PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones

Duccio Malinverni[1], Simone Marsili[2], Alessandro Barducci[1]*, Paolo De Los Rios[1]*

1 Laboratoire de Biophysique Statistique, École Polytechnique Fédérale de Lausanne, Faculté de Sciences de Base, Lausanne, Switzerland, **2** Structural Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

* alessandro.barducci@epfl.ch (AB); paolo.delosrios@epfl.ch (PDLR)

CrossMark
click for updates

OPEN ACCESS

Citation: 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 Comput Biol 11(6): e1004262. doi:10.1371/journal.pcbi.1004262

Editor: Marco Punta, Pierre and Marie Curie University (UPMC), FRANCE

Received: December 11, 2014

Accepted: April 1, 2015

Published: June 5, 2015

Copyright: © 2015 Malinverni et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: DM thanks the Swiss National Science Foundation (http://www.snf.ch/) for financial support under grant 2012_149278. AB thanks the Swiss National Science Foundation (http://www.snf.ch/) for financial support under grant PZ00P2_136856. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Hsp70s are a class of ubiquitous and highly conserved molecular chaperones playing a central role in the regulation of proteostasis in the cell. Hsp70s assist a myriad of cellular processes by binding unfolded or misfolded substrates during a complex biochemical cycle involving large-scale structural rearrangements. Here we show that an analysis of coevolution at the residue level fully captures the characteristic large-scale conformational transitions of this protein family, and predicts an evolutionary conserved–and thus functional–homo-dimeric arrangement. Furthermore, we highlight that the features encoding the Hsp70 dimer are more conserved in bacterial than in eukaryotic sequences, suggesting that the known Hsp70/Hsp110 hetero-dimer is a eukaryotic specialization built on a pre-existing template.

Author Summary

Molecular chaperones are a class of proteins that are crucial for the correct functioning of cells. They play central housekeeping roles in the normal cell cycle, and are major actors of the protection system of the cell against cell stress conditions. In this study, we apply statistical inference methods to analyse the structure and function of the Hsp70 molecular chaperone, one of the main members of chaperones. We use the correlated amino acid co-evolutions in protein sequences to identify directly interacting amino acids. Our results show that coevolutions capture an appreciable fraction of native contacts throughout the protein. Furthermore, amino acid coevolution predicts previously hypothesized functional dimer interactions between Hsp70s, thus giving a theoretical contribution to this debate.

Introduction

Molecular chaperones are a broad class of proteins that protect cells against the potentially deleterious effects of denatured and unfolded proteins. They have been shown to play an essential

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

role in multiple proteostasis pathways [1,2]. The 70-kDa heat shock proteins (Hsp70s) are highly conserved and ubiquitous chaperones present in virtually all organisms [3–5]. Besides the canonical roles of chaperones under stressful conditions, Hsp70s have been identified playing several housekeeping roles in the cell under normal conditions such as assisted folding [6,7], oligomeric complex assembly [8], cell cycle regulation [9], import of unfolded polypeptides in the mitochondria and endoplasmic reticulum [10,11], as well as ubiquitin-mediated protein degradation [12,13] and prion propagation [14].

These tasks are all supported by the ability of Hsp70s to bind substrate proteins through an ATP-consuming, non-equilibrium biochemical cycle [15] that involves several conformational transitions at different scales due to nucleotide and substrate binding. The Hsp70 cycle is further regulated by the cooperative action of co-chaperones: J-domain proteins (JDPs) strongly stimulate ATP-hydrolysis [16], whereas nucleotide exchange factors (NEFs) catalyse the release of ADP [17,18].

Hsp70s are composed of two domains, connected by a flexible linker (Fig 1). The N-terminal ATPase nucleotide-binding domain (NBD), is composed of four lobes, and hosts the active site where ATP and ADP molecules bind. The C-terminal substrate-binding domain (SBD) is subdivided into a β-sandwich subdomain and an α-helical lid; it binds exposed hydrophobic stretches of target substrates in non-native conformations [19–22]. When Hsp70s are bound to ATP, the α-lid and the β-sandwich of the SBD dock onto opposite lobes of the NBD. Upon ATP hydrolysis, the NBD undergoes an intra-domain allosteric transformation resulting into a slight rotation of the lobes with respect to each other [23–25]. Concomitantly, a larger-scale inter-domain allosteric change takes place, whereby the two SBD subdomains undock from the NBD and bind to each other, clamping any substrate that was bound to the β-sandwich.

Beyond these functional conformational rearrangements, oligomerization has been reported for several members of the Hsp70 family [26–29]. Recently, Hsp70 oligomers have been observed by means of electron microscopy [30] and mass spectroscopy [31]. Notably, the results of the latter study indicated that the intermolecular interaction between the linker and the SBD might be responsible for the assembly. Unfortunately neither the functional relevance of these oligomeric states nor the structural details of the quaternary arrangements have yet been clarified.

In contrast, the interaction of Hsp70s in eukaryotes with members of the related Hsp110 family has been well characterized functionally as well as structurally. Hsp110s have been

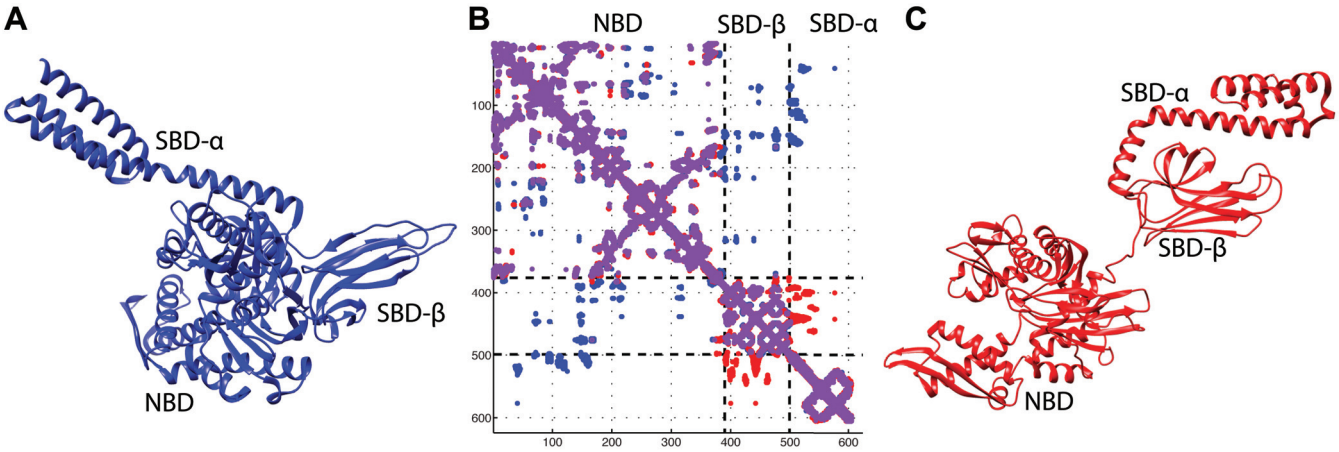

Fig 1. ATP/ADP states of DnaK and the associated contact map. Dashed lines in the map represent the limits of the domains. **A)** Crystal structure of DnaK in ATP state (PDB ID 4jne). **B)** Contact map of the union of both structures. In purple are contacts present in both structures, in red contacts only in ADP, in blue contacts only in ATP. All contacts refer to a threshold of 8.5 Å. **C)** Crystal structure of DnaK in the ADP state (PDB ID 2kho).

doi:10.1371/journal.pcbi.1004262.g001

shown to act as a NEFs in the Hsp70 cycle [32,33] and later studies have indicated that in the Hsp70/110 dimer, the two proteins act both as bona-fide chaperones and as mutual NEFs [34]. Indeed, when Hsp70s are bound to ADP, the binding of Hsp110 induces a slight opening of the lobes forming the NBD of Hsp70, thus leading to a facilitated release of the nucleotide [8]. The crystal structure of this complex has been determined [8,35], revealing that the two chaperones associate through both NBD/NBD and NBD/SBD contacts. Intriguingly, similar inter-molecular arrangements have been observed in crystals of DnaK, an *E. coli* bacterial Hsp70 [21,22] although their functional relevance has not been determined.

