COMBINING SEQUENCE AND STRUCTURE INFORMATION TO MODEL BIOLOGICAL SYSTEMS DYNAMICS

by

Ying Liu

BS, Wuhan University, 2003

ME, Tsinghua University, 2006

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Ying Liu

It was defended on

September 23rd, 2011

and approved by

Dr. Panayiotis V. Benos, Associate Professor, Department of Computational and Systems Biology

Dr. Lila M. Gierasch, Professor, Department of Biochemistry and Molecular Biology, University of Massachusetts Amherst

Dr. Christopher J. Langmead, Associate Professor, Department of Computer Science, Carnegie Mellon University

Dr. Daniel M. Zuckerman, Associate Professor, Department of Computational and Systems Biology

Dissertation Advisor: Dr. Ivet Bahar, Professor, Department of Computational and Systems Biology

COMBINING SEQUENCE AND STRUCTURE INFORMATION TO MODEL BIOLOGICAL SYSTEMS DYNAMICS

Ying Liu, PhD

University of Pittsburgh, 2011

Biochemical activity and core stability are essential properties of proteins, maintained usually by conserved amino acids. Structural dynamics emerged in recent years as another essential aspect of protein functionality, which enables the adaptation of the protein to substrate binding. It also underlies its ability to undergo allosteric transitions, while maintaining its fold. Key residues that mediate structural dynamics would thus be expected to be conserved, or exhibit co-evolutionary patterns at least. Yet, the correlation between sequence evolution and structural dynamics is yet to be established. To this end, we have performed in-depth analyses of a number of representative proteins, using a combined approach of sequence analyses and coarse-grained physics-based models. For the Hsp70 family, we studied the interactions of Hsp70 ATPase domains with four different nucleotide exchange factors (NEFs) and revealed two classes of key residues: (i) those highly conserved residues involved in nucleotide binding, which mediate the ATPase domain opening via a global hinge-bending, and (ii) those co-evolving and highly mobile residues engaged in specific interactions with NEFs. The observed interplay between these respective intrinsic (pre-existing, structure-encoded) and specific (co-evolved, sequence-dependent) interactions provides us with insights into the allosteric dynamics and functional evolution of the modular Hsp70 ATPase domain, and inspired a follow-up study that identified a group of key residues mediating the Hsp70 allosteric pathways using perturbation analysis. Along the same lines, a systematic study has been performed on a set of 34 enzymes representing

various folds and functional classes, which generalizes the previous findings and unravels a unique correlation between sequence evolutionary properties and conformational dynamics. Our findings suggest that there is a balance between physical adaptability (enabled by structure-encoded motions) and chemical specificity (conferred by correlated amino acid substitutions), and this balance underlies the selection of a relatively small set of versatile folds by proteins. In another study, HIV-1 protease was investigated as a special case in which short-term evolutionary pressure plays a significant role. With advanced clustering techniques, we differentiated multi-drug resistant mutations from those arising from phylogenetic variations; correspondingly, these mutations exhibit distinctive structural/dynamical features, underlying the role of protein dynamics in conferring drug resistance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my greatest gratitude toward my advisor, Prof. Ivet Bahar, for her consistent support and rigorous training through my Ph.D. work. It was a tough yet highly rewarding experience trying to meet the high standard she has both exemplified and set, and she has always been encouraging and helpful as an advisor. I am very fortunate to have conducted my Ph.D. studies under her guidance.

I also benefited tremendously from the collaboration with Prof. Lila Gierasch. I would like to thank her for guiding my work in the Hsp70 project, and for generously hosting my visit to her lab in July 2011. The help from her lab members are also cordially appreciated.

I am indebted to Prof. Daniel Zuckerman, Prof. Panayiotis Benos, and Prof. Christopher Langmead, for their insightful discussions and suggestions that helped to shape and improve my work over the years. I also had great learning experiences in the courses they taught.

It is my privilege to have the opportunity to work with an excellent group of colleagues, including Drs. Eran Eyal,. Zheng Yang, Tim Lezon, Ahmet Bakan, and many others. My life would have been much more difficult without their help and friendship.

Finally, I would like to thank my beloved wife Zhao Jin, who has accompanied me through all these years with her love and support; and my parents Mr. Xiaowen Liu and Ms. Ping Li, for their unconditional love that has chaperoned me through my life.

1.0 INTRODUCTION

Many proteins are molecular machines. They function because their three-dimensional (3D-) structure allows them to undergo cooperative changes in conformation that maintain the native fold while enabling their biological functions. These collective changes have been pointed out to be structure-encoded, intrinsically accessible to proteins, as deduced from simple physics-based approaches (Bahar et al., 2010). They are predominantly determined by the overall shape or architecture of the protein. On the other hand, amino acid specificity is another important property that selectively mediates the interactions with specific partners and ligands (Tokuriki and Tawfik, 2009). Overall, a subtle balance exists between structure-encoded mechanical properties and sequence-encoded specific properties, and this balance must be evolutionarily optimized to achieve precise functioning.

The interplay between these two effects becomes particularly important in the case of allostery. Allostery is the regulation of the activity by binding another molecule to the "allosteric site". It enables signal transduction across the structure (Changeux and Edelstein, 1998; Kovbasyuk and Kramer, 2004; Gunasekaran et al., 2004; Changeux and Edelstein, 2005). Two classical models have been proposed on the mechanism of allosteric interactions. The first, also known as the Monod-Wyman-Changeux (MWC) model (Monod et al., 1965), hypothesizes a flip-flop machinery in which all subunits of multimeric structures undergo the transition from one conformation to another simultaneously. The Koshland-Némethy-Filmer (KNF) model

(Koshland, Jr. et al., 1966), on the other hand, proposes a different scenario. According to this model, also called sequential allostery model, the conformational change propagates through the structure via an induced-fit mechanism. There is a sequence of events, as opposed to the all-or-none transition of the MWC model. MWC model has found broader support in recent years. Many computational and experimental data collected in recent years point to intrinsic, structure-encoded dynamics, which cooperatively affects the intact structure (Henzler-Wildman et al., 2007; Lange et al., 2008; Bahar et al., 2010). However, some systems are also observed to obey KNF-type motions (Schmeing et al., 2005). It has been proposed that many events may involve a combination of both cooperative (intrinsic to the protein structure) and induced (triggered by substrate binding) events (Tobi and Bahar, 2005; Csermely et al., 2010).

The most important system investigated in this dissertation, the heat shock protein 70 (Hsp70) family of chaperones, is known to be allosterically regulated, as will be shown in the next section. We will elaborate on the structure-encoded dynamics of the Hsp70 ATPase domain in particular, and focus on the collective motions intrinsically favored by the domain architecture. In principle, such motions may have important functional implications, especially if they cooperatively involve a large portion of the structure. Their impairments would thus be consequential and resisted by compensating mutations. The multi-drug resistance in HIV-protease, another important component of our study, will showcase how the interplay between sequence variation and structural dynamics affects the response of the enzyme to external perturbations under well-defined evolutionary pressure.

1.1 HEAT SHOCK PROTEIN 70

Heat shock proteins (HSPs), also known as molecular chaperones, are ATP-regulated machines that perform several housekeeping activities in the cell: they assist in folding newly synthesized peptides, or unfolding and refolding partially folded or misfolded proteins; they regulate the intracellular trafficking of proteins; they facilitate, in particular the recognition of those to be degraded by the proteasome, and most importantly, assist in the correct folding, and prevent the aggregation, of the proteins denatured in response to heat and other environmental stresses (Hartl and Hayer-Hartl, 2002; Hartl and Hayer-Hartl, 2009).

Hsp70 is one of the most ubiquitous members of the HSP family, existing in almost all organisms (Bukau and Horwich, 1998). It is composed of two domains (**Figure 1**): the ATPase domain, also referred as nucleotide binding domain (NBD (Flaherty et al., 1990)), is the major regulatory unit of the molecular machine, and is further divided into four subdomains known as subdomains IA, IB, IIA and IIB; the substrate binding domain (SBD (Zhu et al., 1996)), on the other hand, binds to the client proteins to perform the chaperoning function. The two domains regulate the activity of each other via allosteric communication: ATP hydrolysis at the NBD increases the substrate binding affinity of the SBD, thus lowering the substrate exchange rate of the latter; on the other hand, the replacement of the ADP produced upon ATP hydrolysis by a new ATP (nucleotide exchange) lowers the binding affinity of the SBD thus enhancing the release and exchange of substrates (Bukau and Horwich, 1998).

The regulation of substrate binding affinity during the ATPase cycle is a crucial aspect of the chaperone activity of Hsp70, and notably, of other HSP family members (Ali et al., 2006). The ATPase domain undergoes conformational changes between open and closed forms during the ATPase cycle, which correspond to different nucleotide binding states (**Figure 1**). The open

conformation has been observed in the presence of ATP (Bhattacharya et al., 2009; Bertelsen et al., 2009). Nucleotide exchange efficiency is viewed to be largely dependent on the conformational change to an open state. The alternation between the two states forms the ATPase cycle.

The precise functioning of the Hsp70 ATPase domain involves an interaction with two families of co-factors, also called co-chaperones: the J-domain proteins that catalyze ATP hydrolysis (Craig et al., 2006), and the nucleotide exchange factors (NEFs) that assist in the replacement of ADP with ATP, by significantly increasing the ADP dissociation rate (Kabani, 2009). A molecular understanding of Hsp70 function requires a systematic analysis of the structural basis and mechanism of interaction with these co-chaperones. In Chapter 3, we present our results from the study of the interactions between the Hsp70 ATPase domain and the NEFs. In Chapter 4, we present the results from perturbation scanning analysis to identify allosteric pathways that mediate the interdomain interactions..

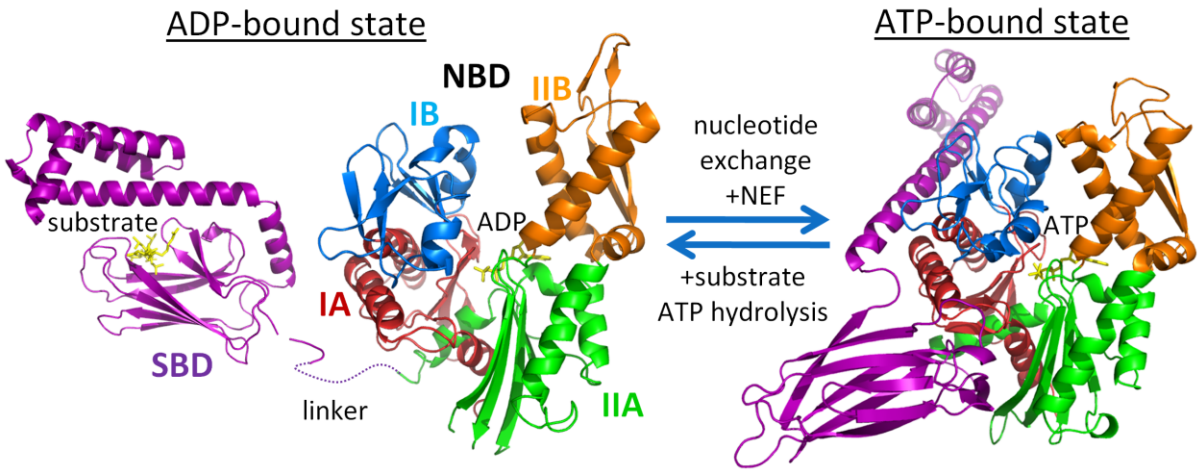

Figure 1. Hsp70 ATPase cycle.

In the ADP-bound state (left), the two domains are loosely connected by the inter-domain linker, and the SBD has a bound substrate. In the ATP-bound state (right), the SBD is docked onto the ATPase domain at its α-helical lid and β-sandwich, leaving the substrate-binding site open. The alternation of the two states is achieved by ATP hydrolysis

or the ADP replacement by ATP. The nucleotide and substrates are shown in stick representation and colored yellow. The four subdomains of the ATPase domain are colored differently as indicated by the label in the ADP-bound state, and the SBD is colored purple. The ribbon diagrams are generated using the PDB files 1DKG (Harrison et al., 1997), 1DKX (Zhu et al., 1996), and a homology model of DnaK (Smock et al., 2010).

1.2 HIV-1 PROTEASE AND ITS DRUG-RESISTANT MUTATIONS

HIV-1 protease is a homodimer with 99 residues in each subunit. It plays an important role in the late stage of viral replication: it cleaves the premature viral polypeptides to peptides that fold into mature virus proteins (Brik and Wong, 2003). HIV-1 protease has been a major drug target for AIDs therapy; however, the ability of HIV-1 protease to rapidly acquire a variety of mutants in response to various protease inhibitors (PIs), known as multidrug resistance (MDR), confers the enzyme with high resistance to anti-AIDS treatments. In addition, a high cooperativity has been documented among drug-resistant mutations observed in HIV-1 protease (Ohtaka et al., 2003).

In the current study, HIV-1 protease is used as model system to study the relation between the sequence, dynamics, and sequence-evolution constraints developed in the presence of highly specific external perturbations (drug treatment). Because of the large sets of sequences available and the observed fast rate of mutations in response to drug treatments, HIV-1 protease is particularly suitable for sequence covariance analysis. The intrinsic dynamics of the molecule can be inferred from its native structure using elastic network models.

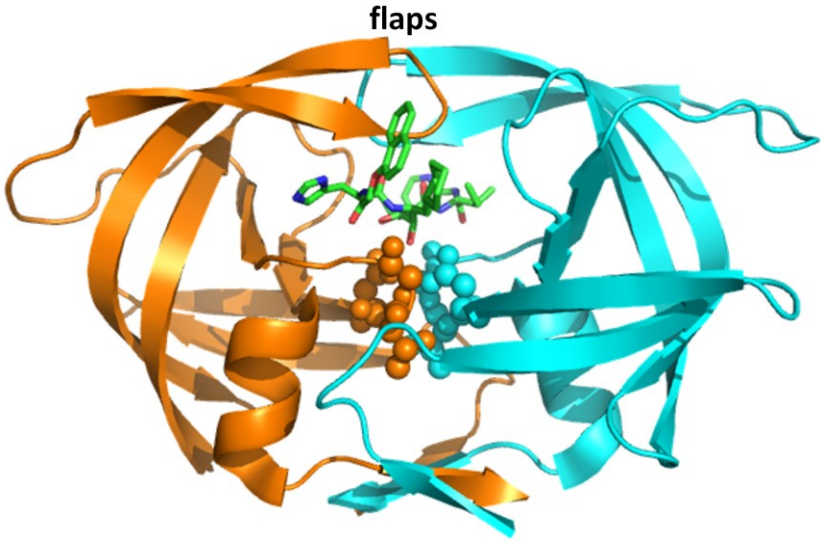

Figure 2. Structure of HIV-1 protease bound to an inhibitor.

The ribbon diagram is generated using the PDB id 1HIV (Thanki et al., 1992). The two monomers are colored orange and cyan. The inhibitor is colored green, and the active site residues (Asp25, Thr26 and Gly27) are shown in sphere representation.

Figure 2 shows the structure of HIV-1 protease in complex with a protease inhibitor (PI). The active site (or catalytic residues Asp25, Thr26 and Gly27) is located at the dimerization interface, and the flaps at the top of the molecule undergo significant conformational fluctuations that allow for the opening/closing of the active site. Extensive studies have been made on this protein structure and dynamics (Cecconi et al., 2001; Zoete et al., 2002; Perryman et al., 2004; Hornak et al., 2006) although the molecular mechanisms of MDR are yet to be elucidated. Our findings on this problem are presented in Chapter 6.

1.3 OUTLINE OF THE DISSERTATION

Figure 3 shows an overview of the work presented in this dissertation. Our work is composed of two groups of topics: the first is to investigate the interplay between sequence, structure, dynamics, and function in different proteins. This part focused on two representative systems, the Hsp70 ATPase domain (Chapters 3) and HIV-1 protease (Chapter 6). In addition, a systematic study of 34 enzymes is presented in Chapter 5, to evaluate and generalize our findings made for the individual proteins. The second group of studies investigates the key residues and their roles in the allosteric communication of Hsp70 domains using a combination of sequence analyses and structural dynamics (Chapter 4). This is a challenging task since much less is known about the mechanism of interdomain interactions in Hsp70. Our studies on Hsp70 structure, dynamics and allostery have been conducted in collaboration with Prof. Lila Gierasch's lab. Finally, in Chapter 7, we discuss potential work that can be pursued in the future.

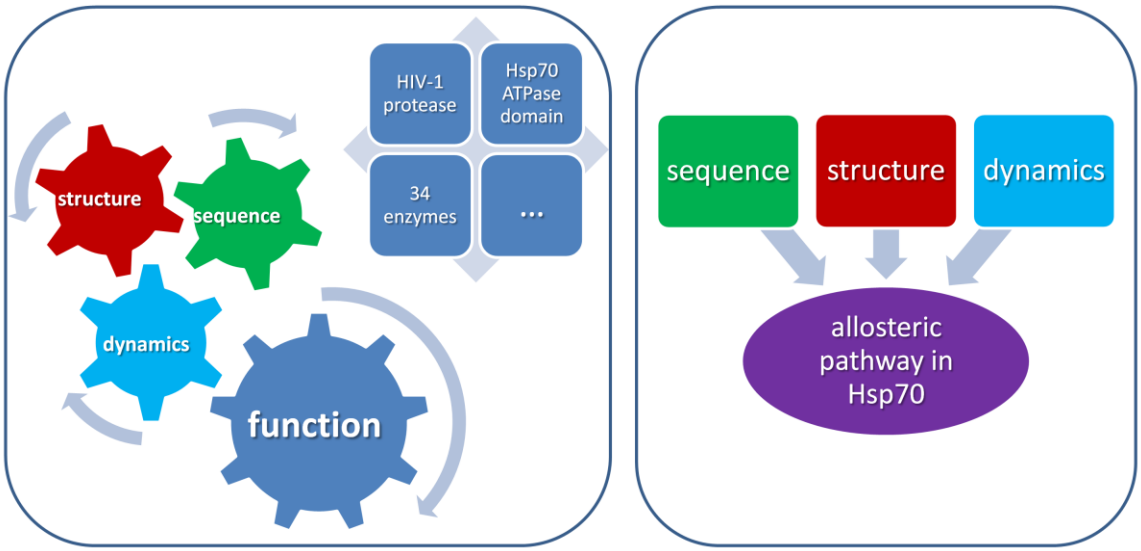

Figure 3. Outline of the dissertation.

Overall, our study led to 3 papers published in peer-reviewed journals or conference proceedings, one book chapter, one submitted manuscript, and another manuscript in preparation. Below is a list of the published and submitted studies.

1. Ying Liu, Lila M.Gierasch, Ivet Bahar. (2010) Role of Hsp70 ATPase Domain Intrinsic Dynamics and Sequence Evolution in Enabling its Functional Interactions with NEFs. *PLoS Computational Biology* **6:** e1000931.
2. Ying Liu and Ivet Bahar. (2010) Toward Understanding Allosteric Signaling Mechanisms in the ATPase Domain Of Molecular Chaperones. *Pacific Symposium on Biocomputing* **2010:**269-80.
3. Ying Liu and Ivet Bahar. (2011) Sequence evolution correlates with structural dynamics. (submitted to *Molecular Biology and Evolution*)
4. Ying Liu*, Eran Eyal*, Ivet Bahar. (2008) Analysis of Correlated Mutations in HIV-1 Protease Using Spectral Clustering. *Bioinformatics* **24:**1243-1250. (* equal contribution)
5. Pemra Doruker, Ying Liu, Zheng Yang, Ivet Bahar. (2012) Coarse-grained methods: Applications to allosteric proteins. Book chapter in *Comprehensive Biophysics*, Elsevier. (to appear in 2012).

2.0 THEORY AND METHODS

The last decade has witnessed a rapid growth in protein sequence and structure data (**Figure 4**) and the development of a broad array of computational tools to refine and analyze this wealth of information. Recently, the integration of information derived from different data sources using multiple approaches invites increasing attention to the emerging field of systems biology, opening the way to a more comprehensive understanding of biological systems dynamics.

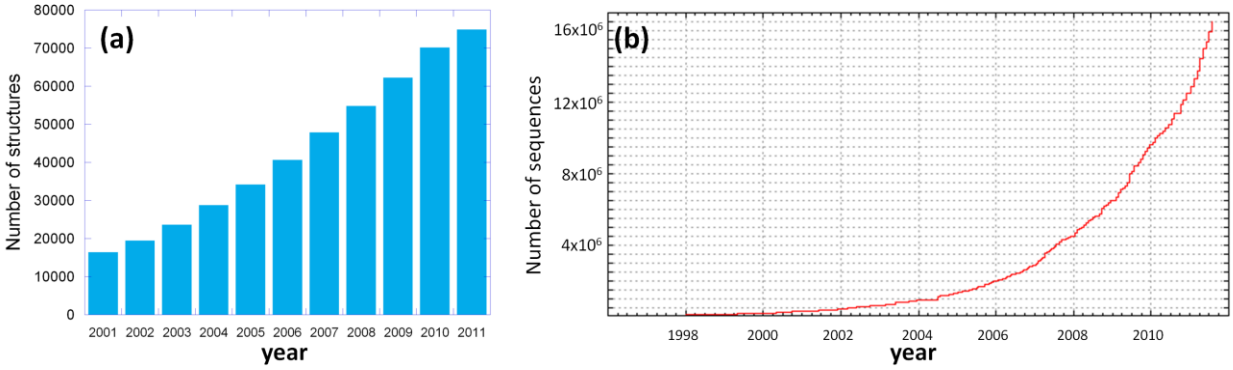

Figure 4. Growth of biological databases over the last decade.

(a) Number of structures in the Protein Data Bank (Berman et al., 2000) and (b) number of sequences in the UniprotKB/TrEMBL database (Jain et al., 2009).

Protein structure is one of the most important sources for understanding the mechanism of protein functions. However, proteins are subject to continuous structural fluctuations under native state conditions, and sometimes they undergo significant conformational changes, or allosteric switches to achieve their function; the static structure can provide limited information

only about the mechanism of protein function. Protein dynamics is recognized as the bridge between structure and function (Smock and Gierasch, 2009; Bahar et al., 2010). Inasmuch as structure and dynamics are closely related, structures-based computational models have found great success in studying protein dynamics.

Protein sequence analysis provides another important approach toward understanding protein function. Sequence variation patterns reflect the evolutionary constraints to maintain the proper function of proteins, and the same constraints underlie the intrinsic correlation between dynamics and sequence. Sequence variations observed in multiple sequence alignments (MSAs) result from an evolutionary process over years, whereas molecular dynamics describe molecular events at the time scale of nanoseconds; hence protein dynamics and sequence evolution entail complementary information regarding the protein function. A combined analysis can help to cross-validate hypotheses using different perspectives, and gain more insights into the sequence → structure → dynamics → function mapping paradigm. The computational tools described in the dissertation are developed to make progress toward this goal.

Here is the outline of this chapter. Section 2.1 is devoted to the description of elastic network models, including the Gaussian Network Model (GNM) and Anisotropic Network Model (ANM). Section 2.2 introduces two methods that follow the line of perturbation analysis, perturbation response scanning (PRS) and residue centrality. Section 2.3 focuses on the techniques of sequence analysis used in the current work. Information-theoretic approaches are introduced in subsection 2.3.1, and phylogeny-based evolutionary trace method is presented in subsection 2.3.2. Finally section 2.4 elaborates on spectral clustering, with its applications to co-evolution analysis.

2.1 ELASTIC NETWORK MODELS (ENMS)

Elastic network models (ENMs) have been widely used to study the collective dynamics of biomolecules (Rader et al., 2006). This coarse-grained models approximate the bio-molecular structure in its equilibrium state as a network composed of a group of beads inter-connected by elastic springs. In the case of protein modeling, the bead serves as the abstraction of an amino acid residue, and the spring stands for inter-residue interactions (**Figure 6a**). The equilibrium state is thus stabilized by a sum of harmonic potentials contributed by the individual springs.

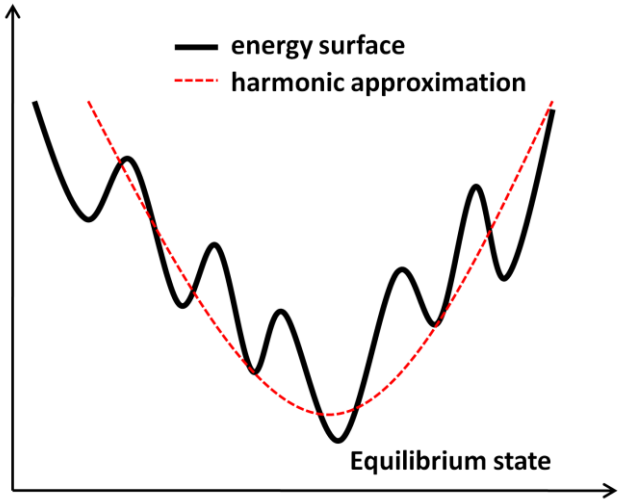

Figure 5. Approximating the vicinity of the equilibrium state by harmonic potentials.

The model could be made more complicated/detailed with different combinations of force constant and criteria for inter-residue interactions. However, the pursuit of every detail of the molecule could render the model mathematically intractable, and also obscure the dominant patterns that govern the functionally relevant motions of the molecule. Hence we employed a minimalist model which adopts a uniform force constant and a single cutoff distance between the C^{α} atoms of residues to determine their interaction.

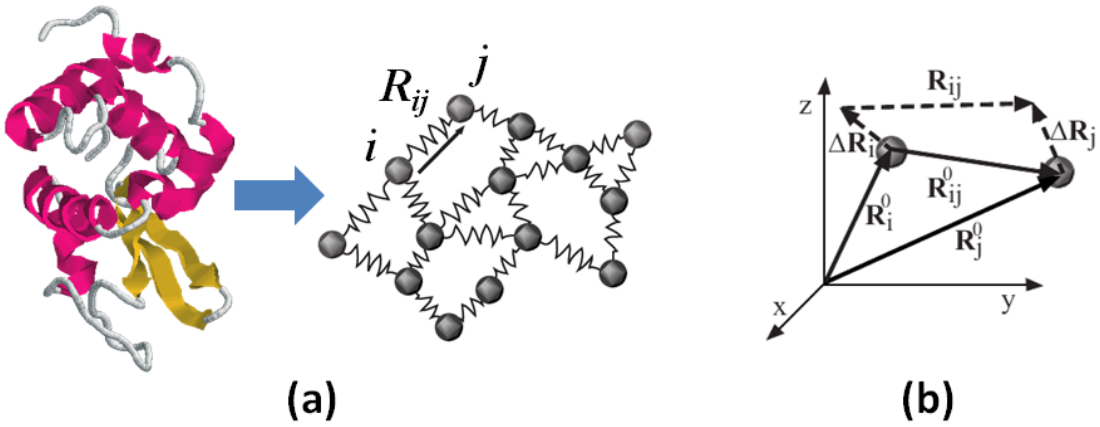

Figure 6. General scheme for constructing an ENM using PDB coordinates (adopted from (Rader et al., 2006)).

(a) Representation of the protein structure as an elastic network. (b) Deformed position vector due to conformational fluctuations. See the text for the definition of the variables.

In our study, two most widely used ENMs are employed, the Gaussian Network Model (Bahar et al., 1997; Yang et al., 2006) and Anisotropic Network Model (Doruker et al., 2000; Tama and Sanejouand, 2001; Atilgan et al., 2001). Their differences arise from the underlying potentials (Rader et al., 2006): the GNM potential penalizes the orientational changes in inter-residue separations in addition to magnitude changes, whereas the ANM potential only considers the change in magnitude of the position vector. This difference also leads to different mathematical treatments of the force constant matrix, as will be demonstrated in the following subsections.

2.1.1 Gaussian network model (GNM)

The potential used in GNM takes the following form

$$V_{\mathrm{GNM}} = \frac{\gamma}{2}\sum_{i,j}^{N}\left\|\mathbf{R}_{ij} - \mathbf{R}_{ij}^{0}\right\|^{2} H(r_c - R_{ij}) \tag{1}$$

where $\mathbf{R}_{ij}^{0}$ is the position vector between residues i and j in the equilibrium state, and $\mathbf{R}_{ij}$ is the deformed position vector due to fluctuations (see **Figure 6b**). N is the total number of beads/residues in the network/structure. H is the heavyside step function which equals (1) if the argument is positive, and 0 otherwise. r_c is the cutoff distance that determines if residues i and j are close enough to interact with each other. Equation (1) can be expressed in terms of the *Kirchhoff matrix* $\mathbf{\Gamma}$**,** defined as

$$\Gamma_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and } R_{ij} \leq r_c \\ 0, & \text{if } i \neq j \text{ and } R_{ij} > r_c \\ -\sum_{j,j\neq i}\Gamma_{ij}, & \text{if } i = j \end{cases} \tag{2}$$

V_{GNM} can thus be written as

$$V_{\mathrm{GNM}} = \frac{\gamma}{2}\sum_{i,j}^{N}\Gamma_{ij}\left\|\mathbf{R}_{ij} - \mathbf{R}_{ij}^{0}\right\|^{2} \tag{3}$$

Note that the change in inter-residue distance vector may be expressed as $\Delta\mathbf{R}_{ij} = \mathbf{R}_{ij} - \mathbf{R}_{ij}^{0} = \Delta\mathbf{R}_j - \Delta\mathbf{R}_i$ (see **Figure 6b**) . Writing $\Delta\mathbf{R}_i$ in the vector form as $(\Delta X_i \quad \Delta Y_i \quad \Delta Z_i)^{\mathrm{T}}$, we obtain

$$\begin{aligned} V_{\mathrm{GNM}} &= \frac{\gamma}{2}\sum_{i,j}^{N}\Gamma_{ij}[(\Delta X_i - \Delta X_j)^2 + (\Delta Y_i - \Delta Y_j)^2 + (\Delta Z_i - \Delta Z_j)^2] \\ &= \frac{\gamma}{2}(\Delta\mathbf{X}^{\mathrm{T}}\mathbf{\Gamma}\Delta\mathbf{X} + \Delta\mathbf{Y}^{\mathrm{T}}\mathbf{\Gamma}\Delta\mathbf{Y} + \Delta\mathbf{Z}^{\mathrm{T}}\mathbf{\Gamma}\Delta\mathbf{Z}) \end{aligned} \tag{4}$$

where $\Delta\mathbf{X} = (\Delta X_1 \quad \Delta X_2 \quad \ldots \Delta X_N)^{\mathrm{T}}$ and so on.

Because of the isotropic assumption inherent to the GNM, the probability distribution of the fluctuations $\Delta\mathbf{R}$ can be decomposed as the product of the distributions for different components, i.e.,

$$P(\Delta\mathbf{R}) = P(\Delta\mathbf{X}, \Delta\mathbf{Y}, \Delta\mathbf{Z}) = p(\Delta\mathbf{X})\,p(\Delta\mathbf{Y})\,p(\Delta\mathbf{Z}) \tag{5}$$

By the Boltzmann's law

$$\begin{aligned} p(\Delta\mathbf{X}) &\propto \exp\left\{-\frac{\gamma}{2k_{\mathrm{B}}T}\Delta\mathbf{X}^{\mathrm{T}}\boldsymbol{\Gamma}\Delta\mathbf{X}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\Delta\mathbf{X}^{\mathrm{T}}\left(\frac{k_{\mathrm{B}}T}{\gamma}\boldsymbol{\Gamma}^{-1}\right)^{-1}\Delta\mathbf{X}\right)\right\} \end{aligned} \tag{6}$$

where k_{B} is the Boltzmann constant and T is the absolute temperature. Note that the expression of $p(\Delta\mathbf{X})$ obtained in equation (6) is actually the probability density function of a multivariate Gaussian distribution where the $N\times N$ covariance matrix $\left\langle \Delta\mathbf{X}\Delta\mathbf{X}^{\mathrm{T}} \right\rangle$ is equal to $(k_{\mathrm{B}}T/\gamma)\ \boldsymbol{\Gamma}^{-1}$. Considering

$$\left\langle \Delta\mathbf{X}\Delta\mathbf{X}^{\mathrm{T}} \right\rangle = \left\langle \Delta\mathbf{Y}\Delta\mathbf{Y}^{\mathrm{T}} \right\rangle = \left\langle \Delta\mathbf{Z}\Delta\mathbf{Z}^{\mathrm{T}} \right\rangle = \frac{1}{3}\left\langle \Delta\mathbf{R}\Delta\mathbf{R}^{\mathrm{T}} \right\rangle \tag{7}$$

we then obtain

$$\left\langle \Delta\mathbf{R}_i^2 \right\rangle = \frac{3k_{\mathrm{B}}T}{\gamma}(\boldsymbol{\Gamma}^{-1})_{ii} \tag{8}$$

to compute the mean square fluctuations (MSF) of residues and

$$\left\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \right\rangle = \frac{3k_{\mathrm{B}}T}{\gamma}(\boldsymbol{\Gamma}^{-1})_{ij} \tag{9}$$

to compute the correlations between the fluctuation of different residues.

A caveat in equation (6) lies in the fact that the Kirchhoff matrix $\boldsymbol{\Gamma}$ has a zero eigenvalue, and is therefore not invertible. However, this can be circumvented by taking the pseudo-inverse of $\boldsymbol{\Gamma}$. Because $\boldsymbol{\Gamma}$ is semi-positive definite, singular value decomposition (SVD) yields $\boldsymbol{\Gamma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues on the diagonal and $\mathbf{U}$ is an orthonormal matrix with the k^{th} column $\mathbf{u}^{(k)}$ being the eigenvector corresponding to eigenvalues λ_k. Then $\boldsymbol{\Gamma}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}}$, and in component form equation (8) becomes

$$\left\langle \Delta \mathbf{R}_i^2 \right\rangle = \frac{3k_{\mathrm{B}}T}{\gamma}(\mathbf{\Gamma}^{-1})_{ii} = \frac{3k_{\mathrm{B}}T}{\gamma}[\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}}]_{ii} = \frac{3k_{\mathrm{B}}T}{\gamma}\sum_{k=1}^{N-1}[\lambda_k^{-1}\mathbf{u}^{(k)}\mathbf{u}^{(k)\mathrm{T}}]_{ii} \tag{10}$$

Note that the summation is performed over non-zero eigenvalues of $\mathbf{\Gamma}$, which are indexed from 1 to N-1.

The obtained eigenvectors $\mathbf{u}^{(k)}$ are designated as collective modes of the dynamics, and the corresponding eigenvalues λ_k serve as frequencies of each mode; thus the mode corresponding to the smallest nonzero eigenvalue is the slowest, or softest mode. The correlation matrix can be decomposed into the contributions from individual modes

$$\left\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \right\rangle = \frac{3k_{\mathrm{B}}T}{\gamma}(\mathbf{\Gamma}^{-1})_{ij} = \frac{3k_{\mathrm{B}}T}{\gamma}[\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}}]_{ij} = \frac{3k_{\mathrm{B}}T}{\gamma}\sum_{k=1}^{N-1}[\lambda_k^{-1}\mathbf{u}^{(k)}\mathbf{u}^{(k)\mathrm{T}}]_{ij} \tag{11}$$

indicating the softest mode makes the largest contribution to the $\mathbf{\Gamma}^{-1}$ among others. The i^{th} element, $[\mathbf{u}^{(k)}]_i$, of $\mathbf{u}^{(k)}$ describes the displacement of residue i along the k^{th} mode axis; the plot of $[\mathbf{u}^{(k)}]_i^2$ as a function of residue index i defines the *mobility profile* $M_i^{(k)}$ in mode k. The mobility profile averaged over a set of m modes is

$$\left\langle M_i \right\rangle|_m = \frac{\sum_{k=1}^{m}\lambda_k^{-1}[\mathbf{u}^{(k)}]_i^2}{\sum_{k=1}^{m}\lambda_k^{-1}} = \frac{\sum_{k=1}^{m}\lambda_k^{-1}M_i^{(k)}}{\sum_{k=1}^{m}\lambda_k^{-1}} \tag{12}$$

where the reciprocal λ_k^{-1} serves as the statistical weight of mode k. The above equation yields the contribution of the first m modes to the overall dynamics, using the probabilistic contribution of each mode, given by

$$\sum_{k=1}^{m}\lambda_k^{-1} \,/\, \sum_{k=1}^{N\text{-}1}\lambda_k^{-1} \tag{13}$$

The soft modes correspond to large motions that are usually beyond the time scale typically accessible to all-atom simulations, and such motions are usually relevant to the biological function of the protein (Yang and Bahar, 2005; Bahar et al., 2010). In particular, residues exhibiting restricted mobility in the slow modes have been shown in numerous applications to play key mechanical roles, serving as hinges, for example, at the interface between domains subject to concerted motions.

2.1.2 Anisotropic network model (ANM)

The total potential of the structure in the ANM is defined as (Atilgan et al., 2001)

$$V_{\mathrm{ANM}} = \frac{\gamma}{2}[(\|\mathbf{R}_{ij}\| - \|\mathbf{R}_{ij}^0\|)^2 H(r_c - R_{ij})] \tag{14}$$

The force constant matrix in ANM is derived from the equilibrium condition of each residue. Let f_{ij} be the elastic force between residues i and j due to their interaction, s_{ij} be the distance between the C^{α} atoms of residues i and j. In the equilibrium state, the net force applied on any residue i is 0 in all directions, i.e.

$$\begin{aligned}
\sum_j f_{ij} \cos\alpha_{ij}^{\mathrm{X}} &= \sum_j f_{ij} \frac{(X_j - X_i)}{s_{ij}} = 0 \\
\sum_j f_{ij} \cos\alpha_{ij}^{\mathrm{Y}} &= \sum_j f_{ij} \frac{(Y_j - Y_i)}{s_{ij}} = 0 \\
\sum_j f_{ij} \cos\alpha_{ij}^{\mathrm{Z}} &= \sum_j f_{ij} \frac{(Z_j - Z_i)}{s_{ij}} = 0
\end{aligned} \tag{15}$$

The relation can be written in matrix form by introducing the $3N \times M$ cosine matrix **B** and the $M \times 1$ force vector **f**, where M is the total number of interactions. Then

$$\mathbf{Bf} = \mathbf{0} \tag{16}$$

In addition, by Hooke's law, the force is related to the deformation by $\mathbf{f} = \mathbf{K}\Delta\mathbf{s}$, where $\mathbf{K}$ is the $M \times M$ diagonal matrix with each diagonal element being the force constant of an interaction, and $\mathbf{K} = \gamma\mathbf{I}$ in the case of ANM. Then

$$\mathbf{BK}\Delta\mathbf{s} = \gamma\mathbf{B}\Delta\mathbf{s} = \mathbf{0} \tag{17}$$

Let $\Delta\mathbf{R}$ be the $3N \times 1$ deformation vector, then $\Delta\mathbf{R}$ and $\Delta\mathbf{s}$ can be related through $\mathbf{B}$

$$\mathbf{B}^{\mathrm{T}}\Delta\mathbf{R} = \Delta\mathbf{s} \tag{18}$$

Substituting $\Delta\mathbf{s}$ in equation (17) with the expression in equation (18) yields

$$\mathbf{BB}^{\mathrm{T}}\Delta\mathbf{R} = \mathbf{0} \tag{19}$$

Since $\mathbf{\Gamma}\Delta\mathbf{R}_{N\times 1} = \mathbf{0}$ in the case of GNM, $\mathbf{BB}^{\mathrm{T}}$ serves as the counterpart of $\mathbf{\Gamma}$ in the anisotropic case. It can also be shown that $\mathbf{BB}^{\mathrm{T}}$ is equivalent to the Hessian matrix $\mathbf{H}$ derived in the context of normal mode analysis (NMA). Let V be the potential between residue i and j, then the Hessian matrix $\mathbf{H}$ is composed of N^2 super-elements each with size 3×3, and the ij^{th} super element $\mathbf{H}_{ij}$ is

$$\mathbf{H}_{ij} = \begin{pmatrix} \partial^2 V/\partial X_i \partial X_j & \partial^2 V/\partial X_i \partial Y_j & \partial^2 V/\partial X_i \partial Z_j \\ \partial^2 V/\partial Y_i \partial X_j & \partial^2 V/\partial Y_i \partial Y_j & \partial^2 V/\partial Y_i \partial Z_j \\ \partial^2 V/\partial Z_i \partial X_j & \partial^2 V/\partial Z_i \partial Y_j & \partial^2 V/\partial Z_i \partial Z_j \end{pmatrix} \tag{20}$$

The evaluation of the normal modes of motions is performed by eigenvalue decomposition of $\mathbf{H}$, similar to the eigenvalue decomposition of $\mathbf{\Gamma}$ in the GNM. Since this time the $3N$-dimensional mode provides directional information about the motions, it can be conveniently compared with the deformation vectors observed in experiments (obtained by comparing different conformations of the same protein resolved experimentally). Suppose the $3N$-dimensional deformation vector is denoted by $\mathbf{d}$, then correlation cosine between the k^{th} ANM eigenvector $\mathbf{v}^{(k)}$ ($k = 1,\ldots, 3N\text{-}6$) and $\mathbf{d}$ is ($|\mathbf{v}^{(k)} \bullet \mathbf{d}| / |\mathbf{d}|$), and the *cumulative overlap* (Yang et al., 2009) achieved by the m softest modes is defined as

$$\mathrm{CO}(m) = \sqrt{\sum_{k=1}^{m} (\mathbf{v}^{(k)} \cdot \mathbf{d} / |\mathbf{d}|)^2} \tag{21}$$

The deformation **d**, is obtained by superposing the two conformations and evaluating the differences in the C^{α}-coordinates. Kabsch's algorithm (Kabsch et al., 1990) is used for optimal superposition that eliminates rigid-body translations and rotations.

