{1}{P_i(q)} & \text{for } i = j. \end{cases} \tag{2.23}$$

The inference problem is then solved, by plugging into Eq. (2.23) the empirical version of the connected correlation matrix $C_{ij}^{emp}(a,b) =$

$f_{ij}(a,b) - f_i(a)f_j(b)$ computed directly from the MSA. The MF approximation is computationally very efficient since it just requires to invert the empirical connected-correlation matrix.

Note that the connected correlation matrix $\mathbf{C}$ is surely rank deficient as it displays N zeros nodes:

$$\sum_b C_{ij}(a,b) = \sum_b f_{ij}(a,b) - f_i(a) = 0 \tag{2.24}$$

and therefore it is not invertible. This is due to the over-parametrization of the model, which can be solved by fixing the *lattice-gas gauge* (see Section 2.3.1).

Unfortunately, in most cases the finite size of the MSA still makes $\mathbf{C}$ rank deficient, as some states may never be observed in finite-size samples. One common solution is to add a pseudocount to the frequencies (see Section 2.3.2).

The MF approximation was the first efficient successful method for contact prediction [25]. However, MF only provides a rough estimate of Potts parameters and the model usually does not satisfy the constraints Eqs. (2.9,2.10).

2.2.3 *Pseudolikelihood Maximization*

The Pseudolikelihood Maximization, first introduced in [43] and later [44] in different contexts, was extended to Direct Coupling Analysis in [45, 46]. Essentialy, the idea behind PLM is to avoid the computation of the partition function, by replacing the $P(\mathbf{a}^m)$ with the conditional probability of observing a variable a_r given the rest of the sequence $\mathbf{a}_{\setminus r} = (a_1, \dots, a_{r-1}, a_{r+1}, \dots, a_N)$:

$$P(a_r|\mathbf{a}_{\setminus r}) = \frac{\exp\left(h_r(a_r) + \sum_{j\neq r} J_{rj}(a_r, a_j)\right)}{\sum_{b=1}^{q} \exp\left(h_r(b) + \sum_{j\neq r} J_{rj}(b, a_j)\right)}. \tag{2.25}$$

The log-likelihood Eq. (2.15), is then replaced by a sum of site-dependent terms called *pseudo-likelihood*:

$$\frac{1}{M}\sum_{i=1}^{M} \log P(\mathbf{a}^m) \mapsto \sum_{r=1}^{N} \mathcal{L}_r(h_r, \mathbf{J}_r) = \sum_{r=1}^{N} \left(\frac{1}{M}\sum_{m=1}^{M} \log P(a_r^m|\mathbf{a}_{\setminus r}^m)\right) \tag{2.26}$$

The inference problem can be solved by maximizing Eq. (2.26). This approach, called *symmetric pseudo-likelihood maximization*, is slower than the *asymmetric pseudo-likelihood maximization* while leading to a virtually identical performance [45, 46]. The latter consists in maximizing independently each $\mathcal{L}_r(h_r, \mathbf{J}_r)$ since each term depends on a different set of parameters h_r and $\mathbf{J}_r = \{J_{rj}\}_{r\neq j}$. This method returns two different values for the same coupling $J_{ij}(a,b)$ respectively from the maximization of $\mathcal{L}_i(h_i, \mathbf{J}_i)$ and $\mathcal{L}_j(h_j, \mathbf{J}_j)$. One simple way to overcome this shortcoming [45, 46] is to replace them by their average.

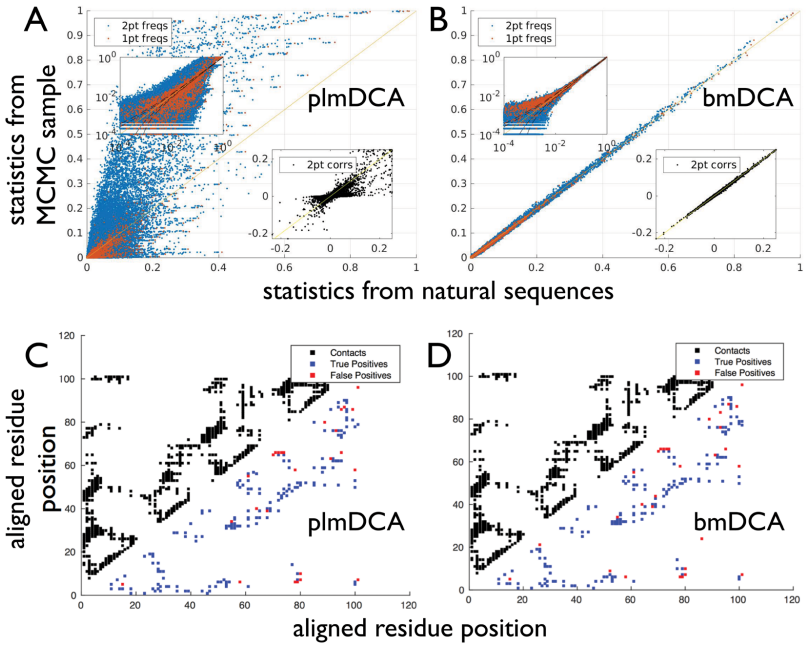

Figure 2.2: *Fitting accuracy and contact prediction for* PLM *and* BML *in the PF00072 protein family*: the DCA model inferred by PLM (panel A) fails to reproduce the one- and two-residue frequencies (orange / blue) and the connected two-point correlations (black). On the contrary the BML (panel B) is very accurate. Despite these differences, the contact predictions using the strongest $2L$ DCA couplings (with $|i-j|>4$) are close to identical. Source: [35].

Note that PLM is somehow in contrast with MaxEnt. Indeed, to compute the $\mathcal{L}_r(h_r, \mathbf{J}_r)$ the full sequences $\mathbf{a}$ are needed while, according to MaxEnt, the average value of the chosen observables - the one- and two-point statistics for DCA - are the only information required for the inference.

As shown in Figure 2.2 the distribution inferred by PLM typically do not satisfy the constrains Eqs. (2.9,2.10). Despite this, the contact prediction accuracies of PLM and BML are close to identical.

2.3 TECHNICAL POINTS

2.3.1 *Gauge Transformation*

The number of parameters of the DCA model Eq. (2.11) is $N_p = Nq + \frac{N(N-1)}{2}q^2$ but the constraints Eqs. (2.9,2.10) are not independent of each other, in particular we have the linear dependencies:

$$\sum_{a=1}^{q} f_i(a) = 1 \quad \text{and} \quad \sum_{a=1}^{q} f_{ij}(a,b) = f_j(b) \tag{2.27}$$

Therefore, the number of independent observables is $N(q-1) + \frac{N(N-1)}{2}(q-1)^2$. This means that the model is over-parametrized: there

are more parameters then observables to fit. This condition leads to a so-called *gauge invariance*: the probability distribution Eq. (2.11) is invariant under a set of transformations of the Potts parameters $\mathbf{h}, \mathbf{J}$. Indeed, for any arbitrary function $K_{ij}(a)$ and for arbitrary constants c_i and c_{ij}, one can easily verify that the following transformation leave the probabilities of the model unchanged:

$$J_{ij}(a,b) \to J_{ij}(a,b) + K_{ij}(a) + K_{ij}(b) + c_{ij} \tag{2.28}$$

$$h_i(a) \to h_i(a) + c_i - \sum_{j=1(j\neq i)}^{N} K_{ij}(a) \tag{2.29}$$

In DCA, especially for contact prediction, a common choice is the so-called *zero-sum gauge* (also known as *Ising gauge*):

$$\sum_{b=1}^{q} J_{ij}(a,b) = \sum_{a=1}^{q} J_{ij}(a,b) = \sum_{a=1}^{q} h_i(a) = 0. \tag{2.30}$$

It is obtained by the transformation:

$$J_{ij}(a,b) \to J_{ij}(a,b) - J_{ij}(\cdot,b) - J_{ij}(a,\cdot) + J_{ij}(\cdot,\cdot) \tag{2.31}$$

$$h_i(a) \to h_i(a) - h_i(\cdot) + \sum_{j=1(j\neq i)}^{N} \Big(J_{ij}(a,\cdot) - J_{ij}(\cdot,\cdot) \Big) \tag{2.32}$$

where $g(\cdot)$ is average of the function g over all states $g(\cdot) = \frac{1}{q}\sum_{a=1}^{q} g(a)$.

Another common gauge is the *lattice-gas gauge* which removes the trivial zero nodes of the connected correlation matrix $\mathbf{C}$:

$$J_{ij}(a,q) = J_{ij}(q,b) = h_i(q) = 0 \tag{2.33}$$

obtained by the transformation:

$$J_{ij}(a,b) \to J_{ij}(a,b) - J_{ij}(q,b) - J_{ij}(a,q) + J_{ij}(q,q) \tag{2.34}$$

$$h_i(a) \to h_i(a) - h_i(q) + \sum_{j=1(j\neq i)}^{N} \Big(J_{ij}(a,q) - J_{ij}(q,q) \Big). \tag{2.35}$$

2.3.2 *Regularization*

A Potts model describing a protein family usually contains between 50 to 500 residues. Therefore, the number of Potts parameters can range from 10^5 to 10^8. Since typical MSAs have a limited number of sequences (usually from 10^2 to 10^5) regularization is essential to avoid overfitting. We refer to [47] for a complete study of regularization

techniques. A common inference regularization for PLM or BML [25, 35] is the ℓ^2 regularization-scheme. In Eq. (2.13) it correspond to imposing a Gaussian prior in our model, yielding a penalty term in the likelihood, which forces a trade-off between optimizing the bare likelihood (or pseudo-likelihood) and absolute values of parameters being small:

$$\ell^2(J,h) = \lambda_h \sum_{i=1}^{N}\sum_{a=1}^{q} h_i(a)^2 + \lambda_J \sum_{i<j}^{N}\sum_{a,b=1}^{q} J_{ij}(a,b)^2. \tag{2.36}$$

The two parameters λ_h and λ_J define the strength of the ℓ^2-regularization (typical values for DCA are $\lambda = 10^{-2}$ or 10^{-3}).

In inference problems, another common regularization-scheme is ℓ^1

$$\ell^1(J,h) = \lambda_h \sum_{i=1}^{N}\sum_{a=1}^{q} |h_i(a)| + \lambda_J \sum_{i<j}^{N}\sum_{a,b=1}^{q} |J_{ij}(a,b)| \tag{2.37}$$

which forces weak parameters to be set to 0, thereby favoring sparse networks. It is less prevalent in the DCA context, at least for contact prediction, where we are usually not interested in the full network but only in large couplings.

Using the Dirichlet distribution as a prior on frequency counts we get another regularization-scheme referred to as *pseudocounts* [9]:

$$f_i(a) \leftarrow (1-\alpha) f_i(a) + \frac{\alpha}{q} \tag{2.38}$$

$$f_{ij}(a,b) \leftarrow (1-\alpha) f_{ij}(a,b) + \frac{\alpha}{q^2} \tag{2.39}$$

Adding pseudocounts is equivalent to adding random $M \cdot \alpha/(1-\alpha)$ sequences to the data with amino acids sampled uniformly. It is frequently used in the context of MF for contact prediction (typically with $\alpha = 0.5$) to assure invertibility of the connected-correlation matrix.

2.3.3 *Sequence re-weighting*

Sequences in MSAs do not represent independent samples of a protein family for at least two reasons. First, sequences are evolutionarily-related via phylogenetic trees, therefore they can have a complicated dependence structure. Second, there is selection bias from sequencing species of special medical or academic interest or sequencing closely related species, thus leading to uneven sampling of a protein family sequences. To reduce both bias a simple reweighting scheme was introduced in [25]. Sequences too similar are considered to carry almost the same information, and they should be down-weighted. Formally, given an alignment with M sequences and N residues, for

each sequence $\mathbf{a}^m$ we determine the number n^m of similar sequences via

$$n^m = |\left\{\mathbf{a^k}|1 \leq k \leq M, seqid(\mathbf{a}^m, \mathbf{a}^k) \geq \theta N)\right\}| \tag{2.40}$$

A weight $w^m = 1/n^m$ is therefore associated to $\mathbf{a}^m$ which reflects its importance in the MSA. Frequencies are changed accordingly:

$$f_i(a) = \frac{1}{M_{\text{eff}}} \sum_{m=1}^{M} w^m \delta_{a,a_i^m} \tag{2.41}$$

$$f_{ij}(a,b) = \frac{1}{M_{\text{eff}}} \sum_{m=1}^{M} w^m \delta_{a,a_i^m} \delta_{b,a_j^m}. \tag{2.42}$$

The total weight $M_{\text{eff}} = \sum_{i=1}^{M} w^m$ is considered the effective number of independent sequences. In [25] the authors showed that contact prediction results are consistently better than without any re-weighting and that values of θ around 0.7-0.9 lead to very similar results (the most commonly used one in DCA inferences is $\theta = 0.8$).

2.3.4 *Frobenius norm*

The statistical modeling of proteins proposed by DCA was originally developed to improve the prediction of the three-dimensional structure of a folded protein [24, 25]. To this end, one needs a score for ranking the $N(N-1)$ possible coupling matrices.

In [24, 25], the authors introduced the so-called Direct information (DI). For each site i and j they define a new probability model:

$$P_{ij}^{dir}(a,b) = \frac{1}{Z_{ij}} \exp\left(J_{ij}(a,b) + \tilde{h}_i(a) + \tilde{h}_j(b)\right) \tag{2.43}$$

where the new fields are tuned such that empirical single-frequencies $f_i(a)$ and $f_i(b)$ are recovered. Z_{ij} is the partition function for the system restricted to positions i and j where the sites are only directly coupled via a single matrix $J_{ij}(a,b)$.

The DI can be understood as the amount of MI between columns i and j, which results from DCA couplings:

$$DI_{ij} = \sum_{a,b=1}^{21} P_{ij}^{dir}(a,b) \log \frac{P_{ij}^{dir}(a,b)}{P_i(a)P_j(b)} \tag{2.44}$$

DI is indepent of the selected gauge and can be considered a real observable.

Nevertheless, increased accuracy in contact predictions [46] were obtained using the Frobenius norm of the coupling matrix:

$$F_{ij} = \sqrt{\sum_{a,b=1}^{21} J_{ij}(a,b)^2} \tag{2.45}$$

Contrarily to the DI, F_{ij} is gauge-dependent, therefore a gauge needs to be fixed before computing it. A common choice is the zero-sum gauge Eq. (2.30) since it minimizes the absolute values of the norm Eq. (2.45) thereby explaining "as much as possible" of the distribution with the fields.

2.3.5 *Average Product Correction*

The Average Product Correction (APC) correction was originally introducted in [23] as a correction for the MI. We here briefly review the derivation of APC following the sketchy explanation of [48]. The idea of the APC is that the mutual information between positions i and j is the sum of the a real mutual influence MI_{ij}^{true} and single-site background dependences B_iB_j -like phylogeny and site entropy- which one aims to minimize:

$$MI_{ij} = MI_{ij}^{true} + B_iB_j. \tag{2.46}$$

If we assume that the true average mutual informations are small, $MI_{ij}^{true} \ll B_iB_j$, we find

$$\begin{aligned} MI_{i.} &\simeq B_iB_. \\ MI_{..} &\simeq (B_i)^2 \end{aligned}$$

from which:

$$MI_{ij}^{true} = MI_{ij} - \frac{MI_{i.}MI_{.j}}{MI_{..}}. \tag{2.47}$$

Even if it was originally introduced as a correction for the MI, the previous argument is still valid for the Frobenius norm:

$$F_{ij}^{APC} = F_{ij} - \frac{F_{i.}F_{.j}}{F_{..}}. \tag{2.48}$$

This correction is common to all implementations of DCA that are used for contact predictions

2.4 APPLICATION OF DCA

DCA turned out to reach a substantial breakthrough in detecting residue-residue contacts from sequence information alone [24, 25]. This success has encouraged its application in other contexts. In the next section, we discuss some remarkable achievements obtained by DCA with real protein sequence data. For a complete review we refer to [49].

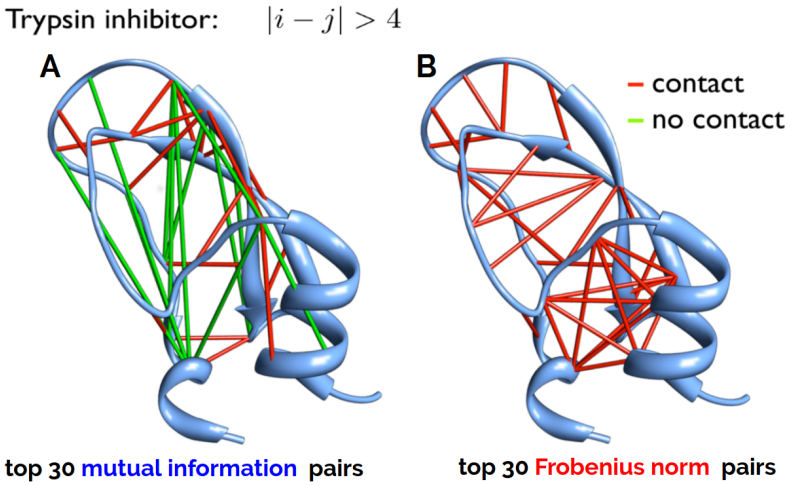

Figure 2.3: Contact predictions for the Pfam family PF00014 (trypsin inhibitor) mapped onto the x-ray crystal structure PDB 5PTI. Panel A shows the top 30 MI predicitons, and Panel B the top 30 DCA shows predicitons. Trivial predictions with sequence separation $|i-j| \leq 4$ are removed. Each pair with distance between heavy atoms$<$ 8Å in the PDB, a true positive, is connected by a red link, and the more distant pairs, false positives, are connected by green links. Source:[49]

2.4.1 *Contact prediction*

De novo protein modeling attempts to find the tertiary structure of the protein given only the amino acid sequence. Residue-residue contacts predicted from the sequence - or the corresponding MSA of homologous sequences - can be used to build three-dimensional models and consequently to predict protein folds from scratch. Here-after we define in "contact" each pair of residues with distance between heavy atoms below 8Å. The DCA contact prediction proceeds by ranking each couple of residues i and j according to the APC-corrected Frobenius norm. Sequentially close residue usually show a strong signal due to gap-streches or their proximity in the secondary structure. Those contacts are not informative about the tertiary structure. For this reason it is common to exclude all pairs $|i-j| \leq 4$, a distance corresponding to one turn in an α-helix.

DCA has been shown to significantly outperform local correlation measures, such as the MI. Figure 2.3, taken from [49], displays the top 30 predictions of the ranked MI score -directly computed from the correlations in the MSA- and DCA-score -obtained after inferring the coupling matrices with PLM approximation and computing the corresponding Frobenius norms.

The currently best unsupervised DCA-based method for contact prediction uses PLM. Note, however, that in order to obtain a good performance it is sufficient to infer the correct ranking of the pairs of

residue. This explains why all the three approximations BML, MF and PLM have a similar accuracy for contact prediction despite the fact that MF and PLM have a larger error in the inference of the exact parameter values than BML.

The PLM approximation is fast enough to allow a large-scale application. As a reference, in [50] Ovchinnikov et collaborators, applied GREMLIN [51], which is equivalent to PLM, on more than 5000 Pfam families with no structural information. By integrating sequence alignments with metagenome sequence data, they were able to successfully predicted quality structural models for 614 protein families with previously unknown structures.

Note that DCA is pure sequence based, meaning that it is completely blind with respect to any form of structure underlying the contact prediction problem.

State-of-the-art methods for contact prediction rely on additional sources beyond the MSA, such as solvent accessibility or predicted secondary structure. In particular, deep learning approaches have become increasingly popular in the last few years [52]. They can be directly trained on large set of PDBs and they are currently the most accurate methods for contact prediction (*cf.* Chapter 4). Their architecture usually consists of many layers of Convolutional Neural Network (CNN)s that can learn and collect features at different levels during training, see Figure 2.4. In Chapter 4 we explain the generic concepts behind CNNs.

2.4.2 *Protein-protein interaction*

The vast majority of proteins need to interact with others for proper biological activity. They form Protein-protein interaction (PPI) networks, and unveiling the PPI organization is one of the most formidable tasks in system biology. PPIs can be studied focusing on two major aspects, (I) identification of interacting protein families, and (II) prediction of the interaction interface within a protein complex. As we will show shortly, DCA has been used to address both points.

2.4.2.1 *Paralog matching*

In order to perform a coevolutionary analysis of PPIs, DCA requires a joint MSA, of homologous protein pairs which are supposed to interact. The generation of such alignments is a complex computational task on its own. Indeed, protein families often contain paralogs, and we generally do not know which paralog from one family interacts with a chosen paralog from the other family. The number of possible matchings can be astronomically large if multiple paralogs are present. As a reference, two MSAs with 250 species with precisely 4 paralogs per species would have $(4!)^{250} \sim 10^{345}$ possible matchings.

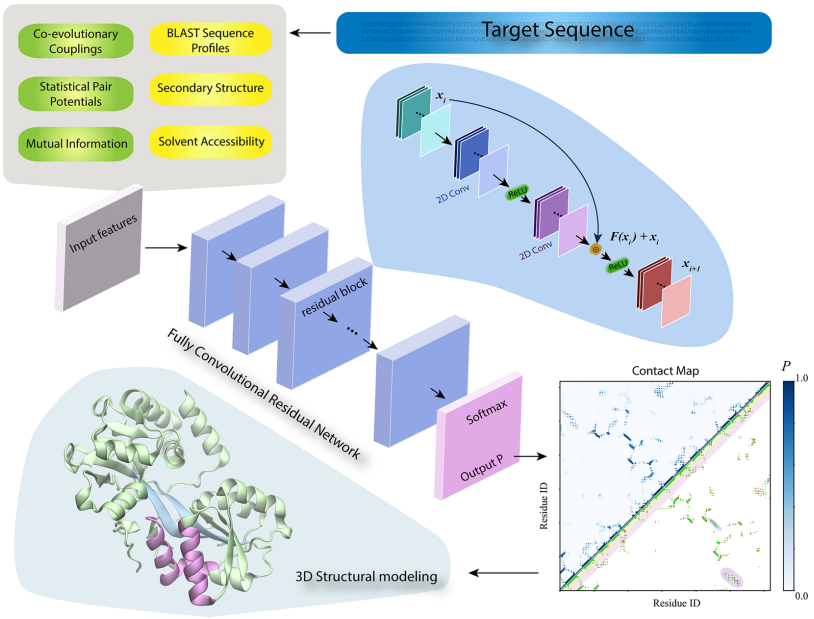

Figure 2.4: Pipeline of the deep learning CNN *DESTINI* [53]. Given an input target protein, sequence genomic and structural features are extracted. These features feed a convolutional neural network composed of multiple identical residual blocks, whose architecture is shown in the blue bubble. The final layer of the network is a softmax activation layer, which outputs the probability score for every pair of residues of the target sequence. The predicted contacts are subsequently employed to derive the 3D model of the target. As common in CNNs, it requires to be trained on a large set of PDBs (7638 in the case of DESTINI). Source: [53].

DCA can actually be used for solving the matching problem [54, 55]. In [54] two algorithms where proposed based on the idea that the correct matching of interacting paralogs should maximize the inter-domain coevolutionary signal. In more detail, for any matching π, we can define the inter-protein log-likelihood:

$$\mathcal{L}^{\pi}_{inter} := \mathcal{L}^{\pi} - \mathcal{L}^{\pi}_{ind}, \tag{2.49}$$

where, from the likelihood of the joint model, we subtract the sum of the likelihoods of the two independent protein models.

π is an additional latent variable, and we look for the maximum likelihood solution:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\left(\mathcal{L}^{\pi}_{inter}\right). \tag{2.50}$$

This is computationally infeasible for realistic cases. First the space search is huge (it grows faster than exponentially). Second $\mathcal{L}^{\pi}_{inter}$ is rather costly to compute. Last but not least, the landscape while varying π is particularly rugged, and classic search strategies such as Simulated Annealing are infeasible. In [54] two heuristic strategies are introduced, that we sketch in Figure 2.5.

2.4.2.2 *DCA for PPI*

Once the joint MSA is constructed, we can define a new model $P(\mathbf{a}, \mathbf{b})$ for sequence $\mathbf{a} = (a_1, \ldots, a_{L_1})$ in MSA_1 and sequence $\mathbf{b} = (b_1, \ldots, b_{L_2})$ in MSA_2 in the same organism. If the members of the two protein families interact in all or most organisms, we expect co-evolution between residues between protein sequences $\mathbf{a}$ and $\mathbf{b}$, i.e. $P(\mathbf{a}, \mathbf{b}) \neq P(\mathbf{a})P(\mathbf{b})$. An intuitive extension of Eq. (2.12) is:

$$H(\mathbf{a}, \mathbf{b}) = H^1(\mathbf{a}) + H^2(\mathbf{b}) + H^{12}(\mathbf{a}, \mathbf{b}) \tag{2.51}$$

where

$$H^{12}(\mathbf{a}, \mathbf{b}) = -\sum_{i=1}^{L_1}\sum_{j=1}^{L_2} J_{ij}(a_i, b_j) \tag{2.52}$$

models the coevolution between residues between proteins sequences.

The strength of these couplings can then be used:

(1) To estimate whether members of the two families are likely to interact. A possible score, introduced in [56], is the average of the n largest interprotein F^{APC}. It takes into account the strongest signals, but averages over a few pairs to be less susceptible to noise. In [56] the average over the largest 4 predictions was used to predict with high accuracy interacting pairs of the small ribosomal subunit (SRU) and the large ribosomal subunit (LRU). Their results are largely robust with respect to the precise choice of n: any value between $n = 1$ and $n = 6$ leads to virtually identical performance [56].

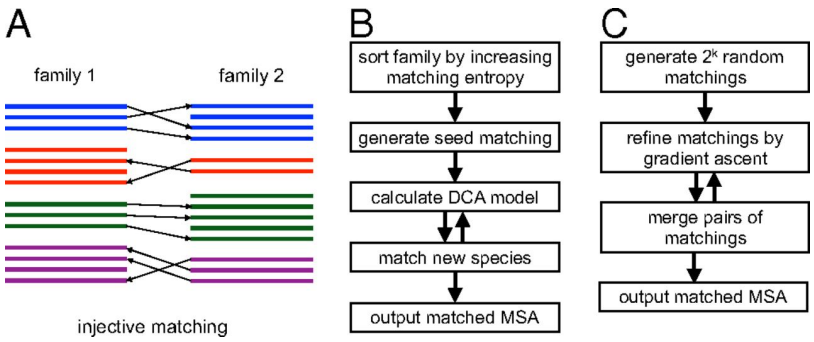

Figure 2.5: Paralog matching procedure of [54]. *Panel A)* For each species (depicted by different colors), each paralog from one species (the one with the lower paralog number) should be matched to a distinct sequence in the other species. *Panel B)* The fast but inaccurate progressive paralog-matching (PPM) algorithm. It starts from a seed MSA, constituted of sequences that can by easily matched because they do not have paralogs. Then the algorithm calculates the DCA model, uses it to add and match a new species, and iterates these two steps until all species are matched. It has limited accuracy in identifying true interaction partners since once a matching error is made, it is not corrected and, furthermore, it influences all subsequently matched species. *Panel C)* The slow but accurate iterative paralog-matching (IPM): k random matchings are generated (in practical applications, k was set to $k = 256$) and each one is independently refined using hill climbing (discrete analog of gradient ascent) of the likelihood. After refinement, pairs of matchings are merged using average matching scores. Refinement and merging are iterated until only a single refined matching is left. Source [54]

(11) To infer residue-residue contacts between proteins. Similarly to the intraprotein case, large interprotein F^{APC} are predicted to be residue-residue contacts in this interface. As shown by Uguzzoni and collaborators [57] in a large-scale analysis of homo-dimers interfaces, only large protein-protein interfaces are widely conserved across species thereby showing reliable coevolutionary signals. On the contray smaller interfaces or those being only conserved in part of a protein family cannot be easily detected by DCA.

In Chapter 3 we introduce PhyDCA which allows a large scale analysis of protein-protein interaction networks. Our approach combines coevolutionary signal from different scales (correlated presence-absence patterns of proteins across species, and correlations in the amino-acid usage) to provide multi-scale evidence for direct but unknown interaction between protein families.

In Chapter 4 we introduce *FilterDCA*, a simple and interpretable supervised machine learning method, which increases the performance of inter-domain contact prediction with respect to unsupervised DCA.

2.4.3 *Mutational landscape*

For a given protein sequence, a task of great biomedical interest is to access the mutational landscape, i.e. to determine the effect of individual mutations on the gene activity. It can help to identify mutations related to antibiotic resistance in bacteria or to the identification of disease-causing mutations in humans. Experimental technologies like *deep mutational scanning* [18] can yield insights about the mutational landscape around the native sequence (the so-called *wild-type*). They allow for the full mapping of the sequence space located one or two mutations away from the native *wild-type* to the phenotype space.

How can the effect of an amino acid change on a protein be computationally inferred?

The simplest scoring system is to compare the difference of energies of the Potts Hamiltonian Eq. (2.12) (or log-probabilities) between the wild-type and the corresponding mutant:

$$\Delta H^{mut} = H(\mathbf{a}^{mut}) - H(\mathbf{a}^{wt}) \tag{2.53}$$

Recent studies [58–62] have demonstrated that experimental measurements of fitness differences are empirically correlated with the change in Potts statistical energy ΔH^{mut} in a number of situations, from viral over bacterial to human proteins. Also, it has been shown [58, 61] that the epistatic DCA model constantly overcomes the Profile model in predicting mutational effects. Indeed, DCA is able to capture epistatic couplings between residues, and therefore to assess the dependence of mutational effects on the sequence context where they

appear. We will discuss the epistasis and context-dependency in more details in Chapter 5.

2.4.4 *Scoring of sequences*

In the DCA framework, low energy sequences are more likely to be statistically similar to the sequences of the MSA. It has been shown [49] that the energy of a sequence in the DCA model is a good predictor of its ability to fold.

In [63], Socolich and collaborators designed new artificial sequences of the WW domain (Pfam PF00397, N = 33 residues), and experimentally tested their ability to fold into the native WW structure.

They generated four groups of sequences:

1. Natural (NAT): natural WW sequences drawn from the original MSA.
2. Random (R): random sequences, generated by random scrambling of the alignment, thus killing all existing statistical patterns.
3. Independent-site conservation (IC): artificial sequences generated by shuffling independently each column of the MSA, thus having same single-site frequencies than the original MSA but no correlations.
4. Coupled conservation (CC): artificial sequences with the same single- and pairwise-frequencies than the original MSA. They are generated using a simulated annealing procedure.

The NAT and CC sequences contains substantial fractions of folding sequence (respectively 67% and 28%), whereas none of the sequence in the R or CC data set is a functional one [63].

As shown in Figure 2.6 the Potts energy can discriminate between folding and non-folding sequences across the NAT, CC, IC and R data sets. The results in Figure 2.6 were obtained using the ACE inference method [64, 65], but other inference techniques (MF, PLM, BML), despite some quantitative differences, have comparable performance in discriminating folding from non folding sequences. In fact the model only needs to rank energies of sequences correctly, thus a more precise inference of the parameters usually does not lead to a better performance, due to the high rank-correlations between methods.

As shown in Figure 2.6, the DCA model can be potentially used for the ambitious goal of generating new and functional protein sequences by Monte Carlo sampling from the inferred model.

A key aspect of the Potts model in this application is its ability to model epistasis, because of the collective effects of pairwise couplings. The energy of a profile model fails both in discriminating folding and non-folding sequences and generating functional sequences, as these would be equivalent to IC sequences.

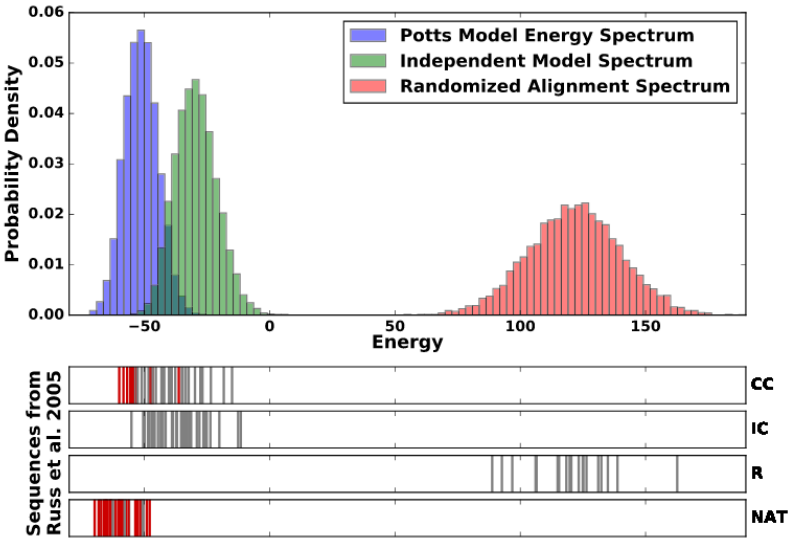

Figure 2.6: *Top*: distribution of Potts energies (parameters inferred by the ACE [64, 65] algorithm) of sequences, which are sampled, via Monte Carlo simulations, from the Potts (blue) and the profile (green) models. The red histogram corresponds to random sequences. *Bottom*: DCA energies for the WW sequences from [63] generated by coupled conservation (CC), independent-site conservation (IC) and random scrambling (R). Each bar indicates a sequence. The red bars are folding sequence while grey bars are non-folding. Note the coherence between energy values for Potts-model generated sequences in the top plot and natural sequences (NAT) in the bottom plot. Source:[49]

2.4.5 *Genome-wide DCA*

So far, DCA has been mainly applied to a number of single exemplary proteins and systems of proteins with known interaction [49]. In this case, the Potts models are inferred using MSAs containing sequences of strongly divergent sequences of homologous proteins (typical sequence identities are from 20 to 30%). However, thanks to rapid advances in next-generation sequencing, whole-genome sequencing data from densely sampled populations are starting to become available. In this context, one is interested in analyzing the genome alignments of different strains of the same species, to detect epistatic interactions between polymorphic loci. It is natural to ask if DCA can be extended to identify patterns of coevolution between single nucleotide polymorphisms (SNPs). This is not obvious. Even fast approximations, like PLM, are not easily scalable to study $10^4 - 10^5$ polymorphisms at once (the amount of genomic variation observed in analyses of many bacterial species [66]). New approximation methods are then required.