The structure and function of proteins is encoded in their amino-acid sequence, which is constantly under the combined evolutionary action of random mutations and selection. Consequently, it has to be expected that a careful analysis of the Hsp70 sequences across the whole family should reveal the presence of residue pairs that coevolve, *i.e.* exhibit correlated mutations, because they are close to each other in the three-dimensional structure and must therefore coordinate their physical properties [36–41]. Based on these premises, the recent Direct Coupling Analysis (DCA) [42–45] has emerged as the most effective algorithm to exploit residue coevolution for the prediction of structural contacts [46,47]. This technique not only can reliably predict the protein native structure [45,48–52], but it can also predict the presence of multiple conformers of the same proteins [44,48,53], homo-multimerization [44,48] and protein-protein interactions [54,55]. DCA, together with similar techniques, has been recently applied to the Hsp70 family by General et al. [56], to study the key residues involved in the allosteric signal propagation in Hsp70 chaperones.

Here we take advantage of the ubiquitous nature of Hsp70s and of the rapid growth of available sequenced proteomes to extensively characterize the function-determining structural features at multiple scales by means of DCA. The structures of the NBD and of the SBD clearly emerge from our analysis, as well as their intra- and inter-domain rearrangements during the chaperone functional cycle. Even more strikingly, DCA predictions show that inter-molecular contacts consistent with crystallographic structures are significantly conserved, thus pointing to their functional relevance.

Results

To investigate residue coevolution in the Hsp70 family, we built a multiple sequence alignment (MSA) containing 3708 sequences, defining 624 residue positions (see Material and Methods). Our MSA covers almost equally bacterial and eukaryotic sequences (1562 Eukaryotes, 1982 Bacteria). DCA was performed on the MSA using the Pseudo-Likelihood method ([52,57], see Material and Methods) and predicted contacts were ranked according to their DCA scores, which denote coevolution strength between pairs of residues. After discarding contacts at sequences separation less than five, mostly related only to local secondary structures, we retained the 624 DCA contacts with the highest scores (corresponding to 0.325% of the total 191890 possible contacts).

DCA predictions can be compared with the structural information available for *E. coli* DnaK, for which high-resolution structures are available for both ATP- (PDB ID 4jne [22] and 4b9q [21]) and ADP-bound states (PDB ID 2kho [20]). Regarding the former state, we use the higher resolution structure (PDB ID 4jne, see S1 Fig for analogous comparison with 4b9q and S1 Text for comments on the supplementary material). The RMSD between the two structures is ~2Å, computed on 597 CA atom pairs, see S8 Fig). Comparisons with partial structures are reported in S3 Fig. We identified 8054 inter-residue native structural contacts (See Material and Methods) in the ATP-bound structure and 7758 in the ADP-bound structure. A close scrutiny of the ATP- and ADP-bound native contact maps of DnaK emphasizes differences

between the two nucleotide-bound conformers, which appear as sets of intra- and inter-domain contacts (Fig 1B). In particular, the contacts associated with the ATP-bound structure docking of the lid on the NBD are mutually exclusive with those characterizing the clamping of the α-lid subdomain on the β-sandwich of the SBD in the ADP-bound structure.

When comparing DCA predictions with the contact maps of the two nucleotide-bound states, we observe a high number of true positives (TPs), *i.e.* correctly predicted contacts (see S2 Fig for TP rates). Indeed, 502 out of 624 predicted contacts correspond to native structural contacts in the ATP-bound state, and 504 in the ADP-bound state (Fig 2), resulting into a TP rate of about 80%. Importantly, all the characteristic structural elements shared by both structures, such as the triple-bundle forming the α-lid subdomain, the β-sandwich and the four-lobe structure of the NBD, are well predicted by DCA. To provide a more refined appraisal of the