2.2 PERTURBATION ANALYSIS

Perturbation analysis is used here as a general term referring to the evaluation of the global effect of locally perturbing/altering the protein molecule. In the current study, we considered two such methods to investigate the allosteric pathway of Hsp70. Both methods are structure-based and involve an iterative procedure: we perturb one residue at a time and evaluate the effect on other residues.

2.2.1 Perturbation response scanning

Perturbation Response Scanning (PRS) method (Atilgan and Atilgan, 2009; Atilgan et al., 2010) is based on linear response theory (Ikeguchi et al., 2005). The goal is to investigate how the force applied on a single amino acid propagates through the entire structure. The protein structure is modeled as an elastic network with the $3N \times 3N$ Hessian matrix **H** as the force constant matrix (Atilgan et al., 2001), where N is the number of residues. Then the $3N$-dimensional force vector **F** exerted on the network model and the resulting displacement vector $\Delta\mathbf{R}$ can be expressed using the Hooke's law

$$\mathbf{F} = \mathbf{H}\Delta\mathbf{R} \tag{22}$$

Or equivalently

$$\Delta\mathbf{R} = \mathbf{H}^{-1}\mathbf{F} \tag{23}$$

Note that $\Delta\mathbf{R}$ is also a $3N$-dimensional vector that represents the displacement of each residue in the 3D-space. Suppose the force is only applied on residue i, i.e., $\mathbf{F}$ takes the following form

$$\mathbf{F}_i = (0\ 0\ 0\ \ldots \quad \ldots \quad , \tag{24}$$

then the corresponding displacement vector $\Delta\mathbf{R}^{(i)}$ represents the response of all the residues to the perturbation $\mathbf{F}_i$.

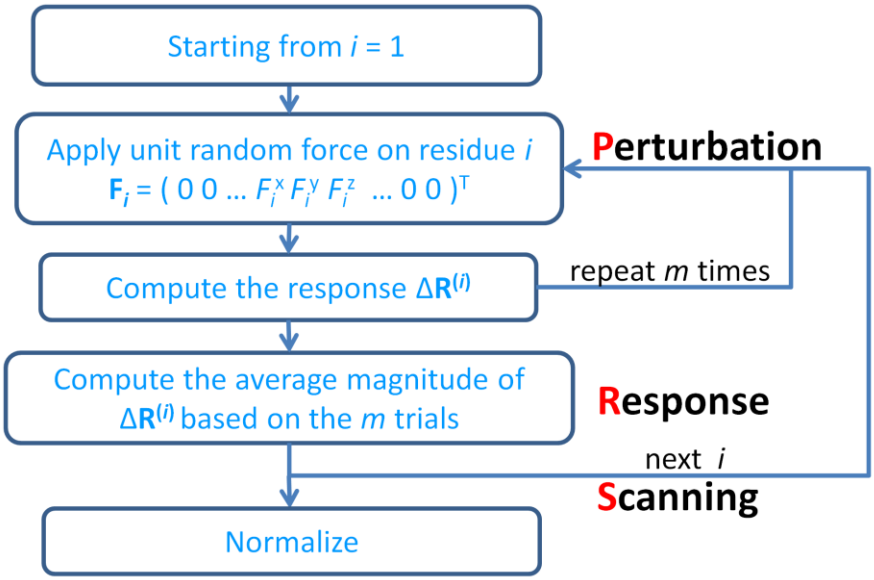

Figure 7. Protocol of perturbation–response scanning (PRS) methodology.

Figure 7 illustrates the workflow of PRS calculations in the current study. At the i^{th} iteration, a random force with unit magnitude is applied on residue i and the $3N$-dimensional displacement vector $\Delta\mathbf{R}^{(i)}$ elicited in all C^{α}-coordinates in response to the perturbation at residue i, $\Delta\mathbf{R}^{(i)} = (r_1^{\mathrm{x}}\ r_1^{\mathrm{y}}\ r_1^{\mathrm{z}}\ \ldots\ r_N^{\mathrm{x}}\ r_N^{\mathrm{y}}\ r_N^{\mathrm{z}})$ is calculated. To avoid the bias in a particular direction,

this procedure is repeated multiple times (e.g., m times). This way, a sphere centered at the C^{α} atom of residue i is almost uniformly sampled. Then given the obtained m response vectors, the average magnitude of the response of the k^{th} residue to a perturbation exerted at position i is

$$\left\langle \left\| \Delta \mathbf{R}_k^{(i)} \right\|^2 \right\rangle = \frac{1}{m} \sum_m \left[(r_k^x)^2 + (r_k^y)^2 + (r_k^z)^2 \right]_m \tag{25}$$

An $N \times N$ matrix is thus obtained by using $< \| \Delta \mathbf{R}_k^{(i)} \|^2 >$ in its ki^{th} element. This matrix is referred to as the PRS matrix $\mathbf{S}_{PRS}$.

The self-response of a residue, i.e., the displacement of a residue when itself is perturbed (the diagonal terms of $\mathbf{S}_{PRS}$) is closely related to its mean-square fluctuations (MSFs). Although this is an intrinsic property of the residue, it can conceal the actual "preference" of different residues in response to the same perturbation. To solve this problem, we normalized the PRS matrix by scaling the response of residue i with its self-response d_i

$$\bar{\mathbf{S}}_{PRS} = \begin{pmatrix} 1/d_1 & & 0 \\ & \ddots & \\ 0 & & 1/d_N \end{pmatrix} \mathbf{S}_{PRS} \tag{26}$$

where d_i is the i^{th} diagonal term of $\mathbf{S}_{PRS}$. Thus the diagonal of $\bar{\mathbf{S}}_{PRS}$ is all 1, and the i^{th} column in $\bar{\mathbf{S}}_{PRS}$ is referred to as the *response profile* generated upon perturbing residue i, denoted as $< \| \Delta \mathbf{R}^{(i)} \|^2 >_{norm}$.

2.2.2 Residue centrality

Residue centrality has its root in the small world network theory (Watts and Strogatz, 1998). The small world network exhibits a combination of high regularity and certain amount of

randomness. It has found application in a broad array of fields such as the social network and world wide web (Barabasi and Albert, 1999). In our study, we employed the method of residue centrality (del Sol et al., 2006), which exploits the small world network theory to investigate the residue's role in mediating the propagation of interactions. The central residues are those residues exhibiting a high probability of participating in shortest-path communication when all such paths between all residue pairs are examined. The protein is modeled as a network to this aim, each node representing a residue. The procedure of calculating the centrality of residue i is by removing it from the protein structure and calculating the average shortest path length of the remaining network, referred to as the characteristic path length L_i. The pairwise shortest path length between residue i and j is computed using Dijkstra's algorithm (Cormen et al., 2001), and the adjacency matrix **A** is defined as

$$(\mathbf{A})_{ij} = \begin{cases} d_{ij} & \text{if } d_{ij} < r_c \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

where d_{ij} is the shortest distance between any two atoms in residue i and j, and r_c is the cutoff distance for interatomic interactions, which is typically taken as 4.5Å. In previous studies, such contact-based network models have been pointed out to exhibit properties of small-world networks (Greene and Higman, 2003). The centrality of residue k is measured by the difference $\Delta L_k = L_{hyp}(k) - L$ in the characteristic path length with respect to the original network, obtained for the hypothetical (or perturbed) network where node k and all connected edges have been removed. If there is a significant increase in the characteristic path length due to removal of residue k, then residue k is considered to be a "central residue" in establishing internode communication. Central residues are hypothesized to play a role in allosteric signal transduction. Those residues with the highest centrality are considered to be most critical to facilitate the communications of the network.

The calculation of residue centrality is solely based on the geometrical properties of the structure and is highly sensitive to side chain orientations, which allows it to capture local interactions, but at the same time makes it less robust than ENM. Central residues are typically located near the active sites or at interdomain interfaces (Liu and Bahar, 2010).

2.3 SEQUENCE EVOLUTION

The main goal of sequence analysis in this study is to add yet another dimension to the information obtained from the 3D protein structure. The techniques used here for sequence analysis are well-established and have been successfully applied to a large number of systems in the literature.

We focused on extracting two types of information from the MSA: sequence conservation level of each residue and the co-evolution of residue pairs. Our approach is based on information theory, which adopts a probabilistic view of the evolutionary process such that sequence variation of a residue is modeled as a random variable. We also incorporated phylogenetic information embedded in the sequence alignment, which is reflected in the evolutionary trace method.

2.3.1 Information entropy and mutual information

Information entropy (Shannon, 1948) was originally proposed to measure the uncertainty of a probability distribution. It is defined for all types of random variables. In our application, it is sufficient to consider the simple case in which the random variable is discrete with finite sample

space. In a MSA, each sequence is considered as a sample, and each of the N columns (representing N residues) is considered as a discrete random variable that takes on one of the 21 amino acid types (gaps are treated as the 21st type) with some probability. Then the entropy of column i is defined as

$$S(i) = \sum_{x_i=1}^{21} P(x_i) \log \frac{1}{P(x_i)} \tag{28}$$

where $P(x_i)$ is the probability of observing amino acid type x_i at the i^{th} position. The widely used sequence logo representation is based on the entropy (Schneider and Stephens, 1990).

Mutual information (MI) (Cover and Thomas, 1991) can be considered as a generalization of entropy; it measures the level of dependence between two random variables. Using the same notation as above, we can calculate the MI between residues i and j using

$$I(i,j) = \sum_{x_i=1}^{21} \sum_{x_j=1}^{21} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \tag{29}$$

Here $P(x_i, x_j)$ is the joint probability of occurrence of amino acid types x_i and x_j at the i^{th} and j^{th} positions respectively, $P(x_i)$ and $P(x_j)$ are the corresponding singlet probabilities. Then an $N \times N$ MI matrix **I** corresponding to the examined MSA can be constructed with $I(i, j)$ being the ij^{th} element. The MI profile for each residue i (denoted as <I (i)>) is calculated by taking the average along the i^{th} row of the MI matrix.

MI is widely used for identifying correlated sites in proteins. Such correlations are usually inferred from the statistical analysis of pairwise amino acid substitutions among the members of the examined family of proteins. Because correlated substitutions are expected to occur between residue pairs directly interacting in the 3D structure, sequence covariance analysis, also referred to as correlated mutation analysis (CMA), has long been used for detecting inter-residue contacts within proteins (Eyal et al., 2007).

The CMA procedure consists of three steps, in general: (i) generation of MSA using homologous protein sequences; (ii) quantifying the covariance between different columns in MSA and (iii) identifying groups of highly covariant positions, also called clustering. The underlying assumption is that co-varying residues reflect essential structural/functional inter-residue couplings.

These techniques have some major limitations. The purpose of the method is to identify inter-residue couplings that are directly relevant to protein structure or function. However, the observed signals may not solely arise from such couplings. In fact sequence data are known to be noisy. A strong covariance may be detected among columns due to evolutionary signals that originate from early random mutation events. (Noivirt et al., 2005) have shown that the signal due to inter-residue interactions is comparable in magnitude to the noise caused by other stochastic evolutionary events.

2.3.2 Evolutionary trace

The evolutionary trace (ET) method (Lichtarge et al., 1996; Lichtarge and Sowa, 2002) identifies conserved residues within protein subfamilies. The procedure starts with the construction of a phylogenetic tree based on the MSA; in the present study the Fitch-Margoliash method (Fitch and Margoliash, 1967; Innis et al., 2000) is used to this aim. **Figure 8** illustrates the application of the procedure to the Hsp70 ATPase domain as an example (Liu et al., 2010). The method consists of three steps:

(1) The constructed phylogenetic tree is marked with multiple levels as indicated by the red vertical bars, where each level corresponds to a certain time point of the evolutionary clock.

(2) for each level,

(2.1) the marking line partitions the tree into different sub-trees; sequences within the same sub-tree are considered as belonging to the same subfamily, and are examined to see if each residue is conserved within the subfamily. The result is summarized as the "class consensus sequence" for each subfamily, as shown in the box in **Figure 8a**.

(2.2) The consensus sequences from different subfamilies at this level are cross-examined to identify fully conserved (across subfamilies) and class-specific or trace residues (conserved within but not across subfamilies), and the resultant ET sequence is written by the single-letter code of conserved amino acids and by the symbol 'X' for the trace residues, and the remaining residues as blank.

(3) The ET sequences generated at each level are organized in rows (**Figure 8b**). An ET rank (leftmost column) is assigned to each residue. A fully conserved residue is assigned the highest rank (rank of 1).

In practice, the conservation of a given residue in all subfamilies is a very strict condition when large sets of aligned sequences are considered. This limitation have been shown to restrict the applications of the ET method to MSAs of 100 and 200 sequences (Yao et al., 2003). To adapt the ET method and its variations to our dataset of >1,500 sequences, we relaxed the condition for defining an ET residue from conservation across all members in a given level to conservation in 90% of the members.

In contrast to information entropy which only considers single columns and treats them independently, the ET method has incorporated phylogenetic information based on the entire

MSA. It has also inspired subsequent methods for detecting conserved sites such as Consurf (Armon et al., 2001).

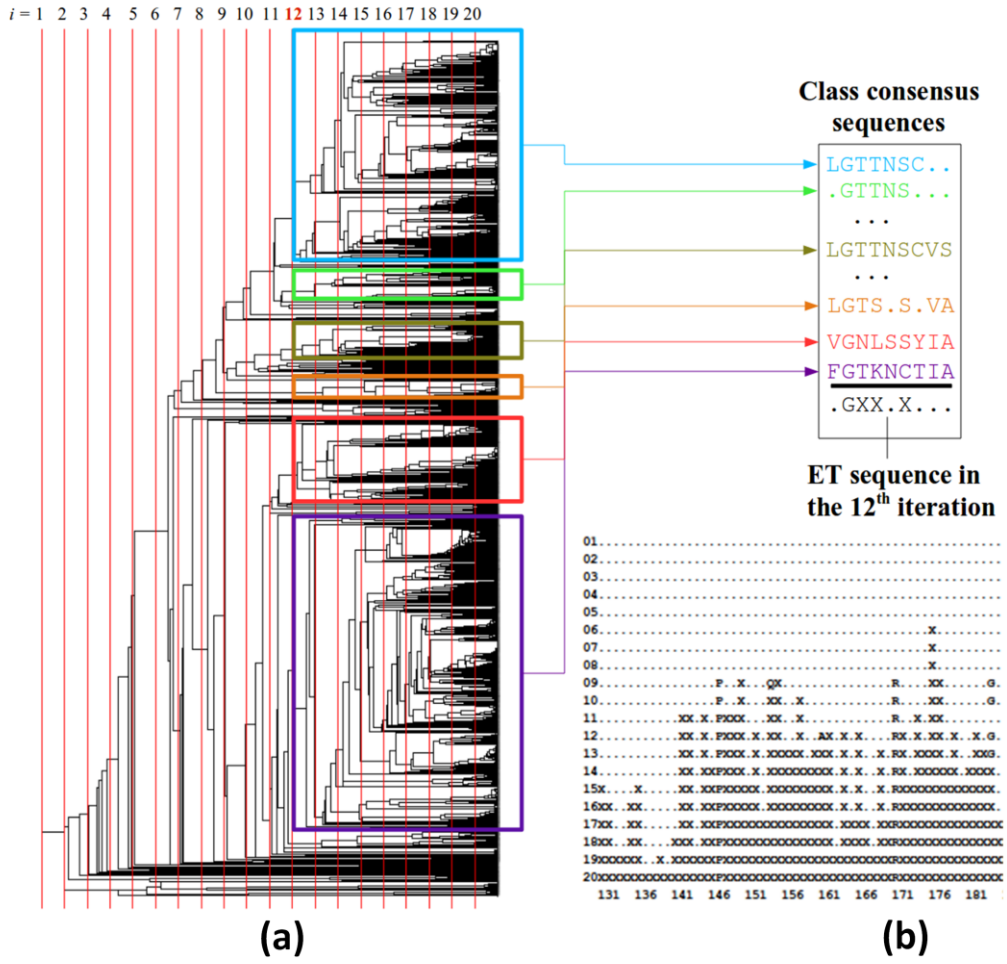

Figure 8. Schematic description of evolutionary trace (ET) method.

(a) The phylogenetic tree is constructed using the ET server and the set of 1627 ATPase domain sequences retrieved from Pfam database (DB) (Finn et al., 2008) for the Hsp70 family. Each vertical line corresponds to a given distance threshold. The boxes in different colors refer to the partitions obtained at the 12th distance threshold (also called level). Each box yields a different consensus sequence. The class consensus sequence for each partitioning level is then identified, as illustrated. Dots therein refer to positions that are sequentially variable between the members of the class. The ET sequence for the particular level is determined by assigning letter code X to all positions that are conserved within classes, but not conserved across classes. Those amino acids conserved across classes are indicated

by their single letter code (e.g., glycine G in the illustrated ET sequence). (b) A portion of the ET map is shown for a 20-level partitioning of the phylogenetic tree. Peaks indicate the most conserved residues, with their conservation level (or ET rank) indicated by the row numbers on the left.

2.4 SPECTRAL CLUSTERING

Spectral clustering (von Luxburg, 2007) has its theoretical root in spectral graph theory (Chung, 1997), to exploit the graph Laplacian and linear algebra theory to optimize its objective function. In this scenario. graph G is a similarity graph; i.e., the weight w_{ij} represents the similarity between node v_i and v_j. The general objective is to partition the nodes such that the similarity is high between nodes within the same partition/cluster and low across different partitions/clusters. There are several ways to define the objective function underlying this criterion, and accordingly several versions of algorithms have been proposed, yet the basic idea is the same. In our study, the algorithm based on normalized cut (Shi and Malik, 2000) is used.

We start with the simplest case where the nodes are partitioned into two clusters A and B. The normalized cut of this "conformation" of partitioning is defined as

$$\mathrm{Ncut}(A,B)=\frac{\mathrm{cut}(A,B)}{\mathrm{assoc}(A)}+\frac{\mathrm{cut}(A,B)}{\mathrm{assoc}(B)} \tag{30}$$

where cut(A, B) is the total weight of edges connecting the nodes in A and B,

$$\mathrm{cut}(A,B)=\sum_{v_i\in A,v_j\in B} w_{ij} \tag{31}$$

and assoc(A,V) is the total weight of connections from A to all nodes in the graph. Shi and Malik have derived an algorithm to approximately solve the optimization problem of minimizing

Ncut(A, B). By adopting a solution for the discrete clustering problem in a continuous space, the problem reduces to solving the generalized eigenvalue problem

$$(\mathbf{D}\text{-}\mathbf{W})\mathbf{y} = \lambda\mathbf{D}\mathbf{y} \tag{32}$$

where $\mathbf{W} = (w_{ij})$ is the matrix of the edge weights, also called similarity or affinity matrix, $\mathbf{D}$ is the diagonal matrix with elements $d_i = \Sigma_j\ w_{ij}$. λ and $\mathbf{y}$ are the generalized eigenvalues and eigenvectors of $\mathbf{W}$, respectively. The difference $\mathbf{D} - \mathbf{W}$, also called the Laplacian matrix, is symmetric and positive semi-definite (Chung, 1997). In order to partition a graph of N nodes into k clusters, we utilize the first k eigenvectors $\mathbf{y}_1, \ldots, \mathbf{y}_k$. For the particular case of bi-partitioning the graph (i.e. $k = 2$), $\mathbf{y}_2$ becomes the only eigenvector used as a criterion, since $\lambda_1 = 0$. In our application, each column in the MSA corresponds to a residue, which, in turn, is represented as a node in the graph. The degree of co-evolution of pairwise residues is considered as a metric of their similarity. Since we adopted mutual information to this end which is non-negative, it can be directly used as weight in the similarity graph, thus the MI matrix $\mathbf{I}$ replaces $\mathbf{W}$, and the graph (protein) is bi-partitioned based on the elements of $\mathbf{y}_2$.

We also performed k-way partitioning of the data in combination with k-means clustering. For each k we performed ten runs, and reported the results for the one with the minimum point-to-centroid distance sums.

3.0 ROLE OF HSP70 ATPASE DOMAIN INTRINSIC DYNAMICS AND SEQUENCE EVOLUTION IN ENABLING ITS FUNCTIONAL INTERACTIONS WITH NUCLEOTIDE EXCHANGE FACTORS

In the present study, we examine the interactions between the Hsp70 ATPase domain and different nucleotide exchange factors (NEFs), using sequence-, structure- and dynamics-based computations and identify their shared features. The Hsp70 ATPase domain is composed of four subdomains: IA and IB in lobe I, and, IIA and IIB in lobe II (**Figure 9a**). ATP binds the central cleft at the interface between subdomains IIA and IIB such that the geometric and energetic effects of its binding and hydrolysis are efficiently transmitted throughout the ATPase domain.

To date, four classes of NEFs have been identified: GrpE in prokaryotes (Harrison et al., 1997), and BAG-1 (Alberti et al., 2003), HspBP1 (McLellan et al., 2003) and Hsp110 (Andreasson et al., 2008) in eukaryotes. Their diverse 3D structures exhibit a variety of binding geometries and interfacial interactions with the Hsp70 ATPase domain. Our analysis provides insights into the generic and specific aspects of ATPase domain-NEF interactions, as well as the molecular machinery and sequence design principles of this highly versatile module, the Hsp70 ATPase domain, thus reconciling robust structure-encoded cooperative dynamics properties and highly correlated amino acid changes that enable specific recognition.

Here is a brief summary of the approach and rationale. First, we examine the structural properties of known Hsp70 ATPase domain-NEF complexes from different organisms to identify

the interfacial residues. Second, we analyze the intrinsic (structure-encoded) dynamics of the ATPase domain using the GNM, with an eye on the dynamic characteristics of the NEF-binding residues, on the one hand, and ATP/ADP-binding residues, on the other. A clear difference emerges between these two groups of functional residues: the former is distinguished by enhanced mobility in the softest modes while the latter is severely restricted. Third, calculations repeated with NEF-bound ATPase domains reveal how the open form of the ATPase domain is stabilized in order to facilitate ADP release, which is enabled by the intrinsic mobility of the NEF-binding regions. Nucleotide-binding sites, on the other hand, are shown to maintain their generic structure and dynamics irrespective of NEF binding, pointing to the robustness of the ATP-regulation by the ATPase domain. Fourth, detailed sequence analysis of Hsp70 family members reveals the distinctive sequence properties of the two regions: NEF-binding sites exhibit highly correlated mutations, consistent with the recognition of specific NEFs. Nucleotide-binding sites on the other hand, are almost fully conserved. In a sense, sequence variability is accompanied by conformation variability and vice versa.

Overall, Hsp70 ATPase domains appear to have been evolutionarily optimized to acquire a dual character: functional variability accompanied by structural variability at the co-chaperone binding sites and conservation/robustness both in terms of sequence and structural dynamics at the nucleotide-binding sites. This dual character is proposed to be essential for adapting to interactions with different co-factors while maintaining ATPase activity.

3.1 STRUCTURAL DYNAMICS

We examined the interface between the Hsp70 ATPase domain and the corresponding NEF in four structurally resolved complexes: with GrpE, BAG-1, HspBP1 or Sse1. We retrieved from the PDB structural data for HSP70 ATPase domains complexed with GrpE (PDB id: 1DKG (Harrison et al., 1997)), BAG-1 (PDB id: 1HX1 (Sondermann et al., 2001)), HspBP1 (PDB id: 1XQS (Shomura et al., 2005)), and Sse1 (Hsp110, PDB id: 3D2E (Polier et al., 2008)), shown in **Figure 9b-e**. Additionally, the structure of bovine Hsc70 ATPase domain resolved at 1.7 Å resolution (PDB id: 1HPM (Wilbanks and McKay, 1995)) has been used for the unbound form, and the PDB structure 1S3X (Sriram et al., 1997) of the human Hsp70 served as a template to reconstruct the lobe I missing in the complex with HspBP1 using the method described in the **Figure 10**.

Despite their structural differences, all four NEFs make contacts with subdomain IIB (**Figure 9b-e**). Subdomain IIB regions making contacts with NEFs include the α-helices 8 and 9, the double-stranded β-sheet E, and the loop connecting the two strands (**Figure 9a**). The NEF-contacting surface also includes small regions in subdomains IA and IB, but rarely IIA. The complete lists of Hsp70 ATPase domain residues that make contacts with each of the four NEFs are presented in **Tables 1** and **2**. **Table 1** is based on atom-atom interactions closer than 4Å separation. **Table 2** is based on the change in solvent-accessible surface areas (SASA), Δ(SASA), induced upon NEF binding. The entries in **Table 2** form a subset of those in **Table 1**, thus helping consolidate the identity of the NEF-binding residues on the Hsp70 ATPase domain. We note in particular Asn57, Arg258, Arg261, Arg262 and Tyr134 shared by both mammalian and bacterial chaperones in their NEF binding activity. **Table 2** also draws attention to the abundance of salt bridges at the mammalian Hsc70/NEF interfaces. In contrast, DnaK-GrpE

contacts are predominantly hydrophobic, consistent with previous observations (Sondermann et al., 2001).

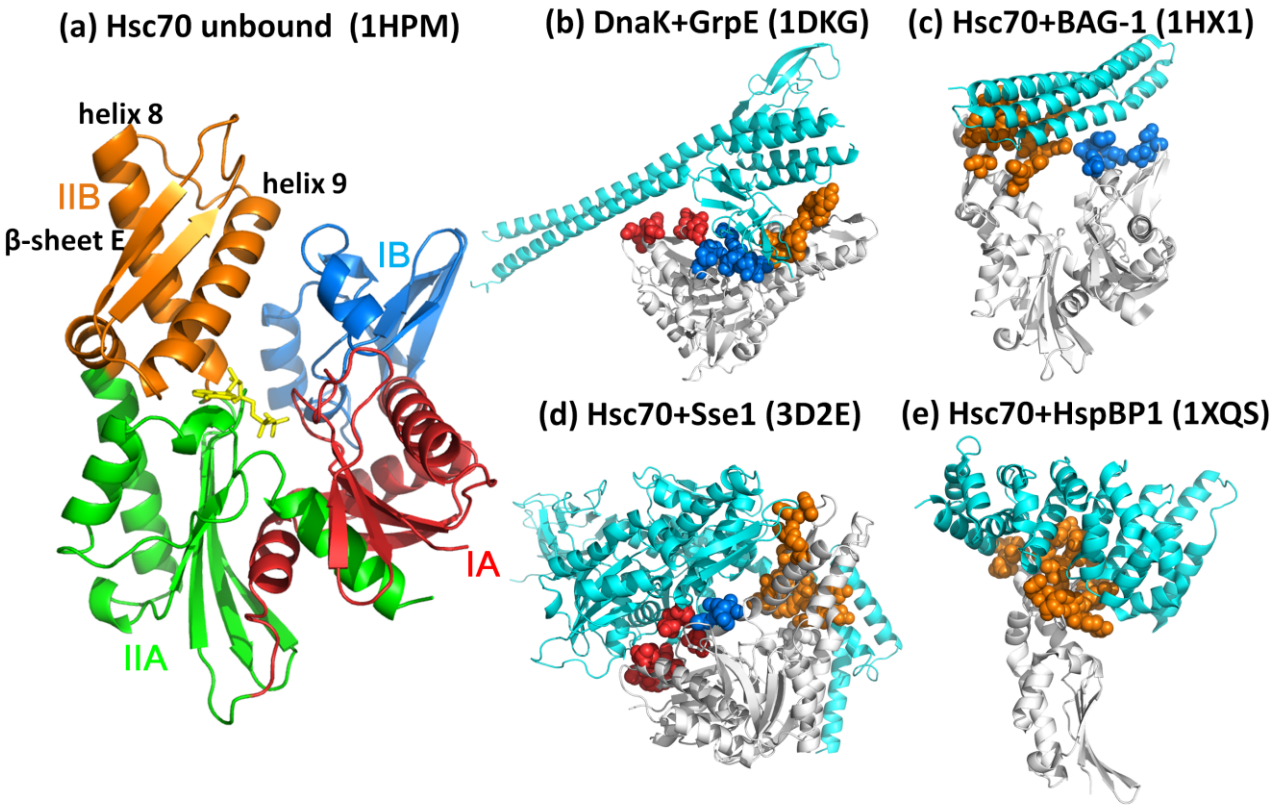

Figure 9. Structure of Hsp70 ATPase domain and its complexes with different nucleotide exchange factors (NEFs).

(a) ATPase domain structure colored by subdomains: IA (red; residues 1–39 and 116–188), IB (blue; residues 40–115), IIA (green; residues 189–228 and 307-C-terminus) and IIB (orange; residues 229–306). Several subdomain IIB residues are involved in NEF recognition and binding, including residues at the C-terminal part of helix 8 (G230-H249), the helix 9 (K257-S275), and the β-sheet E (strands Q279-I284 and F293-T298 connected by a long exposed loop). Residue identifications and secondary structure nomenclature are based on the PDB entry 1HPM. In yellow stick representation is a bound ADP. (b–e) Interactions with four different NEFs. (b) DnaK ATPase fragment from E. coli complexed with GrpE, (c) bovine Hsc70 complexed with BAG-1, (d) human Hsc70 with Sse1, and (e) human Hsc70 with HspBP1. In each case the NEF is colored cyan, ATPase fragment white, and interface residues,

shown in space-filling representation, are colored according to their subdomain locations. See **Table 1** and **Table 2** for more information on the examined complexes, and the identity of NEF-recognition residues in each case.

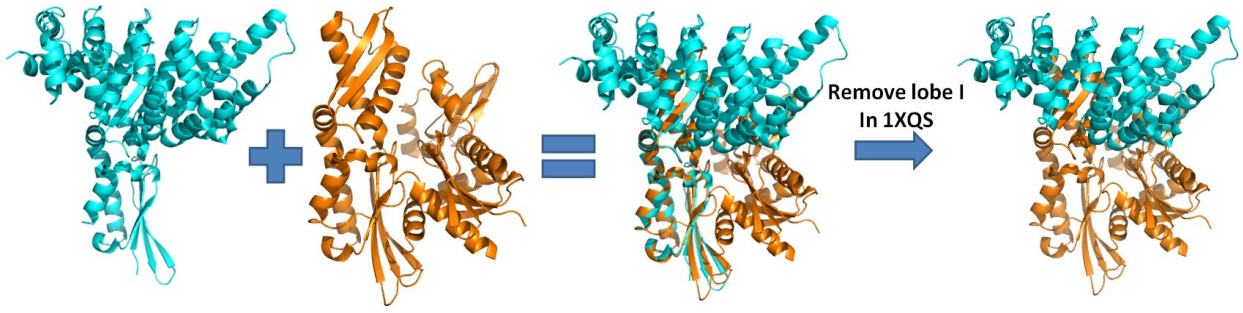

Figure 10. Reconstruction of the Hsp70 ATPase domain complexed with HspBP1.

1XQS is colored cyan, and 1S3X is colored orange. We used 1S3X as a "template" to replace lobe I in the original 1XQS structure. 1S3X was selected because of its highest sequence homology with respect to 1XQS (189 identical amino acids out of 190 in lobe I). The reconstruction procedure was based on the superposition of C^{α} atoms of the matching residues (i.e., lobe I) in 1S3X and 1XQS, since the reconstructed structure is solely intended for analysis with coarse-grained models. Optimal match between the C^{α} atoms was obtained using the C alpha Match webserver (Fischer et al., 1992). After 1S3X was optimally superimposed onto 1XQS, the original lobe I in 1XQS was replaced by the complete ATPase domain. Note that a closed-form was substituted for an open form in this reconstruction, which gave rise to steric clashes between a few atom pairs at the interface between the subdomain IB and HspBP1. These could be alleviated by minor side chain reorientations, and would have minimal, if any, effect on the slow modes obtained with the GNM, because the GNM analysis is based on the backbone topology as a whole (with little effect from inaccuracies in atomic coordinates that fall below the resolution of the model) or the distribution of C^{α}-C^{α} contacts within a cutoff distance of 7-10 Å, and the global modes are robust to minor changes in structural coordinates.

Table 1. Hsp70 ATPase domain residues making close atom-atom contacts with different NEFs [(a)]

PDB ID	***Molecule and Organism***		***NEF***	***NEF-Recognition/Binding Residues*[(b)]**		
				Subdomain IIB	**Subdomain IA**	**Subdomain IB**
1DKG	DnaK	*E.coli*	GrpE	**L257 (R258)**, **Q260 (R261)**, **R261 (R262)**, **E264 (T265)**, N282 (E283), P284 (D285), Y285 (S286) [(c)]	E28 (A30), E31 (Q33), E128 (E132), E129 (A133), **Y130 (Y134)**, L131 (L135), **G132 (G136)** [(c)]	L49 (L50), **P53 (A54)**, **R56 (N57)**, Q57 (Q58), **V59 (A60), T60 (M61)** [(c)]
1HX1	Hsc70	Bovine	BAG-1	**R258**, **R261**, **R262**, **T265**, R269, S281, **E283**, I284, D285, **S286**, G290, **D292**, Y294		F45, **D46**, **N57**, A60-N62
1XQS[(d)]	Hsc70	Human	HspBP1	**R247**, **K248**, K250, **R258**, **R262**, T265, **E268**, **R269**, **R272**, **T273**, **S277**, **Q279**, **S281**, R282, **E283**, **D285**, **D292**, **Y294**		
3D2E [(e)]	Hsc70	Human	Sse1 (Hsp110)	R262, T273, S276, T278, Q279, S281, E283, **D285**, S286, **T298**, R299, **A300**, R301, Glu303, **E304**	Q22, **H23**, K25, E27, D32-G34, R36, A133-**Y134**	**A54**, **N57**, Q58

(a) Close atom-atom contacts are defined as those having interatomic distance less than 4 Å.

(b) Amino acids are grouped according to their subdomain locations; those written in boldface are also detected by SASA calculations (**Table 2**) to exhibit a decrease in their accessible surface upon NEF binding.

(c) The entries in parentheses refer to the aligned residues in the mammalian Hsp70s

(d) The original structure of Hsc70-HspBP1 complex only contains lobe II.

(e) This complex contains four additional interfacial residues, all in subdomain IIA: Lys345, Lys348 and Asp352.

Table 2. Hsp70 ATPase domain residues making contact with different NEFs based on ΔSASA

GrpE	***BAG-1***	***HspBP1***	***Sse1***
Pro53 (Ala54)	*Asp46*	*Arg247*	His23
Arg56 (Asn57)	Asn57	*Lys248*	Ala54
Val59 (Ala60)	*Arg258*	*Arg258*	Asn57
Thr60 (Met61)	*Arg261*	*Arg262*	Tyr134
Tyr130 (Tyr134)	*Arg262*	Glu268	*Asp285*
Gly132 (Gly136)	Thr265	*Arg269*	Thr298
Leu257 (Arg258)	*Glu283*	Arg272	Ala300
Gln260 (Arg261)	Ser286	Thr273	Glu304
Arg261 (Arg262)	*Asp292*	Ser277	*Lys348*
Glu264 (Thr265)		Gln279	
		Ser281	
		Glu283	
		Asp285	
		Asp292	
		Tyr294	

(*) in parentheses are their counterparts in the mammalian Hsp70 ATPase domain.

The solvent accessibility surface area (SASA) of each residue of the Hsp70 NBD is calculated for both the NEF-bound and –unbound forms, using PyMol get_area function. All residues with Δ(SASA) < 0 are listed above. Note that all these residues are a subset of the residues listed in **Table 1**. Residue pairs that form salt bridges at the interface are written in italic. We note the abundance of such interactions in the mammalian chaperones/co-chaperone interfaces.

3.1.1 Intrinsic dynamics of the Hsp70 ATPase domain

Results from the GNM analysis of Hsp70 ATPase domain dynamics are presented in **Figure 11**. Panel a displays the mobility profile $M_i^{(1)}$ in the lowest frequency (global) mode of motion intrinsically favored by the overall ATPase domain architecture, calculated for the unbound Hsc70 ATPase domain (1HPM, (Wilbanks and McKay, 1995)). Subdomain IIB is distinguished by its enhanced mobility (see also the color-coded diagram in the inset of **Figure 11a**).

Interestingly, this region also contains the primary contact surface with NEFs. The symbols on the curve indicate the sequence positions of NEF-contacting residues identified using two methods: atom-atom contacts (blue open circles) and Δ(SASA) (red filled circles). In particular, Glu283, Asp285, Ser286, Asp292 and Tyr294 at β-sheet E loop form the highest peak in the mobility profile, succeeded by Arg247-Lys248 on helix 8 C-terminus, suggesting that these residues play a role in NEF recognition.