In [67], Skwark and collaborators introduced *genomeDCA* to study a large population data sets of the human pathogen *Streptococcus pneumoniae (pneumococcus).* The idea behind this method is first to use a reduced alphabet with only $q = 3$ states[1]. Second, the pneumo-coccal genome was split into about 1500 chunks, and a putative interaction score was defined using an iterative procedure based on PLM [67]. Using *genomeDCA* the authors were able to identify 5,199 putative epistatic interactions between 1,936 sites of the *Streptococcus pneumoniae* genome [67]. Similarly, in [68] Schubert and collaborators used an extend version of PLM to discovered 38 loci and 240 epistatic pairs that influence the minimum inhibitory concentrations of 5 different antibiotics in 1,102 isolates of *Neisseria gonorrhoeae.*

Another possibility is to reduce the dimensionality of the data before the DCA inference. In [38] Gao and collaborators performed a pre-filtering of the data based on empirical correlations, which can be computed directly even for very large problems. This study was performed on data from *S. pneumoniae* and it yields results very similar to [67]. But their method allows for considerable computational speed-up as the inference problem is smaller.

Note that genome wide scan for epistasis are seen as potentially a very fruitful approach to better understand the genetics of these diseases and to identify new therapeutic strategies, therefore it is not surprising that in recent years, there has been an increasing interest in this field. For example, in [69], Cui and collaborators studied strains of *Vibrio parahaemolyticus* (a human gastrointestinal pathogen), finding that the great majority of interactions (85%) were detect between acces-

1 each locus in one sequence can be a major allele (i.e. the most common allele for a given SNP), a minor allele, or the locus can be missing.

sory genes, many involved in carbohydrate transport and metabolism, while only few interarctions involving the genes in the core genome.

In this thesis, we adopt a different point of view to extend the coevolutionary analysis at the genome scale. In Chapter 3, we study the PPI networks exploiting coevolutionary signals at multiple but interconnected scales, ranging from the correlated presence or absence of related proteins (or their genes) across genomes, down to the correlated usage of amino-acids in residues, which are located in different proteins but in contact across the interface.

In Chapter 5, through an extensive genome-wide study of *E.coli* strains, we will be asking to what extent DCA models inferred from divergent homologous, are informative about intra-genic epistasis and context-dependency in recently diverged sequences.

Part II

PROTEIN-PROTEIN INTERACTIONS

Few proteins exert their function in isolation. Instead, the vast majority of proteins interact with others for proper biological activity, forming networks of protein-protein interactions (PPI). PPI can be studied focusing on two major aspects, (i) analysis of protein-protein interaction networks, and (ii) identification of the interaction interfaces within a protein complex. In this second part, we first present *PhyDCA* (Chapter 3) which, by combining proteins coevolution at multiple but interconnected scales, yields valuable insights about the protein interaction network. Second, we introduce *FilterDCA* (Chapter 4), a simple and interpretable supervised machine learning method, to improve the inter-protein contact prediction. The idea behind our approach is that residue-residue contacts follow typical patterns that can be used for constraining the DCA predictions. We demonstrate the effectiveness of *FilterDCA* in terms of interpretability and prediction performance.

3

PHYLETIC DIRECT COUPLING ANALYSIS

3.1 MOTIVATION

Protein-protein interactions (PPIs) play fundamental roles in the vast majority of biological processes. Hence, unveiling the PPI network organization is one of the most formidable tasks in systems biology today. High-throughput experimental technologies, such as large-scale yeast two-hybrid [70] analysis and in protein affinity mass-spectrometry studies [71] allowed to enhance our knowledge of protein interaction networks. However, the reliability of these methods remains problematic due to their high false-positive and false-negative rates [72].

To complement experimental approaches, we propose a genome-wide coevolutionary method, called *PhyDCA*. It is based on the fact that interacting proteins are required to coevolve across several scales, from correlated presence-absence patterns of proteins across species up to correlations in the amino-acid usage. *PhyDCA* bridges these different scales within a common mathematical-statistical inference framework.

At the genome level, we revisit a classical method called *phylogenetic profiling* [73, 74] which uses presence/absence correlations across genomes to predict functionally related protein families. We introduce the concept of *phyletic couplings*: by using a global statistical model, we are able to disentangle direct and indirect correlations in the presence and absence of protein domains across more than 1000 fully sequenced representative bacterial species. Phyletic couplings substantially increase the capacity to find relations between domains beyond correlations: while standard correlation measures used in phylogenetic profiling only reach 30–50% of true positives between the first 1000 predictions, the positive predictive value of phylogenetic couplings reaches about 80%. These relations can be physical interactions, but also genomic co-localization (and thus likely functional relations).

To refine the results and predict physical interactions, we have added a coevolutionary analysis on the scale of residue-residue covariation, as provided by DCA, to identify currently unknown but biologically sensible physical interactions between protein families. We find that 72% of the 500 phylogenetically most coupled pairs correspond to large enough alignments to run DCA, and 12.5% of these have significant DCA scores, meaning that these domain pairs are our strongest candidates for predicted domain-domain physical interactions. Since they are not co-localized in the same protein, they also provide predictions for new protein-protein interactions.

Similarly, negative phyletic couplings appear to be biologically reasonable. They disfavor the joint presence of two domains within the same genome. In our analysis of the pairs of the strongest negative couplings we find many pairs having the same functionality, including documented pairs of convergent evolution. Some pairs actually are of unknown function, and our method might help to transfer functional annotations from one domain to the other.

Supplementary information are provided in A.1.

3.2 ARTICLE

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A multi-scale coevolutionary approach to predict interactions between protein domains

Giancarlo Croce[1], Thomas Gueudré[2], Maria Virginia Ruiz Cuevas[1], Victoria Keidel[3], Matteo Figliuzzi[1], Hendrik Szurmant[3], Martin Weigt[1]*

1 Sorbonne Université, CNRS, Institut de Biologie Paris Seine, Biologie computationnelle et quantitative–LCQB, Paris, France, **2** Italian Institute for Genomic Medicine, Torino, Italy, **3** Department of Basic Medical Sciences, College of Osteopathic Medicine of the Pacific, Western University of Health Sciences, Pomona CA, United States of America

* martin.weigt@upmc.fr

OPEN ACCESS

Citation: Croce G, Gueudré T, Ruiz Cuevas MV, Keidel V, Figliuzzi M, Szurmant H, et al. (2019) A multi-scale coevolutionary approach to predict interactions between protein domains. PLoS Comput Biol 15(10): e1006891. https://doi.org/10.1371/journal.pcbi.1006891

Editor: Sergei Maslov, University of Illinois at Urbana-Champaign, UNITED STATES

Received: February 19, 2019

Accepted: September 27, 2019

Published: October 21, 2019

Data Availability Statement: The code for estimating phylogenetic couplings and data for results (list of positive domain-domain relations, phyletic couplings for bacteria with and without E. coli as reference, for eukaryotes with human reference, DCA-scores for top 500 new predictions by phylogenetic couplings) are provided in the GitHub repository https://github.com/GiancarloCroce.

Funding: MW acknowledges funding by the EU H2020 research and innovation programme

Abstract

Interacting proteins and protein domains coevolve on multiple scales, from their correlated presence across species, to correlations in amino-acid usage. Genomic databases provide rapidly growing data for variability in genomic protein content and in protein sequences, calling for computational predictions of unknown interactions. We first introduce the concept of *direct phyletic couplings*, based on global statistical models of phylogenetic profiles. They strongly increase the accuracy of predicting pairs of related protein domains beyond simpler correlation-based approaches like phylogenetic profiling (80% vs. 30–50% positives out of the 1000 highest-scoring pairs). Combined with the direct coupling analysis of inter-protein residue-residue coevolution, we provide multi-scale evidence for direct but unknown interaction between protein families. An in-depth discussion shows these to be biologically sensible and directly experimentally testable. Negative phyletic couplings highlight alternative solutions for the same functionality, including documented cases of convergent evolution. Thereby our work proves the strong potential of global statistical modeling approaches to genome-wide coevolutionary analysis, far beyond the established use for individual protein complexes and domain-domain interactions.

Author summary

Interactions between proteins and their domains are at the basis of almost all biological processes. To complement labor intensive and error-prone experimental approaches to the genome-scale characterization of such interactions, we propose a computational approach based upon rapidly growing protein-sequence databases. To maintain interaction in the course of evolution, proteins and their domains are required to coevolve: evolutionary changes in the interaction partners appear correlated across several scales, from correlated presence-absence patterns of proteins across species, up to correlations in the amino-acid usage. Our approach combines these different scales within a common mathematical-statistical inference framework, which is inspired by the so-called direct coupling analysis. It is able to predict currently unknown, but biologically sensible interaction, and

MSCA-RISE-2016 under grant agreement No. 734439 INFERNET. HS was funded by Grant GM106085 from the National Institute of General Medical Sciences, NIH. This work undertaken partially in the framework of CALSIMLAB and supported by the public grant ANR-11-LABX-0037-01 overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (ANR-11-IDEX-0004-02). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

to identify cases of convergent evolution leading to alternative solutions for a common biological task. Thereby our work illustrates the potential of global statistical inference for the genome-scale coevolutionary analysis of interacting proteins and protein domains.

Introduction

Essential to life at the molecular level is the interplay of molecules and macromolecules. Interactions contribute to diversity and coordination of reactions to accomplish feats that would be impossible if all parts worked fully in isolation. Proteins are no exceptions and many of them undergo concerted interactions to achieve their full potential. Many interactions have been described in detail, including inter- and intra-protein domain-domain interactions, which will be the focus of this work. However, many more meaningful interactions await to be discovered and explored. An issue with the experimental description of such interactions is that many are transient and that high-throughput technologies to identify such interactions are very error prone [1]. Advances in sequencing technology and the subsequent accumulation of vast sequence databases have fueled the generation of mathematical frameworks which aim to identify protein-protein interactions [2, 3]. Some of these techniques rely on the correlated evolution of interacting proteins [4–10]. Whenever interactions are conserved across many organisms, sufficient sequence examples are now in principle available to computationally identify novel interactions relying on sequences alone.

We suggest a statistical approach based on the *coevolution of interacting protein domains*. Coevolution can be detected at very different scales, ranging from the correlated presence or absence of related proteins (or their genes) across genomes, down to the correlated usage of amino-acids in residues, which are located in different proteins but in contact across the interface. Each scale contains valuable information for detecting and understanding interactions between proteins and their domains, and adapted methods have been designed to unveil this information from data. However, none of the scales contains exhaustive information. Therefore, our work proposes a coherent mathematical-algorithmic framework bridging different scales, thereby combining the information content of the different scales.

The first, largest scale concerns the correlated presence and absence of interacting proteins in genomes. If a biological function depends on two proteins simultaneously (not necessarily via their direct physical interaction, but via any functional relation), we will either observe both proteins in a genome, i.e. the function is present, or none of them, i.e. the function is absent. More rarely we may observe the presence of only one of the two proteins. This idea is at the basis of a classical method called *phylogenetic profiling* [4, 5], which uses presence/absence correlations across genomes to predict interactions. Its accuracy suffers, however, from a number of shortcomings and confounding factors:

1. *Phylogenetic relationships* between considered genomes may introduce correlations unrelated to biological function; single evolutionary events may be statistically amplified when closely related species are included in the data. Evolutionary models taking into account the underlying species tree, have been proposed [11–13] to prune such correlations.
2. Correlations may result from direct couplings, e.g., when two domains or proteins interact physically, but they may be caused by intermediate partners: If A co-occurs with B, and B with C, also A and C will show correlations. Analyses based on partial correlations [14] and spectral analysis [15] have been proposed to *disentangle direct from indirect correlations*.

3. Simple presence/absence patterns cannot *discriminate physical interaction from more general relationships*, like co-occurrence in a biological pathway or genomic co-localization. Here, using full amino-acid sequences instead of presence/absence patterns may help to refine the analysis, e.g. via the comparison of protein-specific phylogenetic trees [6].

This last point actually suggests to change resolution, and to consider coevolution at the residue scale to refine the analysis of phylogenetic profiles. The last decade has seen important progress in this respect [16, 17], related to methods like Direct Coupling Analysis (DCA) [18, 19], Gremlin [20] or PsiCov [21]. DCA-type methods were initially developed to capture the correlated amino-acid usage of residues in physical contact. Concerning interacting proteins, they have triggered a breakthrough in using sequence covariation for inter-protein residue-residue contact prediction [16, 17], which in turn is used to guide computational quaternary structure prediction [22–25].

Beyond structure prediction, DCA was suggested for the identification of interacting proteins [9, 10, 26, 27]. Such analysis requires the construction of a large joint multiple-sequence alignment (MSA) of two protein families, with each line of the MSA containing two potentially interacting proteins. However, when proteins possess numerous paralogs inside the same genome, the matching of potentially interacting paralog pairs becomes computationally hard [8, 28]. In some cases, genomic co-localization (e.g. bacterial operons) helps to identify the interacting paralogs [18, 23, 24]. Residue-residue coevolution itself has recently been proposed as a means to match paralogs, and to identify specific interaction partners [26, 27]. While results for individual protein pairs are promising, the computational cost is prohibitive for genome-wide analysis, i.e., for systematically investigating all pairs of present protein families for signatures of coevolution and thus interaction.

Our work addresses this issue, together with Points 2 and 3 given above. We propose a common statistical-modeling framework, which is applied successively to the genomic and the residue scale (presence/absence patterns and amino-acid sequences) of coevolution. It is intended to extract information from data, which cannot be extracted at each individual scale. Performing the genome-wide analysis on the coarse scale of presence/absence patterns, we can identify promising protein-domain pairs, which are subsequently analyzed using DCA at the fine residue scale. A direct comparison of our genome-wide results with those obtained using a phylogeny-aware method [29] unveils some interesting connections between Points 1 and 2 above.

Results

Phyletic couplings improve the prediction of domain-domain relationships beyond correlations

The analysis starts with a fairly standard construction of phylogenetic profiles [5], as outlined in Fig 1. Multiple-sequence alignments are needed at a later stage to perform inter-protein DCA. Since Pfam MSA have been extensively used in this respect, the analysis is performed on the domain level [30], using Pfam [31] as the input database. Pfam is based on reference genomes and we use the 1041 bacterial ones. The bacterial model organism *Escherichia coli* is used as a reference, i.e. only the 2682 domain families existing inside the K12 strain of *E. coli* are considered (the *Supplement* S1 Text Fig K shows that the results are robust when expanded beyond this choice). Since our method is based on covariation of presence and absence of domains in genomes, only variable domains existing in at least 5% and at most 95% of the considered genomes are considered, leaving 2041 domains. Note that the upper limit removes domains, which are omnipresent in the bacteria–mostly related to central life processes like

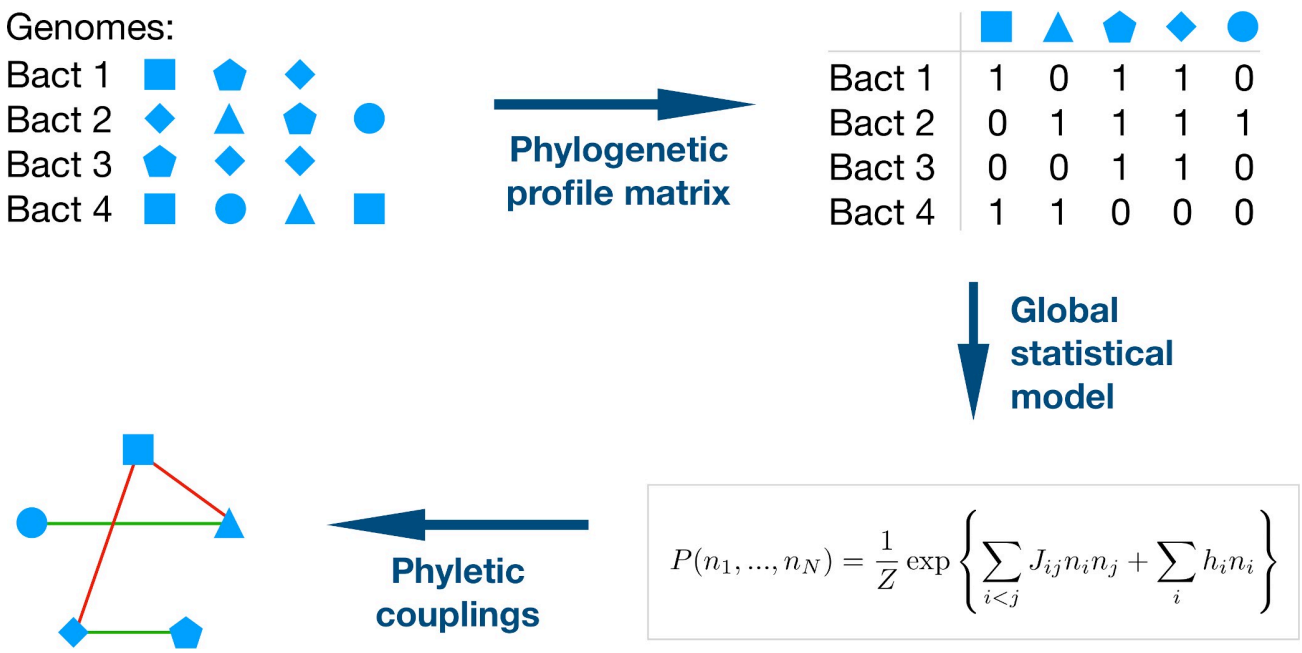

Fig 1. Schematic representation of the inference of phyletic couplings. –The composition of bacterial genomes in terms of protein families is extracted from the Pfam database. The presence and absence of each family is coded into the binary phylogenetic profile matrix (PPM); note that this matrix does not account for the presence of multiple paralogs of a domain. The statistics of the PPM is reproduced by a global statistical model $P(n_1, \ldots, n_N)$ for a full genomic phylogenetic profile, the model corresponds to a lattice gas model in statistical physics. The strongest positive couplings (favored domain-domain co-occurrence) are expected to stand for positive relationships between domains, like domain-domain interactions or genomic co-localization. Negative couplings (avoided co-occurrence) is expected to indicate alternative solutions for the same biological function, like in cases of domain families in a common Pfam clan, or for convergent evolution.

https://doi.org/10.1371/journal.pcbi.1006891.g001

replication, transcription and translation. However, being omnipresent, these domains cannot give any covariation signal within phylogenetic profiling. They could be analyzed using the finer residue-scale of coevolution, which might bring complementary evidence for interactions between these domains, but this analysis is out of scope in the current paper. The final input data are given by a binary phylogenetic profile matrix (PPM) of $M = 1041$ rows (species) and $N = 2041$ columns (domains), with entries 1 if a domain is present at least once in a genome, and zero if it is absent, cf. *Methods* and Fig 1.

An important breakthrough in coevolutionary analysis at the residue level was the step from a local correlation analysis to global maximum-entropy models [16, 32], which are able to disentangle indirect (i.e. collective) effects in correlations, and to explain them by a network of direct couplings. Here we show that the same idea can be adapted to phylogenetic profiling, and leads to a strongly increased accuracy in predicting relationships between domains. The method, which we call *Phyletic Direct Coupling Analysis (PhyDCA)*, infers a statistical model $P(n_1, \ldots, n_N)$ for the phylogenetic profile of an entire species, i.e. for a binary vector $(n_1, \ldots, n_N)$ signaling the presence or absence of all N considered domains in the considered species, cf. *Methods* for details. The PhyDCA model resembles a *lattice-gas model* in statistical physics, describing N coupled particles that can be present or absent. The phyletic coupling J_{ij} between particles / domains i and j can be positive–i.e. the presence of one domain favors the presence of the other. In this case we expect a positive relationship between the two domains, corresponding to biological processes requiring both domains. The coupling J_{ij} can also be negative–i.e. the presence of one domain favors the absence of the other. We would expect that these domains have overlapping functionalities, and one of the two is sufficient to guarantee

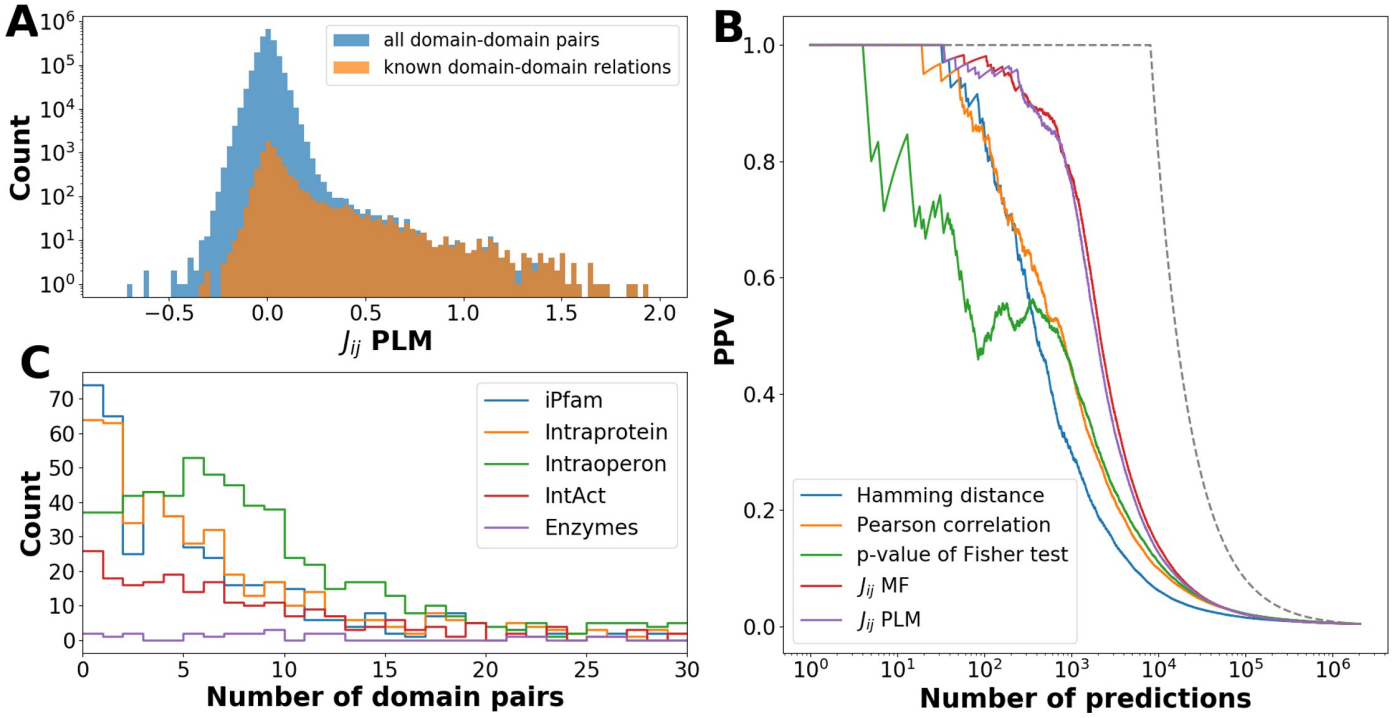

Fig 2. Phylogenetic couplings predict domain-domain relationships. –*Panel A* shows histograms of couplings J_{ij} as inferred using pseudo-likelihood maximization (PLM), cf. *Methods*, for all domain-domain pairs (blue) and for the subset of known positive domain-domain relations (brown). The histogram shows a dominant central peak around zero (note the logarithmic scale of the counts) with a pronounced fat tail for positive couplings. In contrast to the central peak, this tail is strongly dominated by the known positive domain-domain relations. A small tail for negative couplings is visible, too, but much less pronounced. *Panel B* shows the PPV (positive predictive value), defined as the fraction of known domain-domain relations in between the strongest couplings or correlations. A random prediction would correspond to a flat line close to zero; a perfect prediction would follow the dashed black line. Note that the curves corresponding to phylogenetic couplings (inference vis PLM (pseudo-likelihood maximization) or MF (mean field), cf. *Methods*) are substantially higher than those using correlation measures. *Panel C* shows, in bins of 100 domain pairs ordered by their phyletic couplings, the number of pairs belonging to the different parts of the positive-relation list (note that the categories are not exclusive, so the sum of different categories may exceed 100). We find enrichment of co-localized and interacting domain pairs, but not of related enzymes.

https://doi.org/10.1371/journal.pcbi.1006891.g002

this functionality in a species. Fig 2A shows a histogram of the couplings found for the phylogenetic coupling matrix. We observe a clear bulk of small coupling values concentrated around zero, with a broad tail for larger positive values, and a less pronounced tail for negative values.

The performance of PhyDCA can be assessed by comparing the domain pairs of strongest phyletic couplings to a carefully compiled list of 8,091 known domain-domain relations. As is explained in *Methods*, we have included genomic, functional and structural relationships: domains may coexist inside a single protein, they may be co-localized in an operon, they may be in contact in an experimental crystallographic structure or an interaction might be known according to other experimental techniques, or they may belong to enzymes catalyzing related reactions.

The PhyDCA couplings J_{ij} are ordered by size, and the fraction of positive relations in between the highest-scoring domain-domain pairs is calculated (PPV = positive predictive value). Fig 2B shows the results: we observe a strong enrichment in known positive relations in between strongly phyletically coupled domain-domain pairs. This enrichment is much stronger than for local correlation measures like Hamming distance, Pearson correlation or p-value of Fisher's exact test applied individually to two domains (i.e. two columns of the PPM): E.g., for the first 1000 predictions we observed a PPV of about 0.8 for the phyletic couplings, and only 0.3–0.5 for the different correlation measures. Interestingly, the difference between

applied PhyDCA approximations based on mean-field or pseudo-likelihood maximization is much smaller than expected from experience with contact-prediction in standard DCA. As is shown in the *Supplement* S1 Text, Fig C, couplings of both approximations are highly correlated (Pearson correlation 93% for all domain pairs, 97.5% for the known positives), resulting rather in a minor relative reranking of the two predictions than in a different accuracy. Similarly, the effect of applying the average-product correction (cf. *Methods*) has only a limited effect. As is shown in Fig 2C, interacting and co-localized domain pairs are enriched in the predictions of large positive couplings, whereas enzymes from related metabolic reactions are not. Interestingly, pairs with intra-protein co-localization are most enriched in between the strongest PhyDCA couplings (the comparable iPfam enrichment can be traced back to intra-chain co-crystals, i.e., to the same signal), which is confirming their evident functional relationship as compared to, e.g., pairs in distinct proteins coded in a joint operon. However, even inside multi-domain proteins the coupling density remains low, which results from both the sparsity of strong couplings in general, and the fact that the same domain may exist in very different protein architectures, thereby reducing correlation signals related to a specific multi-domain architecture.

From an overlay of the J_{ij}-histograms for all domain pairs and those with known relations in Fig 2A, we immediately see that the fat tail is strongly dominated by the known relations. This domination stops as we leave the tail and enter the bulk of the histogram, as a result we can determine a threshold of 0.3–0.5 for couplings to be significant. This threshold is coherent with the sharp drop in PPV in Fig 2B after about the first 1000 predictions.

Databases of genome-wide protein-protein or domain-domain interactions are currently incomplete. We therefore expect the real PPV to be even higher than the one measured in Fig 2: strongly coupled domain-domain pairs *not* belonging to our list of positives may actually be considered as predictions for new, currently unknown relations. According to the observations in Fig 2C, these relations might be direct physical interactions, but also genomic co-localization (frequently related to joint biological function). Before exploring these possibilities in more detail and on the finer scale of the residue-residue coevolution, we compare the PhyDCA results to phylogeny-aware correlation analysis and investigate the negative tail of the J_{ij} distribution.

Comparison of phyletic couplings to phylogeny-aware analysis of correlated presence/absence patterns

Phyletic couplings are, like simpler correlation measures, based on counting co-presence and co-absence of proteins or domains. However, due to the uneven phylogenetic distribution of species in our dataset, single evolutionary event may be amplified when appearing in an ancestor of several closely related species. More importantly in the context of this study, phylogeny may introduce spurious correlations in the presence and absence of domains, which are not related to biological function.

To remove this bias, several methods have been proposed, cf. [11, 13], which use evolutionary models to decide, if observed correlations can be explained by phylogeny alone (i.e. by independent evolution on a phylogenetic tree), or remain significant even when such phylogenetic effects are removed. Since this idea is complementary to the one behind PhyDCA, it is important to compare the outcome of both approaches.

To this end, we have used the CoPAP (coevolution of presence-absence patterns) server [29]. It uses the same type of binary input matrix of our approach, and is able to efficiently treat matrices of more than 2,000 domains across more than 1,000 species. As an output, CoPAP provides p-values measuring the significance of correlated domain presence and

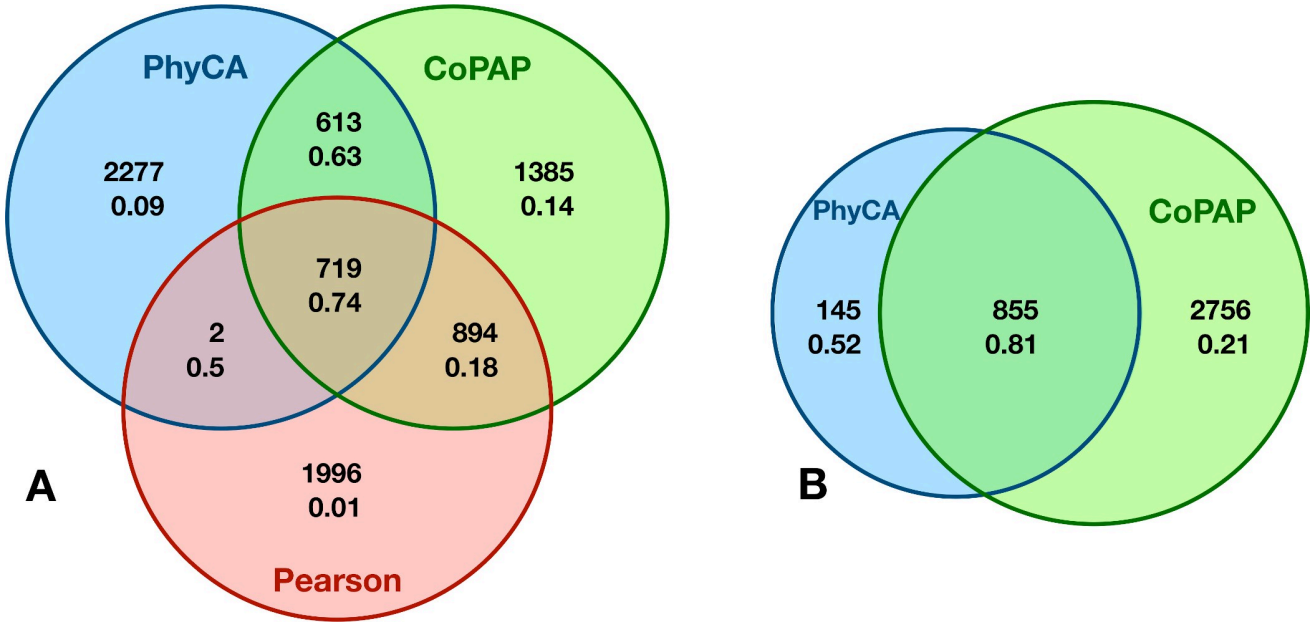

Fig 3. Comparison of simple correlations, phyletic couplings and phylogeny-corrected correlations. –Panel A shows a Venn diagram for the 3,611 first predictions of each of the three coevolution measures as extracted by Pearson correlation (red), PhyDCA (blue) and CoPAP (green). Numbers are the size of the corresponding intersection, and the PPV indicating the fraction of true positives according to our list of positive domain-domain interactions. Panel B compares the first 3,611 CoPAP predictions of highest possible significance, with the most significant 1,000 PhyDCA predictions. Most of them (855) are found to be significant by CoPAP, and of very high PPV (81%). However, not all CoPAP pairs are strongly coupled, and thus PPV is reduced (21%).

https://doi.org/10.1371/journal.pcbi.1006891.g003

absence, as compared to independently evolving domains on the same phylogenetic tree. The group of maximum significance ($log_{10}p < -7.9$) contains 3,611 domain pairs, out of which 1,251 (34.6%) are true positives in our list of known domain-domain relationships.

Since a further sorting of these pairs using CoPAP results is not possible (p-values are calculated using finite simulations), we compare them to the first 3,611 domain pairs extracted by PhyDCA, and to the 3,611 domain pairs of highest Pearson correlation. The Venn diagram in Fig 3 and the numbers given in Table 1 allow for a number of interesting observations:

- While CoPAP and PhyDCA have similar global PPV, with an advantage for CoPAP (34.6%) over PhyDCA (31.2%), Pearson correlation performs substantially worse (PPV 19.7%).
- Very small fractions of the correlated pairs, which are discarded by PhyDCA or CoPAP, are TP: PhyDCA discards 2,890 pairs of PPV 6%; CoPAP discards only 1,998 pairs, but with even lower PPV (1.2%).
- 74% of the 721 correlated pairs, which are retained by PhyDCA, are TP. Note that almost all of them (719/721) show also a significant CoPAP signal.
- Only 43% of the correlated pairs, which are retained by CoPAP, are TP. PhyDCA divides them into two groups of comparable size but distinct PPV. For the 719 pairs retained also by PhyDCA, the PPV rises to 74%. The other 894 pairs have weak phyletic couplings, so their significant correlation has to be interpreted as dominated by indirect effects. Actually only 18% are TP.
- When going to lower Pearson correlations, both CoPAP and PhyDCA decrease their accuracy. However, their intersection shows 613 pairs with a high PPV of 63%.