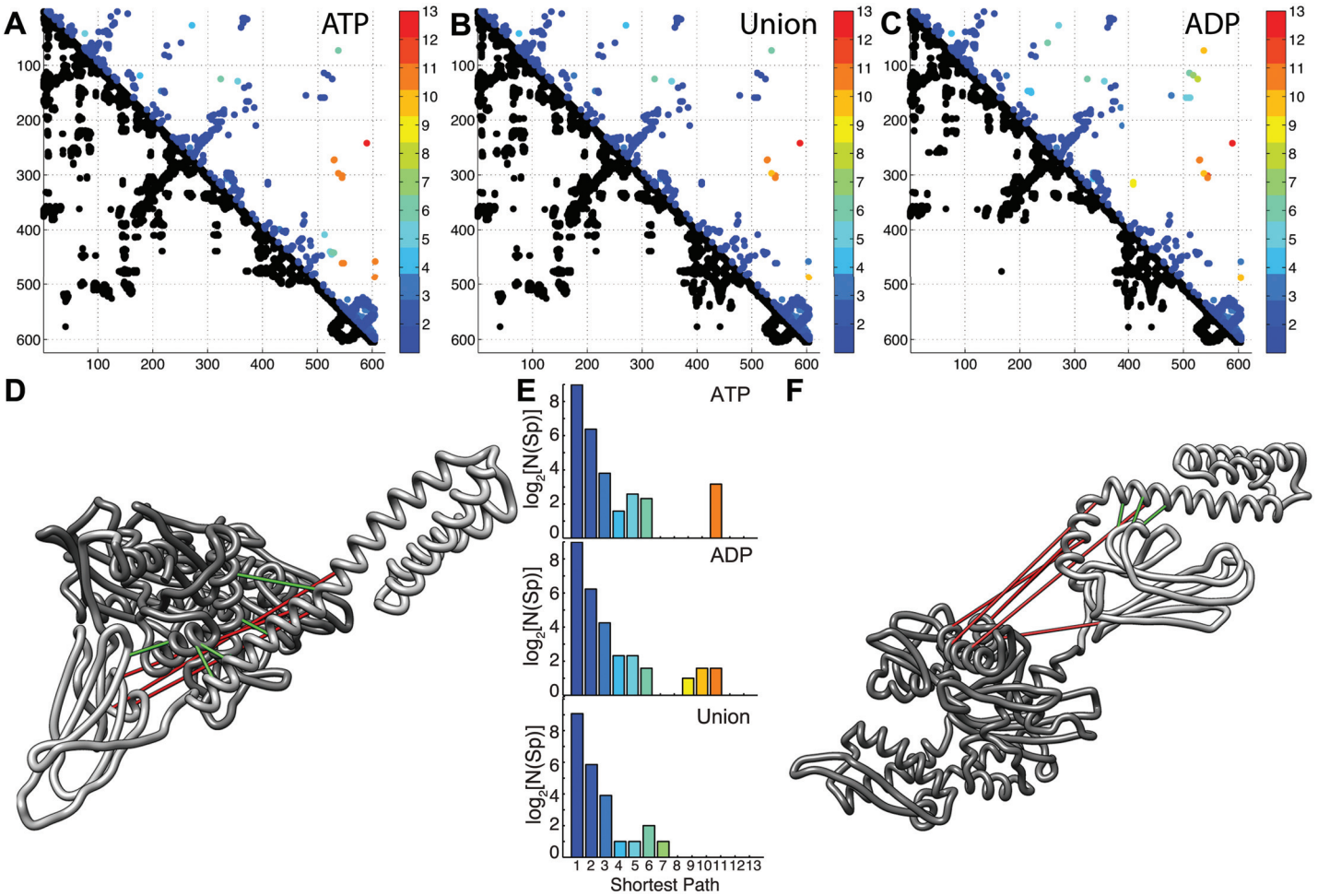

Fig 2. DCA predictions of DnaK allostery. In the contact maps (A-C), the lower triangular parts are the structure contacts at threshold 8.5 Å, the upper parts contain the DCA predictions, coloured by shortest paths. In D and F are shown 8 strongly allosteric contacts. In green are contacts that are true in the conformation, in red contacts that are false in the conformation. The correct contacts in ATP are false in ADP, and vice-versa: The three false positives in the ATP state (red lines) are true positives in the ADP state (green lines). Conversely, the 5 true positives in the ATP state are false positives in the ADP state. **A)** Contact map of DnaK ATP (PDB ID 4jne). **B)** Contact map of the union ATP/ADP. The shortest paths are taken as the minimum between the corresponding shortest paths in the two states. **C)** Contact map of DnaK ADP (PDB ID 2kho). **D)** Set of 8 strong allosteric contacts in the ATP state. 5 correct contacts (green lines), 3 false contacts (red lines) **E)** Histograms of shortest paths of the predicted DCA contacts. Each histogram refers to the corresponding contact map. Counts are reported in log-scale. Bins are coloured corresponding to the colour scheme of the contact maps. **F)** Set of 8 strong allosteric contacts in the ATP state. 3 correct contacts (green lines), 5 false contacts (red lines).

doi:10.1371/journal.pcbi.1004262.g002

DCA results, each predicted contact is associated with the length of the shortest path (SP) between the corresponding residues, computed over the contact map of the crystallographic structure (see Material and Methods). In this context, the SP provides a topological measure of the distance between two residues that further characterizes our prediction. As shown by Burger and van Nimwegen [58] in the context of mutual information coevolutionary networks, the shortest paths efficiently capture the mediation of coevolutions along chains of residues. Indeed, we observe that about half of the predictions not directly compatible with structural contacts have SP = 2, thus involving two residues that have a native contact with the same amino acid. We expect many of these seemingly wrong predictions to be correct, due to the definition of native contacts that depends on an arbitrary threshold (here 8.5Å) and neglects structural fluctuations. Including all the predictions at SP = 2 in the TPs, the TP rate increases to about 93% in the ADP-bound structure and 94% in the ATP-bound structure. We therefore estimate the actual TP rate to be between 80% and 90%.

The comparison of DCA results with the individual contact maps corresponding to the ADP- and ATP-bound states highlights in both cases a small set of significantly incompatible predictions (SP$\geq$6). However since DCA analysis is expected to capture contacts present in all the functionally relevant conformers [48,59], the most appropriate strategy is to compare DCA predictions with a contact map corresponding to the union of those relative to single-states. In this case, the number of TPs grows to 538 (86%) (95% when taking into account contacts with SP = 2) and the number of significantly incompatible predictions (SP$\geq$6) is decreased. This behaviour indicates that DCA predicts those set of contacts that are found exclusively either in the ATP- or in the ADP-bound structures. In the following, we refer to such contacts as allosteric contacts. As seen in Fig 2A–2C, these contacts correspond to the docking of the α-lid on the NBD in the ATP-bound structure and the clamping of the β-basket by the α-lid in the ADP-bound structure. Furthermore, it must be noted that the predicted allosteric contacts appear early in the score ranking (S1 Table), indicating their evolutionary relevance. These results confirm that DCA applied to the Hsp70 family is able to capture the large-scale allosteric transition of this chaperone that involves variations of inter-residue distances larger than 50 Å.

A more accurate inspection of the of X-ray structure of DnaK in ATP-bound state (PDB ID 4jne) reveals that the small subset of coevolving pairs of residues not compatible with the monomeric structures (SP>9, see Fig 2B) accurately corresponds to the inter-monomeric contacts in the crystallographic arrangement (Fig 3). We identify six predicted inter-monomeric contacts among the top 624 predictions (S2 Table) which can be separated in two groups: One set of four contacts is associated to the docking of the SBD α-lid of one monomer onto lobe II of the NBD of the other monomer (Fig 3D). The second set is composed of two NBD-NBD contacts (Fig 3C). Among these first six predicted DCA contacts, four display clear electrostatic interactions. We can evaluate the probability that the observed allosteric contacts are the result of random errors by calculating the corresponding p-value (see Material and Methods). The resulting value of 1.44×10^{-4} is a clear hint that the dimeric arrangement observed in DnaK crystals is evolutionary conserved in the Hsp70 family, thus suggesting a functional role.

As discussed in the introduction, Hsp110s are eukaryotic remote homologues of Hsp70s that have retained high sequence similarity [60] and are known to form functional hetero-dimers with Hsp70s with a dimerization pattern extremely similar to that observed in DnaK crystals. It could be argued that the dimer-compatible DCA predictions are due to the presence of interacting eukaryotic Hsp70s and Hsp110s in our MSA. Several arguments can be brought forward to discard this possibility. First and foremost, the Pseudo-Likelihood algorithm used here does not consider couplings between residues belonging to two different sequences in the MSA by construction (see Materials and Methods and [52]), thus not predicting Hsp70/110 dimers. Furthermore, to exclude the possibility that the observed dimerization pattern is a