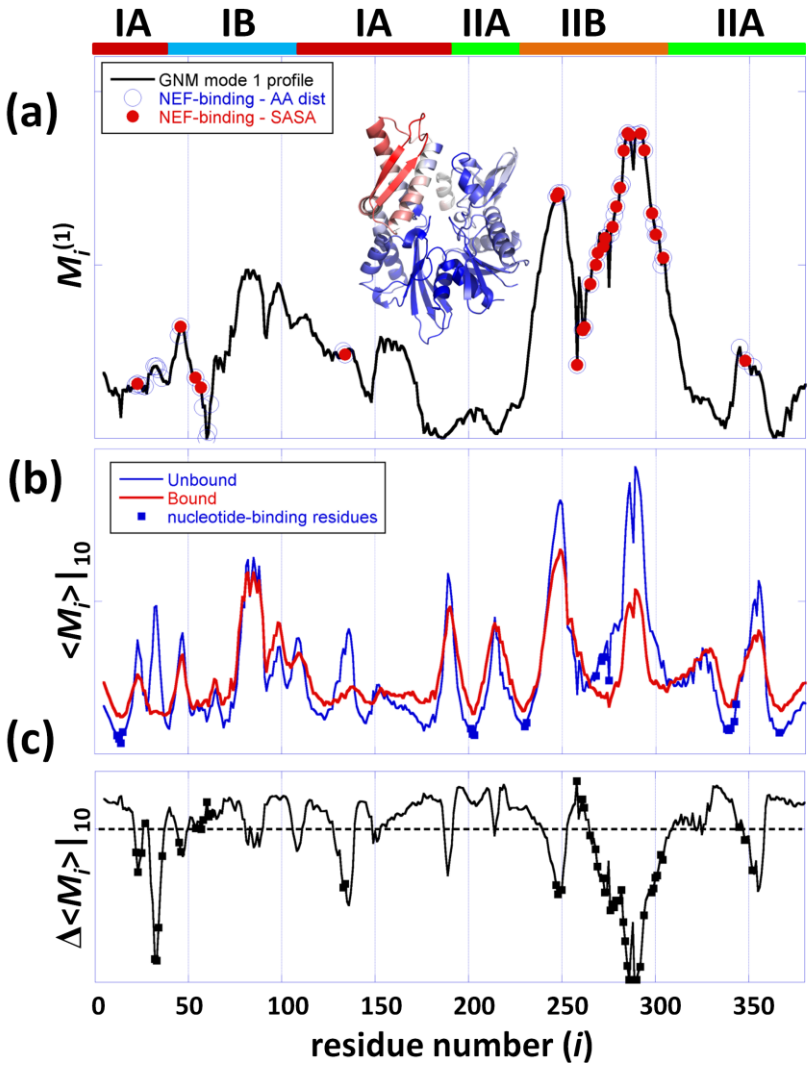

Figure 11. Intrinsic dynamics of the Hsp70 ATPase domain: high mobility of NEF-recognition sites in contrast to restricted mobility of nucleotide-binding residues.

(a) Distribution of residue mobilities $M_i^{(1)}$ in the global mode of motion calculated for the unbound Hsp70 ATPase domain. The horizontal bars on the upper abscissa indicate the ranges of the four subdomains IA, IB, IIA and IIB, colored as in **Figure 9a**. Subdomain IIB is distinguished by its enhanced mobility, with peaks at two regions: the C-terminal part of helix 8, and the β-hairpin loop. NEF-binding residues are indicated by the blue open circles (based on atom-atom distances) and red filled circles (based on ΔSASA). The diagram in the inset is color-coded to illustrate the global mobility profile (red: most mobile, blue: most rigid). (b) Weighted-average mobility profiles based on top-ranking ten GNM modes of motion, calculated using equation (12) for the unbound ATPase domain (blue) and for the NEF-bound structures (red), averaged over three mammalian complexes (**Table 1**). Nucleotide-binding residues (G12-Y15, G201-G203, G230, E231, E268, K271, R272, S275, G338- S340, R342, I343, D366) are indicated by filled squares. (c) Change in mobility between bound and unbound ATPase domain, obtained by taking the difference of two curves shown in panel b. The dashed line corresponds to the zero level. NEF-binding residues are marked by filled squares.

Thus, the subdomain that makes the majority of the contacts with the NEFs (i.e., subdomain IIB) is the one favored by the Hsp70 ATPase domain architecture to enjoy the largest mobility in the most cooperative mode of motion accessible to the ATPase domain. We also note, among NEF-contacting residues, a few exhibiting more restricted mobilities, located at the interface between subdomains IB and IIB, in particular. The tendency of biomolecules to involve their most mobile regions (peaks in the softest modes) in ligand recognition appears to be a design property noted in other applications; the ATPase domain subdomain IIB conforms to this rule. Its intrinsic mobility or conformational adaptability presumably allows for optimal interaction with the bound NEFs. On the other hand, final stabilization of a 'bound' conformer and communication of the conformational change locked upon substrate binding to other functional sites (e.g., nucleotide-binding site, in this case), may require the involvement of spatially constrained regions near the binding site (Yang and Bahar, 2005). The binding site thus

tends to exhibit a dual character, comprising both highly mobile residues that easily reconfigure for optimal binding and spatially constrained residues that efficiently communicate the structural change (from unbound to bound form) to other functional parts of the molecule (Luque and Freire, 2000; Lafont et al., 2007). NEF-binding residues His23 in subdomain IA, Asn57, Ala60 and Met61 in subdomain IB, and Arg258 and Arg261 in subdomain IIB presumably assume this allosteric communication role, as will be further clarified below.

The Hsp70 ATPase domain nucleotide-binding site, on the other hand, coincides with a rotationally flexible but spatially immobile global hinge region. These residues, indicated by the blue squares in **Figure 11b** (and listed in the caption), occupy regions that are severely constrained in the low frequency modes, i.e., they undergo minimal, if any, displacements in the collective movements of the entire domain. They participate in precisely tuned interactions at the global hinge region. The hinge region mediates the concerted movements of the subdomains, and as such the hinge residues need to remain in their key mechanical positions. Their lack of mobility, or displacement/translation in space, does not imply lack of rotational flexibility, however. On the contrary, in the same way as hinges operate, these residues are fixed in space, but have highly rotatable bonds that allow for the relative motions of the adjoining subdomains. Not surprisingly, this set has an abundance of glycines (G12, G201, G202, G203, G230, G338 and G339). The hinge bending role of these residues is critical to enabling the opening of the nucleotide binding pocket in response to NEF binding.

We also note among nucleotide-binding residues three charged residues, K271, R272 and R342, which were distinguished by their ‘central’ role in mediating the communication between the nucleotide-binding site and the other parts of the Hsp70 ATPase domain (for more details, see Chapter 4 and (Liu and Bahar, 2010)). Their central role was deduced from the small-world

network approach introduced by del Sol and coworkers (del Sol and O'Meara, 2005; del Sol et al., 2006).

The co-localization of chemically active sites with the global hinge region is another design feature consistent with previous observations reported for catalytic sites of enzymes (Yang and Bahar, 2005).

It will be shown below that the NEF-contacting and nucleotide-binding residues form two groups fundamentally different in terms of their evolutionary properties, in addition to their contrasting (highly mobile *vs*. highly constrained) dynamics in the global modes intrinsically accessible to the Hsp70 ATPase domain.

3.1.2 NEF binding suppresses the motions of subdomain IIB and stabilizes an open conformer

Figure 11b compares the mobility profiles obtained for Hsp70 ATPase domains in the NEF-free form (blue curve) with the average profile exhibited by three NEF-bound structures (with mammalian homologues 1HX1, 1XQS and 3D2E in **Table 1**). For clarity, the average over these three cases (red curve) is displayed in **Figure 11b,** and the individual curves for each complex may be seen in the **Figure 12**. In each case, the ten top (lowest frequency) modes are used to display the weight-averaged square displacements, which provide an accurate representation of the overall collective dynamics. The results indicate that the NEF-bound form of the Hsp70 ATPase domain closely maintains the intrinsic dynamics accessible to its unbound form, i.e., the loci of peaks and minima remain practically unchanged; however, binding of a NEF alters the relative (quantitative) distribution of mobilities: in particular, a reduction is observed in the mobility of subdomain IIB. As can be seen in more details for each of the four

complexes in **Figure 12**, the peak around the β-hairpin loop (residues 285-292) in subdomain IIB is almost completely depressed in the case of Sse1 and BAG-1, while GrpE and HspBP1 binding suppresses the mobility of the C-terminal end of helix 8 in the same subdomain.

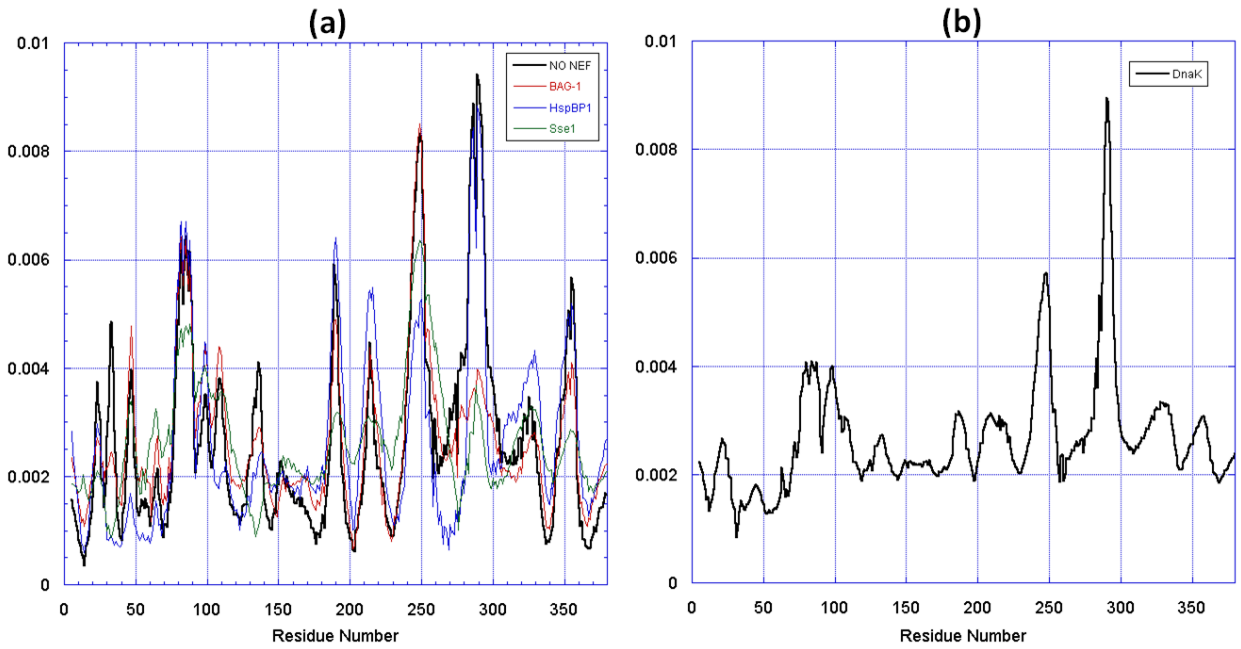

Figure 12. Intrinsic mobilities of residues in the ATPase domain.

(a) The profiles represent the GNM-predicted weighted average mobilities (squared) of all residues, as driven by the first ten slowest modes, calculated for the three structures of mammalian homologs of Hsp70 listed in the inset (see also rows 2–4 in **Table 1**). The profiles are normalized such that the area under each curve is 1. The thick black curve corresponds to the unbound form. (b) GNM-predicted weighted average mobilities of all residues, as driven by the first ten slowest modes, calculated for the structure of DnaK bound with GrpE.

Figure 11c shows the change in the mobility profile of the Hsp70 ATPase domain upon NEF binding. In addition to the suppressed motions at the β-hairpin, we also observe a drop in mobility at a number of NEF-contacting residues in subdomain IA (e.g., D32-G34). Notably, while NEF-binding residues on the ATPase domain experience reduced mobility upon NEF

binding, the NEFs themselves enjoy large conformational freedom, as illustrated in the **Figure 13**, as if their global fluctuations are conferred by the dissipation of those in the ATPase domain.

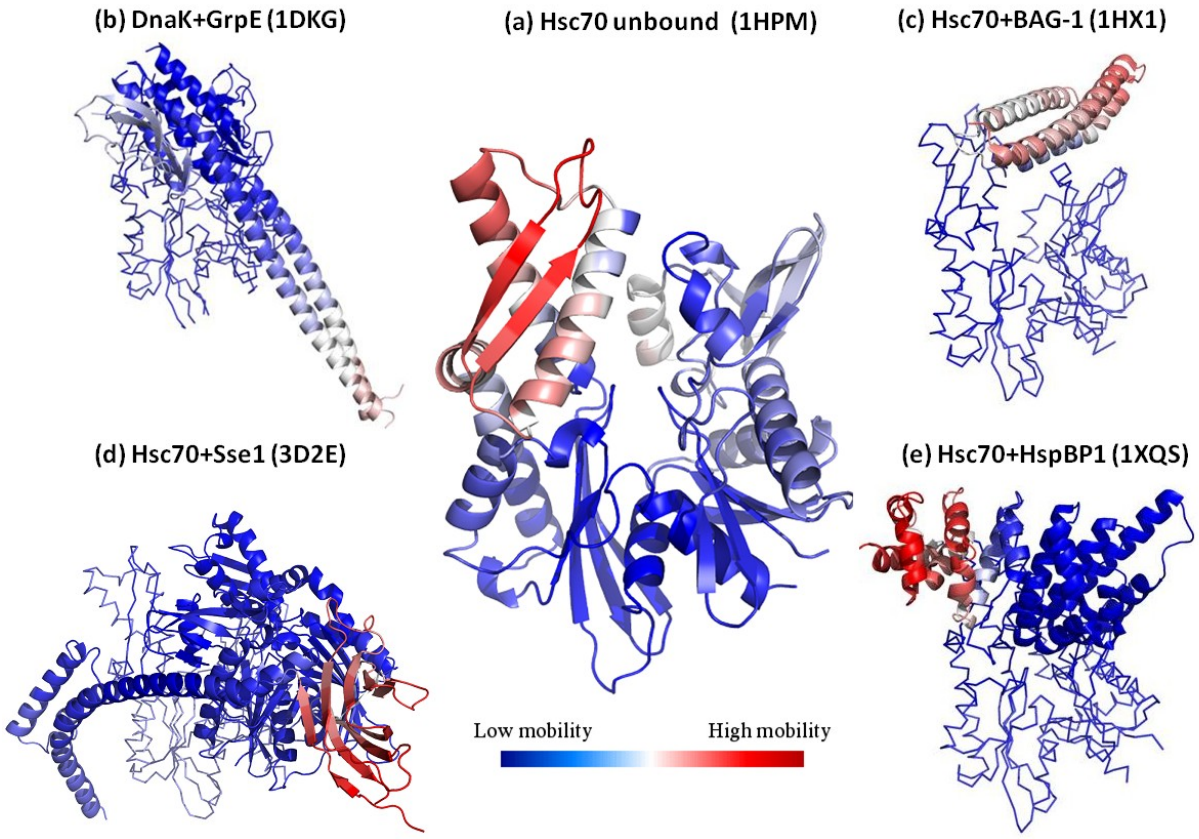

Figure 13. Softest mode of the ATPase domain in bound and unbound forms.

(a) Ribbon diagram of the ATPase domain in the unbound state color-coded by the mobilities in the first (lowest frequency, largest amplitude) GNM mode ($M_i^{(1)}$). Figure generated with PDB entry 1HPM. The slowest mode of the ATPase domain complexed with NEF is displayed for four different cases: (b) DnaK in contact with GrpE. (c) Hsc70 in contact with BAG-1. (d) Hsc70 in contact with Sse1. (e) Hsc70 in contact with HspBP1. Structural diagrams are generated with PDB entries (b) 1DKG (c) 1HX1 (d) 3D2E (e) 1XQS. The ATPase domain backbones are shown in stick representation, all in the same orientation, and the NEFs, as ribbon diagrams. In each case the complex is color-coded according to mobility (see the scale at the bottom).

By and large, all these observations support a common mechanism shared by all NEFs: they bind the most mobile subdomain of the Hsp70 ATPase domain - subdomain IIB, and in essence 'lock' the ATPase domain in a fixed conformation. This newly stabilized conformation is the 'open' form of the ATPase domain, as will be elaborated below. The stabilization of the open form is essential to facilitating nucleotide exchange, which is the co-chaperone activity of NEFs.

3.1.3 Induced *vs*. intrinsic dynamics

Comparison of the NEF-bound and –unbound structures of the Hsp70 ATPase domain shows that the distance between the subdomains IIB and IB is larger in the NEF-bound form. **Figure 14a** illustrates this 'opening' for BAG-1-bound ATPase domain. In this complex, subdomain IIB undergoes a rotation of 14° with respect to the rest of the structure (Sondermann et al., 2001), and in Sse1-bound ATPase domain, subdomain IIB is observed to rotate 27° sideways. The stabilization of an open conformer is a common feature in all NEF-bound structures, although they exhibit slight variations in the detailed geometry of the accompanying conformational changes (Sondermann et al., 2001; Shomura et al., 2005; Polier et al., 2008; Schuermann JP et al., 2008). By stabilizing the open form, NEFs assist in increasing the nucleotide exchange rate and communication with the SBD.

The observed conformational change of the Hsp70 ATPase domain may be explained by three possible scenarios: (i) induced upon NEF binding, (ii) a pre-existing equilibrium/path where the open form is already sampled, or can be readily reached via a soft mode, by the NEF-free ATPase domain, (iii) pre-existing equilibrium/path followed by induced fit. The former would be NEF-specific; the latter, would be intrinsic to the ATPase domain; and the third is an

intermediate behavior, i.e. the original recognition requires the pre-disposition of the suitable 'binding' conformation (pre-existing equilibrium); and binding of NEF induces further rearrangements to optimize the intermolecular interactions. For a more extensive discussion see for example a previous work of our lab (Tobi and Bahar, 2005). Clearly, in the case of intrinsically disordered proteins, folding upon binding is a common phenomena (Uversky et al., 2008; Wright and Dyson, 2009), in line with an induced fit. On the other hand, structural adaptability to increase substrate specificity would be explained by scenarios (ii) or (iii) (Tokuriki and Tawfik, 2009).

In order to examine quantitatively to what extent the observed reconfiguration is an intrinsic property of the Hsp70 ATPase domain (as opposed to a property induced by NEF), we focused on the softest motions predicted by the ANM. The black curve in **Figure 14b** displays the correlation cosine between the ANM modes ($\mathbf{v}(k)$, k=1-20) predicted to be intrinsically accessible to the unbound ATPase domain and the experimentally observed deformation (a $3N$-dimensional vector $\mathbf{d}$; see section 2.1.2 in Chapter 2) between the open and closed forms of the ATPase domain. Note that the complete space of equilibrium motions comprises a collection of $3N$-6 modes in the ANM, and by definition these form an orthonormal basis vector such that a cumulative overlap CO($3N$-6) = 1 is obtained by adding up all modes' contributions (see equation (21)). In the absence of correlations between the predicted modes and the experimentally observed changes, i.e., if the modes were completely random, their correlation cosine with $\mathbf{d}$ would therefore be $(3N\text{-}6)^{-1/2} = 0.029$, using N = 380. In contrast, a single mode alone (k = 3) is found here to exhibit a correlation cosine of 0.62 with the observed deformation, and the cumulative overlap reaches 86% by moving in the subspace spanned by 6 eigenvectors (modes) only (black curve with circles in **Figure 14b**). This result suggests that NEF binding

exploits to a large extent the reconfiguration accessible to the Hsp70 ATPase domain via this particular mode (mode 3) to drive the transition of the Hsp70 ATPase domain from its closed (NEF-free) state to an open (NEF-bound) state. Selection from a pre-existing path appears to be the dominant mechanism, although there is a minor contribution from higher modes selected via induced fit mechanism, in support of scenario (iii).

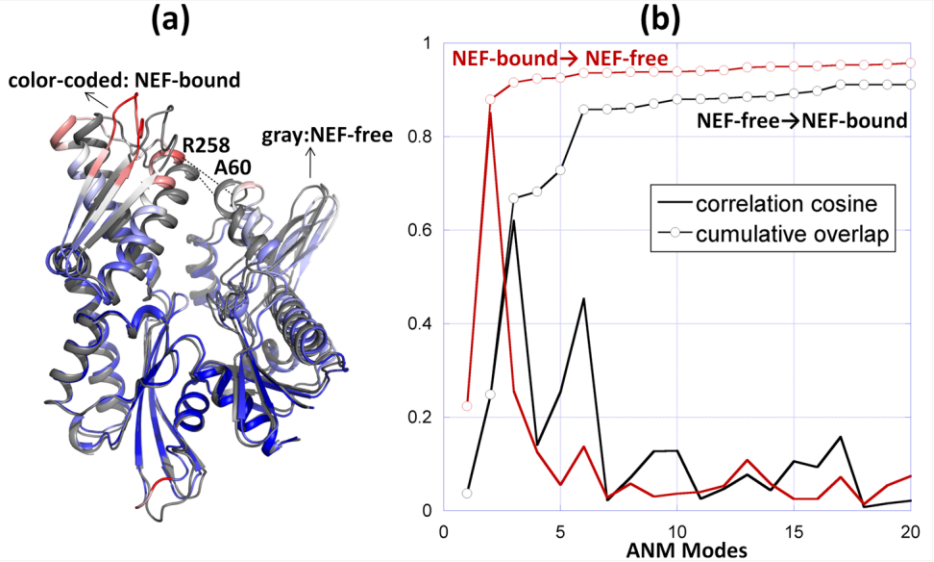

Figure 14. Comparison of experimentally observed and computationally predicted structural changes in the Hsp70 ATPase domain.

Experimental changes are illustrated for BAG-1-bound and free forms of the bovine Hsp70 ATPase domain (respective PDB ids: 1HX1 and 1HPM). Computational results are obtained by the ANM applied to the respective two structures. (a) Structural alignment of NEF-bound and unbound ATPase fragments. The unbound ATPase fragment (1HPM) is colored gray. The NEF-bound ATPase fragment (1HX1) is color-coded according to its extent of deformation with respect to the unbound ATPase, the regions showing the largest deformation being colored red, and those unchanged, blue. The distance between Ala60 and Arg258 C^{α}-atoms is 5.0 Å in the closed form and 10.9Å in the open form. Panel b displays the results for the unbound (black) and BAG-1-bound (red) ATPase domain. The solid curve represents the correlation cosine between the experimentally observed deformation vector **d** and the ANM modes 1–20 accessible to the ATPase domain (either NEF-bound or -free). The curve with circles describes

the cumulative overlap (equation (21)). A subset of 6 slow modes accessible to the unbound form ensures the passage to the NEF-bound conformer with an overlap of 0.86. The NEF-bound form exhibits an even stronger potential to be reconfigured back to its closed form, consistent with the preferred conformation of the ATPase domain in the absence of NEF binding: top ranking two modes yield a cumulative overlap of 0.88 with the experimental deformation **d**.

We further explored the transition between the open and closed forms of the Hsp70 ATPase domain by examining the reverse process, i.e., we examined the ability of the open form of the ATPase domain to restore its conformation back to the closed form in the absence of a NEF (red curves in **Figure 14b**). The results show that the intrinsic tendency to go back to the closed form is even stronger (than the tendency to open up). In fact, the 2^{nd} softest mode in this case exhibits a correlation cosine of 0.85 alone with the experimentally observed deformation d. Therefore, the movement along this single mode coordinate is practically sufficient to restore a significant portion of the conformational perturbation selectively stabilized by NEF. Calculations performed for different NEF-bound forms exhibited similar features. We conclude that the restoration of the NEF-free conformation after the dissociation of NEF is an intrinsic change almost exclusively favored by pre-existing one or two softest modes, in line with scenario (ii).

Notably, this type of intrinsic ability of the Hsp70 ATPase domain to undergo changes in its structure is consistent with the experimental observations made by Zuiderweg and co-workers (Bhattacharya et al., 2009). Zuiderweg and co-workers determined by NMR residual dipolar coupling measurements the ensemble of structures sampled in solution by the ATPase domain of DnaK from *Thermus thermophilus* in the ADP-bound state. Interestingly, the conformational variabilities observed in this ensemble, as noted by the authors, were found to be consistent with the structural change crystallographically observed (Sondermann et al., 2001) in the ATPase

domain upon NEF binding. This provides strong support, and experimental validation, for the intrinsic ability of the ATPase domain, in the absence of NEF, to have access to conformers that are pre-disposed to bind NEF, and for the utility of ANM analysis for accurately predicting the intrinsically favored changes in structure (softest modes).

3.2 SEQUENCE CONSERVATION

We began with 4,839 sequences retrieved from the Pfam DB 22.0 (Finn et al., 2008) for the Hsp70 family of molecular chaperones (Pfam id: PF00012). We refined the generated MSA by using the consensus sequence of the ATPase domain (380 residues) in the bovine cytosolic homolog of Hsp70 (Wilbanks and McKay, 1995). The refinement consists of three steps: (i) iterative implementation of Smith-Waterman algorithm (SW) for pairwise alignment (Smith and Waterman, 1981) using our consensus sequence, and elimination of those sequences below a threshold SW score (or less than 40% sequence identity) to retrieve the closest orthologs to human (Hsc70) and bacterial (DnaK) chaperones; (ii) deletion of MSA columns that correspond to insertions with respect to the consensus sequence, and (iii) removal of the sequences containing more than 10 gaps. These three steps resulted in a MSA of 1627 sequences with N = 380 columns (corresponding to residues 6 to 385 in Hsc70 ATPase domain), which has been subjected to evolutionary trace and mutual information analyses for detecting residue conservation and co-evolution patterns, respectively.

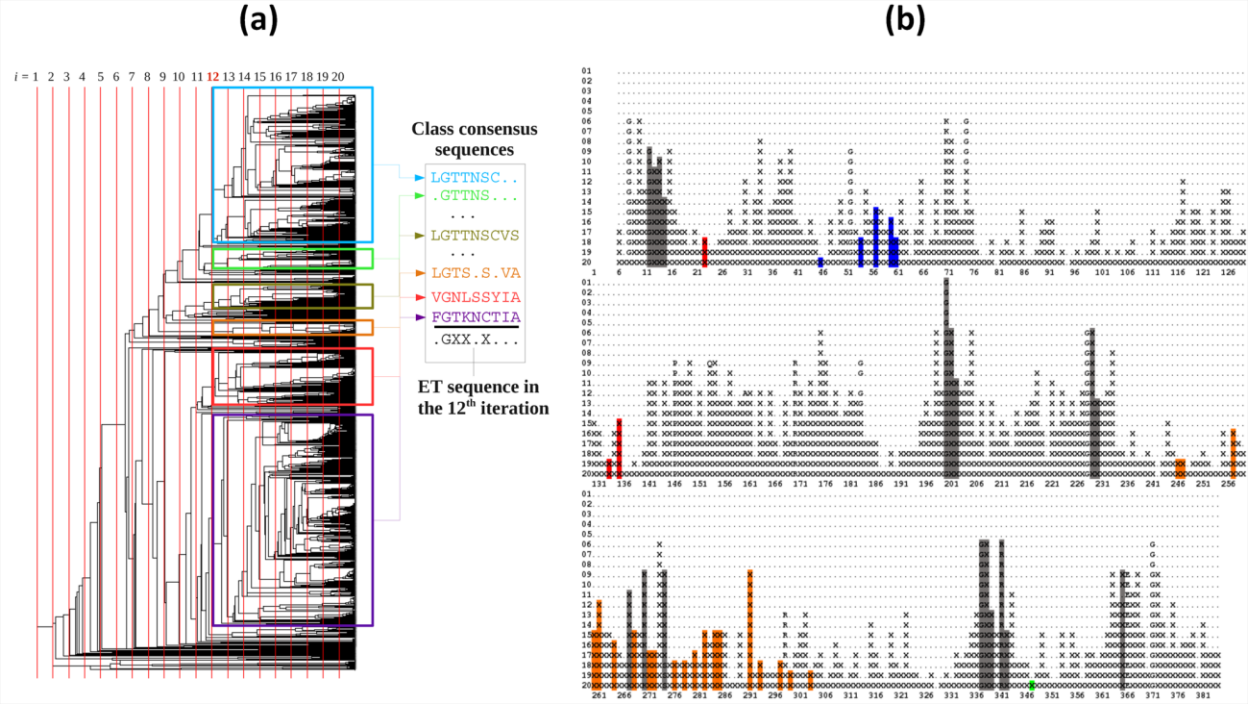

Figure 15. ET calculations for Hsp70 family.

(a) The phylogenetic tree is constructed using the ET server (Innis et al., 2000) and the set of 1627 ATPase domain sequences retrieved from Pfam DB for the Hsp70 family. Each vertical line corresponds to a given distance threshold. The boxes in different colors refer to the partitions obtained at the 12th distance threshold (also called level). Each box yields a different consensus sequence. The class consensus sequence for each partitioning level is then identified, as illustrated. Dots therein refer to positions that are sequentially variable between the members of the class. The ET sequence for the particular level is determined by assigning letter code X to all positions that are conserved within classes, but not conserved across classes. Those amino acids conserved across classes are indicated by their single letter code (e.g., glycine G in the illustrated ET sequence). (b) Results are shown for a 20-level partitioning of the phylogenetic tree. Peaks indicate the most conserved residues (among the 380 amino acids represented in each sequence), with their conservation level (or ET rank) indicated by the row numbers on the left. The columns highlighted in gray correspond to nucleotide binding residues. Those corresponding to the NEF binding residues are colored by the subdomains to which they belong (see **Figure 9a**).

3.2.1 Evolutionary trace analysis

An ET analysis highlights a cluster of conserved residues at the nucleotide-binding site. The results presented above lend strong support to the evolutionary selection/stabilization of a fold (by the Hsp70 ATPase domain) that endows suitable mobility and flexibilities at particular sites so as to favor functional changes in conformation (between open and closed forms), and optimal recognition and binding of the co-chaperones (NEFs). Next, we take a closer look at the evolutionary properties of Hsp70 ATPase domain sequences.

The results from ET analysis are presented in **Figure 15b**. Peaks therein represent the most conserved sites, within subfamilies (indicated by X), or across subfamilies (indicated by the single-letter amino acid code). The large majority, if not all, of the key residues reported in previous studies to be important to Hsp70 activity is captured by the ET peaks, including those participating in the hydrogen bond network proposed to form a proline switch (P147 and R155 in Hsc70; or their counterparts P143 and R151 in DnaK) (Vogel et al., 2006a). Residues known to coordinate the nucleotides are shown in gray shade. As expected, most of these residues are highly conserved. Among them, G201 exhibits ET rank 1, succeeded by G338 and R342, and then G12. We also note among the peaks K71 and E175, two residues identified in our previous studies to play a key role in ATPase domain allosteric communication (Liu and Bahar, 2010). Residues involved in NEF recognition and binding, on the other hand, are colored red, orange, blue and green depending on their subdomains. These residues exhibit low levels of conservation.

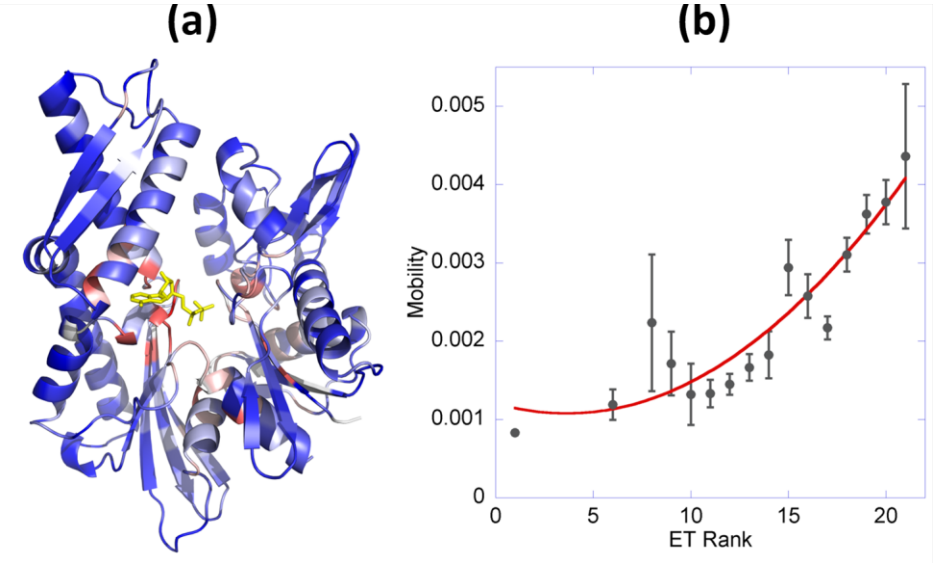

Figure 16. Correlation between residue mobility and its sequential variability.

(a) Ribbon diagram colored by the ET rank of residues, from red (most conserved) to blue (most variable). (b) The average mobility of residues corresponding to different ET ranks. The mobilities are evaluated using equation (12). The bars display the standard error for each ET rank. Best fitting second order polynomial (red curve) is shown to guide the eye (correlation coefficient of 0.92). See also Appendix Figure A1 panel b for a similar plot, based on ConSurf score (instead of ET rank).

3.2.2 Correlation between structural dynamics and sequence conservation

The color-coded ribbon diagram in **Figure 16** (based on the ET displayed in **Figure 15b**) shows that conserved residues (colored red) are mostly located in the nucleotide-binding pocket. The comparison of the weighted-average mobility profile in **Figure 11b** and the ET trace in **Figure 15b** suggests an inverse correlation between the extent of mobility of a given residue and its level of conservation: the ET trace indeed exhibits high peaks not only at nucleotide-binding sites, but also at other sites indicated by the GNM to participate in global hinge motion.

Towards a more detailed examination of this tendency, we have grouped residues based on their ET ranks, starting from the most conserved residues (ET rank = 1), and computed the average mobility profile of residues for each ET rank. **Figure 16b** displays the resulting relation between sequence conservation (ET rank) and global mobility. The ordinate represents the average displacement $\langle M|_{10} \rangle_{ET}$ for all residues that exhibit a given ET rank (abscissa), and the bars display the standard error in each case. The observed decrease in mobility with increased conservation suggests that constraints on the collective mechanics of the molecule may be as important as those associated with chemical activity, such that the residues at key mechanical sites also tend to be evolutionarily conserved.

A closer examination (Appendix Figure A1) shows that G34, D292 and L274 are outliers when comparing their ET rank with their global mobility (they are too mobile for their level of conservation). Their enhanced mobilities may however be explained by their functionalities: G34 is presumably critical to maintaining the loop structure near the nucleotide-binding site; D292 is a class-specific residue recognized to be a key element of the signature loop that differentiates subfamilies (Brehmer et al., 2001). It takes part in conserved salt-bridges with NEF basic residues in mammalian homologues (K238 for BAG; K245 for BP1). L274 is located at the C-terminus of helix 9 near the nucleotide-binding site, and may be playing a key role in stabilizing this long helix in a functional state. This helix indeed appears to be bridging between the NEF-contacting residues on subdomain IIB and the nucleotide-binding residues in the central cleft, hence its high conservation (or high ET rank). E268 and R272 are two other residues on helix 9 in contact with both NEF and nucleotide, and as such, they may be playing a role in initiating the allosteric communication between the bound NEF and the nucleotide-binding pocket. Like most of other nucleotide-binding residues, E268 is relatively conserved (ET rank = 11) (**Figure 15b**);

R272, on the other hand, is highly variable (ET rank of 17) and more exposed. Its high correlation with other NEF-contacting residues (discussed in the next section) and the orientation of its side chain (exposed) support its primary role in NEF recognition rather than nucleotide interaction.

As an additional verification of the relationship detected between collective mechanics and evolutionary conservation, we have examined the mobilities of residues as a function of their ConSurf scores. ConSurf scores provide a measure of the level of conservation, higher scores corresponding to less conserved residues (similar to ET rank) (Glaser et al., 2003). The plot in Appendix Figure A1 also confirms the relation between the extent of restrictions in mobility and the level of conservation, again suggesting that sequential and structural variabilities go hand in hand.

3.3 SEQUENCE CORRELATIONS

3.3.1 Co-evolutionary patterns for NEF-recognition residues

The results from the MI analysis of the 1627 Hsp70 ATPase domain sequences examined here are presented in **Figures 17** and **18**. The ribbon diagram in panel a highlights the residues distinguished by their co-evolutionary patterns. These are determined by analyzing the MI map for the complete sequence shown in **Figure 17b**. Close-up views of the two highlighted regions that contain the large majority of NEF-binding residues are presented in panels c and d. These two regions (residues 246-305 and 16-75) include 90% of all NEF-contacting residues. The bar plots below the MI map in panel b indicate the contribution of individual residues to the most

correlated pairs in the MI matrix (upper plot), and the frequency of NEF-ATPase domain contacts made by these residues in the examined three mammalian complexes (lower plot). The bar plots and enlarged panels c and d clearly show that NEF-contacting residues exhibit high sequence correlations. Residue pairs that exhibit the highest MI values are listed in **Table 3**.

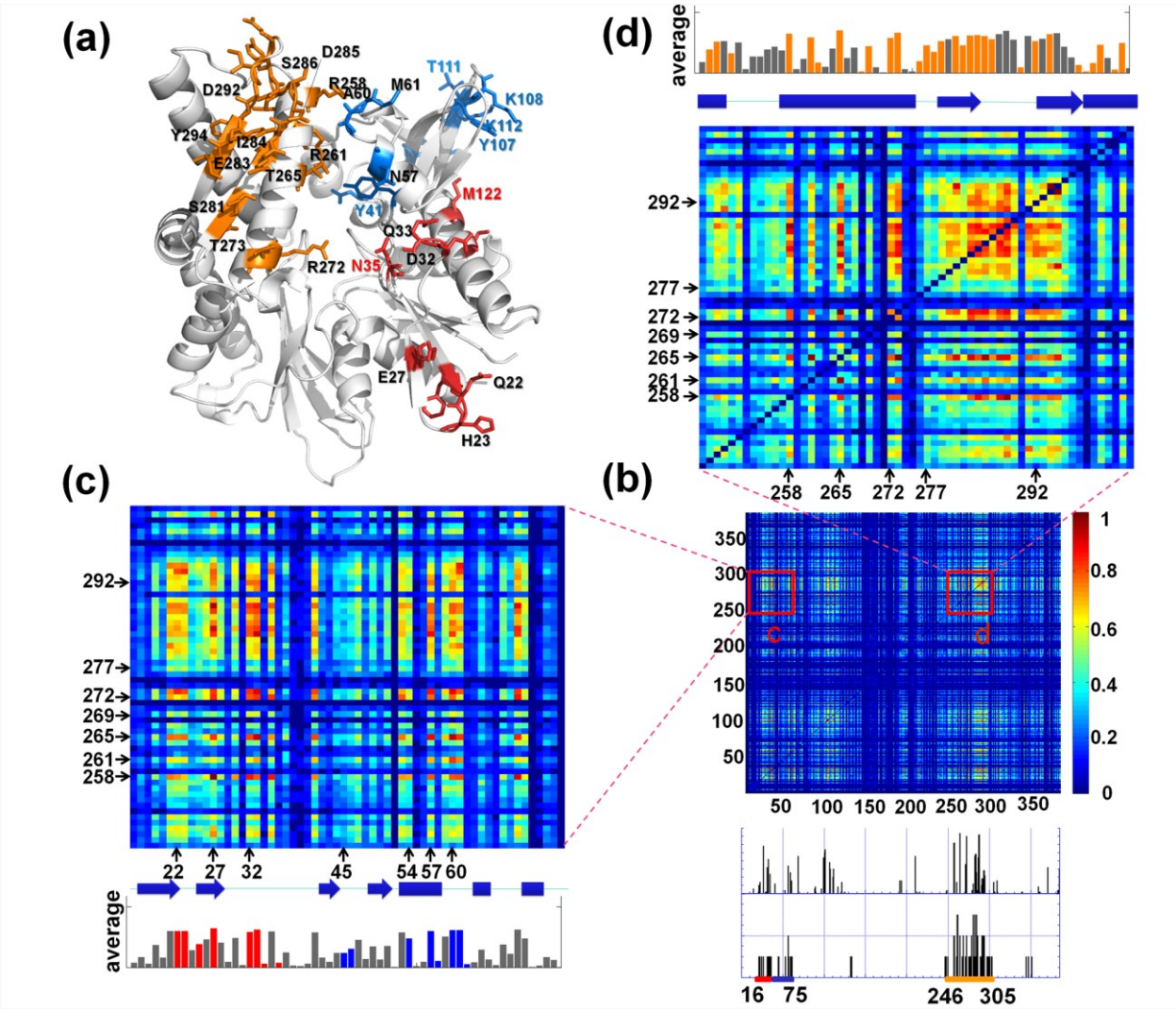

Figure 17. Co-evolution of NEF-binding residues.