Table 1. Comparison of the predictions of Pearson correlation, PhyDCA and CoPAP. –We analyze the different combinations between the 3611 highest scoring predictions according to each of the three scores. In the first three columns, "YES" means that predictions are retained for the concerned score, "NO" means that predictions are discarded by the score, and "–" indicates, that the score is not taken into account.

Pearson	PhyDCA	CoPAP	Elements	TP	PPV
–	–	YES	3611	1251	0.346
–	YES	–	3611	1126	0.312
–	YES	YES	1332	915	0.687
–	YES	NO	2279	211	0.093
–	NO	YES	2279	346	0.152
YES	–	–	3611	713	0.197
YES	–	YES	1613	689	0.427
YES	–	NO	1998	24	0.002
YES	YES	–	721	531	0.736
YES	YES	YES	719	530	0.737
YES	YES	NO	2	1	0.500
YES	NO	–	2890	182	0.063
YES	NO	YES	894	159	0.178
YES	NO	NO	1996	23	0.012
NO	–	YES	1998	572	0.286
NO	YES	–	2890	595	0.206
NO	YES	YES	613	385	0.628
NO	YES	NO	2277	210	0.092
NO	NO	YES	1385	187	0.135

https://doi.org/10.1371/journal.pcbi.1006891.t001

- The 2,277 pairs only identified by PhyDCA have a low PPV of only 9%. This is coherent with Fig 2B, which shows a sharp PPV drop in PhyDCA after the first ca. 1,000 phyletic couplings. We have therefore compared these 1,000 domain pairs separately to CoPAP. A vast majority of 855 pairs have the highest possible significance in CoPAP, this intersection has a PPV of 81%. The other 15% have lower CoPAP scores and lower PPV (52%). Interestingly, only 21% of the 2,756 strongest CoPAP without strong coupling are TP, illustrating again the capacity of PhyDCA to–at least partially–disentangle direct couplings from indirect correlations.

In principle, CoPAP and PhyDCA treat very different confounding factors of coevolutionary analysis–phylogenetic biases and indirect correlations. So, it might appear astonishing that almost none of the correlated pairs, which are strongly coupled in PhyDCA, are actually discarded by CoPAP. The reason might be given by the spectral properties of the covariance matrices of the input data, and their relation to phylogeny and direct couplings. As shown in [33], the phylogenetic bias is most evident in the largest eigenvalues of the data-covariance matrix. These correspond mostly to extended eigenmodes, which in turn give rise to a dense network of small couplings [15, 34]. On the contrary, the strongest pairwise couplings are related to small eigenvalues with more localized eigenmodes, which give rise to strong, sparse couplings. Phylogenetic biases and strong direct couplings are thus related to different tails of the eigenvalue spectrum of the covariance matrix, the strongest PhyDCA couplings are thus robust with respect to phylogenetic biases.

On the other hand, we expect non-phylogenetic but indirect correlations to exist, related to the observation that PhyDCA separates the CoPAP output into strongly coupled pairs of high PPV, and weakly coupled pairs of reduced PPV. To further illuminate these indirect effects, we have introduced Fig H into the *Supplement* S1 Text, which shows a scatter of phyletic couplings vs. Pearson correlations for the CoPAP output. We find a clear triangular shape of this

scatter plot: large couplings imply large correlations, but large correlations exist also between pairs of small coupling. The coupling network is thus sparser than the correlation network. Since the PhyDCA model reproduces all correlations, at least some of them must be induced indirectly. We have also taken the network of the before-mentioned 1,000 strongest phyletic couplings, and studied the correlations as a function of the distance along this network. As is shown again in the *Supplement* S1 Text, Fig I, the strongest correlations appear between directly coupled pairs, and the correlations decay with distance until they saturate at a low but non-zero level. This observation is coherent with the idea, that the empirical correlations found in the data have at least three contributions–direct correlations induced by direct couplings (at distance 1), indirect couplings induced by coupling chains, and a ground level of correlations, which possibly result from phylogenetic correlations between the species. Taking alternatively the network induced by the 1,613 domain pairs of high Pearson correlation and CoPAP score, we find a slower decay of correlations along the network, cf. Fig J in S1 Text. At same distance, pairs on the phyletic coupling network are less correlated than those on the correlation/CoPAP network, demonstrating that the coupling network more parsimoniously explains the connectivity patterns present in the data.

Negative phyletic couplings appear between alternative solutions for the same biological function, including cases of convergent evolution

A smaller tail of negative phyletic couplings can be observed in Fig 2A. A negative coupling disfavors the joint presence of two domains in the same genome, i.e., if one of the negatively coupled domains is present in a genome, the other is less likely to be simultaneously present. Intuitively this suggests similar functionalities, one of the two domains is sufficient, the joint presence unnecessary or even costly for a bacterium. Such pairs, called anti-correlogs in [14] were used in [35] to identify analogous enzymes replacing missing homologs in biochemical pathways.

When using *E. coli* as a reference genome, the number of such negative couplings is limited, since only domain pairs co-occurring in *E. coli* are analyzed. To better understand the meaning of negative couplings, we have therefore extended the original analysis to all 9,358 families containing bacterial protein domains. While results restricted a posteriori to *E. coli* are very robust (96% correlation, cf. *Supplement* S1 Text, Fig K), the extended analysis leads to a substantially higher number of negative couplings.

To explore these in some detail, we analyzed the 20 domain pairs with the strongest negative couplings, cf. Table 2 (an extended list is given in Table C in *Supplement* S1 Text). From their detailed analysis it is evident that protein pairs can be classified into three distinct groups. First, we find several cases of convergent evolution as evidenced by proteins with the same or similar activities but distinct protein structures (rankings 1, 2, 9, 14, 15, 16). Second, we find domain pairs of the same fold and, where known, of similar activity. For various reasons these are not described by the same Pfam HMM (rankings 3, 4, 6, 7,8, 10,11, 17,19), but typically belong to the same Pfam clan indicating distant homology. Lastly, there are several cases of relatively unknown activity, and some domains have no known structure (rankings 5, 12, 13, 18, 20).

Cases of convergent evolution include PF00303 and PF02511, which describe two different thymidylate synthases, the former a 5,10-methylenetetrahydrofolate, the latter a flavin dependent enzyme [36]. Interestingly, PF00186, dihydrofolate reductase is also strongly negatively coupled with PF02511 (but positively to PF00303), since the former is not needed to regenerate 5,10-methylenetetrahydrofolate when the flavin-dependent enzyme is used. Other cases of convergent evolution are PF01220 and PF01487 that describe two classes of dehydroquinases

Table 2. The 20 domain pairs of top negative phyletic couplings.

	Pfam 1	Pfam 2	J_{IJ}	Domain 1 description	Domain 2 description
1	PF00303	PF02511	-0,9978	Thymidylate synthase	Thymidylate synthase complementing protein
2	PF01220	PF01487	-0,9277	Dehydroquinase class II	Type I 3-dehydroquinase
3	PF02834	PF13563	-0,9075	LigT like Phosphoesterase	2'-5' RNA ligase superfamily
4	PF00406	PF13207	-0,8258	Adenylate kinase	AAA domain
5	PF01205	PF02594	-0,7077	Uncharacterized protein family UPF0029	Uncharacterised ACR, YggU family COG1872
6	PF13623	PF13624	-0,7051	SurA N-terminal domain	SurA N-terminal domain
7	PF04816	PF12847	-0,6316	tRNA (adenine(22)-N(1))-methyltransferase	Methyltransferase domain
8	PF00636	PF14622	-0,6281	Ribonuclease III domain	Ribonuclease-III-like
9	PF00186	PF02511	-0,6281	Dihydrofolate reductase	Thymidylate synthase complementing protein
10	PF01227	PF02649	-0,6118	GTP cyclohydrolase I	Type I GTP cyclohydrolase folE2
11	PF06745	PF13481	-0,5844	KaiC	AAA domain
12	PF02677	PF08331	-0,581	Uncharacterized BCR, COG1636	Domain of unknown function (DUF1730)
13	PF02696	PF03190	-0,5651	Uncharacterized ACR, YdiU/UPF0061 family	Protein of unknown function, DUF255
14	PF00311	PF02436	-0,5432	Phosphoenolpyruvate carboxylase	Conserved carboxylase domain
15	PF02502	PF06026	-0,5371	Ribose/Galactose Isomerase	Ribose 5-phosphate isomerase A (phosphoriboisomerase A)
16	PF00245	PF05787	-0,5333	Alkaline phosphatase	Bacterial protein of unknown function (DUF839)
17	PF00075	PF13456	-0,5317	RNase H	Reverse transcriptase-like
18	PF01169	PF02659	-0,5294	Uncharacterized protein family UPF0016	Putative manganese efflux pump
19	PF01321	PF05195	-0,5165	Creatinase/Prolidase N-terminal domain	Aminopeptidase P, N-terminal domain
20	PF02594	PF09186	-0,5139	Uncharacterised ACR, YggU family COG1872	Domain of unknown function (DUF1949)

https://doi.org/10.1371/journal.pcbi.1006891.t002

with similar activity but significantly different primary and secondary structure [37]. PF00311 and PF02436 describe proteins in oxaloacetate biogenesis, the former from phosphoenolpyruvate, the later from pyruvate and ATP. PF00245 and PF05787 describe two classes of bacterial alkaline phosphatases, termed PhoA and PhoX with distinct protein folds [38]. PF02502 and PF02436 distinguish two classes of ribose- or phosphoribo-isomerases with differing enzyme folds.

Structurally similar proteins that are identified by different Pfam families are of less interest and will not be separately described. The fact that they are distinct enough in sequence to be covered by separate Pfam families suggests a level of divergent evolution, i.e. one or the other domain has distinct features such us additional interaction partner, distinct activity regulation etc.

Of special interest are domain pairs with unknown function. Ideally, if the function of one Pfam family becomes available one can infer the function of the other family as well. In addition, the evolutionary importance of a given protein family and its activity is often judged by its conservation across different phyla and organisms. This however neglects cases of unknown convergent evolution. Among the highest negatively coupled pairs, we did not find any, where the function of one has been clearly identified and the function of the other has not. However, there are several instances, where a potential role has been loosely associated with one or the other domain. For instance, PF01205 and PF09186 have been suggested to be involved in countering translation inhibition under starvation conditions [39]. These domains are strongly negatively coupled with PF02594, suggesting that the latter might also serve a role in countering translation inhibition. PF01169 and PF02659 are both putative transporters, the former for calcium [40], the latter for manganese ions [41]. Their coupling suggests overlapping specificities or roles. PF02677 and PF08331 describe two entirely unstudied bacterial proteins. The later appears associated with iron-sulfur cluster domains, suggesting a potential role in redox

regulation. Lastly, we find a negative coupling between domains PF02696 and PF03190. Both proteins are entirely unstudied in bacteria, but they are also common in Eukaryotes where the latter is a proposed redox protein that has been implicated in fertility regulation in mammals [42]. It would be interesting to unveil their function in the bacteria.

It might be interesting to study the context, in which these negative couplings appear in the PhyDCA network. To this end we have taken all couplings of absolute value above 0.3, resulting in a sparse network of 82 negative, and 3173 positive links. We have now studied the triangles in the resulting network, which have at least one negative coupling. From the fact, that positive links are close to 40times more frequent than negative links, we would expect the other two links of the triangle to be typically positive. On the contrary, we do find only triangles with exactly two negative and one positive coupling; they contain 29 out of the 82 negative phyletic couplings. No so-called "frustrated" triangles are found, where both supplementary links in the triangle are positive. This indicates that more likely entire processes are realized by alternative solutions, than single domains are exchanged against each other within an otherwise positively correlated solution.

A residue-scale DCA analysis of phylogenetically coupled domain pairs unveils directly coevolving pairs

As seen in Fig 2C, a large positive phyletic coupling is a strong signal for a positive relationship between two domains, but not necessarily for a direct physical interaction of the two domains within a protein complex. Furthermore, co-localization of two domains either inside the same protein (i.e. an evolutionary conserved protein architecture) or inside the same operon may lead to strong phyletic couplings.

Relying only on the coarse scale of coupled presence and absence in genomes, does not reveal more detailed information. Since the number of domain-domain pairs under question is limited as compared to all domain pairs existing in *E. coli*, we can afford computationally more expensive approaches, which study coevolution of domain pairs at the individual residue scale. To this effect, we use the procedure suggested in Gueudré et al. [27]: Two Pfam MSA for the two domain families are matched using a variant of DCA such that (a) only sequences appearing inside the same species are matched and (b) the inter-domain covariation as measurable by DCA is maximized. In [27] it was shown that this idea allows to identify protein-protein interactions via a large coevolutionary score between the two domains at a sufficiently large joint MSA. DCA scores above 0.2 at an effective sequence-pair number of at least 200 (sequences below 80% sequence identity, cf. *Supplement*) can be considered as a strong indicator for a potential interaction [10, 27]. On the contrary, according to [43], a low DCA score is not necessarily a sign for the absence of a physical interaction. A low score might also originate from a relatively small or structurally not well-conserved interface, both resulting in a weak coevolutionary signal.

We have applied the progressive paralog matching procedure of [27] to the 500 most strongly coupled domain pairs, which are not in our previously constructed test set of positive domain-domain relations, i.e. to the first 500 *predictions* at the scale of phyletic couplings. The results are presented in Fig 4A: 360 domain pairs have an M_{eff} above 200, and DCA results can thus be considered reliable. Of those 45 pairs have an inter-domain DCA score above 0.2 (24 out of the first 200 PhyDCA predictions). This number is significantly larger than for randomly selected protein pairs, cf. Fig 4B: only 10 pairs have a score above 0.2 and M_{eff} above 200, mostly related to short amino-acid sequences. This shows that the preselection by high phylogenetic couplings leads to a subsequent enrichment of high DCA scores also at the residue scale. For comparison, we have also applied the matching procedure to the 200 domain-domain pairs, which are known to interact by iPfam [44], and which have high phylogenetic

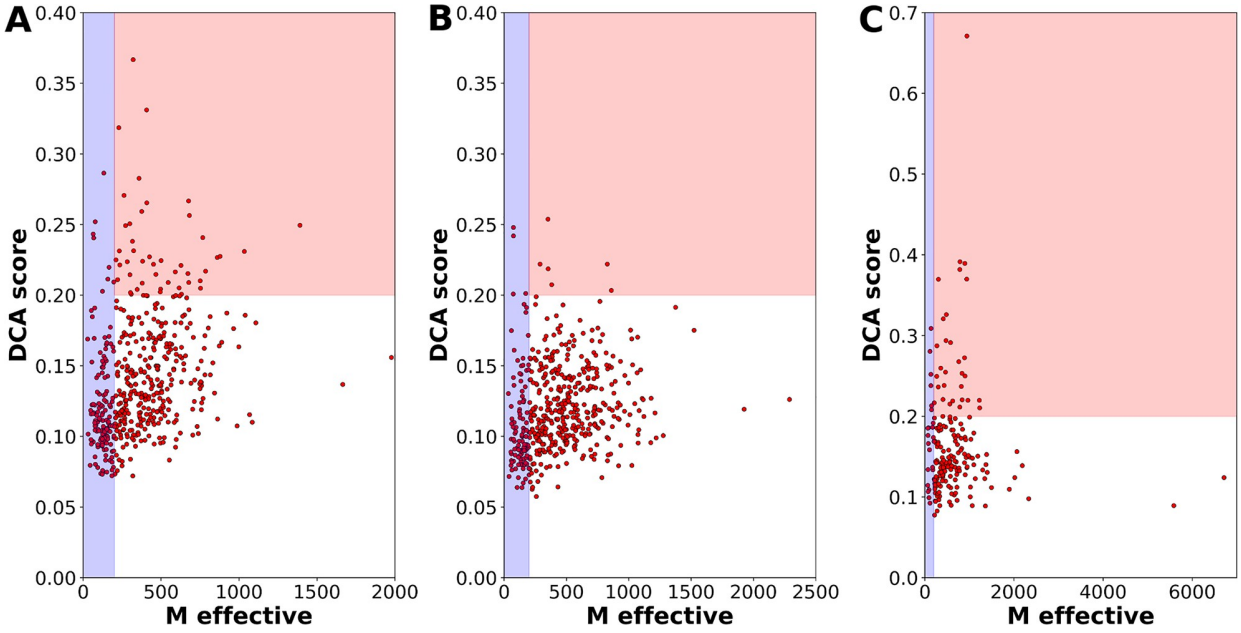

Fig 4. DCA identifies strong residue-scale coevolution between phyletically coupled domain pairs. –Panel A shows the effective sequence number (defined as the sequence number at 80% maximum sequence identity, cf. *Supplement* for the precise definition) and the DCA scores for the 500 domain pairs of strongest phyletic coupling not belonging to the positive-relation set (i.e. the 500 most significant predictions). The interesting region is the red one, where sequence numbers are sufficient to provide reliable DCA results, and DCA scores are beyond 0.2 as established in [10]. Panel B shows, as a comparison, the results for 500 randomly selected domain pairs. Only very few pairs show substantial scores, most of them related to very short peptides. Panel C shows a positive control, the 200 pairs of highest phylogenetic couplings belonging to iPfam are analyzed analogously. The fraction and the amplitude of high DCA scores is slightly increased with respect to Panel A, but the qualitative behavior is similar.

https://doi.org/10.1371/journal.pcbi.1006891.g004

couplings, cf. Fig 4C. 29 have a significant DCA score at large enough sequence number. Interestingly, the signal is only marginally stronger than for the newly predicted relations, which are discussed in more detail below. In Fig G the *Supplement* S1 Text, we analyze also the 200 phyletically most positively coupled domain-domain pairs, which co-occur inside the same protein in *E. coli*. In their case, the DCA score is found to be substantially larger. This is to be expected, since due to the intra-protein co-localization no paralog-matching has to be applied, and therefore the joint MSA of the two domain families are expected to be of higher quality. However, also in this case, some pairs show a low DCA score despite a large sequence number. This is to be expected, since not all domain-domain pairs inside a multi-domain protein have physical interactions, and also small and structurally non-conserved interfaces may lack clear DCA signals, cf. [43].

Many predictions of domain-domain interactions resulting from PhyDCA and residue-level DCA are biologically interpretable

Domain pairs with both strong PhyDCA and residue-level DCA signals are our strongest candidates for predicted domain-domain interactions. Since they are not co-localized in the same protein, they also provide predictions for new protein-protein interactions. We analyze here in detail the 24 pairs with a DCA score larger than 0.2, which result from the first 200 PhyDCA predictions.

Among these 24 pairs we find several examples of known interactions that have not yet been structurally resolved. These include K^+ transporter subunits KdpC (PF02669) and KdpA (PF03814) [45], Sigma54 activator (PF00158) and Sigma54 activator interacting domain (PF00309) [46] and exonuclease VII subunits domains PF02609, PF2601 and PF13742 [47].

For several additional positively coupled pairs an interaction seems functionally very likely but to our knowledge no interaction studies are available. These are all proteins involved in pilus formation or maturation. Domain PF06750 is a putative methyl transferase domain in the prepilin peptidase PppA, and proposed to interact with methylation motif domain PF07963, found in numerous pilin proteins and with PF05157, a type II secretion system protein [48, 49]. PF05157 is also predicted to interact with domain PF05137 found in the PilN fimbrial assembly protein required for mating in liquid culture [50].

Of interest, there are predicted interactions for several members of biosynthetic pathways catalyzing either consecutive or closely following reactions. These include domains PF02542 and PF13288 of isoprenoid biosynthesis enzymes Dxr and IspF, domains PF00885 and PF00926 of riboflavin biosynthesis enzymes RisB and RibB and domains PF01227 and PF01288 of tetrahydrofolate biosynthesis enzymes Gch1 and HppK. A more complex connection is predicted between multiple domains of molybdenum cofactor biosynthesis enzyme MoaC (PF01967), MoeA (PF03453 and PF03454) and MoaA (PF06463). Similarly, scores suggest a protein-protein interaction between domains of hydrogenase maturation enzymes HypF (PF07503) with HybG (PF01455) and HycI (PF01750).

Perhaps most intriguing are the observation of strongly coupled co-occurrence and potential protein-protein interactions of two proteins pairs. Ada (PF02805) and AlkA (PF06029) are two enzymes involved in DNA repair in response to alkylation damage [51, 52]. One of the proteins serves as demethylase of guanosyl residues whereas the other excises alkylated nucleotides. These seemingly complementary functions suggest that an interaction is plausible. The other pair is YoeB (PF06769) with HicA (PF07927). These two proteins constitute two toxins of distinct toxin-antitoxin systems. Both proteins inhibit translation by distinct and complementary mechanisms and an interaction seems plausible. YoeB blocks the ribosome A site leading to mRNA cleavage [53]. HicA interacts with mRNA directly and thus acts independent of the translation apparatus [54].

Additional and perhaps plausible interactions are predicted between domains PF05930 and PF13356 of prophage protein AlpA and several phage integrase proteins as well as between domain PF13518 with PF13817, the former a HTH domain commonly associated with transposase domains and the latter a transposase domain.

Insufficient information on the function of two domain pairs and their associated proteins does not allow us to draw any conclusions on the plausibility of interaction. These are for domains PF02021 and PF13335 of proteins YraN and YifB and domains PF01906 with PF02796, the former a metal binding domain and the latter a domain found in site specific recombinases.

Lastly, we find three proposed interactions between domains found in ribosomal proteins RL36, RL34 and RL32 (PF00444, PF00468, PF01783) and also a protein of unknown function YidD (PF01809). We consider these to be likely false positive predictions since we previously observed spurious results for members of very large macromolecular complexes such as the ribosome [10]. At least the interaction between YidD and RL36 seems plausible, as the former has been suggested to play a role as membrane protein insertion factor [55].

In summary, we are able to recapitulate several known or plausible but structurally unresolved interactions and find several examples of interaction that should be of interest for future experimental studies.

Discussion

In this work, we propose a coevolutionary analysis connecting signals at the phylogenetic level (correlated presence of domain pairs across genomes) with the residue level (correlated

occurrence of amino acids between proteins). At the phylogenetic level, we introduce the concept of *phyletic couplings*: by using a global statistical model, we are able to disentangle direct and indirect correlations in the presence and absence of protein domains across more than 1000 fully sequenced representative bacterial species. Couplings substantially increase the capacity to find relations between domains beyond correlations; these relations can be physical interactions, but also genomic co-localization (and thus likely functional relations). Standard correlation measures used in phylogenetic profiling only reach 30–50% of true positives between the first 1000 predictions. In contrast the positive predictive value of phylogenetic couplings reaches about 80%. The results are very robust: when applying the same methodology to all 9358 Pfam domains appearing in the bacteria, and selecting only later the couplings between domains present in *E. coli*, couplings have 96% correlation with the couplings found by the procedure described before.

The high accuracy of phyletic couplings in predicting domain-domain relations, along with the robustness of these couplings when extensively changing the data set, allows us to hypothesize that large couplings not corresponding to known relations predict novel, unknown relations. A list of the 100 first predictions is provided in Tables A and B in *Supplement* S1 Text.

As mentioned, a large phyletic coupling does not automatically imply a direct physical interaction. Two proteins may have a strong phyletic coupling because they belong to the same multi-protein complex, without touching each other. They may have a strong phyletic coupling, because they act within the same biological process or pathway, again without any direct interaction. To refine the results and predict physical interactions, we have added a coevolutionary analysis on the scale of residue-residue covariation, as provided by DCA, in the version with paralog matching as recently proposed in [27]. We find that 72% of the 500 phylogenetically most coupled pairs correspond to large enough alignments to run DCA, and 12.5% of these have significant DCA scores.

These domain pairs are our strongest candidates for predicted domain-domain interactions. Since they are not co-localized in the same protein, they also provide predictions for new protein-protein interactions. In a detailed discussion, we have shown that most of the 24 pairs with a DCA score larger than 0.2, which result from the first 200 PhyDCA predictions have a sensible biological interpretation and, in principle, could be tested experimentally.

Similarly, negative phylogenetic couplings appear to be biologically reasonable. They disfavor the joint presence of two domains within the same genome. In our analysis of the pairs of the strongest negative couplings, presented above in *Results*, we actually find many pairs having the same functionality, including documented pairs of convergent evolution. Some pairs actually are of unknown function, and our method might help to transfer functional annotations from one domain to the other.

An important extension would be the application of our approach beyond the bacteria. Bacteria, due to their compact genomes, are overrepresented in genomic databases, including the Pfam database, which we used for our analysis. To test the applicability to higher organisms, we have repeated the same procedure, concentrating on eukaryotic genomes and taking humans as the reference species. Data get much less abundant; the phylogenetic profile matrix now contains 5343 domains as compared to only 481 eukaryotic species. Still, phyletic couplings, when compared to a positive list extracted from domain architectures of human proteins (co-localization in one protein), from iPfam [44] and human entries in IntAct [56], show a similar performance as the bacterial case, cf. Fig 5A: 75% of the first 1000 couplings correspond to known domain-domain relations. Entries corresponding to protein-protein interactions (iPfam, IntAct) are again significantly enriched, even if to a lesser extent than in the bacterial case. The most important difference emerges, however, when using paralog matching and DCA on the 200 most coupled predictions (i.e. pairs with strong phylogenetic coupling

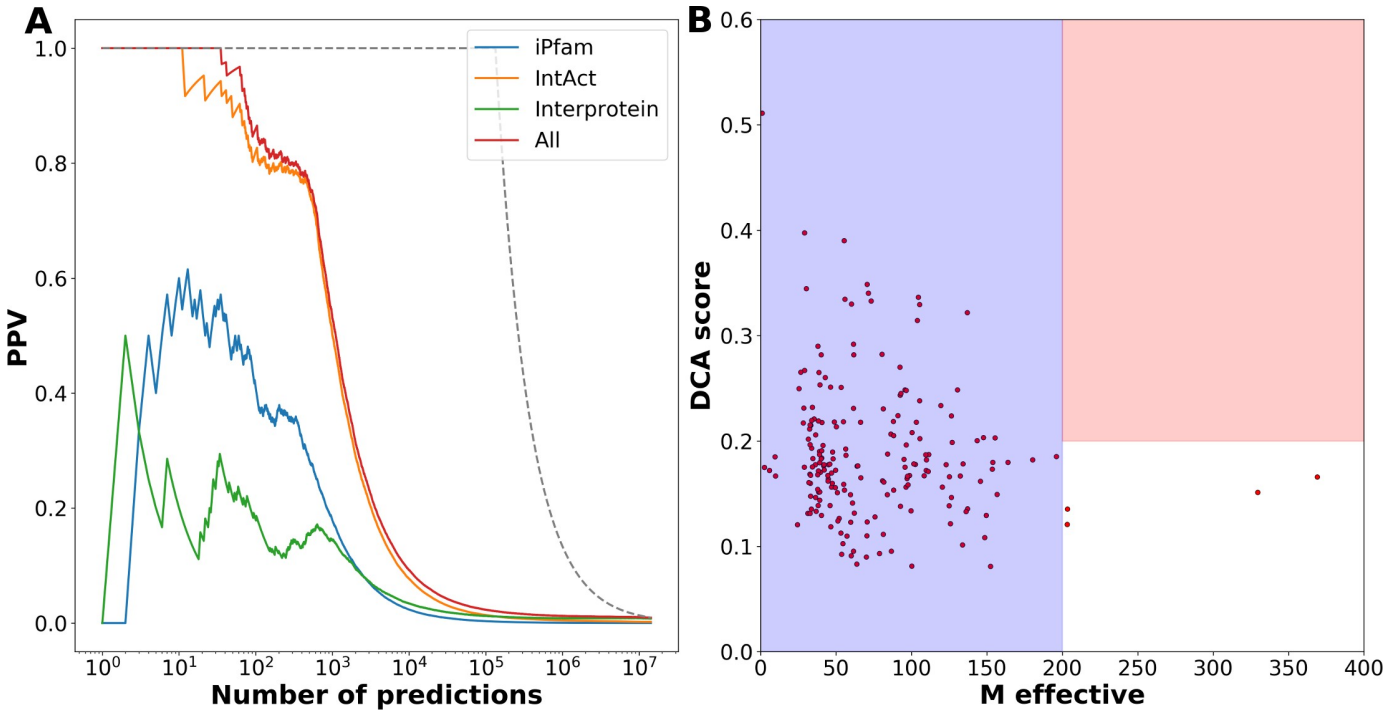

Fig 5. Performance of our multi-scale coevolutionary analysis for human protein domains. –Panel A shows the positive predictive value of the phyletic couplings for predicting positive domain-domain relationships (including protein architecture, iPfam and human IntAct entries). While there is a clear overrepresentation of intra-protein localization in between the highest-scoring domain pairs, also physical interactions as captured by iPfam and IntAct are enriched in particular in the first ca. 10^3 phyletic couplings. The overall performance is coherent with the one found in the bacteria. Panel B shows the paralog-matching and DCA results for the 200 most coupled domain pairs, which are not in the positive-relation dataset. We observe that currently the joint MSA are too small (Meff < 200) to allow for a reliable application of DCA to detect inter-protein residue-scale coevolution.

https://doi.org/10.1371/journal.pcbi.1006891.g005

but not belonging to the positive list), cf. Fig 5B: Only 2–4 have sequence numbers that allow for reliable DCA results. More eukaryotic genomes are urgently needed to carry out our full procedure also in higher species.

To conclude, our work illustrates the potential of combining rapidly growing genomic databases and statistical modeling: the increasing number of fully sequenced genomes allows for extracting rich *samples for the variability in protein content and protein sequences* across hundreds and thousands of species; their statistical analysis helps us to detect multiple scales of coevolution between interacting or functionally related proteins.

The *genomic scale* explores the correlated presence or absence of proteins (in the sense of homologous protein families) across species. This correlation has been used before within phylogenetic profiling to detect functional relations or direct interactions between proteins. Within our work, we propose to infer direct phyletic couplings via global statistical models, and prove that this concept strongly improves our capacity to detect protein relations over local correlation measures.

However, phylogenetic couplings cannot distinguish between functional relations or direct interactions between proteins. This problem can–at least partially–be resolved at the *residue scale* of inter-protein coevolution. Interacting proteins show a correlated usage of amino acids across their interface, and again global statistical modeling approaches like DCA have been used to discriminate between interacting and non-interacting protein pairs.

Since the computational cost of the residue-scale analysis is high, it is possible to analyze all pairings between 10–50 proteins, but not all pairs between thousands of proteins forming a

species' proteome. It is the combination of both scales, which allows us to first explore the genomic scale and then refine promising results at the residue scale. Doing so, we have provided a number of biologically sensible predictions for currently unknown protein-protein interactions. We provide a list of these predictions, which in turn may be tested directly.

Last but not least, we want to mention that the analysis of both scales of coevolution is done independently, i.e., in a modular way, even if using a common mathematical-statistical framework. In principle it is therefore possible to improve each single component on its own. We might, e.g., come up with a phylogenetically better-founded version of PhyDCA (i.e. combining the spirit of CoPAP and PhyDCA), to generate better candidates for novel domain-domain interactions. Similarly, improvement in paralog matching and DCA-based interaction prediction might lead to a more sensitive treatment of these candidates.

Methods

Phylogenetic profiles

Data are extracted from the Pfam 30.0 database [31]. For each of the 1,041 bacterial genomes present in Pfam, we extract all appearing protein-domain families, accounting to a total of 9,358 Pfam families. A restriction to *Escherichia coli* as reference genome (i.e. counting only domains contained in *E. coli*) reduces this to 2682 domain families. Since we are interested in the *correlated* presence / absence of domains across species, we remove all domain families with less than 5% or more than 95%, keeping only domains with at least 53 and at most 988 appearances. This removes in particular omnipresent domains related, e.g., to replication, transcription and translation. The final phylogenetic profile matrix (PPM) is a binary matrix containing M = 1,041 bacteria and N = 2,041 domains. Entries are one if a domain is present in a species (at least once), and zero if it is absent. Note that a zero entry typically indicates a true absence of the domain in a genome, since the profile matrix is entirely built on fully sequenced genomes.

In standard phylogenetic profiling [5], correlations between domains are evaluated via the Hamming distance, Pearson correlation or p-values of Fisher's exact test, cf. the *Supplement* S1 Text for the definitions in the context of our work.

Phyletic couplings

In analogy to the direct-coupling analysis on the level of amino-acid sequences, we model the phylogenetic profiles via the maximum-entropy principle by a global statistical model

$$P(n_1, \ldots, n_N) = \frac{1}{Z} \exp\left\{ \sum_{i<j} J_{ij} n_i n_j + \sum_i h_i n_i \right\}$$

with $(n_1, \ldots, n_N)$ being a binary vector characterizing the presence ($n_i = 1$) or absence ($n_i = 0$) of domain i in a species, and Z is a normalization constant also known as partition function in statistical physics. The *phyletic couplings* J_{ij} and *biases* h_i are to be determined such that the model P reproduces the one- and two-column statistics of the PPM $(n_i^a)_{i = 1,\ldots,N;\ a = 1,\ldots,M}$:

$$f_i = \frac{1}{M} \sum_{a=1}^{M} n_i^a$$

$$f_{ij} = \frac{1}{M} \sum_{a=1}^{M} n_i^a n_j^a$$

with f_i being the fraction of genomes in the PPM carrying domain i, and f_{ij} the fraction of

genomes containing both domains *i* and *j* simultaneously. While the exact determination of the marginal distributions of P requires exponential-time computations, we apply the mean-field (MF) and pseudo-likelihood maximization (PLM) approximations successfully used in the context of DCA [19, 57]; cf. the *Supplement* S1 Text for technical details. Due to the high dimensionality of the problems ($N = 2041-9358$), more precise methods based on Boltzmann machine learning, cf. [32], become computationally prohibitive. Strong positive couplings favor the joint presence or joint absence of two domains, signaling therefore a positive association between the two (genomic colocalization, functional relation, domain-domain interaction). Strong negative couplings favor the appearance of only one out of the two domains, signaling domains of similar function (e.g. convergent evolution).Before analyzing the phyletic couplings, we apply the so-called Average Product Correction (APC) [58], cf. *Supplement* S1 Text. APC is widely used to suppress spurious couplings resulting from the heterogeneous conservation statistics domain families across genomes (cf. [59]) as compared to functional couplings. In the case of PhyDCA, it has a limited effect, as is shown in Fig A of *Supplement* S1 Text.