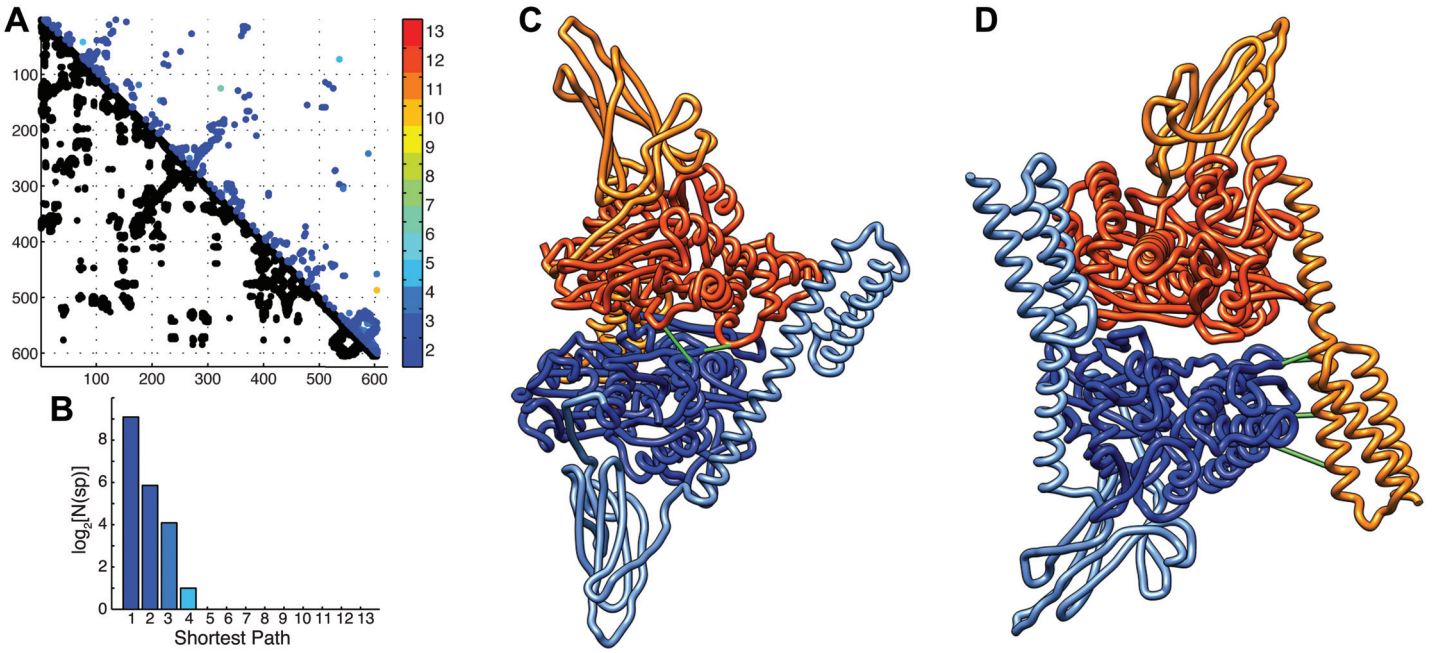

Fig 3. Homo-dimeric complex of *E. coli* DnaK. A) DCA contact map: The lower triangular part is the structure contacts at threshold 8.5 Å, the upper part contains the DCA predictions, coloured by shortest paths. **B)** Histogram of shortest paths of the predicted DCA contacts. **C-D)** The six dimeric contacts predicted by DCA, illustrated on the ATP state dimer (PDB ID 4jne). Each monomer is coloured by domains, with the Nucleotide Binding domain in darker shade and the Substrate Binding Domain in lighter tones. In C, highlight of the NBD-NBD docking contacts, in D, highlight of the SBD-NBD docking contacts.

doi:10.1371/journal.pcbi.1004262.g003

consequence of the presence of Hsp110 sequences, we performed a more stringent filtering of our MSA. To this aim we limit our MSA only to sequences explicitly tagged in Uniprot by the canonical Hsp70 gene names *hspa1a*, *hspa1b*, *hsp70*, *ssa1* and *DnaK*, resulting in a subset containing 1781 sequences. DCA performed on this reduced set (see S5 Fig) resulted in an overall higher noise level, due to lower statistics. However, all the six originally predicted dimeric contacts were retained in the reduced set and an additional dimer-compatible contact appeared in the top 624 predictions. We can therefore safely conclude that coevolutionary analysis predicts Hsp70 homo-dimerization with a quaternary arrangement similar to that observed in the Hsp70-Hsp110 complex (S4 Fig).

We further investigate if this feature of the Hsp70 family is equally present in the different domains of life. To this aim, we performed DCA artificially varying the relative weights of sequences belonging to eukaryotes and prokaryotes in the MSA. Following this approach, we measured the relative strength of the dimeric contacts as the ratio between their average DCA score and that of the original 624 predicted contacts and we report in Fig 4A this quantity as a function of the weight of eukaryotic sequences in the MSA. The dependence of the TP rate on the same quantity is shown in Fig 4B in order to check if sequence reweighting perturbs the overall quality of the structural predictions. We observe that the relative strength of the dimeric contacts decreases as the eukaryotic weight increases, thus suggesting that Hsp70 homo-dimerization has bacterial origin. This behaviour is observed in a range of relative weights (W_E in 0.3–0.7) where the overall quality of prediction is globally unaffected. Moving away from this region, the relative weights are too unbalanced in either direction resulting into poorer statistics and less reliable predictions. The limiting cases of $W_E = 0$ and $W_E = 1$, corresponding to resp. keeping only bacterial or eukaryotic sequences in the MSA, result into high noise levels (see S6 Fig), as the effective number of sequences is strongly decreased. However, we stress that for $W_E = 0$ (only bacterial sequences)

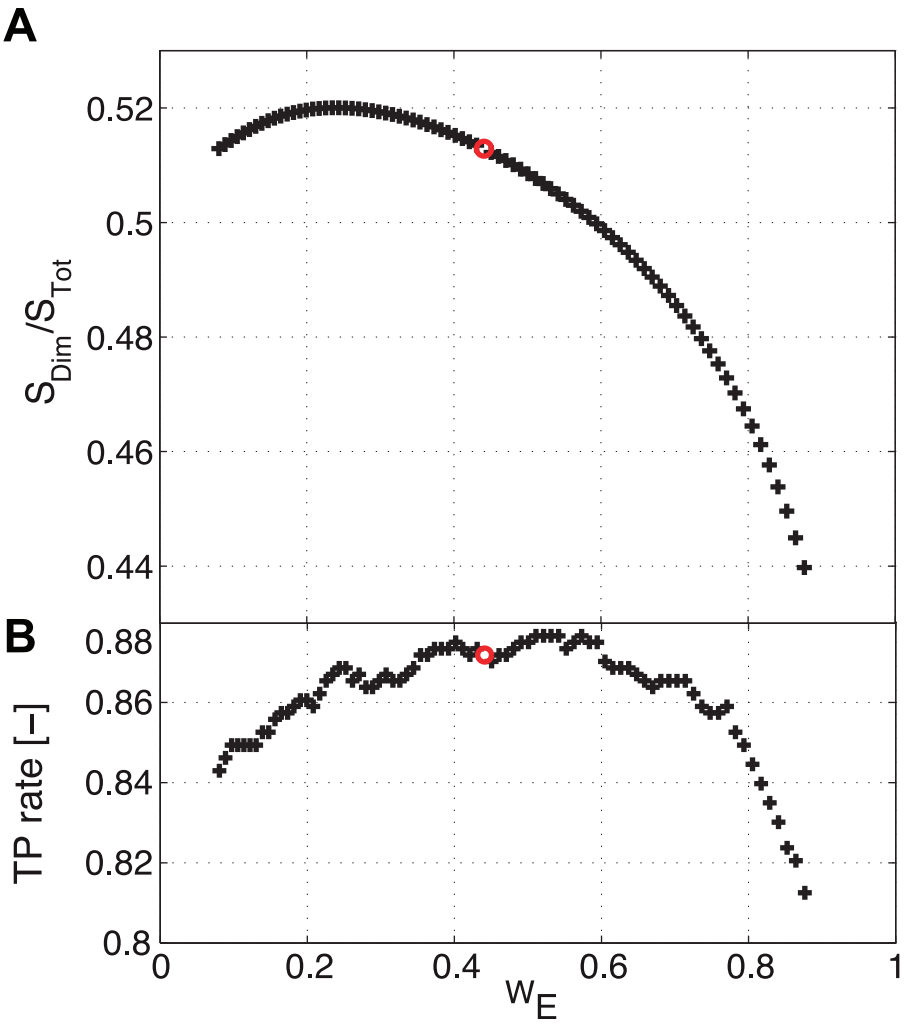

Fig 4. Influence of the relative Bacteria/Eukaryote weights on the dimer contacts. In red are the points corresponding to the unweighted cases. The relative contribution W_E is then dictated by the relative abundance of Eukaryotic to total sequences. **A)** Average DCA score for the 6 predicted dimer contacts, normalized by the average over the top 624 predictions. In abscissa is the relative weight of the Eukaryotic sequences in the total alignment. **B)** True positives for the top 624 predictions with respect to the union of ATP-ADP and dimer contacts.