(a) Amino acids distinguished by their high co-evolutionary patterns in the maps c and d (residues with average MI value greater than 0.32), shown in stick representation and colored by subdomains. Among them, the NEF contacting residues (**Table 1**) are labeled black, and others are colored by subdomain. Note the large proportion of charged or polar residues. (b) MI map for the ATPase domain sequence included in the MSA (residues 6–385). The

color-coded bar on the right indicates the level of correlation between the evolution of residue pairs. Two regions containing a large number of NEF-binding residues are enlarged in panels c and d. The bar plots under the map display the contribution of each residue to the most correlated residue pairs (top 1%, 720 pairs) in the MI matrix (upper plot), and the frequency of NEF-ATPase domain contacts in three mammalian complexes (lower plot). (c) and (d) Close-up views of the MI map portions between residues 246–305 (containing the helix 9 and β-sheet E of subdomain IIB) and residues 16–75 (containing NEF-contacting segments in subdomains IA and IB) (c), and within residues 246–305 (d). The corresponding secondary structural elements are indicated along the abscissa by cylinders (α-helices) and arrows (β-strands). The bar plots display the average MI per residue, NEF binding residues being colored by their subdomain.

Figure 18 provides a broader view of co-evolution patterns for Hsp70 ATPase structural elements. Here average MI values are displayed for all pairs of secondary structural elements and loop regions (panel **a**), and all pairs of subdomains (panel **b**). The strong co-evolutionary property of residues within the subdomain IIB is clearly seen from panel **b**, succeeded by that within subdomain IB. Among the inter-subdomain correlations, we distinguish the pair of subdomains IB and IIB. Consistent with these patterns, four groups of secondary structural elements are distinguished in panel a by their most correlated evolutions: sheet E and connecting loop (Q279-T298) in subdomain IIB, the β-strand R100-Y107, and loop (V59-N62) between helices 1 and 2 in subdomain IB, and a β-strand and preceding turn (His23-Ile28) in subdomain IA.

Table 3. Residue pairs distinguished by their sequence correlation (MI values above 0.8) in Hsp70 ATPase domain. (*)

Residue pair	*MI value*
Asp97---Lys102	1.010
Thr265---Thr273	0.909
Glu27---Arg258	0.885
Arg261---Thr265	0.878
Arg258---Tyr288	0.869
Thr265---Asp285	0.857
Lys102---Thr295	0.856
Thr265---Ser385	0.854
Pro101---Arg258	0.843
Glu27---Asp69	0.842
Glu27---Tyr288	0.835
Arg100---Lys102	0.832
Glu27---Pro101	0.830
Lys102---Lys108	0.829
Gln33---Thr265	0.828
Tyr107---Thr265	0.825
Asp69---Pro101	0.820
Ala60---Thr265	0.820
His23---Lys102	0.818
Lys102---Tyr107	0.817
Ser281---Thr295	0.814
Thr265---Glu283	0.812
Arg100---Tyr288	0.811
Arg100---Arg258	0.811
Gln33---Glu283	0.808
Thr265---Ser281	0.806
Thr273---Tyr288	0.805
Thr265---Tyr288	0.805
Lys102---Tyr294	0.803
Tyr107---Tyr288	0.803
Glu27---Arg100	0.802
Gln33---Thr273	0.801
Asp69---Arg258	0.801

(*) residue pairs separated by at least two amino acids along the sequence

Figure 17d describes in more detail the co-evolutionary properties within the subdomain IIB. The bar plots along the upper abscissa indicate the average MI values corresponding to each residue. The residues involved in NEF recognition are colored by the Hsp70 ATPase domain subdomain to which they belong. We note in particular the remarkably high MI values corresponding to the pairs of residues within the β-sheet E, except for the discontinuity at the loop residue G290. Furthermore, these residues display remarkably high co-evolutionary patterns

with amino acids on helix 9 (K257-S275). Examples of such highly correlated pairs are R258-Y288, T265-D285 and T273-Y288 (ranking 5th, 6th, and 27th respectively among all MI pairs, see **Table 3**). Notably, helix 9 also contains highly conserved residues (E268, K271 and S275) involved in nucleotide binding. This combination of co-evolving (NEF-recognition) and conserved (nucleotide-binding) residues endows helix 9 with a unique mediating role between the NEF-binding region and the nucleotide-binding pocket. Notably, the two β-strands and preceding α-helix on subdomain IIB emerge as a co-evolved structural entity, distinguished by its NEF-recognition and binding role, reminiscent of the functional 'sectors' pointed out by Ranganathan and coworkers (Halabi et al., 2009) for S1A serine proteases.

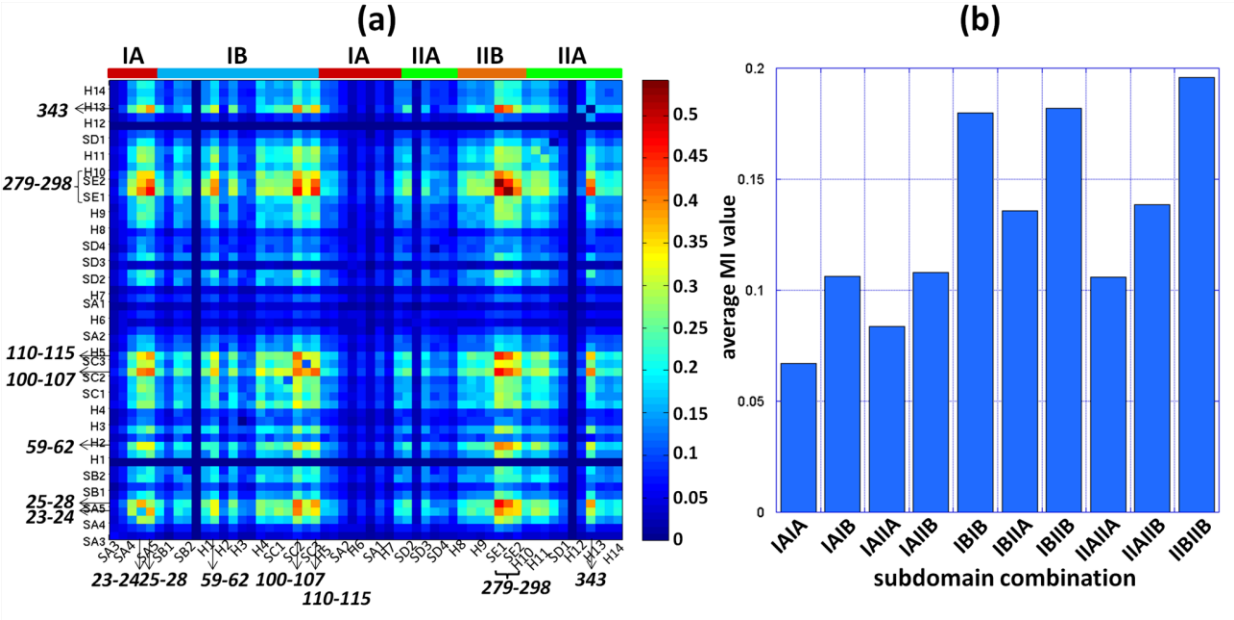

Figure 18. Average MI values calculated for different structural elements (helices/strands) and for different subdomains.

The panels demonstrate the mean MI value between and within (a) pairs of secondary structure elements (the names of helices and sheets are based on the PDB entry 1HPM, H: α-helix, S: β-sheet) and (b) pairs of subdomains.

Figure 17c reveals the cross-correlations between the evolutionary trends of subdomain IIB residues and the residues 16-75 on subdomains IA and IB. The above noted (β-hairpin and α-helix 9) residues of subdomain IIB appear to have co-evolved with well-defined residues on subdomains IA and IB. In particular the pairs E27-R258, E27-Y288, Q33-T273 and Q33-E283, exhibit remarkably high correlations (see **Table 3**), despite their long distance separation on the structure.

The observed sequence correlations may arise from several reasons (Atchley et al., 2000; Noivirt et al., 2005; Halabi et al., 2009) including those originating from the splits between subfamilies (phylogenetic noise). In particular, subdomain IIB has phylogenetic history contributing to its high variability, and subfamilies have evolved to partner with different NEFs. Regardless of the origin of these correlations, the MI map unambiguously shows that NEF-binding residues are distinguished by their co-evolutionary properties. As a further test, we performed a statistical coupling analysis (Lockless and Ranganathan, 1999) of sequence correlations (Appendix Figure A2) which confirmed the results found from MI analysis, while the signals provided by MI are more pronounced due to the weighting strategy employed in SCA.

3.3.2 Complementary information provided by MI maps and ET analysis

The ET diagram (**Figure 15b**) and MI maps (**Figures 17** and **18**) provide complementary information. Those residues distinguished by their high conservation (peaks in **Figure 15b**) cannot usually be detected by the MI map, simply because they exhibit minimal, if any, mutations and it may be hard to capture their co-evolutionary couplings to other residues due to scarcity of data. For example, subdomain IA is known to be relatively more conserved as also

confirmed by ET analysis (see the block of residues 140-185 belonging to this subdomain in **Figure 15b**), and the corresponding region in the MI map (**Figure 17b**) exhibit practically no signals indicative of correlated mutations. The less conserved subdomain IB, on the other hand, has several correlated residue pairs, including in particular those involved in NEF binding, which are furthermore correlated with the NEF binding residues on subdomain IIB.

Therefore, the sets of residues highlighted by these two analyses tend to be mutually exclusive, and involved in different roles, intrinsic (to ATPase domain, per se) vs. specific (to its interaction with different ligands/substrates such as NEFs). The structural regions where these two groups of residues are clustered and/or closely coupled (e.g., α-helix 9; see above) are suggested here to play a key role in reconciling the specific functions (e.g., NEF binding) of the Hsc70 ATPase domain with its intrinsic conserved properties (of nucleotide binding and ATP hydrolysis).

Yet, we note that in some cases some relatively conserved residues are also captured by their MI maps, because their (relatively infrequent but possible) mutations indeed require compensating mutations that can be detected, even if such mutations are rare. NEF binding R262 and D292 (with respective ET rank of 12 and 9) belong to this group of residues, and can sustain mutations provided that these are accompanied by compensating substitutions. As mentioned above R292 is a class-specific residue involved in salt-bridge formation with NEF residues, and likewise, R262 takes part in conserved interactions with acidic residues on NEF in mammalian homologues (D222 for BAG; E132 for BP1). Note that its counterpart in DnaK (R261) makes a contact with M174. This can be explained by the fact that binding of GrpE to DnaK is based on hydrophobic interactions instead of salt bridges (Sondermann et al., 2001).

3.4 DISCUSSION

3.4.1 Interplay between structure-encoded global dynamics and sequence-specific local interactions

Organisms comply with the evolutionary pressure to maintain their phenotype by genotypic variations that are compensated or correlated as needed, conserving certain sequence fragments vital to preserving their functions (Nowak et al., 1997). Understanding the co-evolving and conserved sequence patterns in modular domains is an interesting problem in its own right (Livingstone and Barton, 1993; Olmea et al., 1999). Understanding these patterns in the light of structural data, if available, provides us with further insights into shared mechanisms of interactions that form the molecular basis of the biological function of such modular domains. The Hsp70 ATPase domain is such a modular protein common to functionally diverse actin, hexokinase, and Hsp70 protein families (Bork et al., 1992). The present combined analysis of structure-encoded dynamics and sequence evolution for Hsp70 ATPase domain discloses a subtle interplay between conserved interactions and those involving co-evolved residues. Conserved interactions define generic properties of the Hsp70 ATPase domain: these include the concerted dynamics of its four subdomains, which allow for sampling functional conformations (e.g., that stabilized upon NEF binding, allowing for ADP release; shown in **Figure 15**), and the physicochemical events (ATP hydrolysis) at the nucleotide-binding site. Those residues involved in NEF recognition, on the other hand, show low-to-moderate conservation, but exhibit a remarkably high tendency to co-evolve, or undergo correlated mutations, again to achieve specific NEF-dependent recognition and binding activities.

An observation of interest is the similarity between the interactions of the Hsp70 ATPase domain with different NEFs, in terms of structural dynamics. While Hsp70 ATPase domains are highly conserved both sequentially and structurally, the four NEFs examined have distinct structures and consequently different dynamics. The key point is that their binding to the ATPase domain involves in all cases the subdomain IIB of the ATPase domain, although not in exactly the same arrangement. Their binding to a common interfacial region on the ATPase domain point to a shared mechanism of interaction: The ATPase subdomain IIB is originally distinguished by its high mobility in the slowest mode, especially at the β-sheet E and the exposed loop connecting the two strands of this sheet; and after NEF binding, there is a significant suppression in its mobility. The conserved dynamics of the complexes suggests a role of subdomain IIB as an "adjustable handle", which regulates the Hsp70 chaperone machine, to facilitate other proteins making use of its SBD.

Many applications using the ANM have shown that the substrate recognition involves a region distinguished by its enhanced mobility in the most cooperative (or softest) modes, which enables the molecule to optimize its interactions with the substrate. Here we can see that the C-terminal part of helix 8 and the loop of β-hairpin E enjoy this type of high mobility/adaptability. On the other hand, substrate 'binding' may also involve more constrained residues in the close neighborhood, which may play a role in transmitting allosteric effects. In the opposite case of a binding site composed exclusively of floppy residues, the structural changes induced upon substrate binding could dissipate locally and not efficiently transmitted. In this respect, we propose that the involvement of residues such as Arg258, Arg261 and Arg262 in subdomain IIB, or N57, A60 and M61 in subdomains IB is critically important in establishing the

communication between subdomains and transmitting allosteric signals between NEF-binding and nucleotide binding sites.

A putative communication pathway that couples distant residues in different subdomains of the Hsp70 ATPase domain is suggested here by the structural mapping of correlated and conserved residues, which needs to be further established. **Figure 17a** displays those residues identified to be co-evolving. Notably, we observe several pairs making interdomain contacts, in addition to spatially distant residue pairs (e.g. H23 in subdomain IA and N57, A60 and M61 in subdomain IB correlated with R258, R261, E283 and D292 in subdomain IIB). In a recent study, R272, R261, Y15 and Y41 have been identified to play a central role in establishing the allosteric communication in the unbound Hsp70 ATPase domain, along with highly conserved residues K71, R72, E175 and H227 (Liu and Bahar, 2010). It remains to be seen if these central residues play a key role in mediating between these co-evolving, spatially distant residues. We also note that Smock et al. recently identified a sparse but structurally contiguous group of co-evolving residues at the interface between the ATPase domain and the SBD in Hsp70/110 protein family, which has been proposed to underlie the inter-domain allosteric coupling (Smock et al., 2010), in support of the role of co-evolved residues in mediating allosteric signaling.

3.4.2 Pre-existing paths of reconfiguration intrinsic to Hsp70 ATPase domain and their role in accommodating co-chaperones binding

Many recent studies have pointed out the validity of "pre-existing equilibrium" concept where a substrate or ligand simply selects from amongst an ensemble of conformations already accessible to the protein prior to binding (Tobi and Bahar, 2005; Swain and Gierasch, 2006; Henzler-Wildman et al., 2007; Bahar et al., 2007; Lange et al., 2008; Bakan and Bahar, 2009; Smock and

Gierasch, 2009). The present results, and recent applications of ENMs, suggest that more important than the pre-existence of these 'states', is the existence of energetically accessible 'paths' that provide access to those states, or the intrinsic tendency of the native structure to reconfigure towards such functional states. In terms of energy landscape description, what is needed is not the existence of multiple minima, the depths of which change upon ligand or substrate binding, but the existence of one or more directions of reconfigurations, or paths along the energy landscape, that are easily accessible to the protein and lead to the targeted (functional) conformer. The softest modes provide such paths. They define directions of motion in the space of collective coordinates, which incur a minimal energy ascent as the molecule moves away from its original energy minimum. They also present the best mechanisms of dissipating energy, if the system is perturbed. These are the modes that are being exploited when proteins bind ligands or substrates. Notably these functional conformations accessible near the native state can be observed by NMR residual dipolar coupling, as shown for Hsp70 ATPase domain by Zuiderweg and coworkers (Bhattacharya et al., 2009). **Figure 14** clearly shows that movements along a handful of modes satisfactorily ensures the passage to the alternative (functional) open form, and that the open form itself has a strong tendency to restore its conformation back to the closed form, in the absence of NEF.

3.4.3 Bridging between residue conservation and global dynamics

Protein-ligand binding interfaces and protein-protein contact interfaces are characterized by different sequence variation patterns. The protein-protein contact interfaces usually expose larger contact areas (James et al., 2003) and exhibit high mutation rates. Moreover, if the contact interface is a common recognition site for multiple targets (possibly in different organisms), co-

evolution is likely to occur among the binding residues to preserve specific interactions and conformations at the sequence motif. On the other hand, the protein-ligand interface is usually buried in the folded core of the protein; in contrast to protein-protein interaction, the protein-ligand interaction is usually characterized by higher specificity, requiring sequence conservation (Lichtarge et al., 1996; Lichtarge and Sowa, 2002).

The Hsp70 ATPase domain exhibits patterns in close agreement with these general features: Its ligand (nucleotide) binding site essentially consists of highly conserved residues, which not only precisely coordinate the ligand, but also take part in a global hinge-bending region so that they are both chemically and mechanically required to be highly conserved. NEF recognition sites, on the other hand, exhibit much lower conservation properties; and in addition to their sequence variability, the subdomain IIB, which is observed to be most often involved in NEF binding, enjoys enhanced mobility. Briefly, global dynamics requirements entail residue conservation, and specific recognition entails sequence variation along with enhanced mobility. However, neither the sequence variability, nor the conformational mobility at NEF recognition sites, is random. The sequence variability takes place under unique restrictions, compensating mutations, as unraveled by the MI map. Conformational variability, on the other hand, is uniquely defined by the ATPase architecture, and precisely adept to accommodate the passage to the functional open state that is stabilized upon NEF binding. The ATPase domain uniquely juxtaposes such structure-encoded dynamics and sequence-specific interactions, which underlie its ubiquitous activities.

In general, subdomains IA and IIA are more conserved and more rigid than subdomains IB and IIB (Flaherty et al., 1990), as also indicated by the ET in **Figure 15b**; notably, they also serve as binding site to a number of proteins. For example, subdomain IA accounts for the

binding of J-domain proteins (Jiang et al., 2007); subdomain IIA is reported to contain a putative binding site near its interface with subdomain IA (V189-V195) to the chaperonin-containing TCP-1 (Cuellar et al., 2008), and it is connected to the SBD by an inter-domain linker, which is considered important for the allosteric interactions between the two domains (Vogel et al., 2006b; Swain et al., 2007). It remains to be seen if the correlated sites on Hsp70 ATPase domain emerging from the MI analysis play a role in the functional communication with other co-chaperones or the SBD. Extensive experimental studies have been performed to date with the *E. coli* Hsp70, DnaK, to understand the molecular mechanism of activity of the molecular chaperones in the Hsp70 family. The analysis in the present paper will guide our interpretation of the NMR, FRET, and EPR data on different states accessible to DnaK. Each of these methods gives us a different window into the ensemble of conformational states populated in response to ATP, ADP and NEFs. Excitingly, a detailed chemical shift analysis of six different ligand bound states for the nucleotide-binding domain of DnaK, with and without the linker that connects it to the substrate-binding domain (i.e., 12 NMR samples compared pairwise and as a group) has pointed to the same subdomain interface rearrangements indicated in the present study (Zhuravleva & Gierasch, in preparation). Moreover, the NMR results point to the fundamental feature that subdomain IIB can undergo a hinge-like movement to enable nucleotide entry and release. It is this fundamental movement, intrinsic to Hsp70 ATPase domains, that different NEFs have exploited. They bind in different, sequence-specific ways, but modulate the same fundamental movement. Further detailed analysis of the ensemble distributions and rates of interconversion between states can be achieved using a synergistic battery of computational and experimental tools.

4.0 HSP70 ALLOSTERIC PATHWAY IDENTIFICATION USING PERTURBATION ANALYSIS

The allosteric communication between the two domains, ATPase domain and SBD, of Hsp70 is essential to Hsp70 functioning as a molecular chaperone; understanding the structural/dynamical basis of this allosteric communication is critical to rational design of Hsp70 inhibitors. To this end, we explored the key residues involved in this process, and identified putative pathways of allosteric communication in the molecule using the results from structural dynamics and sequence co-evolution analyses.

Our study consists of two parts, both based on perturbation analysis. Part I concentrates on the Hsp70 ATPase domain, and uses residue centrality concepts to identify the key residues that establish network communication. The 2nd part focuses on the communication between the two domains, using a homology model constructed for the ATP-bound state, and identifies key residues using PRS methodology.

4.1 PART I: PATHWAYS IN THE HSP70 ATPASE DOMAIN

In this part, we examined the type of conformational changes occurring in the ATPase domain, and their influence on inter-residue communication pathways. Our group's previous examination of another ATP-regulated allosteric machine, the bacterial chaperonin GroEL, showed that the

structure has access to intrinsically favored collective dynamics, on the one hand, and to well-defined signal transduction pathways that transmit allosteric effects away from the ATP binding site, on the other (Chennubhotla and Bahar, 2006). Redistribution of on-pathway interactions during the most cooperative (global) modes of motion of the chaperonin has been proposed to be a mechanism of allosteric regulation (Chennubhotla et al., 2008). Toward gaining insights into the dynamic aspects of allosteric regulation, this time in the Hsp70 ATPase domain, we adopted here a multi-pronged approach: First, we identified a number of key residues distinguished by their central role in so far as the allosteric signal transduction across the molecule is concerned. A number of residues lining the cleft between the two lobes of the ATPase domain appear to modulate the opening and closing of the cleft. Second, we analyzed the sequence conservation and co-evolution patterns of these residues. Third, we examined their collective dynamics using GNM.

We used the structures of the bovine homolog of Hsp70 (Hsc70) (PDB id: 1HPM (Wilbanks and McKay, 1995)) for the closed form of the ATPase domain. For the open form, we considered two structures of the same species complexed with mammalian NEFs: a complex with BAG-1, and another with Sse1, with respective PDB identifiers of 1HX1 (Sondermann et al., 2001) and 3C7N (Schuermann JP et al., 2008). The structural alignment in **Figure 19** shows that there is a global change in the relative positions of subdomains IB and IIB, as the structure undergoes a conformational change between closed and open forms.

Figure 19. Superposition of the closed and open conformations of Hsp70 ATPase domain.

The closed form (white) is a structure observed in the absence of nucleotide (PDB id: 1HPM). Two open forms are shown, both observed in the complexes formed with NEFs: in cyan is the structure from the complex with the Sse1 (PDB id: 3C7N); and in orange is that assumed when complexed with BAG-1 (PDB id: 1HX1). The NEFs (Sse1 and BAG) are not shown here. The three structures have been aligned using the Kabsch algorithm ((Kabsch et al., 1990) as implemented in PyMol).

4.1.1 Different conformations of ATPase domain

We adopted the approach proposed by Nussinov and coworkers (del Sol et al., 2006) to identify the central residues of the ATPase domain. The details of the method are presented in section 2.2.2. We calculated the centrality profile for residues in all three structures, including one with the ATPase domain in the closed state and two others, in the open state. The results are shown in **Figure 20** for unbound (panel **a)** and Sse1-bound (panel **b**) ATPase domain. The centrality profile for the BAG-bound form exhibits patterns similar to those observed in panel **b** (not

shown). The characteristic path lengths and the RMSDs calculated after optimal structural alignment indicate that the lobes of the Sse1-bound ATPase domain are further apart than the BAG-bound form.

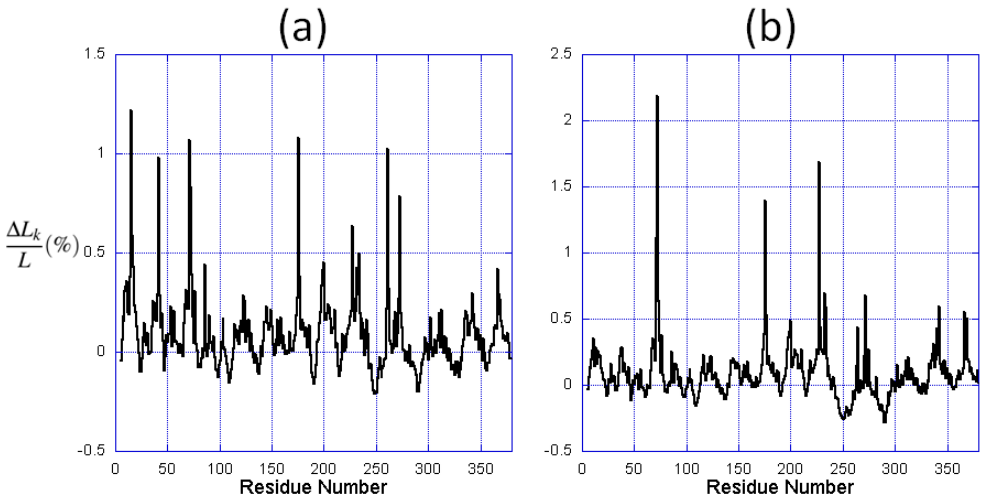

Figure 20. Centrality profile for Hsp70 ATPase domain residues.

The profiles are calculated for (a) the closed form, and (b) the open form in the Sse1-bound ATPase domain. The abscissa represents the fractional change $\Delta L_k/L$ in characteristic path length compared to the original network. Maxima refer to nodes that have a strong impact on communication efficiency, if removed.

Comparison of panels **a** and **b** shows that the two profiles exhibit similar features (i.e., peaks and minima at the same regions), while the relative heights of the peaks vary. In particular, the peaks near residues located at the inter-lobe interface, that is residues 257-276 (helix 9) and residues 10-60, are suppressed in the open form (panel **b**). In contrast, some residues located at the nucleotide binding pocket (e.g., Arg342 and Asp366) display more pronounced centrality properties in the open form compared to the closed form (note that the ordinate scales are different in the two panels). The increase in centrality suggests that they assume an enhanced role in establishing the communication away from the active site in the open form.

We consider the top ranking (top 2%, or equivalently, top eight) residues in the centrality profile in each case, and refer to them as the central residues in the following text. Among them four are distinguished as central residues in all of three structures, regardless of the open or closed state of the ATPase domain: His71, Arg72, Glu175 and His227; in contrast, the other four residues vary with the conformation (**Table 4**).

Table 4. Central residues in the closed and two open conformations of Hsp70 ATPase domain.

PDB id	***L*(Å)**	**Binding NEF**	**conformation**	**Central residues(*)**	
				shared by all	*Specific to examined structures*
1hpm	14.06	None	closed	**Lys71 Arg72 Glu175** His227	Tyr15 Tyr41 Arg261 Arg272
1hx1	14.67	BAG	open		Ala60 Arg261 **Arg342** Asp366
3c7n	15.26	Sse1	open		Leu73 Asp232 **Lys271 Arg342**

(*) fully conserved residues are written in boldface (see **Figure 22a**). Leu73 and Asp232 are also highly conserved.

4.1.2 Structural and sequence variations among central residues

We examined the position of the central residues on the structure (**Figure 21**), and performed sequence analyses to examine their conservation profile and co-evolutionary properties (**Figure 22**).

The four residues that are invariant to conformational changes are colored cyan in **Figure 21**, and are labeled in **Figure 21b**. Interestingly, these (sequentially separated) residues appear to form a (spatially contiguous) communication path across the lobes, starting from Arg227 and ending at Glu175. Indeed, Lys71 and Glu175 serve as catalytic residues (O'Brien et al., 1996; Vogel et al., 2006a) and regulate a proline switch that, in turn, regulates the inter-domain allosteric interactions. The central residues are found to be mediating allosteric communications in a variety of protein families (del Sol et al., 2006). We propose that the residues detected here

also play an important role, not only a catalytic, but also a signaling. The are therefore proposed to be implicated in the communication of the nucleotide exchange events to the other regions of the ATPase domain, including for example the interface with the substrate-binding domain.

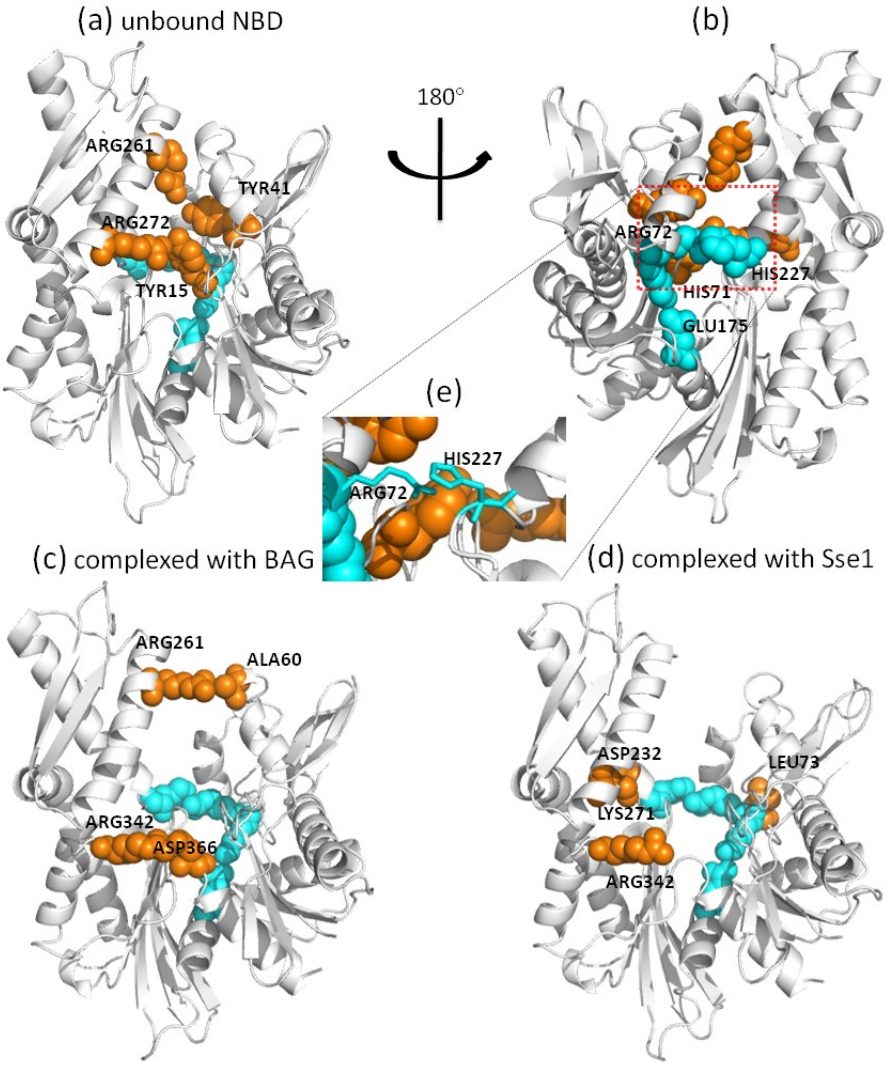

Figure 21. Position of Hsp70 ATPase domain central residues.

Panels a and b display the positions on the ATPase domain closed form, (c) on the ATPase domain open form bound to BAG and (d) on ATPase domain open form bound to Sse1. Panel e highlights the interaction of Arg72 and His227. The conserved central residues are colored cyan; other central residues are colored orange. Panel b is the rotated view of panel a, and the conserved central residues are only labeled in panel b.

The analysis of the MSA generated for the same set of sequences (described in section 3.2) reveals that these four residues are highly conserved. See the sequence logo (Crooks et al., 2004) presented in **Figure 22a**, which clearly indicates that Lys71, Arg72 and Glu175 are fully conserved. His227, although not conserved, can only be substituted by phenylalanine, although histidine probability is much higher, suggesting that a large aromatic group may be functional at this position. The interaction of Arg72 and His227, shown in **Figure 21e**, can be viewed as a highly conserved amino-aromatic interaction (Burley and Petsko, 1986), which is presumably maintained when histidine is replaced by phenylalanine. So even though His227 tolerates a mutation to phenylalanine, its interaction with Arg72 is conserved. In the following text we will refer to these 4 residues as the *shared central residues* (SCR).

The other central residues also exhibit patterns relevant to the functional changes in ATPase domain conformation. In the closed form, these residues (Tyr15, Tyr41, Arg261 and Arg272) are distributed along the cleft formed by lobes I and II to form two closely interacting pairs: Arg272---Tyr15 and Arg261---Tyr41. These pairs serve as two bridges that connect the subdomain IIB with subdomain IA (Arg272---Tyr15) and with subdomain IB (Arg261---Tyr41). Bukau and coworkers (Brehmer et al., 2001) have shown that the salt bridges formed between helices 1 and 9, labeled in **Figure 19**, affect the nucleotide exchange of ATPase domain. We speculate that among the residues located on these two helices, these two pairs arginines and tyrosines, also involved in amino-aromatic interactions, play a key role in controlling the subdomain closure and opening, which in turn ensure nucleotide stabilization or release, respectively. Moreover, since the central residues are supposed to be the most "indispensable" residues in establishing the shortest-path communications, the two pairs we identified might be

the "anchors" that maintain the closed conformation of the ATPase domain. Indeed, this conjecture is reinforced by the collective dynamics of the ATPase domain in the next section.

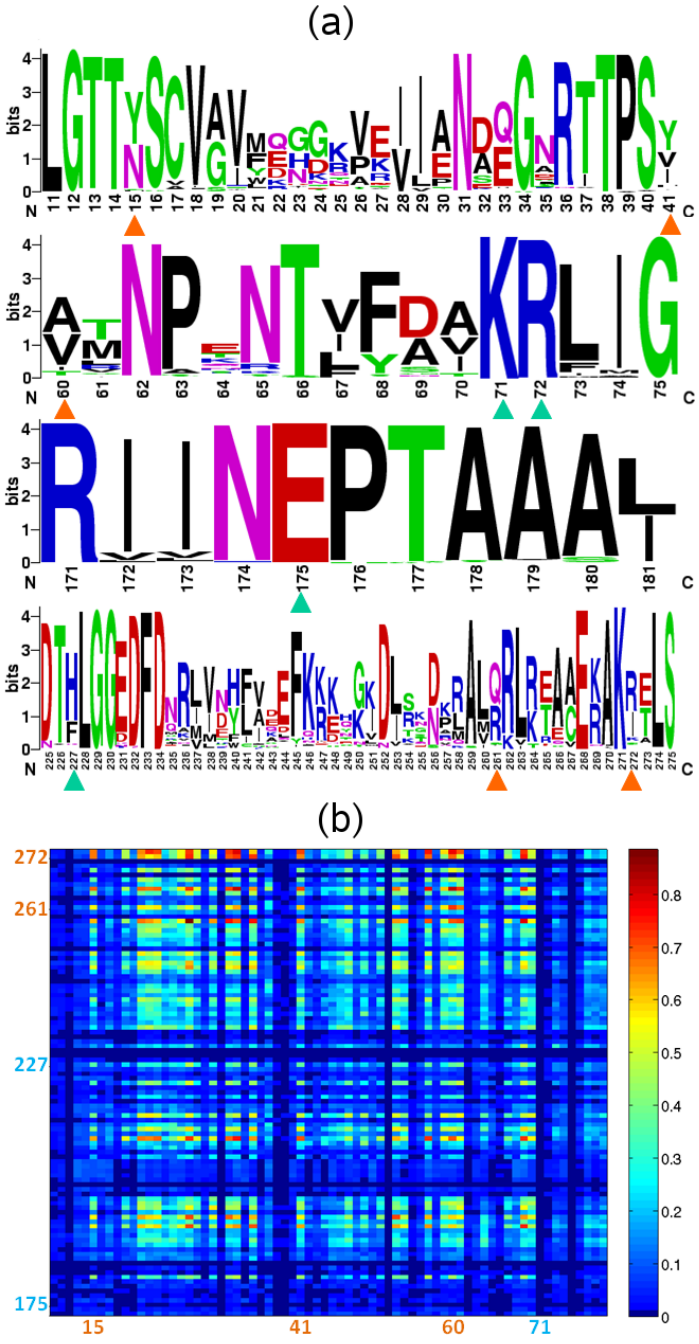

Figure 22. Sequence analysis of central residues.

(a) Sequence logo of the SCRs (marked with cyan triangles) and other central residues (marked with orange triangles) identified in the closed and open (BAG-complexed) ATPase domains. (b) MI map between residues 173-

274 and 10-80, which includes the central residues located at the lobe interface in the closed state. The SCRs' residue numbers are written in cyan, whereas others are written in orange.

Interestingly, residues at these four positions tend to co-evolve, when they are not conserved, as may be seen from the MI map in **Figure 22b**. By examining the sequence logo (**Figure 22a**), we found that the variation of amino acids at these residues primarily arises from the difference between the Hsp70 mammalian homolog Hsc70 and the Hsp70 bacterial homolog Dnak. The interactions between the two lobes of DnaK, as well as the interaction of the Hsp70 ATPase domain with NEF (GrpE in this case), primarily consist of hydrophobic contacts; whereas in Hsc70, there is a prominence of electrostatic interactions. The co-evolution of these central residues is in line with the specificity of their interactions in different organisms.

In the BAG-bound ATPase domain, which assumes a less open conformation between the two NEF-bound structures, there still remains a contacting residue pair between the tips of subdomains IB and IIB (Arg261---Ala60, see **Figure 21c**), but this interaction can hardly account for the interface between the lobes. On the other hand, Arg342 and Asp366 are both conserved and they line the nucleotide binding pocket. Their interactions are crucial for maintaining the conformation of the active site. In the Sse1-bound ATPase domain, because subdomains IB and IIB have undergone a rotation, Asp232 interacts with Lys227, which implies a putative extension of the SCR to subdomain IIB. Similarly, Leu73 extends the SCR to subdomain IB. Lys271 and Arg342 are both conserved residues at the active site.

4.1.3 Global dynamics of ATPase domain and role of central residues

As suggested in previous work (del Sol et al., 2006), the central residues generally relate to the system fragility; that is, these residues ought to remain "stable" to maintain the biological function of the molecule. From the sequence perspective, this requires sequence conservation; from the structural dynamics perspective, one might expect to see little variations, if any, in their spatial positions. In order to critically examine their dynamical characters, we examined the equilibrium dynamics of the ATPase domain using the GNM. We focused in particular on the low frequency end of the spectrum of modes, given that these modes are usually highly cooperative and relevant to function (Bahar et al., 1998; Bahar and Rader, 2005). We compared the centrality profile and the mobility profile resulting from the weighted average of the 10 slowest modes of the closed-form ATPase domain (**Figure 23**). Strikingly, the mobility profile (which represents the normalized distribution of square fluctuations in residue positions driven by these modes) exhibits minima at the peaks of the centrality profile, and vice versa. Minima in the mobility profile represent sites that act as hinges (or anchors) in the collective modes. Notably, all the central residues coincide with minima (**Figure 23a**), which is indicative of their mediating role in the global motions of the ATPase domain. Arg261 and Arg272 are of particular interest: first, their mobility is higher than that of other central residues, suggesting a lower energy barrier for them to dissociate from lobe I to facilitate the cleft opening; second, helix 9 as the linkage between two most mobile regions of ATPase domain, is implicated in functional motions.

