Statistical Analysis of Protein Sequences: A Coevolutionary Study of Molecular Chaperones

THÈSE N^{O} 8463 (2018)

PRÉSENTÉE LE 23 MARS 2018
À LA FACULTÉ DES SCIENCES DE BASE
LABORATOIRE DE BIOPHYSIQUE STATISTIQUE
PROGRAMME DOCTORAL EN PHYSIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Duccio MALINVERNI

acceptée sur proposition du jury:

Prof. P. Ricci, président du jury
Prof. P. De Los Rios, Dr A. Barducci, directeurs de thèse
Dr S. Cocco, rapporteuse
Prof. F. Morcos, rapporteur
Prof. M. Wyart, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2018

À Noupette

Acknowledgements

During these four passed years, I had the chance of sharing my time with great people, both inside the lab and in the hours spent outside of the academic world. These lines are devoted to you all.

I want to thank Paolo De Los Rios who gave me the opportunity to do my thesis at EPFL. As a supervisor, he taught me the importance of optimism in science, to keep believing in a positive outcome and shared his genuine curiosity with me. Paolo also introduced me to the world of biology, which was a daunting and unknown world to me. A great thank you to Alessandro Barducci, who along the road took up the burden of being my co-supervisor. Ale was my absolute guru regarding all aspects of molecular simulations, and continuously answered all questions I had with great patience and humor.

A lab wouldn't be a lab without all of its members. I had the chance to share these four past years with Alberto Sassi and Andrea Martini, with whom I discovered the real meaning of the term *Brothers in arms*. Beyond sharing our office, we went through these passed years in good and bad times, discovering unexpected aspects of the academic world, managing to keep up a good mood in the office through mostly absurd philosophic/scientific/procrastinating discussions. A great thank you to all further members of the LBS, both past and present: Salvo Assenza, who introduced me to being PhD student in the group, Michele Rizzi who is still leading 7-6 in our badminton head-to-head, Stefano Zamuner who is always willing to discuss crazy ideas and to Alessio Cardillo for moral support. Thank you to Riccardo Ravasio for many discussions and proofreading parts of this thesis.

While spending most of my time in a PhD bubble, I was lucky to have friends on the outside, who took it upon them to remind me that a real world still exists. Un merci spécial à Fred et Morgane qui ont partagé mon parcours Lausannois dès les premiers jours. A David *Walter* Forchelet, tu as toujours été dispo pour un café et pour refaire le monde. Merci encore à Dom et Marica et à Charles pour les moments passés ensembles.

Un pensiero forte va alla mia famiglia che mi ha sempre sostenuto e appoggiato durante questi quasi dieci anni a Losanna. Un grossissimo grazie ai miei genitori e alle due sorelle. VVTB.

Last but not least, je dois mes remerciements les plus profonds à ma femme Mélanie, qui m'a porté et supporté durant toute cette thèse. Merci d'avoir été à mes cotés durant cette aventure. Madokken.

Lausanne, December 2017 D. M.

Abstract

Recent advances in DNA sequencing technologies led to the accumulation of enormous quantities of genetic information available in public databases. This rapid growth of available biological datasets calls for quantitative analysis tools and concomitantly opens the doors for new analysis paradigms. Particularly, the analysis of correlated mutations and their structural interpretation have witnessed a second youth in the last years. A natural formulation for such approaches is provided by the statistical physics of disordered systems. This thesis is articulated around different projects aimed at studying particular biological systems of interests, the Hsp70 molecular chaperones, through the lens provided by methods rooted in statistical physics. In a first project, we focus on correlated mutations within the Hsp70 family. Our analysis reveals the existence of a biologically important macro-molecular arrangement of these chaperones and we investigate its phylogenetic origin. A second project investigates the interactions between the Hsp70 chaperones and one of their main co-chaperones, J-proteins. Through the combined use of coevolutionary analysis and molecular simulations at both coarse-grained and atomistic levels, we construct a structural and dynamical model of this interaction which rationalizes previous experimental evidence. In a subsequent study, we specifically focus on the J-protein co-chaperones. Through phylogenetic and coevolutionary methods, we investigate the origin of recently discovered interactions which form the basis of the disaggregation machinery in higher eukaryotes. Finally, in a fourth project, we shift our attention to the analysis of proteins involved in the iron-sulfur cluster assembly pathway. Analysis of residue coevolution in the different proteins composing this pathway reveals multiple structural insights at several scales.

Key Words: Correlated Mutations | Direct-Coupling Analysis | Statistical Inference | Probabilistic modeling | Molecular Chaperones | Hsp70 | Iron-Sulfur Cluster Assembly

Résumé

Des avancées technologiques récentes dans le domaine du séquençage d'ADN ont mené à l'accumulation d'énormes quantités d'informations génétiques accessibles dans des bases de données publiques. Cette rapide croissance des banques de données biologiques requiert des outils d'analyse quantitatifs, tout en ouvrant la porte à de nouveaux paradigmes d'étude. En particulier, l'analyse de mutations corrélées et leur interprétation structurelle ont récemment regagné en popularité. Une formulation naturelle de ces approches est trouvée dans la physique statistique des systèmes désordonnés. Cette thèse est articulée autour de différents projets visant à étudier des systèmes biologiques spécifiques, les chaperons moléculaires Hsp70, à l'aide de méthodes trouvant leur origine en physique statistique. Dans un premier projet, nous inspectons les mutations corrélées dans la famille de protéines Hsp70. Nos résultats révèlent l'existence d'un complexe macro-moléculaire de cette protéine, et nous étudions son origine phylogénétique. Un second projet est axé sur l'étude des intéractions entre le chaperone Hsp70 et l'un de ses co-chaperons, les J-protéines. En combinant l'analyse de mutations corrélées avec des simulations moléculaires au niveau atomistique et gros-grain, nous construisons un modèle structurel et dynamique de cette intéraction, qui rationalise des données expérimentales existantes. Dans un projet suivant, nous nous concentrons sur les co-chaperons J-protéines. A l'aide de méthodes de coévolution et phylogénétiques, nous étudions l'origine d'une intéraction récemment découverte qui forme la base de la machine de disaggregation des eukaryotes supérieurs. Finalement, un quatrième projet est consacré à l'étude de protéines impliquées dans l'assemblage de cluster Fer-Soufre. L'analyse de coévolutions de diverses protéines impliquées révèle des détails structurels à différentes échelles.

Mots-clés : Mutations corrélées | Analyse de Couplages Directs | Inférence Statistique | Modélisation probabiliste | Chaperons moléculaires | Hsp70 | Assemblage de Cluster Fer-Soufre

Contents

Acknowledgements **v**

Abstract (English) **vii**

Abstract (French) **ix**

List of figures **xii**

List of tables **xiv**

Introduction **1**

1 Methods of coevolutionary analysis **5**
- 1.1 Correlated mutations in homologous proteins . . . 5
- 1.2 Direct Coupling Analysis . . . 6
  - 1.2.1 Maximum Entropy Modeling . . . 6
  - 1.2.2 Scoring functions . . . 8
  - 1.2.3 Free parameters and Gauge invariance . . . 8
  - 1.2.4 Sequence reweighting . . . 9
- 1.3 Parameter Inference . . . 10
  - 1.3.1 General overview . . . 10
  - 1.3.2 Mean-Field solution . . . 10
  - 1.3.3 Pseudo-Likelihood maximization . . . 12
  - 1.3.4 Boltzmann Machine Learning . . . 15
  - 1.3.5 Other approaches . . . 16

2 Molecular chaperones **19**
- 2.1 Overview of Molecular chaperones . . . 19
- 2.2 The Hsp70 chaperone system . . . 20
  - 2.2.1 J-proteins . . . 22
  - 2.2.2 Nucleotide Exchange Factors . . . 23

3 Coevolutionary analysis of the Hsp70 family **25**
- 3.1 Introduction . . . 25
- 3.2 Dataset preparation . . . 26

3.3 Direct-Coupling Analysis of the Hsp70 family 27
3.4 Hsp70 Homo-Dimerization 28
3.5 Experimental evidence of Hsp70 homo-dimerization 33
3.6 Discussion 34

4 Hsp70-Hsp40 interactions 37
4.1 Introduction 37
4.2 Experimental models of the Hsp40-Hsp70 interaction 38
4.3 Extending DCA for protein-protein interactions 39
  4.3.1 The Random Matching Strategy 41
4.4 Random matching on the Hsp40-Hsp70 system 43
  4.4.1 RM predictions of the complex 43
  4.4.2 Comparison with IPA and PPM 45
4.5 Coarse-grained modeling of the interaction 46
  4.5.1 Simulation protocol 46
  4.5.2 CG modeling of the Hsp40-Hsp70 interaction 48
4.6 Atomistic Simulations 51
4.7 Role of the SBD in the Hsp40-Hsp70 interactions 54
4.8 Differential binding of bacterial and eukaryotic pairs 55
4.9 Discussion 58
4.10 Supplementary Figures and Tables 60

5 Synergistic action of class A/B Hsp40s 63
5.1 Introduction 63
5.2 Sequence extraction and pre-processing 65
5.3 Phylogenetic analysis of the synergistic interaction 66
  5.3.1 Taxonomic distribution of J-proteins 66
  5.3.2 Phylogenetic signature of the J-protein synergy 67
5.4 Coevolutionary Analysis 74
5.5 Discussion 75

6 Coevolutionary study of Iron-Cluster pathway proteins 77
6.1 Introduction 77
6.2 Structural insights into the IscU family 79
6.3 Multimerization and cluster coordination of IscA 82
6.4 Interactions between Fe-S cluster proteins 83
6.5 Conclusions 86
6.6 Supplementary Figures 88

7 Conclusions and Outlooks 89

Bibliography 103

Curriculum Vitae 105

List of Figures

1.1 Correlated mutations and evolution of the Uniprot database 6

2.1 ADP and ATP structures of *E. coli* DnaK 21
2.2 Biochemical cycle of Hsp70 22

3.1 Multiple conformations of Hsp70 and their contact maps. 26
3.2 Precision of DCA predictions on the isolated Hsp70 family. 28
3.3 DCA predictions of multiple Hsp70 conformations. 29
3.4 DCA predicted homo-dimeric arrangement of Hsp70. 30
3.5 DCA predictions restricted to canonically tagged Hsp70 sequences. 31
3.6 PCA of the Hsp70 family and phylogenetic reweighting. 33
3.7 Experimental evidence of the Hsp70 dimerization 34

4.1 Structural organization of canonical J-proteins 38
4.2 Principle of DCA applied to protein-protein interactions. 40
4.3 Distribution of Hsp40 and Hsp70 paralogs per organism. 43
4.4 Random Matching results on the Hsp40-Hsp70 interaction. 45
4.5 Analysis of lower ranked RM predictions. 46
4.6 Comparison of the RM, IPA and PPM methods on Hsp40/Hsp70. 47
4.7 Coarse-grained modeling of the Hsp40-Hsp70 interaction. 50
4.8 HPD-IN and HPD-OUT conformations predicted by the CG model 51
4.9 Comparison of the RM and CG predicted Hsp40-Hsp70 interface. 51
4.10 1μs time-scale atomistic simulations of Hsp40-Hsp70. 53
4.11 GBSA Analysis of the Hsp40-Hsp70 complex 56
4.12 Random matching results on Bacterial and Eukaryotic subsets. 57
4.13 Random matching including Full-Length Hsp70 sequences 60
4.14 DRMS of atomistic simulations of the Hsp40-Hsp70 complexes 61
4.15 Angular fluctuations of atomistic simulations of the Hsp40-Hsp70 complexes . 62

5.1 Biochemical evidence for synergistic action of class A and B J-proteins. 65
5.2 Phylogenetic trees of the J-domains and CTDs of class A and B J-proteins. . . . 68
5.3 PDA results on class A and B J-proteins. 71
5.4 Structural mapping of PDA and DCA predictions on class A and B J-proteins. . 73
5.5 Robustness analysis of the PDA approach. 74

5.6 Random Matching results on class A and B J-protein interactions. 76

6.1 Schematics of Fe-S cluster assembly. 78
6.2 N-terminal interactions of Isc. 80
6.3 Fe-S cluster coordination of IscU. 81
6.4 IscA multimerization. 83
6.5 Inter-protein interactions of Fe-S cluster pathway proteins. 86
6.6 IscU Tetramer and Decamer arrangements. 88

List of Tables

4.1 Taxonomic distributions of Hsp40 and Hsp70 sequences. 44
4.2 Most frequently appearing contacts in the RM procedure. 45
4.3 Hsp40-Hsp70 dissociation constants computed by the CG model. 48
4.4 Comparison of the RM, IPA and PPM methods on Hsp40/Hsp70 60

Introduction

Proteins are fundamental building blocks upon which living cells are constructed, existing in a broad variety of shapes and sizes. The human genome, for instance, contains genes encoding approximatively 20'000 different proteins, with molecular weights ranging from ~ 20 to 4000 kDA. This large variety is contrasted by the low number of their constituent components. Proteins are linear heteropolymers, formed by the concatenation of 20 different types of amino-acids, which form their monomers. The linear arrangement of amino-acids forms the proteins *primary sequence*, which is encoded in the DNA of each cell. Under normal conditions, proteins generally fold into well defined three-dimensional structures, the precise form of which is specifically determined by their primary sequences. The functional role played by proteins is mostly determined by their three-dimensional structure, which dictates the biochemical activity and interactions with other molecular components of the cell. Understanding the functioning and organization of living organisms at a molecular level thus strongly relies on the knowledge of proteins native folds.

The experimental determination of protein structure, which forms the field of structural biology, is a longstanding and challenging task. Experimental methods to determine protein structures, such as X-ray crystallography, NMR spectroscopy or Cryo-electron microscopy form todays golden standard. These have been extremely efficient in obtaining three-dimensional structures of proteins with up to atomistic resolution. However, the experimental determination of protein structures is a time consuming and difficult task. Particularly difficult cases, which encompass membrane proteins, low affinity complexes or highly dynamic structures, still present experimental difficulties which cannot systematically be overcome. To partially circumvent these limitations, the last decades have witnessed the emergence of computational structural biology, which aims at complementing the experimental structural determination with *in silico* methods. While the computational *de novo* prediction of protein structures from their sequence alone, considered as the holy grail of the field, is still out of reach of current methods, a plethora of semi-empirical approaches have been developed to tackle this problem, with increasing success.

In contrast to protein structure determination, the sequencing of DNA has seen huge improvements in the past years. Indeed, technological breakthroughs in high-throughput sequencing have led to the availability at relatively low cost of high-performance sequencing facilities, which resulted in an exceptional growth of sequenced proteins deposited in publicly available databases. The emergence of extremely large databases, which projected computational biology in the era of Big-Data, opened the door to new analysis paradigms. In particular, the availability of large *protein families*, i.e collections of homologous protein sequences across multiple organisms, allowed the possibility to precisely measure *correlated mutations* which could be used to infer structural knowledge.

This availability of large sequence datasets calls for quantitative analysis tools. In particular, the intrinsically heterogeneous nature of proteins, both in terms of their constituent amino-acids and in terms of their interactions, finds a natural formulation in physics, specifically in the field of statistical physics of disordered systems. Indeed, recent years have seen the appearance of extremely efficient computational methods, based on statistical physics models, aimed at modeling correlated mutations in protein families (see chapter 1). These methods are illustrative of a strong phenomena of convergence of the biological and physical fields. Combined with a steady increase in computational power and the development of novel algorithmic approaches to tackle complex inverse problems, the statistical physics approach to model biological complex systems has become a well anchored domain of biophysics. It is noteworthy to underline that the application of statistical physics methods to analyze biological data goes well beyond the field of molecular biology. Indeed, similar approaches have been successfully applied to study a wide spectrum of biological systems, ranging from the collective behavior of flocking birds [1, 2] to the analysis of firing patterns of interacting neurons in neuronal recordings [3, 4]. These seemingly distant subjects underline a core property of the physics approach to biological problems, which consists in making abstraction of the first layer of details of a system, in order to focus on the universal basic properties of interacting complex systems (see e.g. [5] for a broad review of the field).

This coarse-grained view of physicists contrasts with the traditional approach in biology, which focuses attention at a much finer-grained level of details, through the precise study of particular systems of interest. Embracing these two approaches is a challenging task faced by biophysicists. In this thesis, we propose to combine these two approaches, throughout the use of models drawn from statistical physics to the detailed analysis of a particular class of ubiquitous proteins. Specifically, we will focus on the study of molecular chaperones, a class of proteins involved in the regulation and control of the cellular protein population. Cells must in general cope with external stresses, may they be chemical, thermal or osmotic, which can induce deregulations in the cellular protein populations. Typical consequences, such as protein denaturation, misfolding and aggregation can have potentially deleterious effects on the survival of organisms. It is thus not surprising that early on in evolution, all organisms have acquired specialized protein machineries specifically targeted at *proteostasis*. i.e. the control and maintenance of the cellular protein population in a functional state. Since their discovery in the seventies, the ubiquity and central importance of molecular chaperones stimulated abundant experimental and theoretical studies in order to understand the molecular basis of these machines. However, despite four decades of active research on chaperones, their complexity left many fundamental questions unanswered.

The main objective of this thesis is the study of molecular chaperones through the use of coevolutionary methods. This objective will be tackled by a series of projects, through each of which we will investigate a different aspect of a particular chaperone, namely Hsp70 (see chapter 2 for an overview of molecular chaperones). Our results presented in the following chapters, taken together, extend our current structural and functional knowledge of this class of proteins, thereby contributing towards the quest of the overall better understanding of the molecular processes of proteostasis, a fundamental mechanism shared by all organisms.

The following of this thesis is organized as follows: Chapters one and two form an extended introduction. In Chapter 1, we will review the theoretical basis of the coevolutionary methods used throughout this manuscript. Emphasis will be given on the mathematical formulation of the methodological tools employed in subsequent chapters. Chapter 2 introduces a broad overview of molecular chaperones with particular emphasis on the Hsp70 machinery, and will give a deeper introduction to the biological role played by molecular chaperones. Chapter 3 deals with the coevolutionary analysis of the Hsp70 chaperone, particularly investigating the nature of the homo-dimeric arrangement of this protein. In Chapter 4, we analyze the interactions between the Hsp70 chaperones and their co-chaperones Hsp40, by combined means of coevolutionary methods and molecular simulations at multiple scales. The fifth Chapter treats about the synergistic inter-class interaction between multiple Hsp40 proteins, which forms the basis of the disaggregation machinery in higher eukaryotes. In the last Chapter, the focus is shifted onto the analysis of proteins involved in the Iron-Sulfur cluster assembly pathway.

1 Methods of coevolutionary analysis

1.1 Correlated mutations in homologous proteins

In order to perform their biological function, proteins must fold to their native three dimensional structure[1]. Under the process of evolution, natural mutations will occur in the primary structure of the proteins, thereby modifying the amino-acid composition of protein sequences. However, in order to maintain a functional fold, deleterious mutations will be suppressed by natural selection, whereas beneficial and neutral mutations will be conserved and propagated. A potentially deleterious mutation can in principle be compensated by one or several mutations of amino-acids interacting in the three dimensional structure. Thus the presence of compensatory mutations of interacting amino-acids in proteins results in patterns of pairwise correlated mutations. This principle lies at the heart of coevolutionary methods. The idea of inverting the observed correlation patterns to infer structural contacts in protein families dates back at least to the end of the eighties [6, 7, 8], where it had been observed that pair of residues showing strongly correlated mutations were weakly enriched in structural contacts [7] (Fig.1.1a).

Initial approaches to infer structural contacts from residue covariation focused on measures of pairwise correlations, analyzing pairs of residues independently. A prototypical such approach is based on the use of mutual information (MI) to quantify the strength of covariation between a pair of residues [10, 11, 12]. While this approach benefitted from a remarkable success, partially pertaining to both its conceptual and practical simplicity, its performance in terms of contact prediction was only moderate. It has indeed been recognized quite early that models aiming to infer structural contacts focusing on independent pairs of residue would strongly suffer from the effect of mediated correlations [13], thus limiting their predictive power. Going beyond pairwise models required the conjunction of several factors: The rapid growth of available sequences in the last fifteen years (Fig.1.1b), combined with new theoretical insights led to the development of such new computational methods designed to account for chains of correlations, leading to a second youth of coevolutionary contact prediction methods.

---

[1] The case of intrinsically disordered proteins will not be discussed

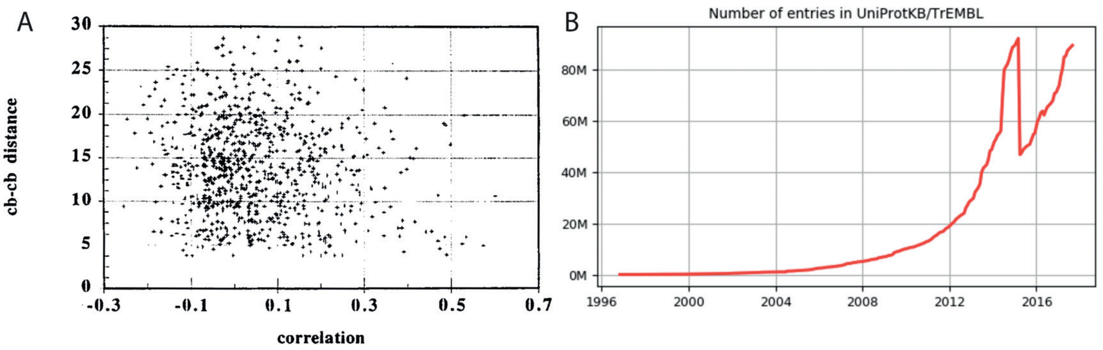

Figure 1.1 – **A** Relation between correlated mutations and inter-residue distances (Cβ - Cβ distance in Å). The mutation correlations are computed as the correlation coefficient between two positions in the MSA (Figure adapted from [7]). **B** Evolution of the number of deposited protein sequences in the UniprotKB/TrEMBL database [9].

1.2 Direct Coupling Analysis

To circumvent the inherent limitations of single-pair correlation based methods to infer structural contacts in proteins, Direct Coupling Analysis (DCA) has recently been introduced [14]. It has been noticed that the only way around the problem of mediated correlations was through the use of global statistical models [13, 14, 15]. In contrast to pairwise correlation based methods, global statistical model approaches aim to infer a joint probability distribution over the sequence space which best reproduces the observed correlations, thus naturally reproducing the mediated chains of correlations. There are *a priori* innumerable ways one can construct a global probabilistic model over the sequence space. In the following, we will overview a rational approach commonly used in statistical physics, namely the maximum entropy modeling.

1.2.1 Maximum Entropy Modeling

The approach followed in [14] to build a global statistical model on the sequence space is based on the maximum entropy principle [16], which seeks the least biased distribution subject to some constraints. In the case of DCA, the constraints are such that the inferred probability distribution must reproduce the single site and two-site marginal distributions of amino-acid compositions. The general procedure is outlined hereafter.

Let $\mathbf{X} = \{X_i\}_{i=1}^{N}$ denote a protein sequence of length N, where each residue position X_i can take one of $q = 21$ symbols (20 natural amino-acids + 1 alignement gap). Let $\{\mathbf{X_b}\}_{b=1}^{B}$ be a collection of aligned homologous proteins forming a multiple sequence alignment (MSA) of the protein family. We start by building the single- and two-site empirical marginal distributions

$$f_i(A) = \frac{1}{B}\sum_{b=1}^{B}\delta_{X_i^b,A} \qquad\qquad f_{ij}(A,B) = \frac{1}{B}\sum_{b=1}^{B}\delta_{X_i^b,A}\delta_{X_j^b,B} \tag{1.1}$$

where i, j denote the residue index position and A, B denote any of the q possible symbols. The maximum entropy principle seeks the least constrained joint probability distribution $P(\mathbf{X})$ that reproduces the marginals 1.1. In other words, the goal is to maximize

$$S = -\sum_{\mathbf{X}} P(\mathbf{X}) \log P(\mathbf{X}) \tag{1.2}$$

with the marginal constraints discussed above and a global normalization constraint. Putting all these ingredient together yields the constrained entropy to be maximized

$$\begin{aligned}
\tilde{S} = &-\sum_{\mathbf{X}} P(\mathbf{X}) \log \mathbf{P}(\mathbf{X}) \\
&+\sum_{i=1}^{N} \sum_{A_i=1}^{q} h_i(A_i) \left(\sum_{\{A_k | k \neq i\}} P(A_1, A_2, ..., A_N) - f_i(A_i) \right) \\
&+\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{A_i, A_j=1}^{q} J_{ij}(A_i, A_j) \left(\sum_{\{A_k | k \neq i, j\}} P(A_1, A_2, ..., A_N) - f_{ij}(A_i, A_j) \right) \\
&+\lambda \left(\sum_{\mathbf{X}} P(\mathbf{X}) - \mathbf{1} \right)
\end{aligned} \tag{1.3}$$

where $h_i(A_i)$ (resp. $J_{ij}(A_i, B_i)$) are the Lagrange multipliers enforcing the single- (resp. two-) site marginals and λ is the Lagrange multiplier enforcing the normalization of the distribution $P(\mathbf{X})$. Maximizing 1.3 with respect to $P(\mathbf{X})$ leads to

$$P(\mathbf{X}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} h_i(X_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(X_i, X_j) \right) \tag{1.4}$$

where $Z \equiv \exp(\lambda - 1)$ is the partition function imposing the normalization

$$Z = \sum_{\mathbf{X}} \exp\left(\sum_{i=1}^{N} h_i(X_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(X_i, X_j) \right) \tag{1.5}$$

Let us further define the Hamiltonian of the system as

$$\mathcal{H}(\mathbf{X}) = -\sum_{i=1}^{N} h_i(X_i) - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(X_i, X_j) \tag{1.6}$$

such that the distribution 1.4 takes a classical Boltzmann form $P(\mathbf{X}) = \frac{1}{Z} e^{-\mathcal{H}(\mathbf{X})}$. The model 1.4 is known in statistical physics as a generalized q-state Potts model [17] and forms the core of DCA methods. The parameters of the model are the local biases $h_i(A_i)$ which control the amino-acid compositions at sites i and the coupling parameters $J_{ij}(A_i, B_i)$ controlling the statistical coupling strengths between pair of sites i, j. This formulation allows chains of correlated mutations to be naturally generated by a potentially reduced set of inter-protein couplings J_{ij}, thus intrinsically accounting for their presence in the sequence data.

1.2.2 Scoring functions

It has been shown hat the direct couplings J_{ij} are efficient predictors of physical inter-residue contacts, largely outperforming correlation based predictors [14, 15, 18] . In order to rank the predicted contacts, the local $q \times q$ couplings matrices $J_{ij}(A,B)$ must be reduced to a scalar score characterizing the coupling strength of the contact, irrespective of the particular amino-acid types A, B in the contact. In [14, 15], the authors introduced the Direct Information (DI) score which measures the mutual information generated by the direct couplings J_{ij}. In [18], the authours have shown that the use of a simpler score, based on the Frobenius norm of the $q \times q$ local coupling matrices improved the predictive power. They defined the interaction score as

$$S_{ij} = \sqrt{\sum_{A,B}^{q,q} J_{ij}(A,B)^2} \tag{1.7}$$

Furthermore, a simple modification to the Frobenius norm 1.7 was introduced in [19], which significantly improved the predictions. The modification consists in neglecting the couplings with the gap symbol in the double summation in 1.7. Indeed, long stretches of gaps usually found at the two ends of MSAs introduce strong correlations between these positions, thus contributing artificially strong couplings, which are non informative about structural contacts. The simple exclusion of the couplings with the gap symbol thus partially removes this source of error in DCA predictions.

Furthermore, an empirical correction term called Average Product Correction [20] has been shown to additionally improve the quality of the predictions by partially removing some compositional biases of the interaction scores 1.7

$$S_{ij}^{APC} = S_{ij} - \frac{S_{i,\cdot} S_{\cdot,j}}{S_{\cdot,\cdot}} \tag{1.8}$$

where · denotes the averaging over the relevant dimension of the matrix. If not explicitly stated otherwise, we will always use the corrected Frobenius norm score 1.8 as contact predictor in the following sections of the manuscript.

1.2.3 Free parameters and Gauge invariance

In the Potts model formulation 1.4, there are a total of $\frac{N(N-1)}{2} q^2 + Nq$ parameters, which appear as Lagrange multipliers enforcing the marginals of the model. However, due to normalization, not all one- and two-site marginals are independent quantities. Indeed, each single-site marginal $P_i(A)$ is normalized, hence resulting in only $q-1$ independent components. Similarly, each two-site marginal has only $(q-1)^2$ independent components. This reduced number of independent components is reflected in the Lagrange multipliers, which therefore have only

$\frac{N(N-1)}{2}(q-1)^2 + N(q-1)$ independent components. The remaining dependent components can thus be arbitrarily fixed without changing the probability distribution. Such a freedom in fixing parts of the parameters is known in physics as a gauge-invariance. This freedom is characterized by the fact that by fixing a gauge, one can shift weights between the contribution of the local biases and the coupling parameters without affecting the probabilities.

Indeed, the following class of transformations 1.9 leave the Hamiltonian invariant up to a constant [21], and do hence not modify the probabilities 1.4

$$\begin{aligned} \tilde{J}_{ij}(A,B) &= J_{ij}(A,B) + K_{ij}(A) \\ \tilde{h}_i(A) &= h_i(A) - \sum_{j(j>i)} K_{ij}(A) \end{aligned} \tag{1.9}$$

where K_{ij} is an arbitrary function of A. The choice of a gauge is crucial in the inference of the Potts model and varies depending on the inference method. Details of the gauges induced by the different inference schemes will be discussed in the following sections.

1.2.4 Sequence reweighting

The empirical marginals 1.1 estimated from the data should in principle represent the true statistical properties of the protein family's MSA. This would in principle be true if the sequences were identically and independently drawn from some underlying distribution. However, due to several reasons, the observed proteins are far from being independent and generally present some strong biases.

Part of the inter-dependencies stem from experimental bias, due to over-representation in the dataset of some experimentally relevant organisms. This is particularly the case for some selected model organisms (e.g. *E.coli, H. sapiens, S. cerevisiae*) which have been extensively studied and sequenced. Thus several strains, variants and mutants of these organism are present in the dataset resulting in an unbalanced phylogenetic distribution of organisms.

Furthermore, protein sequences are not random objects independently drawn from a distribution, but are the result of a sequential evolutionary process. Thus homologs are intrinsically correlated through the sharing of common ancestors. Such phylogenetic biases have been long recognized [13, 20] and still present a challenge in statistical inference on homolog families.

In order to partially correct such biases, a simple reweighting scheme has been introduced in [14, 22, 15]. Each sequence in the dataset is assigned a weight w_b inversely proportional to the number of other sequences in the dataset closer than some user-defined threshold

$$\omega_b = |\{\mathbf{X}^{\mathbf{a}},\ \ a = 1, \ldots, B,\ \ \mathrm{Id}(\mathbf{X}^a, \mathbf{X}^b) > \tau_{Id}\}|^{-1} \tag{1.10}$$

where $\mathrm{Id}(\mathbf{X}^a, \mathbf{X}^b)$ denotes the fraction of identical residues shared by two sequences a and b and τ_{Id} is a user-defined similarity threshold. In practical applications, τ_{Id} is chosen of the order of 70-90 % [15, 23, 24].

We can thus define an effective number of sequences in the dataset, as

$$B_{eff} = \sum_{b=1}^{B} \omega_b \tag{1.11}$$

B_{eff} controls the effective number of pseudo-independent sequences in the dataset, and is thus a more suitable quantity than the raw number of sequences B to assess the quality of the MSA. In the following sections, these sequence weights will be used to build reweighted frequency counts, as well as directly incorporated into the Pseudo-likelihood method.

1.3 Parameter Inference

1.3.1 General overview

The task of inferring the parameters of the Potts model 1.4 from data is already computationally intractable for protein families of moderate size. In fact, the computation of the partition function Z involves a summation over the sequence space which grows exponentially with the length N of the sequences. To solve this computational problem, several approximation schemes have been developed, which can roughly be split in two categories: Those who seek an exact (at least numerically exact) solution to a tractable approximation of the Potts model and those who seek an approximate solution to the exact problem. Tree of the most used approximations are presented hereafter. The first two, the Mean-Field solution and the Pseudo-Likelihood maximization fall in the former category, while the Boltzmann Machine Learning approach lies in the latter.

1.3.2 Mean-Field solution

Several different approaches exist to obtain the mean-field solution of 1.4. We will here use an approach based on the cluster variation method [25], which slightly varies from the original derivation in [15]. We start with the canonical definition of the variational Helmotz free energy

$$F[P] = U[P] - S[P] \tag{1.12}$$

where U, S denote respectively the variational internal energy and entropy and P the trial distribution for which F will extremized.

In the variational formulation of the mean-field solution, we factorize the trial distribution P

in single site trial marginals $P_i(A)$

$$P(\mathbf{X}) \approx P_{MF}(\mathbf{X}) = \prod_i P_i(X_i) \tag{1.13}$$

Note that in the cluster variation framework, this factorization corresponds to truncating the entropy expansion at clusters formed by isolated spins [25].

Combining equations 1.6, 1.12, 1.13, and adding normalization constraints on the trial distributions $\{P_i(A)\}$ yields

$$\begin{aligned} F(\{P_i(A)\}) = &-\sum_i \sum_A h_i(A) P_i(A) - \sum_{i<j} \sum_{A,B} J_{ij}(A,B) P_i(A) P_j(B) \\ &+ \sum_i \sum_A P_i(A) \log P_i(A) + \sum_i \lambda_i \left(\sum_A P_i(A) - 1 \right) \end{aligned} \tag{1.14}$$

where we readily recognize the third term as the factorized Shannon entropy for N independent variates. The complete variational free energy can now be extremized with respect to the trial distributions $P_i(A)$, yielding the well-known self-consistent mean-field equations

$$P_i(A) = \frac{1}{Z_i} \exp\left(\sum_{i \neq j} \sum_{B=1}^{q} J_{ij}(A,B) P_j(B) + h_i(A) \right) \tag{1.15}$$

with $Z \equiv e^{\lambda_i - 1}$.

Due to the gauge invariance discussed in section 1.2.3, there is a freedom in the choice of some parameters. In practice, this consists in expressing $P_i(C)$ as a function of the $q-1$ other $P_i(A \neq C)$ for any arbitrary chosen symbol C, and setting the corresponding parameters $h_i(C), J_{ij}(A,C), J_{ij}(C,A)$ to zero. For consistency with [15], we here set all the couplings involving the gap symbol ($C = q$) to zero, i.e. $h_i(q) = J_{ij}(A,q) = J_{ij}(q,A) = 0 \quad \forall i,j,A$. With this gauge choice, the local biases can be extracted from 1.15, yielding

$$h_i(A) = \log\left(\frac{P_i(A)}{P_i(q)} \right) - \sum_j \sum_{B=1}^{q-1} J_{ij}(A,B) P_j(B) \tag{1.16}$$

We can now apply linear response theory [26] to compute the inverse correlations in the Mean-field approximation using

$$(C^{-1})_{ij}(A,B) = \frac{\partial h_i(A)}{\partial P_j(B)} = \begin{cases} -J_{ij}(A,B), & i \neq j \\ \frac{\delta_{A,B}}{P_i(A)} + \frac{1}{P_i(q)}, & i = j \end{cases} \tag{1.17}$$

with $C_{ij}(A,B) = f_{ij}(A,B) - f_i(A) P_j(B)$ the connected correlations.

Equations 1.17 and 1.16 yield the Mean-field solution to the inference of the local biases and couplings of the Potts model.

In the solution 1.17, the coupling parameters are obtained by simple matrix inversion of the empirical correlation matrix C_{ij}. In case of insufficient sampling, C_{ij} can become rank deficient, resulting in the problem becoming ill-defined. In order to compensate this, the authors in [15] made use of pseudo-counts to regularize the empirical correlation matrix. Conceptually, this accounts in adding a fixed number of completely random sequences to the empirical MSA. In practice, this is obtained by introducing a pseudo-count parameter λ in the definitions of the single- and two-site frequencies as

$$f_i(A) = \frac{1}{B_{eff}+\lambda}\left(\frac{\lambda}{q} + \sum_{b=1}^{B} \omega_b \delta_{X_i^b,A}\right) \qquad f_{ij}(A,B) = \frac{1}{B_{eff}+\lambda}\left(\frac{\lambda}{q^2} + \sum_{b=1}^{B} \omega_b \delta_{X_i^b,A}\delta_{X_j^b,B}\right) \tag{1.18}$$

where we introduced the sequence weights ω_b discussed above (see 1.2.4). As seen in 1.18, the pseudo-count parameter effectively controls the amount of random uniform sequences added to the dataset to compute the first two moments of the distributions. In practice, it has been observed that optimal results are obtained for a high pseudo-count amount, i.e $\lambda \sim B_{eff}$ [15]. In [27], the authors have shown that the use of large pseudo-counts is not only useful to regularize the rank-deficiency of empirical correlation matrices, but is essential to compensate for systematic biases of the Mean-field solution, thus elucidating the need for large pseudo-counts for the inference of Potts models in the Mean-field approximation.