doi:10.1371/journal.pcbi.1004262.g004

DCA still predicts the six dimeric contacts albeit the higher noise levels, whereas the same are absent in the predictions for $W_E = 1$ (only eukaryotic sequences). All these observations strongly suggest that the predicted homo-dimerization of Hsp70s emerges mainly from bacterial sequences whereas this feature is absent or significantly less conserved in eukaryotes.

Discussion

The function of Hsp70s depends on multiple conformational changes. The structure of the NBD varies with the nature of the bound nucleotide, and the conformational changes induced by nucleotide binding or hydrolysis are propagated to the SBD, thus modulating the Hsp70 interactions with the substrate during the chaperones biochemical cycle. The functional necessity of this orchestrated gymnastic has left a profound footprint in the evolutionary history of these chaperones. It is thus not surprising that sequence analysis methods based on coevolution can

effectively provide structural information on all the functional conformers, thanks to the taxonomic breadth of the available Hsp70 sequences.

Indeed, we have found here that out of the first 624 contacts predicted by the Pseudo-Likelihood based method, 75% are compatible with the structures of ATP- and ADP-bound Hsp70s and another 11% can be explained by the intra- and inter-domain allosteric transformations. It is noteworthy that even though DCA has already been used to detect multiple conformers in proteins [44,48,53,56], the analysis of the Hsp70 family presented here resulted into an unprecedented characterization of a large-scale conformational transition due to the identification of an appreciable fraction of relevant allosteric contacts.

Our DCA predictions show thus a remarkable matching (86%) with the contacts between residues in the experimental structures of functional Hsp70 conformers. Furthermore, we introduced here a topological measure inspired from graph theory to further characterize the quality of DCA predictions. This measure allows a finer appraisal than the binary true or false classification of contacts based on a hard cut-off. The shortest path analysis highlights that as many as 95% of our predictions may reasonably correspond to real contacts. These results suggest that the remaining apparently wrong predictions may actually correspond to yet uncharacterized structural features. In this respect, we observe that DCA predicts a group of 6–7 contacts that are compatible with the interface between the two Hsp70 molecules in the DnaK crystal. While it has been noted that the two DnaK monomers in the crystal possess an interface reflecting a molecular dimer, the weak propensity for dimerization *in vitro* questioned the functional relevance of the dimer in the chaperone cycle [22]. Our results indicate that in the Hsp70 family this dimerization interface is evolutionary conserved in a statistically significant way, thus strongly suggesting an important role for the homo-dimer in the cellular function of Hsp70s.

The remarkable similarity between the intermolecular arrangement of the Hsp70/70 homodimer and that observed in the functional Hsp70/110 hetero-dimer suggests that the latter might represent an eukaryotic functional specialization of a pre-existing Hsp70/70 homodimer. This intriguing hypothesis is actually corroborated by our finding that the co-evolutionary conservation of the dimer interfaces is stronger if bacterial sequences are assigned more statistical weight than eukaryotic ones. Indeed, because bacterial genomes are likely under evolutionary pressure to remain short [61,62], we make the hypothesis that bacteria cannot afford having too many specialized versions of the same protein. As a consequence Hsp70 monomers in the homo-dimer may have to play the same role that the specialized Hsp110s perform in the eukaryotic hetero-dimer.

In this work, we based our analysis on the *a priori* knowledge of the existence of multiple conformers, and the availability of their respective structures. Furthermore, we had at hand a crystallographic homo-dimeric arrangement of two Hsp70 monomers. These data allowed us an in-depth analysis of coevolutionary conservation of structural contacts in DnaK and the prediction of the functional relevance of such a quaternary complex of two bacterial Hsp70s. In general, the blind prediction of multiple conformations or multi-meric arrangements of proteins based solely on coevolutionary information remains an important and challenging problem, whose solution would greatly improve the predictive capabilities of DCA and other coevolutionary methods to the study of previously uncharacterized protein families.

DCA has already shown its impressive potential in reproducing known structural information both at the single protein and at the protein-protein interaction level. The rapid growth of the number of available protein sequences, combined with the improvement of inference algorithms, foreshadow a near future when the use of evolutionary information will be fully exploited as a powerful predictive tool. Thanks to its ubiquity and its evolutionary conservation, together with the state-of-the-art Pseudo-Likelihood optimization method, the Hsp70 family offers a glimpse of these opportunities.

Materials and Methods

Multiple Sequence Alignments

Starting from the PFAM seed of the Hsp70 family (*PF00012*), we manually curated it by adding sequences and suppressing gapped positions. The added sequences were chosen to cover a wide range of organisms, stemming from different taxonomy (see S3 Table). The seed was then aligned using MAFFT. The aligned seed was used to build a Hidden Markov Model (HMM) of the family, using the HMMER utility hmmerbuild. The multiple sequence alignment (MSA) was built by running a HMMER search on the Uniprot database. We used the union of the Uniprot Tremble (un-annotated sequences) and Swissprot (annotated) databases for the extraction of the Hsp70 MSA. All utilities were run with default parameters. Our Hsp70 family MSA is available online as supplementary material (S1 Dataset).

Sequence Filtering

We filtered sequences in the MSA based on their gap contents, keeping only sequences with a maximal gap content of 25% in the final alignment. In order to correct for phylogenetic bias, the MSA was filtered by sequence identity using the HHblits hhfilter utility, allowing a maximum of 90% pairwise sequence identity. An alternative consisting of reweighting the sequences based on their mutual identity leads to nearly identical results. As the DCA computation time grows linearly with the number of sequences in the MSA, we chose to filter the sequences by identity rather than reweighting them.