Overall, the centrality profile and the slow modes curve are negatively correlated, which can be observed in **Figure 23b**. **Figure 23a** indicates the correspondence between the peaks of one curve and the valleys of the other, in most cases. In **Figure 23b**, the residues with high

centrality (≥ 0.05) are shown to be characterized with low mobility, except for Asp86 (labeled in italic in **Figure 23b**). Indeed, Asp86 is located in an exposed helix that accounts for the rotation of subdomain IB and forms a salt bridge with Arg72, which in turn is one of the shared central residues presently identified. It appears that the salt bridge between Asp86 and Arg72 is critical to the motion of the exposed helix. On the other hand, the residues with negative centrality are usually located at the ends or tips of the structure, consistent with their high mobility.

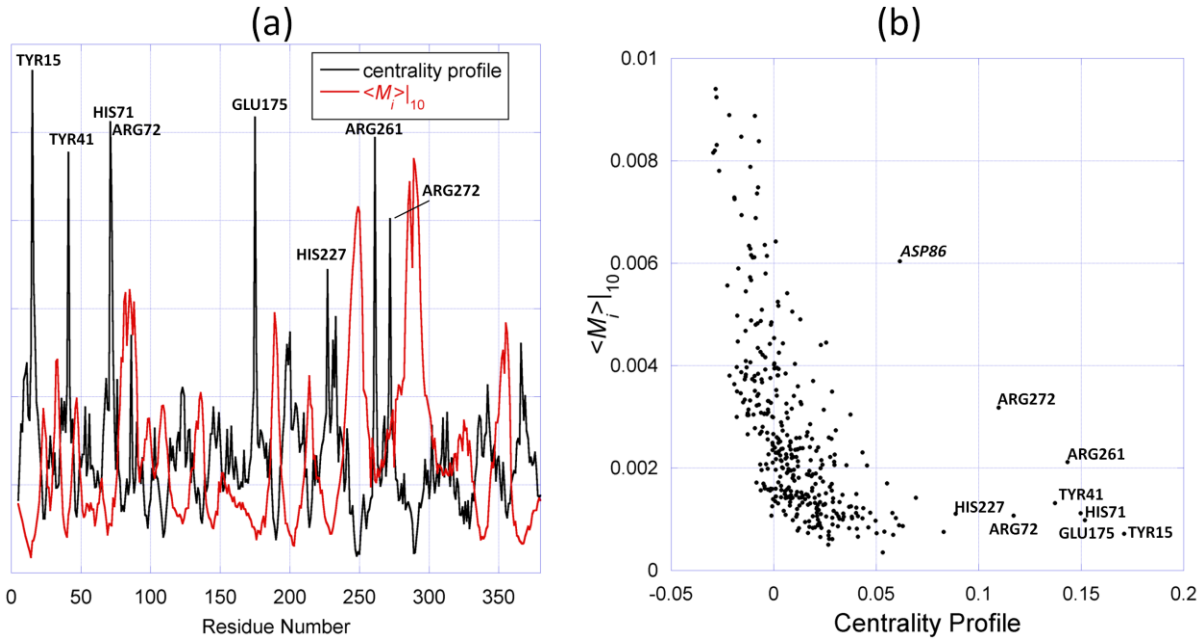

Figure 23. Comparison of the slowest modes and the centrality profile.

(a) Mobility profile resulting from the weighted average of the 10 slowest modes and the re-scaled centrality profile of the ATPase domain closed form. The centrality profile is re-scaled for visual comparison. The peaks corresponding to central residues are labeled, in addition to another outlier, Asp86, distinguished by its high mobility. (b) Mobility versus centrality for all residues in the ATPase domain closed form. The points corresponding to central residues are labeled. The ordinate and abscissa values are taken from the two curves in panel a, for each residue.

4.1.4 Summary of different types of central residues

We can group the central residues into three categories depending on their location on the structure and/or their role in the structural dynamics: (i) The hinge point, (ii) bridging point at the interface near the cleft (or contact interface), and (iii) stretched linker (or long helix/loop stretching out). These three categories are illustrated in **Figure 24.**

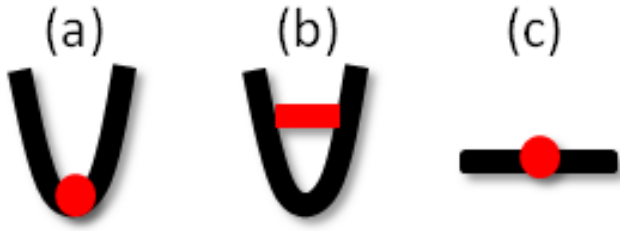

Figure 24. Three scenarios for the central residue's location on the structure.

In the first case, the central residues (e.g., SCR) connect two parts, at least one of which is highly mobile. These residues mediate the communications between different parts of the molecule, and transmit the information necessary for the proper functioning of the molecule. Perturbations at these residues are most likely to impede function. These residues are also highly conserved and serve as hinge points not only with respect to the two structural elements that are directly connected, but in the global dynamics of the entire ATPase domain. In the second case, the central residues serve as linkages at the interface between substructures that have intrinsically access to alternative (e.g. open and closed) conformations. They act as the "anchoring point" of the interface, and can be the determinants of the motions of the moving parts. These residues are more exposed to the environment and more tolerant to mutations compared to the first case. Yet, their important role is signaled by correlated mutations that take place which presumably aim at restoring the key role (that of locking the closed form in this case). For residues in the third

category, although we did not observe any such residue in this study, they have been observed in other systems. For example, the inter-domain linker between the ATPase domain and SBD of the Hsp70 possesses such residues, which evidently play a key role in establishing the allosteric communication between the two domains (Swain et al., 2007).

4.2 PART II: PATHWAYS IN THE TWO-DOMAIN MODEL

There is no structural data available to date for the intact Hsp70 composed of the two domains, ATPase domain and SBD, except for a structure where the two domains are connected by a loose linker (ADP-bound state) (Bertelsen et al., 2009). In order to examine the allosteric interactions between the two domains of Hsp70, we utilized the homology model of DnaK, the *E. coli.* homolog of Hsp70 (Smock et al., 2010), where the two domains are in close contact (ATP-bound state). This structural model was generated using the conformation of Sse1 (a member of the Hsp110 family) (Liu and Hendrickson, 2007; Schuermann JP et al., 2008) as template in Modeller (Sali and Blundell, 1993). The resulting structure, representative of the ATP-bound state, is shown in **Figure 25b** (and **Figure 1**).

We also evaluated the sequence conservation and co-evolution properties within the full-length DnaK. To this aim, a MSA of 2608 sequences (and 601 representative columns/sequence positions, corresponding to residues 4-604 in DnaK) was generated by refining the data retrieved from Pfam for Hsp70 family members (Pfam id: PF00012, Pfam version 24.0). The acquisition of the MSA follows the same protocol and parameters from another study of ours (section 5.1). Briefly, the DnaK wild type sequence (Bertelsen et al., 2009) was used to search against the Pfam MSA to identify therein a reference sequence with one-to-one residue mapping to the

query sequence, then the columns of the MSA corresponding to the reference sequence residues were retained to represent the DnaK residues. The MSA was then subjected to further refinement, including removal of redundant sequences and those with extensive gaps. A detailed description can be found in section 5.1. The conservation profile (information entropy) is calculated using equation (28) for each residue (**Figure 26**). The results based on ET analysis (see section 2.3.2) are shown in the Appendix Figure A3.

As a second step, we performed a GNM analysis. To this aim, we considered the first 530 residues denoted as $DnaK_{530}$. The remaining C-terminal portion of the structure has been truncated because of its high mobility, which might obscure the collective motions, in accord with previous examination of the same model (Smock et al., 2010). **Figure 25a** displays the global mobility profile based on the weighted average of the m = 10 slowest modes (which account for 40% of the overall dynamics). Three major observations are made. First, the portion of the profile corresponding to Hsp70 ATPase domain (residues 4-388 in the examined structure) is highly similar to that previously obtained for the Hsp70 ATPase domain alone (**Figure 11** in Chapter 3). This indicates that this domain maintains its intrinsic dynamic character in the Hsp70. Second, the linker residues 389-392 are located at low mobility regions (indicated by the red dots). The linker region thus serves as a hinge that modulates the concerted motions of the two domains. Third, we distinguish three interfacial residues that occupy an important mechanical position (hinge-site indicated by the minima of the global mobility profile) in the SBD: Thr417, Asp481 and Gly506. These are all clustered at the interface with the ATPase domain, presumably playing a role in mediating interdomain interactions (**Figure 25b**). These residues are not conserved, but exhibit high co-evolution tendencies with other functional residues, as will be shown below. Interestingly, the loop containing Thr417 has been shown to

exhibit a large structural reorientation in an apo form of Hsp70 SBD (Pellecchia et al., 2000), suggesting that the observed constrained mobility in the homology model is due to the interaction with the ATPase domain.

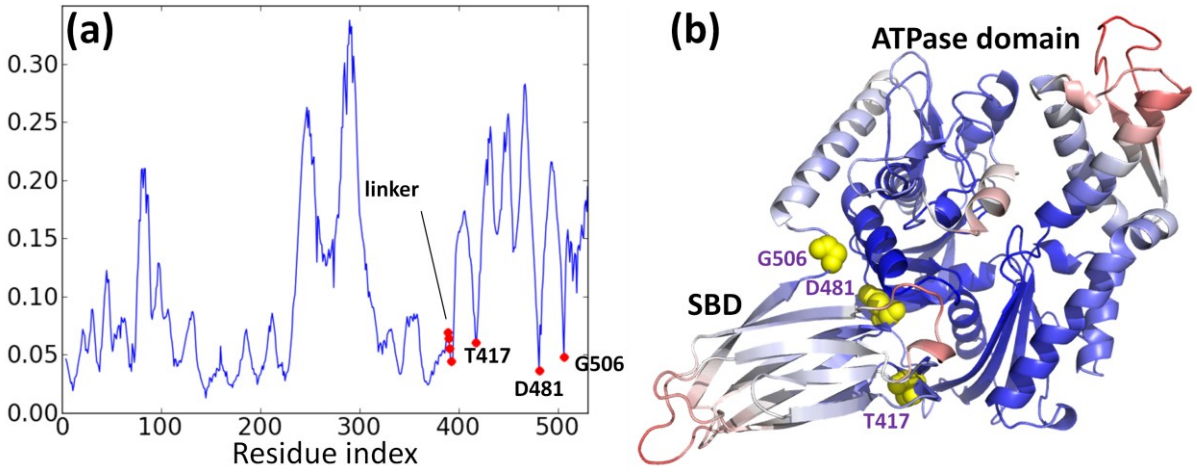

Figure 25. Mobility profile of DnaK$_{530}$.

(a) Global mobility profile of DnaK$_{530}$ plotted as the function of the residue number. The linker and the key mechanical residues are marked with red dots. (b) Color-coded ribbon diagram of the DnaK$_{530}$ based on the extent of mobility (red: most mobile; blue: list mobile). The three key interfacial SBD residues, which appear to be highly stable (participating in hinge-bending region) are highlighted in yellow sphere representation. Note that the most mobile region on the ATPase domain (colored pink-red) is the NEF-binding subdomain IIB, consistent with previous observations (Liu et al., 2010).

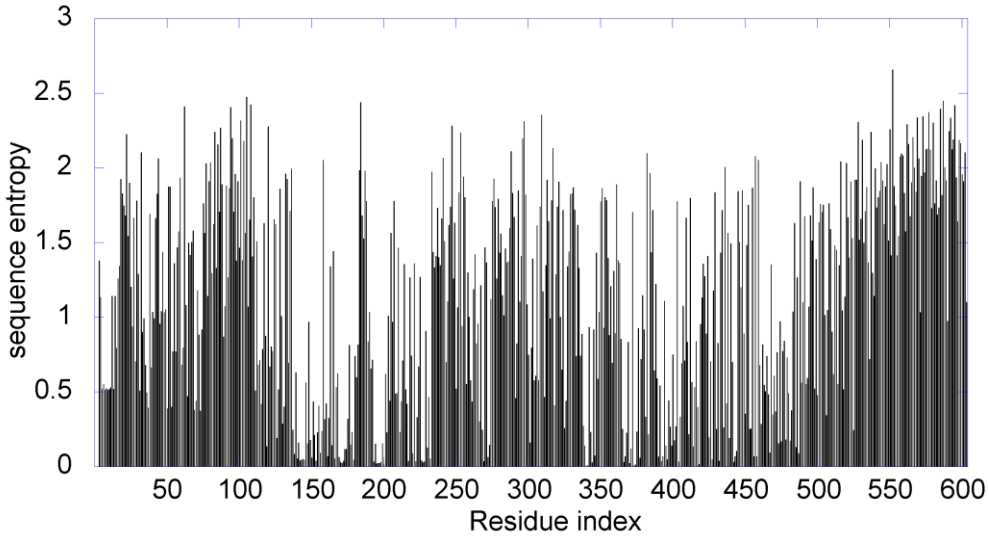

Figure 26. Conservation profile of DnaK residues 4-604.

4.2.1 Sensitivity and influence profiles derived from the PRS matrix

The PRS analysis (see section 2.2.1) was then performed on $DnaK_{530}$. **Figure 27a** shows the obtained normalized PRS matrix. Note that the diagonal of the matrix has been set to 0 (originally all 1) for visual clarity.

The bar graphs in **Figure 27a** indicate the average taken along the rows and columns of the map. The average value of the i^{th} row of $\bar{\mathbf{S}}_{PRS}$ is the average response of residue i, representing its *sensitivity* to external forces applied to the structure. Here a numerical technique is used to identify peaks on the response profiles: a cubic spline with smoothing parameter 0.1 (Wahba, 1990) is used to approximate the response profile, on which the stationary points are identified. Then the local maxima around these stationary points are considered (within 5 neighboring sites) as the peaks. In this way 13 such sites are identified among the top 50 highly sensitive residues (**Figure 27b**). These residues tend to be located at three regions of different

subdomains/domains: the NEF-binding site (subdomain IIB, colored orange), the substrate binding site (colored purple), and four residues close to the interface (subdomain IA, colored red). As we shall see in the following sections, many of them exhibit high responses to perturbations at the active site and the inter-domain interface.

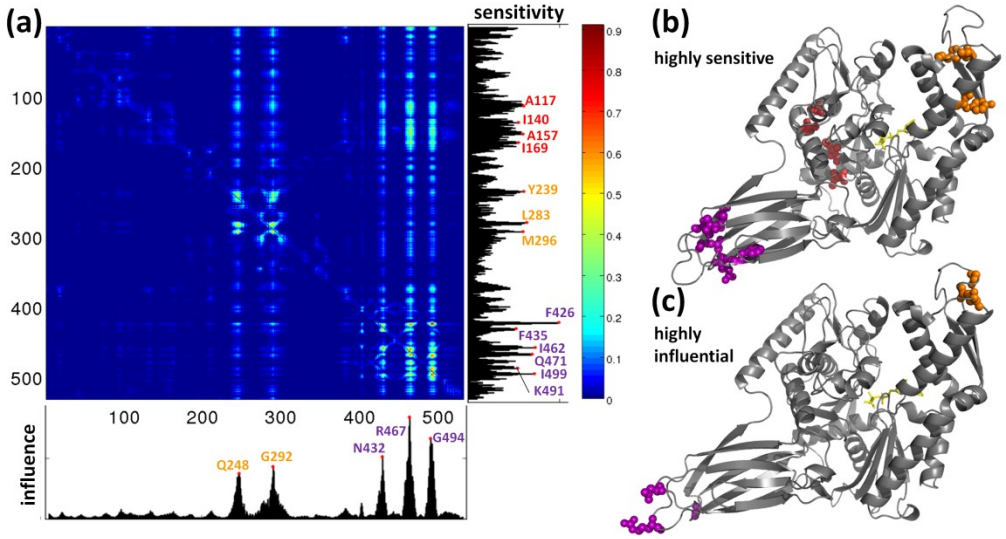

Figure 27. Results of Perturbation Response Scanning (PRS) analysis.

(a) Normalized PRS matrix. The bar graph on the right hand side shows the average of each row (sensitivity profile), and the one below the map shows the average value along each column of the map (influence profile). In the former case the peaks among the 50 top-ranking residues are marked with red dots and labeled, whereas in the latter case the 5 most distinguishable peaks are highlighted. The labels of these residues are colored according to the subdomain locations shown in **Figure 1**. (b) Ribbon diagram of $DnaK_{530}$ with the highly sensitive residues (labeled in the same color as in the bar graph in panel a) shown in spheres. (c) Ribbon diagram of $DnaK_{530}$ with highly influential residues (labeled in the same color as in the bar graph in panel a) shown in spheres.

On the other hand, the average value of column j in $\bar{\mathbf{S}}_{PRS}$ indicates the average response of other residues to the perturbations at residue j, reflecting how "influential" residue j is to account for the conformational changes of the protein. The obtained results for all residues are collectively referred to as the influence profile. Five residues are easily distinguishable in the

influence profile: Q248, G292, N432, R467, and G494. These residues tend to be located at the exposed loop regions of the molecule (**Figure 27c**). Among them, R467 has been shown to form salt bridge with the α-helical lid (Liebscher and Roujeinikova, 2009) in the ADP-bound state.

4.2.2 Perturbing the ATP γ-phosphorus atom

ATP hydrolysis provides the driving potential for cooperative structural transitions in many allosteric proteins. Toward a mechanistic understanding of the effect of structural changes at the ATP-binding site, on the collective dynamics of Hsp70, we adopted a recently introduced methodology, PRS, and examined the response of the molecule to perturbations introduced at the γ-phosphate group of ATP. ATP was placed in the active site of the homology model by structural alignment against the ATP-bound Hsc70 structure resolved by McKay and coworkers (PDB id: 1NGF, (Flaherty et al., 1994)). In the present coarse-grained (ANM) representation of the structure, a node was identified with the position of the γ-phosphorus atom; other ATP atoms included in the network as additional nodes were two carbon atoms, C4' and C2, and the α- and β-phosphorus atoms.

The bar graph in **Figure 28a** shows the response profile generated upon perturbing the γ-phosphorus atom on the ATP bound to the Hsp70 ATPase domain, denoted as an *N*-dimensional array, $<\|\Delta\mathbf{R}^{(\gamma)}\|^2>_{norm}$. The residues that exhibit high responses to the perturbation, shortly referred to as *high-susceptibility* (HS) sites, tend to form clusters composed of 5-10 sequential residues. We identified the top-ranking HSs in the computed $<\|\Delta\mathbf{R}^{(\gamma)}\|^2>_{norm}$ profile (some of which are labeled in the panels a and displayed by the same color sphere representation in label b). We note these residues are distributed across all subdomains, indicating the multiple directions in which the perturbation propagates. Notably, the residues close to the active site exhibit strongest

responses, such as Pro143, Glu171, Thr199, and Thr12. Both Pro143 and Glu171 have been shown to play an important role in relaying the activity at the nucleotide binding pocket to the SBD (Vogel et al., 2006a), and Glu171 is close to the interfacial residue Asp481 distinguished above to occupy a key position at the interface between the two domains (**Figure 25**).

Two residues in the Hsp70 ATPase domain subdomain IIB (Leu283 and Met296), on the other hand, exhibit distant, yet strong, responses to the perturbation. Both of these residues are located at the tip of a β-hairpin involved in NEF recognition (**Figure 28b**). As shown in **Figure 25b**, this region also enjoys an enhanced mobility in the global modes. The strong response of this region to perturbations at the ATP-binding site suggests that the ATP hydrolysis is exploiting the intrinsic high mobility of this region to regulate its interaction with NEFs. In our previous work, we have shown that NEFs interact primarily with the subdomain IIB, and lock this subdomain in an open conformer to facilitate nucleotide exchange (Liu et al., 2010). The current observation complements our previous findings by suggesting an effective two-way signal transduction mechanism between the NEF- and nucleotide-binding sites.

The inset diagram in **Figure 28a** displays the mapping of the response profile $<\|\Delta\mathbf{R}^{(\gamma)}\|^2>_{norm}$ onto the 3D structure. The dynamical coupling between the two functional sites, ATP-binding and NEF-binding, is clearly seen. The examination of the figure suggests that the strong response of the NEF-binding site is due to the mediating role of the long helix 8 on the ATPase domain. It can also be seen that the perturbation propagates toward the interdomain interface, including the regions in contact with both the α-helical lid and β-sandwich on the SBD.

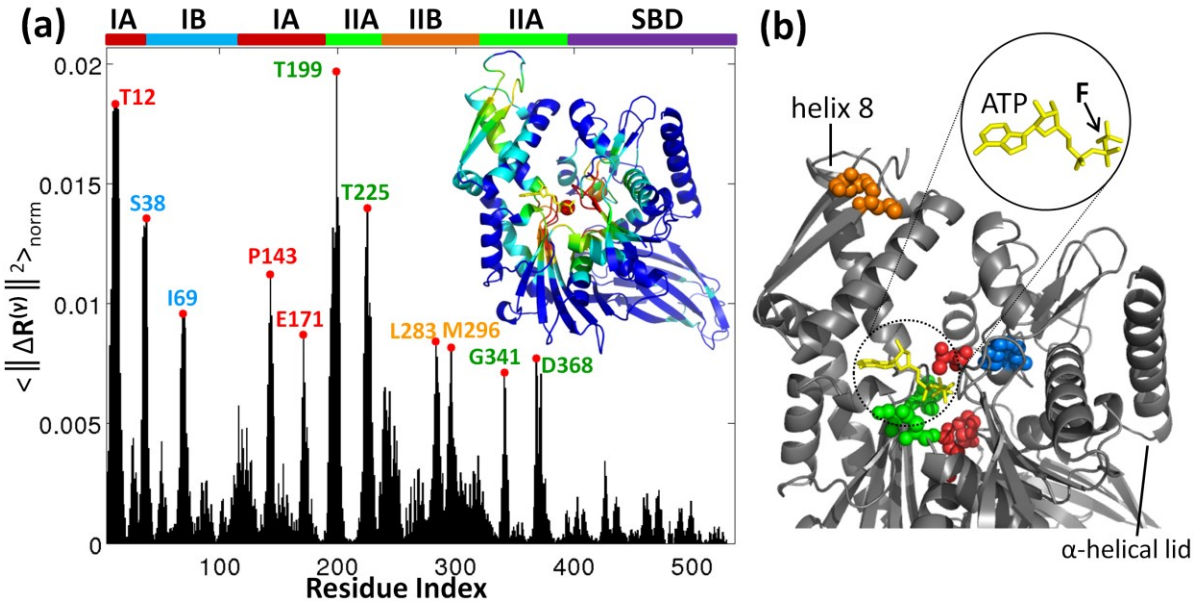

Figure 28. Responses to perturbation at the γ-phosphorus atom of the ATP.

(a) Response profile obtained upon perturbing the γ-phosphorus of the ATP ($<\|\Delta\mathbf{R}^{(\gamma)}\|^2>_{norm}$). The highly susceptible sites (HSs) are labeled; the labels are colored according to their subdomain location on the ATPase domain, as shown in the colored bar above (also in **Figure 9**). The inset diagram is color-coded according to $<\|\Delta\mathbf{R}^{(\gamma)}\|^2>_{norm}$, and the red sphere represents the atom being perturbed. (b) Ribbon diagram highlighting the HS residues. The HS residues labeled in panel (a) are shown in spheres and colored according to their subdomain location in the ATPase domain and SBD. The ATP molecule is shown in stick representation and colored yellow. The force exerted on the ATP is illustrated by the encircled diagram.

4.2.3 Perturbing the interfacial residues between the two ATPase and SBD domains of Hsp70

Next, we perturbed the three interfacial key mechanical residues identified above (**Figure 25**), and the interdomain linker, and examined the resulting response profiles. We begin with Val389 at the inter-domain linker. The resulting profile is shown in **Figure 29**. Note that the bars

corresponding to the perturbed residue and three sequential neighbor on both sides (residues 386-392) are not shown in the figure for visual clarity as their high response is trivial (see Appendix Figure A4), and would obscure the cooperative response of the overall molecule. The HS residues that emerge in response to Val389 perturbation are mainly located at five regions on the structure, three in the ATPase domain, and two in the SBD, as displayed in **Figure 29b.** These regions are: (i) the hydrophobic pocket adjacent to the linker in the ATP-bound state (Ile174, Leu181, Ile205, Ile215, and Val377 in subdomains IA and IIA; colored green and red); (ii) the close neighborhood of the nucleotide-binding site (Ile5, Ile18, and Arg25 ; colored red); (iii) a distal site centered at Glu306 (orange) at the interface between subdomains IIA and IIB; (iv) the close neighborhood of the interdomain global hinge site on the SBD (Asp393, Ile418, and Ala480; colored purple); and (v) another distal region, this time on the SBD, located at the exposed end of the SBD β-barrel (Phe426, Ile462, and Ile472; colored purple).

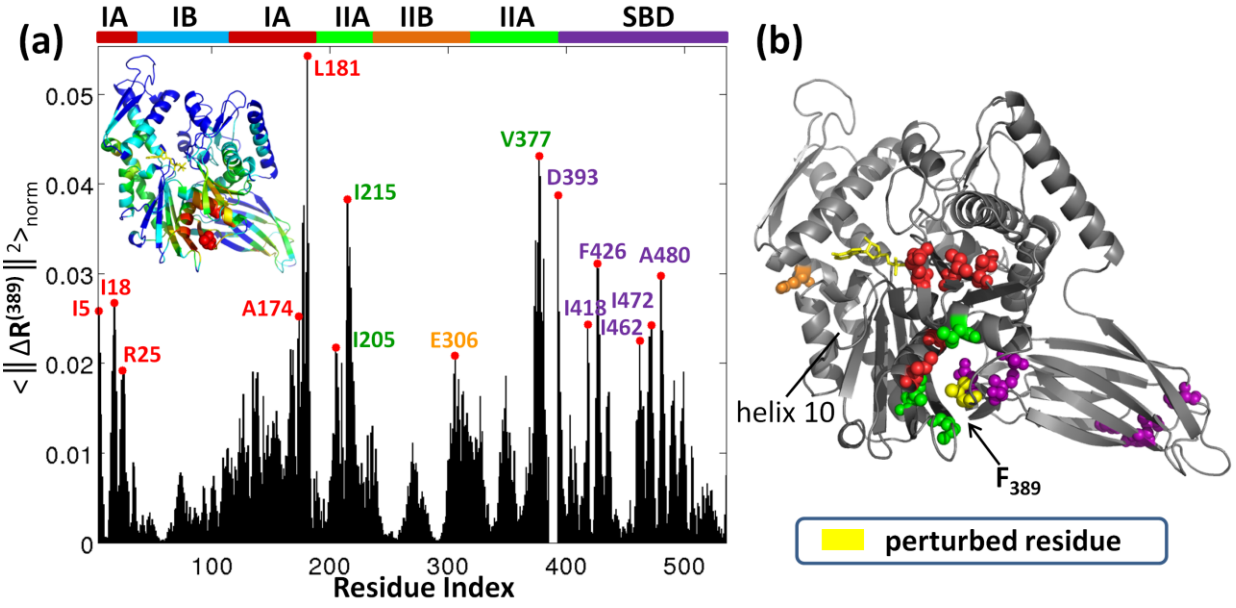

Figure 29. Responses to perturbation at the linker residue Val389.

(a) Response profile of perturbing Val389 ($<\|\Delta\mathbf{R}^{(389)}\|^2>_{norm}$). Peaks highlight the HS residues in the presence of this perturbation. Labels are colored according to the subdomains in which the labeled residues participate. The inset ribbon diagram is color-coded in the order of decreasing response $<\|\Delta R^{(389)}\|^2>_{norm}$, from red to blue. (b) Location of five clusters of HS residues. The HS residues are shown in sphere representation and colored according to their subdomain. Val389 is shown in yellow spheres, and the ATP molecule is shown in yellow stick representation.

The residues in the groups (i) and (iv) appear to act as *sensors* that detect the perturbation and initiate its propagation toward the nucleotide binding site (cluster (ii)), and even to distal regions on the ATPase domain and SBD (clusters (iii) and (v)). Note that the cluster (ii) that comprises several residues at the ATP-binding site serves as an efficient *effector* for transmitting signals, given its tight packing properties complemented by highly specific interactions. Indeed, the coupling between the linker-neighboring site and the nucleotide-binding site has been noted in previous studies (Swain et al., 2007; Zhuravleva and Gierasch, 2011). **Figure 29b** suggests that the Glu306, located near helix 10, presumably plays a mediating role between the linker site and the subdomain IIB.

The linker thus transmits signals to both domains, and even to the distal regions in these domains, if subjected to an external perturbation. As we will see in **Figure 30**, the cluster (v) is highly sensitive to perturbations at other key mechanical residues as well. Note that our response profiles have been normalized with respect to the equilibrium MSFs of residues. As such, they essentially reflect the *changes* in fluctuations elicited in response to perturbations. Like their counterpart at the NEF-binding site (Leu283 and Met296 in **Figure 28**), the distal residues, Phe426, Ile462, and Ile472, on the SBD may potentially serve as recognition/binding site for the substrate. Previous work has shown that mutations at F426 and I462 reduce the substrate-binding affinity, and in particular mutation at I462 can impair the DnaK function *in vivo* (Davis et al.,

1999; Montgomery et al., 1999). Their functional relevance to substrate binding is also supported by their high conservation.

Figure 30 shows the ribbon diagrams of the Hsp70 colored-coded by the responses profiles triggered upon perturbing the three key interfacial residues on the SBD, Thr417, Asp481 and Gly506. Panels a and b indicate that $<\|\Delta\mathbf{R}^{(417)}\|^2>_{norm}$ and $<\|\Delta\mathbf{R}^{(481)}\|^2>_{norm}$ exhibit similarities, while $<\|\Delta\mathbf{R}^{(506)}\|^2>_{norm}$ presents new features.

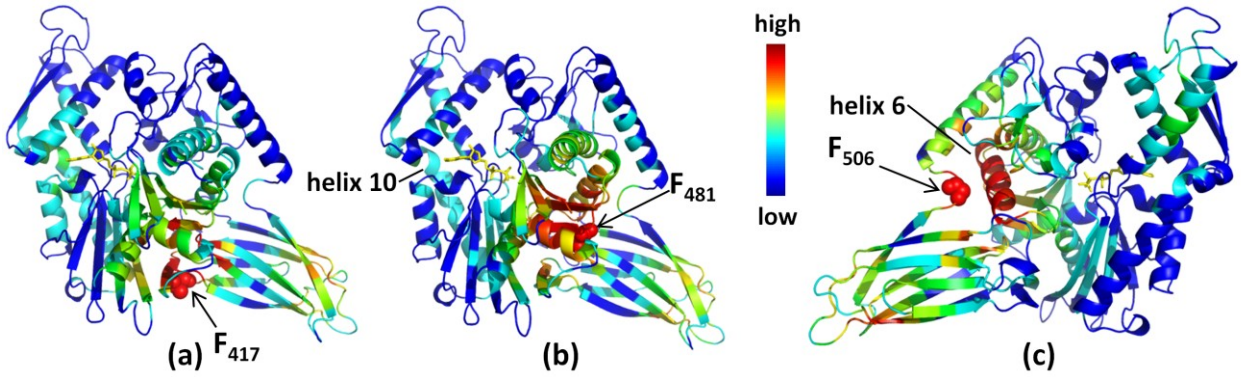

Figure 30. Responses to perturbation at key mechanical residues.

The ribbon diagram is color-coded according to the response profile of perturbing (a) Thr417, (b) Asp481, and (c) Gly506. In each diagram the residue being perturbed is shown in spheres and colored in red. The scale of the heat map is shown between panels b and c. The ATP molecule is shown in yellow sticks in all panels.

Let us first examine more closely $<\|\Delta\mathbf{R}^{(417)}\|^2>_{norm}$ and $<\|\Delta\mathbf{R}^{(481)}\|^2>_{norm}$. Their SBD profile is similar to that observed for $<\|\Delta\mathbf{R}^{(389)}\|^2>_{norm}$, and the linker-binding pocket and helix 10 again appear to be involved in mediating the interactions between subdomain IIB and the inter-domain interface. On the other hand, we note that the perturbation of Thr417 causes significant response from both the linker and Asp481, whereas perturbing Asp481 induces most significant response from Arg167 on the ATPase domain, close to the core region of this domain (see **Figures 31** and **32**). Therefore, Asp481 displays a more direct communication with the

nucleotide binding pocket. We note in particular that three charged residues R167, K155 and D393 appear to act as sensors via electrostatic interactions (**Figure 32a**) and R167 closely interacts with Q378, which, together with Val139 apparently serve as effectors for signal transduction.

Previous studies invited attention to the involvement of exposed conserved, polar and charged residues in substrate binding (Hu et al., 2000; Ma et al., 2003). Our previous work suggests that while coordinating residues at substrate binding site are usually conserved, those at 'recognition' sites may (and are apparently functionally required to) undergo correlated mutations to maintain a balance between substrate specificity and structural adaptability (Liu and Bahar, 2011). The propenderance of co-evolving amino acids in the subdomain IIB of Hsp70 ATPase domain was indeed attributed to the adaptability to specific NEF recognition (Liu et al., 2010). In the same way, it is of interest to examine the sequence conservation, or co-evolution, properties of amino acids that emerge here as highly susceptible sites. Indeed, the residues within the neighborhood of Asp481 on the ATPase domain is populated with highly conserved polar/charged residues (see **Figure 32b**), suggesting a functional role in establishing interdomain communication. As will be further elaborated in the next subsection, the residues that serve as sensors and effectors exhibit distinctive co-evolutionary properties as well, in support of their role in facilitating the allosteric response of Hsp70.

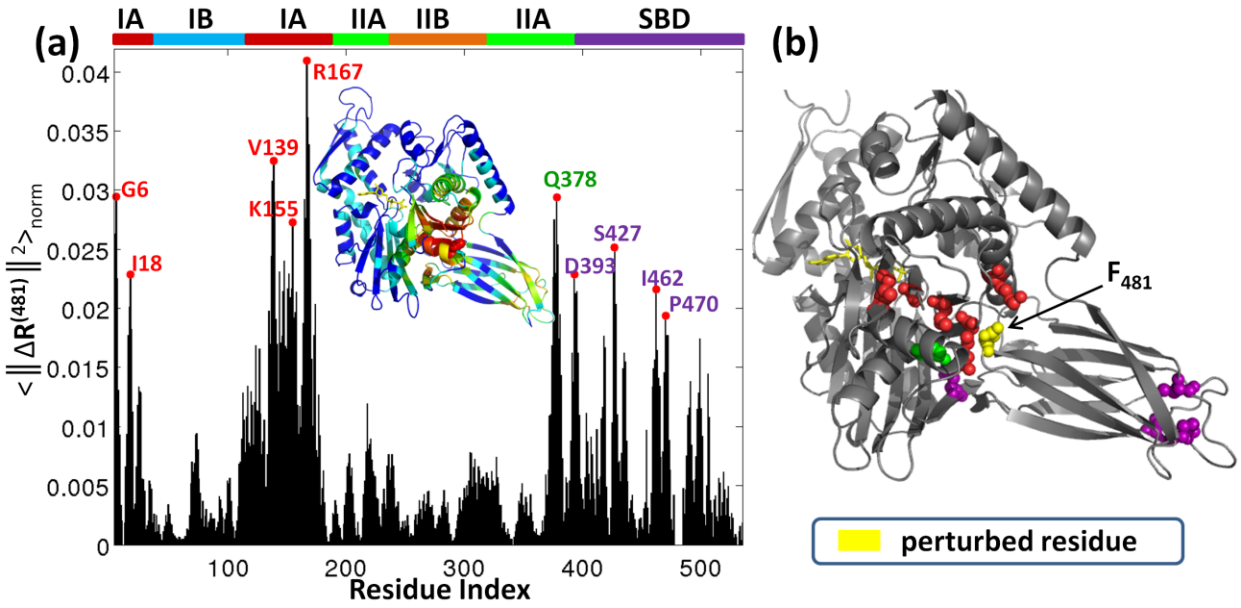

Figure 31. Responses to perturbation at residue Asp481.

(a) Response profile of perturbing Asp481 ($<\|\Delta R^{(481)}\|^2>_{norm}$). Peaks highlight the HS residues in the presence of this perturbation. Labels are colored according to the subdomains in which the labeled residues participate. The inset ribbon diagram is color-coded in the order of decreasing response $<\|\Delta R^{(481)}\|^2>_{norm}$, from red to blue. (b) Ribbon diagram highlighting the HS residues. The HS residues are shown in sphere representation and colored according to their subdomain location in the ATPase domain and SBD, and the ATP molecule is shown in yellow stick representation.

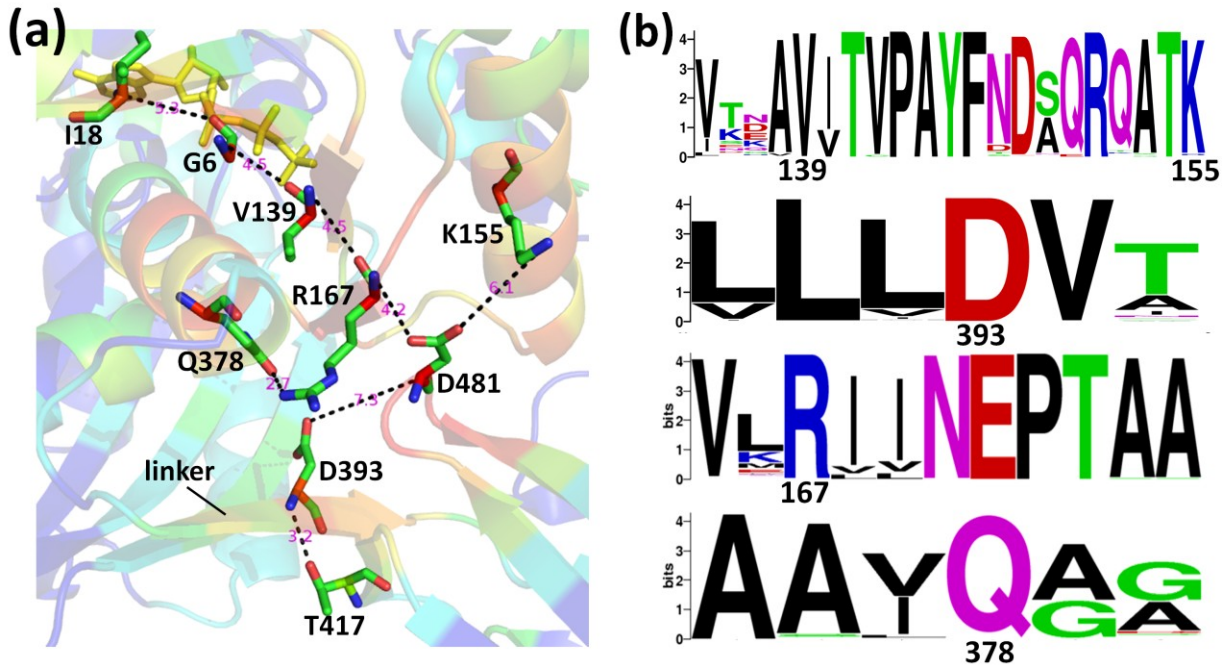

Figure 32. Interactions between the high-susceptibility (HS) residues identified upon perturbing Asp481.