1.3.3 Pseudo-Likelihood maximization

The Pseudo-Likelihood (PL) approximation was introduced in the context of DCA in [18] and allows to approximate the Potts model 1.4 by a tractable model. We start by noticing that the complete joint distribution can be factorized in terms of its conditional probabilities as

$$P(\mathbf{X}) = P(X_1, X_2, ..., X_N) = \prod_i^N P(X_i | X_1, ..., X_{i-1}) \tag{1.19}$$

The Pseudo-Likelihood approximation consists in extending the partial conditioning in 1.19 to all other sites $j \neq i$, i.e.

$$P(\mathbf{X}) \approx P_{PL}(\mathbf{X}) \equiv \prod_i^N P(X_i | X_1, ..., X_{i-1}, X_{i+1}, ..., X_N) = \prod_i^N P(X_i | \mathbf{X}_{\setminus i}) \tag{1.20}$$

where $\mathbf{X}_{\setminus i}$ denotes the set of all sites except i. We can now use this approximation to build an approximation to the negative log-likelihood of the Potts model

$$\begin{aligned} l_{PL}(h,J) &= -\frac{1}{B_{eff}}\sum_{b=1}^{B}\omega_b \log P_{PL}(\mathbf{X}^b|h,J) = -\frac{1}{B_{eff}}\sum_{b=1}^{B}\omega_b \log\prod_{i}^{N} P(X_i|\mathbf{X}_{\setminus i}) \\ &= -\frac{1}{B_{eff}}\sum_{b=1}^{B}\omega_b \sum_{i=1}^{N}\log P(X_i|\mathbf{X}_{\setminus i}) = -\frac{1}{B_{eff}}\sum_{b=1}^{B}\omega_b \sum_{i=1}^{N}\log\frac{P(X_i,\mathbf{X}_{\setminus i})}{P(\mathbf{X}_{\setminus i})} \\ &= -\frac{1}{B_{eff}}\sum_{b=1}^{B}\omega_b \sum_{i=1}^{N}\log\left(\frac{\exp\left(h_i(X_i^b)+\sum_{j\neq i}J_{ij}(X_i^b,X_j^b)\right)}{\sum_{s=1}^{q}\exp\left(h_i(A_s)+\sum_{j\neq i}J_{ij}(A_s,X_j^b)\right)}\right) \\ &= -\frac{1}{B_{eff}}\sum_{b=1}^{B}\omega_b \log\left(\frac{\exp\left(\sum_i h_i(X_i^b)+\sum_{i<j}J_{ij}(X_i^b,X_j^b)\right)}{\prod_i\sum_{s=1}^{q}\exp\left(h_i(A_s)+\sum_{j\neq i}J_{ij}(A_s,X_j^b)\right)}\right) \end{aligned} \tag{1.21}$$

where we used the shorthand notation (h,J) to denote the complete set of all local biases and couplings and used the sequence weights ω_b discussed in 1.2.4. In the PL approximation, the complete likelihood of the data is replaced by the approximated form 1.21, which is now tractable. Indeed, we see in the denominator of 1.21 that the PL approximation essentially factorizes the partition function in a product of N "local partition functions" Z_i. It is of central importance in the PL approach to notice that in the expression 1.21 the "local partition functions" explicitly depend on the data and not only on the parameters. Furthermore, l_{PL} does not only depend on the data through their marginals $f_i(A)$ and $f_{ij}(A,B)$, but explicitly depends on the complete set of sequences. Thus, formally, the PL approximation is not considered a truly Maximum Entropy inference, as it implicitly incorporates higher moments of the empirical data.

The negative-pseudo likelihood 1.21 being computationally tractable and differentiable, it can be easily minimized by standard gradient based methods, yielding the PL solutions for the local biases and coupling parameters. Notice that due to the explicit dependence on the sequences in the expression of the negative log-likelihood, the numerical cost of computing the gradient now scales linearly with the number of sequences used to perform the inference.

As discussed in Sec.1.2.3, the gauge-invariance of the Potts model results in a freedom of choice of some parameters. In the optimization procedure of the negative-log pseudo-likelihood 1.21, this translates to the existence of infinitively many global minima. In order to fix the gauge and set a unique global minimum, a L_2 regularization term is added to 1.21 [23] yielding

$$F(\{h_i\},\{J_{ij}\}) = l_{PL} + \lambda_h\sum_i ||h_i||_2^2 + \lambda_J\sum_{i<j}||J_{ij}||_2^2 \tag{1.22}$$

It can be easily shown that the use of L_2 regularization implicitly induces the following gauge choice

$$\begin{aligned}
&\lambda_J \sum_A J_{ij}(A,B) = \lambda_h h_i(B) \\
&\lambda_J \sum_B J_{ij}(A,B) = \lambda_h h_j(A) \\
&\sum_A h_i(A) = 0
\end{aligned} \tag{1.23}$$

To efficiently use the scoring function 1.8 discussed above, the parameters of the Potts model must be shifted to the zero-sum gauge

$$\sum_A J_{ij}(A,B) = \sum_B J_{ij}(A,B) = \sum_A h_i(A) = 0 \quad \forall i, j, A, B \tag{1.24}$$

which can be easily obtained by the following transformation

$$\begin{aligned}
&\tilde{J}_{ij}(A,B) = J_{ij}(A,B) - J_{ij}(A,\cdot) - J_{ij}(\cdot,B) + J_{ij}(\cdot,\cdot) \\
&\tilde{h}_i(A) = h_i(A) + \sum_{j|j>i} \sum_B J_{ij}(A,B) + \sum_{j|j<i} \sum_B J_{ij}(B,A)
\end{aligned} \tag{1.25}$$

A further approximation of the Pseudo-likelihood has been introduced in [23], which substantially accelerates the numerical minimization of eq. 1.22. By inverting the summation over the sequences and over the nodes, the third line of 1.21 can be rewritten as

$$l_{PL}(h,J) = \sum_{i=1}^{N} f_i(\mathbf{X}) \tag{1.26}$$

The asymmetric Pseudo-likelihood method consists in performing N independent minimizations of the integrands f_i. The advantage of this approximation is that the f_i only dependent on the local biases h_i and on the parameters J_{ij} directly incident to node i. There are thus two main numerical advantages of this formulation. First, one large optimization problem in $\frac{N(N-1)q^2}{2} + Nq$ parameters is decimated in N smaller optimizations problems over $Nq + q$ parameters. Second, the optimizations being performed independently, the procedure can be trivially parallelized over the N problems, yielding an essentially linear parallelization speed-up.

Care must be taken when scoring the parameters inferred by the asymmetric Pseudo-likelihood method. As the parameters J_{ij} and J_{ji} are solutions of two different optimizations problems, they will generally not be consistent. In order to circumvent this problem, the authors in [23] considered the averaged couplings $\tilde{J}_{ij}(A,B) = \frac{J_{ij}(A,B)+J_{ji}(B,A)}{2}$ after shifting them in the zero-sum gauge 1.24.

1.3.4 Boltzmann Machine Learning

In contrast to the previously discussed inference methods, the Boltzmann Machine Learning tackles the direct maximization of the likelihood. Given an MSA $\{\mathbf{X_b}\}_{b=1}^{B}$ of B homologous sequences, let the likelihood function be defined as

$$\mathscr{L}(\{h_i\},\{J_{ij}\}) = \prod_{b=1}^{B} P(\mathbf{X^b}) \tag{1.27}$$

where $P(\{\mathbf{X_b}\})$ denotes the Potts distribution 1.4. We seek the optimal parameters $\{h_i\},\{J_{ij}\}$ that maximize the likelihood 1.27 . Equivalently, we can seek the minimum of the negative-log likelihood

$$l(\{h_i\},\{J_{ij}\}) = -\frac{1}{B_{eff}} \sum_{b=1}^{B} \omega_b \log P(\mathbf{X^b}) \tag{1.28}$$

where we introduced the sequence weights ω_b discussed in 1.2.4 and the averaging $\frac{1}{B_{eff}}$ has been added for convenience. As discussed above, due to the gauge-invariance of the Potts model, their is no unique optimal solution to this problem. A L_2 regularization term is thus added to 1.28, yielding a convex function to be optimized

$$l_R(\{h_i\},\{J_{ij}\}) = -\frac{1}{B_{eff}} \sum_{b=1}^{B} \omega_b \log P(\mathbf{X^b}) + \lambda_h \sum_i ||h_i||_2^2 + \lambda_J \sum_{i<j} ||J_{ij}||_2^2 \tag{1.29}$$

with $||h_i||_2^2 = \sqrt{\sum_A h_i(A)^2}$ the usual L_2 norm.

Ignoring the reweighting for simplicity, note that 1.29 can be rewritten as

$$\begin{aligned} l_R(\{h_i\},\{J_{ij}\}) &= -\frac{1}{B} \sum_{b=1}^{B} \left(\log P(\mathbf{X^b}) - \tilde{\lambda}_h \sum_i ||h_i||_2^2 - \tilde{\lambda}_J \sum_{i<j} ||J_{ij}||_2^2 \right) \\ &= -\frac{1}{B} \sum_{b=1}^{B} \log \left(P(\mathbf{X^b}) \exp\left\{ -\tilde{\lambda}_h \sum_i ||h_i||_2^2 \right\} \exp\left\{ -\tilde{\lambda}_J \sum_{i<j} ||J_{ij}||_2^2 \right\} \right) \\ &= -\frac{1}{B} \sum_{b=1}^{B} \log \Big(P(\mathbf{X^b}) P(h_i) P(J_{ij}) \Big) \end{aligned} \tag{1.30}$$

This last form highlights the Bayesian interpretation of the regularization term. Indeed, from 1.30, one sees that the regularization terms are interpreted as prior distributions $P(h_i), P(J_{ij})$ on the parameters. In the case of L_2 regularization, these priors correspond to gaussian distributions over $\{h_i\}$ and $\{J_{ij}\}$, where the hyper-parameters λ_h and λ_J control the (inverse) variance of the priors. In Bayesian terms, the maximization of the regularized likelihood 1.29 corresponds thus to the maximum a posteriori estimator.

Rewriting the negative log-likelihood 1.29 by inserting 1.4 yields

$$l_R(\{h_i\},\{J_{ij}\}) = -\frac{1}{B_{eff}}\sum_{b=1}^{B}\omega_b\left(\sum_{i=1}^{N}h_i(X_i)+\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}J_{ij}(X_i,X_j)\right)+\log Z+\lambda_h\sum_i||h_i||_2^2+\lambda_J\sum_{i<j}||J_{ij}||_2^2 \tag{1.31}$$

Noting that $\frac{\partial \log Z}{\partial h_i(A)} = f_i(A)^{Model}$ and $\frac{\partial \log Z}{\partial J_{ij}(A,B)} = f_{ij}(A,B)^{Model}$, maximizing 1.31 we obtain

$$\begin{aligned}\frac{\partial l_R(\{h_i\},\{J_{ij}\})}{\partial h_i(A)} &= 0 \Rightarrow f_i(A)^{Model} - f_i(A)^{Data} + 2\lambda_h h_i(A) = 0\\ \frac{\partial l_R(\{h_i\},\{J_{ij}\})}{\partial J_{ij}(A,B)} &= 0 \Rightarrow f_{ij}(A,B)^{Model} - f_{ij}(A,B)^{Data} + 2\lambda_J J_{ij}(A,B) = 0\end{aligned} \tag{1.32}$$

where $f_i(A)^{Data}$ and $f_{ij}(A,B)^{Data}$ are the reweighted empirical frequencies measured from the data, while $f_i(A)^{Model}$ and $f_{ij}(A,B)^{Model}$ are the model marginals computed with the inferred parameters. In the absence of the regularization term, the optimal parameters are thus such that the reproduced marginals coincide with the empirical marginals.

The Boltzmann Machine Learning inference strategy is based on the numerical evaluation of $f_i(A)^{Model}$ and $f_{ij}(A,B)^{Model}$ by Monte-Carlo sampling [28] and iteratively updating $h_i(A)$ and $J_{ij}(A,B)$ until the relations 1.32 are satisfied. Thus, the negative log-likelihood of the model, eq. 1.31 is numerically minimized by means of gradient descent techniques, evaluating its gradient (eq. 1.32) at each iteration by MC sampling.

In the field of DCA, Boltzmann Machine Learning has been used by several authors. In [24], the authors used a combination of standard and accelerated gradient descent for the optimization of the parameters, coupled with distributed Metropolis-Hastings Monte-Carlo simulations to estimate the model marginals. In [29], BoltzmannBoltzmann Machine Learning is used as a refinement step after inferring the model parameters using an Adaptative Cluster Expansion (ACE) approximation (see next section) .

1.3.5 Other approaches

Several other methods use global statistical models to infer structural contacts in protein families, all of which share many conceptual similarities with the three DCA flavors described above.

In [29], an Adaptative Cluster Expansion (ACE) of the Potts model is built and the parameters inferred by iteratively computing the contributions to the cross entropy of clusters of increasing size. The authors have shown that ACE inference outperforms other methods (excepted Boltzmann Machine Learning) when inferring generative models, and performs on par with the Pseudo-Likelihood method when benchmarked agains contact prediction. This is generally achieved with improved computational efficiency compared to a full Boltzmann learning approach. However, the computational cost of including clusters of increasing size might limit the applicability of ACE to the analysis of moderately sized protein families.

Several approaches have implicitly built global models using partial correlations [30, 31]. Interestingly, these approaches lead to coupling estimates which have forms very similar to the Mean-field solution discussed above.

Bayesian networks were successfully used in [32, 33] to account for mediated correlations in residue coevolution networks.

2 Molecular chaperones

2.1 Overview of Molecular chaperones

Protein quality control in the cell, also known as proteostasis, consists of multiple regulated pathways whose goal is to maintain the cellular protein population in a functional state. Under several stress conditions (heat, chemical or osmotic shocks among others) native proteins tend to unfold and/or misfiled, potentially leading to cytotoxic aggregates [34]. Furthermore, the continuous turnover of protein synthesis, degradation and translocation under normal conditions leads to a sizable proportion of proteins transiently being in non-native state in a living cell [35].

In order to control the population of non-native proteins, organisms have since long acquired a large class of proteins known as molecular chaperones. The broad common function of all chaperones consists in regulating the population of non-native proteins and form the core machinery of the cell to drive them towards their native states [35], under both stress and normal conditions. Historically, molecular chaperones were initially observed in the early seventies in cells responding to heat shock and were thereafter referred to as heat-shock proteins (Hsps) [36]. However, it has since been recognized that the chaperone function extends beyond the heat-shock response, and that a sizable fraction of molecular chaperones are not heat-inducible, thus carrying the name of Hsps for historical reasons.

Molecular chaperones form a complex and dense protein network consisting of multiple families, which can be broadly categorized in 5 machineries based on the core proteins molecular weight: Small heat-shock proteins (sHsp), Hsp60, Hsp70, Hsp90 and Hsp100. These different chaperone machines are involved at several steps of proteostasis, from the early protection against aggregation of misfolded proteins, to protein folding and assisting proteolysis at final stages of proteins life cycle. In the complex environment of living cells, most chaperone families collaborate, forming a complex protein quality control system [37]. Indeed, a typical proteostasis pathway involves the sequential action of different chaperone families, where the product of the action of upstream chaperone systems are transferred to downstream chaperones for further processing [38, 39]. Given the importance of protein quality control for cell viability, it is not surprising that most chaperones are present throughout the whole tree-of-life and make up a large fraction of the dry mass of living cells [40]. This ubiquity of molecular chaperones results in a generally large number of paralogs (i.e. related proteins belonging to the same family in an organism).

2.2 The Hsp70 chaperone system

The Hsp70 chaperone family is one of the main components in the proteostasis network. It is, together with Hsp90, the largest chaperone family, both in terms of number of paralogs and in terms of abundance in cells [40]. Hsp70s are present in virtually all known organisms, with numbers of paralogs ranging from a single Hsp70 in some primitive bacteria, to 13 in *H.sapiens*.

Hsp70s functionally act both under stress and normal conditions. They have been shown to be involved in disaggregating potentially cytotoxic protein aggregates [41], assisted folding and refolding of non-native substrates [42], co-translational folding at the ribosome [43], sub-cellular transport of newly synthesized polypeptides through membrane pores [44], disassembly of macro-molecular complexes [45] and assisting the assembly of iron-sulfur clusters [46]. This large functional spectrum is based on the generic ability of Hsp70s to bind to client substrates in non-native conformations. This is achieved through the recognition by Hsp70 of typical motifs formed by patches of solvent exposed hydrophobic residues, flanked by regions enriched in basic residues [47]. These generic characteristic motifs are indicative of substrates being in non-native conformations. In fact, in a natively-folded protein, hydrophobic residues tend to be buried in the proteins core, whereas the surface residues are generally enriched in charged and polar residues [48]. This results in the recognition of substrates by Hsp70 to be generally unspecific, targeting a broad range of client proteins.

Upon binding, Hsp70s lead to the expansion of the client substrates [49]. Although, the exact mechanism involved is still a matter of debate, a generally accepted basic principle of the induced substrate expansion is based on the entropic pulling mechanism [50, 44, 51]. This mechanism is best understood in the illustrative case of the action of Hsp70 in protein translocation through membrane pores [44]. mtHsp70s, particular members of this chaperone family, are mitochondrial chaperones located at the exit of the membrane pore in the mitochondrial matrix. Upon emergence of nascent imported polypeptides at the pore exit, mtHsp70s bind to their specific binding site motifs and are detached from the mitochondrial membrane. This leads to an effective increase of the size of the imported polypeptide, now in complex with attached mtHsp70s. Excluded volume interactions, between the bulky chaperone near the pore and the membrane, lead to a decrease in conformational entropy of the polypeptide. This last effect strongly depends on the distance between the chaperone and the membrane, so that a free energy gradient of entropic origin appears. Thus, this gives rise to an effective pulling force generated by the binding of the chaperone to the unfolded protein being translocated. In analogy, the binding of multiple Hsp70s to a client substrate in non-native conformation can lead to a decrease of entropy, by chaperone-substrate or chaperone-chaperone excluded volume interactions. The resulting free-energy gradient again effectively acts as an entropic force, leading to an expansion of the misfolded substrate. Thus, the entropic pulling mechanism forms an appealing theoretical framework to explain substrate expansion and macromolecular disassembly induced upon Hsp70 binding [52].

Structurally, Hsp70s are approximatively 70 kDa proteins, formed by two domains (Fig.2.1). The nucleotide binding domain (NBD, blue in Fig.2.1) forms a globular domain where ATP is bound and hydrolyzed, whereas the substrate binding domain (SBD, purple in Fig.2.1), composed of the α and β sub-domains, interacts with client substrates in a clamp-like fashion. The two domains are joined by a highly conserved flexible linker [53]. Upon ATP binding and hydrolysis, the chaperone undergoes a large-scale allosteric conformational change. In a simplified view, the ATP bound state ("open state") is characterized by the two SBD sub-domains docked on the NBD and the inter-domain linker forming a β-strand with a sheet on the NBD [54, 55]. In the ADP bound state ("closed state"), the two domains are detached and only joined by the undocked flexible linker [56]. This simplified view must however be considered with care. Indeed, recent experimental results showed that Hsp70s are present in both the open and closed state when bound either to ATP or ADP, with nucleotide-dependent relative populations of the two states [57].

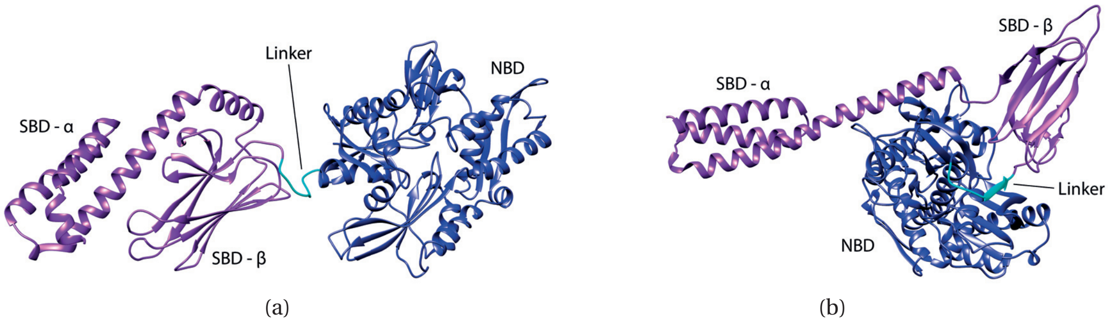

Figure 2.1 – Experimental structures of the *E.coli* Hsp70 DnaK. Blue: NBD, Cyan: Linker, Purple: SBD. **A** ADP bound structure (PDB ID: 2KHO [56]). **B** ATP bound structure (PDB ID: 4JNE [55]).

The biochemical cycle of Hsp70s is composed of the continuous switching between the open (ATP-like) and closed (ADP-like) conformations, resulting in alternating binding and unbinding of client substrates (Fig.2.2). The chaperone substrate binding/unbinding rates depend on the bound nucleotide, with the ATP-bound conformation displaying 100-1000 faster substrate exchange compared to the ADP-bound conformation [58]. The resulting Hs70-substrate affinity is thus a subtle interplay between the kinetic rates and the populations of the chaperone in the different states. It has recently been recognized that the effective affinity for client substrate is strongly enhanced by the non-equilibrium nature of the ATP-driven cycle, a phenomena dubbed ultra-affinity [58]. These observations, combined with novel experimental evidence in their support, raised central questions on the energetic use of molecular chaperones targeted at modifying the folding free energy landscape of proteins [58, 59, 60].

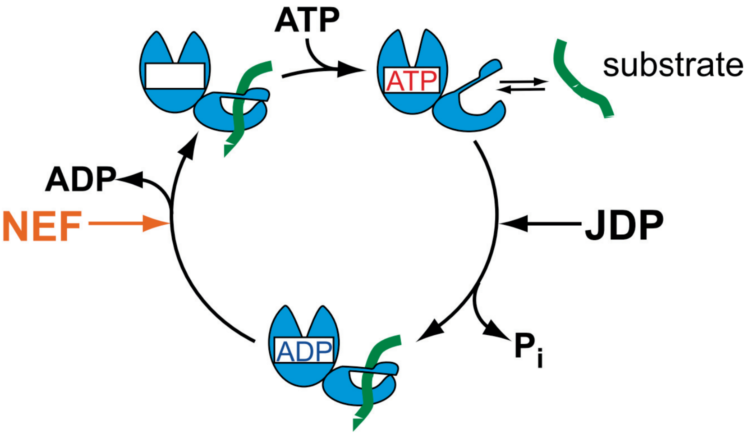

Figure 2.2 – Simplified biochemical cycle of Hsp70. NEF: Nucleotide Exchange Factor, JDP: J-domain Proteins (Hsp40s). Figure adapted from [53].

In the simplified cycle represented in Fig.2.2, the ATP hydrolysis and nucleotide exchange are regulated by the interactions of Hsp70s with co-chaperones and client substrates, which altogether form the Hsp70 chaperone machinery. Particularly, two classes of co-chaperones are of central interest: J-proteins and nucleotide exchange factors.

2.2.1 J-proteins

Hsp70 have a rather low basal ATP hydrolysis rate (0.02min^{-1} [61]). J-domain proteins (JDP), also known as Hsp40s, are a class of co-chaperones which have been shown to stimulate the ATP hydrolysis rate of Hsp70s [62, 63]. Hsp40s form a very broad class of proteins, generally present in large numbers of paralogs in higher eukaryotes [64]. Their common structural feature is the presence of a highly conserved J-domain, which is thought to be the main interacting domain with Hsp70s [65, 66, 67, 68]. Beyond the presence of the J-domain, which *de facto* defines the family, Hsp40s display large variability in the architecture of their other domains. This heterogeneity is a hallmark of the functional differentiation of this family. Indeed, our current knowledge is based on Hsp40s being the specificity determinants of Hsp70s function, by targeting Hsp70s towards the cellular location where they are required. This targeting can be of geographic origin (e.g. the localization of mtHsp70s at the exit of membrane pores, through their interaction with membrane anchored Hsp40s) or of more functional nature, as exemplified in the recruitment of Hsp70s by substrate-bound Hsp40s [64].

A canonical classification, based on the architecture of the C-terminal domains of Hsp40s has been proposed by Kampinga et al. [69, 64]. Class A and B Hsp40s are characterized by a conserved C-terminal domain architecture, which is composed of two repeats of C-terminal substrate binding domains (CTD I & II), connected to the N-terminal J-domain through a glycine-phenanaline rich region (G/F region). The main difference between class A and B Hsp40s is the additional presence of a zinc-finger domain between the J-domain and the G/F region in class A J-proteins, the exact function of which is still unclear [70]. These canonical Hsp40s are known to form constituent homo-dimers through a short dimerization domain located at the extreme C-terminus. Functionally, class A and B Hsp40s act as substrate recruiters for Hsp70s. They bind Hsp70 substrates through their multiple CTDs, which they then deliver to the main chaperone, by direct interaction through the J-domain [71]. The transient formation of Hsp70-Hsp40-substrate complexes, and the induced effect on the cycle of Hsp70 are still only marginally understood, both at the structural and functional level (see chapter 4).

Unlike the canonical class A and B Hsp40s, class C J-proteins are not characterized by any particular architectural features of their C-domain [64]. Beyond the presence of the J-domain, no evolutionary conserved domains cover all class C Hsp40s. Functionally, members of this class possess several different roles, among which the recruitment of Hsp70s at membrane pore exits (Tim44 in the mitochondrion, Sec63 in the ER-lumen [72]) or the interaction with specific Hsp70 substrates in specialized pathways, such as in the bacterial HscA/HscB system, which are particular Hsp70/40 members specialized in the Iron-Sulfur assembly pathway [73] (see chapter 6).

2.2.2 Nucleotide Exchange Factors

Upon ATP hydrolysis to ADP by Hsp70, the spontaneous release of ADP towards an apo state is a slow process [74, 75, 76]. Nucleotide Exchange Factors (NEF) are a class of functionally related proteins which stimulate the release of ADP from the NBD, thus allowing new ATP to be loaded and the cycle to complete. In contrast to Hsp40, which are phylogenetically well defined by the presence of the J-domain, proteins belonging to multiple unrelated families are known to act as NEFs for different Hsp70s, thus defining NEFs only through a functional relationship [76]. In bacterial organisms, one main family of NEFs is known, GrpE, which has homolog members in eukaryotic mitochondria and chloroplasts. In contrasts, in eukaryotic cytosol and ER, three unrelated NEF families are present, Hsp110 (Grp170), HspBP1 (Sil1) and BAG-domains. These different families are further composed of multiple paralogs which act as NEFs for several different Hsp70s [76]. Interestingly, Hsp110 are a sub-family which share a phylogenetic origin with Hsp70 [77], and experimental evidence suggests that Hsp110s might have an intrinsic chaperoning activity beyond their NEF function [78]. In addition to NEFs, eukaryotes also possess proteins with antagonist nucleotide exchange effects. In particular, proteins belonging to the Hip family have been shown to inhibit the Hsp70 release of ADP, by a competitive binding with Hsp70 on a similar interface as NEFs [79].

3 Coevolutionary analysis of the Hsp70 family

The main results presented in this chapter have been published in [80]. The majority of the figures presented in this chapter are directly reproduced from the published article, in accordance with the Creative Commons Attribution License used by PLOS.

3.1 Introduction

We start our coevolutionary study of the Hsp70 system by an analysis of the isolated Hsp70 protein. The availability of high-resolution experimental structures for both the ADP- and ATP-bound conformations allows a detailed comparison of DCA predictions with experimental ground-truth and will serve as a baseline benchmark for subsequent coevolutionary analysis. An interesting feature of Hsp70s is that they present a large set of mutually exclusive contacts in the two main conformations. As seen in Fig.3.1, the α and β sub-domains of the SBD form two separated sets of contacts with the NBD in the ATP-bound conformation, while they directly interact in the ADP-bound conformation, forming a closed clamp-like domain. We can also observe some smaller differences in the intra-NBD contacts between the two conformations, reflecting the subtle conformational rearrangements of the NBD upon ATP hydrolysis to ADP.

While a large amount of experimental data on the Hsp70 system is available, comparatively little has been analyzed from a sequence perspective. In this chapter, we will investigate whether statistical sequence analysis can bring new insights into this molecular machine. Several questions will be addressed: Are residues involved in the large scale allosteric transformation of this chaperone detected by DCA and is the coevolutionary signal concentrated on a small subset of the interface? Are there coevolutionary signals of higher order macro-molecular assemblies in the Hsp70 family? Can we modify the canonical DCA procedure to highlight the phylogenetic origin of particular sets of predicted contacts? These questions will serve as a guideline throughout this chapter, which is aimed at presenting a complete coevolutionary analysis, from the dataset preparation to the biological interpretation of the results.

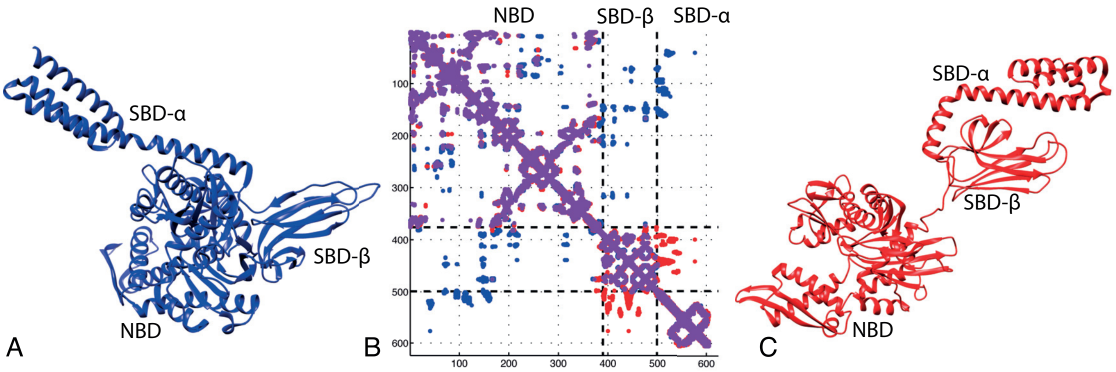

Figure 3.1 – Multiple conformations of *E.coli* DnaK and their contact maps. The three subdomains are denoted as: NBD: Nucleotide Binding Domain, SBD-α: Substrate Binding Domain (α helical sub-domain), SBD-β: Substrate Binding Domain (β-sheet sub-domain) **A** ATP-bound structure [55]. **B** Contact maps of the two conformations and their overlap. Contacts are defined as pairs of residues having at least one pair of heavy atoms separated by less than 8.5 Å. Blue: Contacts only in ATP structure, Red: Contacts only in ADP structure, Purple: Contacts present in both ADP and ATP structures. Dashed lines depict the limits of the three domains. **C** ATP-bound structure [56]. Figure adapted from [80].

3.2 Dataset preparation

Given the wide phylogenetic distribution and large number of paralogs in this family, the overall number of available sequences should be large, thus making Hsp70 an excellent candidate for DCA. We will hereafter describe in detail the sequence extraction protocol used to build the Hsp70 family MSA.

An initial small set of Hsp70 sequences (seed) was built, starting with the PFAM seed of the Hsp70 family [81]. The seed was manually enriched and curated to cover a broad taxonomic distribution of organisms. The seed was then aligned using the MAFFT software package [82] using default parameters. To search the Uniprot database containing publicly available protein sequences [9], we used the HMMER software package, which is based on constructing a Hidden Markov Model (HMM) of the protein family [83]. We thus used the *hmmbuild* utility of this package to build an HMM of the Hsp70 family based on the aligned seed, resulting in a model defining 624 residue positions of the Hsp70 family. The *hmmsearch* utility was then used to search the complete Uniprot database (union of the Uniprot TrEMBL and Swissprot databases, release 2014_04), using the default inclusion threshold of the HMMER package, resulting in a raw MSA of the Hsp70 family. This MSA was then filtered, discarding all sequences containing more than 25% of gaps. Finally, we pruned the resulting MSA, keeping only sequences having at most 90% sequence identity using the hhblits utility. This identity filtering is a similar alternative to the reweighting scheme presented in Sec. 1.2.4 and aims to reduce the phylogenetic bias in the dataset. This protocol resulted in an MSA of width $N = 624$ of the Hsp70 family containing $B = 3708$ sequences with a uniformly covered phylogenetic distribution (1562 Eukaryotes, 1982 Bacteria).

3.3 Direct-Coupling Analysis of the Hsp70 family

We performed DCA using the symmetric version of the pseudo-likelihood approximation as described above on the Hsp70 family MSA and ranked the predicted contacts according to their APC-corrected Frobenius norm scores (see Sec.1.2.2 and Sec.1.3.3). In all the subsequent analysis, DCA predictions were only performed on residue pairs with distances along the sequence of at least 5. In fact, local coevolving pairs separated by less than five residues along the chain pertain mostly to trivial local structural elements and are not particularly informative of the global tertiary structure [15]. In order to compare the DCA predictions with the ground truth of the experimental structures, we defined residue pairs as being in contact if at least one pair of heavy-atoms was separated by less than 8.5 Å in the experimental structures. This binary definition of structural contacts allows to quantitatively assess the quality of the DCA prediction, by measuring the precision of the prediction (also called True Positive Ratio (TP)), defined as the fraction of correctly predicted contacts.

The DCA results benchmarked against both the ATP- and ADP-bound structures showed overall very high quality predictions (see Fig.3.2), and were uniformly distributed over the contact map. Among the $N = 624$ top ranked DCA predictions, 502 (resp. 504) corresponded to structural contacts in the ATP-bound (resp. ADP-bound) experimental structures, corresponding to precisions of 80% (resp. 81%). If the predictions were benchmarked against the union of the two structural contact maps (defined as the joint set of contacts in either of the two conformations), the precision significantly improved to 86%. This indicated that approximately 5% of the DCA predictions were exclusively present in either one of both conformations, thus involving residues participating in the allosteric transition. We hereafter refer to such pairs as allosteric contacts. The inspection of the predicted contact maps (Fig.3.3A,C) highlighted that a substantial fraction of DCA predictions not coinciding with structural contacts lied in very close proximity of native contacts. Such minor discrepancies could easily be explained by thermal fluctuations around the native state, minor structural variations between homolog structures or experimental artifacts induced by the crystallization procedure [84]. This motivated us to associate to each predicted contact a score representing the length of the shortest path (SP) between the two residues, computed with respect to the experimental structures. In practice, the binary contact map associated to a structure defines an unweighted undirected graph, for which we computed the shortest path between any two pairs of residues. This score allowed to give a topological measure of the "wrongness" of a predicted contact, in the sense that it measured the minimal number of native contacts over which a coevolutionary signal should be mediated to explain the observed prediction. By construction, native contact have SP=1, while predictions directly adjacent to native contacts have SP=2 and so on. Applied to the Hsp70 case, we observed that the DCA predictions (Fig.3.3E) were mostly concentrated at low SP values. We argue that the minor discrepancies discussed above could easily be compatible with most DCA predictions at SP 2-3. Indeed, if considering contacts at SP=2 as correct,

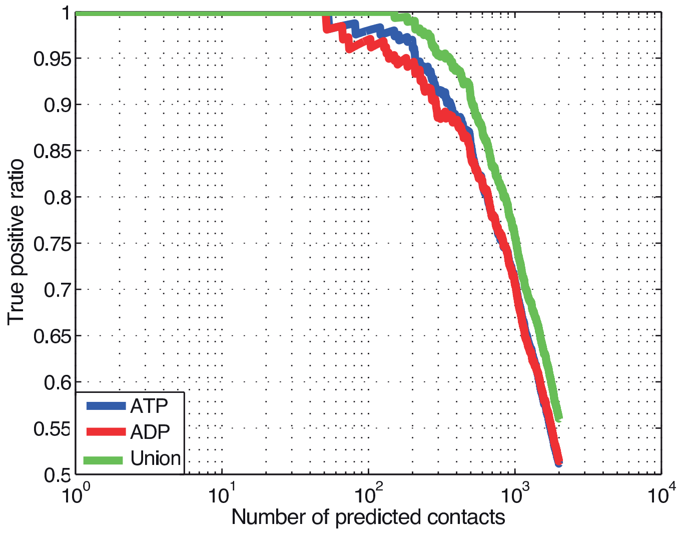

Figure 3.2 – Precision of DCA predictions on the Hsp70 family, for both the ATP- and ADP-bound structures. Union denotes the joint set formed by contacts present in either the ATP- or ADP-bound conformation. Figure adapted from [80].

the precision of the N top ranked predictions on the union of ATP and ADP contact maps increased to a striking 95%. The SP analysis clearly highlighted the prediction of allosteric contacts (Fig.3.3E). Indeed, when the SP distributions were computed over the union of the contact maps of both conformations, a small but significant shift of the SP towards lower values occurred. This effect was visually confirmed in the contact maps (Fig.3.3A,C), where we observed a set of coevolving contacts in the NBD - SBD interface having SP>4 in the ADP conformation but SP=1-2 in the ATP state. Conversely, we identified a set of predicted contacts with SP>6 in the SBD-α - SBD-β interface in the ATP state, while these contacts satisfied SP=1 in the ADP conformation. These results illustrate that DCA not only predicts overall common structural features of the Hsp70 chaperones, but also extracts information about multiple conformations, separated by large-scale structural transitions, and allows the quantification of the important contacts differentially involved in multiple conformations.