Direct coupling analysis of inter-protein residue coevolution

To assess the coevolution on the finer scale of residue-residue coevolution, we have applied exactly the progressive matching and analysis procedure recently published by part of us in [27], details about the procedure are given in the *Supplement* S1 Text. It starts with two domain alignments, containing only bacterial protein sequences. It matches sequences between the domain families, such that (a) only sequences from the same species are matched and (b) the total inter-family covariation signal is maximized. Results are considered positive if (i) the effective number of matched sequences (at 80% seq ID) exceeds 200 and (ii) the covariation score exceeds 0.2. It has been established in [10, 27] that larger scores are rarely obtained by unrelated protein families. Note that a smaller score may be related to a functional relationship rather than a physical protein-protein interaction, or also to a small or non-conserved interaction interface [43].

Known domain-domain relationships

To assess the accuracy of our predictions, we have compiled a number of known relationships (provided in *Supplement* S1 Text). They come from different databases, the same domain-domain pair may appear multiple times, but it is counted only once in the final list of positives:

1. *Intra-protein localization*: From the Pfam database [31], we have extracted a list of domain pairs, which co-occur inside single proteins in *E. coli*. Out of 3,116 proteins, 952 contained multiple domains, giving rise to 799 distinct domain-domain relations.
2. *Intra-operon localization*: Proteins, which are co-localized inside operons, frequently share at least part of their biological function. Using a list of operons from *E. coli* [60], we compiled a list of 4,087 colocalized domain pairs.
3. *Protein-protein interaction*: The IntAct database [56] contains 5,318 pairs of experimentally found protein-protein interactions. At the domain level, we pair all domains in one protein with all domains in the second protein (adding possibly unrelated domain pairs to those interacting), obtaining 3,070 domain pairs.
4. *Domain-domain contacts in 3D structures*: The iPfam database [44] contains domain-domain interactions extracted from structural domain-domain contacts in experimentally determined complex structures in the PDB. We included intra- and inter-chain contacts,

i.e. domain-domain contacts inside a protein or between two proteins. Note that this list does not refer to *E. coli* as reference genome. In total, this accounts to 545 know relationships.

5. *Metabolic relationships between enzymes*: Using the reconstruction iJR904 of *E. coli's* metabolic network [61] and filtering out "currency" metabolites involved in more than 50 reactions (such as water, ATP etc.), we considered three relationships:

   a. *common substrate*– pairs of enzymes catalyzing reactions with at least one common substrate;

   b. *common product*–pairs of enzymes catalyzing reactions with at least one common product;

   c. *reaction chains*–pairs of enzymes catalyzing subsequent reactions, i.e., one product of one reaction is substrate of the second.

   This lead to a total of 677 known relationships.

The total list contains 8,091 domain-domain pairs, as compared to the 2,081,820 possible pairs, which can be formed out of the 2,041 domains in our PPM.

Supporting information

S1 Text. Supplementary information. This text contains technical details about the data, the computational analysis tools, and supporting results and figures.
(PDF)

Author Contributions

Conceptualization: Matteo Figliuzzi, Martin Weigt.

Data curation: Giancarlo Croce, Thomas Gueudré, Maria Virginia Ruiz Cuevas, Victoria Keidel, Matteo Figliuzzi, Hendrik Szurmant.

Investigation: Giancarlo Croce, Thomas Gueudré, Maria Virginia Ruiz Cuevas, Victoria Keidel, Matteo Figliuzzi, Hendrik Szurmant.

Methodology: Giancarlo Croce, Maria Virginia Ruiz Cuevas, Matteo Figliuzzi, Martin Weigt.

Software: Giancarlo Croce.

Supervision: Martin Weigt.

Writing – original draft: Giancarlo Croce, Hendrik Szurmant, Martin Weigt.

References

1. Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, et al. An experimentally derived confidence score for binary protein-protein interactions. Nat Methods. 2009; 6(1):91–7. https://doi.org/10.1038/nmeth.1281 PMID: 19060903
2. Harrington ED, Jensen LJ, Bork P. Predicting biological networks from genomic data. FEBS Lett. 2008; 582(8):1251–8. https://doi.org/10.1016/j.febslet.2008.02.033 PMID: 18294967
3. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol. 2002; 12(3):368–73. PMID: 12127457
4. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999; 96 (8):4285–8. https://doi.org/10.1073/pnas.96.8.4285 PMID: 10200254

5. Pellegrini M. Using phylogenetic profiles to predict functional relationships. Methods Mol Biol. 2012; 804:167–77. https://doi.org/10.1007/978-1-61779-361-5_9 PMID: 22144153
6. Juan D, Pazos F, Valencia A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Natl Acad Sci U S A. 2008; 105(3):934–9. https://doi.org/10.1073/pnas.0709671105 PMID: 18199838
7. Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins. 2002; 47(2):219–27. PMID: 11933068
8. Burger L, van Nimwegen E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol. 2008; 4:165. https://doi.org/10.1038/msb4100203 PMID: 18277381
9. Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. PLoS One. 2011; 6(5):e19729. https://doi.org/10.1371/journal.pone.0019729 PMID: 21573011
10. Feinauer C, Szurmant H, Weigt M, Pagnani A. Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. PLoS One. 2016; 11(2):e0149166. https://doi.org/10.1371/journal.pone.0149166 PMID: 26882169
11. Spencer M, Sangaralingam A. A phylogenetic mixture model for gene family loss in parasitic bacteria. Mol Biol Evol. 2009; 26(8):1901–8. https://doi.org/10.1093/molbev/msp102 PMID: 19435739
12. Cohen O, Pupko T. Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony—a simulation study. Genome Biol Evol. 2011; 3:1265–75. https://doi.org/10.1093/gbe/evr101 PMID: 21971516
13. Cohen O, Ashkenazy H, Burstein D, Pupko T. Uncovering the co-evolutionary network among prokaryotic genes. Bioinformatics. 2012; 28(18):i389–i94. https://doi.org/10.1093/bioinformatics/bts396 PMID: 22962457
14. Kim PJ, Price ND. Genetic co-occurrence network across sequenced microbes. PLoS Comput Biol. 2011; 7(12):e1002340. https://doi.org/10.1371/journal.pcbi.1002340 PMID: 22219725
15. Rivoire O. Elements of coevolution in biological sequences. Phys Rev Lett. 2013; 110(17):178102. https://doi.org/10.1103/PhysRevLett.110.178102 PMID: 23679784
16. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet. 2013; 14 (4):249–61. https://doi.org/10.1038/nrg3414 PMID: 23458856
17. Szurmant H, Weigt M. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. Curr Opin Struct Biol. 2017; 50:26–32. https://doi.org/10.1016/j.sbi.2017.10.014 PMID: 29101847
18. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A. 2009; 106(1):67–72. https://doi.org/10.1073/pnas.0805923106 PMID: 19116270
19. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A. 2011; 108(49):E1293–301. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262
20. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A. 2013; 110(39):15674–9. https://doi.org/10.1073/pnas.1314045110 PMID: 24009338
21. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28 (2):184–90. https://doi.org/10.1093/bioinformatics/btr638 PMID: 22101153
22. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. High-resolution protein complexes from integrating genomic information with molecular simulation. Proc Natl Acad Sci U S A. 2009; 106(52):22124–9. https://doi.org/10.1073/pnas.0912100106 PMID: 20018738
23. Hopf TA, Scharfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife. 2014; 3.
24. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife. 2014; 3:e02030. https://doi.org/10.7554/eLife.02030 PMID: 24842992
25. Rodriguez-Rivas J, Marsili S, Juan D, Valencia A. Conservation of coevolving protein interfaces bridges prokaryote-eukaryote homologies in the twilight zone. Proc Natl Acad Sci U S A. 2016; 113:15018–23. https://doi.org/10.1073/pnas.1611861114 PMID: 27965389
26. Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS. Inferring interaction partners from protein sequences. Proc Natl Acad Sci U S A. 2016; 113(43):12180–5. https://doi.org/10.1073/pnas.1606762113 PMID: 27663738

27. Gueudre T, Baldassi C, Zamparo M, Weigt M, Pagnani A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. Proc Natl Acad Sci U S A. 2016; 113(43):12186–91. https://doi.org/10.1073/pnas.1607570113 PMID: 27729520

28. Yeang CH. Identifying coevolving partners from paralogous gene families. Evol Bioinform Online. 2008; 4:97–107. PMID: 19204811

29. Cohen O, Ashkenazy H, Levy Karin E, Burstein D, Pupko T. CoPAP: Coevolution of presence-absence patterns. Nucleic Acids Res. 2013; 41(Web Server issue):W232–7. https://doi.org/10.1093/nar/gkt471 PMID: 23748951

30. Pagel P, Wong P, Frishman D. A domain interaction map based on phylogenetic profiling. J Mol Biol. 2004; 344(5):1331–46. https://doi.org/10.1016/j.jmb.2004.10.019 PMID: 15561146

31. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44(D1):D279–85. https://doi.org/10.1093/nar/gkv1344 PMID: 26673716

32. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse Statistical Physics of Protein Sequences: A Key Issues Review. arXiv preprint arXiv: 1703:01222. 2017.

33. Qin C, Colwell LJ. Power law tails in phylogenetic systems. Proc Natl Acad Sci U S A. 2018; 115 (4):690–5. https://doi.org/10.1073/pnas.1711913115 PMID: 29311320

34. Cocco S, Monasson R, Weigt M. From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. PLoS Comput Biol. 2013; 9(8): e1003176. https://doi.org/10.1371/journal.pcbi.1003176 PMID: 23990764

35. Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, et al. Systematic discovery of analogous enzymes in thiamin biosynthesis. Nat Biotechnol. 2003; 21(7):790–5. https://doi.org/10.1038/nbt834 PMID: 12794638

36. Myllykallio H, Lipowski G, Leduc D, Filee J, Forterre P, Liebl U. An alternative flavin-dependent mechanism for thymidylate synthesis. Science. 2002; 297(5578):105–7. https://doi.org/10.1126/science.1072113 PMID: 12029065

37. Herrmann KM. The shikimate pathway as an entry to aromatic secondary metabolism. Plant Physiol. 1995; 107(1):7–12. https://doi.org/10.1104/pp.107.1.7 PMID: 7870841

38. Sebastian M, Ammerman JW. The alkaline phosphatase PhoX is more widely distributed in marine bacteria than the classical PhoA. ISME J. 2009; 3(5):563–72. https://doi.org/10.1038/ismej.2009.10 PMID: 19212430

39. Okamura K, Hagiwara-Takeuchi Y, Li T, Vu TH, Hirai M, Hattori M, et al. Comparative genome analysis of the mouse imprinted gene impact and its nonimprinted human homolog IMPACT: toward the structural basis for species-specific imprinting. Genome Res. 2000; 10(12):1878–89. https://doi.org/10.1101/gr.139200 PMID: 11116084

40. Demaegd D, Colinet AS, Deschamps A, Morsomme P. Molecular evolution of a novel family of putative calcium transporters. PLoS One. 2014; 9(6):e100851. https://doi.org/10.1371/journal.pone.0100851 PMID: 24955841

41. Waters LS, Sandoval M, Storz G. The Escherichia coli MntR miniregulon includes genes encoding a small protein and an efflux pump required for manganese homeostasis. J Bacteriol. 2011; 193 (21):5887–97. https://doi.org/10.1128/JB.05872-11 PMID: 21908668

42. Shi HJ, Wu AZ, Santos M, Feng ZM, Huang L, Chen YM, et al. Cloning and characterization of rat spermatid protein SSP411: a thioredoxin-like protein. J Androl. 2004; 25(4):479–93. PMID: 15223837

43. Uguzzoni G, John Lovis S, Oteri F, Schug A, Szurmant H, Weigt M. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. Proc Natl Acad Sci U S A. 2017; 114(13):E2662–E71. https://doi.org/10.1073/pnas.1615068114 PMID: 28289198

44. Finn RD, Miller BL, Clements J, Bateman A. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. Nucleic Acids Res. 2014; 42(Database issue):D364–73. https://doi.org/10.1093/nar/gkt1210 PMID: 24297255

45. Greie JC. The KdpFABC complex from Escherichia coli: a chimeric K+ transporter merging ion pumps with ion channels. Eur J Cell Biol. 2011; 90(9):705–10. https://doi.org/10.1016/j.ejcb.2011.04.011 PMID: 21684627

46. Siegel AR, Wemmer DE. Role of the sigma54 Activator Interacting Domain in Bacterial Transcription Initiation. J Mol Biol. 2016; 428(23):4669–85. https://doi.org/10.1016/j.jmb.2016.10.007 PMID: 27732872

47. Vales LD, Rabin BA, Chase JW. Subunit structure of Escherichia coli exonuclease VII. J Biol Chem. 1982; 257(15):8799–805. PMID: 6284744

48. Abendroth J, Murphy P, Sandkvist M, Bagdasarian M, Hol WG. The X-ray structure of the type II secretion system complex formed by the N-terminal domain of EpsE and the cytoplasmic domain of EpsL of

Vibrio cholerae. J Mol Biol. 2005; 348(4):845–55. https://doi.org/10.1016/j.jmb.2005.02.061 PMID: 15843017

49. Strom MS, Nunn DN, Lory S. Posttranslational processing of type IV prepilin and homologs by PilD of Pseudomonas aeruginosa. Methods Enzymol. 1994; 235:527–40. https://doi.org/10.1016/0076-6879(94)35168-6 PMID: 8057924

50. Sakai D, Komano T. The pilL and pilN genes of IncI1 plasmids R64 and ColIb-P9 encode outer membrane lipoproteins responsible for thin pilus biogenesis. Plasmid. 2000; 43(2):149–52. https://doi.org/10.1006/plas.1999.1434 PMID: 10686134

51. Labahn J, Scharer OD, Long A, Ezaz-Nikpay K, Verdine GL, Ellenberger TE. Structural basis for the excision repair of alkylation-damaged DNA. Cell. 1996; 86(2):321–9. https://doi.org/10.1016/s0092-8674(00)80103-8 PMID: 8706136

52. Mielecki D, Grzesiuk E. Ada response—a strategy for repair of alkylated DNA in bacteria. FEMS Microbiol Lett. 2014; 355(1):1–11. https://doi.org/10.1111/1574-6968.12462 PMID: 24810496

53. Zhang Y, Inouye M. The inhibitory mechanism of protein synthesis by YoeB, an Escherichia coli toxin. J Biol Chem. 2009; 284(11):6627–38. https://doi.org/10.1074/jbc.M808779200 PMID: 19124462

54. Jorgensen MG, Pandey DP, Jaskolska M, Gerdes K. HicA of Escherichia coli defines a novel family of translation-independent mRNA interferases in bacteria and archaea. J Bacteriol. 2009; 191(4):1191–9. https://doi.org/10.1128/JB.01013-08 PMID: 19060138

55. Yu Z, Laven M, Klepsch M, de Gier JW, Bitter W, van Ulsen P, et al. Role for Escherichia coli YidD in membrane protein insertion. J Bacteriol. 2011; 193(19):5242–51. https://doi.org/10.1128/JB.05429-11 PMID: 21803992

56. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012; 40(Database issue):D841–6. https://doi.org/10.1093/nar/gkr1088 PMID: 22121220

57. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys. 2013; 87(1):012707. https://doi.org/10.1103/PhysRevE.87.012707 PMID: 23410359

58. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008; 24(3):333–40. https://doi.org/10.1093/bioinformatics/btm604 PMID: 18057019

59. Talavera D, Lovell SC, Whelan S. Covariation Is a Poor Measure of Molecular Coevolution. Mol Biol Evol. 2015; 32(9):2456–68. https://doi.org/10.1093/molbev/msv109 PMID: 25944916

60. Taboada B, Ciria R, Martinez-Guerrero CE, Merino E. ProOpDB: Prokaryotic Operon DataBase. Nucleic Acids Res. 2012; 40(Database issue):D627–31. https://doi.org/10.1093/nar/gkr1020 PMID: 22096236

61. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol. 2003; 4(9):R54. https://doi.org/10.1186/gb-2003-4-9-r54 PMID: 12952533

4

FILTER DCA

While in the previous chapter we have investigated the PPI network organization at the genome scale, here we shift our focus on the problem of predicting inter-protein residue-residue contacts.

The extension of DCA to inter-protein contact prediction is *in principle* straightforward (*cf.* Section 2.4.2). However, only moderate success [24, 30, 75, 76] has been achieved by DCA in this context. This is mainly because residues that make contacts across protein interfaces usually display a weak coevolutionary signal which cannot be easily detected by a global statistical model [57].

In the light of the astonishing success of Convolutional Neural Networks (CNNs) for intra-protein contact prediction (*cf.* Section 2.4.1 and Section 4.1), it is natural to ask if structure-based supervision can be used to improve inter-protein contact prediction. This is a challenging task since CNNs require to be trained on a large number of PDB structures which are not available for protein complexes [77]. Furthermore, CNNs rely on a large number of parameters and layers, making it difficult to comprehend what they are actually learning.

In this chapter, we introduce *FilterDCA*, a simple and interpretable supervised machine learning method, which can achieve results comparable to much more complex, data-hungry and hardly interpretable CNNs. Even if not out-performing these methods in applications, we think that interpretability is important to understand how the contact information is hidden in sequence data.

FilterDCA is based on the fact that the distribution of residue-residue contacts found in the proximity of other contacts shows characteristic patterns due to secondary structure, *cf.* Figure 4.2 and Section 4.2.3. Inspired by this consideration, we develop and benchmark a supervised classifier which allows incorporating patterns of secondary structures with DCA predictions.

The chapter is organized as follow. Our method was partially inspired by CNNs, therefore in Section 4.1 we explain the generic concepts behind them. In Section 4.2.1 we introduce the dataset we have used to develop and benchmark *FilterDCA*. In Sections 4.2.3 and 4.2.4, we explain how to compute typical patterns of contacts, and how to integrate them with DCA predictions. The performances of our predictor are presented in Section 4.3. Last, Section 4.4 contains discussion and outlook for future work.

4.1 INTRODUCTION ON CNN

As shown in Figure 4.1, the advent of coevolutionary methods - DCA or similar - allowed to significantly improve the prediction of residue-residue contacts. These methods are pure sequence-based, unsupervised method. This means that they take an MSA as input and predict residue-residue contacts by detecting coevolving residues. They never use the PDBs of experimentally solved protein in the inference procedure.

However, as in many other research fields, higher performance for the *intra-protein* contact prediction have been achieved with Deep Learning (DL) methods trained on proteins with experimentally solved structures. In particular, Convolutional Neural Networks (CNNs) [78], brought to impressive advancements in CASP12[1](*cf.* Figure 4.1). The last CASP13 even made headlines when it was won by AlphaFold, created by the industrial laboratory DeepMind [79].. Their approach led to what CASP organisers have called "unprecedented progress in the ability of computational methods to predict protein structure," but it is still built on two ideas developed in the academic community during the preceding decade: (i) the use of coevolutionary analysis, and (ii) structure-based supervision via CNNs [79, 80].

CNNs have been initially developed for object recognition tasks [82], and later found applications in other domains, such as object tracking, pose estimation, text detection and recognition [82]. They broke into protein structure prediction by treating contact prediction as pixel-wise classification problems. Indeed, if we consider a protein contact map as an image, then protein contact prediction is similar to pixel-level image labeling with classes "contact" or "not-contact".

How does deep-learning improve the accuracy of contact prediction? A comprehensive explanation is still lacking, due to the complex architecture of CNN, and to the huge number of interdependent parameters. However, it is clear that protein contact maps are not random matrices: distribution of contacts follows characteristic patterns due to secondary structure. Residue–residue contacts are generally not isolated and they display peculiar clusters (*cf.* Figure 4.2). In [83] and later [53], it was suggested that CNNs can automatically learn the most relevant patterns from the contact maps in the training set and later use them for constraining new predictions. If coevolutionary methods are capable of discovering some elements of the clusters, then CNNs increase the coverage and coherence of predicted contacts within these clusters [53], thus boosting the prediction accuracy.

1 CASP (Critical Assessment of protein Structure Prediction) is a competition for protein structure prediction based solely on a protein's amino acid sequence. Participants are invited to submit models for a set of proteins for which the experimental structures have been solved, but they are not yet public.

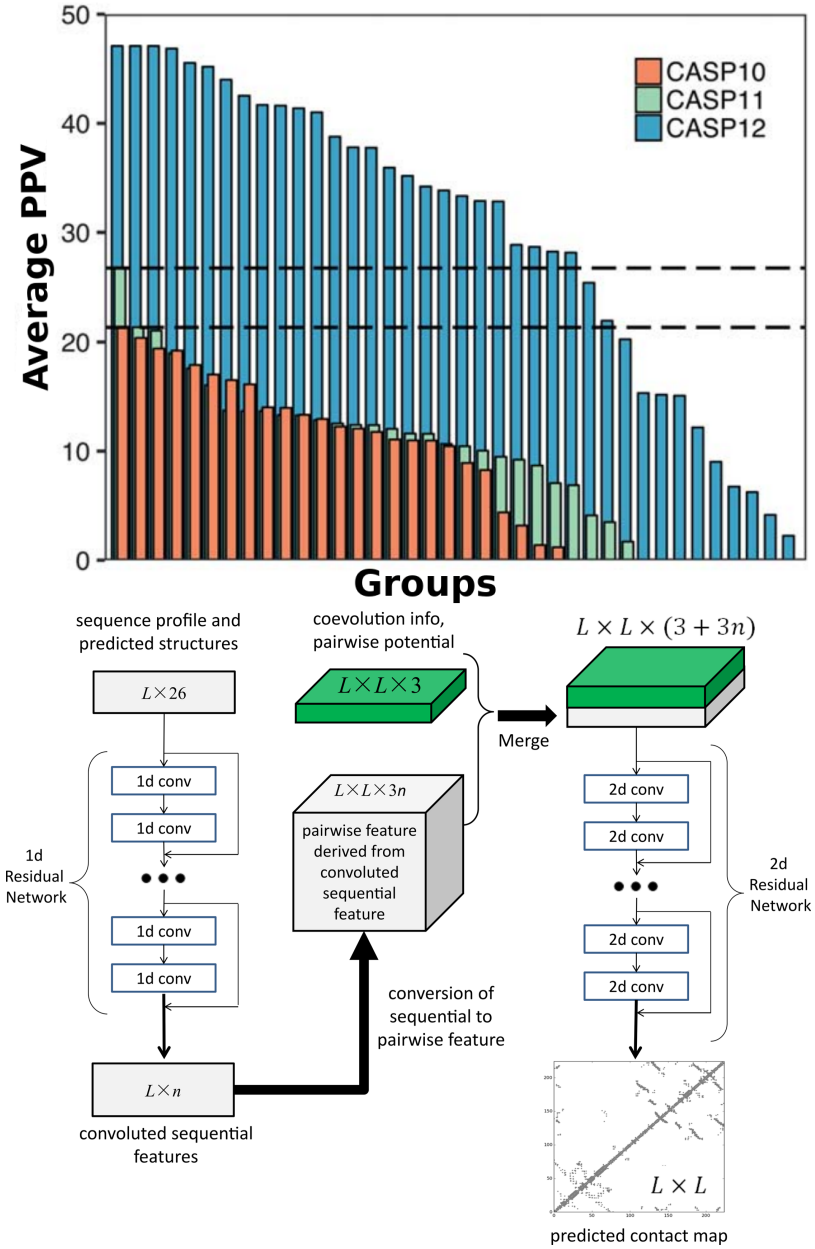

Figure 4.1: *On the top:* Average PPV on L/5 (with L being the length of the protein), normalized on 0-100 scale of long-range contacts in CASP10 (red), CASP11 (green), and CASP12 (blue) sorted by rank. One group (MetaPSICOV [81], based on coevolutionary methods) showed a significantly better average precision than all the others in CASP 11 compared to CASP10. In CASP12, 26 groups (the majority of which using coevolutionary analysis and deep learning techniques) showed an improved average precision compared to the best performing group of CASP11. Source: [52]. *On the bottom:* the architecture of the deep learning method RaptorX [78], winner of CASP 12. It employs dozens of hidden layers. The input features include protein sequence profile predicted 3-state secondary structure and 3-state solvent accessibility, coevolutionary information, mutual information and pairwise potentials. In the training set there are a total of 6767 protein structures.

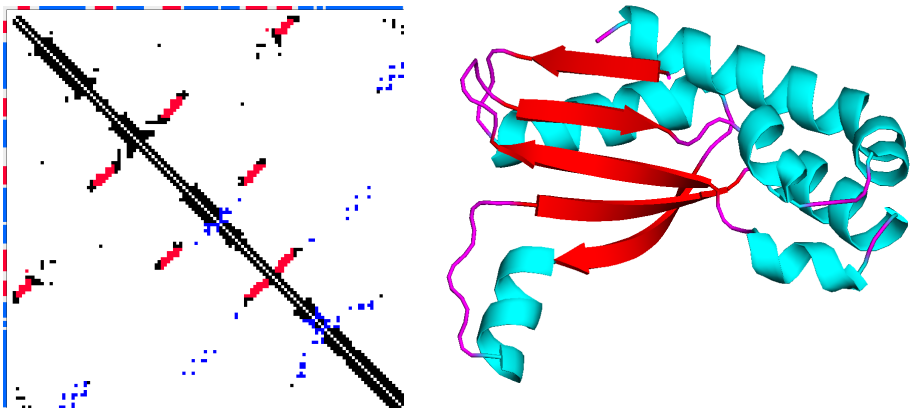

Figure 4.2: *On the left:* the contact map for the protein TT1751 from *T.thermophilus Hb8* (PDB code 1J3M). The contact map is a two-dimensional binary matrix representing the distance between all possible amino acid residue pairs. For two residues, i and j the entry (i, j) is 1 if the two residues i and j are in contact - the minimum atomic distances is < 8Å - and 0 otherwise. All contacts between two α-helix are displayed in blue, while the red points are contacts between extended β-strand. Characteristic patterns of contacts emerge due to secondary structure elements. *On the right*: The PDB 1J3M, with secondary structure elements depicted with different colors; red for β-strand, light blue for α-helix and purple for all residues that do not have easily observable regular patterns in their structure, referred to as "loops".

4.1.1 *Deep learning for inter-protein residue-residue contacts*

DL methods attained unprecedented accuracy levels in 3D structure prediction of individual proteins. However, the learning procedure is highly demanding on the number of PDB structures to learn from. This limits the generalization of those methods to the assembly of complexes of interacting proteins. Solving the 3D structure of a protein complex by experimental techniques remains indeed very challenging [77]. The resulting scarcity of co-crystallized structures poses an essential problem to deep learning approaches and explains why much fewer methods are dedicated to inter-protein contact prediction.

A possible solution, adopted by *RaptorX Complex* [77], is to do "transfer-learning", meaning to train a CNN with intra-protein contacts and then apply it to inter-protein contact prediction .

4.2 FILTER DCA: METHODS

Inspired by this discussion on CNNs, we aim at gaining understanding on the properties of the typical contact patterns of inter-protein contacts, and to employ the new insights to improve the DCA predictions. To develop and benchmark our approach, we decided to focus on interactions between domains in single multi-domain proteins. Protein domains are autonomous folding units, thereby domain-domain interaction can be thought as a proxy for protein-protein interaction.

4.2.1 *Dataset*

A database of 3D interacting domains, called *3did*, was realized by Stein and collaborators [84]. They selected all proteins with a high-resolution 3D structure which contains multiple PFAM domains. Then they computed the number of residue-residue contacts between pairs of contacting domains either within the same chain (intra-chain) or between two different chains (inter-chain). We use the Aug 5, 2017 version which is based on Pfam v.30.0 and contains a total of 11.200 structurally resolved domain-domain interactions. To get the joint MSA we exclude homodimeric cases, and match sequences of domains co-localized on the same protein chain, i.e. we consider exclusively intra-protein inter-domain interactions, see Figure 4.3. Finally, we map each residue in the MSA to the corresponding positions in the PDB. The mapping is done by aligning the PDB sequences to the profile HMMs of the PFAM domain through *hmmalig*n introduced in Section 1.3.1, and allows to associate residue-residue distances to any pair of alignment columns.

In case the same residue-residue pair is associated with multiple PDBs, we assign the minimum distance between all possible copies. This assumes that any predicted contact, which is present in at least

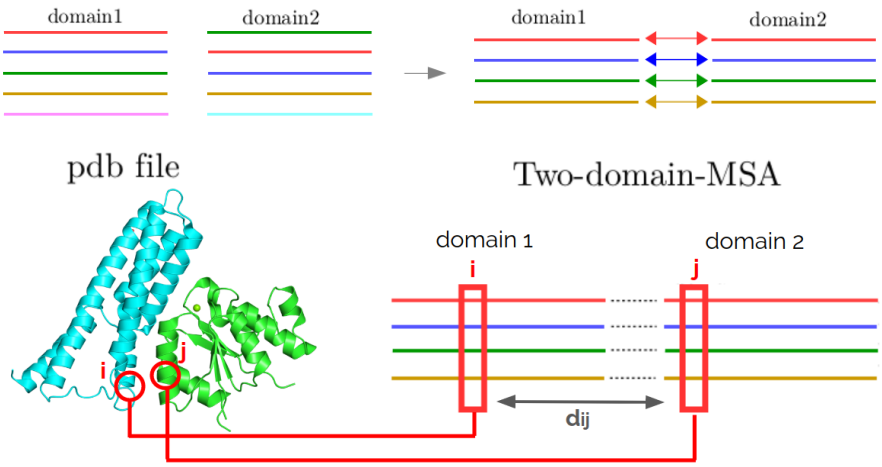

Figure 4.3: *Construction of the domain-domain interaction datasets.* We extract a list of the interacting domains from *3did* [84] and construct a joint MSA by matching domains co-localized on the same protein chain (upper left: domains inside the same protein chain are symbolized by identical color, upper right: the joint MSA). We then map each residue in the MSA to the corresponding positions in the PDB to calculate the 3D distance between pairs of residues (minimum distance between heavy atoms). In case multiple PDBs can be associated with the same residue-residue pairs, we keep only the minimum distance over all PDBs.

one PDB structure, is a true positive prediction. Often only a part of the MSA can be mapped on the corresponding PDBs. We keep only MSAs with domain mapping coverage greater than or equal to 40. We further clean our dataset by requiring at least 10 and at most 2000 residue-residue interactions and removing few cases of coiled-coil structures which, due to repeated motifs, can lead to spurios coevolutionary signal. At the end, we keep a total 2548 joint MSA of pairs of contacting domains, on which we run PLM [45, 46].

4.2.2 *DCA performance on data*

In Section 2.3 we introduced M_{eff}, which is defined as the effective number of independent sequences. It is considered to be an indicator of the accuracy of the DCA predictions - the higher the better (Figure 4.4). DCA predictions are comparable only for MSA having similar M_{eff}. Thereby, we decide to split the 2598 MSA of contacting domains in 3 datasets according to M_{eff}, see Table 4.1, and to analyze them independently.

First, we study the performance of PLM on the three datasets by computing the mean Positive Predictive Value (PPV), Figure 4.4. As expected, a larger M_{eff} leads to better predictions. The mean PPV of cases with $M_{\text{eff}} < 50$ is close to that of a random predictor, meaning that almost no coevolutionary signal is contained in this dataset. This

	$M_{\text{eff}} > 200$	$50 < M_{\text{eff}} \leq 200$	$1 < M_{\text{eff}} \leq 50$
Num joint MSA	842	758	998
Num contacts	274587 (1,7%)	193936 (1,3%)	204752 (1,1%)

Table 4.1: We split the 2598 MSAs of interacting domains in 3 datasets according to the effective number of independent sequences, M_{eff}. They contain approximately the same number of MSAs (first row) and inter-domain contacts (second row). In brackets, the percentage of inter-domain contacts is given with respect to the total number of inter-domain residue pairs.

becomes evident also from the distributions of the DCA scores for contacts and non-contacts, *cf.* Figure 4.5. For cases with $M_{\text{eff}} > 50$ the two distributions are clearly different and dominated by contacting pairs for DCA scores larger than 0.3. For MSAs with $M_{\text{eff}} < 50$, the two distributions completely overlap.

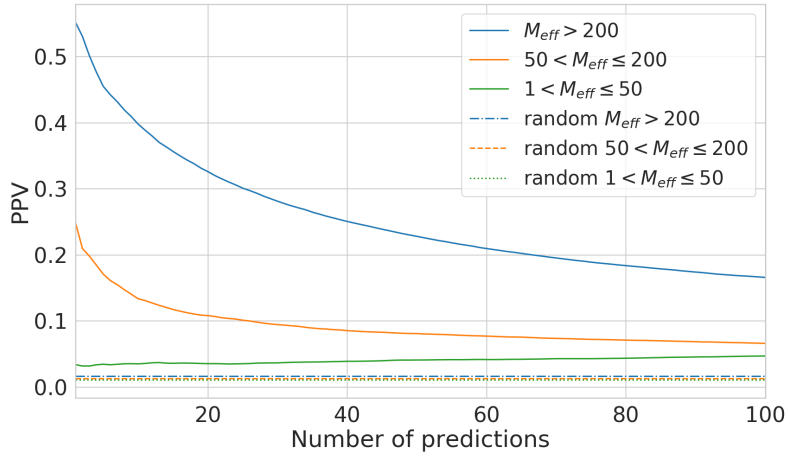

Figure 4.4: *Mean* PPV *using only* DCA *score for the three datasets.* For MSA with $M_{\text{eff}} < 50$, the mean PPV is close to the PPV of a random predictor, meaning almost no coevolutionary signal is contained.