(a) Stick diagram illustrates the HS residues acting as sensors of perturbation in the neighborhood of Asp481 labeled in **Figure 31a**, along with Thr417. The inter-residue interactions are shown as dashed line with distance measure. The ATP is shown in yellow stick representation. The background ribbon diagram of $DnaK_{530}$ is color-coded by $<\|\Delta R^{(481)}\|^2>_{norm}$. (b) Sequence logo plots describing the conservation level of these HS residues. The height of the entire column indicates the level of conservation of the residue, and the size of individual symbols indicates the relative frequency of the corresponding amino acid type(s).

The perturbation of Gly506 yields a different pattern of responses as shown in **Figure 30c**. The most strongly responding structural element is helix 6 (residues Asn147-Ala161) in the ATPase domain. Notably, helix 6 also responds strongly to events at the ATP binding site (**Figure 28a**). In addition, the α-helical lid on the SBD exhibits a strong response, suggesting the perturbation at Gly506 may also affect its docking onto the ATPase domain. It is worth noting that although the α-helical lid makes contribution to the allosteric interactions (Moro et al.,

2003), the allosteric communication can be retained in its absence (Pellecchia et al., 2000), which may explain the less prominent response compared to the other HS sites. However, previous studies have noted that there is a dynamical coupling between the α-helical lid and the inter-domain linker (Liebscher and Roujeinikova, 2009). It is also interesting to note that the subdomain IIB again shows a detectable response, despite its spatial distance. Overall perturbation of G506 appears to elicit a more cooperative and stronger response, compared to the other two key mechanical residues (see Appendix Figures A5 and A6).

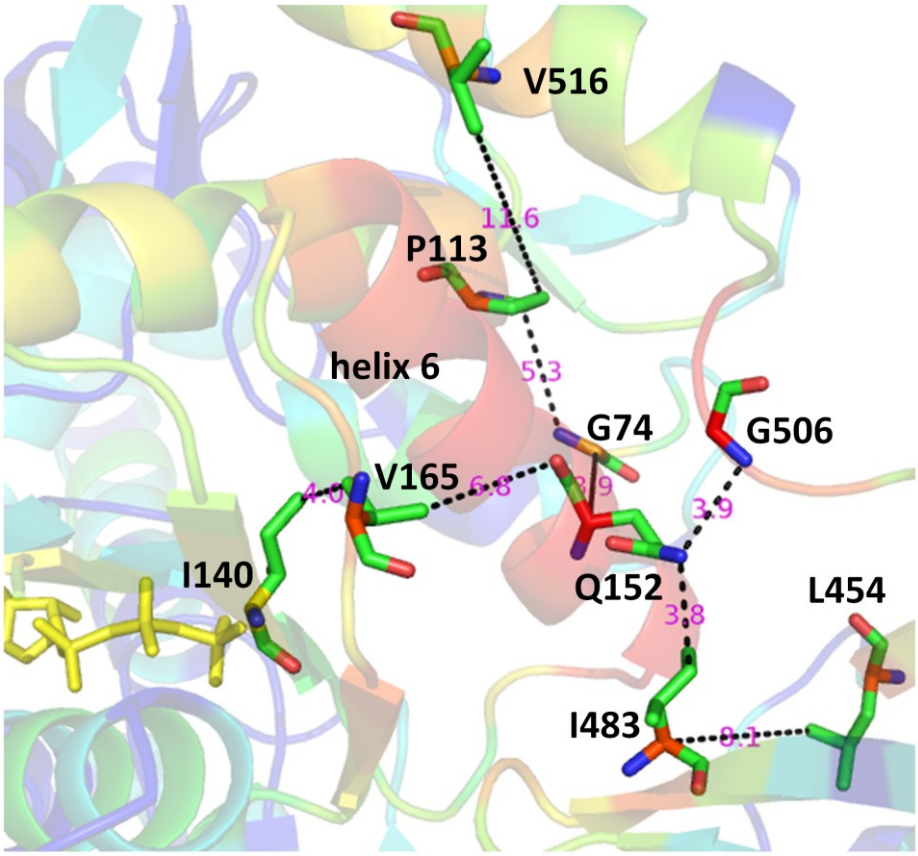

Figure 33. Interactions between the high-susceptibility (HS) residues identified upon perturbing G506.

Stick diagram illustrates the HS residues acting as sensors of perturbation in the neighborhood of G506 labeled in Appendix Figure A6. The inter-residue interactions are shown as dashed line with distance measure. The ATP is

shown in yellow stick representation. The background ribbon diagram of $DnaK_{530}$ is color-coded by $<\|\Delta R^{(506)}\|^2>_{norm}$.

4.2.4 Sequence co-evolution analysis

The analysis of Hsp70 MSAs suggests the involvement of co-evolving residues, in addition to those pointed out above to be highly conserved, in establishing inter-domain allosteric interactions (**Figure 32**). Our approach has been to evaluate the MI maps using equation (29), applied to the same MSA obtained for calculating the conservation profile. While the resulting MI map (shown in Appendix Figure A7) appears to be highly diffuse, the pairs of amino acids which have undergone correlated mutations are easily distinguished by focusing on portions of the MI map. For instance, **Figure 34** displays the inter-domain portion of the MI map. **Figure 35** and **Table 5** give an overview of highly co-evolving residue pairs across the two domains. Of particular interest is the distinctive co-evolutionary propensities of residues 503-505 which are adjacent to the key mechanical residue G506.

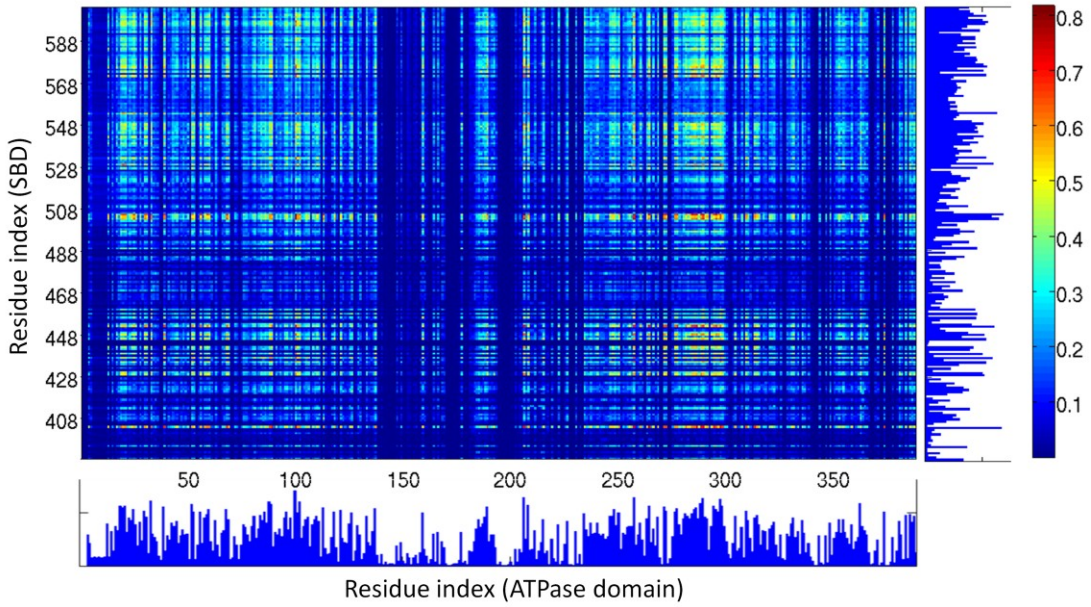

Figure 34. Cross-domain portion of the DnaK MI map.

The bar graphs at the bottom and right hand side of the MI map correspond to the average MI values I(*i*) calculated along the columns and rows of the current MI map, respectively.

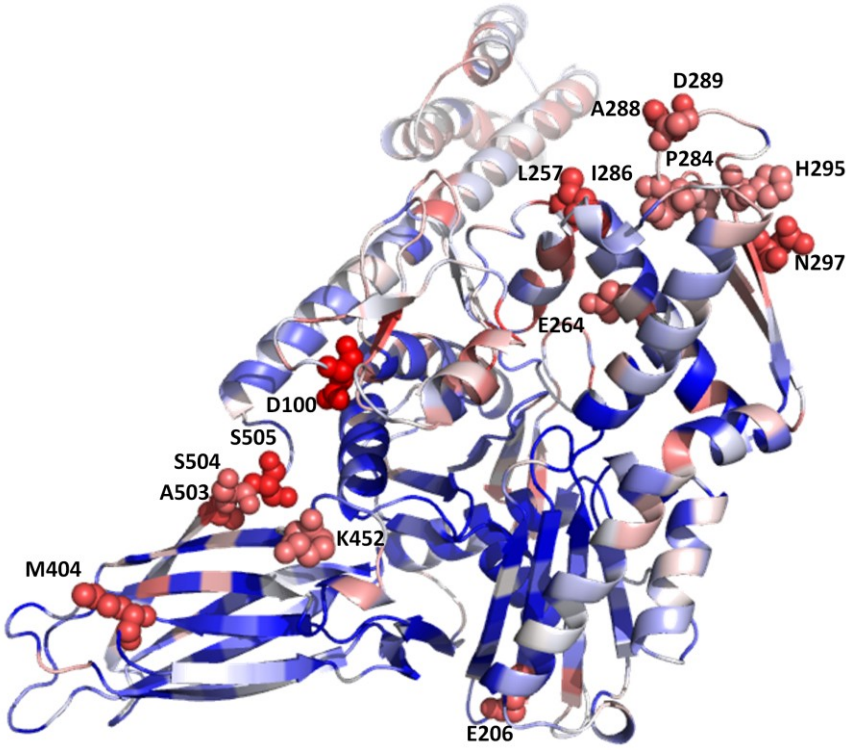

Figure 35. Residues contributing to the top-ranking interdomain I(*i*, *j*).

The ribbon diagram of DnaK is color-coded by the average MI value <I(i)>, increasingly from blue to red. The residues contributing to the top 20 inter-domain I(i, j) pairs are shown in spheres and labeled.

Table 5. Top-ranking 100 pairs inter-domain co-evolving amino acids* in DnaK.

residue pair (i, j)	***I(i,j)***	***residue pair (i, j)***	***I(i,j)***	***residue pair (i, j)***	***I(i,j)***
Glu206-Ser505	0.817	Ile39-Ser505	0.678	Asp289-Ser504	0.645
Asp100-Ser505	0.811	Glu272-Ser505	0.677	Glu206-Ala448	0.645
Asp100-Ser504	0.766	Asp33-Ser505	0.675	Glu272-Met404	0.644
Asp100-Ala503	0.761	Lys294-Ser505	0.674	Ile298-Met404	0.643
Leu257-Ala503	0.759	Ile373-Ser505	0.672	Asn297-Phe529	0.643
Asp100-Met404	0.756	Val313-Met404	0.672	Val59-Ser505	0.643
Glu264-Met404	0.750	Arg56-Ser505	0.670	Asn297-Ala503	0.642
Leu257-Ser505	0.744	Ala68-Ser505	0.669	Ala276-Ser505	0.642
Leu257-Met404	0.724	Ile39-Met404	0.667	Ala68-Ala503	0.641
Asp289-Ser505	0.723	Gln277-Lys452	0.667	Arg25-Ala503	0.641
Glu206-Lys452	0.722	Lys294-Met404	0.666	Asp100-Ala448	0.641
Glu264-Ser505	0.721	Tyr285-Lys452	0.665	Ile373-Met404	0.641
Ile286-Ser505	0.721	Asp100-Lys597	0.663	Asp100-Ala575	0.640
His295-Ser505	0.715	Gln277-Ser505	0.662	Ile373-Lys452	0.640
Asn297-Met404	0.715	Asp289-Lys452	0.662	Val309-Met404	0.638
Ile286-Met404	0.711	His295-Lys452	0.661	Glu206-Ala571	0.638
Asp100-Lys452	0.708	Asp100-Leu532	0.660	Met89-Ser505	0.637
Pro284-Ser505	0.701	Glu272-Lys452	0.657	Asp20-Phe529	0.637
Glu264-Lys452	0.698	Asp100-Ala571	0.657	Thr189-Ser505	0.635
Ala288-Met404	0.698	Ala68-Met404	0.656	Glu206-Ser504	0.635
Glu206-Met404	0.695	His295-Ala503	0.655	Leu257-Ala448	0.634
Ala288-Ser505	0.695	Asp20-Ser505	0.654	Ala288-Ala503	0.634
Asp289-Met404	0.693	Ile39-Lys452	0.653	Ile271-Ala503	0.633
Ile286-Lys452	0.690	Arg25-Ser505	0.653	Met19-Met404	0.632
Leu257-Ser504	0.688	Pro284-Lys452	0.652	Asp208-Ala553	0.631
Pro113-Ser505	0.688	Asn297-Ser504	0.652	His295-Met404	0.630
Glu206-Ala503	0.684	Gly21-Ser505	0.650	Glu31-Ser505	0.630
Asp100-Ala553	0.683	Ala30-Ser505	0.649	Trp102-Ser505	0.629
Met89-Met404	0.683	Asn297-Ser505	0.647	Ile286-Ala503	0.629
Asp289-Ala503	0.680	Ile271-Met404	0.647	Trp102-Phe529	0.629
Leu257-Lys452	0.679	Ile271-Lys452	0.647	Ile88-Met404	0.626
Tyr285-Ser505	0.679	Thr189-Met404	0.647	Arg56-Lys452	0.625
Ile271-Ser505	0.678	Asp100-Thr437	0.646	Val281-Lys452	0.623
				Met89-Lys452	0.623

(*) The rank is based on I(i, j) values (see **Figure 34**)

Examination of the positions and conformations of these pairs in the Hsp70 structural model helps us rationalize these correlations among sequentially distant amino acids on the basis of close tertiary contacts. These contacts are proposed to potentially underlie the close coupling of the two domains. In particular, the linker residues are found to be highly co-evolving with residues from both domains (**Figure 36**). At the ATPase domain side, Val389 appears to co-evolve with Leu177, one of the hydrophobic residues at the linker-binding pocket. Leu177 position, in turn, is highly correlated with that of Ile373 at the pocket adjacent to the linker. Interestingly, both Leu177 and Ile373 have been experimentally shown to be highly sensitive to domain binding (Swain et al., 2007). Their co-evolution may thus be explained by their close hydrophobic interactions. In the core region of the ATPase domain, the sequence evolution at the position of Ile373 is found to be highly correlated with those of Asn13, Ala17, and Met19 on one of the β-hairpin loops that participates in the nucleotide-binding pocket. Closer examination of the correlated substitutions shows that these sites (except for Asn13) retain their hydrophobic character despite the mutations. They essentially form the hydrophobic core of the ATPase domain. Leu177, Ile373 and the linker are involved in transmitting signals upon J-domain protein binding (Jiang et al., 2007). Their sequential and spatial proximity to the HS residues of $<\|\Delta\mathbf{R}^{(389)}\|^2>_{\text{norm}}$ (e.g., to the highly conserved I174 and intermediately conserved V377) suggests a mechanism of allosteric mediation in which these co-evolving residues are implicated.

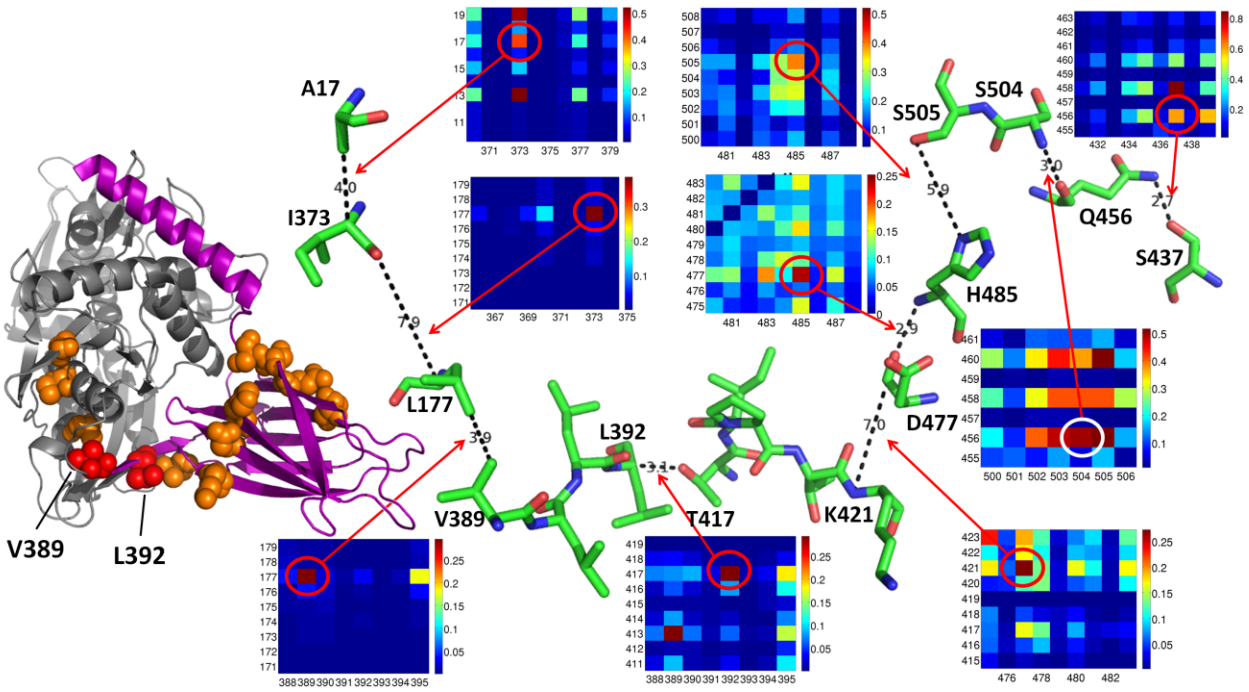

Figure 36. Highly co-evolving residues between the nucleotide-binding site and the substrate binding site, mediated by the inter-domain linker and the key mechanical residue Thr417.

The ribbon diagram on the left displays the highly co-evolving residues in orange spheres, except the linker residues Val389 and Leu392 that are colored red. The co-evolving pairs of amino acids and their relative positions are shown by stick representation. Notably, they form an interdomain communication path. The inter-residue interactions are shown as dashed line with corresponding closest inter-atomic distances. The panels display the regions of the MI matrix corresponding to the identified highly co-evolving residue pairs. The pairs illustrated are marked with red circles and connected to the corresponding dashed line with red arrows.

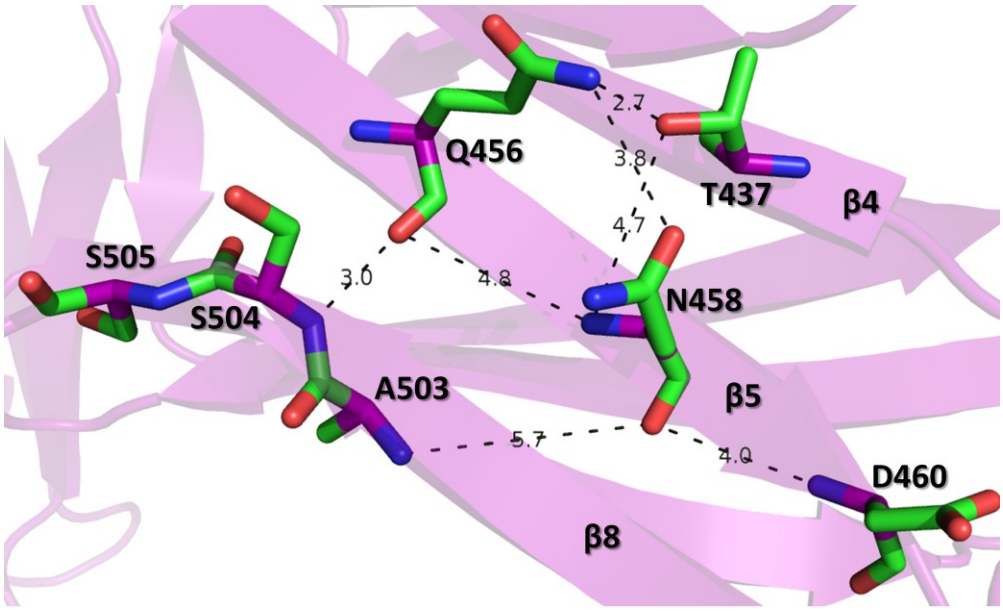

Figure 37. Critical secondary structural contacts at the boundary of the α-helical lid and β-sandwich involve highly co-evolving residues.

The highly co-evolving residues (see **Figure 36**) are shown in stick representation and labeled. The ribbon diagram in the background is the β-sandwich.

At the SBD side, we notice the strong co-evolution between the linker residue Leu392 and key mechanical residue Thr417. These residues were both identified in the sector that mediates inter-domain interactions (Smock et al., 2010). We also notice a number of charged residues across the β-sandwich, namely, Lys421, Asp477, and His485. These residues show relatively weaker response to the interfacial perturbation compared to the exposed end of the SBD β-barrel (**Figure 30**), which may be attributed to their high conformational flexibility/adaptability. But they appear to be instrumental in transmitting the signals to Ser504 (across the β-strands β3, β6, and β7, see **Figure 36**). These residues are located along the loci of the key mechanical residues and obviously Ser504/Ser505 is bonded to Gly506. Perturbation of Lys421 is thus expected to induce a strong response as indicated by the Appendix Figure A8 for $<\|\Delta\mathbf{R}^{(421)}\|^2>_{norm}$. The connection

between the α-helical lid and the β-strand leading to the substrate binding site is manifested by the closely interacting (and highly co-evolving) residue pair Ser504 and Gln456, along with several other highly co-evolving residues on the secondary structure β5 (**Figure 37**).

4.2.5 Summary

In the present work, we investigated the allosteric signal propagation mechanisms within the Hsp70 through physics-based perturbation analysis combined with sequence co-evolution analysis. Our computational strategy relies little on the side chain orientation of residues, and is therefore insensitive to possible inaccuracies in atomic positions, justifying the use of the homology model. Several key mechanical residues at the interdomain interface as well as the linker are shown to have restricted mobility, suggesting their important role in serving as a hinge center in for the concerted motions of the two domains of Hsp70 in the ATP-bound state. Physical perturbations of these residues' coordinates revealed the highly susceptible sites and probable patterns of signal propagation via perturbations of co-evolving residue pair contacts.

Our results point to a number of key residues on the ATPase domain that propagate the interfacial perturbation to the nucleotide binding site. Sequence analysis also indicates specific interactions in this region. Certain secondary structure elements are found to mediate distant communications in the ATPase domain. For instance, helix 10 couples subdomains IA and IIB (**Figure 30a** and **30b**), and helix 8 in subdomain IIB mediates the coupling of the nucleotide binding site to NEF-binding site, indicated by the results obtained for perturbing the γ-phosphate (**Figure 28**).

In the SBD, it is interesting to observe that the interfacial perturbation propagates all the way to the exposed end of the β-barrel despite the relatively weaker responses of the β-strands.

Sequence co-evolution patterns among residues participating in the β-sheet disclose close interactions which may be important to maintaining the long-range coupling. These secondary structure elements (the loop connecting the α-helical lid and the β-sandwich, and the strands β3, β5-β7 in the β-sandwich) are not particularly flexible, presumably due to the need to stabilize the conformation of the molecule; yet their small displacement may induce effects on distal regions that are less constrained, as observed here at the exposed end of the SBD.

5.0 SEQUENCE EVOLUTION CORRELATES WITH STRUCTURAL DYNAMICS

The role of structural dynamics in enabling protein function has been underlined in recent work (Bhabha et al., 2011). In some cases, dynamics is manifested by large-scale collective motions of intact substructures. Examples are the opening/closing of domains around a catalytic cleft, or the allosteric switches that cooperatively engage multiple subunits in multimeric structures. Many enzymes and molecular machines such as the bacterial chaperonin or the ribosome undergo such concerted motions triggered by substrate binding (Tama and Brooks, 2006; Yang et al., 2009; Bahar et al., 2010). These are usually referred to as *global motions* due to their collective nature. In other cases, the motions are *local*, e.g., rearrangements of recognition loops or rotational isomerizations of side chains.

Global motions are predominantly encoded by the architecture of the protein. Models based exclusively on native contact topology, such as ENMs, have proven to closely reproduce the structural variabilities observed in experiments for proteins resolved in multiple substrate-bound forms (Bakan and Bahar, 2009; Bahar et al., 2010). The fact that these motions are uniquely and robustly defined by the architecture, suggests that native folds may have evolved to favor functional motions. This also suggests that there are key mechanical sites that control the global movements while preserving the stability of the fold. To date, no systematic study of the evolutionary conservation properties of amino acids in relation to the structure-encoded dynamics of proteins has been performed to our knowledge.

Local motions, on the other hand, may facilitate the recognition of substrates, optimize binding interactions, usually complementing global motions (e.g., domain closure) or accompanying structure formation upon substrate binding (Wright and Dyson, 2009). Substrate recognition sites tend to exhibit suitable sequence variations so as to enable specific recognition (Liu et al., 2010); and at the same time, they may enjoy structural flexibility, consistent with conformational adaptability required for mediating substrate specificity (James et al., 2003). In contrast, conserved residues are highly ordered, as evidenced by NMR relaxation experiments (Mittermaier et al., 2003). Our examination of the collective dynamics of catalytic sites (Yang and Bahar, 2005) and metal-binding proteins (Dutta and Bahar, 2010) also showed that residues involved in biochemical activities exhibit minimal fluctuations.

All these observations suggest that sequence variability and structural dynamics go hand in hand, i.e., the need to sample functional motions may underlie the evolutionary selection of amino acids that encode the 'proper' fold which lends itself to those required motions as its softest modes of conformational change naturally accessible under physiological conditions. Yet, the prevalence of such a relationship remains to be analytically investigated and established.

In the present study, we present the results from the analysis of 34 enzymes that represent a diversity of protein families, functional classes and sizes (**Table 6**). For each enzyme, we determined the relative mobility each residue enjoys in the collective dynamics, on the one hand, and the amino acid conservation or correlated mutation propensities at the corresponding sequence position, on the other. Our analysis shows that (i) conserved residues have minimal fluctuations in the global modes, their high stability being a prerequisite for their precise functioning, (ii) increase in sequence variability is accompanied with increase in conformational mobility, this feature being most distinctive at intermediate levels of conservation/mobility

typical of co-evolving pairs of amino acids, (iii) the co-evolving residues fall into two groups: those at highly flexible/mobile regions in the global modes, involved in substrate recognition; and those in the close neighborhood of catalytic or ligand-binding sites assist in stabilizing the ligands and/or transmitting signals from/to the active site, (iv) it is possible to define an intrinsic mobility scale for the twenty types of amino acids, which is inversely proportional to the conservation propensity of amino acids, and may be utilized for customizing protein dynamics.

5.1 OVERVIEW OF THE PROCEDURE

Figure 38 illustrates the method of approach. We adopted a two-pronged analysis for each enzyme: (i) perform a GNM analysis of collective dynamics using the PDB structure, and (ii) analyze the residue conservation and co-evolution properties using the MSA retrieved from the Pfam DB.

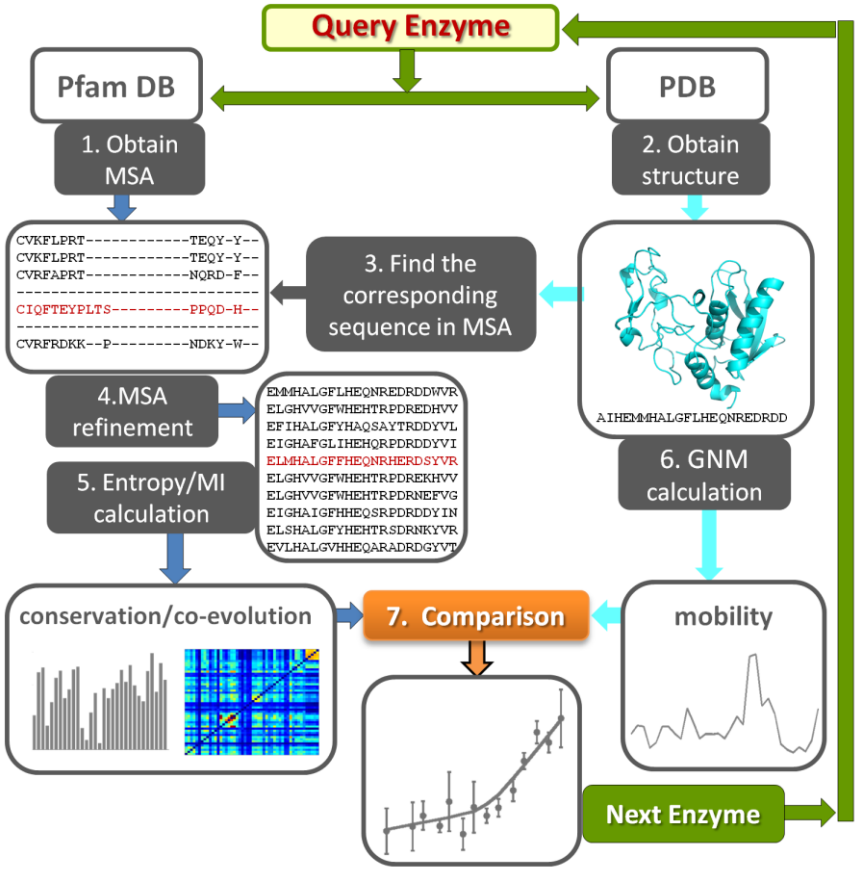

Figure 38. Workflow of the study of 34 enzymes.

For each query enzyme in the dataset, we retrieve the structure from the PDB and the MSA from Pfam DB. These are used as input for (1) GNM evaluation of residue mobilities (*right* branch), and (2) generation of conservation profile and co-evolution (MI) maps (*left* branch), respectively. Comparison of the outputs indicates that sequence entropy is accompanied by conformational mobility (enhanced dynamics), correlated mutations exhibit a broad range of mobilities depending on the type of underlying evolutionary pressure, and conserved sites are practically immobile. Results are consolidated by compiling the results for all 34 studied enzymes.

The dataset used in a previous study (Zen et al., 2008) was adopted as starting point. This dataset contained 76 enzymes with a broad range of functions. Among them, we focused on the monomeric X-ray structures that contained at least 120 structurally resolved residues in the PDB.

For each enzyme, the MSA for the corresponding protein family in the Pfam DB was retrieved, and further refined using the following procedure:

(1) iteratively align the primary sequence from the PDB structure (of the query enzyme) with each sequence in the MSA using the Smith-Waterman algorithm (Smith and Waterman, 1981), and identify a matched sequence with the highest score. If the matched sequence has less than 95% sequence identity with respect to the PDB sequence, it is discarded from the dataset since the PDB sequence is not well represented in the MSA.

(2) based on the residue mapping between the PDB sequence and the matched sequence, truncate the columns of the MSA so as to retain those in the PDB sequence.

(3) remove the redundant sequences in the refined MSA using a threshold of 99%, and eliminate the sequences that have more than 20% gaps.

The procedure described above corresponds to steps 3 and 4 in **Figure 38**, and yielded 34 proteins that formed our final dataset (see **Table 6**). The numbers of rows (sequences) and columns (residues) in each refined MSA are also presented in **Table 6**.

The GNM analysis yields a *mobility profile* for each enzyme. The mobility profile obtained with the contribution of all modes scales with the MSFs of residues. *N*-1 GNM modes of motion contribute to MSFs for an enzyme of *N* residues. Among them, the low frequency modes, also called the *soft* modes or global modes, play a dominant role in defining the most cooperative events. We examined the contribution made by these modes to MSFs. To this aim, we considered subsets of m_1 and m_2 modes at the low frequency end of the mode spectrum, which make fractional contributions of 0.1, and 0.4, respectively, to collective dynamics (**Table 7**).

The MSAs are utilized to generate the *conservation profile* as a function of residue index, and the correlated mutation maps for each pair of residues. The level of conservation is

expressed by the Shannon information entropy S(i) for each sequence position i; and the co-evolutionary propensities are evaluated using the mutual information (MI) theory (Dunn et al., 2008; Liu et al., 2010). Comparison of mobility profiles and conservation/co-evolution trends for each enzyme, consolidated over the entire dataset, discloses three different classes of residues based on their mobility/evolution behavior. Conserved residues distinguished by S(i) values below a threshold undergo minimal changes in their positions in the 3D structure. Conversely, the sites that exhibit uncorrelated variations in their amino acid identity display enhanced mobilities, although the extent of mobility broadly varies. In the intermediate regime, which includes the majority of co-evolving residues, there is a linear increase in mobility with increasing sequence entropy. These results highlight the importance of structural adaptability in sustaining the functional dynamics of the enzyme notwithstanding sequence variations that confer specificity.

Table 6. Dataset of 34 enzymes, their Protein Data Bank (PDB) and Pfam identifiers, and the properties of MSAs

PDB id	*protein name and reference*	*# of matched residues (range)* [a]	*Pfam id*	*# of seqs in MSA*	*EC code (Webb, 1992)*
1.3cd2	dihydrofolate reductase	199 (6-204)	PF00186	91	1.5.1.3
2. 1k03	nadph dehydrogenase 1	353 (15-367)	PF00724	2712	1.6.99.1
3. 1fp9	4-α-glucanotransferase	487 (11-497)	PF02446	642	2.4.1.25
4. 1ajz	dihydropteroate synthase	206 (20-225)	PF00809	1930	2.5.1.15
5. 1u32	Ser/Thr prot phosphatase 1, γ catalytic subunit	196 (57-252)	PF00149	3658	3.1.3.16
6. 2f6f	Tyr protein phosphatase, non-receptor type1	237 (40-276)	PF00102	1157	3.1.3.48
7. 2ffz	phospholipase C	241 (1-241)	PF00882	80	3.1.4.3
8. 1ako	exonuclease III	266 (1-266)	PF03372	4019	3.1.11.2
9. 1vas	T4 endonuclease V	135 (2-136)	PF03013	39	3.1.25.1
10. 1goc	ribonuclease H	141 (2-143)[b]	PF00075	4172	3.1.26.4
11. 1bol	ribonuclease Rh	206 (1-206)	PF00445	96	3.1.27.1
12. 1k2a	eosinophil-derived neurotoxin	130 (5-134)	PF00074	328	3.1.27.5
13. 1kab	staphylococcal nuclease	109 (33-141)	PF00565	1057	3.1.31.1
14. 1b1y	β-amylase	423 (13-435)	PF01373	133	3.2.1.2
15. 2fba	glucoamylase GLU1	457 (27-483)	PF00723	191	3.2.1.3
16. 3eng	endoglucanase V cellobiose complex	199 (2-200)	PF02015	81	3.2.1.4
17. 1bhe	polygalacturonase	341 (36-376)	PF00295	607	3.2.1.15
18. 2ayh	1,3-1,4-β-d-glucan 4-glucanohydrolase	185 (26-210)	PF00722	1061	3.2.1.73
19. 1dy4	cellobiohydrolase I	431 (2-432)	PF00840	170	3.2.1.91
20.4skn[c]	uracil-DNA glycosylase	162(131-292)	PF03167	1599	3.2.2.3
21. 8cpa	carboxypeptidase A	279 (18-296)	PF00246	854	3.4.17.1
22. 3pbh	procathepsin B	250 (1-250)	PF00112	1970	3.4.22.1
23. 1avp	adenoviral proteinase	183 (20-202)	PF00770	43	3.4.22.39
24. 1qjj	astacin	192 (8-199)	PF01400	590	3.4.24.21
25. 1f82	botulinum neurotoxin type B	416 (2-417)	PF01742	50	3.4.24.69
26. 1lba	T7 lysozyme	128 (6-133)	PF01510	1685	3.5.1.28
27. 1lqy	peptide deformylase 2	170 (4-173)	PF01327	1581	3.5.1.88
28. 1ko3	VIM-2 metallo-β-lactamase	179 (62-240)	PF00753	8255	3.5.2.6
29. 1rgy	β-lactamase	350 (12-361)	PF00144	2742	3.5.2.6
30. 2had	haloalkane dehalogenase	231 (75-305)	PF00561	8679	3.8.1.5
31. 1v9i	carbonic anhydrase II	254 (6-259)	PF00194	618	4.2.1.1
32. 1vbl	pectate Lyase	220(111-330)	PF00544	362	4.2.2.2
33. 2plc	phosphatidylinositol-specific phospholipase C	140 (39-178)	PF00388	568	4.6.1.13
34. 1h0p	peptidyl-prolyl *cis-trans* isomerase 5	158 (31-188)	PF00160	3594	5.2.1.8

a. The residue range corresponds to the residues in the PDB file that were aligned with the matched sequence in the MSA.

b. Residue 81 of 1goc was not present in the matched sequence; therefore the number of residues is 141 instead of 142.

c. Here and in the following tables, the entries for the example enzyme (see **Figures 38** and **40**) are highlighted to let the reader easily locate the corresponding data.

Table 7. Number of GNM modes included in generating the mobility profiles for the 34 enzymes

PDB id	*Number of modes*[(a)]			*PDB id*	*Number of modes*[(b)]		
	m_1	m_2	N-1 [(b)]		m_1	m_2	N-1
3cd2	1	9	205	2ayh	2	13	213
1k03	2	20	398	1dy4	2	23	432
1fp9	2	18	499	4skn	1	11	222
1ajz	2	14	281	8cpa	2	16	306
1u32	2	15	292	3pbh	2	15	254
2f6f	2	13	301	1avp	1	12	214
2ffz	1	12	244	1qjj	1	10	202
1ako	2	13	267	1f82	2	16	423
1vas	1	6	136	1lba	1	7	145
1goc	1	7	155	1lqy	1	11	183
1bol	1	11	221	1ko3	2	14	229
1k2a	1	8	135	1rgy	2	15	359
1kab	2	10	135	2had	2	16	309
1b1y	2	23	499	1v9i	2	16	260
2fba	2	21	491	1vbl	1	19	415
3eng	2	15	212	2plc	2	17	273
1bhe	2	21	375	1h0p	2	15	181

[(a)] m_1, m_2 refer to the subset of modes that account for 10% and 40%, respectively of the overall dynamics.
[(b)] N is the number of residues in the PDB file

5.2 SEQUENCE ENTROPY VS. CONFORMATIONAL MOBILITY

5.2.1 An illustrative example

Some of the basic steps and outcomes are illustrated for a DNA repair enzyme, uracil-DNA glycosylase (UDG), in **Figure 39.** Panel a displays the mobility profiles based on m_1, m_2 and N-1 GNM modes. In UDG, m_1 = 1, i.e., the softest mode alone accounts for >10% of the dynamics (see highlighted entry in **Table 6**). $\langle M_i \rangle|_{m1}$ shows the distribution of square displacements of residues in this softest mode. $\langle M_i \rangle|_{N\text{-}1}$ scales with the MSF profile of residues, and contains contributions from both global and local motions; yet the shape of the curve is dominated by

slow/soft modes as the close resemblance to $<M_i>|_{m2}$ reveals. The gray bars in **Figure 39a** represent the Shannon entropy profile. Peaks represent the most variable sites, and minima, the most conserved. Notably, mobility and entropy distributions exhibit similarities, as also evidenced by the color-coded ribbon diagrams displayed in panel b.