3.4 Hsp70 Homo-Dimerization

The high quality results obtained on the Hsp70 monomer in both ATP- and ADP-bound conformation sparked the curiosity to investigate the origin of the apparently incorrectly predicted contacts. In the ATP/ADP union map (Fig.3.3B), the major source of false-positive predictions (SP>9) were located in the NBD-SBD interaction region. Inspection of the crystal structure of the ATP-bound conformation (PDB ID: 4JNE [55]) revealed that the unit cell contained a symmetrical arrangement of two Hsp70 monomers. The comparison of the DCA predictions with a contact map formed by the union of the ATP and ADP structures, and incorporating these homo-dimeric contacts, revealed a quasi-perfect overlap between the

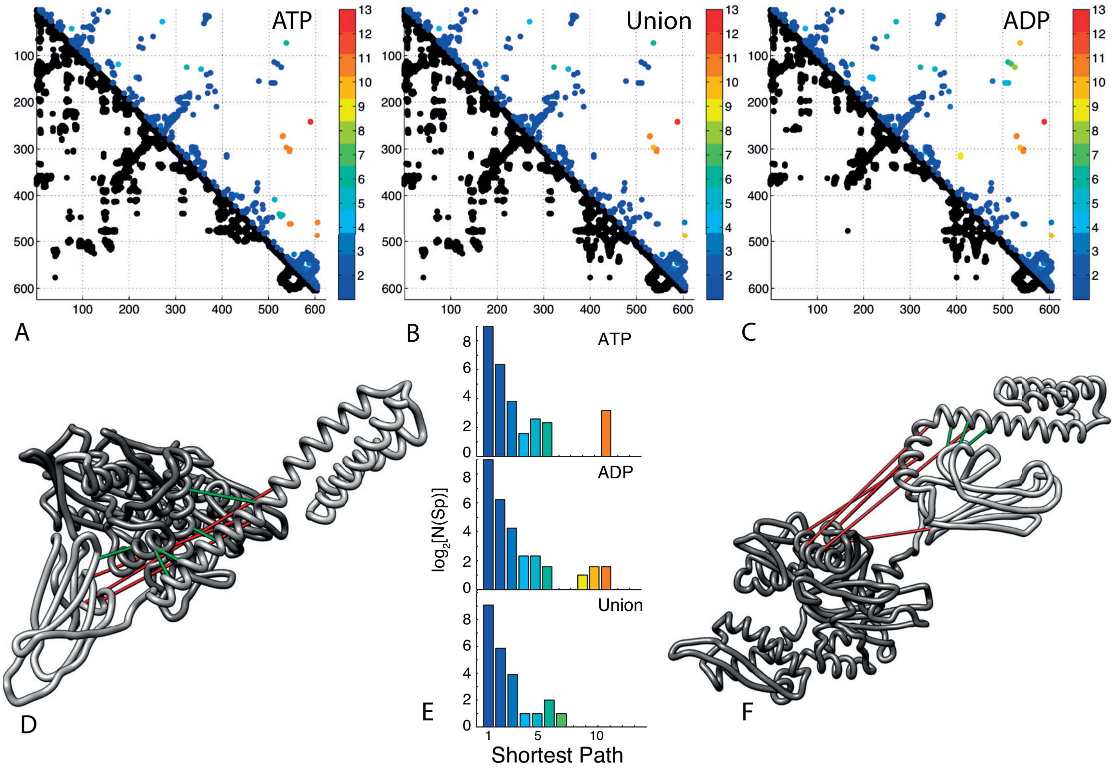

Figure 3.3 – DCA predictions of multiple Hsp70 conformations. **A-C** Lower triangular parts: Structural contact maps, defined using a contact threshold of 8.5 Å. Upper triangular parts: top *N* DCA predicted contacts (excluding contacts involving residue pairs separated by less than five residues along the chain). The DCA predictions are colored according to their Shortest Path length. **A)** ATP conformation, **B)** Union of ATP+ADP conformation, **C)** ADP conformation. **D,F** 8 strongest allosteric predicted contacts in ATP (**D**) and ADP (**F**) conformations. Correct (resp. false) contacts are depicted with green (resp. red) lines. **E** Histograms of shortest path lengths for the top N predicted contacts in the 3 cases. Figure adapted from [80].

set of strongly outlying predictions and the dimeric interface formed in the crystal structure (Fig.3.4A). This was strikingly visible in the shift towards lower values of the SP distribution (Fig.3.4B, compare to Fig.3.3E). This analysis revealed the presence of six strongly coevolving residue pairs in the Hsp70 homo-dimeric interface. Four of these involved the docking of the SBD of one monomer onto the NBD of the other (Fig.3.4C), while the remaining two predictions were located at the dimeric NBD-NBD interface (Fig.3.4D).

Given the low number of dimeric predicted contacts, a precise statistical significance test of these findings was necessary to assert their biological relevance. In this context, a valuable statistical test is given by Fisher's exact test. This consists in evaluating the P-value under a random null model, i.e. the probability of observing six such contacts in the crystal dimer interface, given the number of predictions and the number of crystal dimeric contacts. We thus built a null model consisting in randomly distributing the *N* DCA predictions among all possible $M = \frac{N(N-1)}{2}$ pairs and evaluating the probability that among these *N* predictions, *k* would fall in the dimeric interface consisting of the K contacts observed in the crystal structure. The sought probability was given by a hyper-geometric distribution

$$p(N,k,M,K) = \frac{\binom{K}{M}\binom{M-K}{N-k}}{\binom{M}{N}} \tag{3.1}$$

for which the P-value was then defined as the probability of observing an effect of greater or equal amplitude under the null hypothesis. In this case, we evaluated the one-tailed P-value, given by

$$P = \sum_{k'=k}^{N} p(N,k',M,K) \tag{3.2}$$

Evaluating 3.2 with $N = 624$ total predictions, $k = 6$ dimeric predictions, $K = 241$ dimeric contacts in the crystal and $M = \frac{N(N-1)}{2}$ the total number of possible contacts resulted in $P = 1.44 \cdot 10^{-4}$. This low P-value underlined the statistical significance of the six predicted dimeric contacts. Hence the homo-dimeric arrangement has an evolutionarily conserved interface, hinting at a functional role played by the oligomeric form of Hsp70.

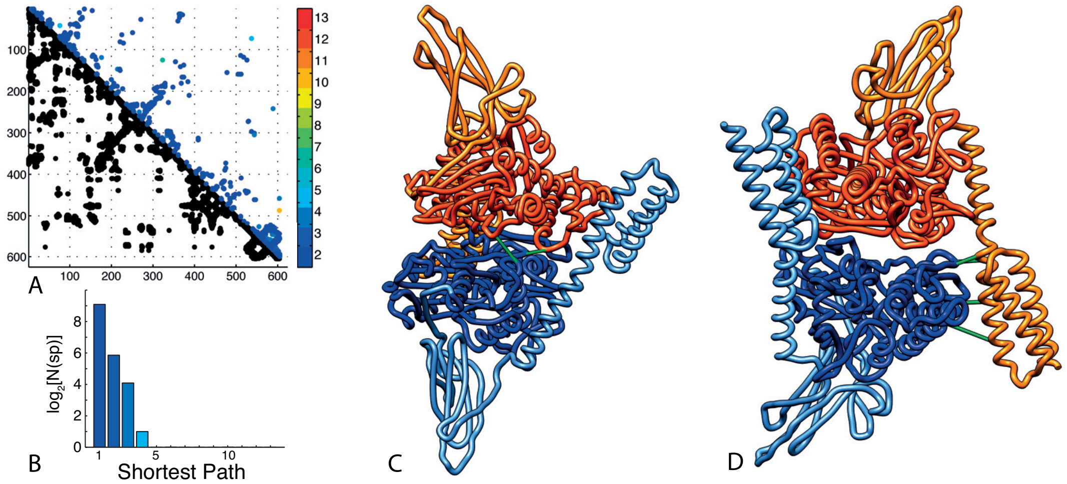

Figure 3.4 – Hsp70 homo-dimeric DCA predictions. **A** DCA predictions compared to the union of ATP and ADP contact maps, including homo-dimeric contacts observed in the crystal structure [55]. Lower triangular part: Structural contact maps, defined using a contact threshold of 8.5 Å. Upper triangular part top N DCA predicted contacts (excluding contacts involving residue pairs separated by less than five residues along the chain). The DCA predictions are colored according to their SP length. **B** Histogram of SP lengths for the top N predicted contacts in the Union+Dimer contact map. **C-D** Structural views of the homo-dimeric arrangement as observed in the crystal structure (PDB ID:4JNE, [55]) and the six DCA predicted oligomeric contacts. The domains of the two Hsp70 monomers are shaded differently (Dark tones:NBD, light tones: SBD). **C** Highlights the NBD-NBD contacts, while **D** shows the NBD-SBD contacts. Figure adapted from [80].

As discussed in Chapter 2, Hsp110 are remote homologs of Hsp70 which can act as nucleotide exchange factors in the biochemical cycle of Hsp70. It is known that Hsp110 form heterodimers with Hsp70, inducing an opening of the NBD resulting in the facilitated release of ADP [85]. Given the similarities between the homo-dimeric arrangement observed in the crystal structures of ATP-bound *E. coli* DnaK (PDB ID: 4JNE) and the Hsp70-Hsp110 heterodimer structure (PDB ID: 3C7N), it could be argued that the presence of Hsp110 sequences in the MSA could introduce spurious traces of the oligomeric state in DCA predictions. To verify whether this was the case, we repeated the DCA analysis on a reduced set, containing only sequences explicitly bearing canonical Hsp70 gene names (*hspa1a, hspa1b, hsp70, ssa1 and DnaK*, extracted from the Uniprot database), resulting in 1781 sequences. Note that this does not imply that only 1781 of the original 3708 sequences were canonical Hsp70s. Indeed, we here relied on the gene name annotations of Uniprot, which are approximate at best, particularly in the case of automatically annotated sequences. Although this resulted in an overall decreased prediction quality due to a decreased number of sequences available to infer the model (Fig.3.5), the six previously predicted dimeric contacts were still predicted among the top *N* ranked contacts. Furthermore, an additional dimeric prediction appeared in the top *N* ranked DCA contacts, strongly supporting that the predicted homo-dimerization pattern is coevolutionarily conserved and thus functional in canonical Hsp70 chaperones and not a consequence of the presence of Hsp110 sequences in the dataset.

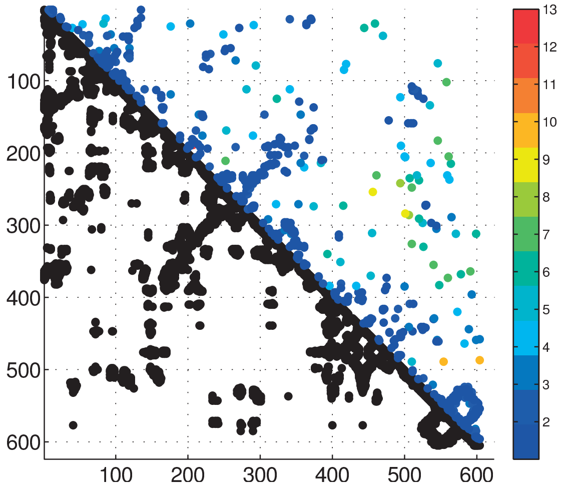

Figure 3.5 – DCA predictions restricted to canonically tagged Hsp70. Lower triangular parts: Structural contact maps, defined using a contact threshold of 8.5 Å. Upper triangular part top *N* DCA predicted contacts (excluding contacts involving residue pairs separated by less than five residues along the chain). The DCA predictions are colored according to their SP length. Figure adapted from [80].

To further characterize the phylogenetic distribution of this oligomeric form, we investigated the taxonomic origin of the coevolutionary signal in the dimeric predictions. To this aim, we started by investigating which phylogenetic groups were responsible for the most variation in the Hsp70 MSA. This can efficiently be highlighted by Principal Component Analysis (PCA)

[86], which consists in projecting sequences onto the low-dimensional subspace conserving the highest fraction of variance. It can easily be shown that this maximum-variance subspace is spanned by the eigenvectors of the positional covariance matrix of the sequences associated to the largest eigenvalues, named principal components. Projecting the Hsp70 sequences onto the first two principal components and tagging the sequences as either bacterial, eukaryotic or archaeal highlighted a number of interesting features (Fig.3.6A). First, the main source of variation in the dataset (variation along the first principal axis, PCA 1) was generated by the divergence of bacteria and eukaryotes, as seen in the large separation between the main eukaryotic and bacterial clusters in Fig.3.6A. Close inspection of the eukaryotic sequences lying in the bacterial region (blue dots in the left part of Fig.3.6A) revealed that these eukaryotic Hsp70s were mainly found in mitochondria and chloroplasts. The close proximity of organellar sequences to bacterial sequences is in agreement with the probable bacterial origin of these organelles [87, 88]. Second, the two dense clusters of sequences at the top of Fig.3.6A were strongly enriched in Hsp110 for the eukaryotic cluster (top-left, green sequences) and in HscA (top-right, blue sequences). HscA are specialized bacterial Hsp70s, involved in the iron-sulfur cluster pathway (see Chapter 6). These have the particularity of having putatively a single client substrate (the scaffolding protein IscU) and only interact with a single Hsp40 cochaperone (HscB). Interestingly, eukaryotic homologs of HscA, which are present in mitochondria, did not appear to form a well defined cluster in the PCA subspace, as was the case for HscA (at least not at the level of the first couple of principal components). This phylogenetic splitting between HscA and their eukaryotic homologs, the latter seemingly closer to other bacterial Hsp70s, currently remains an unexplained curiosity. Thus, PCA of the Hsp70 family not only highlights phylogenetic divergences such as the bacterial-eukaryotic split, but also indicates which functional divergences of sub-families carry most variation in sequence space, naturally complementing the contact-level analysis allowed by DCA.

Having identified the bacteria-eukaryotic split as the main source of variation in the Hsp70 MSA, we investigated whether the coevolutionarily conserved homo-dimeric arrangement was differentially conserved in these two kingdoms. To this aim, we introduced artificial relative sequence weights in the Pseudo-likelihood DCA method, which allowed to control the relative importance of sequences belonging to bacteria or eukaryotes. We repeated the DCA analysis varying the normalized weight ω_E in a range of $[0.3-0.7]$ which ensured a high precision of the top $N = 624$ predictions (Fig.3.6B bottom). Note that the extreme cases $\omega_E = 0$ and $\omega_E = 1$ correspond to performing DCA on subsets of sequences containing only Bacterial, resp. Eukaryotic sequences. Going to these extreme cases however strongly increased the error rate as the number of sequences decreased approximately by half. For each value of ω_E we recorded the normalized DCA score of the dimeric predictions, defined as

$$\frac{S_{Dim}}{S_{Tot}} = \frac{\langle S_{ij}^{Dimer} \rangle_{Dimer}}{\langle S_{ij}^{All} \rangle_{All}} \tag{3.3}$$

where S_{ij} denotes the APC corrected Frobenius norm score of contact i, j. Note that $\langle S_{ij}^{Dimer} \rangle_{Dimer}$

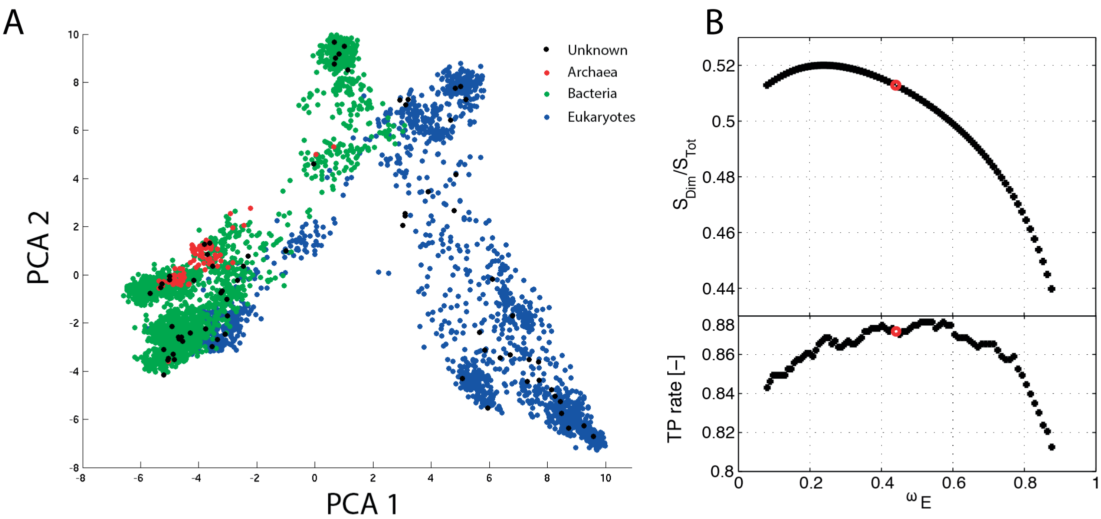

Figure 3.6 – **A** PCA of the Hsp70 family. Each dot represents a sequence projected onto the first two principal components and is colored according to their phylogenetic clade. **B** Top: Normalized dimer score versus relative reweighting. Low values of ω_E correspond to giving more weight to bacterial sequences. Bottom: Precision computed over the top N DCA predictions versus relative reweighting. The red dots correspond to the original case where no relative weights are introduced. Figure adapted from [80].

is averaged only over the scores appearing on the 6 originally predicted dimeric contacts, whereas $\langle S_{ij}^{All}\rangle_{All}$ is averaged over the top N predicted DCA contacts. The trend of the normalized dimer score (Fig.3.6B, top) indicated that the phylogenetic origin of the dimeric signal stemmed from bacterial sequences. This was further confirmed by repeating the DCA analysis on separate bacterial and eukaryotic sub-sets (corresponding to $\omega_E = 0$ and $\omega_E = 1$). While the noise level was to high to reliably exploit contact information, all six dimeric contacts were predicted in the bacterial sub-set, whereas none was predicted using the eukaryotic subset. This phylogenetic analysis of the HSP70 homo-dimerization interaction thus revealed that its origin lies in bacterial sequences. While the biological function of this arrangement can not readily be deduced from this analysis, the structural and phylogenetic results obtained here allow to propose several biological hypotheses which will need further experimental and theoretical investigations (see Discussions below).

3.5 Experimental evidence of Hsp70 homo-dimerization

Several experimental studies observed the presence of Hsp70 dimers and higher order oligomers. As the majority of these studies were performed in vitro, no clear functional role could be assigned to the dimeric form of Hsp70s. Of particular interest are three recent studies. In [89], the authors showed that a small population of DnaK dimers are present at physiological concentrations. They mutated two residues on the NBD-SBD interface to cysteines (A303C, H541C) in order to monitor the formation of homo-dimeric arrangements and found a clear dimerization signal in the presence of ATP. Interestingly, their double mutations fall in close vicinity to one of the six coevolutionarily conserved interface contacts predicted by DCA (E306-H541). They further tested both in vitro and in vivo the functional relevance of the dimer formation by mutating five residues at the dimer interface (one at a time, G28A, R56A, T301A, N537A, D540A) and observed a growth defect at 37 °, while neither the ATPase activity, nor the substrate binding activity was strongly affected by these mutations. It is again interesting to note that their five mutations lie extremely close to residues involved in the six DCA predicted dimer-contacts (Q277-Q534, E306-H541, D129-K363, E310-K548, A30-A276, Q178-V533). Their most intriguing finding was significantly reduced interactions of the dimer-defective mutants with Hsp40 co-chaperones (Fig.3.7A). In [90], authors of the same group showed that the chaperoning function in vitro is compromised for cysteine cross-linked DnaK dimers (Fig.3.7B). They furthermore noted that while the non dimer forming DnaK mutants have reduced interactions with Hsp40 co-chaperones, forcing DnaK to continuously be in a dimeric form through cysteine cross-linking completely abolished the chaperoning activity of DnaK. Interestingly, the authors in [91] showed that the human heat inducible Hsp70 (hspA1A) also formed dimers in vitro and showed by analysis of truncated Hsp70 that the C-terminal part is involved in the dimerization. This last observation hints at a more complex role played by the homo-dimeric form, which might still be present in a subset of eukaryotic Hsp70s.

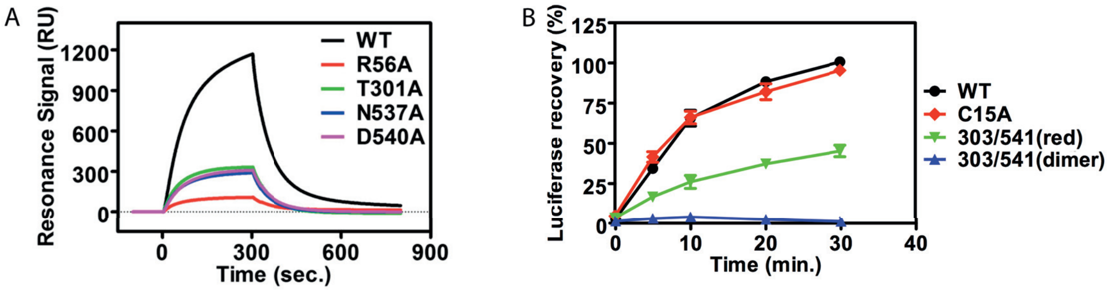

Figure 3.7 – **A** Point alanine mutations in the dimer interface strongly decrease the interactions with Hsp40 co-chaperones. The vertical axis indicates the Surface Plasmon Resonance signal with DnaJ immobilized to the sensor chip. Figure adapted from [90]. **B** Chaperone activity of the dimer-defect DnaK mutant (303/541(dimer), blue) is strongly reduced compared to wild type DnaK (WT, black). The vertical axis indicates the fraction of recovered luciferase after denaturation by heat-shock. Figure adapted from [89].

3.6 Discussion

The coevolutionary analysis of the Hsp70 family revealed several interesting features. First, mutually exclusive contacts resulting from a large scale transformation are fully encoded in the sequence co-variations, and can be easily detected by DCA. Although at the time our analysis, traces of multiple conformations in DCA analysis had be observed [15], the analysis of Hsp70 displayed the ability of DCA to detect tertiary rearrangements on a very-large scale. A natural question emerges in this context: Given that contacts pertaining to multiple conformations are encoded in the DCA predicted contact maps, is it possible to predict the presence of multiple conformational states from sequence data alone? This broad question is of high experimental interest, as the determination of dynamically populated states remains experimentally complicated. In this work, we relied on the knowledge of two known Hsp70 conformers to investigate the allosteric contacts and dimeric contacts. In general, if one conformation is known, contacts formed in an alternative form can be detected by analyzing the differences between the experimental and predicted contacts maps [92]. In the absence of such a-priori knowledge, the classification of contacts belonging to different conformations remains however an unresolved task, which calls for new methodological developments. Whether the detection and prediction of multiple conformers based solely on a predicted contact map is possible, without the explicit construction of a three-dimensional structural model, is a topic of current interest.

The coevolutionarily conserved dimeric complex of Hsp70 predicted by DCA is a strong indication of the physiological relevance of such an arrangement. While this fact had been partially observed experimentally, sequence coevolution gives an orthogonal insight into the functional necessity of the oligomeric form. DCA is based on extracting the inter-residue contacts that bear the strongest evolutionary pressure. During evolution, there has thus been a strong pressure to preserve the formation of the Hsp70 homo-dimer, clearly indicating that it bears a functional importance for the survival of organisms. While powerful for predicting structural contacts and highlighting their biological importance, coevolutionary analysis does not give direct insight into the functional role of the predicted interactions. It is however a valuable tool that allows to draw functional hypothesis. In the case of the Hsp70 homo-dimerization, we hypothesize two possibly complementary functions: In the first place, the homo-dimerization of Hsp70 increases the local concentration of chaperones. Given that in general multiple Hsp70s bind to a substrate protein [49], this co-localization could increase the efficiency with which Hsp70s can quickly target substrates in non-native states. Secondly, the similarities in the binding interface with the Hsp70-Hsp110 heterodimer leads to a second hypothesis. As Hsp110 are remote eukaryotic Hsp70 homologs, which appeared later in evolution, and act as nucleotide exchange factors in eukaryotic Hsp70 cycles, we hypothesize that the emergence of Hsp110 as NEFs could be a specialization of pre-existing Hsp70 which already formed homo-dimers. Thus, the Hsp70-Hsp110 interactions would be based on the template of Hsp70-Hsp70 interactions, already present in bacteria. Given that bacteria have a general pressure for maintaining a smaller genome size compared to eukaryotes [93, 94], eukaryotes could have had the opportunity to acquire specialized members of Hsp70 acting (among other functions [78]) as more efficient NEFs, while bacteria might have relied on homo-dimerization for a rudimentary NEF function. This scenario is fully compatible with the observation of some eukaryotic Hsp70s still bearing signs of homo-dimerization [91]. Note that the presence of other NEFs found in bacteria, notably GrpE [95], does not preclude this scenario, as either multiple different NEF actors could be involved in the Hsp70 cycle, or later addition of GrpE like proteins could have appeared in some bacterial clades.

4 Hsp70-Hsp40 interactions

The main results presented in this chapter have been published in [68]. The majority of the figures presented in this chapter are directly reproduced from the published article, in accordance with the Creative Commons Attribution License used by eLife.

4.1 Introduction

Hsp70 forms the core of the chaperones machinery, however its function in vivo is strongly dependent on interactions with partner co-chaperones. Of particular importance is the interaction of Hsp70s with J-domain containing proteins, also called Hsp40s (see Chapter 2). In fact, Hsp70s have very low basal ATP hydrolysis rates (0.02min^{-1} [61]). The first known role played by the Hsp40-Hsp70 interaction is the stimulation of Hsp70s ATPase activity (10-200 fold increase [61]), thus actively promoting the functional cycling of the Hsp70 chaperone. Furthermore, canonical Hsp40s, such as DnaJA2 and DnaJB1 in *H. sapiens*, drive the substrate specificity of Hsp70s, by binding candidate substrates and delivering them to Hsp70s, effectively targeting Hsp70 to specific substrates [64]. Interestingly, the concomitant binding to substrate proteins and to Hsp40s synergistically increases the ATPase activity of Hsp70s, largely exceeding the ATPase stimulation predicted by the product of the stimuli of isolated J-domains or substrates [96]. Additionally, Hsp40s can promote specific localization of Hsp70s towards particular cellular locations. This is observed in specific members of the Hsp40 family, which bind to particular cellular locations where Hsp70 are required. Examples include the yeast scHlj1, an Hsp40 bound to the cytosolic side of the endoplasmic reticulum which recruits cytosolic Hsp70s to assist in protein degradation at the ER exit [97] or human DnaJC2 which recruits the cytosolic Hsc70 (a constitutive human Hsp70 paralog) to the ribosome, forming a machinery that assists the co-translational folding of nascent protein chains [98, 99].

All these versatile functions are supported by the interaction of Hsp70 with the J-domain of Hsp40s. This approximatively 70 residue long domain, which *de facto* defines the Hsp40 family, is the common structural element present in all Hsp40s. Whereas there is a large variation in the architecture of other domains of Hsp40s, the J-domain presents high sequence and structural conservation among Hsp40 homologs [64]. Structurally, the J-domain is composed of a packing of four alpha helices (denoted I - IV) in a well defined tertiary structure (Fig.4.1 and Chapter 2). The central importance of the Hsp40-Hsp70 interaction in the chaperones biochemical cycle has long been recognized, and a substantial amount of experimental studies have appeared over the last decades. However, despite the considerable experimental efforts targeted at characterizing the nature of this transient complex, several crucial questions at multiple levels of details remain open. At the molecular level, is there a unified structural model of the Hsp40-Hsp70 interactions, compatible with all major experimental evidence? Can we characterize the dynamical nature of this transient interaction? At an organizational level, what are the specificity determinants and the interaction networks for organisms with a high number of Hsp40/Hsp70 paralogs? At a functional level, how can a structural model of the interaction shed light on the role played by the J-domain?

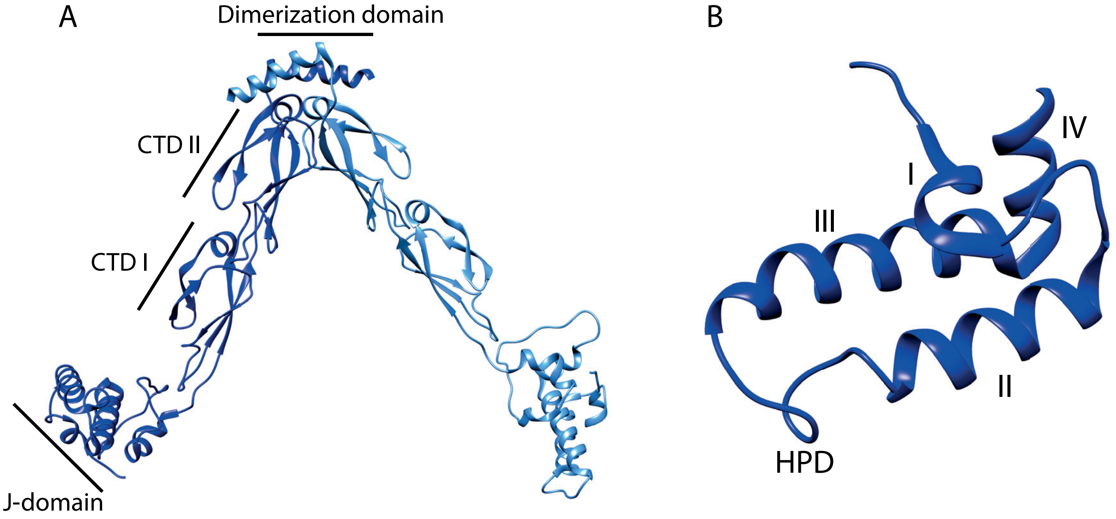

Figure 4.1 – Structural organization of canonical J-proteins. **A** Domain architecture of a canonical (class B) J-protein of *T. thermophilus* (PDB ID: 4J80). The two protomers forming the symmetric dimer are depicted in different shades of blue. **B** Close-up view of the J-domain of panel A. The four helices and the conserved HPD tripeptide locations are highlighted.

In this chapter, we will address several of these questions through the combined use of coevolutionary analysis, coarse grained modeling and atomistic simulations, after briefly reviewing the available experimental data and models of the Hsp40-Hsp70 interaction. The coarse-grained simulations discussed in Sec. 4.5 were performed in collaboration with G.Hummer and A. Jost-Lopez in Frankfurt.

4.2 Experimental models of the Hsp40-Hsp70 interaction

Given the central role played by the interaction with Hsp40 in the Hsp70 cycle, many experimental projects have focused on characterizing the complex. Multiple studies have highlighted the irrefutable necessary presence of the J-domain for a functional interaction with Hsp70s

[100, 101, 65, 102]. Greene and coworkers first determined by NMR spectroscopy that helix II and the HPD motif of the J-domain were directly involved in the interaction with Hsp70, focusing on the bacterial DnaK-DnaJ system (resp. Hsp70 and Hsp40 homologs in *E.coli*) [101]. More specifically, the authors of [65] have identified an intriguing double mutation in the interaction between DnaK and DnaJ: While the R167H mutation on the DnaK NBD resulted in growth defects at 42°, they observed that the simultaneous mutation D35N in the DnaJ J-domain restored growth under heat-shock to normal levels. Furthermore, surface plasmon resonance (SPR) analysis confirmed the lack of strong interactions between wtDnaK and DnaJD35N and between wtDnaJ and DnaKR167H, while SPR clearly showed strong interactions between DnaJD35N and DnaKR167H. Interestingly, the D35H mutation on the J-domain lies in the highly conserved 33HPD35 tripeptide present in the vast majority of J-domains. This observation thus highlighted the central role played by the Hsp70 NBD and the HPD motif of the J-domain in the interaction. Later NMR studies of the DnaK-DnaJ system in the presence of ADP however detected no direct involvement of the HPD motif in the interaction [67]. This observation was supported by the use of paramagnetic resonance relaxation enhancement measurements, showing no NMR signal originating from the HPD motif of DnaJ. The authors further noted the highly dynamic nature of the interaction.

In contrast to the large amount of biochemical data, only little structural models have been obtained for the complex. This is most likely due to the dynamic nature of the interface, the transient low affinity interaction ($K_D \approx 16\mu M$ [67]) and the continuous switching between multiple conformations of the Hsp70 chaperone. To date a single high-resolution structural model of the complex has been deposited [66]. Based on the R167H-D35N mutation described above, the authors built a disulphide cross-linked complex of bovine Hsc70 and Auxillin (a member of the Hsp40 family), by mutating these two residues to cysteines forming a cross-link. In their crystallographic structure, the HPD motif and helix III of the Auxilin J-domain appear to be the main interacting segments. The authors also noted that the ATPase rate of the cross-linked complex was substantially higher than for the wild type interaction, provided the construct was composed of the NBD and the inter-domain linker. A linker truncated version of the complex displayed wild type basal ATPase rates and no significant stimulation upon addition of the J-domain. The differences observed between the crystallographic structure of the bovine Hsc70/Auxilin complex and NMR data on the bacterial DnaK/DnaJ system might have several origins: The nature of the bound nucleotide might modify the binding interface of the complex, the crystallization and/or disulphide cross-linking procedure might have introduced artifacts or trapped the complex in a sparsely populated conformation, or there might be a genuinely different binding interface in bacterial and eukaryotic systems [103, 104].

4.3 Extending DCA for protein-protein interactions

Our first approach to characterize the Hsp40-Hsp70 interaction was based on the analysis of coevolving residue pairs across the interaction interface. The extension of DCA to inter-protein contacts prediction is *in principle* straightforward. For two interacting protein families A and

B, if knowledge of which pairs of proteins A and B interact specifically is available, a paired MSA can be constructed by simply concatenating the pairs of interacting sequences. This concatenated MSA can then be analyzed by DCA and strongly conserved inter-protein contacts identified in the relevant quadrant of the predicted contact map [14, 105, 106] (Fig.4.2).

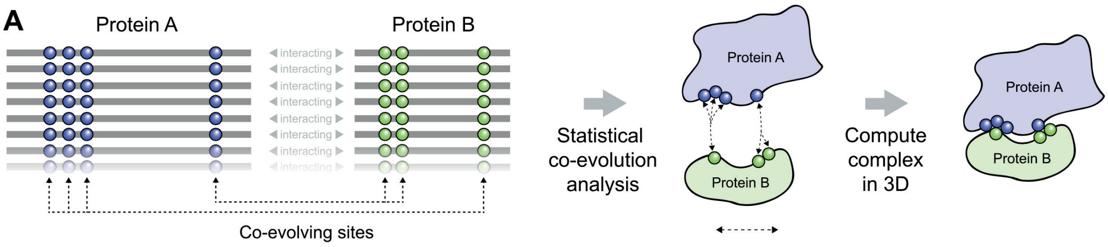

Figure 4.2 – Principle of DCA applied to protein-protein interactions. Pairs of interacting proteins are stitched and DCA is performed on the concatenated MSA. Coevolving inter-protein contacts can be identified in the relevant quadrant of the predicted contact map and further analyzed. Figure adapted from [105].

The a-priori knowledge of interacting pairs of protein sequences however puts a very strong constraint on the application of DCA to predict protein-protein interactions. This is particularly the case when numerous paralogs are present in organisms for both families A and B, and potentially cross-talk can occur [22]. Several approaches have been proposed to tackle this question. In the simplest case where no paralogs are present (i.e. all organisms have a single copy of proteins A and B), the matching is trivial. This allowed the study of coevolving inter-protein contacts for example in the large and small subunits of the ribosome [106, 107]. For the matching of bacterial sequences, one can exploit the operon structure of interacting genes. In operons, sets of interacting genes are adjacently co-localized and can be co-expressed through the transcription of the complete operon. The genomic proximity of two genes encoding for proteins in families A and B can thus be an excellent indicator of the putative interactions of the gene products in bacteria, thus allowing the discrimination of interacting paralogs. This gene-proximity strategy has been successfully used to pair sequences in bacterial organisms [106, 107, 105].

More recently, two new approaches have emerged that self-consistently employ DCA to simultaneously find the near-optimal paralog matching and extract coevolving residue pairs between interacting protein families [108, 109]. Both methods are based upon iteratively building a concatenated MSA, such as to implicitly maximize the coevolutionary signal measured on the protein-protein interface. Both authors start from an initial concatenated MSA and infer the Potts hamiltonian through mean-field DCA (gaussian DCA for [109]). The concatenated MSA of the next iteration is built by selecting the pairs of sequences to be matched based on the inter-protein part of the interaction energy as computed by the current Hamiltonian. The two methods differ in the selection criterion controlling which sequences will be added to the MSA of the next iteration. The Progressive Paralog Matching (PPM) method introduced in [109] uses a linear programming approach to find the organism-specific optimal paralog matching such as to maximize the log-likelihood of the matching. In [108], the Iterative Paralog

Algorithm (IPA) uses a gap criterion to score all possible intra-organism paralog matchings, based on the hypothesis that an optimal match should be paralog-specific, i.e. it should have a low coevolutionary energy and display a large energy gap with the second best match. The two methods further diverge by the number of selected pairs at each iteration and the total number of iterations. Both methods display excellent results on the benchmark cases used in the original publications. Due to the necessity to have a reference gold-standard to benchmark the methods, both cases were benchmarked against known optimal matches, as defined by the operonic structures of the investigated families. This resulted in the testing of the methods against bacterial families, showing high specificity (i.e. one-to-one matching, little paralog cross-talk).

To tackle the matching problem for the Hsp40 and Hsp70 families, we concomitantly developed an alternative strategy. Our approach, named Random Matching Strategy, does not aim at finding an optimal matching of paralogs, but only focuses on extracting contact information for interacting protein families. Thus, in contrast to IPA and PPM, the Random Matching Strategy does not offer any insights into the paralog-matching, cross-talk or specificity. Details about the random matching strategy will be given in the next section.

4.3.1 The Random Matching Strategy

The problem setup of the random matching approach is as follows. We are given two MSAs of protein families A and B, containing a total of B_A and B_B sequences. Each family is composed of O_A and O_B organisms, among which O_{AB} are shared. Each organism contains a variable number of paralogs, i.e. $\{n_o^A\}_{o=1}^{O_A}$ and $\{n_o^B\}_{o=1}^{O_B}$ denote the sets of number of paralogs in family A/B for all organisms.

The goal of the Random Matching Strategy (RM) is to extract pairs of inter-protein coevolving residues from the two protein families described above. The main idea of RM is to randomly match paralogs in all O_{AB} shared organisms, creating a set of stochastically concatenated MSAs, perform DCA on each MSA and record the frequency at which each potential inter-protein contact is predicted. The rationale behind this approach is that if inter-protein contacts are predicted in a high fraction of cases, these are the most robust contacts subject to the random matching noise, and should thus be indicative of strong underlying coevolutionary pressure. For any organism, we enforce that each paralog be matched only once, such that there are $\min(n_o^A, n_o^B)$ matched sequences per organism for a total depth of the randomly matched MSA of $B_{RM} = \sum_{o=1}^{O_{AB}} \min(n_o^A, n_o^B)$. This selective matching per organism has two main advantages: Allowing multiple matches per paralog for each MSA could quickly grow the size of the resulting MSAs up to a potential limit of $B = \sum_{o=1}^{O_{AB}} n_o^A n_o^B$, for which it would become computationally expensive to perform multiple DCAs. Furthermore, the quadratic growth of the number of sequences would inevitably dilute the coevolutionary signal, as we expect that even in very promiscuous interactions, not all possible paralog pairs are biologically interacting. Note however that restricting a unique paralog match for each MSAs does not preclude the inclusion of coevolutionary signal from cross-talks, as the random matching performs independently on all stochastically matched MSAs. A pseudo-code of the matching is shown in Algorithm1.

```
interfaceContacts = []
for $i \in \{1, ..., N_{rep}\}$ do
    MSA = []
    for $o \in \{1, ..., O_{AB}\}$ do
        $n_{Org} \leftarrow \min(n_o^A, n_o^B)$
        $\text{seqs}^A, \text{seqs}^B \leftarrow$ getOrgSequences($o$, $\text{MSA}^A$, $\text{MSA}^B$)
        $\text{p}^A, \text{p}^B \leftarrow$ selectRandomParalogs($n_{Org}$, $\text{seqs}^A$, $\text{seqs}^B$)
        matchedOrg ← concatenate($\text{p}^A$, $\text{p}^B$)
        MSA.$append$(matchedOrg)
    end
    ppiContacts ← plmDCA(MSA)
    interfaceContacts.$append$(ppiContacts)
end
```

Algorithm 1: Pseudo-Code for the Random Matching Strategy.