Direct Coupling Analysis

Direct Coupling Analysis (DCA) was performed using the symmetric version of the pseudo-likelihood method described in [52], which was first introduced in the context of protein contact prediction by Balakrishnan et al. [57]. DCA is based on the use of the maximum entropy principle, constrained to reproduce the observed single- and two-site amino-acid frequencies, leading to a 21-state (corresponding to the 20 natural amino acids and the gap state) Potts model defined by

$$P(S) = \frac{1}{Z} e^{\sum_{i=1}^{N} h_i(s_i) + \sum_{i=1,j>i}^{N-1,N} J_{i,j}(s_i, s_j)}$$

where $\boldsymbol{S}$ is a sequence in the MSA, s_i the amino acid at position i, N the sequence length, and h_i and J_{ij} the model parameters to be optimized. The parameters h_i and J_{ij} are efficiently (but approximately) learned through the numerical optimization of the induced Pseudo-Likelihood with respect to the observed sequences in the MSA [52]. The use of the approximate Maximum Pseudo-Likelihood, in contrast to full Maximum-Likelihood method, allows avoiding the computation of the intractable full partition function **Z**. DCA results come under the form of NxN matrices. Each entry S_{ij} is computed as the Frobenius norm of the local 21x21 coupling matrix J_{ij} and represents the intensity of the evolutionary coupling between residues i and j. An average product correction [63] is finally applied to correct for entropic effects. In our analysis, we retained the N top pairs having the highest coupling scores S_{ij}. The list of the top 624 predicted DCA contacts is available as SI (S2 Dataset).

The optimal regularization parameters of the original method by Ekeberg et al. [52] were used in our study ($\lambda_h = 0.01$, $\lambda_J = 0.01$). As the identity filtering is performed in the pre-processing of the MSA, the reweighting was disabled, setting the maximal sequence identity to 100%.

Backmapping on Structures

The top N DCA predictions are compared to the binary contact maps of the available crystal structures. Contact maps are built by considering two residues in contact if the smallest distance between their heavy (non-hydrogen) atoms is lower than 8.5 Å. As the MSA sites are defined only where the HMM defines relevant positions, adjacent columns in the MSA are not necessarily adjacent in the real sequences (gaps/insertions/deletions are present). The DCA predictions are thus aligned to the contact maps of the structures, considering only DCA scores where the crystal structure contains residues. This implies that not all residues in the structures have corresponding positions in the DCA predictions. Conversely, not all DCA predictions correspond to residues in all structures. In the case of comparison between multiple conformations of DnaK (ATP/ADP states), we considered DCA predictions only where both structures have defined residues.

Shortest Path

In order to assess the quality of predicted contacts, we compute the shortest path between the two residues of the contact. The shortest path (SP) is computed considering the binary contact map as an adjacency matrix of an unweighted and undirected network. Each residue corresponds to a node in the graph, and a link connects two nodes if the corresponding residues are in contact in the protein structure. The shortest path between two residues is the smallest number of links in the graph needed to join two nodes. By definition, physical contacts have SP of 1, while higher SPs indicate a higher topological separation between the residues.

The use of the SP analysis helps highlighting the number of intermediary contacts that would be needed to explain an observed DCA prediction. DCA may not be fully capable of disentangling all indirect correlations in the data, and consequently some residual strong co-evolutionary correlations between residues not in contact in the structure might be found in the predictions. The shortest paths of such predictions give a natural measure of the number of contacts in the native structure through which such a mediated coevolution should propagate in order to be observed. For completeness, we report in S7 Fig the same results using Euclidean distances instead of the shortest paths.

P-value

To quantify the probability of predicted DCA contacts of being random errors, we used the following p-value computation. We introduced a null model where DCA contacts are randomly distributed among all possible pairs. The probability of predicting a set of k correct contacts is thus given by the probability of finding k contacts that exist in the structure, when predicting n total contacts, from a total set of N possible pairs, of which K are contacts in the structure. The proposed null model is thus equivalent to the statistical significance test known as Fisher's exact test. Mathematically, this is modelled by the hypergeometric distribution, given by

$$p(n,k,N,K) = \frac{\binom{K}{N}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where $\binom{N}{n}$ denotes the binomial coefficient. The p-value is defined as the probability of the null model to predict k or more true contacts. For a set of n chosen candidate contacts, with k

predicted contacts, the p-value is thus defined as

$$P_k = \sum_{k'=k}^{n} p(n, k', N, K)$$

In the case of the predicted dimeric contacts, we have thus $k = 6$ dimeric predictions among the $K = 241$ dimeric contacts present in the ATP-state homo-dimeric structure (PDB ID 4jne). We make $n = 624$ total predictions, which can potentially take any of the $N = 624^{*}623/2$ values of the possible contacts. We notice that this is a rather conservative null model, as it does not take in account the fact that among the 624 first predictions, more than 80% are actually correct predictions in the monomeric arrangement of DnaK. Taking this fact in account would drastically decrease the number of predictions n, and would thus lead to a sensitively smaller p-value.

During the review process, two important additional experimental studies regarding Hsp70 homo-dimerization were published.

In the first one, Boateng et al. [64] have confirmed the presence of a DnaK homo-dimer with an interface similar to the one reported here.

The second work, by Marcion et al. [65], highlights the fundamental role of the C-terminal region for human Hsp70 homo-dimerization.

Supporting Information

S1 Fig. Top 624 DCA predictions on the ATP-bound structure of Kityk et al. In the lower triangular part are the structure contacts at threshold 8.5 Å, the upper part contains the DCA predictions, coloured by shortest paths.
(EPS)

S2 Fig. True positive (TP) rates of the DCA predictions. TPs are defined as the ratio between the number of correctly predicted DCA contacts and the total number of DCA predictions. In red are the predictions mapped on the ADP-bound structure, in blue on the ATP-bound structure, in green on the union. The union map is defined as the minimum of each residue pairs distance between the two states.
(EPS)

S3 Fig. DCA predictions on partial structures of Hsp70. In the two representative cases, we considered the top N contacts, where N is the number of residues in the structure. In the lower triangular part are the structure contacts at threshold 8.5 Å, the upper part contains the DCA predictions, coloured by shortest paths. The true positive ratios are computed on the 76 partial structures of Hsp70 in the PDB (41 SBD, 35 NBD). **A)** Top 380 predictions of the NBD of Hsp70 (PDB ID 1s3x). **B)** Top 213 predictions of the SBD of Hsp70 (PDB ID 4hyb). **C)** For the structures of the NBD, we considered the top 400 contacts. **D)** For the structures of the SBD the top 150.
(EPS)

S4 Fig. Hsp70 family DCA predictions projected on Hsp70-Hsp110 hetero-dimers. In the lower triangular part are the structure contacts at threshold 8.5 Å, the upper part contains the DCA predictions, coloured by shortest paths. A) Top 534 DCA contacts of the yeast SSE1 (Hsp110 homologue)—Bovine Hsc70 dimer (PDB ID 3c7n). B) Top 624 of the yeast SSE1 (Hsp110 homologue)—Human Hsp70 dimer (PDB ID 3d2e).
(EPS)

S5 Fig. Top 624 DCA contacts, using only the Hsp70 tagged sequence of the MSA (resulting in 1781 sequences). In the lower triangular part are the structure contacts at threshold 8.5 Å, the upper part contains the DCA predictions, coloured by shortest paths.
(EPS)

S6 Fig. Top 624 DCA contacts, using only the bacterial or eukaryotic sequence of the MSA. In the lower triangular part are the structure contacts at threshold 8.5 Å, the upper part contains the DCA predictions, coloured by shortest paths. A) Bacterial MSA (1982 sequences). B) Eukaryotic sequences (1562 sequences).
(EPS)

S7 Fig. DCA analysis reported using Euclidean Distances. From top to bottom: ADP-bound state, ATP-bound state, Union of ADP+ATP bound states, Union of ADP+ATP bound states and ATP-state homo-dimeric contacts.
(EPS)

S8 Fig. Alignment of the two ATP state PDB structures 4jne and 4b9q. The two views show a 180° rotated version of the structural alignment between the two structures. The RMSD, computed on 597 overlapping CA atoms is of ~2Å.
(TIF)

S1 Table. Allosteric DCA predicted contacts among the first top 624 predictions.
(DOCX)

S2 Table. The six dimeric contacts predicted among the top 624 DCA contacts in the Hsp70 family.
(DOCX)

S3 Table. Uniprot sequences IDs used to build the initial seed of the Hsp70 family MSA.
(DOCX)

S1 Dataset. Multiple Sequence Alignment of the Hsp70 family.
(FASTA)

S2 Dataset. Top 624 predicted DCA contacts, sorted by decreasing coevolutionary strength.