4.2.3 *Secondary structure and contact patterns*

Similarly to the intra-protein case, detailed in Section 4.1, the neighborhood around an inter-domain contact is informative for contact determination. Contacts between secondary structure elements form characteristic patterns in inter-domains contact maps, which we aim to identify and exploit to improve contact prediction. We use the *DSSP* algorithm [85, 86] to assign secondary structure to each amino acids in our dataset. *DSSP* identifies 7 possible types of secondary structure, but in order to reduce the complexity of the problem we convert this result into a 3-letter alphabet: H of all type of helices, E for all type of

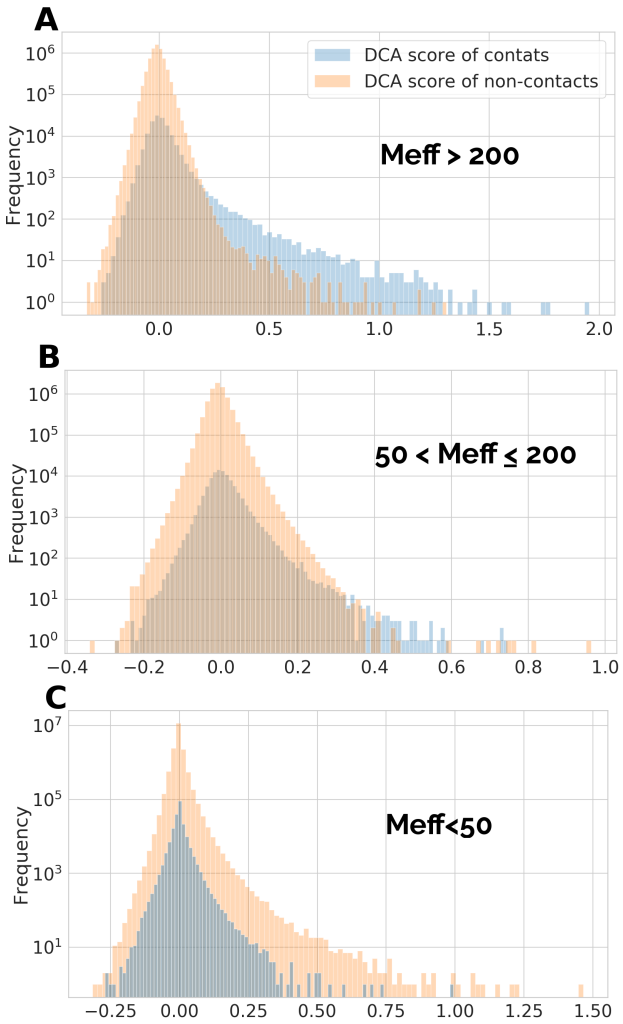

Figure 4.5: *The distribution of* DCA *score for the three datasets.* For MSA with $M_{\text{eff}} > 200$ (Panel A) and $50 < M_{\text{eff}} \leq 200$ (Panel B). Note that the enrichment of true positive predictions (contats) is very high in the tail of large DCA score. In fact, the majority of the pairs with DCA score larger than 0.3 corresponds to contacts. This is not the case for MSA with $M_{\text{eff}} < 50$ (Panel C) where the two distributions completely overlap.

β-strands and C for everything else. The conversion table is detailed in Table 4.2.

H	α-helix	$\rightarrow$ H
B	residue in isolated β-bridge	$\rightarrow$ E
E	extended strand, participates in β ladder	$\rightarrow$ E
G	3-helix (3_{10} helix)	$\rightarrow$ H
I	5-helix (π-helix)	$\rightarrow$ H
T	hydrogen bonded turn	$\rightarrow$ C
S	bend	$\rightarrow$ C

Table 4.2: The conversion table from the *DSSP* alphabet for secondary structure to our simplified alphabet three-letter alphabet.

	H	E	C
H	40705	13046	41274
E	13046	13851	20971
C	41274	20971	62100

Table 4.3: The number of residue-residue inter-domain contacts in our full dataset classified according the secondary structure.

As we consider 3 possible secondary structures for a residue; there are 6 possible different states for an inter-domain contact: HH, HE, HC, EE, EC, CC, see Table 4.3. Henceforward, we consider explicitly only the HH and EE cases which give rise to the most characteristic contact patterns. Figure 4.6 shows the mean contact matrix where the central pair is an HH and EE contact, obtained by averaging over all PDBs of our dataset and using a 7×7 window.

Figure 4.6 shows that the distribution of neighborhood contacts surrounding the central pair is not random: for instance if residue i^A and j^B are in contact and both belong to α-helices, then residue $i^A + 2$ and $j^B + 2$ most likely will not be in contact. The EE mean contact map is apparently less informative. In fact, contacts between two α-helix or two β-strand can both be parallel, anti-parallel or mixed. To disentangle them, we perform a k-means clustering imposing 3 clusters. Let us call $\mathcal{S}$ the set of the 6 resulting centroids (3 for HH and 3 for EE) - hereafter *filters*. Figure 4.7 shows an example using a 21×21 window.

The DCA predictions show slightly similar patterns. For each pair of interacting domains A and B of length n and m, we define a DCA score-matrix $D = (F_{i^A j^B})$ of size $n \times m$, whose entries are the *DCA score*, i.e. the APC corrected Frobenius norm Eq. (2.48) of amino acids i^A and j^B belonging respectively to the first and second domain.

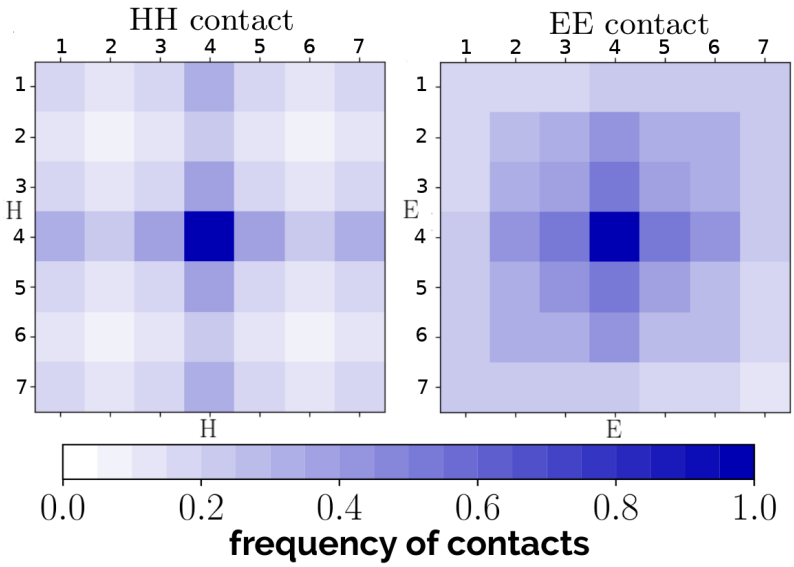

Figure 4.6: *Patterns of secondary structure in PDBs*. The relative frequencies of the number of contacts in a 7×7 mean contact map around a HH or EE contact. The average is done over the 40705 HH and 13851 EE contacts, *cf.* Table 4.3.

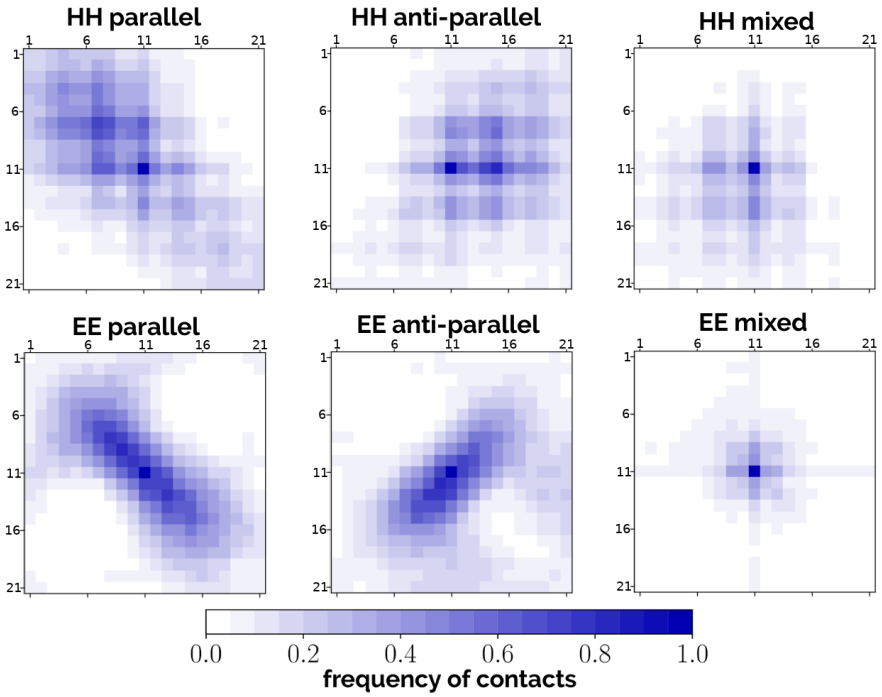

Figure 4.7: *Set S of filters*. The mean contact matrices are a combination of parallel, anti-parallel or mixed HH and EE contacts (here computed using a 21×21 window). We disentangle them with a k-means clustering with $k = 3$. The 6 resulting centroids, 3 for a central HH inter-domain contact (upper figure) and 3 for a EE contact (lower figure), are the set S of *filters* that we will be using in the following.

Then, similarly to what we have done in Figure 4.6, we compute the mean DCA score matrix with the central position being an inter-domain contact in the PDB structure. They are displayed in Figure 4.8. By considering the mean DCA matrix, we average out site-specificities and noise and, as a result, only the most frequent local contact structures are prominently displayed. On a single DCA matrix, the secondary structure signal is usually lower and noisy. Our method is based on the idea that we can compute typical patterns between secondary structures from the PDB and subsequently apply them on the DCA matrix for constraining the prediction of nearby contacts by coherency with predictions in a local window around the residue pair of interest.

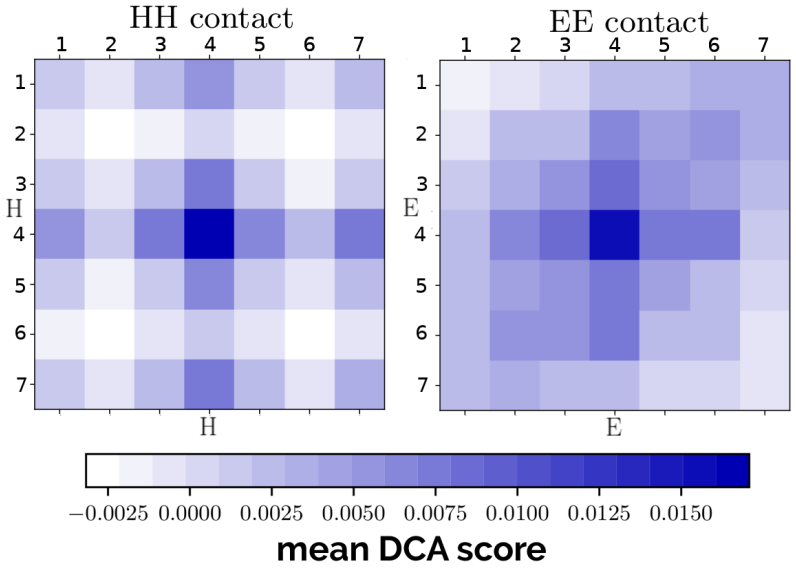

Figure 4.8: *Patterns of secondary structure in DCA predictions*. Patterns of secondary structure that arise from the mean DCA matrix. The average is done over the 2548 joint alignments contained in our dataset.

4.2.4 *Filter score*

Inspired by this discussion, we define a new score by applying patterns between secondary structures on DCA predictions, *cf.* Figure 4.9. Let $D = (F_{ij})$ be the matrix of size $n \times m$ of inter-domain DCA scores and $\mathcal{S}$ the set of the 6 filters of size k. For each pair (i, j) of D, let D_k be the DCA submatrix of size k centered in (i, j): $D_k = (F_{i'j'})$ with $i' \in [i - \frac{k-1}{2}, i + \frac{k-1}{2}]$ and $j' \in [j - \frac{k-1}{2}, j + \frac{k-1}{2}]$. We always choose k to be an odd number in order to get a square matrix centered around the central pair.

For each of the 6 filters in $\mathcal{S}$ of size $k \times k$, we compute the Pearson correlation between D_k and the filter. The central pair is removed from the calculation since, for the filters in $\mathcal{S}$, it is a contact by construction

and since the DCA score of the central pair (i, i) will be used directly. The new score, hereafter *Filter score*, is the maximum between the 6 Pearson correlations.

A problem arises from pairs of residues closer than $\frac{k-1}{2}$ to the border of the DCA matrix: the matrix D_k is smaller than the filter matrix, and the procedure displayed in Figure 4.9 can not be applied. In this case we compute the Pearson correlations only for pairs which are both contained in D_k and in the filter matrix.

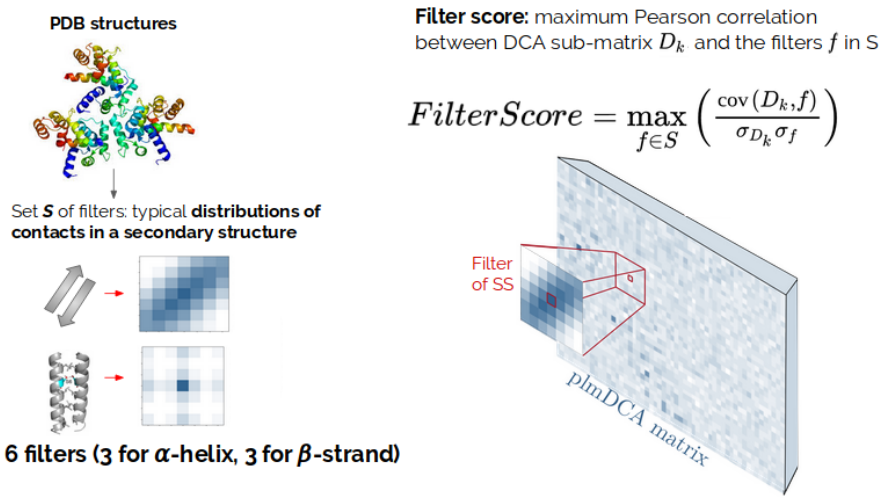

Figure 4.9: *Pipeline to compute the Filter score*: we first fix a window size k and compute the set $\mathcal{S}$of 6 filters of size k. Then for each pair (i, j) of D, we consider D_k the DCA submatrix of size k centered in (i, j). We remove the central pair (i, i) and then we compute the pearson correlation between each filter f in $\mathcal{S}$ and D_k. The *Filter Score* is the largest of the 6 correlations.

4.2.5 *Learning procedure*

From Section 4.2.2 it is clear that MSAs with $M_{\text{eff}} < 50$, do not show reliable coevolutionary signal. Consequently, we exclude them from the following analyisis. For each of the two remaining datasets, we divide the data into training and test sets, with a $50 - 50$ split. We first compute the local contact maps around a HH or EE contacts for all PDBs belonging to the training set. We compute the 6 filters by performing a k-means clustering with $k = 3$ (parallel, antiparallel, and all the rest). An important parameter to fit is the filter matrix size. Since we can not determine a priori the optimal size, we train different models with filter sizes ranging from 5 to 69.

For each pair of residues (i, j), we consider two features: x_1 the *DCA score* of a residue pair (i, j), and x_2 the *Filter score* of the DCA matrix around (i, j), which we use to train a logistic regression classifier. The

probabilities to belong to the class *contact* ($\oplus$) or *non-contact* ($\ominus$) are given by:

$$P(\oplus|\mathbf{x}) = \frac{e^{\mathbf{w}\mathbf{x}+w_0}}{1+e^{\mathbf{w}\mathbf{x}+w_0}} \qquad \text{and} \qquad \mathrm{P}(\ominus|\mathbf{x})=\frac{1}{1+e^{\mathbf{w}\mathbf{x}+w_0}} \tag{4.1}$$

where bias w_0 and the weights $\mathbf{w} = (w_1, w_2)$ are optimized using the 'liblinear' solver of the *sklearn* library [87].

Pairs of residues forming a contact are only a small fraction of all possible pairs (see table 4.1). Thus, the training set is strongly imbalanced: the incidence of class *non-contact* is dominant, being found in 99% of cases. We found that the performance of our classifier is improved when we restrict the training set on residues pairs with DCA score larger than zero (see Figure A.13 in the Appendix). In such a way, the classifier concentrates on cases which show a more reliable coevolutionary signal. Another further slight improvement has been achieved by scaling the Filter Scores, in both training and test set, with the MinMaxScaler of the *sklearn* library [87]:

$$x_2 \rightarrow \frac{x_2 - min(\mathbf{x_2})}{max(\mathbf{x_2}) - min(\mathbf{x_2})} \tag{4.2}$$

where $max(\mathbf{x_2})$ [$min(\mathbf{x_2})$] is the maximum [minimum] Filter Score in the training set.

4.2.6 *Filter size*

To further analyze the importance of the filter size on the quality of the prediction, in Figure 4.10 we plot the decision boundary of each logistic regression. The decision boundary is the line $\mathbf{w}\mathbf{x} + w_0 = 0$ - or, in other words, $P(\oplus|\mathbf{x}) = P(\ominus|\mathbf{x}) = 1/2$ - which partitions the feature space into two sets, one for each class. The logistic regression will classify all the points above decision boundary as *contact* and all those on the other side as *non-contact*.

The lines of Figure 4.10 have slope $-w_2/w_1$. For small filters, such as 5×5, w_2 is close to 0. This means that there is no performance improvement adding the Filter Score: the DCA score is the only variable used by the classifier for the predictions. Note that in this case the classifier assigns $P(\oplus|\mathbf{x}) > 0.5$ for pairs with DCA. A score larger than 0.3, coherently with what observed in Figure 4.5. Small filters do not show clear patterns between secondary structures, thereby they can not be used for constraining the predictions.

Increasing the filter size, the weight w_2 grows until a maximum is reached at $k = 45$. Beyond this size, filters lead to spurious signal; the Filter Score becomes less meaningful and, consequently, w_2 decreases again.

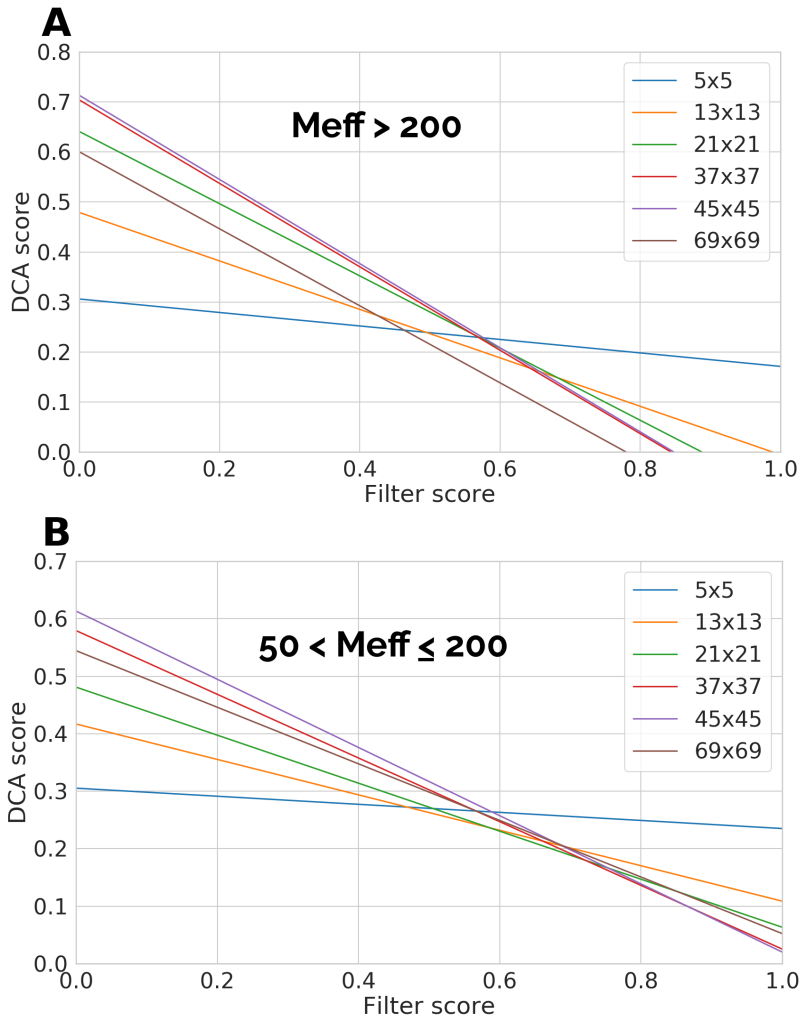

Figure 4.10: *Decision boundary of the logistic regressions.* Panel A and B show the decision boundary, the line $\mathbf{wx} + w_0 = 0$, for the two training sets, $M_{\text{eff}} > 200$ and $50 < M_{\text{eff}} \leq 200$ respectively. In both cases we see that the importance of the second feature - the Filter Score - is increasing with the filter size. Small filters, like 5×5, do not show clear patterns of secondary structure. Therefore, the Filters score is essentially neglected by the logistic regression which classifies as "contact" pairs with DCA score larger than 0.3, coherently with Figure 4.5. Filters bigger than 45×45 lead spurious signal, thus the classifier assigns a smaller weight to the Filter score.

4.3 RESULTS

To access the performance of our method we use the test sets. There we sort pairs of residues according to the estimated probabilities for each MSA. Figure 4.11 shows the mean PPV for cases with $M_{\text{eff}} > 200$ and $50 < M_{\text{eff}} \leq 200$.

FilterDCA is a supervised Machine Learning (ML) method which uses two features - the DCA score and the Filter score. Hence, it is important to compare its performance with at least two other methods. First, PLM [45, 46], which is the reference algorithm. It is unsupervised and sort pairs according to the DCA score, i.e. according to the first of the features we used. Second, we use a DL algorithm, Pconsc4[88], which is based on a far more complicated and thus less transparent parameterization as compared to the logistic regression, Eq. (4.1). We could not fairly compare our results with *RaptorX Complex* [77] since the latter is trained on all single-chain proteins available on the PDB, thereby making impossible to avoid overfitting when studying intra-protein inter-domain contacts (we discuss this issue in Section A.2).

Figure 4.11 shows that FilterDCA always outperforms PLM, proving that the filters add useful information to the local value of the DCA score F_{ij}, Note that the performance of our predictor with small filters (size 5×5) is only slightly better than PLM, as expected in the light of the above discussion. Larger filters lead to better performance, and the classifier reaches the best performance - comparable to that of PconsC4 - with astonishingly large filters 37×37, i.e. looking a 13 positions before and after position i and j.

4.4 CONCLUSION AND DISCUSSION

Despite some success [24, 30, 75, 76], the accuracy of DCA for inter-protein contact prediction remains limited, mainly because the interface coevolutionary signal is weak and difficult to detect.

Supervision by structure could *in principle* be used to boost contact prediction from weak coevolutionary signals. However, deep learning algorithms are highly demanding on the number of structures to learn from, and the scarcity of experimentally solved protein complexes poses an essential problem. Furthermore, they rely on a large number of parameters and layers, which makes it obscure to comprehend in details what they are learning.

In this chapter, we proposed FilterDCA which at the same time does not rely on an extensive training set, and it is easily interpretable since it requires to fit only 3 parameters with respect to thousands or even millions of parameters of deep learning methods.

The idea behind FilterDCA is, in some sense, opposite to the one behind CNNs: instead of letting a CNN automatically learn the most

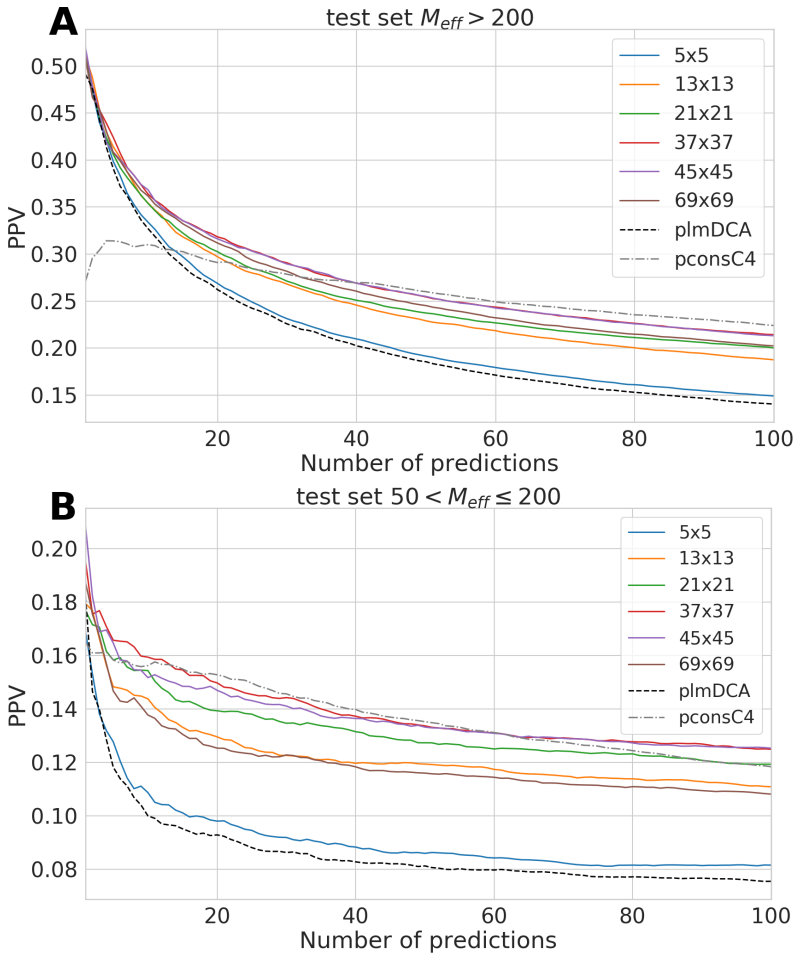

Figure 4.11: *Performance of FilterDCA on the test set.* The mean positive predictive value for the tests sets with $M_{\text{eff}} > 200$ (Panel A) and $50 < M_{\text{eff}} \leq 200$ (Panel B) as a function of the number of predictions. The average was performed over the 412 and 379 Pfam domains of the respective tests sets. For small filters, the performances are comparable to unsupervised PLM. They progressively increase using larger and larger filters. The best performance, comparable to that of PconsC4, is reached with filters of size 37.

relevant patterns from an extensive training set, we exploit our prior biological knowledge about protein structure, and compute patterns between HH and EE contacts directly from the PDB. Via the Filter score, we incorporate them into the DCA predictions. More precisely, we combine the DCA and Filter score with a logistic-regression, which is a simple, easy to interpret model. Our results indicate that large patterns between secondary structures enhance significantly the performance of DCA for inter-domains contact prediction. Our classifier reaches performance comparable to that of PconsC4 with filters 37×37. Even if not out-performing DL methods in applications, we think that simple and interpretable models, like *FilterDCA*, can be useful to understand how the contact information is hidden in sequence data.

So far, we considered exclusively intra-protein inter-domain contacs. A natural extension would be to assess the performance of FilterDCA for inter-protein contact prediction. We then plan to apply out method on the dataset composed of 36 bacterial protein complexes of [30] for which GREMLIN [51], which is equivalent to PLM, has previously been shown to yield accurate predictions.

Also, we aim at understanding if FilterDCA can improve the accuracy of docking prediction, thereby be useful to assemble unknown protein complexes. To this end, we need to find out if FilterDCA is increasing the coverage of predicted contacts between secondary structure elements or if it is removing spurious isolated contacts relative to DCA predictions.

Part III

FITNESS LANDSCAPE

In this final chapter, we show how DCA can be used to model the fitness landscape. Through an extensive genome-wide study of *E.coli* strains, we investigate the fitness landscape properties at the local and global scale. First, we quantify the strength of the epistatic interaction within genes of *E. coli* strains. Second, we introduce the context-dependent entropy to quantitatively and qualitatively characterize how the variability of a residue is influenced by the sequence context, i.e. the amino acids present in all other positions of the protein.

5

LOCAL FITNESS LANDSCAPE

In the very last chapter of this dissertation, we will introduce the topic of fitness landscapes. We will show how the machinery of DCA can be used to investigate the connection between global and local fitness landscape. To this end, we performed a genome-wide analysis of a set of 61160 *E. coli* strains to address two different questions:

(I) Do recently diverged strains exhibit intragenic epistatic interactions between mutations, e.g. via the compensation of effects of the single mutations?

(II) Can we quantitatively and qualitatively characterize how the variability of a residue is influenced by its context, i.e. by the amino acids present in all other positions of the protein?

The chapter is organized as follow. Section 5.1 introduces the concept of global and local sampling of the fitness landscape in a way suitable to our presentation. In Section 5.2 we briefly described the findings of [58, 89] which explored the connection between DCA and the fitness landscape. We then introduce the dataset we used for our analysis (Section 5.3) and the results we obtained on epistasis (Section 5.4) and context-dependency (Section 5.5). Last, Section 5.6 contains discussion and outlook for future work.

5.1 FITNESS LANDSCAPE

In essence, the fitness landscape is a genotype-phenotype mapping, which associates a quantitative phenotype $\Phi(a_1, ..., a_L)$ to each possible amino-acid sequence $(a_1, \ldots, a_L)$. The fitness landscape was initially proposed as "a metaphor" [90], a simplification that makes evolution tractable and potentially predictable. The fitness landscape may not always an appropriate descriptive tool for all biological systems (for example, it is well known that coevolving ecosystems can be characterized as multi-player games [91] and the fitness landscape model would not be applicable).

However, within this framework, one can visualize evolution as the process of sampling sequences in the fitness landscape. Natural selection prunes dysfunctional sequences, while amplifying those that have large fitness compared to other protein sequences, i.e. perform their function efficiently. However, a characterization of the fitness landscapes appeared infeasible for several reasons. First, the number of possible genotypes is astronomically high making impossible an exhaustive experimental measurement of the fitness for each variant.

Second, the pervasiveness of epistasis - the phenomenon by which the effect of a mutation depends on its genetic background - which may lead to a rugged landscape with many local optima. The advent of deep mutational scanning, i.e .the large-scale experimental determination of the fitness effects of mutations around the same wild-type sequence, offers the opportunity to exhaustively experimentally explore a local region of the fitness landscape. The Wright's idea of genotype-fitness landscapes started to move from a metaphor to an object of quantitative experimental studies.

To what extent DCA can yield insight about the fitness landscape? To answer this question one needs to keep in mind that the proteins in an MSA are the results of the evolutionary pathways, which are constrained by the shape of the fitness landscape - its set of hills, valleys, plains and ridges. If the DCA model can actually grasp the evolutionary constraints contained in the MSA, it is not surprising that is can be used to model the underlying fitness landscape.

Due to the high dimensionality of the genotype space, our knowledge of the landscape is restricted to the subspace of sequences - either naturally occurring or artificially generated - that have been sampled and, in rare cases, phenotypically characterized. We can distinguish:

1. *Global sampling of the landscape.* It contains homologous sequences present across different species. Being widely-distributed throughout the sequence space, they are strongly divergent (typical sequence identities are from 20 to 30%). The exact value of their fitness is usually unknown. However, being present in nature, they are folded and functional proteins. We can therefore assume that they are located near a local maximum of the fitness landscape, *cf.* the blue dots in the schematic representation of Figure 5.1.

2. *Local sampling of the landscape.* It contains orthologous sequences belonging to different strains of the same species. The genomic context is highly conserved. Indeed, they usually display very low variability ($>$ 90% sequence identity). As above, they are folded and functional even if the exact value of their fitness is usually unknown. We then expect strongly deleterious mutations to be absent, *cf.* the green dots in Figure 5.1.

3. *Exhaustive local quantification.* Recently developed deep-sequencing approaches, like deep mutational scanning [18], allow for an exhaustive quantification of a tiny region of the fitness landscape, *cf.* the red line in Figure 5.1. A deep mutational scan proceeds by identifying a reference sequence (the so-called *wild type*) and then generating thousands of sequence variants which are one or two mutations away from the wild type. Mutations that enhance protein activity are enriched following an appropriate selection or experimental screen, whereas deleterious mutations are de-

pleted. The enrichment ratio for each sequence variant is a proxy for the protein fitness.

Since few sequences of the space can be explored, there is a need of statistical modeling of the fitness landscape. The strategy adopted by DCA is to start from a *global sampling* (1.) of homologous proteins, arrange them in the MSA, infer the parameters $\mathbf{h}$ and $\mathbf{J}$ of an epistatic Potts model $P(a_1, \ldots, a_L) \sim \exp\big(-H(a_1, \ldots, a_L)\big)$, and use the statistical DCA energy $H(a_1, \ldots, a_L) = -\sum_i h_i(a_i) - \sum_{i<j} J_{ij}(a_i, a_j)$ as a proxy for the mathematical form of the fitness function (low energy sequences are considered more likely to be functional, and thus of high fitness), see Figure 5.1. The underlying assumption of this approach is that the natural sequence variability between homologous proteins can be used to model the fitness landscape. This is far from obvious: homologous sequences are strongly divergent and sparsely distributed throughout the sequence space, so it is unclear to what extent such statistical model can be used to make predictions about the observable sequence statistics in the *local sampling* (2.) or the fitness in a *deep mutational scan* (3.). To be more specific, it is not obvious that $10^3 - 10^4$ sequences at low average sequence identity of $20 - 30\%$ can provide quantitative information about the effect of a single mutation in any specific sequence belonging to the same protein family.