To extract inter-protein coevolving contacts after DCA has been performed on the MSAs, we employed a criterion introduced in [105], which normalizes the inter-protein DCA scores as

$$\tilde{S}_{ij} = \frac{S_{ij}}{\left|\min\left(S_{ij}^{Inter}\right)\right|\left(1 + \sqrt{\frac{N}{B_{eff}}}\right)} \tag{4.1}$$

where S_{ij} denotes the APC-corrected Frobenius norm (see Chapter 1) and the minimum $\min\left(S_{ij}^{Inter}\right)$ is taken over all inter-protein DCA scores. This normalization approximatively allows comparison of inter-protein DCA scores between different protein families with varying widths and sequencing depths. We used the selection criterion originally proposed by the authors, which consists in retaining inter-protein residue pairs with $\tilde{S}_{ij} > 0.8$, which empirically corresponds to a precision of approximatively 80% [105]. Furthermore, the APC-correction of the scores is adapted following [106], where the two averages appearing in the APC are restricted to the two isolated families. This last modification equalizes variations introduced by inhomogeneous evolutionary rates between different families.

In contrast to IPA and PPM, RM does not rely on any energetic information extracted from the coevolutionary hamiltonian, but only on contact prediction by DCA. RM is computationally efficient, as the construction of the N_{rep} randomly matched MSAs and the DCA contact predictions are all independent, and can thus be trivially parallelized with linear speedup. This also allows us to employ the Pseudo-Likelihood DCA method, which is known to give the best performances on contact predictions.

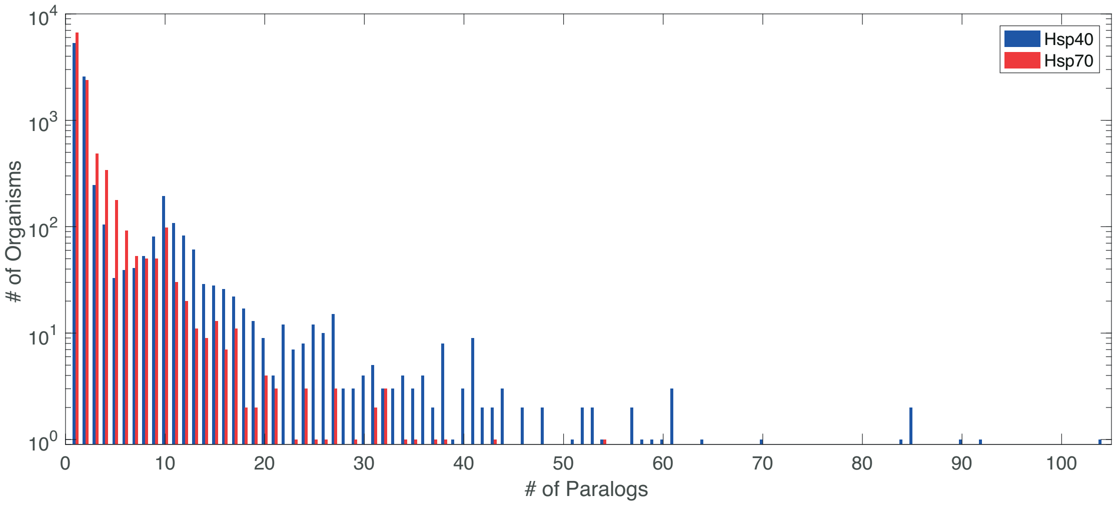

Figure 4.3 – Distribution of Hsp40 and Hsp70 paralogs per organism. Figure adapted from [68].

4.4 Random matching on the Hsp40-Hsp70 system

4.4.1 RM predictions of the complex

To investigate the Hsp40-Hsp70 interaction by means of the RM strategy, we built two separate MSAs of the Hsp40 and Hsp70 families, using the same extraction procedure as presented in Sec. 3.2 (applied to the Uniprot TrEMBL+Swissprot release 2015_08). This resulted in MSAs with large taxonomic coverages (Tab.4.1). The distribution of number of paralogs per organism displayed broad distributions of paralogs, ranging up to ~ 50 (resp. 100) paralogs for the Hsp70 (resp. Hsp40) families (Fig.4.3). Noteworthy is the expansion of number of Hsp40 paralogs, compared with the number of Hsp70 paralogs, which is consistent with their role of specificity determinants for the Hsp70 machinery. The small fraction of organisms having a very large number of Hsp40 paralogs (>80) is essentially composed of plants, which are known for having highly developed proteostasis networks. Surprinsingly, the number of organisms for which Hsp40/70 are found in a given kingdom present large differences (number in parentheses in Tab.4.1). These variations have two origins: On the one hand, organisms having partially sequenced genomes can have members of only one family sequenced and thus available in the database. On the other hand, the finite homology-search sensitivity of current state-of-the art sequence search tools (*hmmer* in our case) do not extract *all* possible homologs from the databases. As the effects of sequence length and domain architecture on the homology search sensitivity can not be fully controlled, it is expected that not all paralogs will be recovered for a given organism. This will inevitably introduce organisms for which only one member of an interacting paralog pair is present, thus contributing to an intrinsic matching noise. Combined, these effects lead to an effective decrease of the number O_{AB} of organisms having sequences sampled for both protein families.

	Eukaryotes	Bacteria	Archaea	Viruses	Other	Total
Hsp40	14369 (1093)	11379 (7837)	311 (273)	36 (22)	159 (13)	26254 (9238)
Hsp70	7881 (1933)	11819 (8272)	273 (258)	25 (17)	63 (13)	20061 (10493)

Table 4.1 – Taxonomic distributions of Hsp40 and Hsp70 sequences. Entries denote the number of sequences found in each taxonomic kingdom. Number in parenthesis denote the number of organisms in which the sequences are distributed.

Using the two MSAs for the Hsp40 and Hsp70 families discussed above, we performed RM by generating 1000 stochastically matched MSAs, and analyzed the fraction of appearance of the inter-protein contacts (Fig.4.4A). On average, there were 15.6 inter-protein contacts predicted per random realization. Interestingly, these were far from being uniformly distributed, but displayed some very sharp peaks. In order to select which of these peaks we considered significant for further analysis, a decision criterion was needed. We here chose to take advantage of the knowledge of the isolated structures of the two interacting proteins and used a conservative selection criterion based on the surface accessibility of residues involved in the predicted inter-protein contacts. We thus selected all the contacts having the highest fractions of appearance until the first predicted contact involving a buried residue (Solvent Accessible Surface Area< $1Å^2$, Tab.4.2). This resulted in the prediction of three coevolving residue pairs in the Hsp40-Hsp70 interface, depicted by colored dots in Fig.4.4A, corresponding in the *E.coli* DnaK/DnaJ numbering to N187-K23, D208-K26 and T189-R19. Mapping these coevolving pairs on the experimental structures of the *E.coli* DnaK/DnaJ members highlighted that the three predicted residues on both proteins clustered in close proximity, forming two well defined interaction patches (Fig.4.4B,C).

We further investigated whether RM predicted contacts below the first appearance of a buried residue were informative of the complex. Fig.4.5 displays the locations of less frequently appearing RM contacts. As seen, contacts appearing less frequently than the first appearance of a buried residue were still consistent with the three highest ranked contacts. Among the nine first contacts analyzed, appearing in more than 20% of the random matchings, two contained buried residues, while five clustered in near proximity (among which the three first) on the NBD and on the J-domain. The two remaining exposed contacts were distant from the predicted binding site. These observations strengthened our choice of a conservative selection criterion, focusing on the three most frequently appearing contacts with high confidence, but highlighted the fact that useful information was present in lower ranked RM contacts.

The RM strategy thus identified three strongly coevolving and robust residue pairs across the Hsp40-Hsp70 interface, for which the surface exposure allowed putative inter-protein interactions. To build a structural model based on these predictions, two approaches could be followed. The predicted contacts could be translated into distance restraints and introduced in molecular simulations to perform a constraint search for a binding pose approximatively respecting the DCA restraints as proposed in [111]. Alternatively, unconstrained docking simulations could be performed, comparing the predicted contacts to the outcome of the

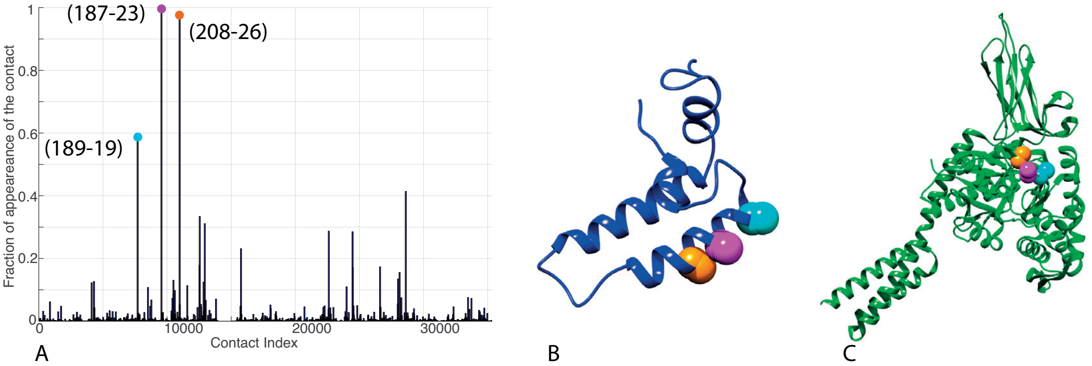

Figure 4.4 – Random Matching results on the Hsp40-Hsp70 interaction. **A** Fraction of appearances of the inter-protein predicted contacts in the RM approach. The abscissa is an arbitrary contact index, mapping each inter-protein residue pair to the natural numbers. The colored dots highlight the selected contacts before the first contact involving a buried residue. The numbering of the highlighted contacts refer to the *E.coli* DnaK/DnaJ proteins. **B/C** Mapping of the residues involved in the selected contacts on the experimental structures of *E.coli* DnaJ (**B**, PDB ID: 1XBL, [110]) and DnaK (**C**, PDB ID: 4JNE, [55]). The colors of the highlighted residues follow the scheme of panel A. The C-α and C-β atoms of the six residues are depicted by spheres. Figure adapted from [68].

Contact	SASA 1 [Å^2]	SASA 2 [Å^2]	Selection Frequency
N187 - K23	102.7 (0.64)	119.8(0.60)	0.996
D208 - K26	41.6 (0.27)	144.9 (0.72)	0.976
T189 - R19	17.3 (0.12)	177.9 (0.79)	0.587
A176 - A64	0.6 (0.005)	34.6 (0.30)	0.414
L392 - A29	89.5 (0.52)	11.1 (0.10)	0.334
L219 - E55	44.7 (0.26)	69.8 (0.37)	0.311
D224 - K51	40.7 (0.27)	95.2 (0.48)	0.287
L320 - M30	0.7 (0.004)	129.2 (0.70)	0.285
F356 - R36	48.5 (0.23)	214.6 (0.95)	0.232

Table 4.2 – Most frequently appearing contacts in the RM procedure. The numbering in the first column pertains to the *E.coli* DnaK/DnaJ pair. SASA denotes the Solvent Accessible Surface Area. Predictions involving buried residues are highlighted. Table adapted from [68].

independent simulations and using them for selecting between multiple candidate model complexes. We pursued the latter approach, which is the subject of section 4.5.1.

4.4.2 Comparison with IPA and PPM

We compared our results obtained using the RM strategy to the two state-of-the art DCA based methods to predict inter-protein contacts, namely IPA and PPM. We performed both IPA and PPM on the Hsp40-Hsp70 dataset with the parameters used in their original publications [108, 109]. For a fair comparison regarding the contact predictions, we performed the same

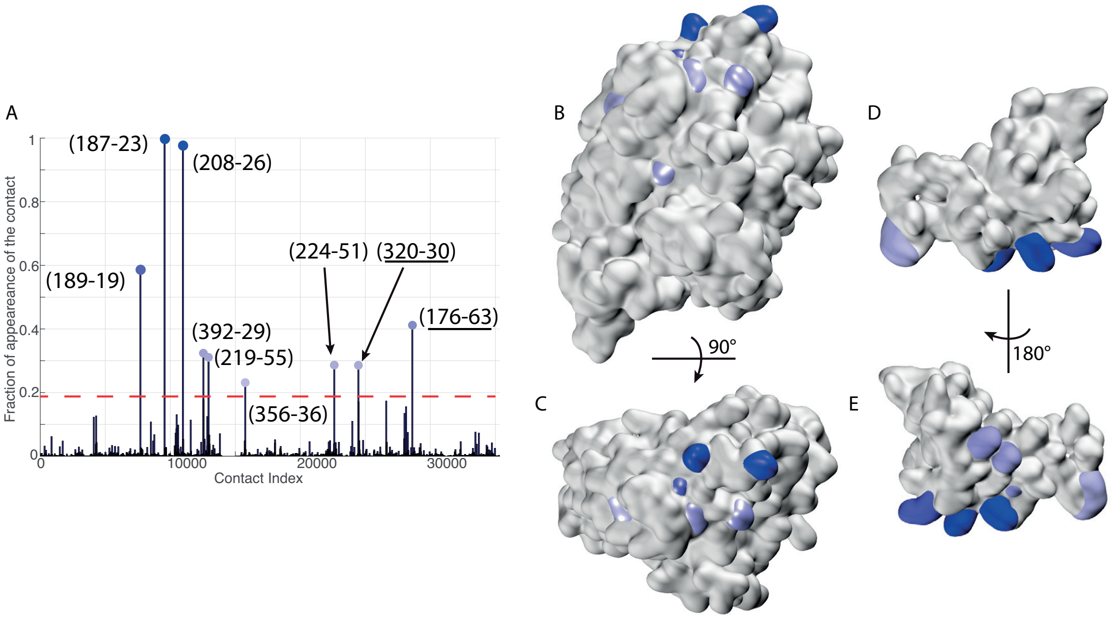

Figure 4.5 – Analysis of lower ranked RM predictions. **A** Fraction of appearances of the predicted inter-protein contacts in the RM approach in 1000 realizations. The abscissa is an arbitrary contact index, mapping each potential inter-protein contact to the natural numbers. The dots are shaded linearly in the appearance fractions. RM contacts appearing in more than 20% of random matchings (red line) are highlighted. **B,C** The nine RM contacts identified in **A** depicted on the *E.coli* DnaK NBD. **D,E** The nine RM contact identified in **A** depicted on the *E.coli* DnaJ J-domain. The colors in **B,C,D,E** follow the color scheme in **A**. Figure adapted from [68].

plm-DCA procedure as the one used in RM on the matched MSAs obtained by IPA and PPM and analyzed the highest ranked contacts. Comparison of the 5 most frequently appearing contacts in the RM procedure with the five top ranked interface contacts predicted using IPA and PPM revealed a large overlap of the three methods (Fig.4.6, Tab.4.4, see Supplementary Figures and Tables). The strong overlap of the three methods highlighted the ability of the RM procedure to identify strongly coevolving inter-protein contacts and further strengthened our confidence in the identified binding spot for the Hsp40-Hsp70 interaction.

4.5 Coarse-grained modeling of the interaction

4.5.1 Simulation protocol

To further investigate the binding modes of the Hsp40-Hsp70 complex, we performed coarse-grained (CG) docking simulations of the complex using a methodology introduced in [112, 113], specifically designed for low affinity complexes. The CG model is based on a rigid body approximation of the two monomers, represented by one bead per residue, centered on the Cα atoms. The residue specific pairwise potential is composed of long-range screened

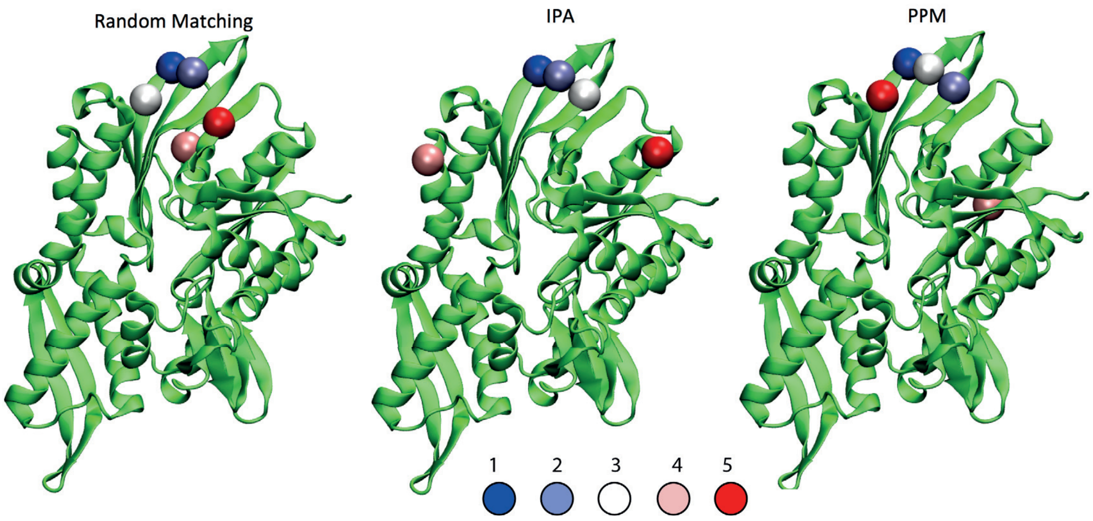

Figure 4.6 – Comparison of the RM, IPA and PPM predictions of the Hsp40/Hsp70 interactions. **A** Five most appearing interface contacts in the RM procedure. **B** Five top ranked interface contacts predicted by IPA. **C** Five top ranked interface contacts predicted by PPM. The ranks of the contacts are depicted from blue (first rank) to red (fifth rank).Figure adapted from [68].

electrostatics and short-ranged interactions

$$U_{ij}(r) = U_{ij}^{El}(r) + U_{ij}^{VdW}(r) \tag{4.2}$$

where r denotes the distance between Cα beads. The long-range electrostatic energy is modeled by a Debye-Hückel potential

$$U_{ij}^{El}(r) = \frac{q_i q_j}{Dr} e^{-\frac{r}{\xi}} \tag{4.3}$$

where q_i, q_j are the unit charges of the residues ($q_i = +1$ for Arg and Lys, $q_i = +0.5$ for His, $q_i = -1$ for Asp and Glu). The dielectric constant of water is set to $D = 80$ and the Debye screening length to $\xi = 10$Å. The residue specific Van der Waals potential is given by a modified 12-6 Lennard-Jones potential

$$U_{ij}^{VdW}(r) = \begin{cases} 4|\epsilon_{ij}|\left(\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right), & \text{if } \epsilon_{ij} < 0 \\ 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right) + 2\epsilon_{ij}, & \text{if } \epsilon_{ij} > 0 \ \& \ r < r_{ij}^{0} \\ -4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right), & \text{if } \epsilon_{ij} > 0 \ \& \ r \leq r_{ij}^{0} \end{cases} \tag{4.4}$$

where ϵ_{ij} is a residue-pair specific coefficient, based on shifted and rescaled Miyazawa-Jernigan coefficients e_{ij} [114]

$$\epsilon_{ij} = \lambda(e_{ij} - e_0) \tag{4.5}$$

where $\lambda = 0.159$ and $e_0 = -2.27 k_B T$ are scaling coefficients which have been fitted against experimental data [112, 113]. The scale parameters σ_{ij} of Eq.4.4 are given by the average of the standard radiuses of amino-acids i and j [112]. The canonical ensemble is sampled by replica-exchange Monte-Carlo simulations, and bound configurations are extracted from the ensemble by selecting frames having $E_{Pot} < -2 k_B T$ and at least one inter-protein residue pair having $r < 8$Å.

4.5.2 CG modeling of the Hsp40-Hsp70 interaction

Using the coarse-grained model presented above, we studied the Hsp40-Hsp70 interaction, focusing on the bacterial DnaK-DnaJ system. We took advantage of the availability of high-resolution structures of *E.coli* ATP-bound DnaK (PDB ID: 4JNE [55]) , ADP-bound DnaK (PDB ID: 2KHO [56]) and the DnaJ J-domain (PDB ID: 1XBL [110]), and built three constructs of DnaK : The isolated NBD either bound to ADP or ATP (NBD(ADP) and NBD(ATP)) and Full-Length DnaK bound to ATP (FL(ATP)). These various DnaK constructs were then used to investigate the influence of the nucleotide and the presence of the SBD at the CG level.

	NBD(ADP)	NBD(ATP)	FL(ATP)
K_D [μ M]	540 ± 60	370 ± 35	23 ± 3

Table 4.3 – Hsp40-Hsp70 dissociation constants computed by the CG model. Uncertainties are standard deviations computed over 5 replicates simulated at different box sizes.

We first numerically estimated the binding affinity of the different constructs (Tab.4.3), which were compatible with experimental determinations [67, 101]. The presence of the docked inter-domain linker and of the SBD of DnaK significantly increased the affinity, hinting to their stabilizing role in the interaction. In contrast, the nature of the bound nucleotide only marginally influenced the affinities. After extracting the bound conformations from the sampled ensemble, cluster analysis performed with the Gromos algorithm [115] identified two main clusters containing over 91% of the bound conformations (Fig.4.7A). Focusing the analysis on these two main clusters highlighted that the J-domain predominantly bound to DnaK at one location, situated near an upper cleft between lobes II and III of the NBD (Fig.4.7C). The main interacting region on the J-domain was composed mainly of helix II, with some contacts present on helices I and III (Fig.4.7B).

A finer analysis was achieved by considering the free energy surfaces of the bound conformations in spherical coordinates. Denoting by $\vec{I}_i^J, \vec{I}_i^K$ the normalized inertia axes of the J-domain and the NBD, computed over all Cα atoms, we introduced the set of three Euler angles defining the relative orientation of the J-domain with respect to the NBD , given by

$$\begin{aligned}
\Theta &= \text{asin}\left(\vec{I}_1^J \cdot \vec{I}_2^K\right) \\
\Omega &= \text{atan2}\left(\frac{\vec{I}_2^J \cdot \vec{I}_2^K}{\cos\Theta}, \frac{-\vec{I}_3^J \cdot \vec{I}_2^K}{\cos\Theta}\right) \\
\Psi &= \text{atan2}\left(\frac{\vec{I}_1^J \cdot \vec{I}_3^K}{\cos\Theta}, \frac{\vec{I}_1^J \cdot \vec{I}_1^K}{\cos\Theta}\right)
\end{aligned} \tag{4.6}$$

where atan2 is the quadrant-checking arctangent function. Additionally, we computed the spherical angles (Φ_{COM}, Ω_{COM}) defining the position of the J-domain center of mass (CoM) with respect to the NBD CoM , which characterized the overall location of the binding site on the DnaK NBD.

The free energy surface in CoM coordinates (Fig.4.7D,G,J) confirmed that there was one main binding spot of the J-domain on the DnaK NBD. The two main clusters detected by the Gromos algorithm were confirmed and appeared as two distinct clusters in the orientational free energy surface (Fig.4.7E-L). Several interesting observations could be drawn from these results. While there was a significant influence of the SBD and inter-domain linker on the binding affinities (Tab.4.3), the conformational ensembles were only very slightly perturbed by the presence of the SBD and linker. This indicated that their influence on the binding was predominantly of energetic nature and did not significantly perturb the geometrical arrangement of the complex. Furthermore, the nature of the bound nucleotide did not influence the binding mode, at least at a coarse-grained level. Structurally, the conformations falling in the two clusters were characterized by the same binding site, but a relative orientation of the J-domain approximatively rotated by 180° (Fig.4.8). The orientation of the J-domain lead to two opposite positioning of the characteristic HPD motif of the J-domain. In the first ensemble, the HPD motif pointed outwards with respect to the NBD, especially it was orientated opposite to R167 of the NBD, the co-mutated residue observed in [65]. In the second ensemble, the HPD motif pointed inwards with respect to the NBD, and particularly in the direction of the R167 residue on DnaK. We therefore denoted these two bound ensembles as HPD-OUT and HPD-IN (Fig.4.7E and Fig.4.8).

Comparison between the coarse-grained simulations and the RM results discussed above (see Sec.4.4.1) showed an excellent overlap of the predicted binding interfaces (Fig.4.9). Indeed, the three coevolving residue pairs identified by RM perfectly lied in the binding region predicted by the coarse-grained model. This strong overlap of the binding sites predicted independently by RM and CG modeling permitted the use of the coevolutionary predictions to assess the validity of the two CG clusters displaying opposite relative orientations. Close inspection highlighted a slightly better agreement between the HPD-IN conformation and the three coevolving contacts. The better fit of the HPD-IN conformation was further underlined by

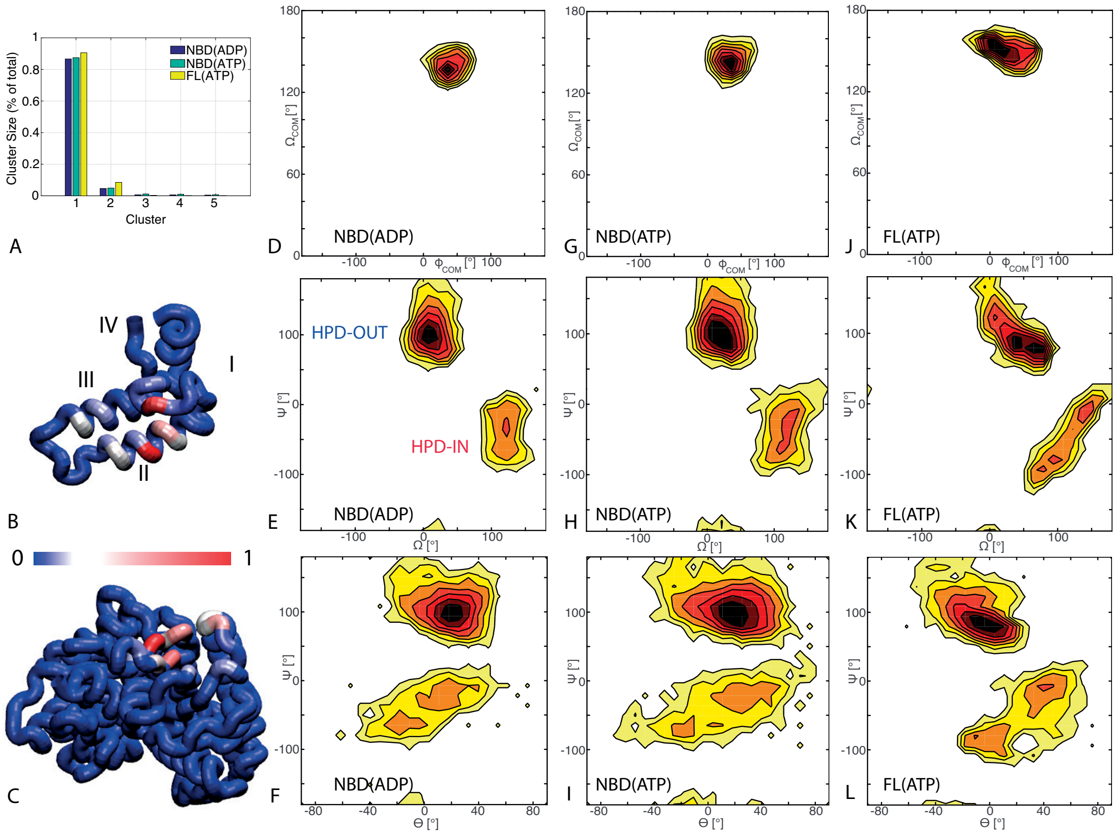

Figure 4.7 – **A** Distribution of the cluster sizes identified in the bound conformations. Clusters are sorted in decreasing size order. Only the first five clusters are displayed **B** . Per-residue contact probability mapped on the CG model of the NBD(ATP) construct. A residue is considered in contact if it has a distance with any residue in the partner protein less than 8Å. **D,G,J** Orientational free energy as a function of spherical coordinates (Φ_{CM}, Ω_{CM}) defining the angular position of the center of mass of the J-domain, centered around the center of mass of the NBD. **E,H,K** Free energy surface as a function of the two Euler angles (Ω, Ψ), defined by the inertia axes of the J-domain and of the NBD. **F,I,L** Free energy surface as a function of the two Euler angles (Θ, Ψ), defined by the inertia axes of the J-domain and of the NBD. Iso-levels of all free energy surfaces are plotted at 1 $k_B T$ intervals. The locations of the minima corresponding to the HPD-IN and HPD-OUT ensembles are marked in **E** (see main text). Figure adapted from [68].

the better positioning of the D35 residues of the J-domain with respect to the R167 residue of the NBD. Hence, the HPD-IN conformation was in qualitative agreement with the results of the coevolutionary analysis, with experimental results of the DnaJ:D35N-DnaK:R167H double mutant [65] and with the PRE-NMR measurements which highlighted the involvement of helix II of the J-domain in this interaction [67]. In contrast, the HPD-OUT ensemble could hardly be reconciled with these results. Based on these promising features of the HPD-IN conformation, we investigated its stability and dynamical properties by performing atomistic simulations of the Hsp40-Hsp70 system, which is the subject of the next section.

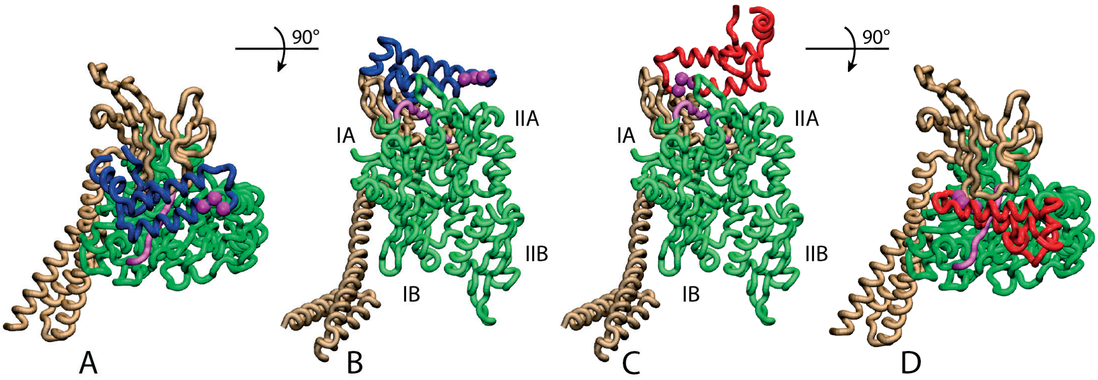

Figure 4.8 – HPD-IN and HPD-OUT conformations of the Hsp40-Hsp70 complex. The central conformations of the two ensembles of bound conformations are depicted on the FL(ATP) construct . **A,B** HPD-OUT. **C,D** HPD-IN. The three beads of the characteristic HPD motif of the J-domain are depicted by magenta spheres. The sub-domains of DnaK are highlighted: NBD: Green, SBD: Ochre, Linker: Magenta. The four lobes of the NBD are marked by IA,IB,IIA,IIB. Figure adapted from [68].

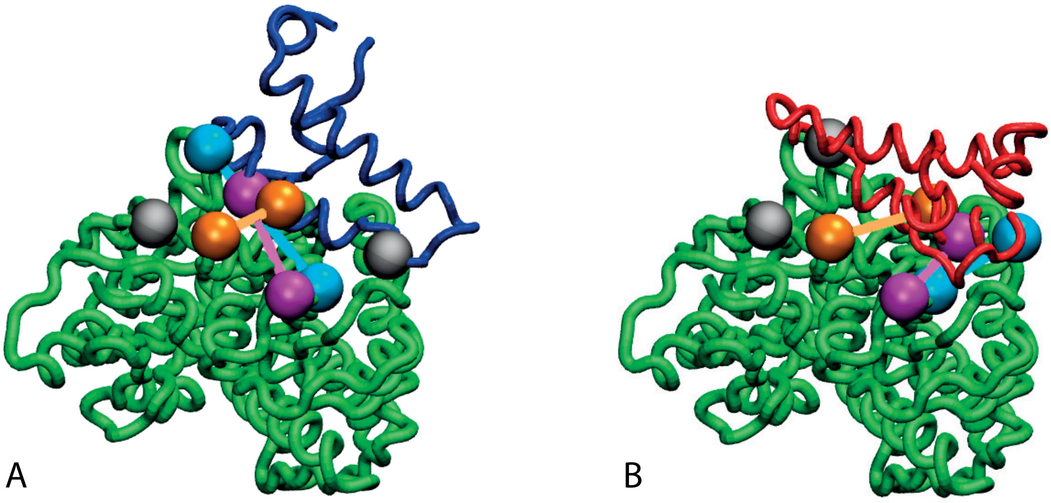

Figure 4.9 – Comparison of the RM and CG predicted Hsp40-Hsp70 interface. The three coevolving interface contacts are depicted as spheres centered on the Cα atoms. The shown complexes are the central conformations of the HPD-OUT (**A**, blue) and HPD-IN (**B**, red) ensembles for the NBD(ATP) case. D35 of the HPD motif and R167 are depicted by grey spheres. Figure adapted from [68].

4.6 Atomistic Simulations

To gain insights at a finer structural level inaccessible to coarse-grained descriptions, we performed explicit-solvent atomistic simulations of the complexes identified in the previous sections. We made use of detailed atomistic force-fields to investigate the structural stability, dynamical properties and energetics of the interaction.

To obtain all-atom representations of the candidate complexes, a mapping between the CG representation and fully atomistic models was required. To this aim, we relied on the RosettaDock package to build all-atom complexes based on the Cα backbone positions obtained by the CG model using a standard protocol [116, 117, 68]. We thus built models for each of the NBD(ATP), NBD(ADP) and FL(ATP) constructs, both in the HPD-IN and HPD-OUT ensembles. Starting from the central configurations of the CG ensembles, we generated 1000 all-atom configurations using RosettaDock and selected the 10 best scoring models, resulting in a total of 60 atomistic models.

These starting structures were then solvated in dodecahedral boxes with the TIP3P water model. Simulations were performed using the Gromacs 5 packages [118], with the Amber14 force-field [119]. In order to focus on the inter-protein dynamics, the intra-protein dynamics were first restrained using harmonic restraints on the backbone atoms of the constructs. We performed short runs of 30 ns for all the 60 constructs and analyzed the relative stability of the complexes using two measures: The distance root mean square (dRMS), defined by

$$dRMS(t) = \sqrt{\frac{1}{N_J N_K} \sum_{i,j}^{N_j,N_k} \left(d_{ij}(t) - d_{ij}(0)\right)^2} \tag{4.7}$$

where N_j, N_K denote the number of residues in the J-domain and DnaK and d_{ij} denotes the distance between the Cα atoms of residues i, j, and the angular orientation defined by

$$\Theta(t) = \mathrm{acos}\left(\vec{I}_1^J \cdot \vec{I}_1^K\right) \tag{4.8}$$

where $\vec{I}_1^{J/K}$ denotes the principal inertia axis computed on the Cα atoms. Although we did not observe any detachment event on this time-scale, the HPD-IN simulations showed significantly reduced dRMS and angular fluctuations compared to the HPD-OUT conformations (Fig.4.14, Fig.4.15, see Supplementary Figures and Tables), which further supported the relevance of the HPD-IN conformation as representative of the Hsp40-Hsp70 interaction. Notably, the presence of the SBD had a stabilizing effect on the interaction, as displayed by the reduced variability. Based on these premises, we concentrated our computational efforts on the FL(ATP) HPD-IN case, and investigated the stability and dynamics of the system through three independent 1μs simulations. While this longer time-scale did not allow full equilibration of the system, which remained computationally out of reach with atomistic details for this system, useful information could still be extracted from the ensemble of the simulations. Indeed, the three trajectories displayed a certain degree of structural variability (Fig.4.10A), transiently populating multiple sub-ensembles of bound conformations in the HPD-IN conformation, thus forming a highly dynamic interface, in agreement with the dynamic nature of the interface reported by Ahmad et al. [67] based on PRE-NMR measurements. Interestingly, as seen in Fig.4.10B, no single pose satisfying all coevolutionarily predicted contacts was found, but these were all transiently satisfied, again hinting to the biological relevance of a dynamic interaction interface. Of important experimental interest were the transient contacts formed by the D35-R167 residues, which were shown to be important for the correct functioning of the Hsp40-Hsp70 complex. Similarly to what was observed for the coevolving contacts, the dynamics did not show configurations permanently satisfying this contact, but multiple transient contact events were observed. These results show that the proposed HPD-IN conformation is compatible with coevolutionary predictions, respects experimental evidence based on NMR data (main interacting region on the J-domain is helix II) and mutagenesis (the D35-R167 contact is transiently populated). Our proposed model is thus computationally self-consistent at the sequence, coarse-grained and atomistic level and reconciles some divergent experimental evidences.

Interestingly, the pose of the J-domain in the HPD-IN conformation displayed an appreciable number of residues directly interacting with the SBD of DnaK. Whether these interactions are relevant for the interaction or are simply a consequence of the geometric arrangement of the complex is an important question, as most of the previous experimental efforts have been focused on the interactions of the J-domain with the NBD. This question will be addressed in the following section.