(DAT)

S1 Text. Supporting information text.
(PDF)

Acknowledgments

We thank Andrea Pagnani, Pierre Goloubinoff, Andrija Finka, Jacques Rougemont and David Fabrice for useful discussions.

Author Contributions

Conceived and designed the experiments: DM AB PDLR SM. Performed the experiments: DM AB PDLR SM. Analyzed the data: DM AB PDLR SM. Contributed reagents/materials/analysis tools: DM AB PDLR SM. Wrote the paper: DM AB PDLR SM.

References

1. Mayer MP Gymnastics of molecular chaperones. *Mol Cell* 2010; 39 (3):321–31. doi: 10.1016/j.molcel.2010.07.012 PMID: 20705236

2. Hartl FU, Bracher A, Hayer-Hartl M Molecular chaperones in protein folding and proteostasis. *Nature* 2011; 475 (7356):324–32. doi: 10.1038/nature10317 PMID: 21776078

3. Daugaard M, Rohde M, äättelä M The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. *FEBS Lett* 2007; 581 (19):3702–10. PMID: 17544402

4. Kampinga HH, Hageman J, Vos MJ, Kubota H, Tanguay RM, Bruford E a, et al. Guidelines for the nomenclature of the human heat shock proteins. *Cell Stress Chaperones* 2009; 14 (1):105–11. doi: 10.1007/s12192-008-0068-7 PMID: 18663603

5. Mayer MP Hsp70 chaperone dynamics and molecular mechanism. *Trends Biochem Sci* 2013; 38 (10):507–14. doi: 10.1016/j.tibs.2013.08.001 PMID: 24012426

6. Koplin A, Preissler S, Ilina Y, Koch M, Scior A, Erhardt M, et al. A dual function for chaperones SSB-RAC and the NAC nascent polypeptide-associated complex on ribosomes. *J Cell Biol* 2010; 189 (1):57–68.

7. Albanèse V, Reissmann S, Frydman J A ribosome-anchored chaperone network that facilitates eukaryotic ribosome biogenesis. *J Cell Biol* 2010; 189 (1):69–81.

8. Schuermann JP, Jiang J, Cuellar J, Llorca O, Wang L, Gimenez LE, et al. Structure of the Hsp110:Hsc70 nucleotide exchange machine. *Mol Cell* 2008; 31 (2):232–43.

9. Truman AW, Kristjansdottir K, Wolfgeher D, Hasin N, Polier S, Zhang H, et al. CDK-dependent Hsp70 Phosphorylation controls G1 cyclin abundance and cell-cycle progression. *Cell* 2012; 151 (6):1308–18. doi: 10.1016/j.cell.2012.10.051 PMID: 23217712

10. Rapoport TA Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* 2007; 450 (7170):663–9. PMID: 18046402

11. Cassina L, Casari G The Tightly Regulated and Compartmentalised Import, Sorting and Folding of Mitochondrial Proteins. *Open Biol J* 2009; (2:):200–221.

12. Kalia L V, Kalia SK, Chau H, Lozano AM, Hyman BT, McLean PJ Ubiquitinylation of α-synuclein by carboxyl terminus Hsp70-interacting protein (CHIP) is regulated by Bcl-2-associated athanogene 5 (BAG5). *PLoS One* 2011; 6 (2):e14695. doi: 10.1371/journal.pone.0014695.Ubiquitinylation PMID: 21358815

13. Muller P, Ruckova E, Halada P, Coates PJ, Hrstka R, Lane DP, et al. C-terminal phosphorylation of Hsp70 and Hsp90 regulates alternate binding to co-chaperones CHIP and HOP to determine cellular protein folding/degradation balances. *Oncogene* 2013; 32 (25):3101–10. doi: 10.1038/onc.2012.314 PMID: 22824801

14. Xu L, Hasin N, Shen M, He J, Xue Y, Zhou X, et al. Using steered molecular dynamics to predict and assess Hsp70 substrate-binding domain mutants that alter prion propagation. *PLoS Comput Biol* 2013; 9 (1):e1002896. doi: 10.1371/journal.pcbi.1002896 PMID: 23382668

15. Rios PDL, Barducci A Hsp70 chaperones are non-equilibrium machines that achieve ultra-affinity by energy consumption. *Elife* 2014;1–10.

16. Kampinga H, Craig E The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nat Rev Mol Cell Biol* 2010; 11 (8):579–92. doi: 10.1038/nrm2941 PMID: 20651708

17. Shaner L, Morano KA All in the family: atypical Hsp70 chaperones are conserved modulators of Hsp70 activity. *Cell Stress Chaperones* 2007; 12 (1):1–8. PMID: 17441502

18. Rampelt H, Kirstein-Miles J, Nillegoda NB, Chi K, Scholz SR, Morimoto RI, et al. Metazoan Hsp70 machines use Hsp110 to power protein disaggregation. *EMBO J* 2012; 31 (21):4221–35. doi: 10.1038/emboj.2012.264 PMID: 22990239

19. Rüdiger S, Buchberged A, Bukau B Interaction of Hsp70 chaperones with substrates. *Nat Struct Biol* 1997; 4 (5):342–349. PMID: 9145101

20. Bertelsen EB, Chang L, Gestwicki JE, Zuiderweg ERP Solution conformation of wild-type E. coli Hsp70 (DnaK) chaperone complexed with ADP and substrate. *Proc Natl Acad Sci U S A* 2009; 106 (21):8471–6. doi: 10.1073/pnas.0903503106 PMID: 19439666

21. Kityk R, Kopp J, Sinning I, Mayer MP Structure and dynamics of the ATP-bound open conformation of Hsp70 chaperones. *Mol Cell* 2012; 48 (6):863–74. doi: 10.1016/j.molcel.2012.09.023 PMID: 23123194

22. Qi R, Sarbeng EB, Liu Q, Le KQ, Xu X, Xu H, et al. Allosteric opening of the polypeptide-binding site when an Hsp70 binds ATP. *Nat Struct Mol Biol* 2013; 20 (7):900–7. doi: 10.1038/nsmb.2583 PMID: 23708608

23. Zuiderweg ERP, Bertelsen EB, Rousaki A, Mayer MP, Gestwicki JE, Ahmad A in *Molecular Chaperones*, pp 99–153.

24. Jiang J, Prasad K, Lafer EM, Sousa R Structural basis of interdomain communication in the Hsc70 chaperone. *Mol Cell* 2005; 20 (4):513–24. PMID: 16307916

25. Zhuravleva A, Clerico EM, Gierasch LM An interdomain energetic tug-of-war creates the allosterically active state in Hsp70 molecular chaperones. *Cell* 2012; 151 (6):1296–307. doi: 10.1016/j.cell.2012.11.002 PMID: 23217711

26. King C, Eisenberg E, Greene L Polymerization of 70-kDa Heat Shock Protein by Yeast DnaJ in ATP. *J Biol Chem* 1995; 270 (38):22535–22540. PMID: 7673245

27. Benaroudj N, Fouchaq B, Ladjimi MM The COOH-terminal Peptide Binding Domain Is Essential for Self-association of the Molecular Chaperone HSC70. *J Biol Chem* 1997; 272 (13):8744–8751. PMID: 9079709

28. Fouchaq B, Benaroudj N, Ebel C, Ladjimi MM Oligomerization of the 17-kDa peptide-binding domain of the molecular chaperone HSC70. *Eur J Biochem* 1999; 259 (1–2):379–84.