The relation between sequence variations and structural dynamics at the level of individual residues is clearly seen by evaluating the *effective mobilities* based on entropy bins of $\Delta S = 0.1$ and compiling the results for all enzymes in our database. To allow for the compilation and combined analysis of the results for all enzymes, the entropy and mobility profiles of residues in each enzyme have been normalized to represent probabilistic distributions, and then uniformly rescaled by the number of residues in that particular enzyme. This way we eliminate the dependence of the resulting mobility/conservation profiles on the size of the proteins. We have further evaluated effective properties using a grid-based mapping scheme. The basic idea therein is to cluster residues with similar entropy (using bin sizes of $\Delta S = 0.1$) and assign an average mobility to each bin. The resulting $<M_i^{eff}>|_{N\text{-}1}$ values yielded a correlation of 0.82 with sequence entropy, while the plot for individual residues gave a correlation of 0.52 (see Appendix Figure A9). This observation underscores the significance of consolidating the outputs with an ensemble of proteins, rather than examining single proteins where the patterns are barely detectable.

The MI map in **Figure 39c** displays the co-evolutionary properties of UDG residue pairs. Yellow/red regions indicate the residue pairs that exhibit high MI values, i.e., the loci of correlated mutations. The upper right portion of the map magnified in panel c reveals the high co-evolutionary properties of residues near DNA-binding site, shown in panel d. The curve under

the map shows the average MI, < I(*i*) >, for each column *i*, a metric of the co-evolution propensity of residue *i*.

Our previous examination of sequence evolution properties of Hsp70 ATPase domain in relation to its intrinsic dynamics suggested that among co-evolving residues those distinguished by high mobility in the global modes serve as substrate recognition sites (Liu et al., 2010). The same observation, recognition enabled by conformationally mobile, sequentially correlated residues, was also made for PDZ domains by Kosik and coworkers (Sakarya et al., 2010). E182, D183, R276 and E282 are such residues in UDG (**Figure 39a**). Notably, as evidenced by the structure shown in **Figure 39d**, the residues R276-G280 do interact with DNA, consistent with these earlier observations (for other systems).

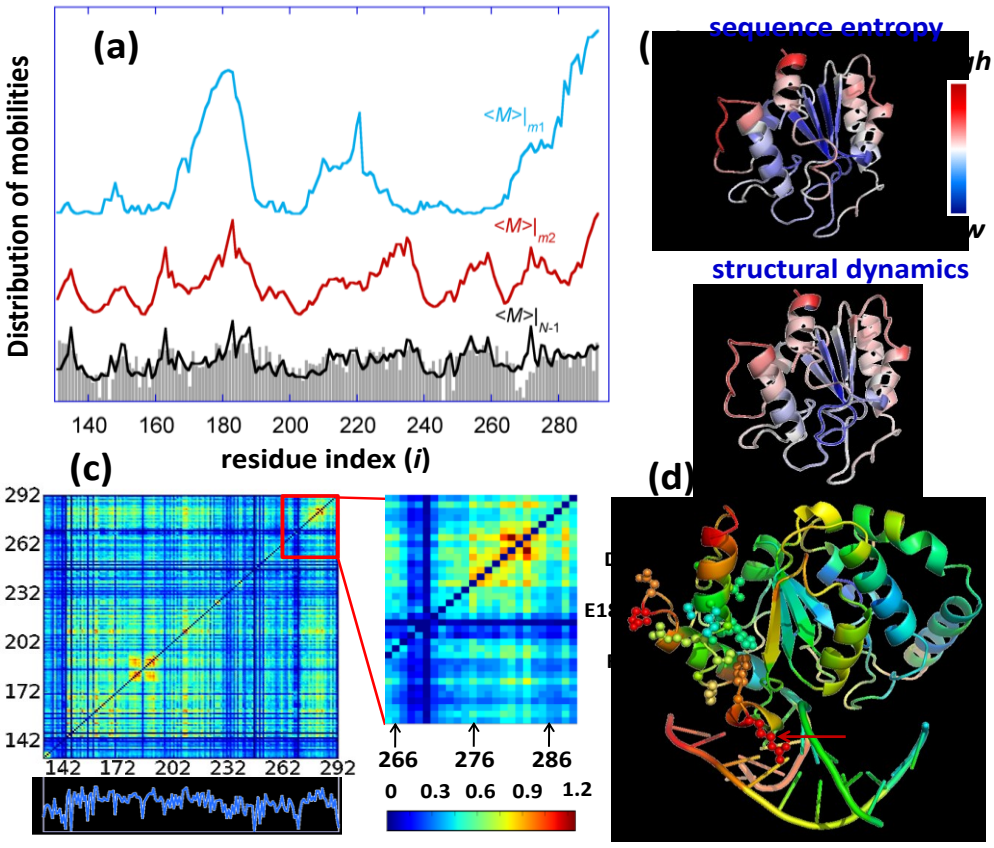

Figure 39. An illustrative example: comparative analysis of residue conservation, conformational mobility and co-evolutionary patterns for uracil-DNA glycosylase (UDG).

(a) Conformational mobility and residue conservation as a function of residue index. Blue, red, and black curves represent the mobility profiles $<M_i>|_{m1}$, $<M_i>|_{m2}$, and $<M_i>|_{N-1}$ (or MSFs) computed using the GNM. The curves are shifted vertically for clearer visualization. The bars represent the information entropy derived from 1599 Pfam sequences (**Table 6**). Results are shown for the structurally resolved residues $131 \leq i \leq 292$ that are fully represented in the MSA. (b) Comparison of conservation (*upper*) and mobility (*lower*) profiles using color-coded ribbon diagrams. (c) MI map for the same family. The lower curve displays the co-evolution propensity of individual residues <I(*i*)>, averaged over all entries in the corresponding column/row of the MI map. The portion of the map corresponding to DNA-binding residues is magnified. (d) Location of highly co-evolving residues (shown in spheres) and their involvement in DNA binding. The diagram is color-coded based on the X-ray crystallographic B-factors (red/blue: most/least mobile) reported for UDG.

5.2.2 Sequence entropy *vs.* conformational mobility for all enzymes

We repeated the comparative analysis summarized for UDG for all 34 enzymes in our dataset. The results, compiled in the **Table 8** confirm that the MSFs of residues and their substitution/mutation propensities exhibit weak but statistically significant correlations; and these correlations become particularly apparent when the results for the complete set of enzymes are consolidated for sequence entropy intervals of $\Delta S = 0.1$. **Figure 40a** shows the results for all the 8,254 residues in our dataset of 34 enzymes. The dots show the effective mobilities $<M_i^{\text{eff}}>|_{m1}$ (red, filled) and the MSFs (or $<M_i^{\text{eff}}>|_{N-1}$) (blue, open). The number distribution of residues in each entropy interval is shown by the histogram (gray bars). $<M_k^{\text{eff}}>|_{m2}$ values (not shown) closely approximate the MSFs.

Several interesting features are observed in **Figure 40a**. First, the coupling between structural dynamics and sequence variability is more pronounced when the global motions driven

by a few soft modes (m_1 = 1-2; **Table 7**) are examined, as opposed to the resultant of all *N-1* modes.

Table 8. Pearson correlation coefficients between sequence-based entropy and structure-based mobility profiles based on m_1, m_2 and N-1 modes

PDB id	$\langle M \rangle_{m1}$			$\langle M \rangle_{m2}$			$\langle M \rangle_{N-1}$		
	correlation coefficient[(a)]		*p-value*	*correlation coefficient*		*p-value*	*correlation coefficient*		*p-value*
3cd2	0.41	0.75	6.04e-10	0.36	0.76	9.62e-08	0.33	0.74	9.91e-07
1k03	0.50	0.73	0.00e+00	0.46	0.76	0.00e+00	0.46	0.85	0.00e+00
1fp9	0.19	0.32	1.36e-05	0.33	0.68	2.44e-14	0.38	0.78	0.00e+00
1ajz	0.33	0.86	4.86e-07	0.54	0.88	0.00e+00	0.54	0.91	0.00e+00
1u32	0.45	0.91	1.32e-11	0.47	0.87	1.22e-12	0.47	0.81	1.90e-12
2f6f	0.36	0.86	6.27e-09	0.47	0.93	8.99e-15	0.51	0.94	0.00e+00
2ffz	0.48	0.68	7.77e-16	0.56	0.88	0.00e+00	0.56	0.92	0.00e+00
1ako	0.33	0.84	1.63e-08	0.47	0.86	4.44e-16	0.48	0.82	0.00e+00
1vas	0.35	0.78	1.23e-05	0.39	0.68	1.81e-06	0.36	0.58	9.03e-06
1goc	0.33	0.55	3.24e-05	0.41	0.78	2.22e-07	0.41	0.85	1.68e-07
1bol	0.12	0.23	4.66e-02	0.19	0.59	2.93e-03	0.25	0.59	1.76e-04
1k2a	0.11	0.43	1.03e-01	0.57	0.83	9.91e-13	0.61	0.86	7.11e-15
1kab	0.23	0.15	8.34e-03	0.39	0.71	1.71e-05	0.44	0.84	6.52e-07
1b1y	0.27	0.81	1.31e-08	0.32	0.79	4.01e-12	0.37	0.90	3.11e-15
2fba	0.34	0.91	8.19e-14	0.37	0.78	2.22e-16	0.35	0.70	1.22e-14
3eng	0.20	0.48	2.49e-03	0.31	0.60	3.00e-06	0.35	0.65	2.25e-07
1bhe	0.41	0.87	6.66e-16	0.45	0.87	0.00e+00	0.44	0.89	0.00e+00
2ayh	0.11	0.16	7.47e-02	0.34	0.73	1.10e-06	0.39	0.80	1.46e-08
1dy4	0.21	0.61	7.96e-06	0.47	0.72	0.00e+00	0.51	0.77	0.00e+00
4skn	0.18	0.72	1.27e-02	0.50	0.80	8.36e-12	0.52	0.82	5.38e-13
8cpa	0.31	0.72	4.41e-08	0.49	0.84	0.00e+00	0.51	0.89	0.00e+00
3pbh	0.36	0.71	2.69e-09	0.47	0.82	1.33e-15	0.52	0.85	0.00e+00
1avp	0.18	0.41	8.01e-03	0.34	0.62	1.43e-06	0.38	0.63	4.70e-08
1qjj	0.14	0.50	2.80e-02	0.33	0.75	1.03e-06	0.38	0.77	1.78e-08
1f82	0.38	0.79	9.99e-16	0.44	0.76	0.00e+00	0.46	0.82	0.00e+00
1lba	0.43	0.69	1.56e-07	0.49	0.85	2.20e-09	0.49	0.95	2.01e-09
1lqy	0.36	0.78	5.07e-07	0.47	0.76	3.11e-11	0.40	0.65	2.17e-08
1ko3	0.29	0.64	5.42e-05	0.39	0.86	4.48e-08	0.35	0.56	5.60e-07
1rgy	0.21	0.62	3.91e-05	0.35	0.79	1.58e-11	0.36	0.83	1.51e-12
2had	0.40	0.75	9.86e-11	0.35	0.83	2.71e-08	0.36	0.84	1.29e-08
1v9i	0.31	0.82	1.80e-07	0.43	0.92	2.56e-13	0.44	0.90	1.26e-13
1vbl	0.48	0.75	1.77e-14	0.60	0.79	0.00e+00	0.63	0.85	0.00e+00
2plc	0.36	0.68	6.94e-06	0.45	0.75	9.97e-09	0.49	0.64	2.90e-10
1h0p	0.21	0.47	3.58e-03	0.37	0.69	7.26e-07	0.39	0.69	2.73e-07
AVG	**0.31**	**0.65**		**0.42**	**0.78**		**0.44**	**0.79**	

[(a)] The two correlation coefficients refer to results obtained for individual residues (left) and those based on entropy intervals of size 0.1.

Second, this dependence is not linear. Higher sequence entropy (or lower conservation) is accompanied by increased mobility as expected, but this increase does not take effect until the entropy reaches a threshold value of $S(i) \approx 0.8$ (orange arrow). In the range $S(i) < 0.8$, the global mobility is minimal with little dependency on the conservation level. About $1/4^{th}$ of residues lie in this regime. Then, there is a sharp increase in mobility tied in with decrease in entropy.

Sequence variability above this threshold value cannot presumably be sustained unless the global dynamics endows suitable structural flexibility. In the other extreme case of high entropy regime ($S(i) > 1.5$, delimited by green arrow), residues exhibit a broad variation in their mobility, partly due to the scarcity of data (9% of residues lie in this regime). Therefore, we distinguish three regimes, with the strictest dependence on mobility manifested at the intermediate level $0.8 \leq S(i) \leq 1.5$ of sequence entropy.

Third, the histogram for entropy (gray bars in **Figure 40a**) exhibits a unique behavior with a peak at the most conserved region (leftmost bar), thus departing from a unimodal distribution. This peak refers to fully conserved residues. The size of this group (322 residues) is much larger than that expected for a normal distribution tail. Calculations confirm that this subgroup of residues exhibit minimal fluctuations (see Appendix Figure A10**)**. In contrast, the most variable group (the rightmost bar in the histogram) contains 117 residues that span a wide range ($1.9 \leq S(i) < 2.9$) of entropy and effective mobility, preferentially sampling larger fluctuations in space (Figure A9**)**.

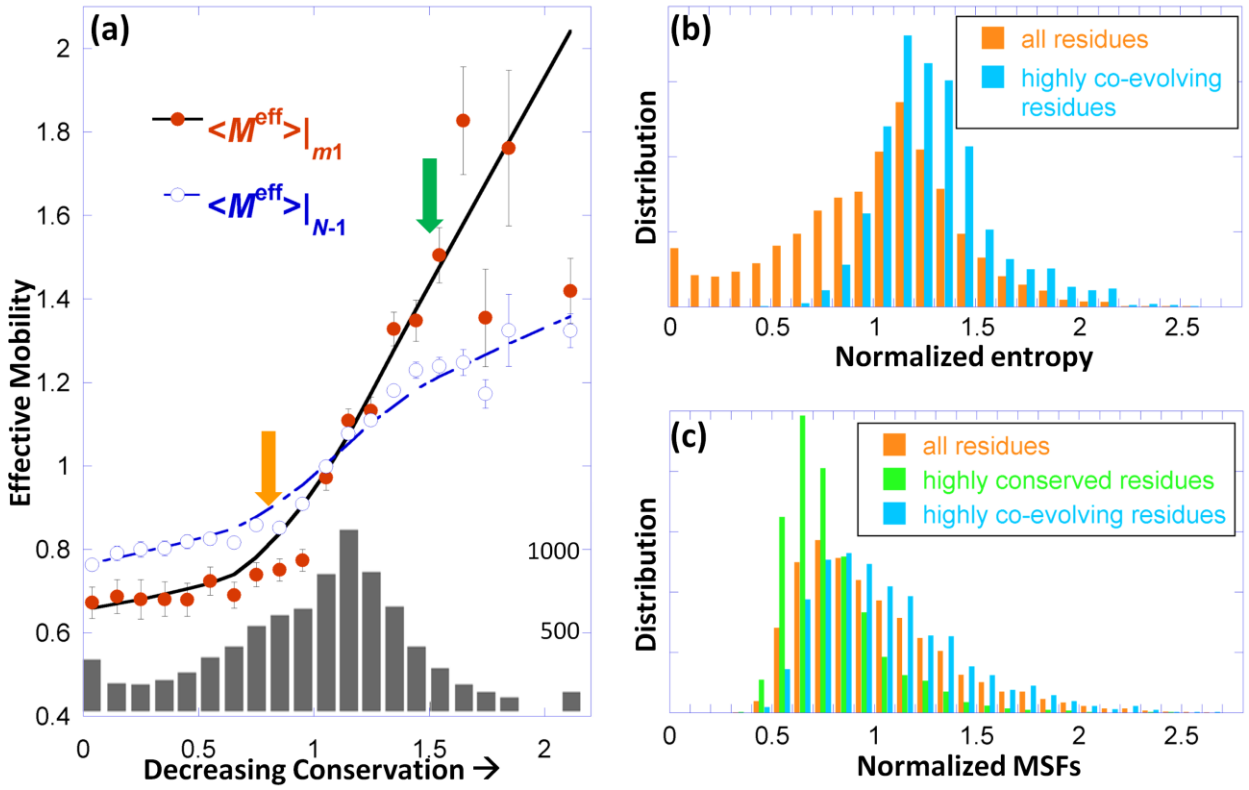

Figure 40. Relationship between structural dynamics and sequence evolutionary properties.

(a) Effective mobility as a function of sequence conservation. Computed data are based on softest modes (red circles) or *N*-1 modes (open circles) as a function of the level of conservation, deduced from the analysis of all 34 enzymes in the dataset. The histogram shows the number of residues in consecutive sequence entropy intervals (right ordinate). The curves are the weighted least square fits, with respective correlation coefficients of 0.90 and 0.95. Entries with $S(i) > 2$ are merged in the last bin. Arrows delimit distinctive mobility *vs.* conservation regimes. (b) Sequence entropy distribution for all residues (orange) and for the highly co-evolving residues identified by MI analysis of all enzymes (cyan). (c) Mobility histograms for three groups of residues, as labeled. Respective mean values and variances are 1.00 ± 0.134, 0.79 ± 0.059, and 1.06 ± 0.127.

5.3 BROAD RANGE OF MOBILITY EXHIBITED BY HIGHLY CO-EVOLVING RESIDUES

We evaluated the MI maps for all the enzymes in our dataset, and identified the residues that yielded the highest MI values. **Figure 40** panels **b** and **c** show the respective conservation and mobility distributions (cyan bars) evaluated for the residues that yielded the top 20% $\langle I(i) \rangle$ values (1,639 of them), referred to as highly co-evolving residues. Panel b compares their sequence entropy distribution to that of the entire set (orange). Notably, a large majority (82%) of highly co-evolving residues fall in the intermediate entropy regime identified above. And the distributions in **Figure 40c** show that these residues tend to enjoy larger mobilities compared to 'all' residues. This tendency may be associated with their substrate recognition role and positioning on the protein surface, as noted for UDG (**Figure 39**). Panel c also displays the histogram (green) for the most conserved sites, referred to as *C*-sites (lowest 20% $S(i)$ values), again showing their lower mobility compared to all residues.

Co-evolution of amino acids appears to enable the adaptability of ubiquitous proteins or their modular domains to cope with diverse substrates (Gotoh, 1992; Liu et al., 2008; Xu et al., 2009; Liu et al., 2010; Smock et al., 2010). Our earlier study invited attention to the enhanced global mobility of such sites involved in substrate recognition (Liu et al., 2010). Observations made here further support this notion. **Figure 41** illustrates the results for procathepsin B (Podobnik et al., 1997). Results for other proteins (staphylococcal nuclease, T7 lysozyme, carbonic anhydrase II and carboxypeptidase A) may be seen in **Figure 42** and in the Appendix Figure A11 and **Table 9**. In all cases, a number of co-evolving residues are detected at the highest peaks in the global mode, and these residues are noted to assist in substrate recognition. **Figure 41** shows that in procathepsin B the residues distinguished by their strong MI values lie

in the occluding loop N113-T125 that is involved in substrate recognition (Illy et al., 1997) and inhibitor binding (Renko et al., 2010). Among fifteen residues that yield the top 0.05% MI values, ten (N113-P117, G121-T125) belong to this loop. **Figure 41b** shows the pronounced mobility of this loop in the softest mode.

It is important to note that not all highly-co-evolving residues are subject to large mobility, i.e., co-evolution is not always confined to substrate recognition sites on the surface. Residues involved in substrate binding near the catalytic site or in signal transduction may also exhibit co-evolutionary trends, if they are not conserved. Binding and signaling are achieved more efficiently in the case of tight packing and minimal energy dissipation or residue fluctuations in the global modes. The inhibitor-bound structure of cathepsin B (Renko et al., 2010) shown in the inset of **Figure 41b**, presents two such sites, C67 and G68 (purple), in close spatial proximity of other highly-co-evolving residues; the restricted mobility of these two residues in the global mode suggests a signal transduction role.

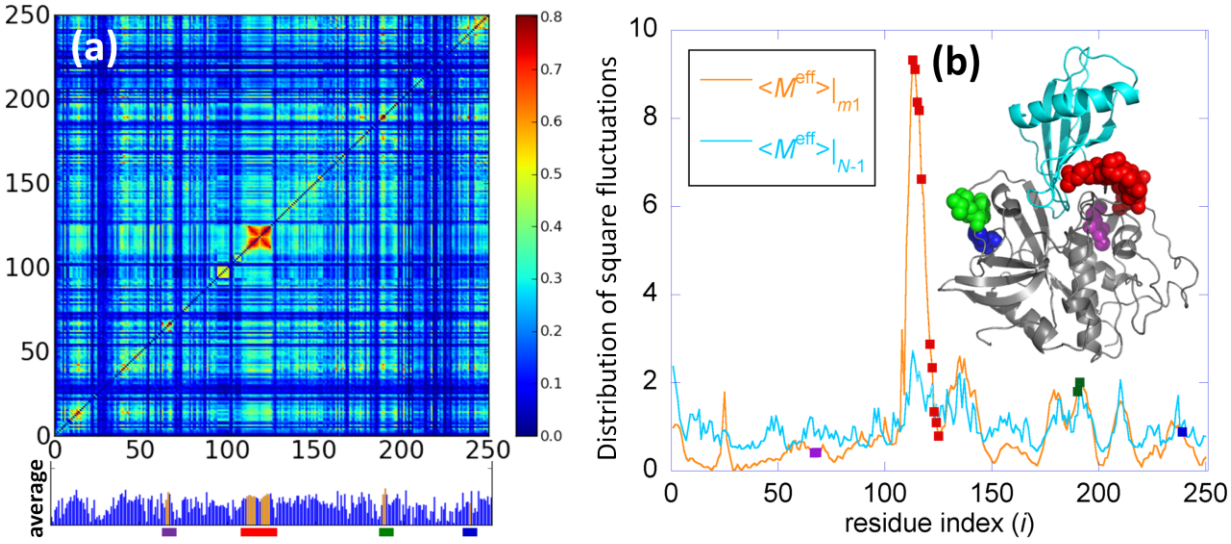

Figure 41. Sequence co-evolution and high mobility properties at the ligand recognition site of procathepsin B catalytic domain.

(a) MI map, highlighting highly co-evolved amino acid pairs such as (N113, T125) and (P117, G121). The bar plot under the map shows the <I(i)> values. Residues corresponding to the top 0.05% MI values (C67, G68, N113-P117, G121-T125, H190, V191 and H239) are highlighted with orange bars, and labeled by different colors (lines at the bottom). (b) Mobility profiles of cathepsin B. The residues identified in panel a are indicated by squares on the $<M^{\text{eff}}>|_{m1}$ curve, color-coded after the lines at the bottom of panel a. They are shown by color-coded spheres in the ribbon diagram for the complex formed with stefin A (cyan).

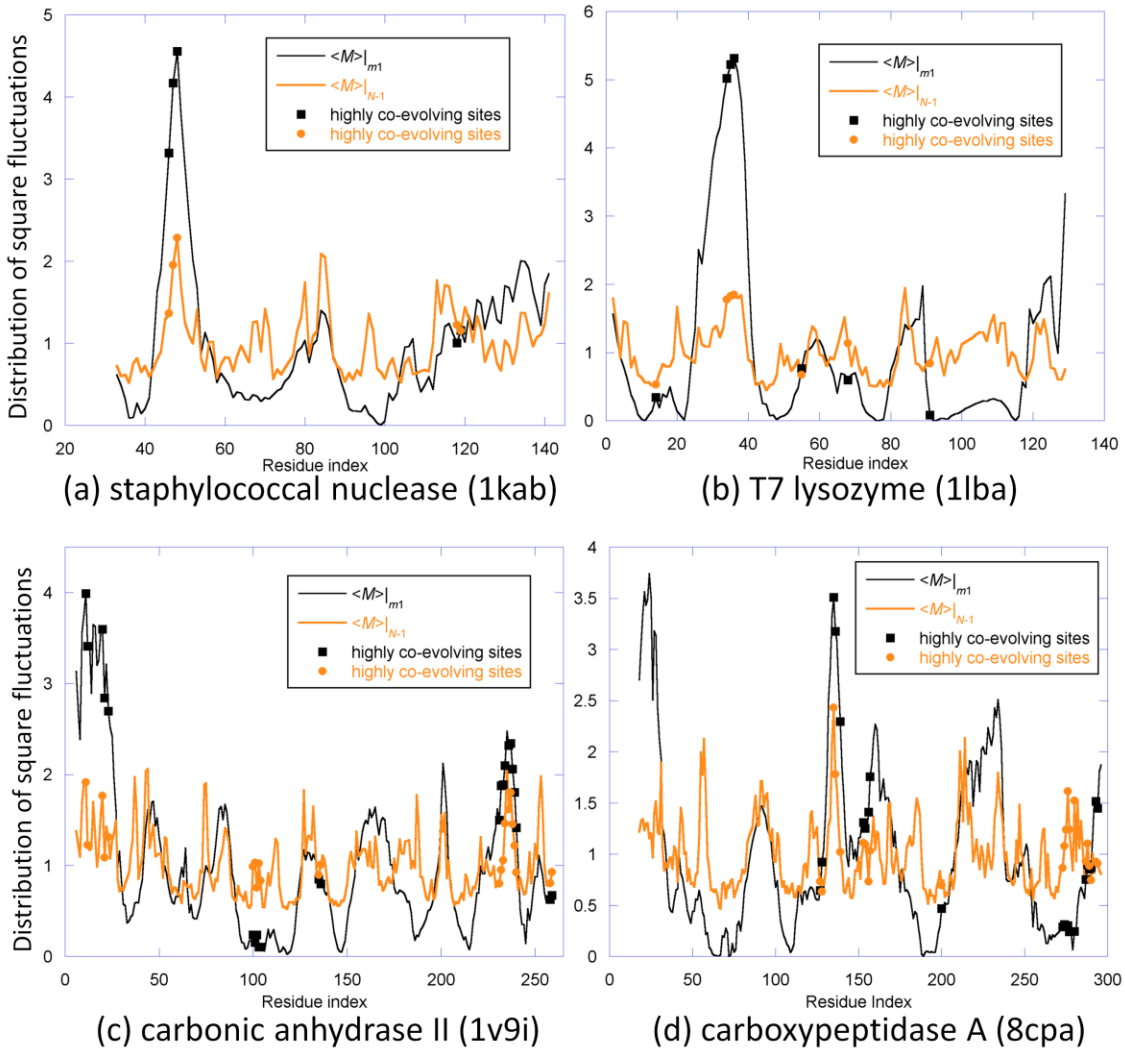

Figure 42. Detection of highly co-evolving amino acids in the regions distinguished by enhanced global mobility.

Global mode shapes and MSF profiles are presented for (a) staphylococcal nuclease (b) T7 lysozyme (c) carbonic anhydrase II (d) carboxypeptidase A. The residues participating in the top ranking 0.05% MI pairs are marked as squares on the mobility profiles. The corresponding PDB id's are shown in each panel. The most prominent peaks in these profiles include: the near neighborhood of Glu43, a residue that promotes the nucleophilic attact required for catalytic activity of staphylococcal nuclease (Cotton et al., 1979); the helix Val28-Glu38 of T7 lysozyme involved in T7 RNA polymerase recognition/binding (Jeruzalmi and Steitz, 1998) (see Appendix Figure A11); two highly flexible solvent-exposed regions (around residues 5-23 and loop 231-241) flanking from both sides the entrance of the hydrophobic ligand-binding pocket in carbonic anhydrase II (Nair et al., 1991); and several residues in the substrate recognition loop R127-S157 of carboxypeptidase A (Kimura, 2001). See **Table 9** for the complete list of residues identified to be highly co-evolving, indicated by the symbols on the curves.

Table 9. List of highly co-evolving residues identified for selected enzymes (*).

PDB id	Residues distinguished by high mutual information values (*)
3pbh	C67, G68, N113, G114, S115, R116, P117, G121, E122, G123, D124, T125, H190, V191, H239
1kab	H46, P47, K48, N118, N119
1lba	I14, Q34, W35, H36, T55, H68, F91
1v9i	H11, N12, D20, F21, I23, S100, S101, D102, D103, Q104, A135, Q136, F231, N232, A233, E234, E236, P237, E238, L239, L240, G258, F259
8cpa	R127, K128, S135, S136, V139, K153, A154, A156, S157, Q200, D273, T274, G275, R276, Y277, L280, I287, P288, W289, L290, T293, W294

(*) residues with I(i, j) values ranking in the top 0.05 percentile (based on the complete dataset of enzymes). See **Figures 41** and **42** for the corresponding mobility profiles.

5.4 MOBILITY, CONSERVATION AND CO-EVOLUTION PROPENSITIES OF AMINO ACIDS

We developed automated procedures to identify in the entire ensemble of 8,254 amino acid, three different subgroups, distinguished by their high mobility (*M*-sites), high conservation properties (*C*-sites) and high co-evolutionary (correlated mutations) propensities (*E*-sites). After identifying the subgroups composed of such *X*-sites (where $X = M$, C or E) we evaluated the propensity P_M of different types of amino acids to take part in these subgroups. The propensity of amino acid type i in the subgroup of *X*-sites is defined as the ratio of the frequency of i in that particular subgroup to that in the entire dataset, i.e.,

$$P_{\mathrm{X}}(i) = \frac{N_{i,\mathrm{X}} \,/\, N_{\mathrm{X}}}{N_i \,/\, N_{\mathrm{total}}} \tag{33}$$

Here $N_{\mathrm{i,X}}$ is the number of occurrences of amino acid type i in the subgroup, N_{X} is the total number of residues in the subgroup. N_i is the number of occurrence of i in all sequences included in the MSAs, and N_{total} is the summation over all $N_{i..}$ The value of $P_X = 1$ indicates that the probabilistic participation in *X*-sites is not different from that expected from *a priori* frequency of amino acids; $P_X > 1$ refers to amino acids that exhibit a high propensity for the examined property (conformational mobility, evolutionary conservation, correlated mutations); and $P_M < 1$, to those exhibiting low propensities.

M-sites were selected by approximating the mobility profiles with cubic splines (Wahba, 1990), and identifying the local maxima. For higher accuracy, two rounds of calculations were performed: first, we identified the "global peaks" on the curve, and a large smoothing parameter (0.99) was adopted; second, more detailed descriptions of mobility profiles were adopted, using a much smaller smoothing parameter (0.3). Finally, local maxima in the sequential neighborhood

(±5 residues) of global maxima were retained as additional highly mobile sites.. The resulting ensemble of *M*-sites comprises 309 (m_1), 457 (m_2), or 641 (*N*-1) residues depending on the mobility profile used as metric.

C-sites were those residues that yielded the lowest 20% Shannon entropy values. A total of 1,706 such residues have been extracted.

E-sites were based on two criteria: the signals observed in the MI maps and the <I(*i*)> values derived from the maps. In the former case, we selected the residue pairs that yielded the strongest 0.05% signals. In the latter case, the top ranking 20% were selected.

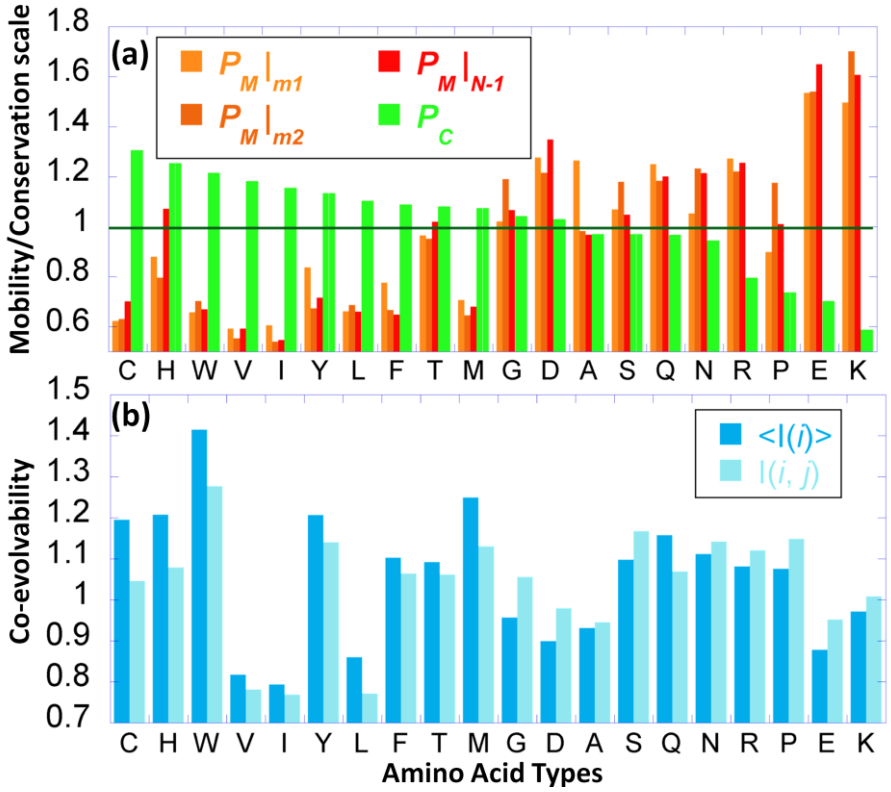

Figure 43. Mobility, conservation and co-evolution propensities of amino acids.

(a) Distributions of amino acids within the subgroups composed of highly conserved (*C*-) (green bars) and highly mobile (*M*-) sites (orange-red bars, based on m_1, m_2 or *N*-1 modes, as labeled). The bars represent the propensities with respect to those expected *a priori* based on the frequency of occurrence of the particular amino acid types in the

dataset. (b) Co-evolution propensities based on mutual information I(i, j) or average MI (<I(i)>) values, as labeled. Amino acid types (shown by 1-letter codes) are listed in the order of decreasing entropy in both panels.

Figure 43 displays the resulting propensities. The light orange bars in panel a describe the propensity of amino acids to undergo high mobilities in the global modes (based on m_1 modes at the lowest frequency end of the mode spectrum). Calculations repeated with m_2 and N-1 modes yielded similar propensities, as shown by the respective dark orange and red bars. The same panel also displays the distribution of amino acids among the most conserved (C-) sites (green). Higher bars indicate higher conservation propensity. Amino acids are ordered along the abscissa according to their conservation propensity. Cysteines are most conserved, followed by His and Trp; and Lys is least conserved. The high level of conservation of histidines (and the occurrence of compensating mutations, if they are not conserved; see below) is presumably due to their unique multi-directional proton transfer capability (Rebek, 1990) , which also makes them the most common amino acid at active sites (Betts and Russell, 2007). Their lowest P_M value among charged amino acids is probably due to aromatic stacking interactions that restrain their flexibility, like other aromatic residues (Trp, Phe, and Tyr). In contrast, Lys and Glu are distinguished by high mobilities (in both global and local motions); whereas Cys is one of the least mobile residues, along with Val, Ile and Leu. The latter group usually lies in the hydrophobic core. The mobility ranking of amino acids is reminiscent of hydrophobicity scales, consistent with the tendency of hydrophobic residues to be buried in the core and thereby have limited motions.

The most striking observation in **Figure 43** is the converse mobility and conservation propensities of amino acids: an amino acid type with high conservation propensity P_C generally

has low propensity P_M for large movements and *vice versa*. These opposite propensities are most pronounced at the two ends of the spectrum.

The co-evolution propensities, P_E, of amino acids are presented in **Figure 43b.** The propensities based on the strongest signals observed in the MI maps (*dark blue*), and based on the highest <I(*i*)> values (*light blue*) exhibit similar features, suggesting that the strong signals in the MI maps also dominate the <I(*i*)> values. For ease of comparison, the amino acids along the abscissa are listed in the same order as panel a. Comparison with the histograms in panel a reveals that co-evolutionary propensities are practically independent of their conservation or mobility scales. Trp is distinguished by its highest propensity, i.e. highest tendency to take part in correlated mutations. This is presumably due to its large size and its ability, along with other aromatic residues such as tyrosine, to make specific interactions (e.g., aromatic-guanidinium interactions with Arg) at protein-protein interfaces (Crowley and Golovin, 2005). Other residues distinguished by their high co-evolutionary tendencies are Met, Cys and His, which, similarly to Trp, are usually conserved and/or highly constrained (see panel a), thus unable to sustain substitutions unless compensated by a correlated mutation.

Polar residues, on the other hand, represent a unique group because of their relatively high co-evolvability and high mobility. Ser, Gln, and Asn, and despite their slightly lower mobility Thr, Pro and Arg (a charged but versatile residue that has hydrophobic and polar moieties) lie in this group. Their combined co-evolution propensity and conformational mobility suggests that they are suitably recruited by proteins at substrate recognition sites being at the same time specific and flexible enough to mediate substrate selectivity.

5.5 DISCUSSION

The present study of the collective dynamics of enzymes in relation to amino acid conservation and co-evolution propensities exposes how the evolution of sequence is tied to structural motions intrinsically accessible to enzymes. Several recent studies have highlighted the significance of collective dynamics in achieving biological functions, or enabling biochemical activities. It is not surprising to see therefore that the sequence evolution or correlated mutations go hand in hand with structure-encoded dynamics. Yet, in previous studies, emphasis has been usually on the evolutionary pressure originating from structure stabilization requirements. For example, mutations that can be accommodated without altering the structure have been pointed out to be evolutionarily selected. Or, alternatively, a designable protein has been viewed as one that can sustain many substitutions while maintaining its structure (Li et al., 1996). In a recent excellent review, the need to retain functional interactions, in addition to conserving the architecture has been pointed out (Worth et al., 2009). Our study further shows that it is equally important to design a structure that enables the required conformational flexibility, or collective dynamics; It further demonstrates that the conservation and co-evolution trends of amino acids correlate with the intrinsic dynamics of the structure: regions severely constrained in global modes are either conserved, or undergo correlated mutations. Conversely, the most mobile regions exhibit the larger sequence variabilities.

The intermediate regime is of interest, in particular, where there is a net proportionality between effective mobility and sequence entropy. Many co-evolving amino acids lie in this regime. Among them, those enjoying enhanced mobility in the global modes appear to be particularly suitable for substrate recognition. This feature noticed in previous studies (Liu et al., 2010; Sakarya et al., 2010) supports the notion that substrate binding entails the conformational

adaptability and physicochemical specificity of recognition sites (Luque and Freire, 2000; Dobbins et al., 2008) prior to stabilization by conserved interactions at the binding epitope. It is widely accepted that the stabilization of the bound ligand is primarily achieved by residues conserved within families or subfamilies. However, prior to binding, the first step is recognition; and the mobility/co-evolution of the recognition sites appears to be a design principle required to accommodate the geometry and chemistry of the substrate (Mittag et al., 2010; Lovell and Robertson, 2010). Our analysis reveals which amino acids have high co-evolution propensities along with enhanced mobilities to satisfactorily fulfill these requirements. Arg and polar residues are distinguished in this respect as versatile mediators of interactions with specific substrates. We also note that there is another, somewhat less prominent, group of co-evolving amino acids, which appear to be assisting conserved residues in either binding the substrate or coordinating cooperative responses, and this group has, in contrast to the former group, significantly suppressed mobilities in the global modes.

The correspondence between residue rigidity (or spatial confinement) and sequence conservation reflects in a sense a functional requirement. The reaction at the active site of an enzyme usually requires high precision: s catalytic residues need to be accurately positioned and oriented, and highly conserved, to achieve chemical specificity (Sacquin-Mora and Lavery, 2006; Dutta and Bahar, 2010). Conserved residues that serve as folding nuclei also need to be highly stable (Mirny and Shakhnovich, 1999). On the other hand, surface-exposed residues are generally involved in substrate recognition or intermolecular interactions, and their high mobility is, not only easily afforded from structural perspective, but even required to accommodate sequence variations that confer substrate specificity.