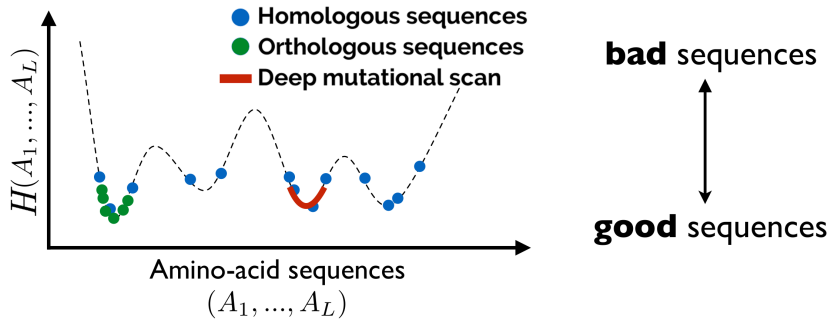

Figure 5.1: DCA models the fitness landscape via the statistical energy $H(a_1, \ldots, a_L) = -\sum_i h_i(a_i) - \sum_{i<j} J_{ij}(a_i, a_j)$. The parameters $\mathbf{h}$ and $\mathbf{J}$, are inferred from a set homologous sequences (blue dots) which are well-distributed throughout the sequence space. Therefore, it is unclear to what degree DCA can accurately predict the fitness values of orthologous sequences (green dots) or deep mutational scan (red line).

5.2 DCA AND FITNESS LANDSCAPE

A connection between the *global sampling* and *exhaustive local quantification* was established by Figliuzzi et al. [58] by proving that the predictions of a DCA model inferred from the homologous Pfam family beta-lactamase TEM-1 (PF13354) are significantly correlated with the

experimental results of a deep mutational scan. Further studies confirmed and expanded these results [58, 61, 62]. Sequences generated by deep mutational scan are confined in a tiny region of the fitness landscape, thereby scoring mutations can be thought of as a local measurement of the fitness landscape. It has been shown (see Figure 5.2) that the negative log odds ratio:

$$\Delta H^{mut} = -\log\left(\frac{P(\mathbf{a}^{mut})}{P(\mathbf{a}^{wt})}\right) = H(\mathbf{a}^{mut}) - H(\mathbf{a}^{wt}) \tag{5.1}$$

can successfully assess the experimental fitness of the variant $\mathbf{a}^{mut}$ with respect to the wild-type sequence $\mathbf{a}^{wt}$, when $\mathbf{a}^{mut}$ and $\mathbf{a}^{wt}$ differ by a single substitution. Negative scores correspond to predicted beneficial mutations, meaning that mutated sequences are more probable than the reference one, whereas positive scores correspond to predicted deleterious mutations.

In Couce et al. [89] DCA was used to investigate the connection between *global sampling* and *local sampling*. They considered a set of *E. coli* strains obtained from three different sources with distinct selective regimes: naturally occurring sequences (low mutation rate, high selection pressure), mutation accumulation experiments (MAEs) (high mutation rate, weak selection), and Long-Term Evolution Experiment (LTEE) (high mutation rate, strong selection). It has been shown that the DCA scores of single-point mutations observed in *E. coli* strains, Eq. (5.3), can be used to discriminate different evolutionary scenarios, *cf.* Figure 5.2.

It is important to point out that DCA constantly overcomes profile models both in predicting mutational effects and in discriminating selective regimes [58, 89]. The reason is to be searched in the epistatic couplings $\mathbf{J}$ of the Potts model. It allows DCA, contrarily to the Profile model, to assess the dependence of mutational effects on the sequence context where they appear. As an example, a single-point mutations substituting the wild-type amino acid a_i at position i with amino acid b, it is scored:

$$\Delta H^{\mathbf{Profile}}(a_i \to b | a_1, \ldots, a_{i-1}, a_{i+1}, a_L) = h_i(b) - h_i(a_i) \tag{5.2}$$

$$\Delta H^{\mathbf{DCA}}(a_i \to b | a_1, \ldots, a_{i-1}, a_{i+1}, a_L) = \tag{5.3}$$

$$= h_i(b) - h_i(a_i) + \sum_{j \neq i}^{L} \Big(J_{ij}(b, a_j) - J_{ij}(a_i, a_j) \Big).$$

In the profile model $h_i(a_i) = \log(f_i(a_i)) + const$ (see Section 2.1.1). This implies that the mutational score depends only on the difference between the logarithm of the frequencies of the original and the new amino acid in the MSA.

On the contrary in the DCA model, the term $\sum_{j \neq i}^{L} \Big(J_{ij}(b, a_j) - J_{ij}(a_i, a_j) \Big)$ of Eq. (5.3) explicitly connects the mutated site i to the rest of the se-

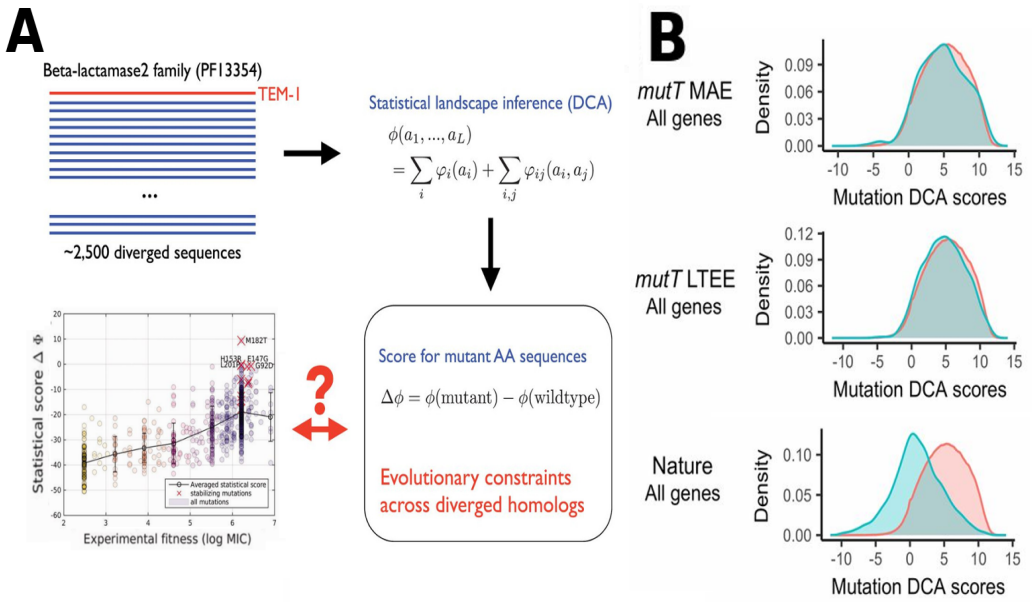

Figure 5.2: *Panel A)* From *global sampling to exhaustive local quantification*. The homologous Pfam family PF13354 is used to fit a DCA model which, in turn, is used to score mutations between the mutant and the wild-type amino acid sequence. A correlation of 74% has been found between experimental large-scale mutagenesis data and the prediction of the epistatic DCA model. Please note that the statistical score $\Delta\Phi$ of Panel A is defined with opposite sign with respect to our definition Eq. (5.1). Source: [58]. *Panel B)* From *global sampling to local sampling*. A set of *E. coli strains* was obtained from three different sources: Long-Term Evolution Experiment (LTEE) where a strain of *E. coli* evolved for more than 60,000 generations in a minimal glucose-limited medium [92], mutation accumulation experiments (MAEs) , in which bacteria are repeatedly propagated through single-cell bottlenecks [93, 94], and naturally occuring variants. In [89], for each *E. coli* protein, the domain architecture was extracted from Pfam, and a DCA model was fitted using the corresponding Pfam MSA. They compared the negative log odds ratio of single-point mutations observed in strains (blue lines) with random single point mutations (red lines). While the two histograms are clearly different for naturally occurring sequences (low mutation rate, high selection pressure), this is not the case for LTEE (high mutation rate, strong selection) or MAEs strains (high mutation rate, weak selection). These findings suggest that DCA can discriminate between distinct selective regimes by analyzing the genome-wide variability between different strains. Source: [89].

quence, i.e. it incorporates its context-dependence via the epistatic couplings $\mathbf{J}$.

The evidence for epistatic interactions and condition-dependent effects is abundant [95–98]. Yet, it still remains unclear to what extent those interactions contribute in determining the phenotypic impact of a mutation.

In the next section we analyze a set of 61160 *E. coli* strains (a local sampling of the fitness landscape) to quantify the importance of intragenic epistasis and of context dependency for recently diverged sequences.

5.3 THE DATASET

Here we sketch the procedure we adopted to construct the dataset (*cf.* Figure 5.2), while a detailed description of each step is given in the next sections.

First, we choose as reference the strain GA4805AA. Being of medical and biochemical interest, there is an interest in sequencing closely related strains and a large number of genome sequences are available. Second, for each of its proteins we construct an MSA of homologous sequences (the global sampling) with which we train a DCA model. The latter was used to investigate the corresponding MSA of orthologous sequences (the local sampling) obtained from the other 61159 strains. Due to the broad variability in size and gene content of the *E. coli* genome, we decided to restrict our analysis only on highly conserved genes.

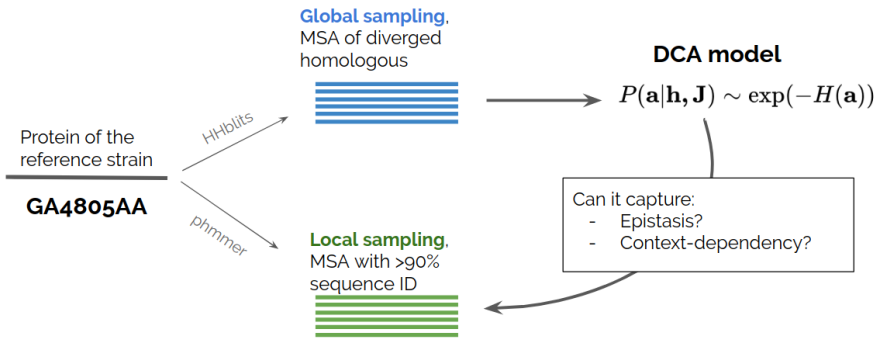

Figure 5.3: For each gene of the reference strain, we obtain a global (MSA of homologous sequences) and local sampling (MSA of orthologous sequences). We train the DCA model on the first, which we use to investigate the latter.

5.3.1 *Local sampling*

The 61160 considered *E. coli* strain are all naturally occurring strains, therefore subject to comparable selective pressure. The reference strain

GA4805AA contains a total of 5051 protein-coding genes and, for each of the translated proteins, we aim to obtain an MSA of orthologous sequences characterizing its sequence variability across strains. To this end, we constructed a target dataset by concatenating all proteins of all strains. Using phmmer [13], we searched each gene of the reference strain against this target dataset keeping for each genome, only the best hit (to avoid paralogs). Actually, this straightforward approach was too time-consuming. We therefore included a intermediate step, where we clustered the target database by grouping identical sequences, and then run phmmer on this clustered dataset. This allowed to reduce the size of the target dataset more than one order of magnitude, from $\sim 1,27 \times 10^8$ to $\sim 9 \times 10^6$ sequences. To further clean the MSAs, we removed sequences with sequence identity lower than 90% with the reference, and contains more than 10 gaps. Sequences which do not satisfy these conditions are usually fragments, which were aligned by adding a large number of gaps.

5.3.2 *Global sampling*

For each of the 5051 proteins extracted from the reference strain, we identified a set of homologous sequences in the UniProt database. A first attempt was done using jackhmmer [13]. Jackhmmer iteratively searches each query sequence - in our case, a protein of the reference strain - against the UniProt database. We soon realized that this approach was computationally infeasible for large scale analysis since it iteratively searches sequence against the protein database (*cf.* Section 1.3.1). We decided therefore to use HHblits [11]. The main advantage of the latter is that the HHblits suite provides a pre-clustering of UniProt (sequences alignable over at least 80% of their length and down to 30% pairwise sequence identity) and an HMM for each cluster. Given a query sequence, HHblits first converts it to an HMM. Second, it searches the query profile against the HMM database. Sequences from the clusters which are below a default E-value threshold 10^{-3} are added to the query MSA. Thanks to the pre-clustering of the Uniprot database, HHblists is much faster than Jackhmmer while leading to virtually identical MSAs. For each MSA with more than 100 sequences (4753 out of 5051 total genes) we learn a DCA model using the pseudo-likelihood approximation. Importantly, DCA was performed without including sequences from *E. coli* or with sequence identity higher than 90% to the reference strain. In such a way we avoid overfitting when studying the fitness landscape around the reference sequence, or a bias towards the specific *E. coli* strain present in Uniprot.

5.3.3 *Core genome*

The *E. coli* genome shows a broad variability in size and gene content. Even closely related strains are subject to repeated events of gene acquisition and loss. In our case, only two genes are preserved in all strains. To achieve statistically significant results, we decided to restrict our analysis to the set of 1520 proteins which are found in at least 61,000 strains (hereafter the *core-genome*). Note that, there is no single universally accepted definition of core genome. A common strategy [99] is to include only genes conserved in 95% of genomes (in our case, $\sim$ 58000 strains). Here we use a more stringent cutoff of $\sim$ 99.7% to minimize erroneous core genes due to over-representation of very similar genomes, which in turn provides a more stringent core. Our results are largely robust with respect to the precise choice of the cutoff.

5.4 EPISTASIS

In this section, we study if DCA is able to detect epistatic signal in our strain dataset. Before this, let us fix some notations. We denote the amino acid sequence of the reference strain GA4805AA (the wild type) by $\mathbf{a} = (a_1, \ldots, a_L)$.

The cost of a single mutation at some position i is defined as

$$\Delta H_i := \Delta H(a_i \to b_i | \mathbf{a}_{\setminus i}) = H(a_1, \ldots, a_{i-1}, b_i, a_{i+1}, a_L) - H(a_1, \ldots, a_L) \tag{5.4}$$

where $\mathbf{a}_{\setminus i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_L)$.

Similarly, we denote by

$$\Delta H_{i_1,\ldots,i_N} := \Delta H(a_{i_1} \to b_{i_1}, \ldots, a_{i_N} \to b_{i_N} | \mathbf{a}_{\setminus\{i_1,\ldots,i_N\}}) \tag{5.5}$$

the cost of having N mutations in positions $\{i_1, \ldots, i_N\}$.

Figure 5.4 shows the distribution $\Delta H_{i_1,\ldots,i_N}$ between mutated sequences observed in our local sampling and the wild type. As a comparison, we plot the $\Delta H_{i_1,\ldots,i_N}$ distribution for sequences where we inserted random mutations. While randomly mutated sequences lie at very high and positive $\Delta H_{i_1,\ldots,i_N}$, the distribution of those observed in the strains remains approximately centered close to zero even for sequences that are 10 mutations away from the reference. Natural variants are still considered "good" sequences by DCA. Substitutions present in the strains are predicted to be close to neutral, while random mutations would be predominantly deleterious. This is consistent with what observed by Couce et al. [89]: DCA correctly detects that the sequences of our dataset are under strong selection regime.

The total epistasis between positions $\{i_1, \ldots, i_N\}$ can be defined as

$$\Delta\Delta H_{i_1,\ldots,i_N} := \Delta H_{i_1,\ldots,i_N} - \sum_{k=1}^{N} \Delta H_{i_k} \tag{5.6}$$

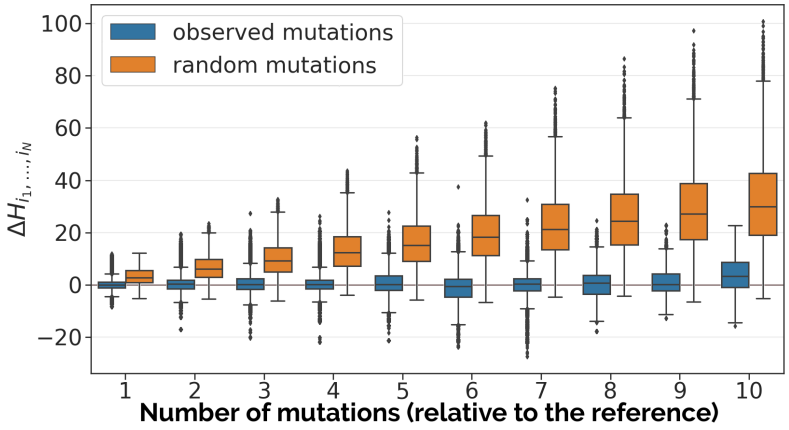

Figure 5.4: Distributions of $\Delta H_{i_1,\ldots,i_N}$ for sequences with random mutations (orange) and of those observed in the *E.coli* strains (blue). Even sequences that are 10 mutations away from the reference sequence are considered as "good" sequences by DCA (distributions remains approximately centered close to zero). On the contrary, randomly mutated sequences have very high and positive $\Delta H_{i_1,\ldots,i_N}$ (which grows approximately with the number of mutations N), i.e. they are "bad" sequences according to DCA.

i.e. the cost of having N mutations minus the sum of the costs of the single-site mutations. Panel B of Figure 5.5 displays the distribution of $\Delta\Delta H_{i_1,\ldots,i_N}$ as a function of the number of mutation N. First, we can note that epistasis is higly sparse and has surprisingly little effect. Second, distributions are symmetric around zero, thereby showing sign of neither positive nor negative epistasis [1].

In [16, 17] it was shown that epistatic residues pairs often are in contact in the 3D structure. In Panel A of Figure 5.6 we checked if sequentially close residues, which are usually in contact in 3D structure due to their proximity in the secondary structures, display higher epistasis. We plot the sum of the two single-residue mutations $\Delta H_i + \Delta H_j$ versus the cost of a double mutation $\Delta H_{i,j}$ for all residues i and j distinguishing those that are sequentially close ($|i - j| \leq 4$, red points) or far (blue points). Once again, no pairs shows significant epistatic signal, i.e. it is far from the diagonal of the scatter plot. A similar plot is shown in Panel B for three-point epistasis.

1 Positive [negative] epistasis occurs when the combined cost of carrying multiple mutations is less [more] than what would be expected if the mutations had independent and simple additive effects on fitness.

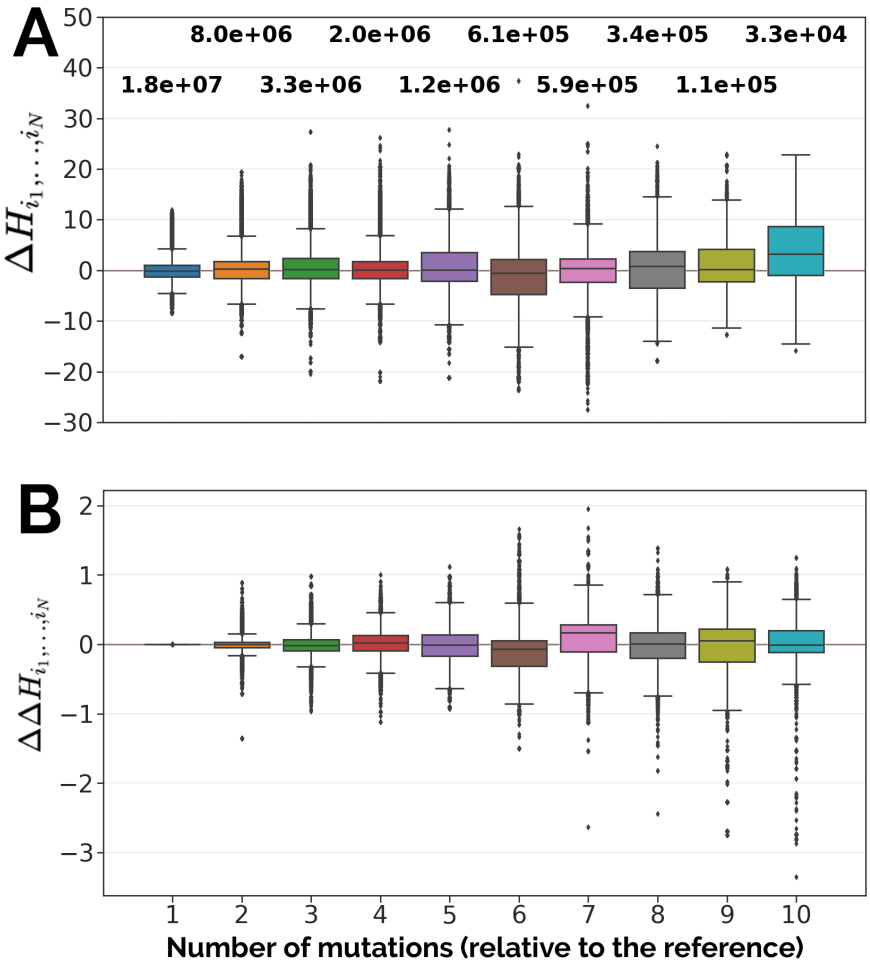

Figure 5.5: *Panel A)* The distributions of the statistical scores $\Delta H_{i_1,\ldots,i_N}$ as a function of the number of mutations N. The number of observations for each group are displayed in the upper part. *Panel B)* The distribution of the epistatic scores $\Delta\Delta H_{i_1,\ldots,i_N}$. Epistasis is highly sparse and has little effect. Note in particular the scales of the vertical axes. We observed that the orthologous alignments often contain stretches of gaps which can lead to spurious epistatic signal. Therefore, in our analysis we excluded $\Delta H_{i_1,\ldots,i_N}$ which display gaps in positions $\{i_1,\ldots,i_N\}$. Also, we excluded residues which mutate to unspecified or unknown amino acid (letter X in the MSA).

Overall, Figures 5.5 and 5.6 indicate that intragenic epistasis is virtually absent for recently diverged *E. coli* strains. This suggests that there may be a nonobvious correlation between epistasis and genetic distance from the reference strain. Therefore, it is natural to ask the following questions: at which genomic distance from the reference epistasis starts to become relevant? If we extended our analysis to *E. coli* close-by species *(e.g. Salmonella enterica* or all *Gammaproteobacteria*), could we observe epistatic signal? As a future work, we plan to tackle these questions using the data in our possession.

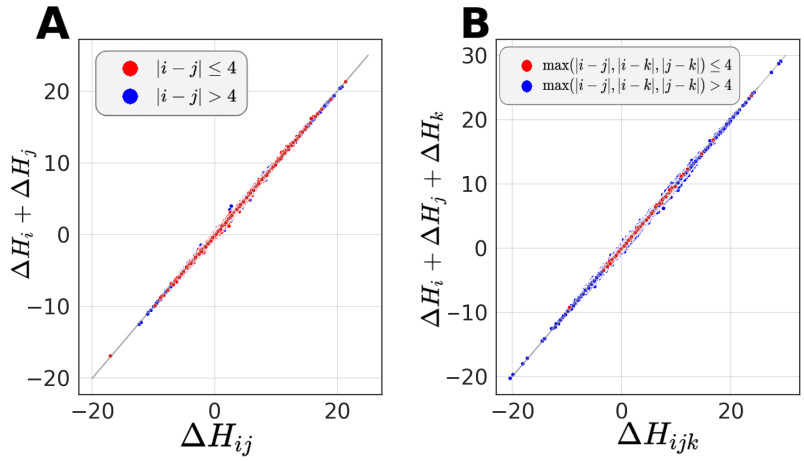

Figure 5.6: The cost of having multiple mutations versus the sum over single mutation for double (Panel A) and triple mutants (Panel B). Red dots are sequentially close residues, which are usually in contact in 3D structure. No significative epistasis effect is observed.

The fact that epistasis remains rare even when up to 10 mutations are observed, may indicate that the corresponding residues can mutate almost independently from the sequence background. This calls for a deeper understanding of how the context (i.e the amino acids present in all other positions of the protein) influences the variability (i.e. the propensity to mutate) of a residue.

5.5 QUANTIFYING CONTEXT DEPENDENCE

To quantify how the variability of a residue is influenced by its context, we adopt the following strategy. First, for a residue in position i we compute the *context-independent* site entropy:

$$s_i = -\sum_{a_i=1}^{21} f_i(a_i) \log_2 f_i(a_i) \tag{5.7}$$

where the frequencies $f_i(a_i)$ are directly computed from the MSAs of diverged homologous. The problem with Eq. (5.7) is that a column full of gaps may cause low entropy spurious signals. Therefore, we decided to exclude gaps from the counting of the frequencies[2].

Next, we define the *context-dependent* site entropy, using the DCA models inferred from MSAs of diverged homologs. We fix as "context" the sequence $\mathbf{a} = (a_1, \ldots, a_L)$ of the reference strain GA4805AA. For each position i of the MSA we can define the conditional probability

2 meaning that we compute f_i as a 21-vector, then we remove the entry corresponding to the gap-frequency, and re-normalize f_i.

of observing the amino-acid a_i given the rest of the sequence $\mathbf{a}_{\setminus i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_L)$:

$$P(a_i|\mathbf{a}_{\setminus i}) = \frac{\exp\left(h_i(a_i) + \sum_{j\neq i} J_{ij}(a_i, a_j)\right)}{\sum_{b=1}^{q} \exp\left(h_i(b) + \sum_{j\neq i} J_{ij}(b, a_j)\right)}. \tag{5.8}$$

From which we can define the *context-dependent* entropy $s_i(\mathbf{a}_{\setminus i})$:

$$s_i(\mathbf{a}_{\setminus i}) = -\sum_{a_i} P(a_i|\mathbf{a}_{\setminus i}) \log_2 P(a_i|\mathbf{a}_{\setminus i}) \tag{5.9}$$

As before, we remove the entry corresponding to the gap-conditional probability $P(a_i = "-"|\mathbf{a}_{\setminus i})$ and then re-normalize P.

Note that the context-independent s_i and -dependent $s_i(\mathbf{a}_{\setminus i})$ entropies were both computed from MSAs of diverged homologous (global sample).

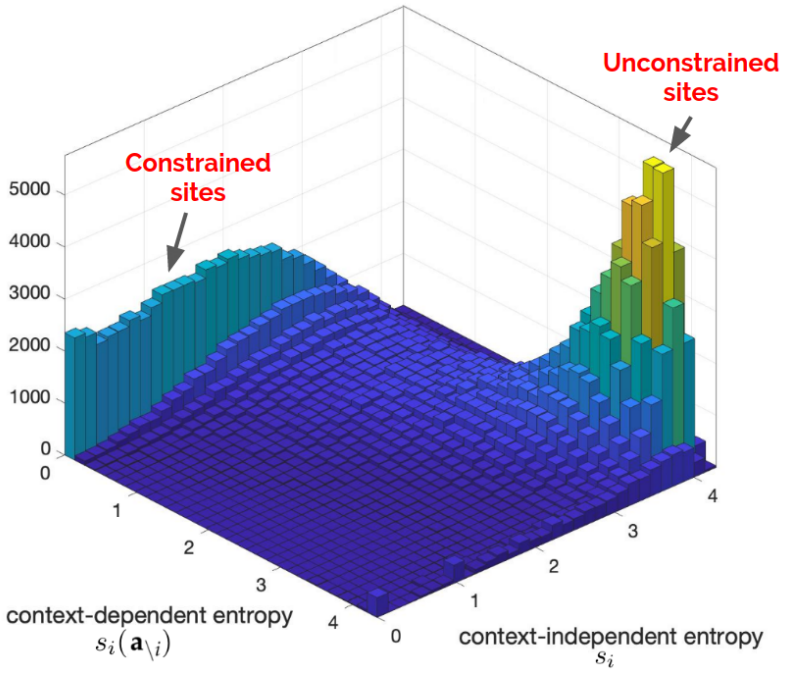

Figure 5.7: Density plot of the context-independent versus context-dependent entropy. Two sub-populations of residues are present: *unconstrained sites* which show high entropy in MSA, but small context dependence and *constrained sites* which display restricted entropy in MSA, but they are conserved in the GA4805AA context.

Figure 5.7 shows a density plot of the context-independent versus context-dependent entropy. Two conclusions can be drawn from it. First, the vast majority of sites are located in the upper triangular part, meaning that there is an entropy reduction by fixing the sequence context. Residues are in general constrained by their context. Amino acids possible in the MSA may be excluded in a specific context. Second, two sub-populations emerge:

1. *unconstrained sites* (right corner) - which show high entropy in the MSA, but small context dependence;

2. *constrained sites* (left part of the plot) – which display restricted entropy in the MSA, but they are close to conserved given the context.

Is there any relation between these quantities, both calculated from the MSA of distant homologous, and the inter-strain sequence variability? In this context, it seems natural to ask if the context-dependent entropy $s_i(\mathbf{a}_{\setminus i})$ can better catch up the variability that is observed in the MSAs of *E. coli* strains (local sample). To answer this, we plot in Figure 5.8 the two entropies considering separately sites which do or do not display polymorphisms in our dataset[3].

The vast majority of strain-polymorphic sites (Panel A of Figure 5.8) are unconstrained (located in the right corner). This may explain the low level of observed epistasis in *E. coli* strains of Figure 5.5. Indeed, unconstrained sites are not subjected to epistatic interactions since they tend to mutate independently from the rest of the protein sequence.

On the contrary the distribution of strain-conserved sites (Panel B of Figure 5.8) is bimodal. However, the pick in the right corner may be interpreted as incomplete sampling of polymorphic sites in the 61160 *E. coli* strains.

Figure 5.9 shows that the context-dependent entropy $s_i(\mathbf{a}_{\setminus i})$ can be used to better predict which sites are more prone to mutations in a given context as compared to the total entropy s_i. It displays the distributions of the context -dependent (Panel A) -independent (Panel B) entropies distinguishing between strain-polymorphic and strain-conserved sites. Although neither of them can perfectly discriminate between polymorphic and non-polymorphic sites, only the two histograms of panel B are well-separated.

We can even quantify more precisely the importance of the context for a residue in position i, by introducing the context dependent information gain:

$$I_i = s_i - s_i(\mathbf{a}_{\setminus i}) \tag{5.10}$$

which measures the reduction in entropy by including the context. To be more precise, the higher I_i the more the site i is constrained by the context, i.e. less variable. As a reference, 1 bit of information gain means that the effective number of amino acids is reduced by 2; 2 bits indicate a reduction by 4 and so on.

Figure 5.10 shows that the enrichment in the positive tail is particularly pronounced for strain-conserved sites, thereby indicating that sites which are mutable in diverged homologous can become highly constrained in a specific context.

3 To minimize erroneous polymorphic sites due to sequencing errors, we removed from the analysis sites that mutate just once.

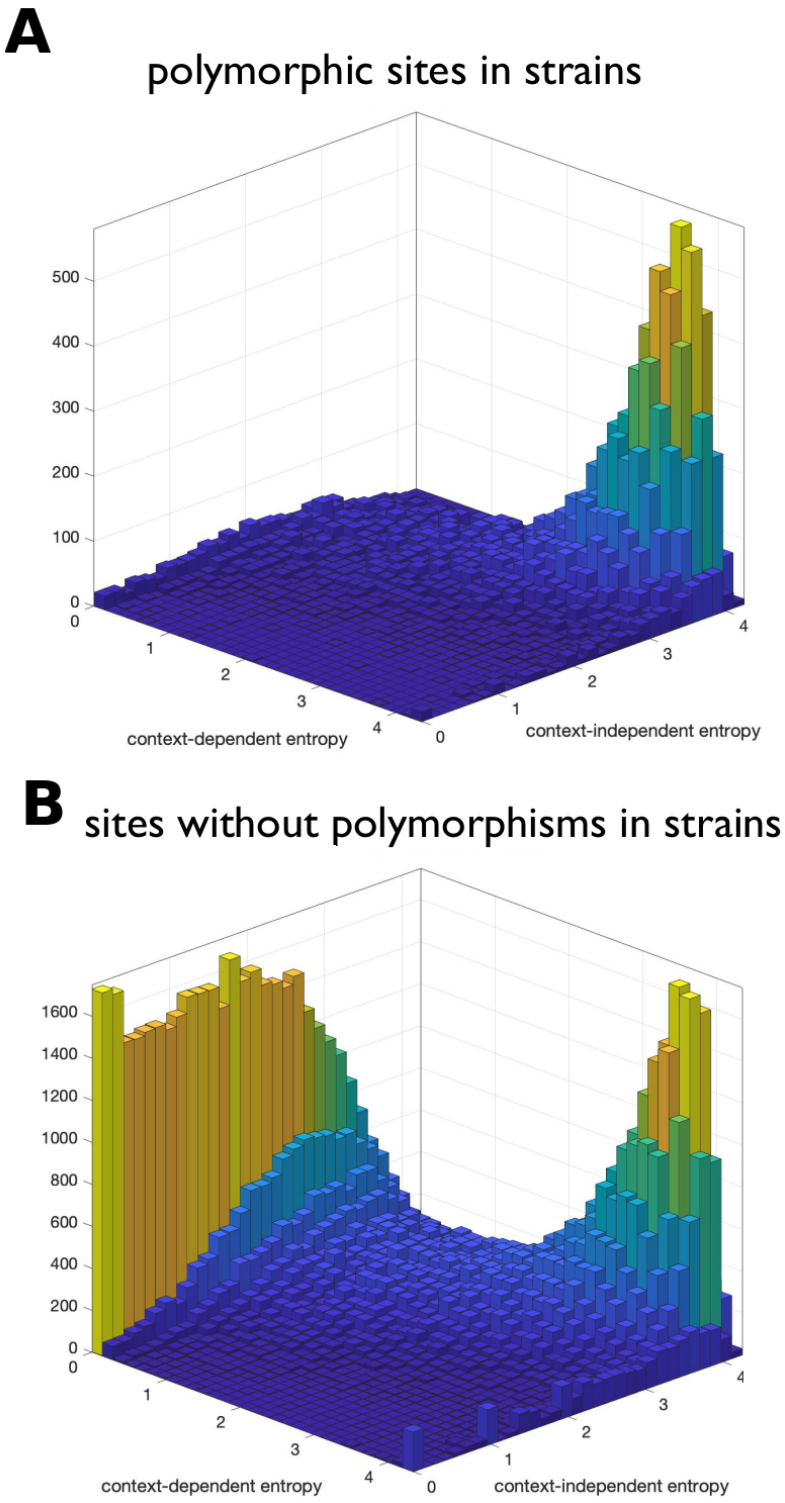

Figure 5.8: Density plot of the context-independent versus context-dependent entropy, considering the 26077 strain-polymorphic position (Panel A) and the 181958 site which are always conserved in the local sample of strains (Panel B). The vast majority of polymorphic residues are *unconstrained sites*, therefore non-epistatic. The pick in the right corner of Panel B (strain-conserved residues) can be interpreted as incomplete sampling of polymorphic sites in the *E. coli* strains of our dataset.