29. Angelidis CE, Lazaridis I, Pagoulatos GN Aggregation of hsp70 and hsc70 in vivo is distinct and temperature-dependent and their chaperone function is directly related to non-aggregated forms. *Eur J Biochem* 1999; 259 (1–2):505–12.

30. Thompson AD, Bernard SM, Skiniotis G, Gestwicki JE Visualization and functional analysis of the oligomeric states of Escherichia coli heat shock protein 70 (Hsp70/DnaK). *Cell Stress Chaperones* 2012; 17 (3):313–27.

31. Aprile F a, Dhulesia A, Stengel F, Roodveldt C, Benesch JLP, Tortora P, et al. Hsp70 oligomerization is mediated by an interaction between the interdomain linker and the substrate-binding domain. *PLoS One* 2013; 8 (6).

32. Dragovic Z, Broadley S A, Shomura Y, Bracher A, Hartl FU Molecular chaperones of the Hsp110 family act as nucleotide exchange factors of Hsp70s. *EMBO J* 2006; 25 (11):2519–28. PMID: 16688212

33. Raviol H, Bukau B, Mayer MP Human and yeast Hsp110 chaperones exhibit functional differences. *FEBS Lett* 2006; 580 (1):168–74. PMID: 16364315

34. Mattoo RUH, Sharma SK, Priya S, Finka A, Goloubinoff P Hsp110 is a bona fide chaperone using ATP to unfold stable misfolded polypeptides and reciprocally collaborate with Hsp70 to solubilize protein aggregates. *J Biol Chem* 2013; 288 (29):21399–411. doi: 10.1074/jbc.M113.479253 PMID: 23737532

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

36. Altschuh D, Lesk AM, Bloomer AC, Klug A Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol* 1987; 193 (4):693–707. PMID: 3612789

37. Göbel U, Sander C, Schneider R, Valencia A Correlated Mutations and Residue Contacts in Proteins. *Proteins Struct Funct Genet* 1994; 18 (4):309–317. PMID: 8208723

38. Shindyalov IN, Kolchanov NA, Sander C Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* 1994; 7 (3):349–358. PMID: 8177884

39. Neher E How frequent are correlated changes in families of protein sequences ? *Proc Natl Acad Sci U S A* 1994; 91 (1):98–102. PMID: 8278414

40. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A Correlated Mutations Contain Information About Protein-protein Interaction. *J Mol Biol* 1997; 271 (4):511–523. PMID: 9281423

41. Taylor WR, Hatrick K Compensating changes in protein multiple sequence alignments. *Protein Eng* 1994; 7 (3):341–8. PMID: 8177883

42. Lapedes A, Giraud B, Jarzynski C Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. *arXiv Prepr* 2012;1–29.

43. Weigt M, White R a, Szurmant H, Hoch J a, Hwa T Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 2009; 106 (1):67–72. doi: 10.1073/pnas.0805923106 PMID: 19116270

44. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 2011; 108 (49):E1293–301. doi: 10.1073/pnas.1111471108 PMID: 22106262

45. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 2011; 6 (12). doi: 10.1371/journal.pone.0028001 PMID: 22162996

46. De Juan D, Pazos F, Valencia A Emerging methods in protein co-evolution. *Nat Rev Genet* 2013; 14 (4):249–61.

47. Kamisetty H, Ovchinnikow S, Baker D Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci* 2013; 110 (39):15674–15679. doi: 10.1073/pnas.1314045110 PMID: 24009338

48. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 2012; 149 (7):1607–21. doi: 10.1016/j.cell.2012.04.012 PMID: 22579045

49. Marks DS, Hopf TA, Sander C Protein structure prediction from sequence variation. *Nat Biotechnol* 2012; 30 (11):1072–80. doi: 10.1038/nbt.2419 PMID: 23138306

50. Cheng RR, Morcos F, Levine H, Onuchic JN Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc Natl Acad Sci U S A* 2014; 111 (5):E563–71. doi: 10.1073/pnas.1323734111 PMID: 24449878

51. Dago AE, Schug A, Procaccini A, Hoch J a, Weigt M, Szurmant H Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc Natl Acad Sci U S A* 2012; 109 (26):E1733–42. doi: 10.1073/pnas.1201301109 PMID: 22670053

52. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys Rev E* 2013; 87 (1):012707 1–16. PMID: 23410359

53. Jana B, Morcos F, Onuchic JN From structure to function: the convergence of structure based models and co-evolutionary information. *Phys Chem Chem Phys* 2014; 16 (14):6496–507. doi: 10.1039/c3cp55275f PMID: 24603809

54. Ovchinnikov S, Kamisetty H, Baker D Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* 2014; 3:1–21.

55. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* 2014;e034030.

56. General IJ, Liu Y, Blackburn ME, Mao W, Gierasch LM, Bahar I ATPase subdomain IA is a mediator of interdomain allostery in Hsp70 molecular chaperones. *PLoS Comput Biol* 2014; 10 (5):e1003624. doi: 10.1371/journal.pcbi.1003624 PMID: 24831085

57. Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ Learning generative models for protein fold families. *Proteins* 2011; 79 (4):1061–78. doi: 10.1002/prot.22934 PMID: 21268112

58. Burger L, Van Nimwegen E Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* 2010; 6 (1).

59. Morcos F, Jana B, Hwa T, Onuchic JN Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci U S A* 2013; 110 (51):20533–8. doi: 10.1073/pnas.1315625110 PMID: 24297889

60. Easton DP, Kaneko Y, Subjeck JR The Hsp110 and Grp170 stress proteins : newly recognized relatives of the Hsp70s. *Cell Stress Chaperones* 2000; 5 (4):276–290. PMID: 11048651

61. Maniloff J The minimal cell genome: "on being the right size". *Proc Natl Acad Sci U S A* 1996; 93 (September):10004–10006.

62. Moran N a. Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* 2002; 108:583–586. PMID: 11893328

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

64. Boateng EB, Liu Q, Tian X, Yang J, Li H, Wong JL et al. A functional DnaK Dimer Is Essential for the Efficient Interaction with Hsp40 Heat Shock Protein. J. Biol. Chem 2015; 290 (14):8849–8862.

65. Marcion G, Seigneuric R, Chavanne E, Artur Y, Briand L, Hadi T et al. C-terminal amino acids are essential for human heat shock protein protein 70 dimerization. Cell Stress Chaperones 2015; 20 (1):61–72.