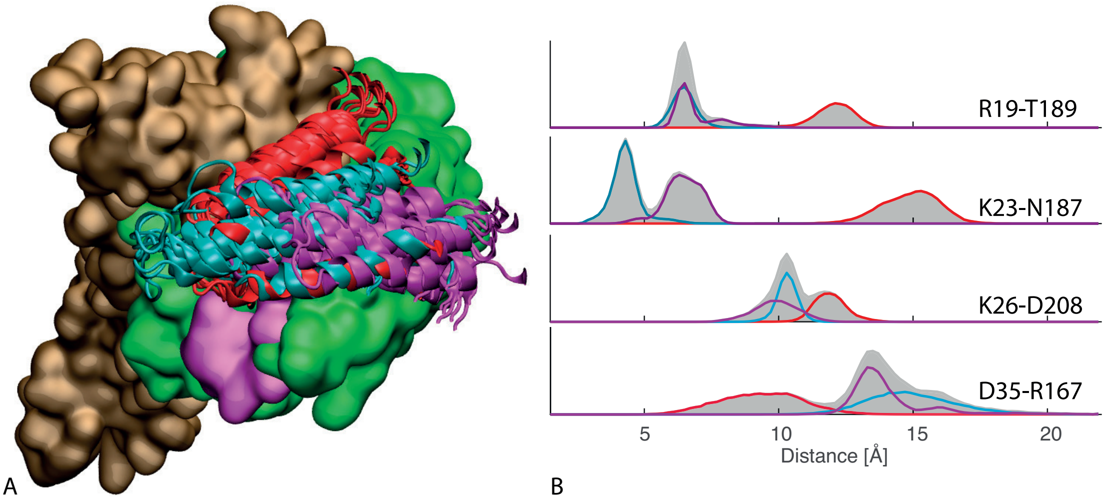

Figure 4.10 – 1μs time-scale atomistic simulations of Hsp40-Hsp70. **A** 10 snapshots of the DnaJ J-domain bound to FL(ATP) of DnaK in the three 1μs simulations. NBD: green, SBD: ochre, Linker: magenta, J-domain: Red/Cyan/Purple. For ease of visualization, only helices II and III of the J-domain are depicted. **B** Distance distributions of the three coevolving interface contacts and of the D35-R167 pair. The distribution for each of the three trajectories is depicted in the same color-scheme as panel A. The shaded distribution depicts the normalized sum of the three histograms. Distances are computed over all heavy-atoms. Figure adapted from [68].

4.7 Role of the SBD in the Hsp40-Hsp70 interactions

To investigate the active involvement of the DnaK SBD in the interaction with the J-domain, we analyzed the energetic contributions of the varying domains of DnaK and of the J-domain to the binding energy of the complex, using a generalized Born surface area (GBSA) approximation [119]. The GBSA consists in estimating the binding free energy of a protein complex by computing the polar contribution to the solvation free energy using a Generalized Born approximation (see [120] for details). For the 1μs trajectories, we computed the contribution ΔE of each residue to the total binding energy in the GBSA approximation, averaging over the three long trajectories of the HPD-IN conformation. This highlighted four main regions along the DnaK sequence strongly involved in the binding to the J-domain, lying on the NBD, linker and SBD, which formed a nearly continuous patch on the DnaK surface where the main interactions with the J-domain take place (Fig.4.11A-B). While the main energetic contribution to the binding unsurprisingly stemmed from the NBD, the SBD and linker significantly contributed to the energetic stabilization of the complex. These results show that the SBD thus plays an active stabilizing role in the interaction, compatible with the increased affinity computed by the coarse-grained model (Tab.4.3) and the reduced variability of the FL(ATP) construct observed in the short atomistic runs (Fig.4.14, Fig.4.15, see Supplementary Figures and Tables). Interestingly, as discussed above, the J-domain - SBD stabilizing interactions do not seem to significantly change the geometry of the bound complex at the coarse-grained level (Fig.4.7). The energetic analysis on the J-domain displayed as expected a main contribution from helix II, with a secondary signal coming from the coiled region linking helices II and III where the HPD motif is located (Fig.4.11D-E).

To address whether this active involvement of the SBD in the interaction was confirmed by coevolution, we repeated the RM protocol using an Hsp70 MSA containing full-length sequences. The results showed that two conserved inter-protein contacts involving the Hsp70 SBD (yellow spheres in Fig.4.13, see Supplementary Figures and Tables) were predicted among the most appearing contacts. We also noted that introducing the full-length sequence reduced the B_{eff}/N ratio, and resulted in a rank swap between the first contact involving a buried residue and the third selected contact in the RM approach using only Hsp70 NBDs. This would have resulted in not having selected the (189-19) contact if we had initially analyzed the full length sequences. This underlined the importance of previous experimental knowledge of the interaction when investigating systems by coevolutionary analysis. In this case, we could first focus on the J-domain - NBD interaction, which was the putative location of the main interaction, and then refine the results by extending the DCA analysis taking in account the presence of the SBD.

Thus both atomistic simulations and coevolutionary analysis predicted that the Hsp70 SBD is actively involved in the interaction with Hsp40 co-chaperones. Our results suggests that the role of the SBD is mainly to stabilize the bound conformation by contributing energetically favorable residues to the interaction interface. However, given the allosteric transition induced by ATP hydrolysis, their ought to exist a subtle conformational interplay between the conformation of the NBD containing the bound nucleotide, the docked SBD and the inter-domain linker. The exact mechanisms by which the conformational changes induced by ATP hydrolysis in the NBD are conveyed to the SBD are still unclear. We here showed that the presence of the bound J-domain must be taken in account for mechanistically understanding the allosteric mechanism of Hsp70, as its location in the interface forms energetic interactions involving three allosterically coupled sub-domains of Hsp70s.

4.8 Differential binding of bacterial and eukaryotic pairs

Our simulation results focused on the *E.coli* DnaK-DnaJ system, due to the availability of high-resolution structures for these bacterial proteins. In contrast, the coevolutionary analysis performed through RM was based on the complete Hsp40-Hsp70 families, containing both bacterial and eukaryotic sequences. A taxonomic gap thus exists between these two approaches, which calls for a phylogenetic assessment of the results obtained on the Hsp40-Hsp70 system. We thus asked whether our convergent results obtained by both sequence analysis and simulations pertained to both bacteria and eukaryotes or explain only the former.

Interestingly, the two experimental models for the Hsp40-Hsp70 interaction were based on members from different kingdoms. In [66], Jiang et al., analyzed bovine proteins, whereas Ahmad et al. [67] focused their NMR analysis on the bacterial DnaK-DnaJ system. The debate concerning the differences observed in the two models was partially settled by conjecturing that eukaryotes and bacteria might have differential binding modes of the Hsp40-Hsp70 interaction [103, 104]. Additional experimental evidence further suggested such differences in the regions involved in the interaction between bacterial and eukaryotic members [121]. One difficulty with interpreting such experimental evidences stems from the fact of observing particular pairs of Hsp40-Hsp70 paralogs, which might *in principle* have different binding modes, irrespective of the kingdom to which they belong. The experimental challenges faced when analyzing this complex (transient interaction inducing ATP hydrolysis, multiple conformations, nature of the bound nucleotide) adds to the difficulty of comparing models resulting from different experimental protocols.

Coevolutionary analysis of proteins family has the appealing property of working at a sequence ensemble level, namely extracting information about coevolving residue pairs pertaining to large ensemble of sequences. We thus decided to apply the RM approach to two sub-ensembles composed of bacteria and eukaryotes separately, in order to analyze whether kingdom specific differences were predicted. As seen in Fig.4.12A,B, while the results restricted to bacterial sequences were in full agreement with the predictions obtained on the complete dataset, the RM results on eukaryotic sequences lacked any significantly frequently selected inter-protein contacts. This clearly indicated that the overall results obtained on the full dataset originated in bacterial sequences, whereas no clear contribution from eukaryotic sequences was present. The fact that eukaryotic sequences did not show RM signal was an interesting effect *per se*. Several hypothesis for this effect can be conjectured. On the one hand, the number of

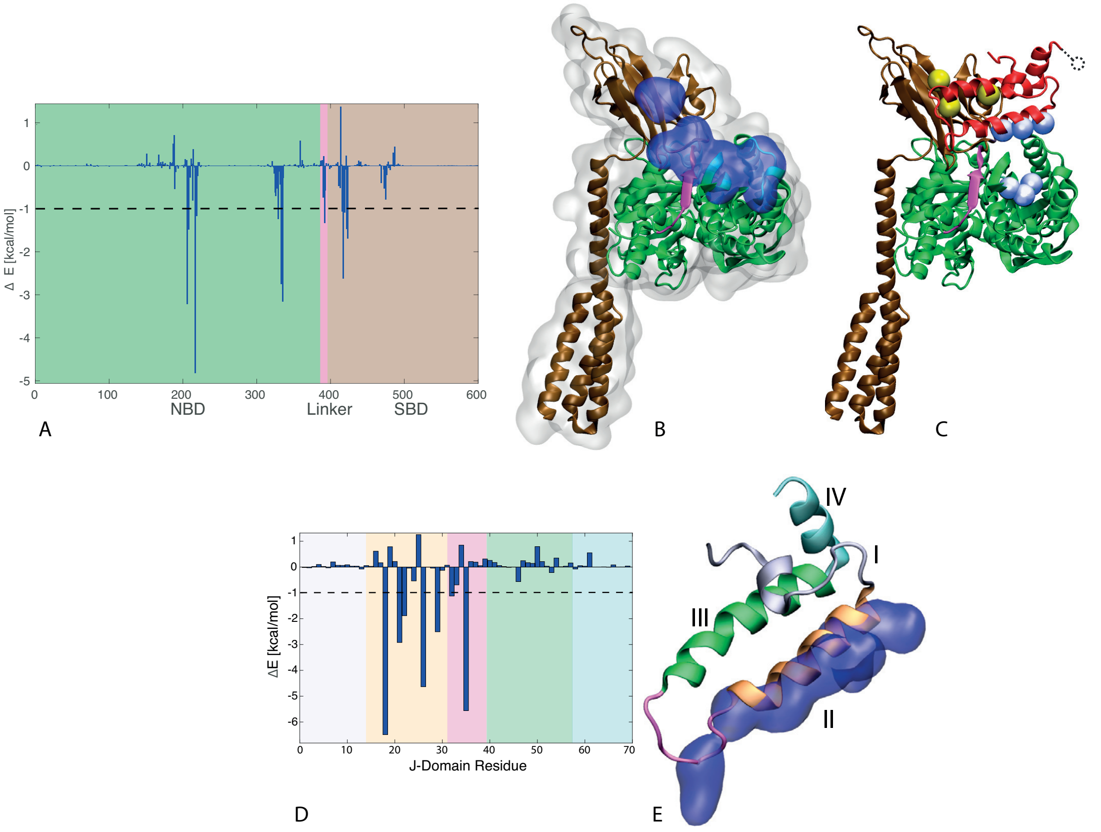

Figure 4.11 – GBSA analysis of the three 1μs trajectories of the HPD-IN conformation. **A** Per residue contribution ΔE of DnaK residues to the binding energy, averaged over the three 1μs trajectories. The dashed line represents the threshold used to depict the strongly contributing residues in panel B (-1$k_\mathrm{B}T$). The background coloring highlights the different DnaK domains: Green: NBD, Magenta: Linker, Ochre: SBD. **B** Most contributing residues to the binding energy ($\Delta E < -1 k_\mathrm{B}T$, see panel A), depicted in blue surface representation on ATP-bound DnaK . The domains of DnaK are colored following the same coloring scheme as in panel A. **C** The five inter-protein coevolving contacts predicted by RM. In blue are the three contacts involving the NBD, in yellow the two involving the SBD. The dotted circle represents the approximative location of residue E75 on the J-domain, which is not present in the NMR structure (PDB ID: 1XBL, [110]). **D** Per residue contribution ΔE of the J-domain residues to the binding energy, averaged over the three 1μs trajectories. The dashed line represents the threshold used to depict the strongly contributing residues in panel E (-1$k_\mathrm{B}T$). The background coloring highlights the different J-domain sub-domains: White: Helix I, Orange: Helix II, Magenta: HPD carrying loop, Green: Helix III, Cyan: Helix IV. **E** Most contributing residues to the binding energy ($\Delta E < -1 k_\mathrm{B}T$, see panel D), depicted in blue surface representation on the J-domain structure. The sub-domains of the J-domain are colored following the same coloring scheme as in panel A. Figure adapted from [68]

paralogs, and especially Hsp40 members, is much larger in eukaryotic organisms compared to bacteria (e.g. 6 Hsp40 members in it E.coli vs 50 in humans). The resulting combinatorial explosion of possible matchings might overthrow the ability of the RM approach to find robust contacts. Interestingly, the more involved IPA and PPM approaches did not perform better at contact predictions on the eukaryotic set, see [68]. A second hypothesis is that whereas the relatively simpler cellular organization of bacteria calls for strong selectivity and specificity for protein-protein interactions, eukaryotic cells have a number of alternative means by which protein-protein interactions can be regulated. Sub-cellular localization is the easiest way of ensuring selective interactions, without the explicit need to tune the physical affinity at the sequence level. Differential tissue-specific expressions of different paralogs might be an alternative way of regulating the interaction network. Finally, the temporal expression levels of the different chaperones and co-chaperones might also be regulated to maximize the interactions between specifically selected paralogs. All these different specificity selecting mechanisms do not, in principle, rely on the explicit optimization of physical interactions between particular pairs of paralogs, which would have a further advantage. In fact, if the specificity has to be fine tuned at the sequence level to promote specific pairs through high affinity, and selectively tune down cross-talk through lower affinity tuning, the sequence-level constraints on proteins having a large number of paralogs would become increasingly less evolutionarily viable. This could lead to sequences being evolutionarily frozen if the number of constraints to satisfy becomes too large. If these hypothesis are valid, our current coevolutionary analysis tools are not tailored to tackle such complex questions in eukaryotic organisms. These observations however raise interesting questions regarding the if and how such higher-level cellular organization principles can be combined with coevolutionary analysis tools to improve our understanding of eukaryotic protein-protein interactions and networks.

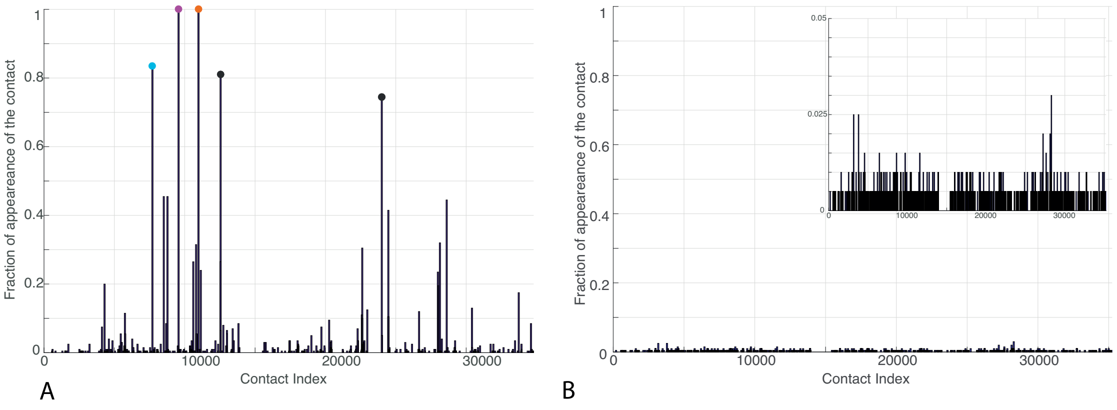

Figure 4.12 – Random matching results on Bacterial and Eukaryotic subsets. Fraction of appearances of the inter-protein predicted contacts in the RM approach in 1000 realizations. The abscissa is an arbitrary contact index, mapping each inter-protein residue pair to the natural numbers. **A** Bacterial dataset. The three colored circles correspond to the three selected contacts in the complete dataset. **B** Eukaryotic dataset (inset: Zoom on the vertical axis at the 0-5% level). Figure adapted from [68].

4.9 Discussion

The integration of coevolutionary analysis and molecular modeling at the coarse-grained and atomistic level allowed us to build a structural and dynamical model of the bacterial Hsp40-Hsp70 reconciling previously divergent experimental observations. The HPD-IN conformation previously discussed is in excellent agreement with the PRE-NMR data showing the main involvement of helix II of the J-domain and the ^{206}EIDEVDGEKTFEVLAT^{221} segment on DnaK NBD [67]. Furthermore, the transient interaction of D35 of the J-domain with R167 agrees with the large amount of experimental evidence pointing to the importance of this residue pair in the complex [65, 101]. Interestingly, we found that the binding interface is not influenced by the nature of the bound nucleotide, nor by the presence of the docked SBD. This latter domain is however actively involved in the interaction, by increasing the stability of the bound J-domain through energetically favorable contacts. In our structural model of the complex, the J-domain directly interacts with the docked inter-domain linker, which has been shown to be crucial for the correct allosteric signal transduction in Hsp70 chaperones [122, 123, 124].

The structural model of the Hsp40-Hsp70 complex proposed here is only valuable if it brings insights in the mechanistic functioning of the chaperone system. In this regards, it has to be interpreted in the light of the current understanding of the machinery. For this aim, two experimental studies stand out for the understanding of the allosteric mechanics of Hsp70s. In [57], Mapa et al. have analyzed the conformational ensembles of Ssc1, a yeast mitochondrial Hsp70, in either ATP- or ADP-bound states. Their findings show that Ssc1 is present in both the open (ATP-like) and closed (ADP-like) conformations when both bound to ATP and ADP. The relative weights of the populations are shifted depending on the nature of the bound nucleotide, thus leading to a dynamical equilibrium of Hsp70 fluctuating between the two experimentally observed conformations. In [124], Zhuravleva et al. predicted by NMR measurements that a third intermediate conformation should exist. In this conformation, which they dubbed allosterically active state, the two subdomains of the SBD are closed in a ADP-like fashion, whereas the inter-domain linker is still docked into the groove located on the NBD, as in the ATP-bound state. They argued that this allosterically active state should be an intermediate step leading to ATP hydrolysis. Thus, the emerging picture is that of Hsp70 exploring two characteristic conformations (ATP- and ADP-like, i.e. open or closed) with an intermediate allosterically-active state leading to ATP hydrolysis. Our structural model of the complex could neatly fall in this picture: As the docked J-domain covers the inter-domain linker, it could stabilize a conformation with docked linker, whereas its interactions with the SBD might lead to a rearrangement of the latter. Concurrently, the binding of a client substrate by the SBD can favor the closing of the two SBD sub-domains by energetically competing with the SBD-NBD and SBD-J-domain interfaces. Taken together, the concurrent binding of the J-domain and the substrate could synergistically shift the population of the Hsp70 chaperone towards the allosterically active state, thus favoring the hydrolysis of ATP. This functional interpretation of our structural model is of course at the moment hypothetical, and will require experimental verifications.

The use of coevolutionary methods, incarnated in our approach by the RM strategy, allows for computationally investigating structural aspects of protein-protein interactions. Whereas for analysis of monomers, the methodology is robust enough, so that the quality bottleneck is essentially dictated by the number of sequences, in the protein-protein analysis case, a further limitation arises due to the lack of knowledge of which (putatively) interacting paralogs to match. Recent progress has been made by the introduction of coevolution based methods [108, 109], however these methods do not perform better than the RM on the Hsp40-Hsp70 case. Of particular importance is the case of eukaryotes, where the number of paralogs is generally larger and where no operonic gene organization is found. The negative results obtained by RM, IPA and PPM on the eukaryotic set of Hsp40s and Hsp70s seems to indicate that deeper organizational principles are at play in eukaryotes. Further research will be needed to address this question, in order to fully exploit the predictive capabilities of currently available coevolutionary analysis tools. Beyond the intrinsic structural and functional interest of the Hsp40-Hsp70 interaction, its coevolutionary analysis also shed light on these methodological aspects that will require further investigations and refinements.

4.10 Supplementary Figures and Tables

Random Matching	PPM	IPA
K23 - N187	K23 - N187	K23 - N187
K26 - D208	E55 - T215	R19 - T189
R19 - T189	K26 - D208	K26 - D208
A64 - A176	R63 - A17	D59 - I39
A29 - L392	R19 - T189	A64 - A176
E55 - L219	Y25 - A191	A29 - L382
K51 - D224	A29 - L382	K50 - Y193
M30 - L320	Y25 - I338	A24 - G358
R36 - F356	Y54 - A376	I21 - I338

Table 4.4 – Nine top ranked interface contacts predicted by RM, IPA and PPM. Green cells highlight overlapping predictions. Yellow cells denote similar predictions, defined as predicted contacts involving one shared residue and the second one in proximity. Table adapted from [68].

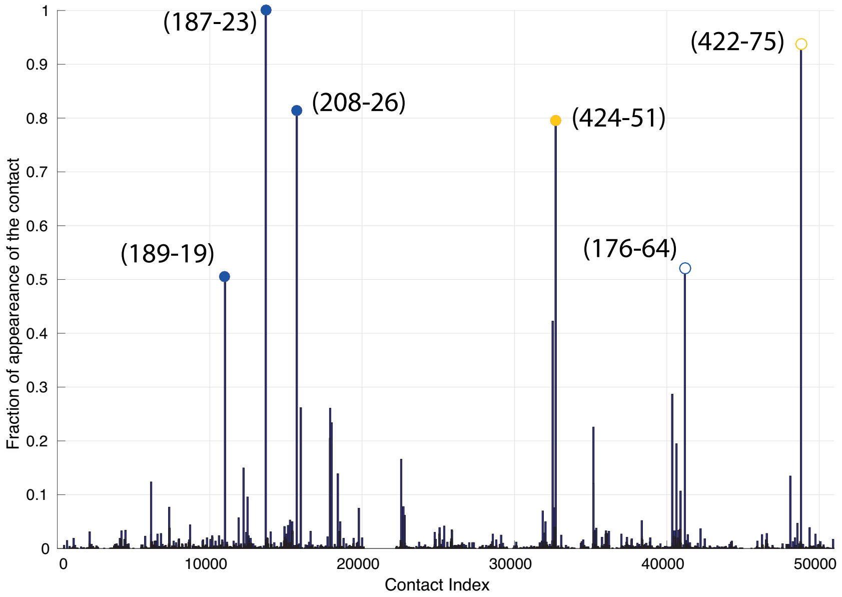

Figure 4.13 – Random Matching results on the Hsp40-Hsp70 interaction with full-length Hsp70 sequences. **A** Fraction of appearances of the inter-protein predicted contacts in the RM approach in 1000 realizations. The abscissa is an arbitrary contact index, mapping each potential inter-protein contact to the natural numbers. The filled blue-colored dots highlight the contacts before the first contact involving a buried residue, as selected by the NBD only RM procedure. The hollow blue circle depicts the first predicted contact involving a buried residue. The yellow circles depict J-domain - SBD predicted contacts. The empty circle denotes the contact involving E75 on the J-domain, which is not present in the crystal structure of the J-domain (PDB ID: 1XBL, [110]). The numbering of the highlighted contacts refer to the *E.coli* DnaK/DnaJ proteins. Figure adapted from [68].

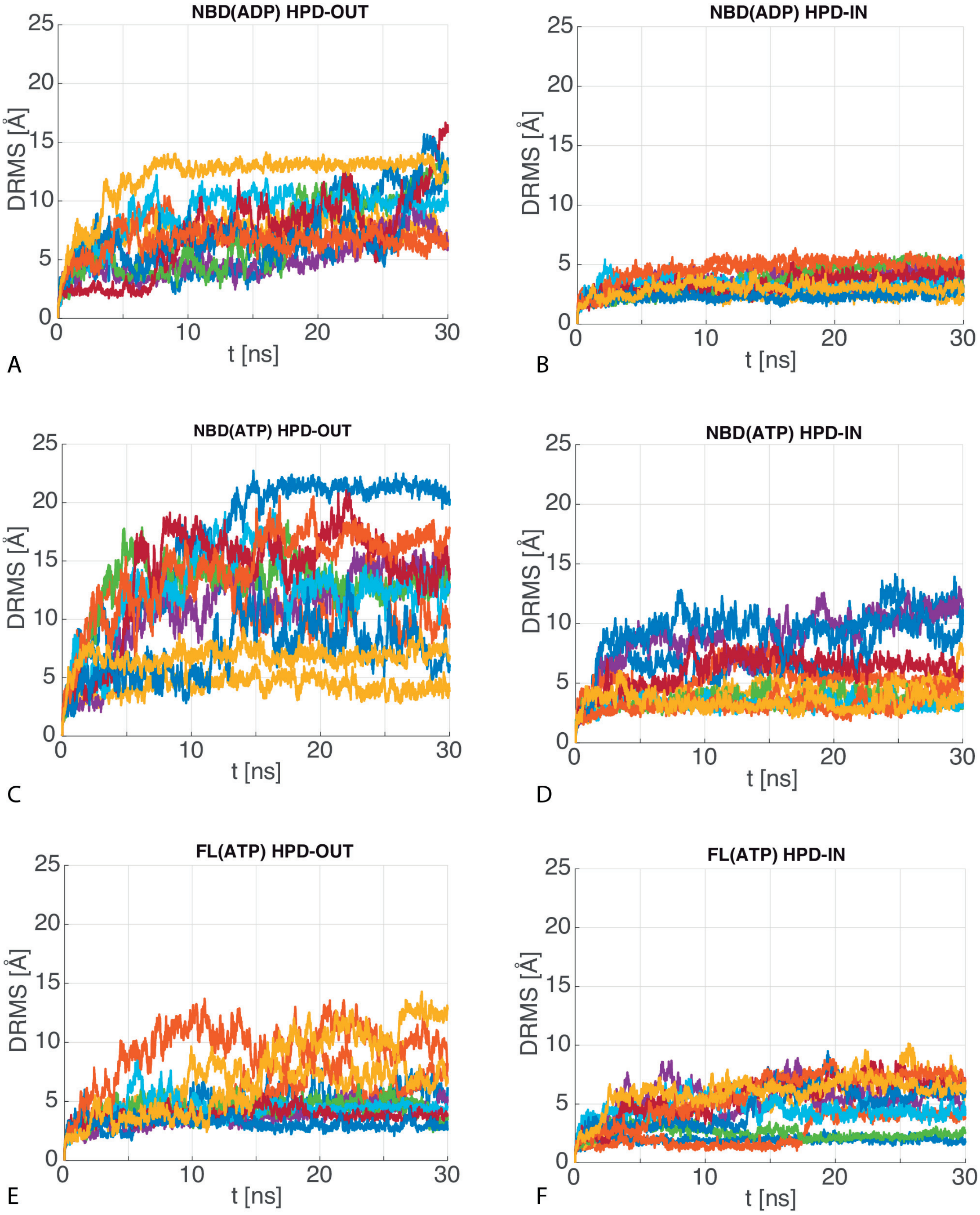

Figure 4.14 – DRMS of atomistic simulations of the Hsp40-Hsp70 complexes. Each colored time series represents a 30 ns simulation of the Hsp40-Hsp70 system. Figure adapted from [68].

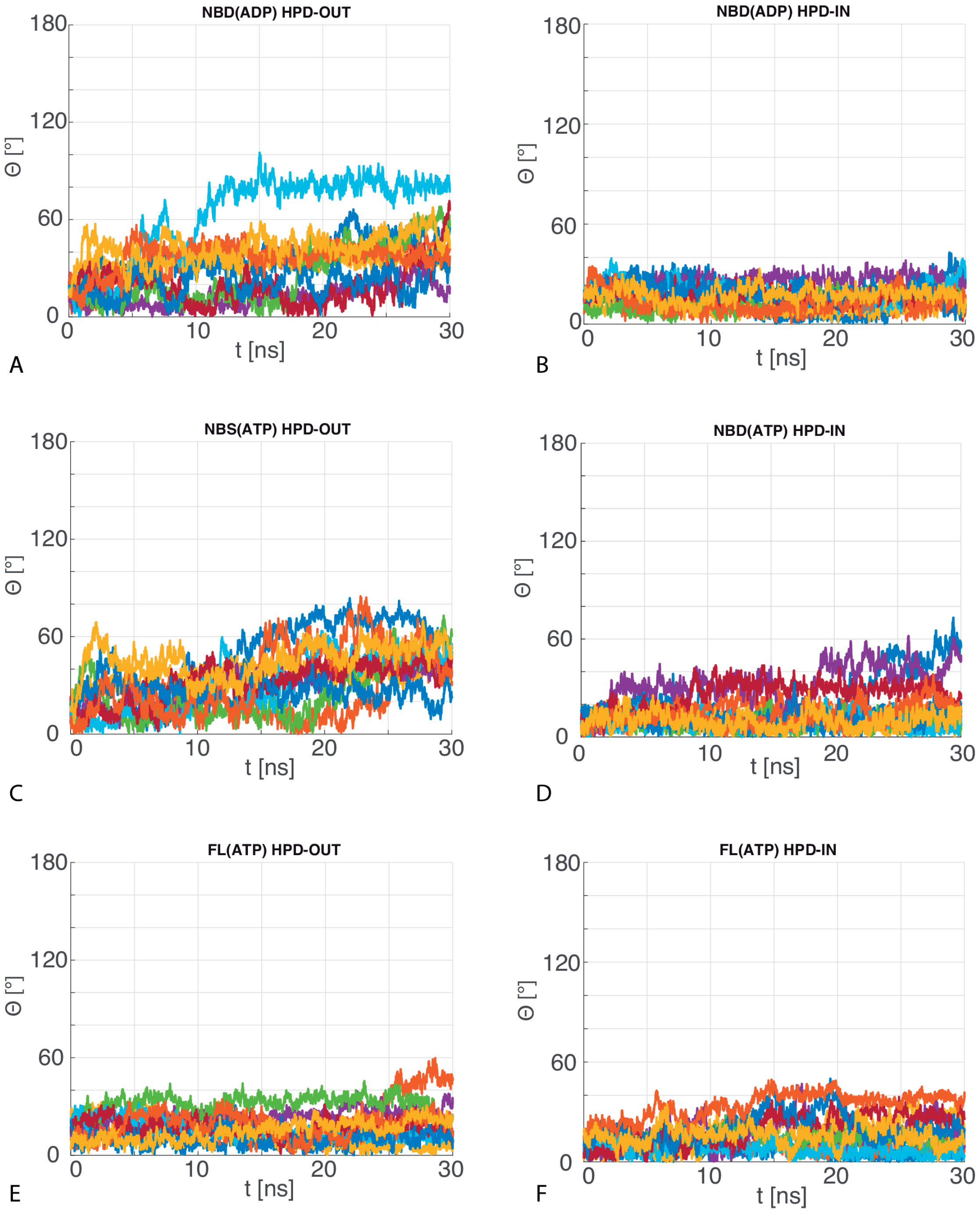

Figure 4.15 – Angular fluctuations of atomistic simulations of the Hsp40-Hsp70 complexes. Each colored time series represents a 30 ns simulation of the Hsp40-Hsp70 system. Figure adapted from [68].

5 Synergistic action of class A/B Hsp40s

The main results presented in this chapter have been published in [125]. The majority of the figures presented in this chapter are directly reproduced from the published article, in accordance with the Creative Commons Attribution License used by eLife.

5.1 Introduction

During the cells life cycle, a consequence of protein denaturation is the potential appearance of protein aggregates, which can have severe cytotoxic effects. Protein aggregates with dramatic consequences are indeed hallmarks of several human neurodegenerative diseases such as Parkinson's and Alzheimer's disease [126]. Multiple different effects can lead to aggregation: Shock conditions (heat or chemical), the aging of the cell or deleterious mutations destabilizing a proteins native fold. Among the principal roles of the cellular proteostasis network, carried out by molecular chaperones, is the prevention and disaggregation of such aggregates. It is thus not surprising that during evolution, all organisms have acquired chaperone machineries tailored for efficient protein disaggregation.

While Hsp70s have been shown to be efficient in the prevention of aggregation, their intrinsic efficiency in protein disaggregation is low [127]. Lower organisms, such as bacteria and yeast, and some eukaryotes such as plants, possess a potent chaperone machinery, called Hsp100, which is specialized in protein disaggregation [35]. In collaboration with the Hsp70 system, the Hsp100-Hsp70 bichaperone system has been shown to be extremely efficient at disassembling large protein aggregates, forming the basis of the disaggregation machinery in these organisms [128]. One of the major mysteries in the field of cellular protein quality control was how metazoans, and other higher eukaryotes which lack the Hsp100 disaggregase system, could efficiently control their aggregate population, thus avoiding their resulting cytotoxic effects.

This fundamental question has been recently addressed by Nillegoda et al. [129]. The authors have shown that in metazoans (*H.sapiens* and *C.elegans*), the Hsp70 system alone can target and disaggregate substrates of all sizes, with efficiencies comparable to the Hsp100-Hsp70 system in other organisms, provided it works in cooperation with a particular set of J-protein co-chaperones. J-proteins are traditionally classified in three classes A, B and C, depending on their C-terminal domain architectures (see Chapter 2). Specifically, the two classes A and B of canonical J-proteins share the same C-terminal domain (CTD) architecture, with the presence of two substrate binding domains (CTD I and II). Their main structural difference lies in the presence of a zinc-finger element in between the J-domain and the CTD I domain in class A J-proteins. Both class A and B J-proteins form stable homo-dimers, by interacting through the dimerization domain, located at the C-terminal end of the proteins. Functionally, these two classes bind through their CTD to non-native client proteins, which they then present to Hsp70, through their interaction mediated by the J-domain (see chapter 4). Class A and B J-proteins functionally differ by their substrate-binding affinities at the CTDs, with class A J-proteins targeting preferentially smaller aggregates, while class B J-proteins have higher affinities for larger aggregates. It was thought until recently that the different J-protein classes acted essentially independently, by targeting different substrates and localizations, thus driving Hsp70 specificity [64]. In their recent work, Nillegoda et al. showed that the simultaneous action of class A and class B J-proteins, together with the Hsp70 chaperone and Hsp110 nucleotide exchange factor, strongly increased the disaggregation power of the Hsp70 machinery compared to the efficiency of the Hsp70 machinery in the presence of solely class A or B J-proteins (Fig.5.1). Using a mixture of class A and B J-proteins at 1-to-1 stoichiometric ratios, they not only observed a strong synergistic acceleration of the disaggregation (Fig.5.1A,B), but also the targeting of a broad spectrum of aggregate sizes (Fig.5.1C,D). Furthermore, biochemical, cross-linking, FRET and mutagenesis experiments showed evidence that the increased disaggregation power of the A/B cocktail was likely due to a direct physical interaction between the J-domain of one class with a hinge region situated between CTD I and II of the other class J-protein. Particularly, the authors of [129] observed that the inter-class cooperativity was strongly dependent on ionic concentrations, thus hinting at interactions dominated by electrostatics. This was further strengthened by the observation that a triple charge reversal mutant on positions located on helices I and IV of the J-domain abolished the cooperative effect, without affecting the correct functioning of the J-proteins in isolation. This triple charge reversal mutant involved the mutation of three negatively charged residues to positively charged arginines and will be denoted as RRR triple mutant in the following. Interestingly, they observed that the inter-class interaction was symmetrical, namely that the J-domain of class A J-proteins (JD^{A}) could directly interact with the CTD region of class B J-proteins (CTD^{B}) and conversely, JD^{B} could also directly interact with CTD^{A}.

These new insights onto the disaggregation capabilities of the metazoan Hsp70 lead to new questions. Beyond metazoans, what is the phylogenetic distribution of the synergistic action of class A and B J-proteins? When did it appear in evolution, and are there organisms possessing both the Hsp100 disaggregase and J-protein synergy? Is there coevolutionary evidence for the inter-class direct physical interaction between the J-domain and CTDs? In collaboration with part of the authors of [129], we tackled these questions using phylogenetic and coevolutionary methods, which are the focus of the next sections.

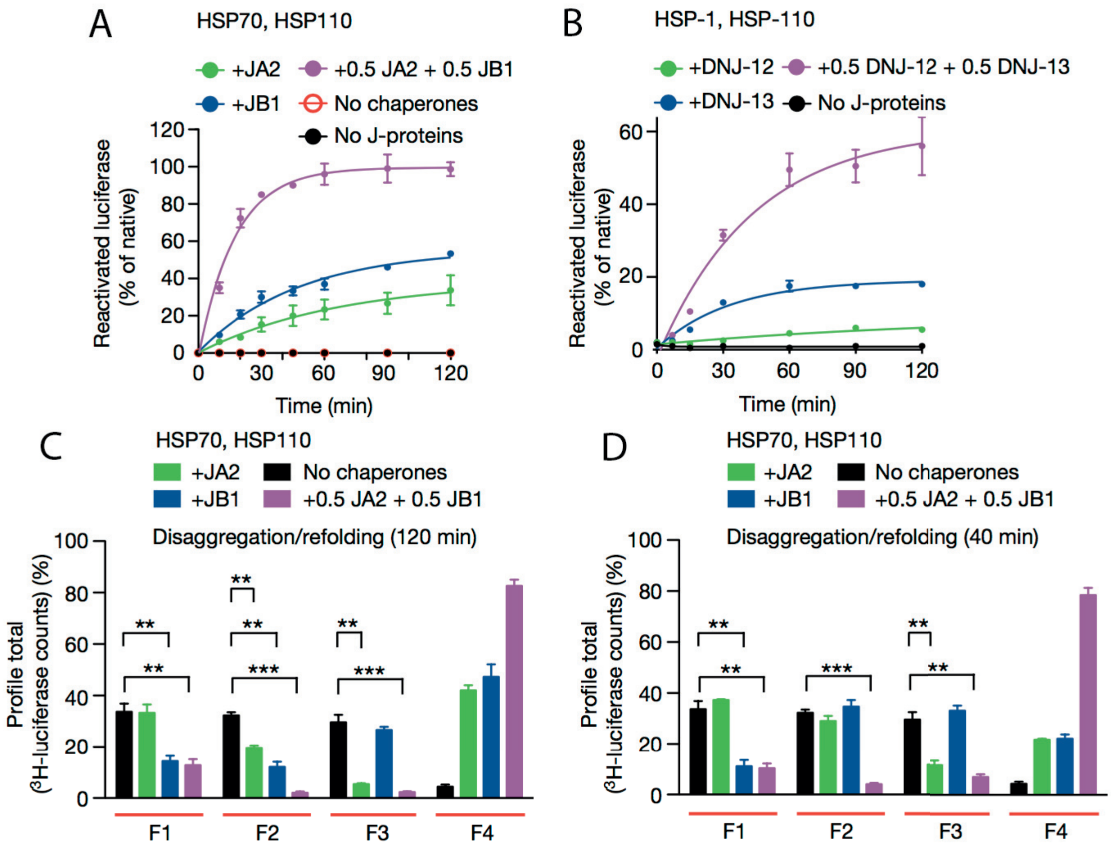

Figure 5.1 – Biochemical evidence for synergistic action of class A and B J-proteins. **A,B** Luciferase disaggregation and reactivation assay, using Hsp70, Hsp110 and class A and/or class B J-proteins. Luciferase aggregates were prepared by thermal denaturation. **A**: *H.sapiens*, **B**: *C.elegans*. **C,D** Aggregates size distribution after assays with class A, class B or class A+B J-proteins. The vertical axis shows the fraction of luciferase in the different size bins after the disaggregation/reactivation assay was performed (120 minutes for *H.sapiens*, 40 minutes for *C.elegans*). F1: ≥ 5000kDa, F2: 700-4000 kDA, F3: 200-700 kDa, F4: Disaggregated monomers (≈ 63 kDa).**C**: *H.sapiens*, **D**: *C.elegans*. Figure adapted from [129].