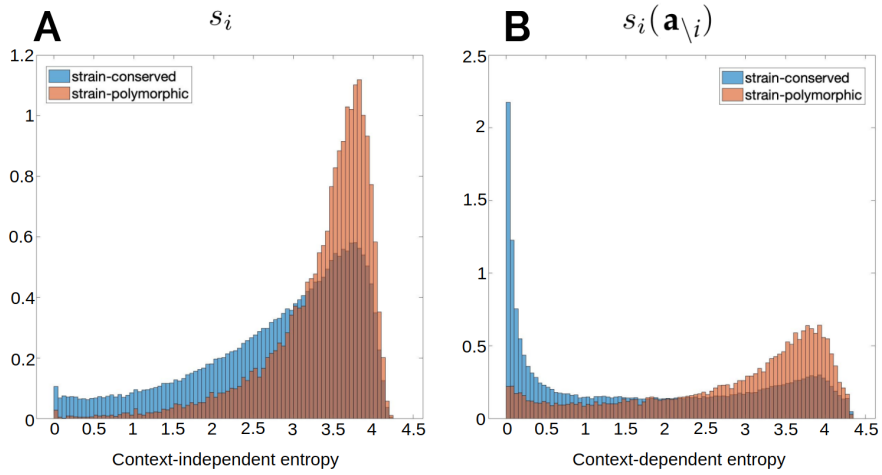

Figure 5.9: The distributions of the context-independent (Panel A) and context-dependent (Panel B) entropies for strain-polymorphic (red) and strain-conserved sites (blue). The separation between the two is cleaner for the context-dependent entropy. This suggests that it can be used to better predict which sites tend to be more variable in a given context.

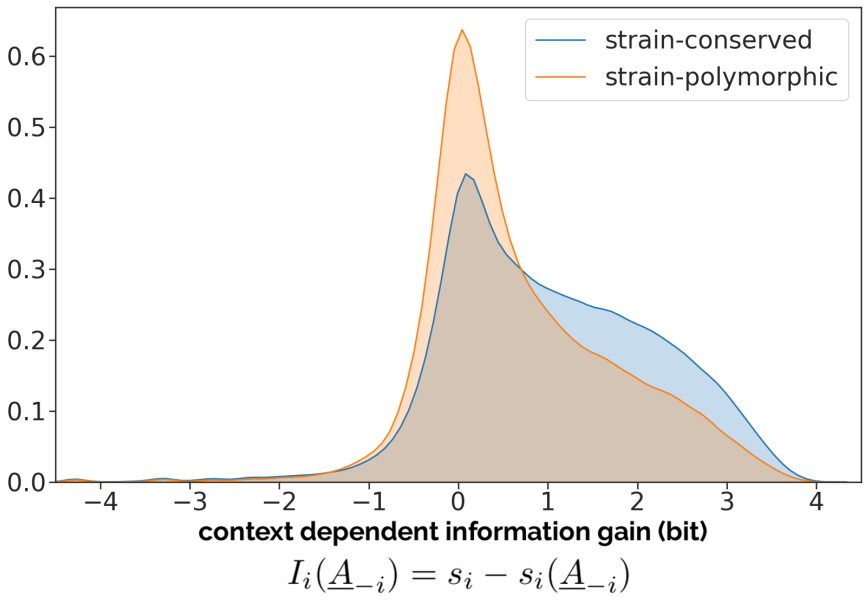

Figure 5.10: Distribution of the information gain from the context for strain-conserved and strain-polymorphic sites. Both display positive tails, thereby indicating that the context is informative about the variability of a residue. In particular, strain-conserved residues display a large enrichment in the positive tail. This indicates that sites which are mutable in diverged homologous can become highly constrained in a specific context.

5.6 SUMMARY AND OUTLOOK ON FUTURE WORK

Observations and hypotheses presented in this chapter provide intriguing insights towards understanding to what extent DCA models trained on distant homologous can be informative about the local fitness landscape. Using 61160 *E. coli* naturally occurring strains, we showed that intragenic epistasis remains rare even when up to ten mutations were combined. This suggests that *E. coli* strains in our dataset did not diverge enough to display a significative epistasis.

This result, maybe surprising at first sight, becomes more consistent at the light of the analysis we have done on the context-dependent entropy $s_i(\mathbf{a}_{\setminus i})$ Eq. (5.9). It quantifies how the variability of a site (i.e. its tendency to mutate) is influenced by the the amino acids present in all other positions. We showed that the vast majority of polymorphic sites observed in our dataset are unconstrained sites, meaning that they tend to mutate independently from the rest of the protein sequence. We can thus conclude that epistasis is very weak in the local landscape explored by the polymorphisms between strains, but that the shape of this local landscape is strongly dependent on the joint epistatic couplings to the entire background sequence.

This may explain the low level of observed epistasis in *E. coli* strains and propose at the same time a line of future research: there should be a nonobvious correlation between epistasis and genetic distance between sequences which is worth being investigated further.

Also, by introducing the information gain Eq. (5.10) we were able to quantify the reduction in entropy by including the context. We showed that the information gain is particularly pronounced for strain-conserved sites, meaning that residues which are mutable in diverged homologous can become highly constrained in a specific context.

5.6.1 *Towards DCA as evolutionary model?*

Usually, in models of protein sequence evolution, the details of the site-specific selective constraints that governs sequence evolution are not known a priori, making it challenging to create predictive evolutionary models. If DCA can accurately model how the selective pressure at a given site depends on the its context, there is hope that it may also be used to more accurately understand - and eventually predict - the evolutionary pathways of a protein.

Our dataset allows to explore in detail the short-term evolution of natural *E. coli* strains and our long-term goal is to understand if DCA could reproduce it. In the spirit of constructing an explicit protein evolution model based on DCA, we can think of evolution as a sampling process, where mutation are proposed randomly one by one, and then selected according to their DCA score ΔH_i.

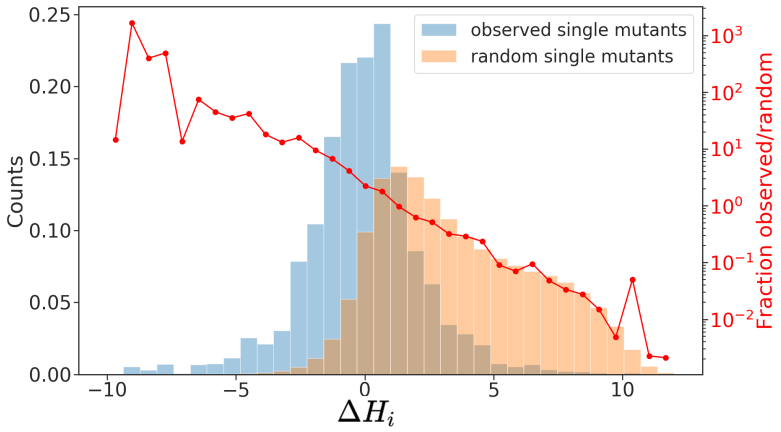

Figure 5.11: *(Left axis)* The distribution of DCA score ΔH for single mutants observed in strain (blue) and randomly generated (orange). While naturally occurring variants are approximately symmetrically distributed around zero, the vast majority of random mutations have a positive score (deleterious mutation). This is a clear signature of natural selection. *(Right axis)* The ratio of the number of observed single mutants to the number of random single mutants *(the acceptance rate of a mutation)* which displays an exponential decay.

In Figure 5.11 we compare the distribution of ΔH_i for observed mutations with random mutations. Similarly to [89] the two histograms are clearly different. The right red axis shows the acceptance rate of a mutation as a function of ΔH_i: for each bin, we compute the ratio between the number of observed single mutants and the number of random single mutants. It displays an exponential decay (over almost 6 orders of magnitude).

This suggests that, for short-term evolution, the probability of accepting a substitution of a the wild-type amino acid a_i at position i with amino acid b is given by:

$$P^{acc}(a_i \to b|a_1, \ldots, a_{i-1}, a_{i+1}, a_L) \sim \\ \sim \exp\Big(- \beta^{acc} \Delta H(a_i \to b|a_1, \ldots, a_{i-1}, a_{i+1}, a_L)\Big) \quad (5.11)$$

with $\beta^{acc} \simeq 0.7$. Note that, the distribution of random single mutants is slightly different from that of Couce et al. [89], Panel B of Figure 5.2. It contains more neutral mutations. Indeed, we performed our analysis on the entire protein sequence of the reference strain, while in [89] only the Pfam domains contained in each sequence were considered. Therefore we mutate also residues belonging to linking regions, which usually do not have a considerable impact on fitness.

Next, we studied the relative multiplicity of single mutants within the population of *E. coli* strain as a function of the ΔH_i. In Figure

5.12 we compare the distribution of ΔH_i of single mutants with the distribution of the same set of single mutants where we removed all duplicates, i.e. considering every single mutation only once. It shows that deleterious single mutations are present in the dataset, but they are exponentially depressed in frequency (over almost 3 orders of magnitude).

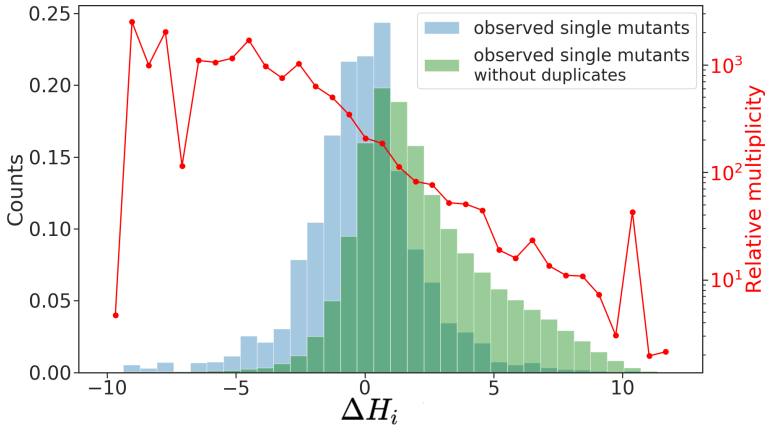

Figure 5.12: *(Left axis)* The distribution of DCA score ΔH for single mutants observed in strain (blue) and the same set of single mutants removing all duplicates (green). *(Right axis)* The ratio between the distribution shows that the frequencies of single mutations in strains depends exponentially by their DCA energies.

While these results are still preliminary, they support the idea that DCA can be used to develop more accurate evolutionary models. Indeed, models of protein sequence evolution usually neglect the details of the site-specific selection, making it challenging to create predictive evolutionary models. If DCA can accurately grasp how the selective pressure at a given site depends on its context, this may open the door to a data-driven evolutionary model.

Part IV

CONCLUDING REMARKS

CONCLUDING REMARKS

Direct-coupling analysis was originally proposed as a tool to help contact prediction from sequence information alone. The basis of DCA is a pairwise statistical model, taking the form of a generalized Potts model, whose parameters are inferred from a set of phylogenetically related biological sequences belonging to a multiple sequence alignments. The great success of DCA for protein structure prediction, encouraged its application in other contexts, from prediction of mutational effects to modeling of fitness landscapes, which we presented in the first part of this thesis.

In this second part, we have shown how the machinery of DCA can help the prediction of protein–protein interactions. PPI can be studied focusing on two major aspects, (i) large scale analysis of protein-protein interaction networks, and (ii) identification of the interaction interfaces within a protein complex.

The first was investigated in Chapter 3. The underlying idea of our approach is that interacting proteins coevolve on multiple but interconnected scales: from correlated presence/absence across species, to correlations in amino-acid usage. Our approach combines these different scales to predict currently unknown, but biologically sensible interactions.

In Chapter 4, we focused on protein interaction surfaces. In principle, contacts across the interface can be identified by analyzing the coevolutionary signals between residues which are located on the protein surface. However, the interface coevolutionary signal is weak and difficult to detect with a global statistical modeling without using structural supervision. We have shown that we can improve the predictive performance of DCA by integrating typical patterns of contacts (due to secondary structure) into the DCA predictions. We demonstrated the effectiveness of our approach in terms of interpretability and prediction performance. It can achieve results comparable to much more complex, data hungry and hardly interpretable, deep-learning methods. Even if not out-performing these methods in applications, we think that interpretability is important to understand how the contact information is hidden in sequence data.

Finally, in the third part of this thesis (Chapter 5), we investigated the fitness landscape properties at the local and global scale through an extensive genome-wide study of *E.coli* strains. We have shown that intragenic epistatic signal is virtually absent for recently diverged *E. coli* strains. At the same time, via the context-dependent entropy, we quantitatively and qualitatively characterized how the variability (i.e.

tendency to mutate) of a residue is influenced by the amino acids present in all other positions of the protein.

We may thus conclude that epistasis is very weak in the local landscape explored by the polymorphisms between strains, but that the shape of this local landscape is strongly dependent on the joint epistatic couplings to the entire background sequence.

To conclude, two lines of future research are proposed following the ideas presented in Chapter 5. First, there seems to be a non trivial connection between genomic distance and epistasis, which is worth investigating further. The ultimate goal would be to determine to what extent the epistatic effects are related to genetic distances between strains or species.

The second one concerns the exciting field of modeling protein evolution. Usually, in models of protein sequence evolution, the details of the site-specific selection that govern sequence evolution are not known a priori, making it challenging to create predictive evolutionary models. If DCA can accurately grasp how the selective pressure at a given site depends on its context, it might also be used to gain insights into the evolutionary pathways of a protein.

In analogy to the sequence-structure relationship, this might be a field where the abundant genomic data can lead to a breakthrough in our understanding, by the consequent use of data-driven modeling techniques.

Part V

APPENDIX

A

APPENDIX

A.1 PHYDCA: SUPPLEMENTARY INFORMATION

A.1.1 *Input data*

The starting point of our analysis is the phylogenetic profile matrix (PPM): a binary matrix $(n_i^a)_{i=1,\dots,N}^{a=1,\dots,M}$ whose entries capture the presence ($n_i^a = 1$) or absence ($n_i^a = 0$) of a domain i in genome a, with $a = 1,\dots,M$ (M being the number of genomes) and $i = 1,\dots,N$ (N being the number of domains). As discussed in chapter 3, the domains (the columns of the PPM) are then compared with each other to look for functionally related domains. The data we use are extracted from the Pfam 30.0 database (version of July 2016), and assigned to bacterial or eukaryotic species using the Uniprot species list available on (http://www.uniprot.org/docs/speclist).

A.1.2 *Similarity measures*

In standard phylogenetic profiling the correlations between the columns $(\mathbf{n}_i, \mathbf{n}_j)$ describing a pair of domains are usually evaluated via the Hamming distance, Pearson correlation or the p-value of the Fisher's exact test. We briefly describe each below.

1. *Hamming distance:* counts the number of bits which differ between two binary strings $\underline{n}_i, \underline{n}_j$ divided by the total number of domains, i.e. the number of species containing exactly one of the two domains,

$$d_H(\underline{n}_i, \underline{n}_j) = |\{n_i^a \neq n_j^a, \quad a = 1,\dots,M\}|/M$$

2. *Pearson Correlation:* measures the linear dependence between two domains $\underline{n}_i, \underline{n}_j$. It is defined as

$$r(\underline{n}_i, \underline{n}_j) = \frac{\sum_{a=1}^{M}(n_i^a - \bar{n}_i)(n_j^b - \bar{n}_j)}{\sqrt{\sum_{a=1}^{M}(n_i^a - \bar{n}_i)^2}\sqrt{\sum_{a=1}^{M}(n_j^a - \bar{n}_j)^2}}$$

3. *p-value of Fisher Test:* for each couple $\underline{n}_i, \underline{n}_j$ we construct an auxiliary 2×2 matrix:

$$\begin{pmatrix} M_{1,1} & M_{1,2} \\ M_{2,1} & M_{2,2} \end{pmatrix}$$

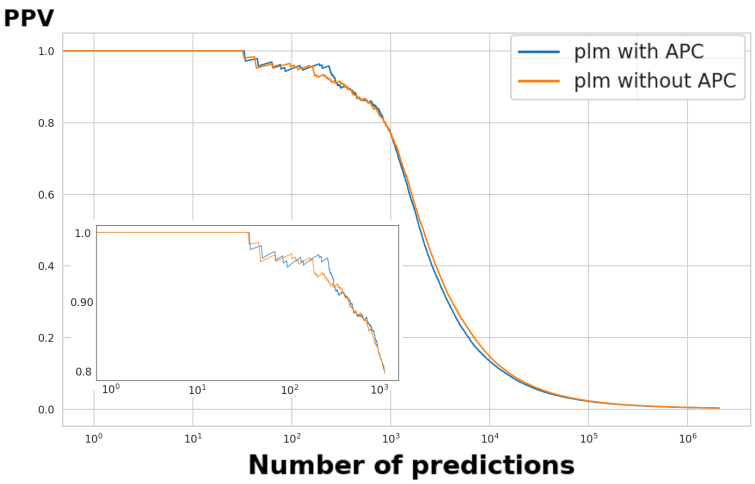

Figure A.1: APC correction. The plots show the PPV curves with and without APC correction for the PLM approximations.

with
$M_{1,1}$: Number of species that do not have neither domain i nor j;
$M_{1,2}$: Number of species that have i but not j;
$M_{2,1}$: Number of species that do not have i but have j;
$M_{2,2}$: Number of species that have both i and j.
We define $R_i = \sum_j M_{i,j}, C_j = \sum_i M_{i,j}$ and we have $M = \sum_{i,j} M_{i,j}$. We calculate the conditional probability of getting the actual matrix given the particular row and column sums:

$$P_{cutoff} = \frac{\binom{R_1}{M_{11}}\binom{R_2}{M_{21}}}{\binom{M}{C_1}} = \frac{R_1!R_2!C_1!C_2!}{M!M_{1,1}!M_{1,2}!M_{2,1}!M_{2,2}!}$$

which is a multivariate generalization of the hyper-geometric distribution. Theoretically, we analyse all the matrices of non negative integers consistent with the marginals R_i,C_j and M, and for each of them we calculate the p-value. The p-value of the test is the sum of all p-values which are $P \leq P_{cutoff}$. Small p-values thus indicate atypical cases related to correlations between the distributions of the two domains across species.

A.1.3 *Average product correction*

While for residue-level DCA , APC-corrected scores have a significantly better prediction accuracy, an a posteriori analysis show that the effect is small and almost negligible in PhyDCA (see Figure A.1).

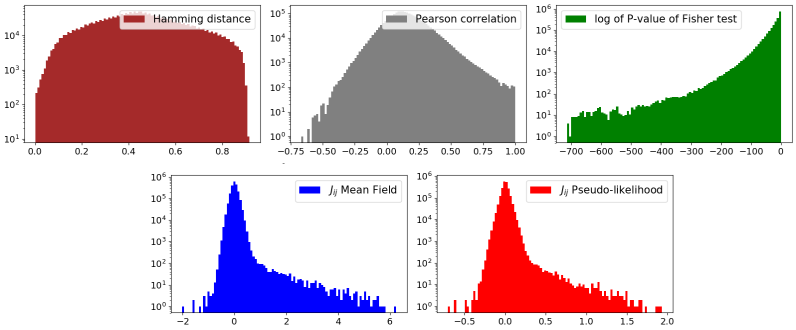

Figure A.2: *Metrics summary.* The plots show the distribution of the metrics (Hamming distance, Pearson correlation, the P-value of the Fisher's exact test mean-field and PLM phyletic couplings) for all the couples of domains existing in the \textit{E. coli} K-12 MG1655 strain..

A.1.4 *Results*

Figure A.2 shows the histograms of the similarity measures (Hamming distance, Pearson correlation, the P-value of the Fisher's exact test and phyletic couplings) for all the pairs of domains considered in the main text. Related domains ought to have profiles of high phyletic couplings, high correlations, low p-value of Fisher's exact test or low Hamming distances, since the first two are similarity, the second two more dissimilarity measure

In Figure A.3 we plot the phyletic-couplings J_{ij} found using the MF and the PLM approximations. Predictions are done by sorting all couplings in decreasing order. They are evidently highly similar, but include a partial reordering. To extract Figure 2 of chapter 3, the blue dots are interpreted as false positives. From Figure A.3 it is evident that these false positives are - even for very large coupling values - similarly distributed for the two approximations in between the true positives (red points), therefore showing that none of the methods has a clear advantage in precision.

In Figure A.4 we plot the PPV as a function of the couplings (not the cumulative PPV, but PPV per bin of coupling values). It shows that the enrichment of true positive predictions is very high in the tail of large couplings ($J_{PLM} > 0.5$ or $J_{MF} > 1.5$), and remains very limited for smaller couplings ($J_{PLM} < 0.3$ or $J_{MF} < 0.5$).

A.1.5 *Paralog matching*

To identify physical interactions we use the procedure introduced in [54], which studies coevolution of domain pairs at the level of the individual residues. The matched MSA is than used to identify interacting protein families: an average of the four highest inter-protein residue-residue PLM scores larger than 0.2 is a strong indicator for

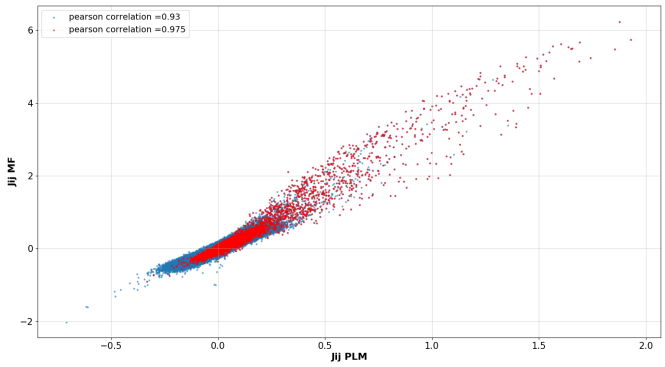

Figure A.3: *Comparison* PLM *and* MF *approximations.* The scatter-plot shows the phyletic couplings found using the MF and the PLM approximations. The blue points are domains pairs are all domain pairs while the red points are those belonging to the positive set of known domain-domain relations (note that the red points form a subset of the blue points). The plot shows that the advantage of PLM over MF (or viceversa) is not visible in the case of domain-domain co-occurrence.}

a potential interaction, at least of the joint MSA has an effective size $M_{eff} > 200$.

Figure A.5 and Figure A.6 show the results of the matching procedure for the domain pairs inside the *E. coli* K-12 MG1655 strain.

Figure A.7 shows the results of residue-level DCA for the 200 domain pairs of strongest phyletic couplings, which are co-localized in one protein in *E. coli*. Due to the co-localization, the generation of a joint MSA is trivial in this case; the paralogs-matching can be avoided. Note that domains can co-occur in the same protein without direct physical interactions. Out of the 200 pairs, 144 of these domain pairs are also listed in *iPfam*, meaning that a direct physical interaction is structurally known.

A.1.6 *Network analysis*

As stated in the main text, CoPAP and PhyDCA treat very different confounding factors of coevolutionary analysis – phylogenetic biases and indirect correlations. Nevertheless from Figure 3 of the main text, it appears that almost none of the correlated pairs strongly coupled in PhyDCA, are actually discarded by CoPAP. But are the correlations of pairs, which are retained by CoPAP as non phyletically coupled, but discarded by PhyDCA, really an indirect network effect of the PhyDCA couplings?

To answer this question, we first introduce in Figure A.8 two scatter plots of the phyletic couplings vs. Pearson correlations between domain pairs, in the first case for the 3611 domain pairs of highest

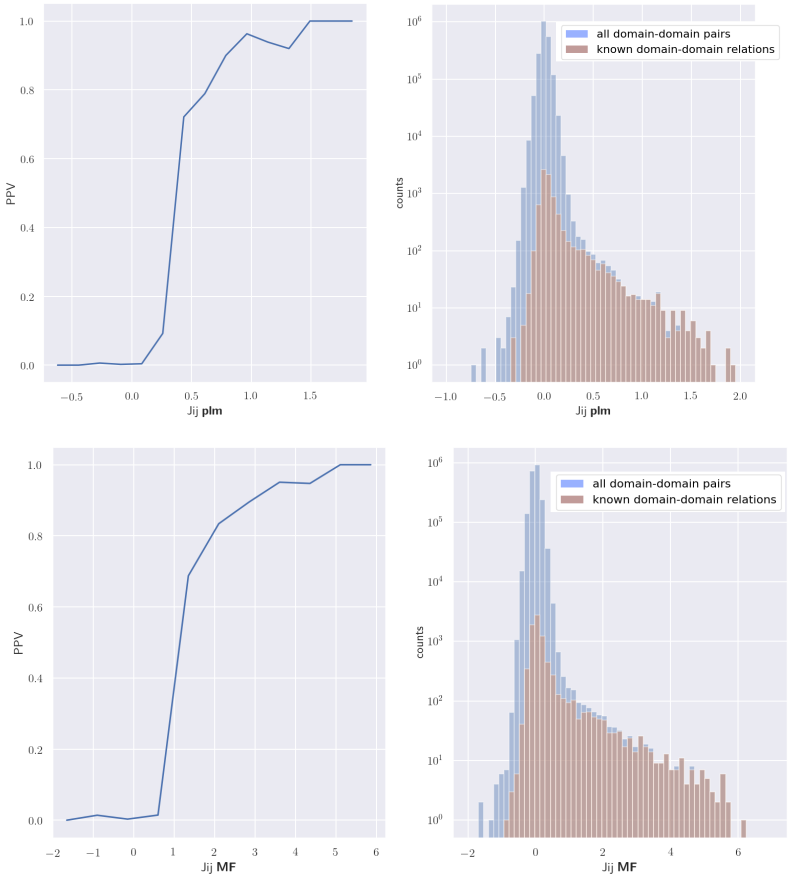

Figure A.4: PPV *as a function of couplings.* The left figures show the sharp drop of PPV from values close to one for $J_{plm} > 0.5$ with the PLM approximation (or $J_{MF} > 1.5$ in the MF approximation) to very low PPV for $J_{plm} < 0.3$ (or $J_{MF} < 0.5$). The right histograms show the distribution of the phyletic couplings for all domains pairs and for known domain-domain relations. The kink in the histograms between the bulk of small J_{ij} and the tail of large J_{ij} is observed to provide a good cutoff value for high-quality predictions. Once again, it is located close to $J_{plm} = 0.5$ or ($J_{MF} = 1.5$).

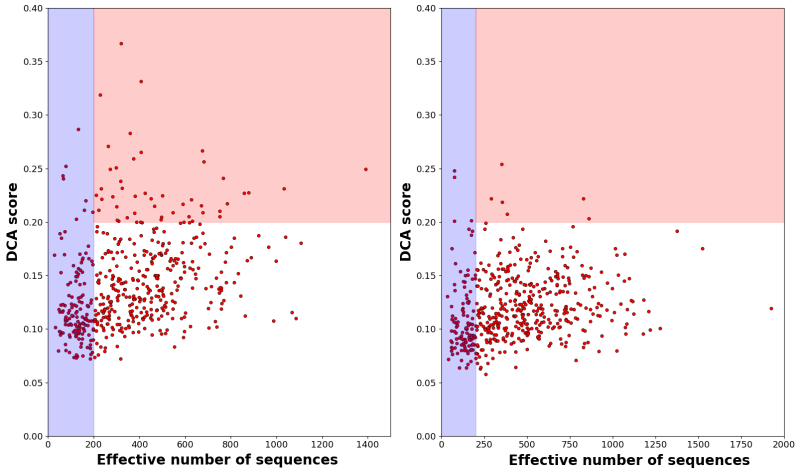

Figure A.5: *Matching procedure for E. coli.* The panel on the left shows the result of the matching procedure for the 500 most significant predictions for domain families existing inside the K12 strain of *E. coli* (the list can be found on the Github page at `results/ECOLI_matching_results.dat`). On the right, as a comparison, a random matching for the same domain pairs

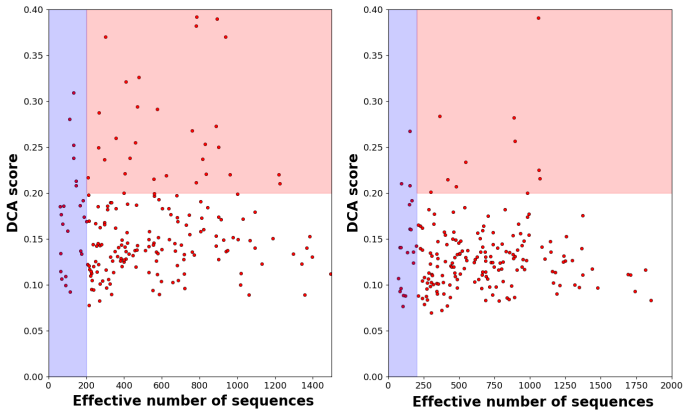

Figure A.6: *Matching procedure for E. coli. domains in iPfam.* The panel on the left shows the result of the matching procedure for the 200 pairs of highest phyletic couplings belonging to the *i*Pfam database (the complete list can be found on the Github page at `results/ECOLI_matching_iPfam_results.dat`) . On the right, as a comparison, a random matching for the same domain pairs.

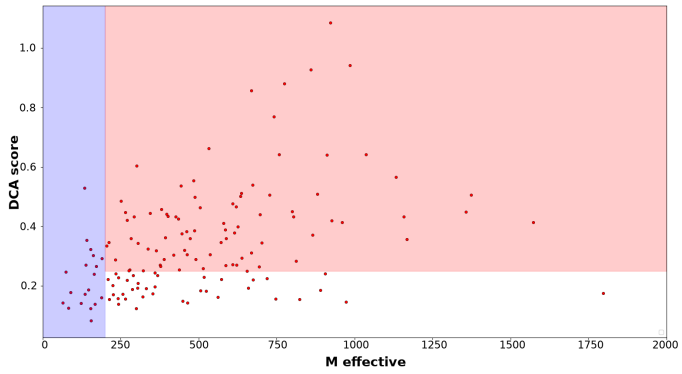

Figure A.7: *Results of residue-level* DCA *for domains co-localized in the same protein*. The panel shows the results of residue-level DCA for the 200 domain pairs of strongest phyletic couplings. Co-localization does not necessarily imply physical interaction. However, out of 200 pairs, 144 are also listed in *iPfam* database as being in physical contact in experimentally resolved PDB structures. Due to their co-localization, the generation of a joint MSA is trivial in this case and the paralogs-matching can be avoided.

CoPAP score, in the second case for all domain pairs. In both cases, we see a clear triangular shape, indicating that large couplings lead to large correlations, but large correlations can exist between weakly coupled pairs. Since our PhyDCA model reproduces correlations using couplings, the latter case must result from indirect correlations. Also as a consequence, the phyletic coupling network is substantially sparser than the correlation network.

To corroborate this, in Figure A.9 , we consider the network of the 1000 strongest phyletic couplings and study the correlations as a function of the shortest-path distance between domains along this network. Correlations decrease with distance until they saturate at a low but non-zero level. This is coherent with the idea that empirical correlations found in the data have at least three contributions - direct correlations induced by direct couplings (at distance 1), indirect couplings induced by coupling chains, and a ground level of correlations, which possibly result from phylogenetic correlations between the species and other sampling effects.

If we take alternatively the network induced by the 1613 pairs, which have large Pearson correlations and are preserved by CoPAP (the intersection of the red and green circles in Figure 3A of chapter 3), we also find a correlation decrease (as to be expected in any sparsely connected graphical model), cf. Figure A.10. However, the decay is slower than on the PhyDCA network, even if the network is denser. Pairs in the PhyDCA network are thus less correlated than pairs at the same distance in the correlation network, which shows

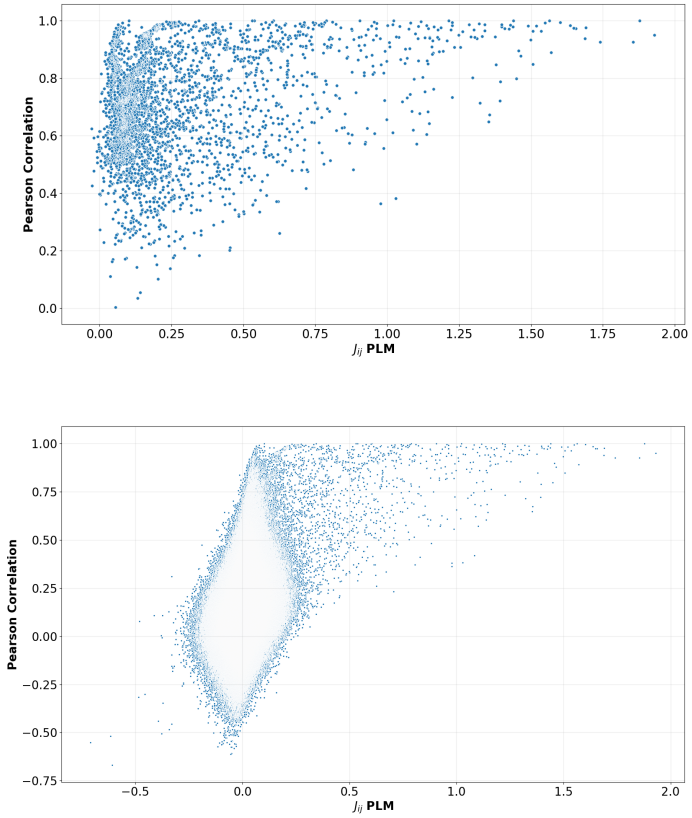

Figure A.8: *Couplings vs. correlations.* The figure shows a scatter plot of PhyDCA couplings vs. Pearson correlations, for the 3611 domain pairs of highest CoPAP score in the upper panel, and for all domain pairs in the lower one.