5.2 Sequence extraction and pre-processing

To perform both phylogenetic and coevolutionary analysis, we built MSAs of class A and B J-proteins. Given the high sequence and structural similarities between these paralogs, additional pre-processing and filtering steps were required to minimize the risk of mixing proteins from the two classes in their respective alignments. We started by building two manually curated alignments, for class A and B J-proteins separately. Both alignments covered the J-domain and the two CTDs. The class A MSA further contained the zinc-finger region characterizing this class. These two alignments were then used to build Hidden Markov

Models, using the hmmer package [83], which were in turn used to scan the complete Uniprot database (union of the Uniprot TrEMBL and Swissprot databases, release 06_2015). We then filtered both extracted MSAs, removing all sequences containing more than 20% gap. We observed that given the high sequence similarity between the J-domains and CTDs of class A and B J-proteins, many canonical class A sequences were present in the alignment of class B (reciprocally, only a small set of class Bs were found in the class A alignments, due to the lack of the zinc-finger in class Bs). To further remove these false-positives, we retrieved the complete sequences for both classes and checked for the presence of the characteristic CxxCxGxG motifs of the zinc-finger. We removed all sequences from the class B alignment for which the unaligned sequence contained at least one such motif. For the class A MSA, we removed all entries for which the unaligned sequences contained less than two CxxCxGxG motifs. Generally, the zinc-finger domain contains four such motifs, however, we observed that many zinc-fingers contained variations on the glycines of the motif, yet forming a full zinc-finger. We thus relaxed the filtering, requiring at least the presence of two such motifs per zinc-finger. This procedure resulted in two MSAs containing respectively 12215 sequences of class A and 4194 sequences of class B J-proteins, covering the whole tree of life. These MSAs were further split in two, by considering the J-domains and CTDs separately. This resulted in four different MSAs JD^A, JD^B, CTD^A and CTD^B.

5.3 Phylogenetic analysis of the synergistic interaction

Based on the experimental and computational evidences for the synergistic action of class A and B J-proteins in protein disaggregation, we aimed at studying this interaction at a phylogenetic level. While in their original work [129], Nillegoda et al. focused their analysis on metazoan organisms (*H.sapiens* and *C.elegans*), we aimed at characterizing this interaction over a broad taxonomic spectrum. With this aim, we first analyzed the phylogenetic distribution of class A and B J-proteins, both at the J-domain and CTD level. We then performed a discriminative analysis to identify the positions most involved in the phylogenetic differentiation.

5.3.1 Taxonomic distribution of J-proteins

To have an overall view of the phylogenetic distribution and evolution of class A and B J-proteins we built phylogenetic trees for the four MSAs. The trees were inferred using the RaxML package [130], using a standard protocol (20 maximum likelihood searches, 100 bootstraps per tree). To decrease the computational burden of the maximum likelihood inference of the trees, we pruned the MSAs, clustering sequences and keeping only sequences having a maximum of 90% sequence identity. This allowed to alleviate the computational burden of inferring large phylogenetic trees while maintaining a broad taxonomic coverage of the system. We then analyzed the inferred trees, by annotating the sequences according their taxonomic groups (Fig.5.2). The overall phylogenetic separation in the inferred trees was highly consistent, clearly separating prokaryotes from eukaryotes. A small group of eukaryotic class A J-proteins were found in branches lying in bacterial regions of the trees (Fig.5.2BC, pink lines). Inspection of these sequences revealed that these were located in chloroplasts and mitochondria, in full agreement with the bacterial origin of these organelles [88, 87]. In contrast, no class B J-proteins were found in these organelles. This observation is intriguing, given that class B J-proteins are commonly found in many bacterial clades. How and when the loss of class B J-proteins in mitochondria and chloroplasts occurred remains an unanswered question. Similarly, no class B J-domains were found in Archaeal organisms in our datasets. The grouping of J-proteins of organisms sharing lower taxonomic groupings (e.g. Fungi, Viridiplantae, Proteobacteria, Firmicutes) furthermore highlighted the consistent hierarchical structure of the trees. Interestingly, groups of sequences of the same taxonomic groups were found in different locations on the tree, hinting to functional specialization events in the Hsp40 family.

The biologically consistent organization of the phylogenetic trees confirmed that although the MSAs were of small widths (compared to the standards of large-scale phylogenetic analysis), they contained sufficient phylogenetic information regarding the taxonomic distribution and differentiation of J-proteins.

5.3.2 Phylogenetic signature of the J-protein synergy

Having determined that there was sufficient phylogenetic information contained in our MSAs to efficiently discriminate between different taxonomic groups, we phylogenetically investigated which residues were responsible for the synergistic interaction of the two J-protein classes. Independent computational investigations performed by the group of Rebecca Wade in Heidelberg determined that the electrostatic potential in the hinge region between the two CTDs was significantly different between bacterial and eukaryotic organisms [125]. The authors observed that there was a significant charge reversal in the electrostatic potentials in the hinge region of the CTDs between organisms of these two kingdoms. Furthermore, in vitro and in vivo experiments confirmed the lack of synergistic action of class A and B J-proteins in *E.coli*. Based on these experimental evidences, we thus investigated whether the clear phylogenetic distinction observed in the trees for the class A and B J-proteins was carried by residues involved in this region. To do so, we developed a discriminatory analysis, named Phylogenetic Discriminant Analysis (PDA), which will be outlined in the following section.

Phylogenetic Discriminant Analysis

The goal of Phylogenetic Discriminant Analysis (PDA) is to identify residues most involved in phylogenetically distinguishing sequences belonging to different taxonomic groups. These groups can in principle distinguish taxonomic clades at any phylogenetic level (i.e. eukaryotes/prokaryotes, fungi/metazoans/plants, mammals/reptiles/birds, etc.). The outline of the method is as follows: We start with an MSA containing annotated sequences, where each

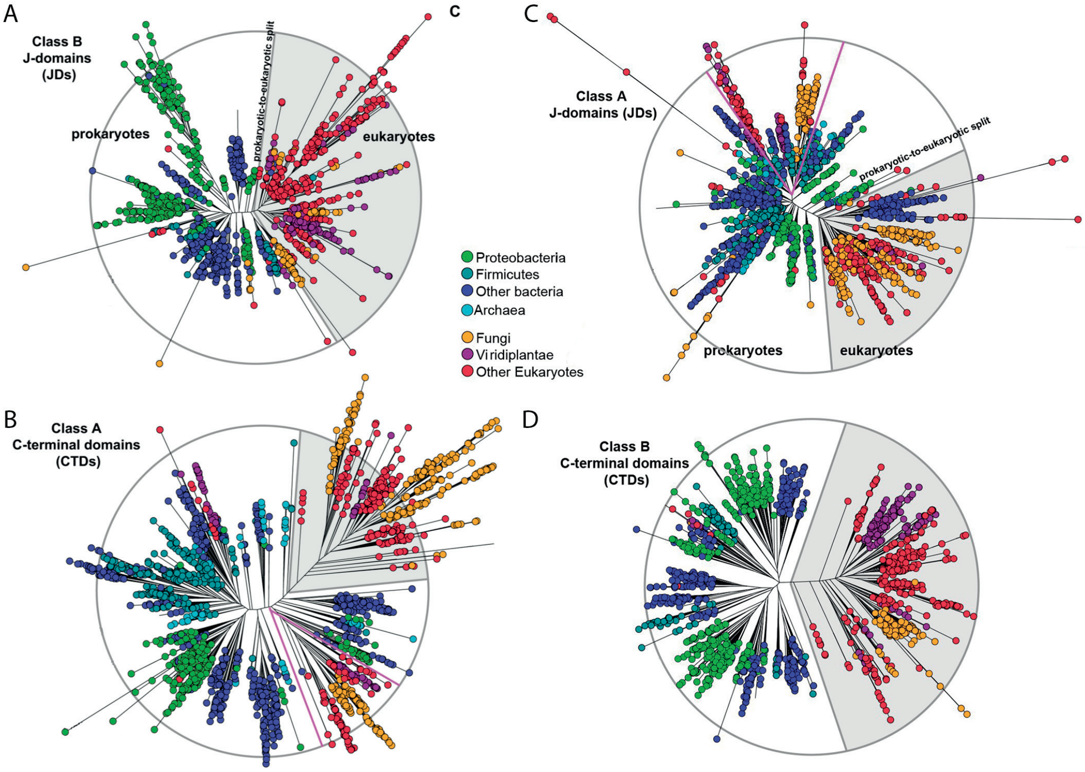

Figure 5.2 – Phylogenetic trees of the J-domains and CTDs of class A and B J-proteins. **A** Class B J-domains. **B** Class A CTDs. **C** Class A J-domains. **D** Class B CTDs. The gray area highlights the separation between main eukaryotic and prokaryotic sequences. Magenta lines approximatively delimit the eukaryotic sequences found in mitochondria and chloroplasts. Figure adapted from [125].

sequence belongs exclusively to one taxonomic group i. A random subset of n columns is extracted from the MSA and forms a new sub-MSA. Principal component analysis is then performed on the sub-MSA to project the sequences onto a lower-dimensional subspace. The sequences are then clustered in the low-dimensional subspace, forming C non-overlapping clusters. For each cluster $c \in C$, the distribution $P^c(i)$ of of the taxonomic groups is then computed, i.e. we build the histogram counting how many sequences in cluster c belong to group i for all taxonomic groups.

The homogeneity of the clusters is then measured by means of the Shannon entropy

$$h(c) = -\sum_i P^c(i) \log P^c(i) \tag{5.1}$$

where the sum runs over all the taxonomic groups i. A global mixing score is then computed as the weighted average of the cluster entropies

$$H(C) = \sum_{c \in C} w_c h(c) \tag{5.2}$$

where w_c denotes the total fraction of sequences falling in cluster c and C denotes the set of all clusters. The average entropy $H(C)$ is thus a measure of the phylogenetic mixing of the clustering performed using a given subset of positions. Subsets of positions for which $H(C)$ is low result in clusters having low internal mixing of taxonomic groups. These positions therefore have high discriminative power, as they generate a partition of the sequences consistent with the taxonomic grouping. The monotonicity of the entropy ensures that H is a good scale to score the discriminative capacity of subsets of positions.

This procedure of randomly extracting subsets of positions is then repeated for all possible subsets of n positions, and we record the entropy $H(C)$ for each sampled subset of positions. We then analyze whether there are single positions in the original MSA that appear frequently in the low-valued tail of the distribution of H(C). Positions which are significantly present in highly discriminant subsets are thus identified as strongly discriminating positions.

We used a hierarchical clustering algorithm to compute the clusters in the low-dimensional subspace [131]. This algorithm requires a definition of a distance cutoff for defining cluster splitting. We here employed the average distance between all sequences in the projected space as cutoff. We verified that our results were robust with respect to the use of an alternative clustering algorithm, based on modularity clustering (Fig.5.5).

In principle, PDA can be applied with an arbitrary number n of positions forming the subsets. However, the number of potential subsets grows exponentially with n, so that increasing this number rapidly leads to an intractable number of subsets. In addition, the number of subsets to sample also increases with the MSA width, making it impossible to sample all possible subsets for large proteins families. In those cases, we resort to random sampling of the subsets, and verify that the distribution of $H(C)$ is stable using cross-validation. This is achieved by performing several independent PDAs with a finite number of drawn subsets and by verifying that their computed distributions of $H(C)$ are statistically indistinguishable within sampling errors.

PDA of the class A and B J-proteins

As discussed previously, in vitro, in vivo and mutagenesis experiments, complemented by electrostatic computations led to the hypothesis that the appearance of the J-protein synergistic behavior appeared at the prokaryote-to-eukaryote split during evolution. We thus applied PDA to investigate which positions of JD^A,JD^B,CTD^A and CTD^B were most involved in discriminating bacteria from eukaryotes. Given the experimental identification of three positions on the J-domain involved in the RRR triple charge reversal mutation (D6R, 61R and E64R in Human DnaJA2 and D4R, E69R and E70R in Human DnajB1), we focused on the analysis of discriminant subsets of $n = 3$ positions. Analyzing triplets furthermore allowed us to exhaustively measure the discriminative power of all position triplets on the two J-domain MSA. For the CTD MSAs, due to their larger width, we had to randomly sample triplets and verified with sixfold cross-validation that our sampling was sufficiently deep.

The first question we addressed was whether the three positions identified on the J-domain triple charge reversal mutant were particularly discriminant between prokaryotes and eukaryotes. We thus computed the distribution of mixing entropies $H(C)$ for all position triplets on the J-domains, as described in the previous section. The mixing entropy of the positions involved in the RRR triple mutant was indeed located in the low entropy tail of the distributions for the JD^B case, while it appeared to be less discriminant for JD^A (Fig.5.3A,B, left panels, vertical red lines, p-value of 4.6% for JD^B, 23% for JD^A). Thus, the experimentally identified positions involved in the RRR mutant did collectively discriminate between prokaryotic and eukaryotic J-proteins in class B J-proteins, while the discriminative power of the triplet was not as well marked in class A J-proteins.

We then analyzed the compositions of the position triplets having stronger discriminative power than the RRR positions (i.e. lower entropies $H(C)$). To do so, we computed how often each position was present in triplets having lower mixing entropies than the reference RRR triplet. We observed that the composition displayed some marked peaks at particular positions that appeared significantly more frequently than others (Fig.5.3A,B, right panels).

To select the significantly outlying positions from these histogram, we employed a uniform prior null model: The average probability of a position to be randomly selected among N positions (dashed red lines in Fig.5.3) is simply given by

$$p_{Null} = \frac{3}{N} \tag{5.3}$$

Denoting by m the number of analyzed triplets, the standard error of the mean for each bin is given by

$$\sigma_p = \sqrt{\frac{p_{Null}(1 - p_{Null})}{m}} \tag{5.4}$$

where m is the number of selection positions triplets. Assuming a gaussian distribution of the errors on the bins, this allows quantifying the p-value of the measured bins in terms of standard deviations from the mean. To select the outlying positions in Fig.5.3, we selected all

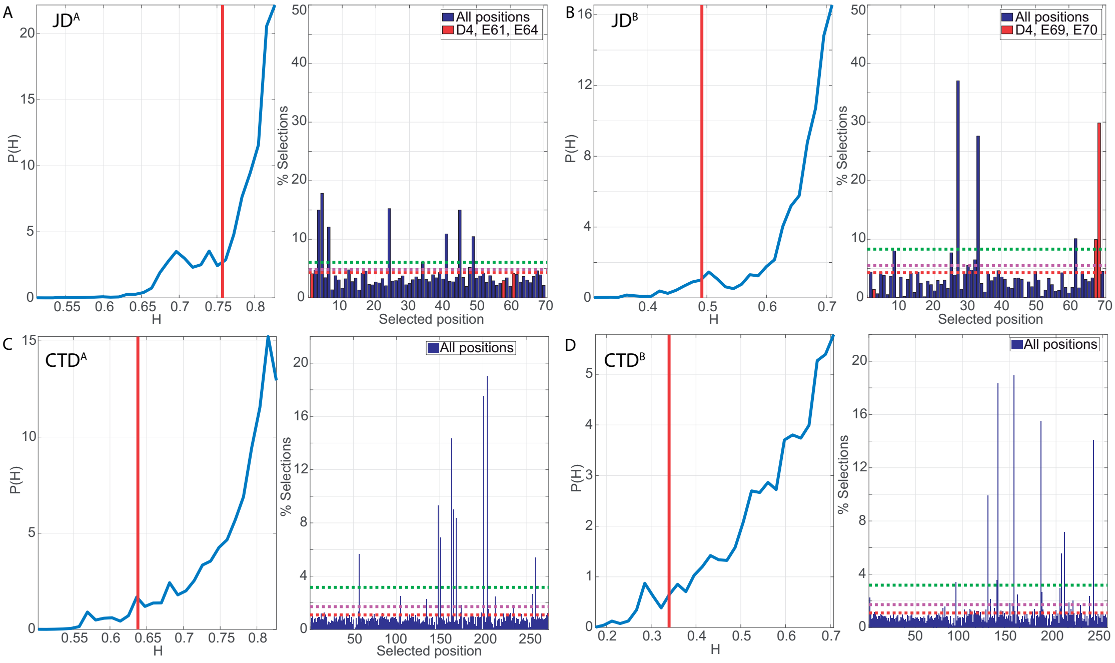

Figure 5.3 – PDA results on class A and B J-proteins. The left panels show the distribution of mixing entropies $H(C)$ as defined in equation 5.2. The red vertical lines denote the entropy of the RRR triple mutant triplet for the J-domains (A,B) or the limit corresponding to a p-value of 5% for the CTDs (C,D). The right panels show the selection probability of all positions. Each bin represents the fraction of times the position has been in the tail of the distribution (left of the vertical red lines). The red dashed lines depict p_{Null} as defined in equation 5.3. The green and magenta dashed lines represent the selection levels of the null model corresponding to 10-σ (green) and 3-σ (magenta) (see equation 5.4). For the J-domains (panels A and B), the three positions mutated in the RRR charge reversal mutant are highlighted by red bins. **A** Class A J-domain. **B** Class B J-domain. **C** Class A CTD. **D** Class B CTD. Figure adapted from [125].

positions above 10-σ (p-value $< 10^{-23}$, green dashed lines in Fig.5.3). We further verified that using a less stringent selection criterion (positions beyond 3-σ (p-value $< 1.5 \cdot 10^{-3}$, dashed magenta liens in Fig.5.3), led to the selection of additional residues in close vicinity of the regions identified using the 10-σ threshold. This procedure allowed the identification of 7 positions in class A and 5 positions in the class B J-domains which strongly participated in the discrimination between prokaryotes and eukaryotes (Fig.5.3A,B).

The analysis of the 5 class B J-domain residues mapped onto Human DnaJB1 showed that these positions were mainly found in two groups (red spheres in Fig.5.4A). One group formed of three positions was located at the C-terminal helix IV of the J-domain. Interestingly, among these, two (E69 and E70 in DnaJB1) were part of the residues which had been experimentally tested in the RRR triple mutant [129], while the third residue of this group (I63 in DnaJB1) was in close proximity to the two previously discussed positions. The second group was found in the coiled region linking helices II and III and was composed of two residues (L29 and K35 in DnaJB1), which directly flanked the highly conserved HPD motif. Thus, this discriminating region appeared to be involved in the interaction between the J-domain and Hsp70s. This last observation echoes to the discussion of chapter 4, and seems to indicate that there might indeed exist structural differences in the Hsp40-Hsp70 interaction between eukaryotes and prokaryotes.

The analysis of the discriminating positions found in class A J-domains (red spheres in Fig.5.4B) were less clear-cut. Here, the discriminating positions were mainly found on helix I, II and III. A cluster of three discriminating positions were situated in vicinity of D6 (in Human DnaJA2), which was part of the RRR triple mutant of class A J-domains. The discriminating residues on helix II flanked K46, which had been identified by chemical cross-linking as directly interacting with class B CTDs [129]. However, given the small spatial extent of the J-domain, the detailed interpretation of the discriminatory signal of such spread out residues remains difficult.

We reciprocally analyzed the most discriminating positions located on the CTDs of both classes of J-proteins (Fig.5.3C,D). As there were no reference triplets to test against for these cases, we selected all position triplets having p-value below 5% (vertical red lines in Fig.5.3C,D). In the class A CTDs, nine positions significantly contributed to the phylogenetic discrimination at the 10-σ level (Fig.5.3C). These were structurally clustered into three regions, two of which were located in the CTD I region, while the third group was found at the dimerization domain (Fig.5.4A). Among the strongly discriminant residues located in the CTD region, D222 and Y129 were found at the CTD I - CTD II hinge region, in close vicinity of K223, which was shown to cross-link with class B J-domains [129]. The other hits lying in the CTD region were clustered near the hinge between the CTD I and the zinc-finger region of class A J-proteins. The third group of residues were all structurally clustered in the dimerization domain. The analogous PDA on the class B CTDs revealed similar features (Fig.5.4B). We found a cluster of discriminant positions lying in the hinge region between CTD I and II, formed by I175 and K209. The latter residue was identify by cross-linking as directly interacting with the J-domain of class A J-proteins [129]. The two remaining groups of residues were clustered in one group located in the CTD I region and one in the dimerization domain.

To verify the robustness of the PDA predictions, we performed several validation tests. We first asserted whether the predictions were robust with respect to a finer taxonomic classification. To this aim, we classified the sequences into lower taxonomic groups at different levels and performed the same PDA analysis on the class B J-domains (Fig.5.5A: Proteobacteria, Other prokaryotes, Fungi and Other eukaryotes. Fig.5.5B: Proteobacteria, Firmicutes, Other prokaryotes, Fungi, Other eukaryotes. Fig.5.5C: Proteobacteria, Firmicutes, Other prokaryotes, Fungi, Viridiplantae, Other eukaryotes.). The PDA results obtained with these finer classifications were in qualitative agreement with the case discussed above, thus showing the robustness of PDA with respect to different description levels. Additionally, we verified if the PDA approach was sensitive to the clustering method used. We thus repeated the PDA analysis of the class B J-domains using a modularity based clustering algorithm [132]. The results obtained using this alternative clustering method were qualitatively in agreement with the results obtained using a hierarchical clustering method (Fig.5.5D). These results highlighted the robustness of

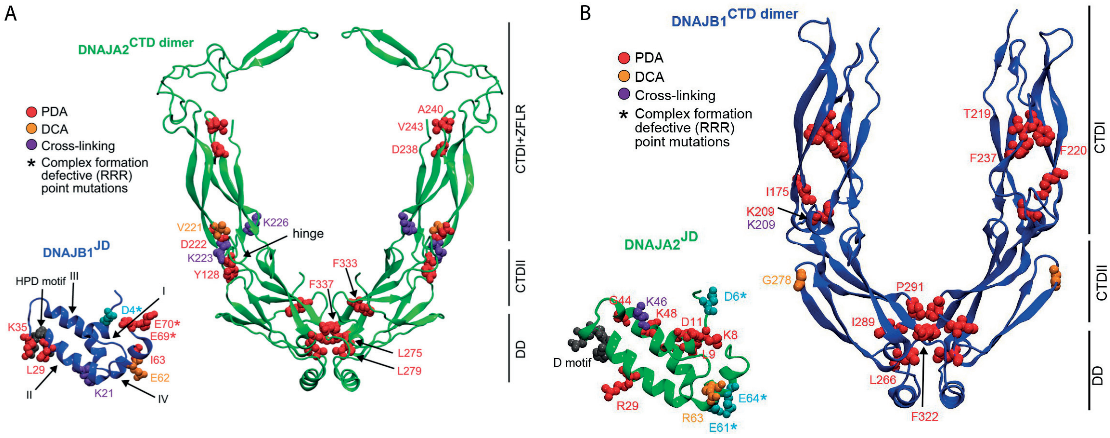

Figure 5.4 – Structural mapping of PDA and DCA predictions on class A and B J-proteins. **A** Class A CTD (green) and class B J-domain (blue) are depicted in ribbon representation. **B** Class B CTD (blue) and class A J-domain (green). Important residues are highlighted in colored spheres: Red: PDA predicted discriminating positions. Orange: DCA predicted inter-protein coevolving residues. Purple: Residues found in direct contact in the JD^B - CTD^A complexes by cross-linking. Cyan and (*): Residues mutated in the RRR charge reversal mutant. Figure adapted from [125].

the PDA method with respect to taxonomic classifications at finer levels and with respect to the clustering method used.

The PDA analysis thus identified a set of positions in the different MSAs which are mostly responsible for the phylogenetic differentiation of J-proteins between prokaryotes and eukaryotes. On the J-domain, the results on the class B J-proteins are readily interpretable: The phylogenetic discrimination is carried by two regions: One involved in the Hsp40-Hsp70 interaction, and the second in excellent agreement with the experimentally observed synergistic J-protein class A/B interactions. The results on class A J-domains are less easily interpretable, as the discriminatory signal is more spread out over larger regions. Whether functional and/or structural features can be directly assigned to these discriminating positions remains an open question. The analysis on the CTDs revealed consistent features between the two classes. In both cases, three main regions were predicted to be strongly discriminant. The first, located at the hinge region between CTDs I and II contains residues showed to directly interact with the corresponding opposite class J-domains in cross-linking experiments, and can thus be directly assigned to the synergistic inter-class J-protein action. The other two regions, predicted in both classes, are located on the CTD I and on the dimerization domain. The functional role of these discriminatory positions can not be readily interpreted. The predicted positions on the dimerization domain might hint to some differences in the dimerization behavior between J-proteins from the two kingdoms, although to the best of our knowledge, no canonical J-protein functioning as monomers have been reported. In summary, the PDA analysis highlighted a strong phylogenetic differentiation between prokaryotes and eukaryotes. This differentiation is mostly carried by the CTD I - CTD II region and by residues located on helix IV of the J, in

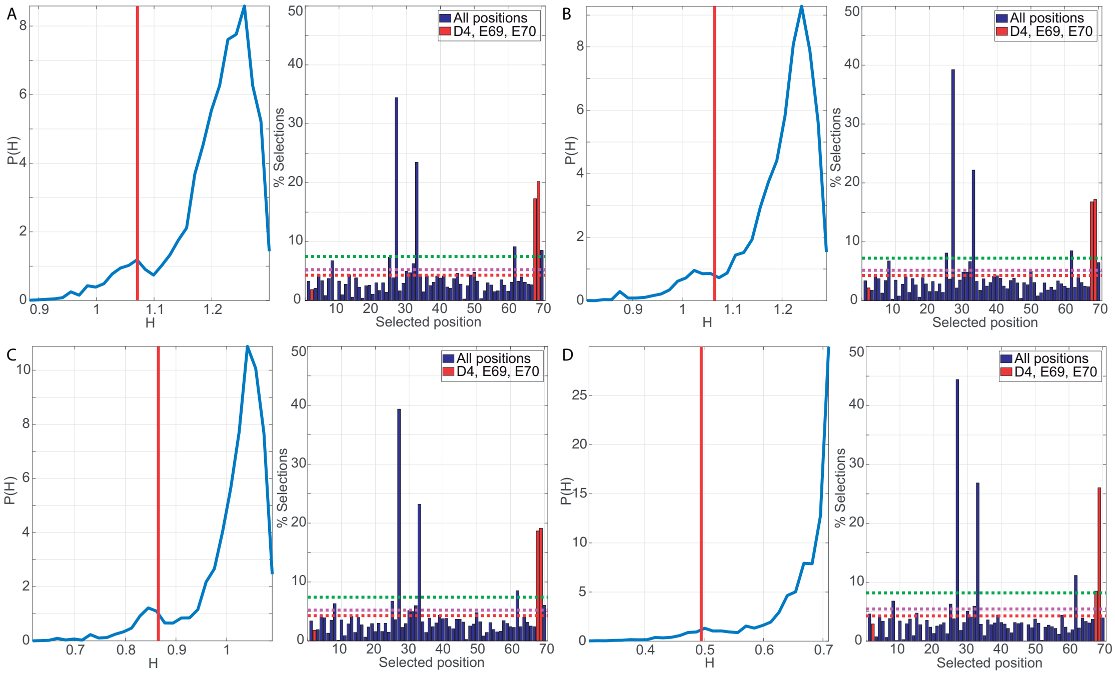

Figure 5.5 – Robustness analysis of the PDA approach. In all panels, class B J-domains are analyzed. Panels A-C show the results for finer taxonomic classifications. Panel D shows the results using a modularity based clustering algorithm (see main text for details). **A** Taxonomic groups: Proteobacteria, Other prokaryotes, Fungi and Other eukaryotes. **B** Proteobacteria, Firmicutes, Other prokaryotes, Fungi, Other eukaryotes. **C** Proteobacteria, Firmicutes, Other prokaryotes, Fungi, Viridiplantae, Other eukaryotes. **D** Taxonomic groups: Bacteria, Eukaryotes. Modularity based clustering algorithm. The same graphical elements as in Fig.5.3 are reported. Figure adapted from [125].

excellent agreement with previous experimental evidence [129]. PDA thus fully supports the cross-linking and FRET data regarding the location of the direct-interaction, and complements the previous experiments by the a large-scale taxonomic analysis, going beyond the study of a limited number of organisms experimentally available.

5.4 Coevolutionary Analysis

To inspect the structural basis of the direct inter-class J-protein interaction, we investigated by DCA whether the synergistic action of J-proteins left evolutionary footprints in the sequences. Given the lack of knowledge regarding which class A members interact with which class B J-proteins, we applied the Random Matching Strategy (RM) presented in chapter 4 to the potential JD^A - CTD^B and JD^B - CTD^A protein pairs. Note that in this case, the number of matchable sequences was severely limited by the size of the class B J-protein family (4194 sequences). Further selecting only eukaryotic sequences would have resulted in poor results already for intra-domain predictions. For this reason, although all evidence suggested that (most) bacteria do not posses synergistically interacting class A and B J-protein pairs, we decided to include bacterial sequences in the MSAs to be randomly matched. This choice was further supported by the observation that in the study of the Hsp40-Hsp70 interaction, although no significant signal was recovered from eukaryotic sequences, the RM results on the complete dataset were robust to their presence (see Chapter 4).

We thus performed 300 realizations of the RM procedure and recorded the selection frequency of all potential inter-protein residue pairs. Compared to the Hsp40-Hsp70 case, we did not observe inter-residue pairs with extremely high selection frequencies (Fig.5.6A,B). In fact, only a limited number of inter-protein contacts were selected during the RM procedure. Although the selection frequencies were overall low, two contacts (R63 - G278 in the JD^A - CTD^B and E62 - V221 JD^B - CTD^A) appeared distinctively more frequently compared to the average selection frequency. Mapping these two contacts on the structures of Human DnaJA2 and DnaJB1 interestingly revealed that they were located in the hinge region between CTD I and II on both class CTDs and onto helix IV on both J-domains (Fig.5.4A,B, orange space-filling spheres). Particularly in the case of the class A CTD, the coevolving residue V221 was in direct contact with K223, showed to interact with class B J-domains by cross-linking [129], and further contacted residue D222 which was identified as strongly discriminating by PDA.

The lack of precise knowledge regarding which class A and B paralogs interact, the low number of class B J-proteins and the high number of paralogs of J-proteins rendered the coevolutionary analysis of this interaction particularly difficult. This was reflected in the relative low quality predictions of the RM procedure. Although these premises, the RM results support a direct inter-class interaction between the CTD I & II hinge region and helices I and IV located on the complementary class J-domain. It will be interesting to re-investigate this inter-class coevolution in the future, when significantly more eukaryotic sequences will have been sequenced and we will potentially have a deeper understanding of the selectivity determinants between class A and B J-proteins.

5.5 Discussion

Trough independent phylogenetic and coevolutionary analysis of class A and B J-proteins, we computationally validate experimental evidences regarding the synergistic interactions between these co-chaperones. These findings support and strengthen the recent model of the Hsp70- JA/JB machinery as principal disaggregation machinery in higher eukaryotes lacking the Hsp100 system.

Using Phylogenetic Discriminant Analysis, we could identify which positions on the different domains were most involved in the phylogenetic differentiation at the sequence level. Our results show strong overlaps with the positions identified by experimental and other computational approaches [125]. Interestingly, beyond the agreement with the experimental data, our approach also identified supplementary regions involved in the prokaryotes-to-eukaryotes split. Whether these signals carry biological relevance, and what it might be, was beyond the

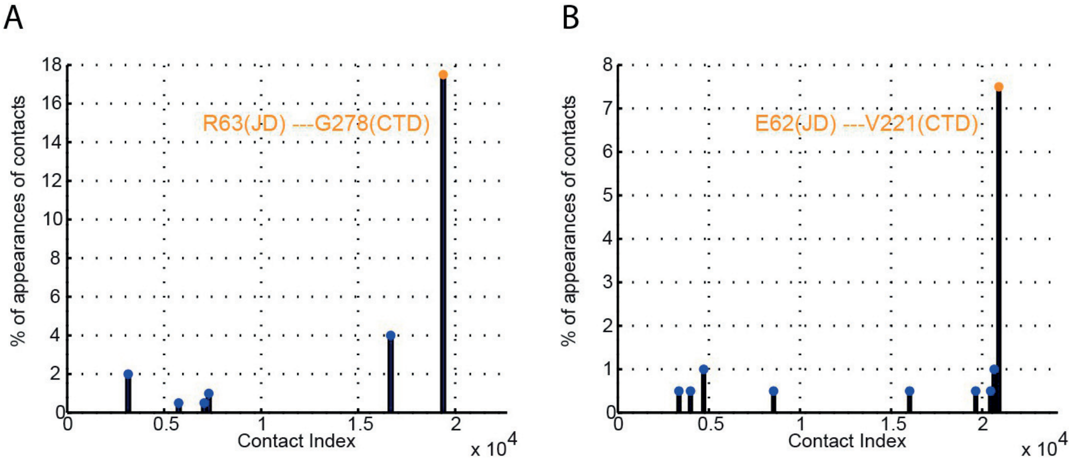

Figure 5.6 – Random Matching results on class A and B J-protein interactions. Fraction of appearances of the inter-protein predicted contacts in RM. The abscissa is an arbitrary contact index, mapping each inter-protein residue pair to the natural numbers. The two contacts depicted in orange are reported on the structural view in Fig.5.4A,B. **A** JD^{A} - CTD^{B}. **B** JD^{B} - CTD^{A}. Figure adapted from [125].

focus of this particular work. However, future investigations should address the question of the origin of these differentiations and analyze their functional and/or structural meaning.

The application of the RM procedure to investigate the direct interaction between the two classes was strongly hindered by the limited amount of available sequence data. In practice, the challenges raised by the class A - class B interactions are probable the worst case scenario for protein-protein interaction prediction based on coevolutionary methods. Nevertheless, although marginal, our results are in good agreement and support the model of a direct interaction between the inter-class J-domain - CTD interaction.

The computational results obtained in this work can only be fully appreciated when incorporated in the global study of the synergistic interaction between class A and B J-proteins. While this chapter mainly focused on the phylogenetic and coevolutionary analysis performed, our results are only informative when interpreted in conjunction with the experimental data gathered in this collaboration, and with the electrostatic computations performed independently [125]. Such an analysis of a highly complex biological interaction through the combined use of experimental, theoretical and computational approaches is indicative of the highly inter-disciplinary nature of the field of molecular biology, far from the historical segregation between wet bench and theoretical work. While none of the approaches could independently claim with high confidence the physiological relevance of our findings, their combined interpretation draws a strong biological picture of this exotic proteostasis pathway.

6 Coevolutionary study of Iron-Cluster pathway proteins

The main results presented in this chapter have been published in [133]. The majority of the figures presented in this chapter are directly reproduced from the published article, in accordance with the Creative Commons Attribution License used by Frontiers.

6.1 Introduction

The project presented in this chapter slightly diverges with respect to the previous discussion of coevolution in molecular chaperones. The system analyzed here is formed by proteins involved in the assembly of Iron-Sulphur clusters (Fe-S clusters, 2Fe2S or 4Fe4S). Fe-S clusters are cofactors which provide electrons in redox reactions and/or are involved in the stabilization of folded proteins [73]. An in-depth discussion of the chemistry of the biosynthesis of Fe-S clusters is beyond the scope of this chapter (and beyond the expertise of the author), the focus of the following discussion will thus be the structural properties of proteins involved in this pathway and their coevolutionary analysis.

The high chemical activity of free iron in the cellular environment is tightly controlled by a set of proteins forming the Fe-S cluster pathway. Indeed, an excess of free iron can form cytotoxic hydroxyl radicals, with potential deleterious effects on the cell [134]. During evolution, organisms have thus acquired specialized proteins involved in the chaperoning of iron and sulfur atoms, through their assembly in Fe-S clusters, and their subsequent transport towards acceptor proteins. Most of these ancient proteins are found throughout the evolutionary tree, displaying high sequence similarity. In eukaryotes, the machinery is located in the mitochondrion, and is tightly coupled to respiratory pathways [134]. In bacteria, several pathways are able to perform these tasks (located in the *nif, isc* and *suf* operons). In the following, we will discuss the case of the *isc* pathway, for which direct orthologs in eukaryotes are known.

The core machinery involved in the *Isc* biosynthesis of Fe-S clusters is composed of four classes of proteins (Fig. 6.1). IscS proteins are desulfurases which convert cysteine to alanine, thereby forming persulfides compounds which are subsequently incorporated in the Fe-S clusters.

The actual clusters are assembled on scaffold proteins IscU. These proteins directly interact with IscS homo-dimers. Upon binding of IscU to IscS, the persulfide is transferred to the IscU scaffold. It is thought that the Fe^2+ atoms are provided by CyaY proteins (Frataxin in eukaryotes), which act as iron chaperones in the process. Furthermore, CyaY also plays a regulatory role, controlling the speed of the cluster formation. The resulting clusters are then putatively transferred to IscA proteins, which can act as carriers of the Fe-S clusters towards acceptor proteins. IscA are also believed to play a role as alternative scaffolding proteins for the assembly of the clusters. Additionally, HscA chaperones and HscB co-chaperones (bacterial members of the Hsp70 resp. Hsp40 families) are directly involved in the *Isc* pathway. While most Hsp70 chaperones are known to have a large spectrum of client substrates, HscA is specifically tailored to interact with the IscU proteins. While the exact role of this interaction is yet to be understood, experimental evidence highlighted that HscA facilitates the transfer of clusters from IscU to acceptor proteins [73]. In addition, direct interactions of the HscB co-chaperones with IscU have also been reported [73]. These interactions, and their stimulation of the Fe-S cluster transfer were further shown to depend on the nature of the bound cluster (2Fe-2S vs 4Fe-4S), suggesting that IscU adopts different conformations depending on the state of the bound cluster. Although we have an overall understanding of the mechanism of the Fe-S biosynthesis, a large number of questions remain unanswered, especially regarding the structural aspects of the mechanism. In collaboration with Annalisa Pastore in London and Marco Fantini in Pisa, we addressed several structural questions regarding the proteins discussed above, using a coevolutionary approach: Is the N-terminal tail of IscU in a structured or unstructured state? What oligomeric arrangement of IscU is compatible with coevolutionary

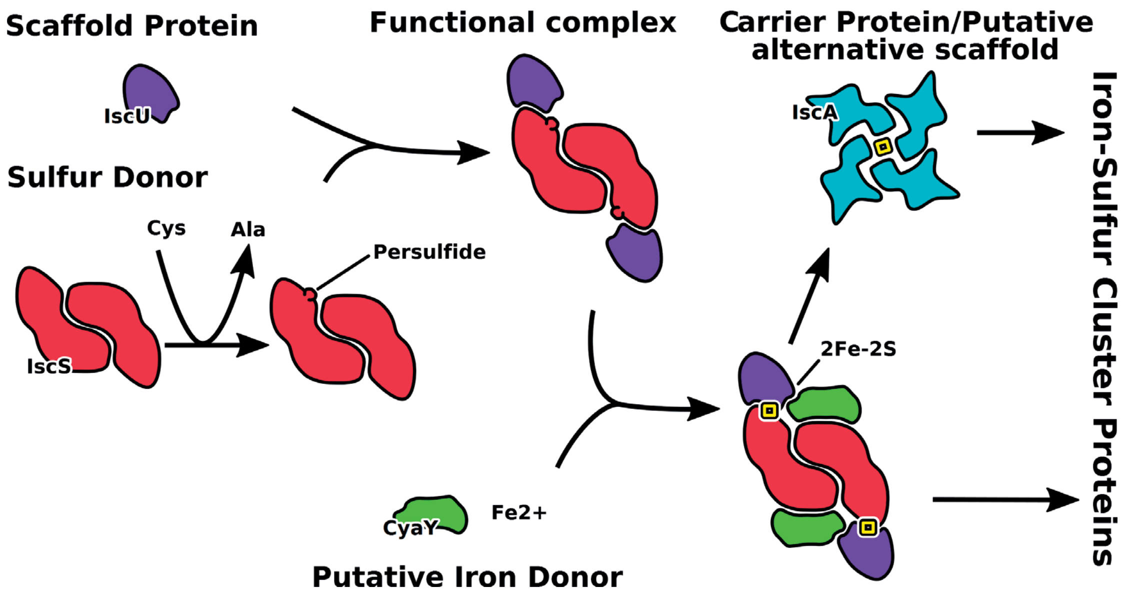

Figure 6.1 – Schematics of Fe-S cluster assembly. The persulfide generated by the desulphurase IscS is transferred onto the scaffold protein IscU. CyaY, a putative iron carrier, regulates the Fe-S cluster assembly. The final cluster is then transferred to IscA proteins, which can also act as alternative scaffold proteins. Figure adapted from [133].

predictions? Does DCA predict a biological functional oligomeric state of the IscA proteins, and what are the consequences on the cluster coordination? Are there coevolutionary traces of the inter-protein interactions among proteins involved in the *Isc* pathway?

6.2 Structural insights into the IscU family

The IscU proteins form the central scaffold on which Fe-S clusters are assembled. IscU are structurally characterized by a globally compact fold, composed of a three-stranded β-sheet packed against an arrangement of α-helices (Fig.6.2A,B). Experimentally, divergences exist regarding the N-terminal state of IscU (residues 1-21). Structural models obtained by X-ray crystallography all display a structured state of the N-terminal of IscU [135]. This tail is packed against the β-sheet region (Fig.6.2A, PDB ID: 3LVL). In contrast, a structural model based on solution NMR displays a completely unstructured and flexible N-terminal (Fig.6.2B, PDB ID: 1R9P) [136].

To investigate the state of the N-terminal tail of IscU by DCA, we built an MSA of the IscU family using the same methodology as presented in chapter 3. After searching the Uniprot database (union of TrEBL and Swissprot, release 11_2015) with a manually curated seed, we kept sequences containing a maximum of 10% of gaps, which resulted in an MSA defined over $N = 119$ positions and containing $B = 13,148$ sequences for the IscU family. We then used the asymmetric pseudo-likelihood version of DCA to predict intra-protein contacts. The high quality of the predictions, resultant of the large number of sequences composing this family, allowed us to analyze the 2N highest ranked contacts (ignoring all predicted contacts separated by less than 5 positions along the chain), with an excellent precision score (88% precision compared to 3LCL on the top N predictions, 78% precision on the top 2N). The comparison of the DCA predicted contacts with two reference structures (3LVL, crystal structure with structured N-terminal and 1R9P, NMR structure with unstructured N-terminal) displayed overall excellent predictions over most of the protein. As expected, the main differences between the two structures were found in the N-terminal segment. A large number of DCA contacts were predicted in the region involving the N-terminal segment (Fig.6.2C,D, boxed region) These contacts corresponded to interactions between the approximatively first 20 residues and the other secondary structures forming the core of IscU and were in excellent agreement with the contacts resulting from the structured state of the N-terminal (Fig.6.2A,C), whereas they displayed high shortest-path scores (SP, see Chapter 3) in the unstructured N-terminal structure (Fig.6.2B,D). These results clearly indicated that the structured form of the IscU N-terminal tail has been strongly evolutionarily conserved, and is thus functionally important in this family. It must however be underlined that this observation does not preclude the existence, nor the functional relevance, of the unstructured conformation of IscU, as it has been observed that DCA generally produces weaker signals from disordered regions of proteins compared to folded regions [137]. Furthermore, the existence of a fully unstructured state of the N-terminal would probably not rely on a set of well defined contacts, consequently not leaving an evolutionary footprint at the residue-pair level.

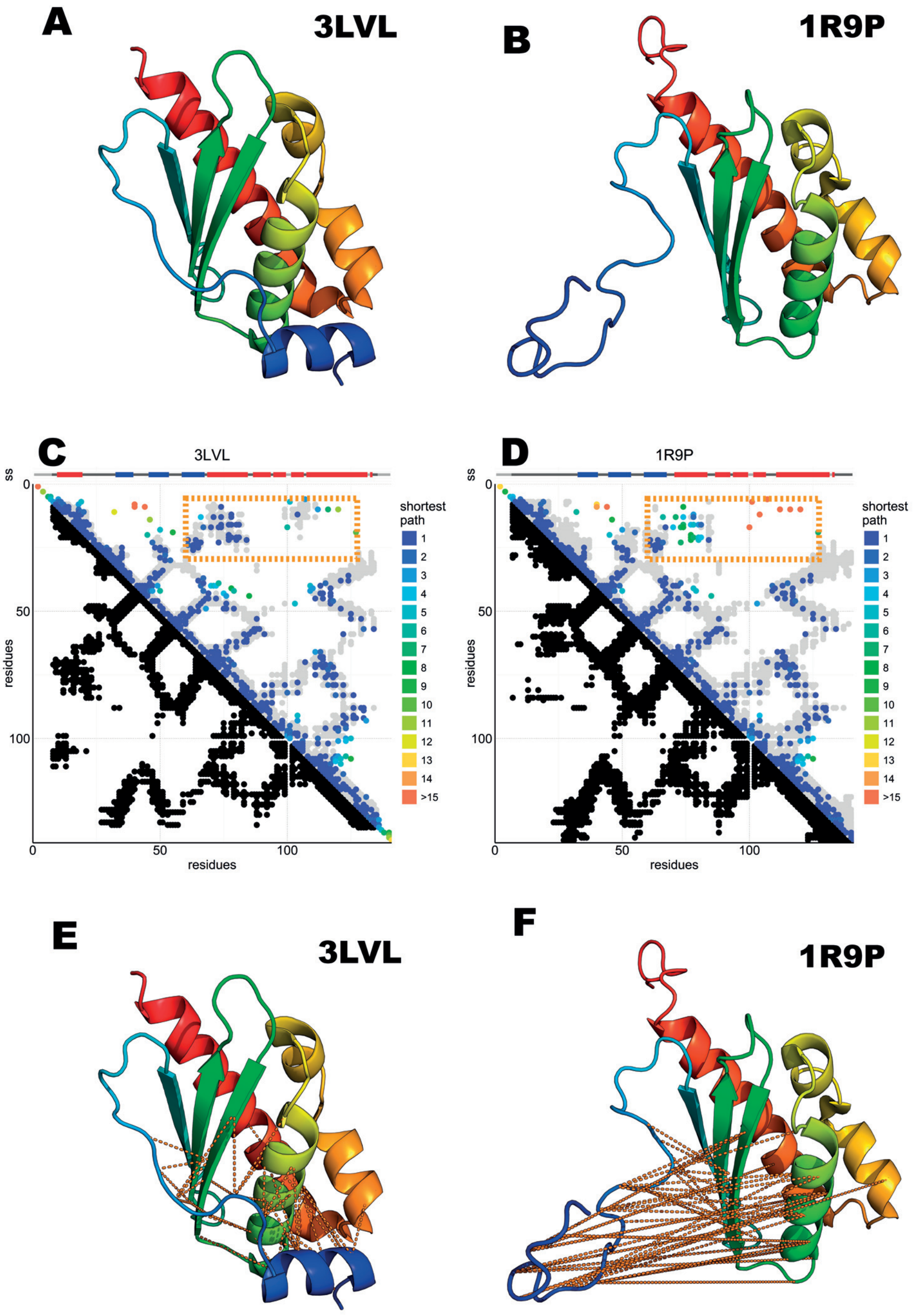

Figure 6.2 – N-terminal interactions of IscU. **A** Crystal structure of the structured N-terminal conformation of IscU (PDB ID: 3LVL, [135]). **B** NMR solution structure of the unstructured conformation of the N-terminal of IscU (PDB ID: 1R9P, [136]). **C,D** DCA predictions overlaid on structural contact maps obtained from **A** and **B**. Lower triangular part: Structural contacts, defined by a contact threshold of 8.5Å between heavy atoms. Upper triangular part: DCA predictions, colored by their SP score. The structural contacts are depicted in grey background. The boxed regions indicate the contacts involving the N-terminal region. **E,F** DCA contacts boxed in **C,D** depicted on the structured and unstructured models. The contacts are drawn between Cα atoms. Figure adapted from [133].

Close inspection of DCA predicted contacts outside the region involving the N-terminus revealed the presence of two small clusters displaying high SP scores (Fig.6.3A, highlighted boxes). These predictions could not be explained by inter-protein contacts as observed in crystal structures of IscU in either trimeric (PDB ID: 2Z7E), nor decameric (PDB ID: 2QQ4) arrangements (Fig.6.6, see Supplementary Figures). As these predicted contacts all lied in close vicinity of the binding site of the cluster on IscU, their predictions by DCA could be an effect of interactions being mediated by a non-proteic substrate. In this scenario, the correct coordination of the cluster must be ensured by residues forming the active site. Mutations of such residues must be compensated in order to ensure the correct localization and binding of the Fe-S cluster at the active site, thus displaying correlated mutations also in the absence of direct physical contacts. An alternative scenario is that these contacts reflect a head-to-head dimerization pattern of IscU. It has been proposed in literature that the correct coordination of 4Fe-4S clusters requires at least the formation of IscU homo-dimers [138]. In this scenario, the homo-dimerization of IscU would allow the co-localization of the two active sites of the monomers, potentially increasing IscUs ability to coordinate 4Fe-4S clusters.

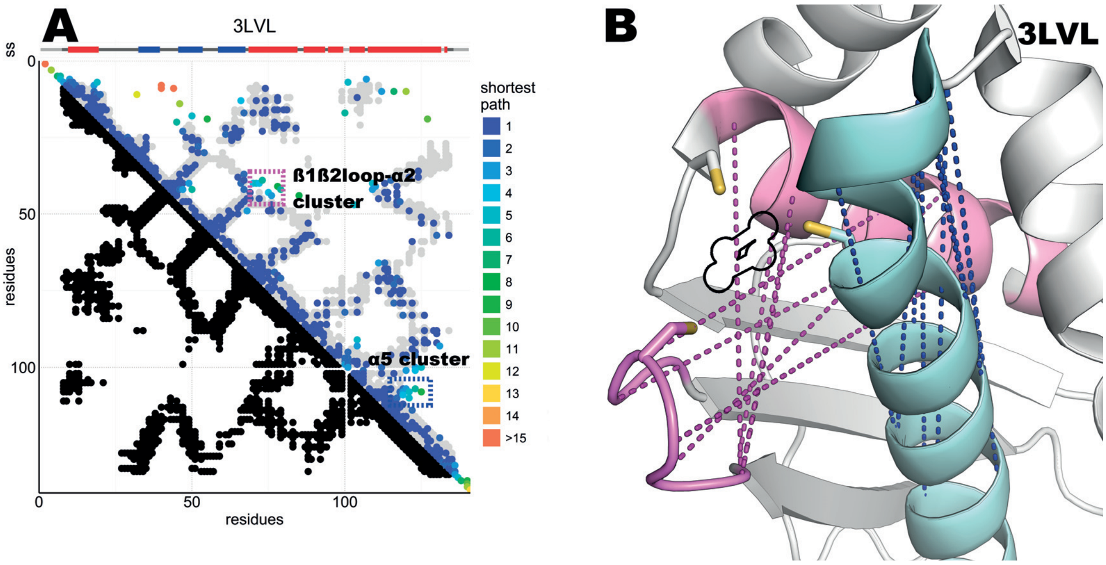

Figure 6.3 – Fe-S cluster coordination of IscU. **A** DCA predictions overlaid on the structural map of PDB ID: 3LVL. Lower triangular part: Structural contacts, defined by a contact threshold of 8.5Å between heavy atoms. Upper triangular part: DCA predictions, colored by their SP score. The structural contacts are depicted in grey background. The two boxed regions highlight the high SP predictions not involving the N-terminal. **B** Boxed contacts in panel A reported on the 3LVL structure. All contacts are between Cα atoms. The colors of the drawn contacts follow the coloring schemes of the boxes in panel A. The coordinating cysteines are shown explicitly. The Fe-S cluster is outlined in black. Figure adapted from [133].

6.3 Multimerization and cluster coordination of IscA

The alternative scaffold protein IscA presents several unresolved questions, among which the understanding of its coordination of Fe-S clusters through multimerization. Several structural models of IscA oligomers have been proposed. The authors of [139] proposed two possible tetrameric IscA arrangements based on crystallographic data (Fig.6.4A, PDB ID: 1R95). Both tetramers are composed of two identical copies of IscA dimers, but they differ in the relative arrangement of the two dimers. Due to unresolved electron density, both models lack the C-terminus of IscA, where two out of the three cysteines of IscA are located. Thus, the resulting cluster coordination could not be asserted based on these structural models. Cupp-Vickery et al. [140] obtained a similar tetrameric arrangement, and modeled the missing C-terminus based on stereochemical parameters. From the resulting model, the authors concluded that the cysteines of an isolated dimer would not allow the coordination of a cluster, and that thus a tetrameric organization was necessary. Furthermore, they argued that only two out of the three cysteines of each IscA would be involved in Fe-S coordination (C99 and C101, but not C35 in . *E.coli*). A similar model for SufA (IscA paralog) was obtained in [141] (PDB ID: 2D2A). Their structure, which comprises the C-terminal tail containing the cysteines showed a similar arrangement to the model obtained in [139]. An alternative structural modeled was obtained for the IscA of *T. elongatus* by Morimoto et al. (PDB ID: 1X0G, [142]). In this structure, the C-terminal was fully resolved and displayed a correct coordination of the Fe-S cluster. This model is also composed of a tetrameric arrangement of IscA, and is formed by a dimer of asymmetric IscA dimers (Fig.6.4B). Intriguingly, two IscA monomers in this model display a domain swap, in which a long anti-parallel beta-sheet is formed between two strands belonging to different monomers. In this arrangement, the cluster is coordinated by the simultaneous interaction with the three cysteines of one monomer and additionally with C103 of another IscA monomer.

We tested whether DCA could discriminate between these different oligomerization patterns and applied the same protocol as for the analysis of the IscU family (see section 6.2). The IscA MSA was defined over $N = 106$ positions and contained $B = 29233$ sequences. Comparing the DCA predictions with the three discussed structures showed that while the overall prediction quality is very high (85% precision over the top N predictions, compared to the 1X0G structure), the main region containing the differences between the models involved interactions between the C-terminal tail with residues located at positions 30-40. We observed that these predictions could not be explained by any intra-molecular contacts and therefore extended our analysis taking in account oligomeric contacts (Fig.6.4C,D,E). While neither of the inter-molecular arrangements observed in the structures of *E.coli* IscA or SufA were compatible with these predictions (Fig.6.4C,D), we observed that the domain swapped tetrameric arrangement present in 1X0G was in excellent agreement with these contacts. Although surprising at first sight, the existence of such domain-swapped dimers has been observed in several different occasions [143]. In contrast, what would be surprising is if these two bacterial orthologs would have drastically different oligomeric structures. Coevolutionary analysis indicates that the inter-chain contacts can be explained by a domain-swapped arrangement of IscA, which is further compatible with cluster coordination. Whether an alternative arrangement lacking the domain swap but still respecting these restraints could be built remains an open question.

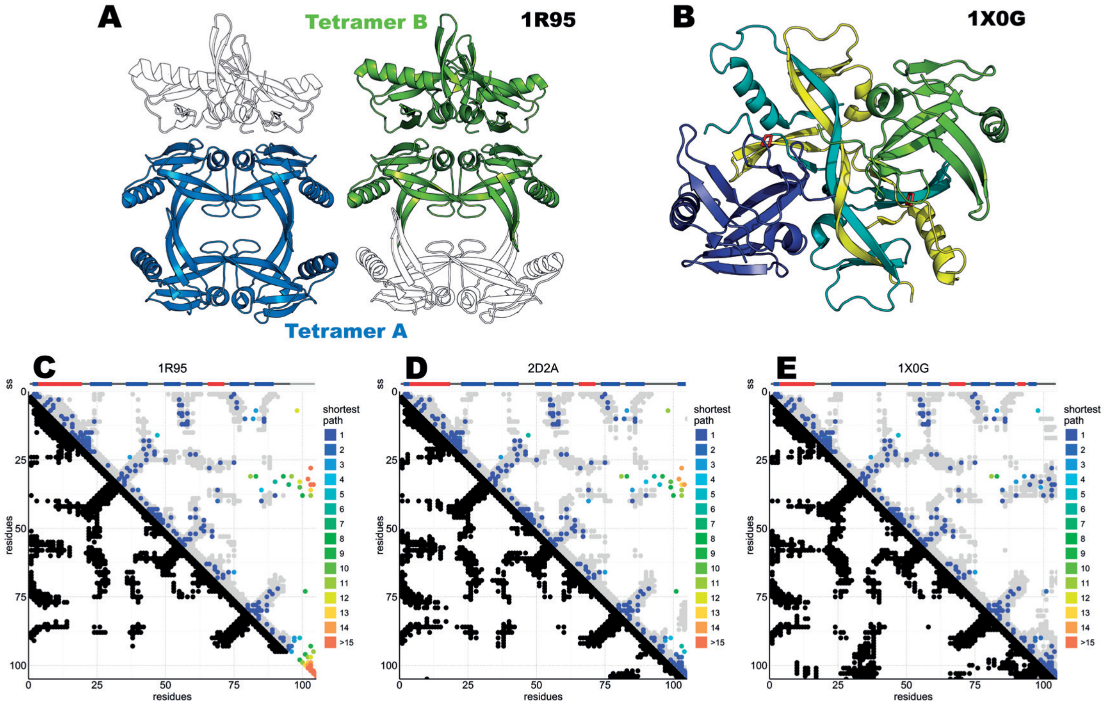

Figure 6.4 – IscA multimerization. **A** The two tetrameric arrangements proposed in [139] (PDB ID: 1R95). The two proposed tetramers are depicted in blue and green. The two complexes are aligned on the central (identical) dimer. The dimers outlined in black depict the positioning of the second dimer in the alternative tetramer. **B** Domain swapped tetramer as proposed in [142] (PDB ID: 1X0G). The four IscA monomers are colored differently. The coordinated Fe-S clusters are shown in red. **C-D** Comparison of the DCA predictions with the contacts maps of the three discussed models, including inter-protein contacts. The same graphical elements as in Fig.6.3A are depicted. Figure adapted from [133].

6.4 Interactions between Fe-S cluster proteins

The assembly of Fe-S clusters is based on the precise interactions between the different proteins of the *Isc* pathway. Among the different binary interactions involved in this pathway, the only available high-resolution structural model is a crystal structure of the complex formed by the desulfurase IscS and the scaffold protein IscU [135]. One of the major unanswered questions regarding these interactions is the role played by the CyaA protein, which is thought to be an iron carrier [46]. A low resolution model, obtain by small angle X-ray scattering, suggests that the binding site of CyaY is distinct of the binding site of IscU on IscS [144]. We thus focused our attention on the interactions of CyaY with both IscU and IscS.

The same sequence extraction as presented in the previous sequences resulted in an MSA of the CyaY family containing 3459 sequences (defined over 109 positions), which was significantly less than what we retrieved for both the IscU and IscA family ($\sim 11'000 - 13'000$ sequences). This was due to the presence of multiple paralogs of both IscU and IscA involved in different pathways (i.e. SufA, SufU, NifU), whereas proteins of the CyaY family are predominantly only present in the *Isc* pathway. In contrast, we retrieved 34632 sequences belonging to the IscS family (defined over 402 positions), highlighting the multiplication of the number of cysteine desulfurases present in a multitude of pathways.

While most of the proteins involved in the *Isc* pathway are contained in the Isc operon in bacteria [73], CyaY is located distantly in the genome. Furthermore, our sequence dataset contained an appreciable number of sequences belonging to the eukaryotic homologs of the Isc proteins. We could therefore not use the genomic-location based matching strategy to pair putatively interacting sequence pairs [106, 105]. As no particular cross-talk between paralogs belonging to different pathways had been reported for the Fe-S assembly proteins, we assumed that the interactions should be highly specific. This led us to favor the use of coevolutionary optimization techniques over the use of random matching to perform the paralog matching (see Chapter 5). In fact, while the random matching strategy gives comparable results on highly promiscuous interactions, in the case of protein-protein interactions presenting high specificity, optimization techniques, such as IPA [108] or PPM [109] lead to significantly better DCA predictions [68]. We therefore used the IPA methodology to investigate the interactions of interest, as briefly outlined below (see [108] for a detailed description of the methodology) .

The Iterative Paralog Algorithm (IPA) is based on iteratively constructing a matched MSA between proteins of two interacting families in the presence of multiple paralogs per organism. At each iteration, DCA is performed on the current MSA, and the resulting Potts-Model is used to score all possible paralog matchings for all organisms. A gap score is then defined for each pair of potentially interacting sequence in an organism, which is based on the absolute energy (lower energies indicate higher coevolutionary coupling between the matched sequences) and taking in account an energy gap, which favors matched sequences having a large energy difference with the second best matched paralog. At each iteration, a fixed number of best scored protein-pairs is then selected and forms the MSA of the next iteration. The number of selected sequences is increased at each iteration until all sequences have been matched. The first MSA is generated by randomly matching pairs of sequences belonging to the same organism. Note that in this scheme, each sequence is only matched once, resulting in an exclusive matching scheme. This fact and the gap scoring used in IPA highlights the design of the method to tackle highly specific interactions.

For each pair of protein families, we performed a given number N_{IPA} of IPA matching simulations using different initial random matchings. For each optimized MSA, we then performed DCA predictions using the asymmetric PLM version. We employed the same inter-protein contact selection criterion as discussed in Chapter 5, namely selecting inter-protein contacts with $\tilde{S}_{ij} > 0.8$, where

$$\tilde{S}_{ij} = \frac{S_{ij}}{\left|\min(S_{ij}^{Inter})\right|\left(\sqrt{1+\sqrt{\frac{N}{N_{Eff}}}})\right)} \tag{6.1}$$

and the symbols of equation 6.1 have been defined in Chapter 4. We then ranked the contacts according to the fraction of times they were selected in the N_{IPA} DCAs (i.e. the acceptance frequencies) and selected the highest ranked contacts for further analysis.

In order to validate the approach presented above, we verified that this procedure correctly predicted inter-protein contacts in the IscU-IscS interaction for which a structural model was available. We could here take advantage of the experimental knowledge of the binding interface to calibrate our selection threshold. The four contacts having the highest acceptance frequency (>85%) lied in the correct interface (Fig.6.5A,B), whereas the fifth contact ranked according to the acceptance frequency was a false positive. We thus selected a stringent threshold of 85% on the acceptance frequency to select robust contacts.

Having fixed a selection criterion for inter-protein contacts, we first analyzed the CyaY-IscS interaction (Fig.6.5C,D). The very low number of sequences retrieved for the CyaY family severely limited the quality of the predictions. Indeed, no inter-protein contact was selected in more than 85% of the DCAs, and thus no significant coevolutionary signal between CyaY and IscS could be recovered according to our selection threshold. We however noted the presence of two contacts having acceptance frequencies of 68% which were significantly more frequent than the lower ranked contacts. Their mapping onto the IscS and CyaY structures revealed that they involved surface exposed residues, and their locations could in principle allow a geometrically acceptable docking. Thus, although they had low frequencies, these contacts pairs could be indicative of an underlying evolutionary origin. However, the low significance level precludes the prediction of a biological relevance of this observed interface.

In contrast, the analysis of the CyaA-IscU interaction revealed three significant DCA predictions (acceptance frequency >94%, Fig.6.5E,F). Interestingly, these three contacts formed a relatively well defined interaction interface between CyaY and IscU. Furthermore, this interface did not overlap with the interface of IscU which binds to IscS (Fig.6.5A). However, comparison of this interface with the low resolution SAXS model of [144] showed significant differences in the binding interfaces. It must be noted that the model presented in [144] is based on SAXS data obtained for a IscS-IscU-CyaY trimer and it is currently not fully understood whether this trimeric arrangement is a biologically active conformation. Hence, our results might indicate an alternative CyaY-IscU binding mode when interacting as dimers.

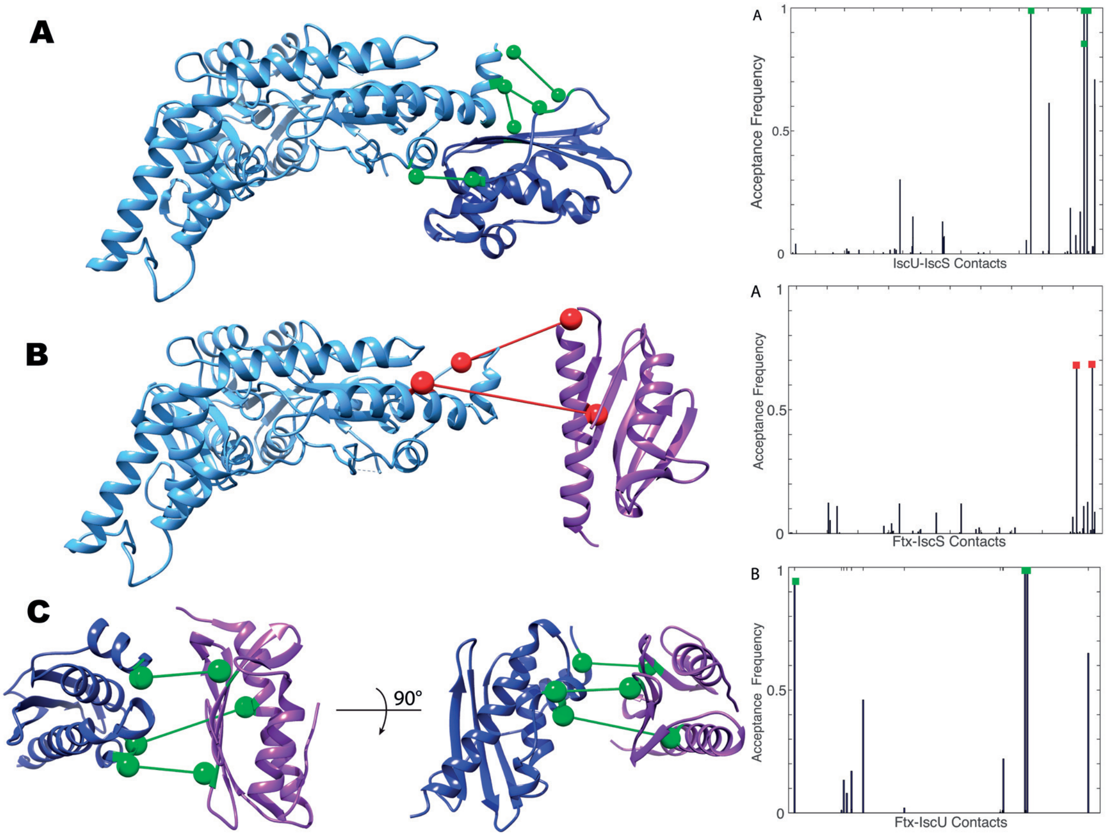

Figure 6.5 – Inter-protein interactions of Fe-S cluster pathway proteins. Structural Views (left panels) depict the highest ranked inter-protein contacts in ball-and-stick representation (spheres centered on the Cβ atoms, Cα for glycines). Colors of the contacts depict their robustness (see text): Red: < 85%. Green: >85%. The positioning of the monomers was manually arranged for ease of visualization and not subject to any docking algorithm. The right panels show the acceptance frequencies of all inter-protein residue pairs. The horizontal axes are an arbitrary indexing, mapping the inter-protein contacts to natural numbers. **A,B** IscU (dark blue) - IscS (light blue). **C,D** CyaY (purple) - IscS (light blue). **E,F** IscU (dark blue) - CyaY (purple). Figure adapted from [133].

6.5 Conclusions

The analysis performed on proteins involved in the *Isc* pathway is illustrative of the power of coevolutionary analysis, and specifically DCA, to investigate structural and functional features at multiple scales [145]. Indeed, our results range from the structural state of a single monomer (the N-terminal state of IscU), up to oligomeric complexes, both at the homo- (IscU and IscA) and hetero-dimeric (CyaY -IscU/IscS) levels. This ability of DCA to fruitfully analyze features at different organizational scales based upon a unique underlying model is sufficiently rare in the field of bioinformatics to be appreciated. The downside of this versatility is the high requirement in terms of number of sequences, as illustrated by the case of CyaY. However, the astonishing growth rate of sequence databases forecasts a bright future, as independently from the methodological improvements that will occur in the field of coevolutionary analysis, the increase of available protein sequences guarantees as steady increase in prediction quality of DCA.

We here focused on the core of the Isc machinery, namely the two scaffolding proteins, the cysteine desulfurase IscS and the iron donor/regulator CyaY. A central question touching upon all proteins involved in the core of the machinery is their ability to correctly coordinate and transfer Fe-S clusters. Interestingly, this coordination appears to be often performed by oligomeric arrangements. This structural complexity is underlined by the number of incompatible structural models available for protein involved in the *Isc* pathway. Much remains to be done to fully understand Fe-S coordination from a structural point of view. Structural analysis based on residue coevolution is a promising candidate to complement structural studies with orthogonal informations.

In this work, we neglected several additional proteins involved in the Fe-s cluster assembly, among which the HscA-HscB chaperone/cochaperone pair and the ferrodoxin Fdx. In particular, the role played by the specialized HscA and HscB chaperones remains to be elucidated. While at the moment, the raw number of sequences of these proteins is too small to be exploited by DCA, this situation will change in the (near) future. Multiple interesting questions will arise. It appears that in contrast to most Hsp70 members, HscA has as unique client substrate IscU. Whether this specialization implies structural rearrangements of the chaperone and/or variations on the allosteric cycle are open questions that will be addressed in future work. In a broader context, it will be interesting to analyze whether features of specific specialized sub-families in large protein families are detected by DCA performed on the whole family. The detection and extraction of such sub-family specific features at the contact level is an exciting and challenging task. The HscA sub-family of Hsp70 chaperones will potentially be an interesting candidate for such finer analyses.

6.6 Supplementary Figures

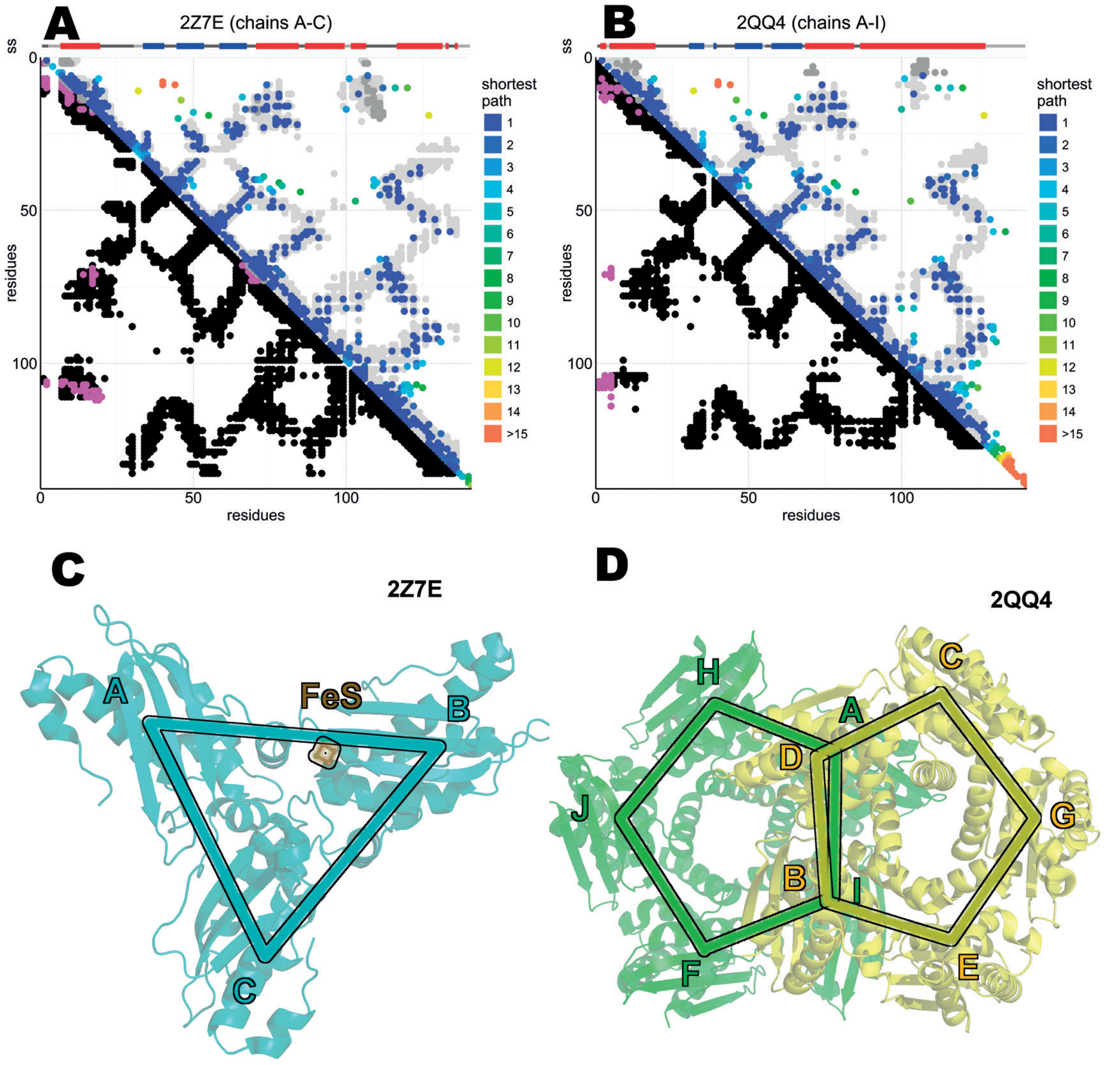

Figure 6.6 – IscU Tetramer and Decamer arrangements. **A,B** Upper triangular par: DCA predictions colored by their SP scores. The structural contacts are depicted in grey background. Lower triangular part: Structural contacts. Intra-chain contacts are depicted in black, inter-chain contacts in magenta. Contacts are defined by a minimal inter-residue distance of 8.5 Å between heavy atoms. **A** IscU trimer (PDB ID: 2Z7E). **B** IscU Decamer (PDB ID:2QQ4). **C** Structural view of the trimeric IscU arrangement. **D** Structural view of the IscU decameric arrangement. Figure adapted from [133].