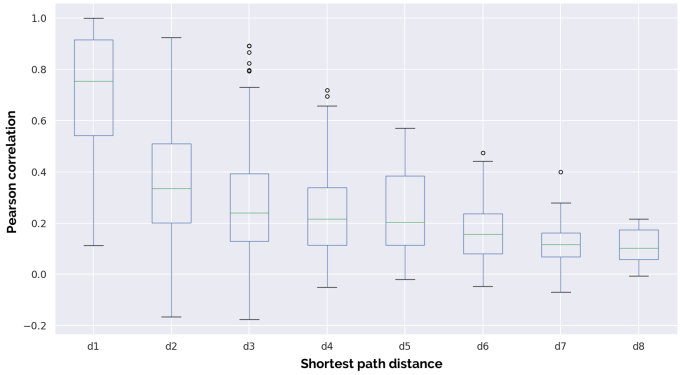

Figure A.9: *Correlation decay on PhyDCA network.* The figure shows the decay of empirical correlations between pairs of domain belonging to the network of the first 1000 strongest phyletic couplings as a function of their shortest-path distance on this network.

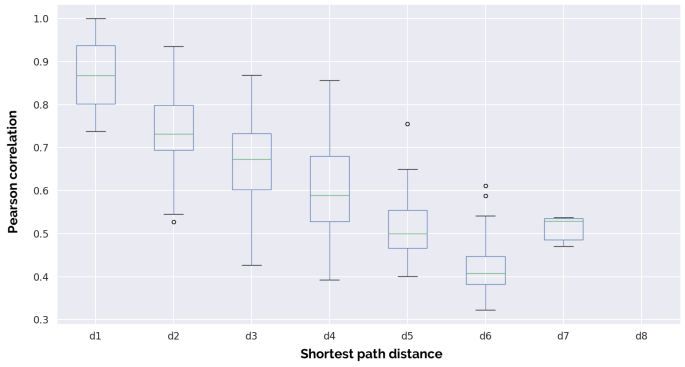

Figure A.10: *Correlation decay on CoPAP-Pearson network.* The figure shows the decay of empirical correlations between pairs of domain belonging to the network of the 1613 domain pairs of strongest CoPAP-preserved Pearson correlations (the intersection of the red and green circles in Figure 3A of chapter 3) as a function of their shortest-path distance on this network.

that the phyletic coupling network more parsimoniously explains the connectivity patterns present in the data.

A.1.7 *All bacteria*

In the main text we use the model organism *E. coli* as reference genome in order to have a large set of known domain-domain relationships. In this section we consider a broader selection of genomes by applying the same methodology to all 9,358 Pfam domains appearing in bacteria. To access the accuracy of our prediction we compile a number of known domain-domain relationships: *intra-protein localization* (out of 2,972,104 proteins 866,591 contain multiple domains, giving rise to 26,381 distinct domain-domain relations), *domain-domain contacts in 3d structures* (from the iPfam database, for a total of 545 known relationships), *protein-protein interaction* (from the IntAct database, obtaining 67,409 domain pairs). This leads to a total of 92,428 known relationship (cf. Figure A.11, Panel A).

We then select the couplings between domains which are only present in *E. coli* genome (cf. Figure A.11, Panel B and C) finding96% correlation with the couplings inferred in the main text, thus proving the robustness of the results with respect to the selection of domains.

We have applied the paralog-matching analysis to the 200 most coupled bacterial domain pairs (see Figure A.12). A list of the domain pairs, their phylogenetic coupling and the DCA score can be found on the Github page at `results/ALLBACTERIA_matching_results.dat`).

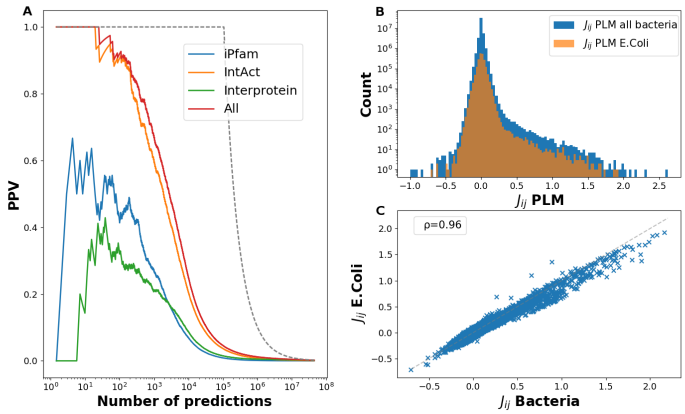

Figure A.11: *Phylogenetic couplings.* Panel A shows the PPV of the phyletic couplings of all bacterial domains for predicting domain-domain relationships (including protein architecture, iPfam and IntAct entries). Panel B shows a histogram of couplings J_{ij}, as inferred by PLM , for the domains present in all bacteria and for those appearing only in *E. coli*. In Panel C we retain from the bacterial phyletic couplings only the couplings between domains present in *E. coli*. Then we compare them with the couplings found by the procedure described in the main text, finding a correlation of 96% between the two.

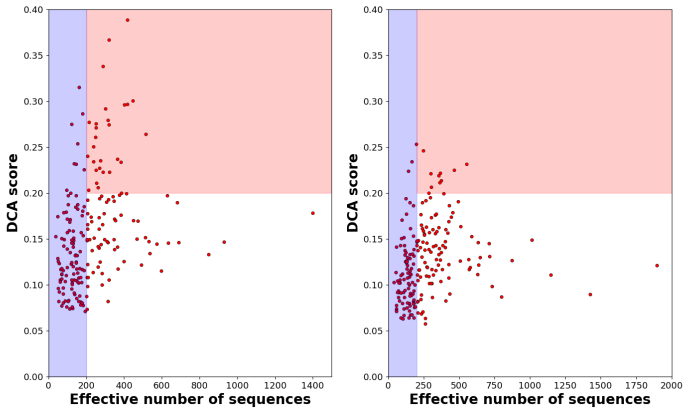

Figure A.12: *Matching procedure for bacteria:* Panel A shows the effective sequence number and the DCA scores for the 200 most significant PhyDCA predictions. Panel B shows a random matching for the same domain pairs.

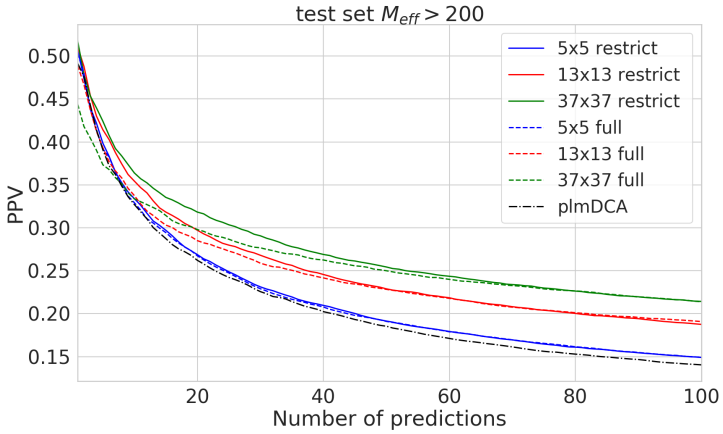

Figure A.13: *Performance of FilterDCA on the test set using different learning strategies.* Training the logistic regression with residues pairs with DCA score larger than zero (full lines) rather than using all pairs of residues (dashed lines), leads to a slight improvement in the performance. In such a way, the classifier concentrate on cases which show more reliable DCA scores.

A.2 FILTERDCA: SUPPLEMENTARY INFORMATION

A.2.1 *Handling imbalanced datasets*

In chapter 4 of the main text, it is mentioned that in the training set the incidence of class *non-contact* is dominant, being found in $\sim 99\%$ of cases. We tried two different strategies for the learning: first using all residue pairs for training, second restricting the training set on residues pairs with DCA score larger than zero. Figure A.13 shows that the second one leads to slight better results.

A.2.2 *Comparison issues*

To access the performance of *FilterDCA* it is important to compare it with at least two methods: unsupervised DCA and a CNN.

To our knowledge the only deep-learning method devoted to inter-protein contact prediction is *RaptorX Complex* [77] which uses the same architecture of *RaptorX* [78] (see Figure 4.1) and then apply "transfer-learning" (see Section 4.1.1). Unfortunately, it does not allows for a fair comparison. Indeed, no code is publicly available, and we can only submit our MSAs to a web server which is trained on all single-chain proteins available on the PDB, thereby making impossible to avoid overfitting when studying intra-protein inter-chain interactions. Therefore, in the main text we have compared our performances with the deep learning method Pconsc4[88]. PconsC4 has been developed to predict intra-protein contacts. However, its training set (which consists of 2891 proteins culled from PDB) contains 9% of multi-domain

proteins (no overlap with our test sets). It adopts the U-net architecture [100], designed for image segmentation, which is composed of a series of CNNs with shortcut connections. 72 features are calculated and fed into PconsC4: 68 one-dimensional sequential features and four pairwise features, the GaussDCA score [101], APC-corrected mutual information, normalized APC-corrected mutual information, and cross-entropy.

BIBLIOGRAPHY

[1] Clare M O'Connor, Jill U Adams, and Jennifer Fairman. "Essentials of cell biology." In: *Cambridge, MA: NPG Education* 1 (2010) (cit. on p. 3).

[2] Rodolfo O Esquivel, Moyocoyani Molina-Espiritu, Frank Salas, Catalina Soriano, Carolina Barrientos, Jesús S Dehesa, and José A Dobado. "Decoding the Building Blocks of Life from the Perspective of Quantum Information." In: *Advances in Quantum Mechanics*. IntechOpen, 2013. DOI: 10.5772/55160 (cit. on p. 4).

[3] Wikipedia contributors. *Protein structure — Wikipedia, The Free Encyclopedia*. 2004. URL: https://en.wikipedia.org/wiki/Protein_structure (cit. on p. 5).

[4] Christian B Anfinsen. "Principles that govern the folding of protein chains." In: *Science* 181.4096 (1973), pp. 223–230. DOI: 10.1126/science.181.4096.223 (cit. on p. 5).

[5] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. "The protein data bank." In: *Nucleic acids research* 28.1 (2000), pp. 235–242. DOI: 10.1093/nar/28.1.235 (cit. on p. 5).

[6] Helen Berman, Kim Henrick, and Haruki Nakamura. "Announcing the worldwide protein data bank." In: *Nature Structural & Molecular Biology* 10.12 (2003), p. 980. DOI: 10.1038/nsb1203-980 (cit. on p. 5).

[7] David De Juan, Florencio Pazos, and Alfonso Valencia. "Emerging methods in protein co-evolution." In: *Nature Reviews Genetics* 14.4 (2013), p. 249. DOI: 10.1038/nrg3414 (cit. on p. 5).

[8] Rolf Apweiler et al. "UniProt: the universal protein knowledgebase." In: *Nucleic acids research* 32.suppl_1 (2004), pp. D115–D119. DOI: 10.1093/nar/gkw1099 (cit. on pp. 5, 6).

[9] Richard Durbin, Sean Eddy, Anders Stærmose Krogh, and Graeme Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998 (cit. on pp. 6–8, 25).

[10] Robert D Finn, Jody Clements, and Sean R Eddy. "HMMER web server: interactive sequence similarity searching." In: *Nucleic acids research* 39.suppl_2 (2011), W29–W37. DOI: 10.1093/nar/gkr367 (cit. on pp. 7, 8).

[11] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." In: *Nature methods* 9.2 (2012), p. 173. DOI: 10.1038/nmeth.1818 (cit. on pp. 7, 9, 91).

[12] Sara El-Gebali, Lorna Richardson, and Rob Finn. *Detection of conserved evolutionary units by profile hidden Markov Models (HMM)*. URL: https://www.ebi.ac.uk/training/online/course/pfam-database-creating-protein-families (cit. on p. 8).

[13] Sean Eddy. *HMMER user's guide*. URL: http://hmmer.org (cit. on pp. 8, 91).

[14] Sean R Eddy. "Where did the BLOSUM62 alignment score matrix come from?" In: *Nature biotechnology* 22.8 (2004), p. 1035. DOI: 10.1038/nbt0804-1035 (cit. on p. 9).

[15] Alex Bateman et al. "The Pfam protein families database." In: *Nucleic acids research* 32.suppl_1 (2004), pp. D138–D141. DOI: 10.1093/nar/gkh121 (cit. on p. 9).

[16] C Anders Olson, Nicholas C Wu, and Ren Sun. "A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain." In: *Current Biology* 24.22 (2014), pp. 2643–2651. DOI: 10.1016/j.cub.2014.09.072 (cit. on pp. 10, 93).

[17] Nathan J Rollins, Kelly P Brock, Frank J Poelwijk, Michael A Stiffler, Nicholas P Gauthier, Chris Sander, and Debora S Marks. "Inferring protein 3D structure from deep mutation scans." In: *Nature Genetics* (2019), p. 1. DOI: 10.1038/s41588-019-0432-9 (cit. on pp. 10, 93).

[18] Douglas M Fowler and Stanley Fields. "Deep mutational scanning: a new style of protein science." In: *Nature methods* 11.8 (2014), p. 801. DOI: 10.1038/nmeth.3027 (cit. on pp. 10, 33, 86).

[19] 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. DOI: 10.1002/prot.340180402 (cit. on pp. 10, 11, 16).

[20] 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 (cit. on pp. 10, 11, 16).

[21] 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 (cit. on p. 11).

[22] 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. DOI: 10.1002/prot.20098 (cit. on p. 11).

[23] 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 (2007), pp. 333–340. DOI: 10.1093/bioinformatics/btm604 (cit. on pp. 11, 27).

[24] 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. DOI: 10.1073/pnas.0805923106 (cit. on pp. 11, 13, 15, 16, 26, 27, 65, 79).

[25] Faruck Morcos et al. "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. DOI: 10.1073/pnas.1111471108 (cit. on pp. 11, 13, 15, 16, 21, 22, 25–27).

[26] Charles Darwin. *On the various contrivances by which British and foreign orchids are fertilised by insects*. John Murray, 1877 (cit. on p. 11).

[27] Ulisse Ferrari, Stéphane Deny, Matthew Chalk, Gašper Tkačik, Olivier Marre, and Thierry Mora. "Separating intrinsic interactions from extrinsic correlations in a network of sensory neurons." In: *Physical Review E* 98.4 (2018), p. 042410 (cit. on p. 13).

[28] Lorenzo Posani, Simona Cocco, Karel Ježek, and Rémi Monasson. "Functional connectivity models for decoding of spatial representations from hippocampal CA1 recordings." In: *Journal of Computational Neuroscience* 43.1 (2017), pp. 17–33 (cit. on p. 13).

[29] Elad Schneidman, Michael J Berry II, Ronen Segev, and William Bialek. "Weak pairwise correlations imply strongly correlated network states in a neural population." In: *Nature* 440.7087 (2006), p. 1007 (cit. on p. 13).

[30] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. "Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information." In: *Elife* 3 (2014), e02030. DOI: 10.7554/eLife.02030 (cit. on pp. 13, 65, 79, 81).

[31] Andrea Cavagna, Alessio Cimarelli, Irene Giardina, Giorgio Parisi, Raffaele Santagati, Fabio Stefanini, and Massimiliano Viale. "Scale-free correlations in starling flocks." In: *Proceedings of the National Academy of Sciences* 107.26 (2010), pp. 11865–11870 (cit. on p. 13).

[32] Andrea Cavagna, Irene Giardina, Alberto Orlandi, Giorgio Parisi, Andrea Procaccini, Massimiliano Viale, and Vladimir Zdravkovic. "The STARFLAG handbook on collective animal behaviour: Part I, empirical methods." In: *arXiv preprint arXiv:0802.1668* (2008) (cit. on p. 13).

[33] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. "Inverse statistical problems: from the inverse Ising problem to data science." In: *Advances in Physics* 66.3 (2017), pp. 197–261. DOI: 10.1080/00018732.2017.1341604 (cit. on pp. 13, 19).

[34] Edwin T Jaynes. "Information theory and statistical mechanics." In: *Physical review* 106.4 (1957), p. 620 (cit. on p. 14).

[35] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. "How pairwise coevolutionary models capture the collective residue variability in proteins?" In: *Molecular biology and evolution* 35.4 (2018), pp. 1018–1027. DOI: 10.1093/molbev/msy007 (cit. on pp. 17, 18, 23, 25).

[36] Erik van Nimwegen. "Inferring contacting residues within and between proteins: what do the probabilities mean?" In: *PLoS computational biology* 12.5 (2016), e1004726. DOI: 10.1371/journal.pcbi.1004726 (cit. on p. 17).

[37] Erik Aurell. "The maximum entropy fallacy redux?" In: *PLoS computational biology* 12.5 (2016), e1004777. DOI: 10.1371/journal.pcbi.1004777 (cit. on p. 17).

[38] Chen-Yi Gao, Fabio Cecconi, Angelo Vulpiani, Haijun Zhou, and Erik Aurell. "DCA for genome-wide epistasis analysis: the statistical genetics perspective." In: *Physical biology* (2019) (cit. on pp. 17, 36).

[39] 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. DOI: 10.1016/S0364-0213(85)80012-4 (cit. on p. 20).

[40] 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. DOI: 10.1073/pnas.1508584112 (cit. on p. 21).

[41] Allan Haldane, William F Flynn, Peng He, RSK Vijayan, and Ronald M Levy. "Structural propensities of kinase family proteins from a Potts model of residue co-variation." In: *Protein Science* 25.8 (2016), pp. 1378–1384. DOI: 10.1002/pro.2954 (cit. on p. 21).

[42] Pierre Barrat-Charlaix, Matteo Figliuzzi, and Martin Weigt. "Improving landscape inference by integrating heterogeneous data in the inverse Ising problem." In: *Scientific reports* 6 (2016), p. 37812. DOI: 10.1038/srep37812 (cit. on p. 21).

[43] Julian Besag. "Statistical analysis of non-lattice data." In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 24.3 (1975), pp. 179–195 (cit. on p. 22).

[44] 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. DOI: 10.1214/09-AOS691 (cit. on p. 22).

[45] 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. DOI: 10.1016/j.jcp.2014.07.024 (cit. on pp. 22, 70, 79).

[46] 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. DOI: 10.1103/PhysRevE.87.012707 (cit. on pp. 22, 26, 70, 79).

[47] John P Barton, Simona Cocco, E De Leonardis, and Rémi Monasson. "Large pseudocounts and l2-norm penalties are necessary for the mean-field inference of Ising and Potts models." In: *Physical Review E* 90.1 (2014), p. 012132. DOI: 10.1103/PhysRevE.90.012132 (cit. on p. 24).

[48] Lukas Burger and Erik Van Nimwegen. "Disentangling direct from indirect co-evolution of residues in protein alignments." In: *PLoS computational biology* 6.1 (2010), e1000633. DOI: 10.1371/journal.pcbi.1000633 (cit. on p. 27).

[49] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. "Inverse statistical physics of protein sequences: a key issues review." In: *Reports on Progress*

in Physics 81.3 (2018), p. 032601. DOI: 10.1088/1361-6633/aa9965 (cit. on pp. 27, 28, 34–36).

[50] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A Pavlopoulos, David E Kim, Hetunandan Kamisetty, Nikos C Kyrpides, and David Baker. "Protein structure determination using metagenome sequence data." In: *Science* 355.6322 (2017), pp. 294–298. DOI: 10.1126/science.aah4043 (cit. on p. 29).

[51] 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 (cit. on pp. 29, 81).

[52] Joerg Schaarschmidt, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Alexandre MJJ Bonvin. "Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age." In: *Proteins: Structure, Function, and Bioinformatics* 86 (2018), pp. 51–66. DOI: 10.1002/prot.25407 (cit. on pp. 29, 67).

[53] Mu Gao, Hongyi Zhou, and Jeffrey Skolnick. "DESTINI: A deep-learning approach to contact-driven protein structure prediction." In: *Scientific reports* 9.1 (2019), p. 3514. DOI: 10.1038/s41598-019-40314-1 (cit. on pp. 30, 66).

[54] Thomas Gueudré, Carlo Baldassi, Marco Zamparo, Martin Weigt, and Andrea Pagnani. "Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis." In: *Proceedings of the National Academy of Sciences* 113.43 (2016), pp. 12186–12191. DOI: 10.1073/pnas.1607570113 (cit. on pp. 31, 32, 111).

[55] Anne-Florence Bitbol, Robert S Dwyer, Lucy J Colwell, and Ned S Wingreen. "Inferring interaction partners from protein sequences." In: *Proceedings of the National Academy of Sciences* 113.43 (2016), pp. 12180–12185. DOI: 10.1073/pnas.1606762113 (cit. on p. 31).

[56] 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. DOI: 10.1371/journal.pone.0149166 (cit. on p. 31).

[57] Guido Uguzzoni, Shalini John Lovis, Francesco Oteri, Alexander Schug, Hendrik Szurmant, and Martin Weigt. "Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis." In: *Proceedings of the National Academy of Sciences* 114.13 (2017), E2662–E2671. DOI: 10.1073/pnas.1615068114 (cit. on pp. 33, 65).

[58] 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 (2015), pp. 268–280. DOI: 10.1093/molbev/msv211 (cit. on pp. 33, 85, 87–89).

[59] 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. DOI: 10.1016/j.immuni.2012.11.022 (cit. on p. 33).

[60] 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. DOI: 10.1073/pnas.1413575111 (cit. on p. 33).

[61] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. "Mutation effects predicted from sequence covariation." In: *Nature biotechnology* 35.2 (2017), p. 128. DOI: 10.1038/nbt.3769 (cit. on pp. 33, 88).

[62] Christoph Feinauer and Martin Weigt. "Context-aware prediction of pathogenicity of missense mutations involved in human disease." In: *arXiv preprint arXiv:1701.07246* (2017) (cit. on pp. 33, 88).

[63] 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), p. 512. DOI: 10.1038/nature03991 (cit. on pp. 34, 35).

[64] Simona Cocco and Rémi Monasson. "Adaptive cluster expansion for inferring Boltzmann machines with noisy data." In: *Physical review letters* 106.9 (2011), p. 090601. DOI: 10.1103/PhysRevLett.106.090601 (cit. on pp. 34, 35).

[65] John P Barton, Eleonora De Leonardis, Alice Coucke, and Simona Cocco. "ACE: adaptive cluster expansion for maximum entropy graphical model inference." In: *Bioinformatics* 32.20 (2016), pp. 3089–3097. DOI: 10.1093/bioinformatics/btw328 (cit. on pp. 34, 35).

[66] Santeri Puranen, Maiju Pesonen, Johan Pensar, Ying Ying Xu, John A Lees, Stephen D Bentley, Nicholas J Croucher, and Jukka Corander. "SuperDCA for genome-wide epistasis analysis." In: *Microbial genomics* 4.6 (2018) (cit. on p. 36).

[67] Marcin J Skwark et al. "Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis." In: *PLoS genetics* 13.2 (2017), e1006508 (cit. on p. 36).

[68] Benjamin Schubert, Rohan Maddamsetti, Jackson Nyman, Maha R Farhat, and Debora S Marks. "Genome-wide discovery of epistatic loci affecting antibiotic resistance in Neisseria gonorrhoeae using evolutionary couplings." In: *Nature microbiology* 4.2 (2019), p. 328 (cit. on p. 36).

[69] Yujun Cui, Chao Yang, Hongling Qiu, Hui Wang, Ruifu Yang, and Daniel Falush. "The landscape of coadaptation in Vibrio parahaemolyticus." In: *bioRxiv* (2019), p. 373936 (cit. on p. 36).

[70] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. "A comprehensive two-hybrid analysis to explore the yeast protein interactome." In: *Proceedings of the National Academy of Sciences* 98.8 (2001), pp. 4569–4574 (cit. on p. 41).

[71] Yuen Ho et al. "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." In: *Nature* 415.6868 (2002), p. 180 (cit. on p. 41).

[72] Pascal Braun et al. "An experimentally derived confidence score for binary protein-protein interactions." In: *Nature methods* 6.1 (2009), p. 91 (cit. on p. 41).

[73] Matteo Pellegrini, Edward M Marcotte, Michael J Thompson, David Eisenberg, and Todd O Yeates. "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." In: *Proceedings of the National Academy of Sciences* 96.8 (1999), pp. 4285–4288. DOI: 10.1073/pnas.96.8.4285 (cit. on p. 41).

[74] Matteo Pellegrini. "Using phylogenetic profiles to predict functional relationships." In: *Bacterial molecular networks*. Springer, 2012, pp. 167–177. DOI: 10.1007/978-1-61779-361-5_9 (cit. on p. 41).

[75] Thomas A Hopf, Charlotta PI Schärfe, João PGLM Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre MJJ Bonvin, and Debora S Marks. "Sequence co-evolution gives 3D contacts and structures of protein complexes." In: *Elife* 3 (2014), e03430. DOI: 10.7554/eLife.03430 (cit. on pp. 65, 79).

[76] Duccio Malinverni, Simone Marsili, Alessandro Barducci, and Paolo De Los Rios. "Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones." In: *PLoS computational biology* 11.6 (2015), e1004262. DOI: 10.1371/journal.pcbi.1004262 (cit. on pp. 65, 79).

[77] Tian-ming Zhou, Sheng Wang, and Jinbo Xu. "Deep learning reveals many more inter-protein residue-residue contacts than direct coupling analysis." In: *bioRxiv* (2018), p. 240754 (cit. on pp. 65, 69, 79, 119).

[78] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. "Accurate de novo prediction of protein contact map by ultra-deep learning model." In: *PLoS computational biology* 13.1 (2017), e1005324. DOI: 10.1371/journal.pcbi.1005324 (cit. on pp. 66, 67, 119).

[79] R Evans et al. "De novo structure prediction with deeplearning based scoring." In: *Annu Rev Biochem* 77.363-382 (2018), p. 6 (cit. on p. 66).

[80] Mohammed AlQuraishi. "AlphaFold at CASP13." In: *Bioinformatics* (2019). DOI: 10.1093/bioinformatics/btz422 (cit. on p. 66).

[81] David T Jones, Tanya Singh, Tomasz Kosciolek, and Stuart Tetchner. "MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins." In: *Bioinformatics* 31.7 (2014), pp. 999–1006. DOI: 10.1093/bioinformatics/btu791 (cit. on p. 67).

[82] Neena Aloysius and M Geetha. "A review on deep convolutional neural networks." In: *2017 International Conference on Communication and Signal Processing (ICCSP)*. IEEE. 2017, pp. 0588–0592. DOI: 10.1109/ICCSP.2017.8286426 (cit. on p. 66).

[83] Marcin J Skwark, Daniele Raimondi, Mirco Michel, and Arne Elofsson. "Improved contact predictions using the recognition of protein like contact patterns." In: *PLoS computational biology* 10.11 (2014), e1003889. DOI: 10.1371/journal.pcbi.1003889 (cit. on p. 66).

[84] Roberto Mosca, Arnaud Ceol, Amelie Stein, Roger Olivella, and Patrick Aloy. "3did: a catalog of domain-based interactions of known three-dimensional structure." In: *Nucleic acids research* 42.D1 (2013), pp. D374–D379. DOI: 10.1093/nar/gkt887 (cit. on pp. 69, 70).

[85] Robbie P Joosten, Tim AH Te Beek, Elmar Krieger, Maarten L Hekkelman, Rob WW Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. "A series of PDB related databases for everyday needs." In: *Nucleic acids research* 39.suppl_1 (2010), pp. D411–D419. DOI: 10.1093/nar/gkq1105 (cit. on p. 71).

[86] Wolfgang Kabsch and Christian Sander. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." In: *Biopolymers: Original Research on Biomolecules* 22.12 (1983), pp. 2577–2637. DOI: 10.1002/bip.360221211 (cit. on p. 71).

[87] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 77).

[88] Mirco Michel, David Menéndez Hurtado, and Arne Elofsson. "PconsC4: fast, accurate and hassle-free contact predictions." In: *Bioinformatics* (2018). DOI: 10.1093/bioinformatics/bty1036 (cit. on pp. 79, 119).

[89] Alejandro Couce, Larissa Viraphong Caudwell, Christoph Feinauer, Thomas Hindré, Jean-Paul Feugeas, Martin Weigt, Richard E Lenski, Dominique Schneider, and Olivier Tenaillon. "Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria." In: *Proceedings of the National Academy of Sciences* 114.43 (2017), E9026–E9035 (cit. on pp. 85, 88, 89, 92, 101).

[90] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*. Vol. 1. na, 1932 (cit. on p. 85).

[91] Michael J Liao, M Omar Din, Lev Tsimring, and Jeff Hasty. "Rock-paper-scissors: Engineered population dynamics increase genetic stability." In: *Science* 365.6457 (2019), pp. 1045–1049 (cit. on p. 85).

[92] Michael J Wiser, Noah Ribeck, and Richard E Lenski. "Long-term dynamics of adaptation in asexual populations." In: *Science* 342.6164 (2013), pp. 1364–1367 (cit. on p. 89).

[93] Heewook Lee, Ellen Popodi, Haixu Tang, and Patricia L Foster. "Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing." In: *Proceedings of the National Academy of Sciences* 109.41 (2012), E2774–E2783 (cit. on p. 89).

[94] Patricia L Foster, Heewook Lee, Ellen Popodi, Jesse P Townes, and Haixu Tang. "Determinants of spontaneous mutation in the bacterium Escherichia coli as revealed by whole-genome sequencing." In: *Proceedings of the National Academy of Sciences* 112.44 (2015), E5990–E5999 (cit. on p. 89).

[95] Michael S Breen, Carsten Kemena, Peter K Vlasov, Cedric Notredame, and Fyodor A Kondrashov. "Epistasis as the primary factor in molecular evolution." In: *Nature* 490.7421 (2012), p. 535 (cit. on p. 90).

[96] Michael J Harms and Joseph W Thornton. "Evolutionary biochemistry: revealing the historical and physical causes of protein properties." In: *Nature Reviews Genetics* 14.8 (2013), p. 559 (cit. on p. 90).

[97] J Arjan Gm De Visser and Joachim Krug. "Empirical fitness landscapes and the predictability of evolution." In: *Nature Reviews Genetics* 15.7 (2014), p. 480 (cit. on p. 90).

[98] Anna I Podgornaia and Michael T Laub. "Pervasive degeneracy and epistasis in a protein-protein interface." In: *Science* 347.6222 (2015), pp. 673–677 (cit. on p. 90).

[99] Kaleb Z Zion Abram, Zulema Udaondo, Carissa Bleker, Visanu Wanchai, Trudy M Wassenaar, and David W Ussery. "What can we learn from over 100,000 Escherichia coli genomes?" In: *bioRxiv* (2019), p. 708131 (cit. on p. 92).

[100] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28 (cit. on p. 120).

[101] 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. DOI: 10.1371/journal.pone.0092721 (cit. on p. 120).

COLOPHON

This document was typeset using `classicthesis` style developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*.
Final Version as of November 20, 2019 (classicthesis v4.6).

Giancarlo CROCE **22 novembre 2019**

Sujet : Vers une analyse co-évolutive à l'échelle du génome

Résumé : Les progrès des technologies de séquençage ont révolutionné les sciences de la vie. L'explosion de données de séquences génomiques a conduit au développement d'une grande variété de méthodes, à l'interface entre la bioinformatique, l'apprentissage automatique et la physique, qui visent à approfondir la compréhension des systèmes biologiques à partir de telles données. Les méthodes co-évolutives par paires, telles que l'analyse par couplage direct (DCA), peuvent extraire une multitude d'informations à partir de données de séquence uniquement, telles que des contacts structurels ou des effets phénotypiques de substitutions d'acides aminés dans des protéines. Bien qu'elles aient été principalement appliquées à un certain nombre de protéines exemplaires, il est maintenant temps de les appliquer au niveau du génome entier.
Dans cette thèse, nous nous appuions sur ces modèles et les développons pour traiter des questions biologiques à l'échelle du génome. Dans un premier projet, nous avons étudié le réseau d'interactions protéine-protéine en combinant des signaux co-évolutifs à des échelles multiples mais interconnectées. Dans un projet ultérieur, nous discutons de la possibilité d'inclure des informations complémentaires aux séquences, telles que des schémas de contacts typiques, afin d'améliorer la prédiction de contacts entre protéines. Enfin, à travers une vaste étude portant sur l'ensemble du génome des souches d'*E. Coli*, nous montrons comment les mécanismes de la DCA peuvent être utilisés pour étudier les propriétés du paysage de la fitness à l'échelle locale et globale.

Subject: Towards a genome-scale coevolutionary analysis

Abstract: Advances in sequencing technologies have revolutionized the life sciences. The explosion of genomic sequence data has prompted the development of a wide variety of methods, at the interface between bioinformatics, machine learning, and physics, which aim at gaining a deeper understanding of biological systems from such data.Pairwise coevolutionary methods, in particular Direct Coupling Analysis (DCA), can extract a multitude of information from sequence data alone, such as structural contacts or phenotypic effects of amino-acid substitutions in proteins. While they have been mainly applied to a number of single exemplary proteins, it is now time for a broader application at the level of the whole genome.
In this thesis, we build upon and extend these models to address biological questions at the genome scale. In a first project, we investigate the protein-protein interaction network by combining coevolutionary signals at multiple but interconnected scales. In a subsequent project, we discuss the possibility of including complementary information to sequences, such as typical patterns of contacts, to improve the inter-protein contact prediction. Finally, through an extensive genome-wide study of *E. coli* strains, we show how the machinery of DCA can be used to investigate the fitness landscape properties at the local and global scales.