7 Conclusions and Outlooks

In this thesis, four different subjects were studied by coevolutionary analysis. While the first three subjects dealt with the in-depth analysis of the Hsp70 chaperone machinery, the last chapter was dedicated to the study of a different system, namely the Fe-S cluster assembly pathway. The treatment of these different subjects lead to both novel biological insights into the systems under study and concomitantly identified several key questions raised in the field of coevolutionary analysis.

The DCA analysis of the Hsp70 family revealed the biological relevance of its dimeric arrangement, which was experimentally verified in parallel work [89, 90]. Beyond this biological finding, our analysis revealed the ability of DCA to predict multiple protein conformations separated by large-scale rearrangements, which result in two sets of mutually exclusive contacts. These results highlighted the encoding of such conformations in sequence variation. Whereas we here took advantage of the structural knowledge of the two conformations of Hsp70, future work should focus on the identification of multiple conformations in predicted contact maps. Furthermore, we introduced a taxonomic reweighting scheme which forms a first step towards the rational study of sub-family specific contact signatures. In this regards, our simple reweighting scheme calls for future analysis, by extending the reweighting to multiple sub-families and automatically detecting the most varying contacts.

The logical next step in the coevolutionary analysis of the Hsp70 machinery was to investigate the interaction between the chaperone and its co-chaperones, as materialized by the Hsp40-Hsp70 interactions. The fundamental challenge presented by this study was the pairing of interacting sequences with little a-priori knowledge of the interaction network. This led us to propose the random matching strategy, which allowed the successful characterization of an evolutionarily conserved interface. By combining coevolutionary analysis with molecular simulations at both coarse-grained and atomistic levels, we could propose a novel model for the elusive Hsp40-Hsp70 complex. These results highlight the complementarity of statistical analysis of proteins sequences with molecular simulation techniques. In addition to the DCA based structural prediction, our matching strategy further revealed intriguing differences in the coevolutionary signals encoded in bacterial and eukaryotic sub-families, which call for future investigation.

Investigating the synergistic action between class A and B J-proteins turned out to be to most challenging project from a DCA point of view. The low number of sequences available, combined with the high degree of promiscuity in this interaction rendered the interpretation of the DCA results difficult. Despite these difficulties, our results fully support the experimental model of an inter-class direct interaction between the J-domain and the CTD of class A/B J-proteins. Our phylogenetic analysis contributed a taxonomic wide view to the question, in complement to detailed data provided by wet bench experiments on a reduced set of model organisms. This point crystalizes the strong interplay between these different approaches, which allowed to propose a coherent model of the interaction, strengthening our knowledge of the disaggregation machinery used by eukaryotes.

Straying away from the analysis of Hsp70 chaperones, our analysis of Iron-Cluster assembly proteins illustrated the power of DCA to investigate structural features at different scales. The complexity of the machinery involved in this pathway is illustrated by the shear number of different structural models available in literature. In this respect, we used sequence analysis to select models from a pool of experimental candidates. A deeper structural understanding of the mechanisms by which Fe-S clusters are assembled and transferred will require an understanding of the dynamical properties of this process. In this aim, further computational and experimental studies will be needed to fully grasp the complexity of this pathway.

Beyond the biological insights which we gained through these different projects, two main questions of general nature emerged during this thesis.

In all four projects, a recurrently encountered theme was the presence and characterization of sub-families in large protein families. Specifically, DCA is generally performed on protein families composed of multiple sub-classes, which can either have purely phylogenetic or functional origin. A central question arising is how can the sub-family specific features be identified and extracted from family wide analysis? These questions were repeatedly encountered in the study of the Hsp70 machinery, in which we investigated the differences emerging from the differentiation between prokaryotes and eukaryotes. While we used different techniques to investigate these differences, a systematic study of this question is necessary to answer both methodological and fundamental questions. It is a-priori not clear how the presence of these different sub-families is reflected in the parameters of the inferred Potts models, namely how are sub-families of varying sizes weighted in the final DCA scores. In addition, the identification of sub-family specific contacts in the overall DCA predictions would be of great interest for functional studies. In this regard, the sub-family analysis could be viewed as a natural extension of the study of specificity determining positions (SPD) for which a large literature exists. Future theoretical and numerical studies will be needed to pursue this line of work.

The second recurring theme was the problem of matching interacting paralogs, which we encountered in the last three chapters. This challenge presents two related aspects: From a practical point of view, the identification of interacting paralogs is needed to perform high quality contact predictions through coevolutionary methods. Additionally, from a systems biology point of view, the knowledge of the interaction network in the presence of multiple paralogs is a basic requirement to understand the cellular organization of complex biochemical processes. While we could satisfy the requirement of the first point through the introduction of the random matching strategy, more advanced recently developed optimization algorithms simultaneously tackled the two aspects [108, 109]. An important point that emerged from our analysis is the fact that all these methods only work marginally for complex eukaryotic systems. Whereas bacterial organisms, which posses simpler cellular organization, seem to strongly encode the interaction specificity in their amino-acid sequences, the matching results on eukaryotic systems did not show such strong specificity. Specifically, it is known for the Hsp40-Hsp70 interactions and for the synergistic class A/B J-protein interactions that the interaction network is highly promiscuous, encompassing a high-degree of cross-talk. This leads to a fundamental question regarding such interaction networks, particularly in the case of eukaryotic organisms: How can a sequence interacting with a large number of partners accommodate compensatory mutations? In a basic scenario, a deleterious mutation would need to be compensated by a large number of complementary mutations on all the interaction partners. The statistical low likelihood of such a scenario would imply a dramatic decrease in observed mutation rates for proteins interacting with multiple partners. Higher organisms however posses multiple means by which protein-protein interactions can be modulated, beyond the physics encoded directly into their amino-acid sequences. Differential temporal and tissue expression levels, as well as cell compartmentalization are possible ways to fine-tune the interaction network of protein-protein interactions. Future work is required to build a comprehensive analysis tool encompassing all these aspects in order to better understand the specificity and selectivity determinants of protein interactions in higher eukaryotes.

As a final remark, this thesis was characteristic of an inter-disciplinary approach to study protein coevolution, which encompasses models emerged from statistical physics and the analysis of specific biological systems. This dichotomy required the study of fields which have traditionally drastically different ways of approaching problems. In such a scenario, major challenges are to marry the theoretical rigor of theoretical physics with more qualitative approaches inherent to the study of extremely complex systems such as living organisms. It is our hope that the material and approaches presented in this thesis will raise interest in audiences from both fields.

Bibliography

[1] Bialek W et al. (2012) Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences of the United States of America* 109(13):4786–4791.

[2] Cavagna A, Giardina I (2014) Bird Flocks as Condensed Matter. *Annual Review of Condensed Matter Physics* 5:183–207.

[3] Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440(7087):1007–1012.

[4] Zhigalov, Alexander. Arnulfo, Gabriele. Nobili, Lino. Palva, Satu. Palva JM (2017) Modular co-organization of functional connectivity and scale-free dynamics in the human brain. *Network Neuroscience* 1(3):275–301.

[5] Nguyen HC, Zecchina R, Berg J (2017) Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics* 66(3):197–261.

[6] Altschuh D, Lesk A, Bloomer A, Klug A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology* 193(4):693–707.

[7] Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated Mutations and Residue Contacts in Proteins. *Proteins: Structure, Function, and Genetics* 18(4):309–317.

[8] Shindyalov I, Kolchanov N, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering* 7(3):349–358.

[9] Bateman A et al. (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Research* 45(D1):D158–D169.

[10] Tillier ER, Lui TW (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 19(6):750–755.

[11] Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21(22):4116–4124.

[12] Gloor GB, Martin LC, Wahl LM, Dunn SD (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 44:7156–7165.

[13] Lapedes AS, Giraud BG, Liu L, Stormo GD (1999) Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. *Proceedings of the IMS/ AMS International Conference on Statistics in Molecular Biology and Genetics: Monograph Series of the Inst. for Mathematical Statistics, Hayward* 33:236–256.

[14] Weigt M, White RA, Szurmant H, Hoch Ja, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America* 106(1):67–72.

[15] Morcos F et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* 108(49):E1293–301.

[16] Jaynes E (1957) Information theory and statistical mechanics. *Physical review* 106(4):620–630.

[17] Wu FY (1982) The Potts model. *Review of Modern Physics* 54(1):235–268.

[18] Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E* 87(1):012707 1–16.

[19] Feinauer C, Skwark MJ, Pagnani A, Aurell E (2014) Improving Contact Prediction along Three Dimensions. *PLoS computational biology* 10(10):e1003847.

[20] Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24(3):333–340.

[21] Coucke A et al. (2016) Direct coevolutionary couplings reflect biophysical residue interactions in proteins. *Journal of Chemical Physics* 145(17).

[22] Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M (2011) Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. *PloS one* 6(5):e19729.

[23] Ekeberg M, Hartonen T, Aurell E (2014) Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics* 276:341–356.

[24] Sutto L, Marsili S, Valencia A, Gervasio FL (2015) From residue coevolution to protein conformational ensembles and functional dynamics. *Proceedings of the National Academy of Sciences* p. 201508584.

[25] Pelizzola A (2005) Cluster variation method in statistical physics and probabilistic graphical models. *Journal of Physics A: Mathematical and General* 38(33):R309–R339.

[26] Kappen HJ, Ortiz FDBR (1998) Boltzmann Machine Learning Using Mean Field Theory and Linear Response Correction. *Advances in Neural Information Processing Systems 10* pp. 280–286.

[27] Barton JP, Cocco S, De Leonardis E, Monasson R (2014) Large pseudocounts and L2 -norm penalties are necessary for the mean-field inference of Ising and Potts models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 90(1):1–19.

[28] Newman M, Barkema G (1999) *Monte Carlo Methods in Statistical Physics.* (Clarendon Press, Oxford).

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

[30] Jones DT, Buchan DWa, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)* 28(2):184–90.

[31] Stein RR, Marks DS, Sander C (2015) Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Computational Biology* 11(7).

[32] Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Molecular systems biology* 4(165):165.

[33] Burger L, Van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Computational Biology* 6(1).

[34] Richter K, Haslbeck M, Buchner J (2010) The heat shock response: life on the verge of death. *Molecular cell* 40(2):253–66.

[35] Saibil H (2013) Chaperone machines for protein folding, unfolding and disaggregation. *Nature Reviews Molecular Cell Biology* 14(10):630–642.

[36] Finka A, Mattoo RU, Goloubinoff P (2016) Experimental Milestones in the Discovery of Molecular Chaperones as Polypeptide Unfolding Enzymes. *Annual Review of Biochemistry* 85(1):715–742.

[37] Kim YE, Hipp MS, Bracher A, Hayer-Hartl M, Hartl FU (2013) Molecular chaperone functions in protein folding and proteostasis. *Annual review of biochemistry* 82:323–55.

[38] Genest O, Hoskins JR, Kravats AN, Doyle SM, Wickner S (2015) Hsp70 and Hsp90 of E. coli Directly Interact for Collaboration in Protein Remodeling. *Journal of Molecular Biology* pp. 1–13.

[39] Żwirowski S et al. (2017) Hsp70 displaces small heat shock proteins from aggregates to initiate protein refolding. *The EMBO Journal* p. e201593378.

[40] Finka A, Goloubinoff P (2013) Proteomic data from human cell cultures refine mechanisms of chaperone-mediated protein homeostasis. *Cell Stress and Chaperones* 18(5):591–605.

[41] Diamant S, Peres Ben-Zvi A, Bukau B, Goloubinoff P (2000) Size-dependent disaggregation of stable protein aggregates by the DnaK chaperone machinery. *Journal of Biological Chemistry* 275(28):21107–21113.

[42] Sharma SK, De los Rios P, Christen P, Lustig A, Goloubinoff P (2010) The kinetic parameters and energy cost of the Hsp70 chaperone as a polypeptide unfoldase. *Nature Chemical Biology* 6(12):914–20.

[43] Albanèse V, Reissmann S, Frydman J (2010) A ribosome-anchored chaperone network that facilitates eukaryotic ribosome biogenesis. *The Journal of cell biology* 189(1):69–81.

[44] Goloubinoff P, De Los Rios P (2007) The mechanism of Hsp70 chaperones: (entropic) pulling the models together. *Trends in biochemical sciences* 32(8):372–80.

[45] Rothnie A, Clarke AR, Kuzmic P, Cameron A, Smith CJ (2011) A sequential mechanism for clathrin cage disassembly by 70-kDa heat-shock cognate protein (Hsc70) and auxilin. *Proceedings of the National Academy of Sciences* 108(17):6927–6932.

[46] Stemmler TL, Lesuisse E, Pain D, Dancis A (2010) Frataxin and mitochondrial FeS cluster biogenesis. *Journal of Biological Chemistry* 285(35):26737–26743.

[47] Rüdiger S, Buchberged A, Bukau B (1997) Interaction of Hsp70 chaperones with substrates. *Nature structural biology* 4(5):342–349.

[48] Munson M et al. (1996) What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. *Protein Science* 5(8):1584–1593.

[49] Kellner R et al. (2014) Single-molecule spectroscopy reveals chaperone-mediated expansion of substrate protein. *Proceedings of the National Academy of Sciences of the United States of America* 111(37):13355–13360.

[50] De Los Rios P, Ben-Zvi A, Slutsky O, Azem A, Goloubinoff P (2006) Hsp70 chaperones accelerate protein translocation and the unfolding of stable protein aggregates by entropic pulling. *Proceedings of the National Academy of Sciences of the United States of America* 103(16):6166–71.

[51] De Los Rios P, Goloubinoff P (2016) Hsp70 chaperones use ATP to remodel native protein oligomers and stable aggregates by entropic pulling. *Nature Structural & Molecular Biology* 23(9):766–769.

[52] Sousa R et al. (2016) Clathrin-coat disassembly illuminates the mechanisms of Hsp70 force generation. *Nature Structural and Molecular Biology* 23(9):821–829.

[53] Mayer MP (2010) Gymnastics of molecular chaperones. *Molecular cell* 39(3):321–31.

[54] Kityk R, Kopp J, Sinning I, Mayer MP (2012) Structure and dynamics of the ATP-bound open conformation of Hsp70 chaperones. *Molecular Cell* 48(6):863–74.

[55] Qi R et al. (2013) Allosteric opening of the polypeptide-binding site when an Hsp70 binds ATP. *Nature Structural & Molecular Biology* 20(7):900–7.

[56] Bertelsen EB, Chang L, Gestwicki JE, Zuiderweg ERP (2009) Solution conformation of wild-type E. coli Hsp70 (DnaK) chaperone complexed with ADP and substrate. *Proceedings of the National Academy of Sciences of the United States of America* 106(21):8471–6.

[57] Mapa K et al. (2010) The conformational dynamics of the mitochondrial Hsp70 chaperone. *Molecular Cell* 38(1):89–100.

[58] De Los Rios P, Barducci A (2014) Hsp70 chaperones are non-equilibrium machines that achieve ultra-affinity by energy consumption. *eLife* 3:e02218.

[59] Barducci A, De Los Rios P (2015) Non-equilibrium conformational dynamics in the function of molecular chaperones. *Current Opinion in Structural Biology* 30:161–169.

[60] Goloubinoff P, Sassi AS, Fauvet B, Barducci A, De Los Rios P (2017) Molecular chaperones inject energy from ATP hydrolysis into the non-equilibrium stabilisation of native proteins. *bioRxiv* 146852.

[61] Russell R, Karzai AW, Mehl AF, McMacken R (1999) DnaJ dramatically stimulates ATP hydrolysis by DnaK: Insight into targeting of Hsp70 proteins to polypeptide substrates. *Biochemistry* 38(13):4165–4176.

[62] Gässler CS et al. (1998) Mutations in the DnaK chaperone affecting interaction with the DnaJ cochaperone. *Biochemistry* 95(December):15229–15234.

[63] Laufen T et al. (1999) Mechanism of regulation of hsp70 chaperones by DnaJ cochaperones. *Proceedings of the National Academy of Sciences of the United States of America* 96(May):5452–5457.

[64] Kampinga H, Craig E (2010) The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nature Reviews. Molecular Cell Biology* 11(8):579–92.

[65] Suh WC et al. (1998) Interaction of the Hsp70 molecular chaperone, DnaK, with its cochaperone DnaJ. *Proceedings of the National Academy of Sciences of the United States of America* 95(26):15223–8.

[66] Jiang J et al. (2007) Structural basis of J cochaperone binding and regulation of Hsp70. *Molecular Cell* 28(3):422–33.

[67] Ahmad A et al. (2011) Heat shock protein 70 kDa chaperone / DnaJ cochaperone complex employs an unusual dynamic interface. *Proceedings of the National Academy of Sciences of the United States of America* 108(47):18966–18971.

[68] Malinverni D, Jost Lopez A, De Los Rios P, Hummer G, Barducci A (2017) Modeling Hsp70/Hsp40 interaction by multi-scale molecular simulations and co-evolutionary sequence analysis. *eLife* 6:e23471.

[69] Kampinga HH et al. (2009) Guidelines for the nomenclature of the human heat shock proteins. *Cell stress & chaperones* 14(1):105–11.

[70] Szabo A, Korszun R, Hartl FU, Flanagan J (1996) A zinc finger-like domain of the molecular chaperone DnaJ is involved in binding to denatured protein substrates. *The EMBO journal* 15(2):408–417.

[71] Rüdiger S, Schneider-Mergener J, Bukau B (2001) Its substrate specificity characterizes the DnaJ co-chaperone as a scanning factor for the DnaK chaperone. *The EMBO journal* 20(5):1042–1050.

[72] Mayer MP, Bukau B (2005) Hsp70 chaperones: cellular functions and molecular mechanism. *Cellular and molecular life sciences : CMLS* 62(6):670–84.

[73] Bandyopadhyay S, Chandramouli K, Johnson MK (2008) Iron-sulfur cluster biosynthesis. *Biochemical Society transactions* 36(Pt 6):1112–1119.

[74] Liberek K, Marszalek J, Ang D, Georgopoulos C, Zylicz M (1991) Escherichia coli DnaJ and GrpE heat shock proteins jointly stimulate ATPase activity of DnaK. *Proceedings of the National Academy of Sciences* 88(7):2874–2878.

[75] Kabani M, Beckerich JM, Brodsky JL (2002) Nucleotide exchange factor for the yeast Hsp70 molecular chaperone Ssa1p. *Molecular and cellular biology* 22(13):4677–89.

[76] Bracher A, Verghese J (2015) The nucleotide exchange factors of Hsp70 molecular chaperones. *Frontiers in Molecular Biosciences* 2(April):1–9.

[77] Dragovic Z, Broadley Sa, Shomura Y, Bracher A, Hartl FU (2006) Molecular chaperones of the Hsp110 family act as nucleotide exchange factors of Hsp70s. *The EMBO journal* 25(11):2519–28.

[78] Mattoo RUH, Sharma SK, Priya S, Finka A, Goloubinoff P (2013) Hsp110 is a bona fide chaperone using ATP to unfold stable misfolded polypeptides and reciprocally collaborate with Hsp70 to solubilize protein aggregates. *The Journal of biological chemistry* 288(29):21399–411.

[79] Li J, Bioucas-dias JM, Plaza A, Member S (2013) Spectral – Spatial Classification of Hyperspectral Data Using Loopy Belief Propagation and Active Learning. *IEEE Transactions on Geoscience and Remote Sensing* 51(2):844–856.

[80] Malinverni D, Marsili S, Barducci A, De Los Rios P (2015) Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. *PLoS Computational Biology* 11(6):e1004262.

[81] Finn RD et al. (2010) The Pfam protein families database. *Nucleic acids research* 38(Database issue):D211–22.

[82] Katoh K, Misawa K, Kuma Ki, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30(14):3059–3066.

[83] Finn RD, Clements J, Eddy SR (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research* 39(SUPPL. 2):29–37.

[84] Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D (2017) Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences* 114(34):9122–9127.

[85] Polier S, Dragovic Z, Hartl FU, Bracher A (2008) Structural basis for the cooperation of Hsp70 and Hsp110 chaperones in protein folding. *Cell* 133(6):1068–79.

[86] Casari G, Sander C, Valencia A (1995) A method to predict functional residues in proteins. *Nature structural biology* 2(2):171–178.

[87] Deloche O, Kelley WL, Georgopoulos C, Deloche O, Kelley WL (1997) Structure-function analyses of the Ssc1p , mitochondrial proteins in Escherichia coli . Structure-Function Analyses of the Ssc1p , Mdj1p , and Mge1p Saccharomyces cerevisiae Mitochondrial Proteins in Escherichia coli. *Journal of Bacteriology* 179(19):6066–6075.

[88] Lu B, Garrido N, Spelbrink JN, Suzuki CK (2006) Tid1 isoforms are mitochondrial DnaJ-like chaperones with unique carboxyl termini that determine cytosolic fate. *Journal of Biological Chemistry* 281(19):13150–13158.

[89] Sarbeng EB et al. (2015) A functional DnaK dimer is essential for the efficient interaction with Heat Shock Protein 40 kDa (Hsp40). *Journal of Biological Chemistry* 290(14):jbc.M114.596288.

[90] Liu Q et al. (2016) A disulfide-bonded DnaK dimer is maintained in an ATP-bound state. *Cell Stress and Chaperones* 22(2):201–212.

[91] Marcion G et al. (2015) C-terminal amino acids are essential for human heat shock protein 70 dimerization. *Cell Stress and Chaperones* 20(1):61–72.

[92] Jana B, Morcos F, Onuchic JN (2014) From structure to function: the convergence of structure based models and co-evolutionary information. *Physical chemistry chemical physics : PCCP* 16(14):6496–507.

[93] Maniloff J (1996) The minimal cell genome: "on being the right size". *Proceedings of the National Academy of Sciences of the United States of America* 93(September):10004–10006.

[94] Moran Na (2002) Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* 108:583–586.

[95] Harrison CJ (1997) Crystal Structure of the Nucleotide Exchange Factor GrpE Bound to the ATPase Domain of the Molecular Chaperone DnaK. *Science* 276(5311):431–435.

[96] Karzai AW, McMacken R (1996) A bipartite signaling mechanism involved in DnaJ-mediated activation of the Escherichia coli DnaK protein. *Journal of Biological Chemistry* 271(19):11236–11246.

[97] Nakatsukasa K, Huyer G, Michaelis S, Brodsky JL (2008) Dissecting the ER-Associated Degradation of a Misfolded Polytopic Membrane Protein. *Cell* 132(1):101–112.

[98] Hundley Ha, Walter W, Bairstow S, Craig Ea (2005) Human Mpp11 J protein: ribosome-tethered molecular chaperones are ubiquitous. *Science (New York, N.Y.)* 308(5724):1032–4.

[99] Otto H et al. (2005) The chaperones MPP11 and Hsp70L1 form the mammalian ribosome-associated complex. *Proceedings of the National Academy of Sciences of the United States of America* 102(29):10064–9.

[100] Tsai J, Douglas MG (1996) A conserved HPD sequence of the J-domain is necessary for YDJ1 stimulation of Hsp70 ATPase activity at a site distinct from substrate binding. *Journal of Biological Chemistry* 271(16):9347–9354.

[101] Greene MK, Maskos K, Landry SJ (1998) Role of the J-domain in the cooperation of Hsp40 with Hsp70. *Proceedings of the National Academy of Sciences of the United States of America* 95(11):6108–13.

[102] Mayer MP, Laufen T, Paal K, McCarty JS, Bukau B (1999) Investigation of the interaction between DnaK and DnaJ by surface plasmon resonance spectroscopy. *Journal of Molecular Biology* 289(4):1131–44.

[103] Sousa R et al. (2012) Evaluation of competing J domain:Hsp70 complex models in light of existing mutational and NMR data. *Proceedings of the National Academy of Sciences of the United States of America* 109(13):E734; author reply E735.

[104] Zuiderweg ERP, Ahmad A (2012) Reply to Sousa et al.: Evaluation of competing J domain:Hsp70 complex models in light of methods used. *Proceedings of the National Academy of Sciences of the United States of America* 109(13):E735–E735.

[105] Hopf TA et al. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3:e03430.

[106] Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* 3:e02030.

[107] Feinauer C, Szurmant H, Weigt M, Pagnani A (2016) Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. *Plos One* 11(2):e0149166.

[108] Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS (2016) Inferring interaction partners from protein sequences. *bioRxiv* p. 050732.

[109] Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A (2016) Simultaneous identification of specifically interacting paralogs and inter-protein contacts by Direct-Coupling Analysis. *Proceedings of the National Academy of Sciences* 113(43):12186–12191.

[110] Pellecchia M, Szyperski T, Wall D, Georgopoulos C, Wüthrich K (1996) NMR structure of the J-domain and the Gly/Phe-rich region of the Escherichia coli DnaJ chaperone. *Journal of Molecular Biology* 260(2):236–250.

[111] Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences of the United States of America* 106(52):22124–9.

[112] Kim YC, Hummer G (2008) Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *Journal of Molecular Biology* 375(5):1416–33.

[113] Kim YC, Tang C, Clore GM, Hummer G (2008) Replica exchange simulations of transient encounter complexes in protein-protein association. *Proceedings of the National Academy of Sciences of the United States of America* 105(35):12855–60.

[114] Miyazawa S, Jernigan RL (1996) Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *Journal of Molecular Biology* 256(3):623–644.

[115] Daura X et al. (1999) Peptide folding: When simulation meets experiment. *Angewandte Chemie (International ed. in English)* 38(1-2):236–240.

[116] Gray JJ et al. (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology* 331(1):281–299.

[117] Chaudhury S et al. (2011) Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS ONE* 6(8).

[118] Abraham MJ et al. (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2:19–25.

[119] Case D et al. (2014) Amber 14. *University of California, San Francisco.*

[120] Genheden S, Ryde U (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery* 10(5):449–461.

[121] Garimella R et al. (2006) Hsc70 contacts helix III of the J domain from polyomavirus T antigens: addressing a dilemma in the chaperone hypothesis of how they release E2F from pRb. *Biochemistry* 45(22):6917–29.

[122] Alderson TR et al. (2014) The specialized Hsp70 (HscA) interdomain linker binds to its nucleotide-binding domain and stimulates ATP hydrolysis in both cis and trans configurations. *Biochemistry* 53(46):7148–59.

[123] Vogel M, Mayer MP, Bukau B (2006) Allosteric regulation of Hsp70 chaperones involves a conserved interdomain linker. *Journal of Biological Chemistry* 281(50):38705–38711.

[124] Zhuravleva A, Clerico EM, Gierasch LM (2012) An interdomain energetic tug-of-war creates the allosterically active state in Hsp70 molecular chaperones. *Cell* 151(6):1296–307.

[125] Nillegoda NB et al. (2017) Evolution of an intricate J-protein network driving protein disaggregation in eukaryotes. *eLife* 6:e24560.

[126] Brent Irvine G, El-Agnaf O, Shankar GM, Walsh DM (2008) Protein Aggregation in the Brain: The Molecular Basis for Alzheimer's and Parkinson's Diseases. *Molecular Medicine* 14(7-8):1.

[127] Liberek K, Lewandowska A, Ziętkiewicz S (2008) Chaperones in control of protein disaggregation. *The EMBO Journal* 27(2):328–335.

[128] Mogk A, Kummer E, Bukau B (2015) Cooperation of Hsp70 and Hsp100 chaperone machines in protein disaggregation. *Frontiers in Molecular Biosciences* 2(May):1–10.

[129] Nillegoda NB et al. (2015) Crucial HSP70 co-chaperone complex unlocks metazoan protein disaggregation. *Nature* 524(7564):247–251.

[130] Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

[131] Murtagh F, Contreras P (2012) Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2(1):86–97.

[132] Granell C, Sergio G, Arenas A (2011) Data clustering using community detection algorithms. *International Journal of Complex Systems in Science* 1:21–24.

[133] Fantini M, Malinverni D, De Los Rios P, Pastore A (2017) New Techniques for Ancient Proteins: Direct Coupling Analysis Applied on Proteins involved in Iron Sulfur Cluster Biogenesis. *bioRxiv* 4(June):1–14.

[134] Schilke B, Voisine C, Beinert H, Craig E (1999) Evidence for a conserved system for iron metabolism in the mitochondria of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* 96(18):10206–11.

[135] Shi R et al. (2010) Structural basis for Fe-S cluster assembly and tRNA thiolation mediated by IscS protein-protein interactions. *PLoS Biology* 8(4).

[136] Ramelot TA et al. (2004) Solution NMR structure of the iron-sulfur cluster assembly protein U (IscU) with zinc bound at the active site. *Journal of Molecular Biology* 344(2):567–583.

[137] Toth-Petroczy A et al. (2016) Structured States of Disordered Proteins from Genomic Sequences. *Cell* 167(1):158–170.e12.

[138] Adrover M, Howes BD, Iannuzzi C, Smulevich G, Pastore A (2015) Anatomy of an iron-sulfur cluster scaffold protein: Understanding the determinants of [2Fe-2S] cluster stability on IscU. *Biochimica et Biophysica Acta - Molecular Cell Research* 1853(6):1448–1456.

[139] Bilder PW, Ding H, Newcomer ME (2004) Crystal Structure of the Ancient Fe-S Scaffold IscA Reveals a Novel Protein Fold. *Biochemistry* 43(1):133–139.

[140] Cupp-Vickery JR, Silberg JJ, Ta DT, Vickery LE (2004) Crystal structure of IscA, an iron-sulfur cluster assembly protein from Escherichia coli. *Journal of Molecular Biology* 338(1):127–137.

[141] Wada K et al. (2005) Crystal structure of Escherichia coli SufA involved in biosynthesis of iron–sulfur clusters: Implications for a functional dimer. *FEBS Letters* 579(29):6543–6548.

[142] Morimoto K et al. (2006) The Asymmetric IscA Homodimer with an Exposed [2Fe-2S] Cluster Suggests the Structural Basis of the Fe-S Cluster Biosynthetic Scaffold. *Journal of Molecular Biology* 360(1):117–132.

[143] Liu Y, Eisenberg D (2002) 3D domain swapping: As domains continue to swap. *Protein Science* 11(6):1285–1299.

[144] Prischi F et al. (2010) Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly. *Nature communications* 1(7):95.

[145] Szurmant H, Weigt M (2017) Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Current Opinion in Structural Biology* 50:26–32.

Duccio **Malinverni**

MSc in Physics, Minor in Computational science

Born 10.12.1988. Nationality: Swiss
Chemin de Pierrefleur 62, 1004 Lausanne, Switzerland
☎ (+41) 79 706 30 88 | ✉ duccio.malinverni@gmail.com

Education

École Polytechnique Fédérale de Lausanne
Lausanne, CH

PhD in Theoretical Biophysics — *2014-2018*

- Thesis title: ”Statistical Analysis of Protein Sequences: A Coevolutionary Study of Molecular Chaperones”

École Polytechnique Fédérale de Lausanne
Lausanne, CH

MSc degree in Physics, Minor degree in Computational science. — *2011-2013*

- Major: Statistical physics and polymer physics. Numerical simulations and Monte Carlo methods
- Minor: Computational Science

École Polytechnique Fédérale de Lausanne
Lausanne, CH

BSc degree in Physics. — *2007-2010*

- Major: Numerical simulations of physical systems, Plasma physics

Gymnase Denis-de-Rougemont.
Neuchâtel, CH

Swiss Academic Baccalaureate (with honors) — *2004-2007*

- Major: Physics and Applied Mathematics

Professional Experiences

Laboratory of Statistical Biophysics, École Polytechnique Fédérale de Lausanne
Lausanne, CH

Doctoral Research Assistant — *Feb. 2014-Mar. 2018*

- Coevolutionary analysis of protein families (Direct-Coupling Analysis, PCA).
- Implementation of high-performance software for protein structure prediction and toolbox for coevolutionary data analysis.
- Coarse-grained molecular docking simulations.
- Large-scale atomistic simulations of macromolecular complexes.
- Biological data mining, processing and curation from databases. Dimensionality reduction of biological data.
- Phylogenetic analysis of protein families. Phylogenetic tree construction. Phylogenetically discriminant positions determination.

Computational Structural Biology Group, Forschungszerntrum Jülich
Jülich, DE

IHRS BioSoft Guest Student — *Jul. - Oct. 2014*

- Guest student in the International Helmholtz Research School of Biophysics and Soft Matter.
- Implementation of a computational approach to fitting low-resolution experimental cryo-EM density maps.

University of Neuchâtel, Center for Hydrogeology and geothermal science
Neuchâtel, CH

Research assistant (Compulsory Swiss Civil Service) — *Mar. - Aug. 2011 & Aug 2012*

- Setup of a hydrogeology laboratory.
- Inverse modeling of heterogenous media in subsurface flows.
- Systematic evaluation of a simulation method based on multiple-points statistics for heterogenous media simulations.

Swiss Institute of Speleology and Karstology (ISSKA)
La Chaux-de-Fonds, CH

Scientific assistant (Compulsory Swiss Civil Service) — *Sep. 2010- Mar. 2011*

- IT support for the research staff: scripting, programming and automation for geological and hydrogeological data analysis.
- Collaboration on the national research project PNR 61, hydrological data analysis and cartography.

Skills

Programming	Fluent use of Python, C/C++, MatLab, LaTeX. Basic knowledge of R, Fortran, Basic
Scientific Software	Gromacs, LAMMPS, HMMER, HHBLITS, BLAST, MAFFT, Chimera, RaxML, Fasttree
Computational Methods	Statistical Data Analysis, Probabilistic Inference, Monte-Carlo Simulations, Molecular Dynamics
Languages	Italian, English, French, German

Research Interests

Protein-Protein Interactions	• Molecular understanding of PPIs • Specificity determinants • Interaction networks of multiple paralogs
Statistical Analysis of Biological Data	• Computational approaches to protein family characterization • Sub-family structures of large protein families • Machine Learning approaches: Feature selection/extraction for PPI • Algorithms for biological data mining

Publications

- **Malinverni D**, Jost Lopez A,De Los Rios P, Hummer G, Barducci A, 2017, Modeling Hsp70/Hsp40 interaction by multi-scale molecular simulations and co-evolutionary sequence analysis, *eLife*,6(e23471)
- Nillegoda NB, Stank A, **Malinverni D**, Alberts N, Szlachcic A, Barducci A, De Los Rios P, Wade RC, Bukau B, 2017, Evolution of an intricate J-protein network driving protein disaggregation in eukaryotes, *eLife*, 6(e24560)
- Fantini M, **Malinverni D**, De Los Rios P, Pastore A, 2017, New Techniques for Ancient Proteins: Direct Coupling Analysis Applied on Proteins involved in Iron Sulfur Cluster Biogenesis, *Frontiers in Molecular Biosciences*,40(40)
- **Malinverni D**, Marsili S, Barducci A, De Los Rios P, 2015, Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones, *PLoS Comp. Biol.*,11(6)
- Straubhaar J, **Malinverni D**, Addressing conditioning data in multiple-point statistics simulation algorithms based on a multiple grid approach, *Mathematical Geosciences*,46(2)
- Fasoli A, Avino F, Bovet A, Furno I, Gustafson K, Jolliet S, Loizu J, **Malinverni D**, Ricci P, Riva F, Theiler C, Spolaore M, Vianello N, 2013, Basic investigations of electrostatic turbulence and its interaction with plasma and suprathermal ions in a simple magnetized toroidal plasma, *Nulcear Fusion*, 53(6)

Presentations

Basel Computational Biology Conference BC^2 *Basel, Switzerland*

Oral presentation *Sep. 2017*

- Title: Coevolutionary Analysis of the Hsp70 chaperone machinery

Structural Biology Super Group Meeting *Lausanne, Switzerland*

Invited oral presentation *May, 2017*

- Title: Coevolutionary Insights into the Hsp70 Chaperone System

First Biology for Physics Conference: Is there New Physics in Living Matter? *Barcelona, Spain*

Poster presentation *Jan. 2017*

- Title: Coevolutionary Analysis of the Hsp70 chaperone machinery

Presentation at the Centre de Biochimie Structurale de Montpellier *Montpellier, France*

Invited oral presentation *Jun. 2016*

- Title: Insights into Protein Structure and Function from Sequence Coevolution

Workshop on Coevolution in Proteins and RNA, Theory and Experiments *Cargese, France*

Poster presentation *Apr. 2016*

- Title: Coevolutionary Insights into Molecular Chaperone Systems

EPFL Bioengineering Day *Lausanne, Switzerland*

Poster presentation *Nov. 2015*

- Title: Evolutionary Information Predicts Multi-Scale Allostery and Dimerization of Hsp70 chaperones

EMBO Conference: Molecular chaperones: From molecules to cells and misfolding diseases *Heraklion, Grece*

Poster presentation *May. 2015*

- Title: Residue Coevolution Predicts Functional Dimerization of the Hsp70 chaperones

Awards

Nov. 2015 **2015 BioE Best Poster Award**, Bio Engineering Day, EPFL *Lausanne, CH*

Teaching Experiences

Analytical Mechanics *EPFL*

Teaching Assistant *2015, 2016, 2017*

- 1 semester BsC course for physicists, covering the fundamentals of Analytical Mechanics.

Statistical Physics of Biomacromolecules *EPFL*

Teaching Assistant *2015, 2016*

- 1 semester MsC course for physicists, covering the statistical physics of polymers.

Advanced Statistical Physics *EPFL*

Teaching Assistant *2016*

- 1 semester MsC course for physicists, covering phase transitions in statistical physics.

Introduction to Classical Mechanics *EPFL*

Teaching Assistant *2017*

- 1 semester BsC course for engineers, covering the fundamentals of Newtonian Mechanics.

Academic duties

Association des Élèves Électriciens EPFL *Lausanne, CH*

EPFL Students association comity member *2008-2010*

- Event management and coordination.

Coaching Association EPFL *Lausanne, CH*

Coach and Super-Coach *2008-2009*

- Welcoming and chaperoning of the first year students at EPFL.

References

Prof. Paolo De Los Rios	École Polytechnique Fédérale de Lausanne, Institute of Physics, +41 21 69 30510, **paolo.delosrios@epfl.ch**
Dr. Alessandro Barducci	Centre de Biochimie Structurale de Montpellier, CNRS, +33 467 4179 11, **barducci@cbs.cnrs.fr**