The evolutionary *vs*. dynamic properties of binding sites may depend on the size and specificity of the substrate, whether it is a small molecule (e.g., ATP) or a biopolymer (e.g., protein). The two types of interactions have been shown to exhibit distinct structural properties: the former is conserved and almost rigid; whereas the latter tend to exhibit correlated mutations and higher mobility (Jones and Thornton, 1997; Liu et al., 2010). Pre-organization of conserved residues with restricted mobility has been suggested to help in stabilizing the bound conformer with minimal entropic penalty (Yogurtcu et al., 2008), while in the opposite case of high mobility the favorable enthalpic interaction with the binding partner may more than compensate the unfavorable entropic contribution provided that the interaction surface is large enough (protein-protein interactions). Insights into such design properties may be gained by performing similar investigations for different classes of complexes. Interfacial residues of obligate pairs are more conserved than that of transient pairs, or alternatively, they contain correlated mutations (Mintseris and Weng, 2005), although the distinctive dynamics of these two classes have yet to be established. Likewise, although the present analysis has been performed for enzymes, it remains to be seen if/how the observations hold for other classes, including in particular membrane proteins whose growing number of structures is expected to soon lend themselves to systematic analyses.

6.0 CORRELATED MUTATIONS ANALYSIS (CMA) OF HIV-1 PROTEASE USING SPECTRAL CLUSTERING

In the present study, we introduce the use of spectral partitioning methods for efficient analysis of the MI matrices derived for HIV-1 protease sequences. Spectral clustering was originally proposed for partitioning the nodes in an undirected weighted graph $G = (V, E)$. The weight w_{ij} of each edge is defined as a measure of similarity between nodes v_i and v_j. This weight matrix **W** is replaced in our work by the MI matrix (see section 2.4), with the objective to examine sequence co-variance and distinguish between correlations of different origin.

We show that the method successfully identifies the residues cooperatively involved in MDR, as well as the mutational patterns arising from different drug treatments. The results suggest that spectral partitioning of the data obtained from correlated mutations analysis (CMA) can help in detecting cooperative functional relations and discriminating to a certain degree between the covariance patterns originating from functional constraints and those associated with neutral/stochastic mutation events that occur early in the evolution of the species/family.

6.1 SPECTRAL CLUSTERING OF CMA RESULTS

To investigate the correlation between drug treatment and mutational patterns, we compiled six datasets of sequences retrieved from the Stanford HIV Drug Resistance DB

(http://hivdb.stanford.edu; (Rhee et al., 2003)) (**Table 9**). This DB includes sequences obtained from isolates along with information on the type of protease inhibitors (PIs) given to the patients (accessible via the 'Detailed Treatment Queries' interface of the DB). We collected sequences of all subtypes and aligned them against the consensus subtype B sequence (Korber and Myers, 1992). Any sequence shorter than 99 residues was excluded, and all residues with ambiguity were treated as gaps. A MI matrix was generated for each dataset of HIV-1 protease sequences listed in **Table 10**.

Table 10. Summary of the HIV-1 protease sequence data subjected to spectral clustering

Dataset	Treatment	Number of sequences
1	Treated	7758
2	Untreated	8761
3	IDV only	1112
4	IDV +	2569
5	NFV only	885
6	NFV +	2131

In the 'Treatment' column, 'treated' means at least one PI is used in the treatment. 'IDV +' and 'NFV +' means that at least one of the other PIs has been used in combination with the one before the '+' sign. IDV and NFV are the respective PI drugs indinavir and nelfinavir.

The result for dataset 1 is illustrated in **Figure 44**. The plot underneath represents the entropy profile. Peaks are distinguished at positions such as 10, 20, 63 and 82, reflecting the high tendency of these residues to undergo substitutions.

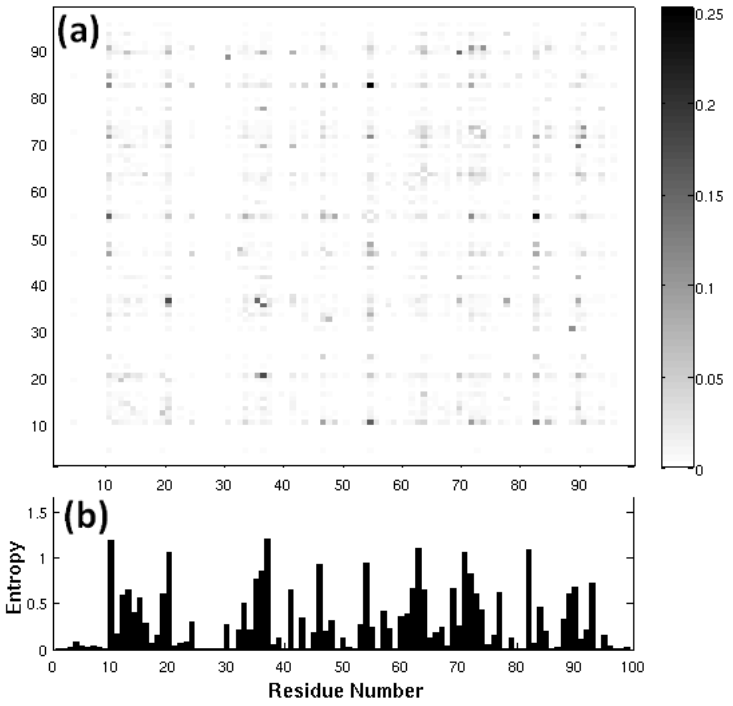

Figure 44. Mutual information (MI) map (a) and entropy profile (b) for HIV-1 protease sequences in Dataset 1.

The entries in the map are calculated using Equation (29) for the 7758 sequences compiled in Dataset 1 (**Table 10**). The MI varies in the range $0 < I(i, j) < 0.25$, as indicated by the gray scale on the right. Panel b displays the entropy profile, with the peaks indicating those sites exhibiting the largest variation among the members of this dataset.

In order to extract more distinctive information, each MI matrix was subjected to spectral graph bi-partitioning as described above, and the elements were re-ordered (i.e. rows/columns were shuffled) according to the rank of residues indicated by the dominant eigenvector $\mathbf{y}_2$ (i.e. by sorting the elements of $\mathbf{y}_2$ in descending order). **Figure 45** displays the MI maps as a function of the re-ordered residues for datasets 1 and 2. Equivalent figures for the other four datasets can be found in **Figure 46**. The exact labeling of residues following rank ordering can be found in the **Table 11**. For visual clarity, the top ranking (highest MI) pairs of amino acids (500 out of a total of 99×99 pairs) are displayed. The bar plots refer to the entropy at each site.

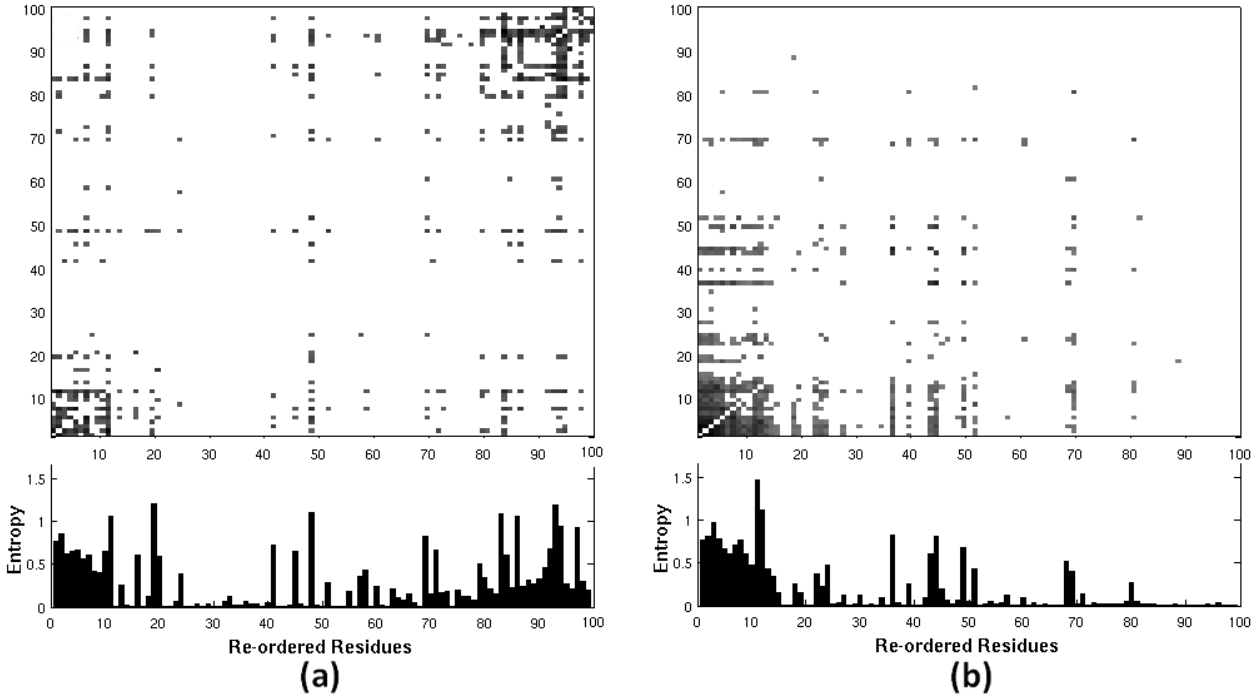

Figure 45. MI maps with residues re-ordered according to spectral graph bi-clustering.

(a) Re-organized MI matrix for treated data (dataset 1). Two distinctive types of correlated mutations can be seen at the lower left and upper right portions of the map. (b) Re-organized MI matrix for untreated data (dataset 2). One of the previous clusters is observed (lower left), while the 2nd (top right) is non-existent. The latter is attributed to correlated substitutions induced in the presence of inhibitors, while the former (upper right) refers to evolutionary changes observed between HIV-1 protease subtypes. See **Figure 47** for the identity of residues belonging to the two clusters, and the **Table 11** for the identity of rank-ordered residues for each dataset listed in **Table 10**. The bar plots refer to the sequence entropy associated with each position.

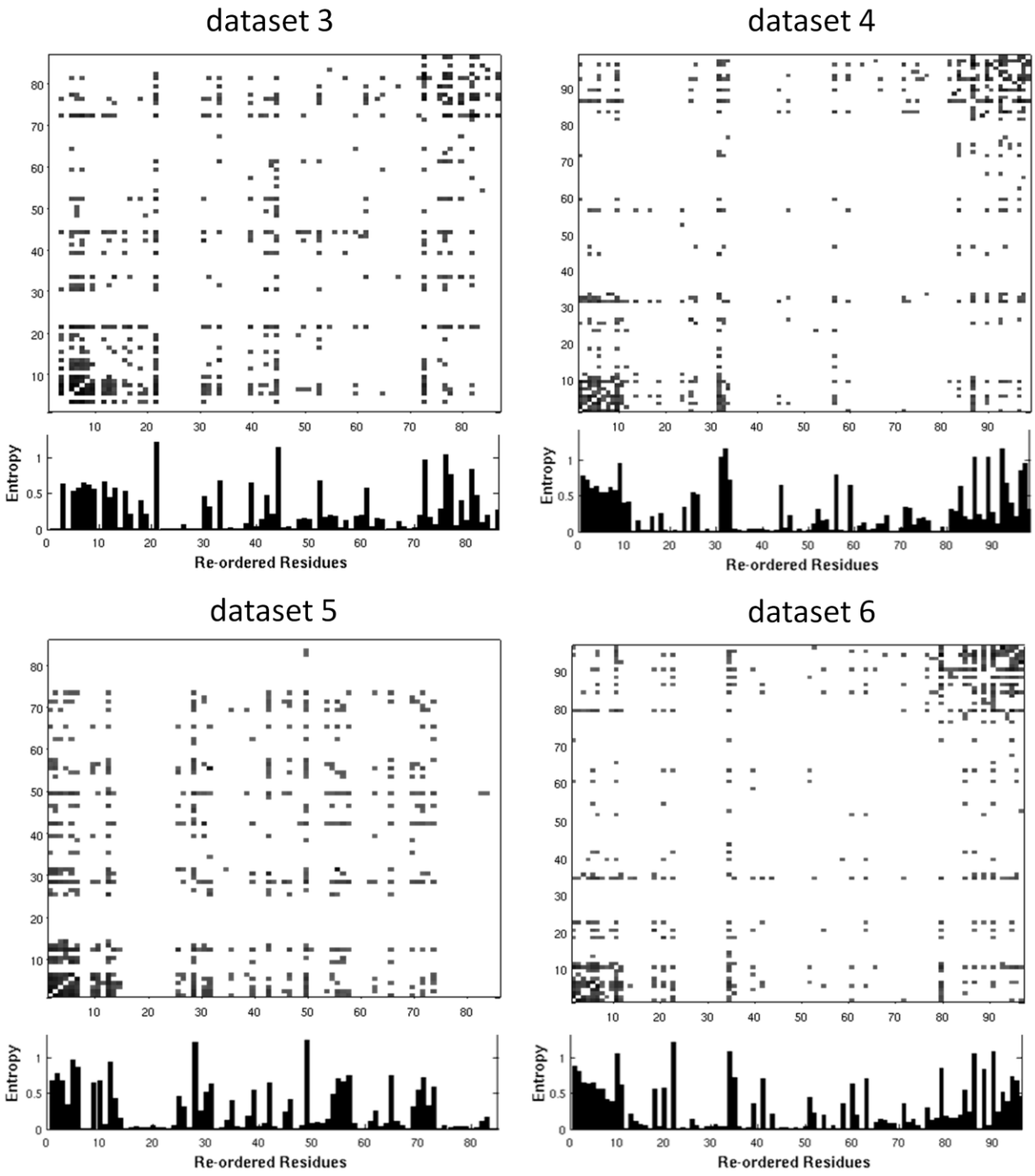

Figure 46. MI maps with residues re-ordered according to spectral graph bi-clustering for datasets 3-6.

Comparison of panels a and b of **Figure 45** reveals that dataset 1 (panel a) contains two distinctive clusters of correlated residues located at the upper right and lower left portions of the map, while dataset 2 does not contain the 2nd cluster (at the upper right) (panel b). The identity of the residues at these two extreme ends of the maps generated for all datasets in **Table 10** can be seen in **Figure 47**. Here we colored in blue and red the first and last 12 residues rank ordered after the spectral bi-partitioning of the MI matrix for each dataset (labeled). Interestingly, all datasets, treated with different regimens or untreated, exhibit similar patterns, with the two groups of residues exhibiting most distinctive correlation behavior clustered at similar sequence positions.

6.1.1 Examination of the two distinctive clusters

Given that the respective datasets 1 and 2 refer to treated and untreated sequences, the cluster at the top right in **Figure 45a**, which does not exist in panel b, is attributed to the substitutions induced by drug treatment. We will refer to these positions as drug resistance cluster (DRC) sites.

The 2nd cluster of residues, on the other hand, is interestingly found to primarily contain positions reported to exhibit sequence variability between different viral subtype isolates (Gonzales et al., 2001). To verify this feature, we collected 5149 untreated non-B subtype sequences from the Stanford DB, and calculated the variation frequency at each position with respect to the consensus subtype B sequence (**Figure 48a**). (More detailed information on the variation for each individual non-B subtype isolates can be found in Figure 2 of Gonzales et al., 2001). This suggests a phylogenetic origin for the observed covariance, which can well be obtained simply based on few neutral substitution events in the evolution of the HIV subtypes.

These residues do not necessarily possess important functional/structural associations (Noivirt et al., 2005). We will refer to this cluster of residues as the phylogenetic variation cluster (PhVC).

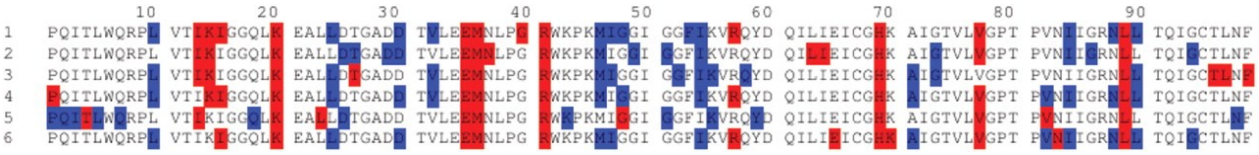

Figure 47. Sequence position of two most distinctive clusters of residues deduced from CMA of HIV-1 protease sequences.

Results are reported for each of the six datasets listed in **Table 10**. The two clusters include the two extreme subsets of 12 residues rank ordered according to the spectral bi-partitioning of the MI matrix computed for each dataset. The DRC residues are colored blue, and the residues belonging to the PhVC are colored red.

Table 11. Reordering of amino acids in each dataset based on spectral clustering.

Residue Number	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6
1	Glu35	Leu89	Thr96	Met36	Leu89	Met36
2	Met36	Met36	Leu97	Glu35	His69	Glu35
3	Val77	His69	Met36	Val77	Arg41	Val77
4	Arg41	Arg41	Phe99	His69	Val82	Arg41
5	His69	Lys20	Leu89	Ile15	Met36	His69
6	Ile15	Lys14	His69	Arg41	Glu35	Leu89
7	Leu89	Ile13	Lys20	Leu89	Gly48	Ile15
8	Arg57	Glu35	Arg41	Lys20	Thr4	Lys20
9	Lys14	Ile64	Ile13	Arg57	Val77	Arg57
10	Ile13	Val77	Thr26	Pro1	Ile13	Lys70
11	Lys20	Leu63	Glu35	Lys14	Leu23	Asn83
12	Gly40	Asn37	Lys14	Ile13	Lys20	Glu65
13	Lys70	Ile62	Val77	Thr96	Lys14	Lys14
14	Phe99	Lys70	Arg8	Lys70	Ile54	Thr4
15	Thr26	Lys45	Ile15	Glu65	Leu76	Ile13
16	Leu19	Pro1	Glu65	Leu38	Asn83	Ile50
17	Leu97	Thr80	Glu21	Gly52	Val11	Ala28
18	Glu65	Val82	Arg57	Gly16	Leu24	Leu97
19	Asn37	Leu33	Gly16	Asn83	Pro9	Gly49
20	Thr12	Val56	Leu23	Gln61	Ala22	Ala22
21	Ala28	Gly52	Leu63	Thr26	Gly40	Phe99
22	Pro1	Gly16	Asn98	Gln2	Thr26	Leu38

23	Asn83	Ala71	Ile50	Thr12	Leu97	Asn37
24	Gln61	Ile72	Leu38	Leu19	Cys95	Thr96
25	Gly49	Leu97	Asp30	Asn37	Leu90	Ile93
26	Asp25	Leu38	Gly17	Gly68	Lys45	Leu19
27	Thr96	Thr74	Val56	Ala28	Arg8	Ile3
28	Val56	Ile50	Val75	Thr31	Leu63	Glu21
29	Gln7	Ile66	Gln2	Gly49	Lys70	Gly52
30	Gln2	Glu65	Leu19	Thr91	Leu19	Gln7
31	Gly52	Thr96	Lys70	Trp42	Asp30	Gly17
32	Ala22	Ala22	Pro9	Glu21	Arg87	Gly86
33	Pro39	Asn83	Ile93	Ile93	Leu38	Leu5
34	Arg87	Cys95	Asn83	Thr4	Val75	Asp25
35	Arg8	Ile47	Asp25	Asp25	Arg57	Thr12
36	Gly17	Leu19	Arg87	Phe99	Trp6	Trp42
37	Gly68	Gln2	Gly86	Ile3	Ile50	Pro39
38	Leu5	Gly78	Gly48	Leu97	Pro39	Gln61
39	Asn98	Cys67	Ile64	Trp6	Ile15	Gly51
40	Leu38	Phe53	Leu5	Val56	Ile84	Thr80
41	Ile93	Trp42	Ile66	Asp60	Ile66	Ile62
42	Tyr59	Gln18	Thr12	Ala22	Thr12	Gly68
43	Thr31	Ile15	Gln61	Arg87	Thr91	Leu63
44	Ile3	Thr12	Asn37	Ile62	Pro81	Cys67
45	Ile64	Pro39	Glu34	Leu63	Gly16	Val56
46	Glu21	Pro79	Val11	Asp29	Thr74	Arg8
47	Gly86	Arg87	Thr4	Cys67	Glu21	Val11
48	Leu63	Glu34	Pro39	Gly27	Gly68	Pro9
49	Trp6	Ile93	Cys67	Val11	Asn37	Ile64
50	Asp29	Ile3	Gln18	Ile64	Ile85	Gly27
51	Gly16	Arg57	Tyr59	Gln7	Gly17	Thr74
52	Pro9	Pro81	Ile72	Leu5	Glu34	Thr91
53	Trp42	Phe99	Lys45	Pro39	Cys67	Arg87
54	Pro44	Gly68	Gln92	Gly78	Met46	Gln92
55	Lys45	Val75	Leu33	Gln92	Asn88	Trp6
56	Gly27	Asn98	Gly51	Glu34	Ile93	Gly16
57	Asp60	Gly17	Leu76	Leu33	Ile64	Asp60
58	Thr74	Ala28	Thr31	Gln18	Gly73	Gln2
59	Gly78	Leu23	Thr74	Ile72	Gly49	Ile72
60	Cys67	Val11	Asp60	Ile50	Thr31	Glu34
61	Gly51	Gly48	Ile62	Pro9	Glu65	Thr31
62	Gly94	Thr4	Gln7	Val75	Leu33	Lys45
63	Gln92	Leu76	Ile85	Lys45	Pro79	Ile85
64	Thr91	Thr31	Ile84	Gly51	Gln58	Leu23
65	Thr4	Asn88	Pro79	Leu76	Ala71	Leu33
66	Gln18	Pro9	Ala28	Arg8	Gln92	Pro1

67	Val75	Pro44	Cys95	Pro79	Val32	Thr26
68	Pro81	Leu10	Ile3	Gly94	Val56	Pro79
69	Ile72	Gln61	Thr80	Pro44	Gln61	Cys95
70	Cys95	Gln58	Trp6	Gly17	Ile62	Lys43
71	Ile62	Lys43	Lys43	Thr80	Leu10	Gln18
72	Val11	Ile85	Val82	Pro81	Asp60	Tyr59
73	Ile66	Thr91	Asn88	Thr74	Ile72	Ile66
74	Thr80	Glu21	Phe53	Cys95	Tyr59	Val75
75	Ile85	Tyr59	Gly73	Tyr59	Lys55	Leu76
76	Ile50	Trp6	Leu10	Lys43	Asn98	Asn98
77	Pro79	Lys55	Ala71	Leu23	Gln2	Gln58
78	Leu23	Gly40	Gln58	Ile85	Gln7	Phe53
79	Leu33	Gln92	Leu90	Ile66	Leu5	Lys55
80	Lys43	Asp60	Lys55	Gln58	Pro1	Gly73
81	Glu34	Gln7	Met46	Gly73	Ile3	Val32
82	Leu76	Leu90	Ile54	Lys55	Gln18	Gly48
83	Val82	Leu5	Ile47	Phe53	Lys43	Ala71
84	Gly73	Arg8	Val32	Ile47	Gly51	Gly94
85	Gln58	Leu24	Gly52	Ala71	Asp25	Ile84
86	Ala71	Val32	Leu24	Val82		Val82
87	Lys55	Gly27		Gly48		Ile47
88	Gly48	Ile54		Val32		Leu10
89	Phe53	Asp29		Leu10		Met46
90	Asn88	Ile84		Ile54		Leu90
91	Ile84	Gly73		Ile84		Ile54
92	Leu90	Gly94		Leu90		Leu24
93	Leu10	Gly51		Met46		Asn88
94	Ile54	Asp25		Asn88		Asp30
95	Asp30	Thr26		Leu24		
96	Val32	Met46		Asp30		
97	Met46	Gly86				
98	Leu24	Asp30				
99	Ile47	Gly49				

*Each column corresponds to the re-ordered sequence (shown for datasets 1 and 2 in the abscissa in **Figure 45**). The two extreme subsets of 12 residues used in **Figure 47** are colored red (PhVC) and blue (DRC).

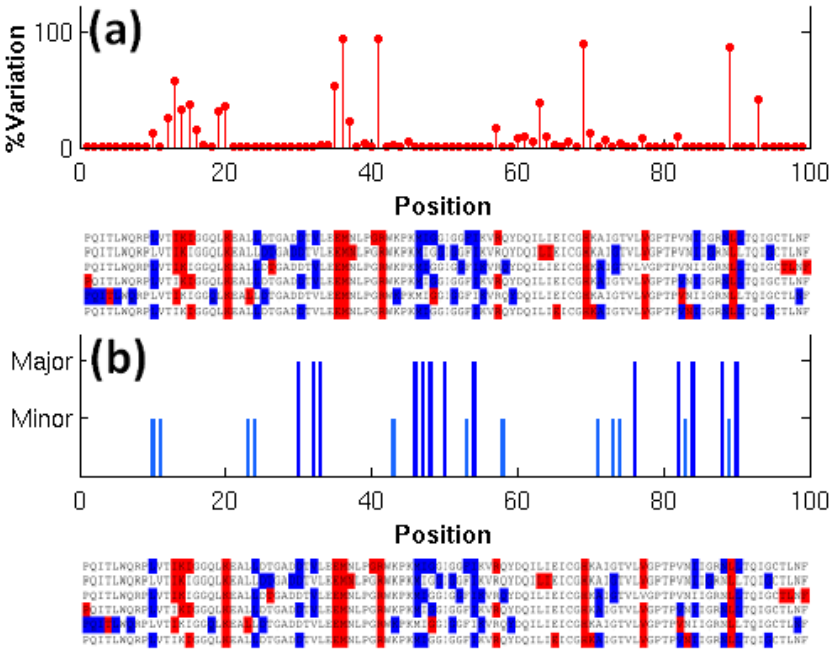

Figure 48. Comparison of computationally predicted sites on HIV-1 protease with experimental data. (a) Sequence variation profile compiled from experimental data for the non-B subtype HIV (from Stanford DB). Note the correspondence between peaks (most variable sites) and the phylogenetic variation sites (red in the alignment) identified in the present study. (b) Comparison with drug resistance profile (based on data in Stanford DB http://hivdb.stanford.edu/cgi-bin/PIResiNote.cgi). Dark blue lines refer to residues that exhibit major drug resistance; light blue, to minor drug resistance sites.

It should be noted, however, that sequence variations between subtypes are not necessarily functionally insignificant. This is reflected for example by the fact that different subtypes have different tendencies for acquisition of resistance mutations (Kantor et al., 2005). Indeed, residues related to drug resistance can be found in this cluster. Positions 20 and 36 exhibit enhanced mutation rates in the presence of PIs (Wu et al. 2003, Table 2; Hoffman et al., 2003, Figure 1a). It is possible that the evolution of HIV subtypes is partially related to the exposure to natural or unnatural PIs. Residue Leu89 in the PhVC is known, for example, as a minor drug-resistant residue (meaning that a mutation at this position contributes to drug

resistance only in the presence of a major resistant mutation, whereas a major resistant mutation reduces drug susceptibility by itself (Shafer, 2002). Yet, overall, the members of the PhVC are best characterized as those demonstrating sequence variability between subtypes with no clear functional relation between them.

In contrast to the PhVC, the DRCs identified for datasets 1, 3, 4 and 6 mostly contain drug-resistant mutations (**Figure 48b**). In particular, some residues belonging to these clusters are associated with mutations involved in multi-drug cross-resistance, such as Leu10, Met46, Ile54, Ala71, Val82, Ile84 and Leu90 (Hertogs et al., 2000; Kozal, 2004). In a previous study (Ohtaka et al., 2003), Leu10, Met46, Ile54, Val82, Ile84 and Leu90 were shown to exert a cooperative effect in lowering the affinity of multiple PIs. Leu10, although not causing resistance alone (it is a minor resistance residue), plays a critical role in eliciting the cooperative response along with Leu90 (Ohtaka et al., 2003), consistent with the high correlation detected here among these residues in the DRC. We also note that some major mutation sites in the DRC are not active in MDR; or say, they are specific to one PI, like Leu24 and Asp30 (Shafer, 2002). Still, their participation in the DRC suggests that the resistance mechanism cooperatively involves several residues.

The DRCs for datasets 2 and 5 contain a number of sites that depart from those shared by other datasets (**Figure 47**). For dataset 2, which contains untreated isolates only, this is clear, and even the observed level of similarity to other datasets is striking. For dataset 5, on the other hand, the result implies that NFV elicits unique responses at specific sites, quite different from that of most other drugs. We note in particular that Asp30 and Asn88 exhibit extraordinarily high MI. As shown before (Rhee et al., 2003), the double mutation D30N and N88D can reduce nelfinavir susceptibility by 50-fold, explaining the selection pressure for their co-variation. When NFV is

used in combination with other PIs (dataset 6), the DRC sites shared with other datasets are observed, indicating that the cooperative effect is related to cross-resistance in this case. Most of the residues of the DRC remain unchanged in the IDV set (dataset 3), suggesting that the correlations revealed in our analysis are not only due to individual resistance mutations developed against different drugs, but reflect real cooperativity.

An exhaustive search for correlated mutations among drug-resistant sites in HIV-1 isolates was performed by (Wu et al., 2003), which yielded small groups of correlated residues, ranging in size from three to six residues. On the other hand, the present study yields one large cluster providing evidence for the high cooperativity of the residues belonging to these small groups. We also note that the presently detected positions 47 and 48 in the flap region do not appear in the study by Wu et al. as prominent drug resistance sites, but they are known to be major resistant mutations. Wu et al. listed other residues, e.g. Ile62, Leu63 and Ile93, together with known drug resistance residues. We have not detected these residues in our DRC, and neither do they appear in the Stanford PI DB drug resistance notes as drug-induced mutations. Note that our study is based on a larger dataset of isolates, and a major merit of the present work is to identify the DRC sites without prior knowledge of drug-resistant mutation sites, while the study of Wu et al. analyzes the mutations at 45 (out of 99) positions that have been significantly associated with protease inhibitor treatment.

Hoffman et al, 2003 analyzed the correlations between 31 positions in HIV-1 protease, which showed the highest variability in their dataset of HIV-1 isolates (from 648 untreated, and 531 treated persons). These were grouped in three clusters based on the comparison of mutation rates between treated and untreated datasets. This criterion is different from the one (based on MI data) adopted in our study, but it is still tempting to compare the two sets of results. Those

residues in Class III therein are similar to those in our DRC, while Class I resembles our phylogenetic cluster PhVC. Notably, residues Lys20, Met36 which are part of our phylogenetic cluster appear in cluster II and cluster III, respectively. These residues exhibit substantial sequence variability between subtypes, and appear to be relevant to drug resistance, but apparently not in a cooperative manner with other residues.

6.1.2 *k*-way clustering using more eigenvectors

The results from *k*-way clustering of dataset 1 using $k = 3$, 4 and 5 are presented in **Table 12**. The most correlated residues identified above take part in the same clusters, consistent with results from bi-partitioning. Notably, Asp30 and Asn88, which originally belonged to the DRC, exhibited a tendency to form a separate cluster together with Val75. This triplet (Asp30, Asn88, Val75) was also reported to form a cluster in previous work (Wu et al., 2003). It has long been known that co-substitutions at Asp30 and Asn88 are most effective in reducing the susceptibility of nelfinavir; however, little attention has been given to date to their possible association with Val75. As indicated in **Figure 49**, the high correlation of Val75 with Asp30 and Asn88 (**Figure 49a**), consistent with their structural proximity (**Figure 49b**), may originate from a cooperative mechanism for drug resistance between these three sites.

Table 12. Results from *k*-way spectral clustering of the HIV-1 protease treated dataset

k	Cluster
3	**C1:** 30, 75, 88 **C2:** 1– 9, 12–15, 17, 19, 20, 22, 25, 26, 28, 31, 35–42, 45, 49, 52, 56, 57, 59, 61, 65, 68–70, 77, 83, 87, 89, 96–99
4	**C1:** 1, 2, 9, 26, 30, 40, 45, 56, 59, 75, 81, 88, 98 **C2:** 13–15, 20, 35–38, 41, 42, 49, 57, 69, 70, 77, 83, 89 **C3:** 10, 23, 24, 27, 32–34, 43, 46–48, 50, 53–55, 58, 71, 76, 80, 82
5	**C1:** 30, 75, 88 **C2:** 1, 2, 9, 26, 40, 45, 59, 87, 98 **C3:** 13–15, 20, 35–38, 41, 49, 57, 69, 70, 77, 83, 89 **C4:** 10, 23, 24, 27, 32–34, 42, 43, 46–48, 50, 53–55, 58, 71,76, 80, 82

For clarity, the largest cluster that includes all the remaining residues in each case is not shown.

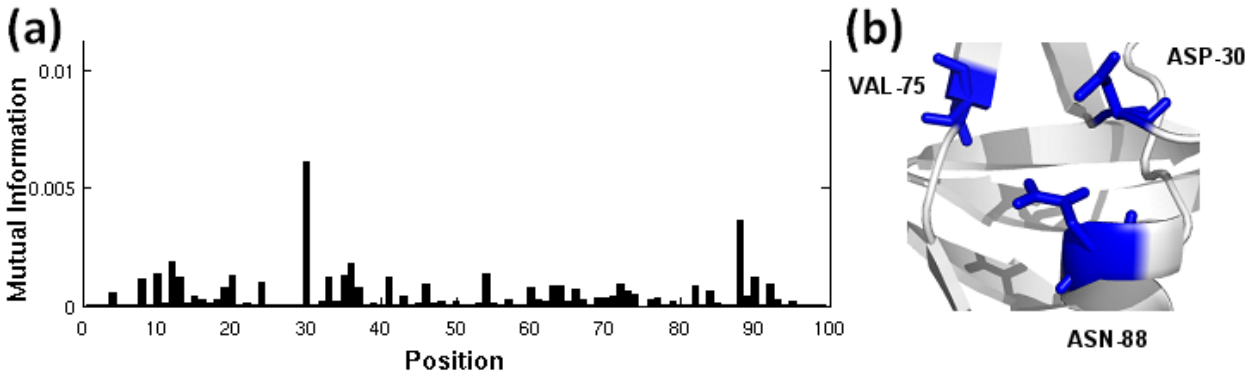

Figure 49. Examination of Asp30, Asn88 and Val75.

(a) The MI profile of Val75 with other residues in the treated dataset (dataset1). (b) The structural vicinity of Asp30, Asn88 and Val75.

6.2 INTERPRETATION WITH RESPECT TO PROTEIN DYNAMICS

The examination of HIV-1 protease 3D-structure reveals that the residues participating in the DRC tend to occupy the flap region (Met46, Ile47, Ile54), the close neighborhood of the active

site (Asp30, Val32, Val82, Ile84), and the dimerization interface (Leu10, Leu90). Most of PhVC residues, on the other hand, are located away from the interface, toward the exterior of the protein (**Figure 50a**). Interestingly, both groups of residues assume regular secondary structures (helices or strands), although their relative positions with respect to the interfacial region differs.

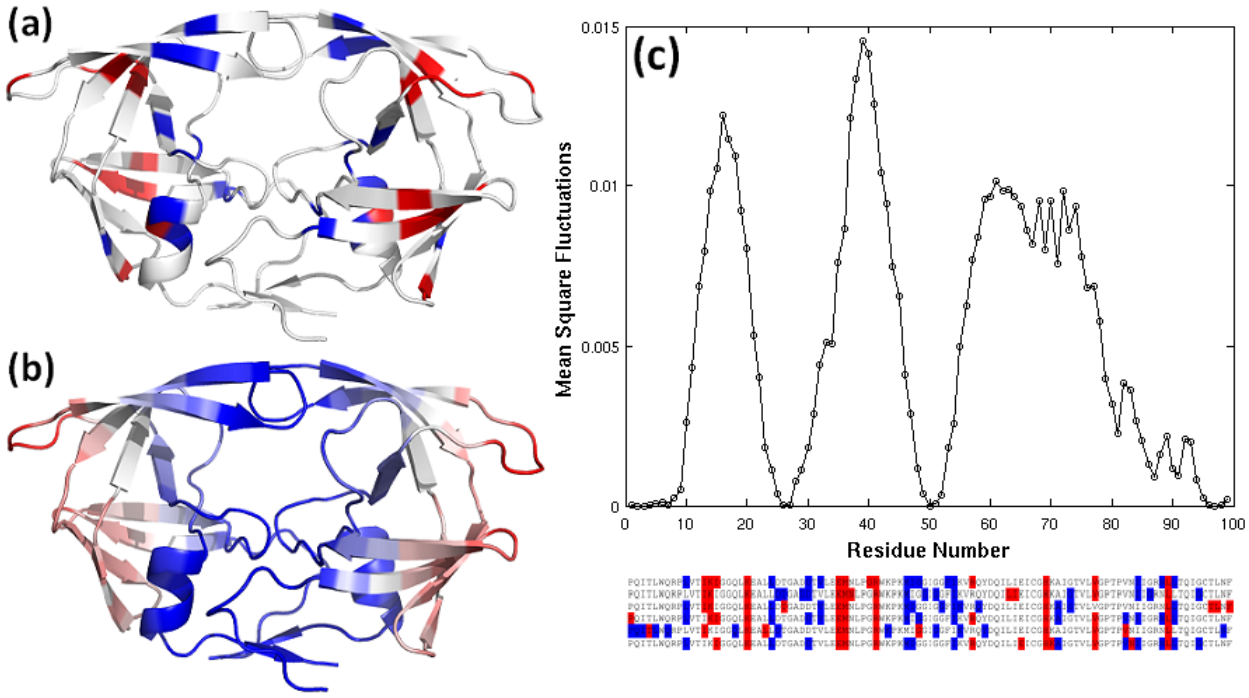

Figure 50. Comparison of results from correlated mutation analysis (CMA) and GNM dynamics.

(a) The location of the two clusters identified for dataset 1 on the 3D-structure of HIV-1 protease. The DRC is colored blue, and the PVC is colored red. We displayed the residues that have appeared at least three times (out of six examined datasets) in the same cluster in **Figure 47**. (b) Ribbon diagram color-coded after the mobilities of residues in the first slow mode predicted by the GNM. The residue mobility increases from blue to red. (c) GNM slow-mode profile as a function of residue index. Note that calculations are performed for the dimer, but results are shown for a monomer, the curves for the two monomers being identical. The HIV-1 protease mutant bound with IDV (PDB id: 2B7Z) was used.

We also examined the distance separation between the closest atoms of residue pairs belonging to the two clusters. For each pair two values have been considered: intra-molecular

(monomers A–A or B–B contacts) or intermolecular (A–B contacts). These data clearly demonstrate that the correlated pairs essentially refer to intra-molecular interactions, rather than inter-molecular. Note that the MI method cannot detect the correlations between the fully conserved residues at the interface between the monomers (e.g. P1-F99 and D29-R8).

A further comparison between the results from CMA and the mobilities of residues predicted by the GNM elucidates the close correspondence between the global dynamics of the enzyme and its function. The lowest frequency GNM mode usually defines the global dynamics of the enzyme accessible under native state conditions, and such cooperative motions intrinsically favored by the structure have been shown to relate to enzymatic function (Yang and Bahar, 2005). In particular, the global hinge regions (minima in the mobility profiles driven by global modes) play a critical role in conferring the mechanical properties of enzymes that complement their chemical (catalytic) activities.

In order to examine the dynamics of residues belonging to the DRCs and PhVCs, we performed GNM calculation for an HIV-1 protease mutant bound to IDV (PDB file 2B7Z). This structure contains 10 mutations, most of which belong to the DRC presently identified for the IDV-treated dataset. The color-coded ribbon diagram in panel b of **Figure 50**, and the slow-mode profile in panel c, display the mobilities in the lowest frequency mode predicted by the GNM for this structure. Comparison of panels a and b shows that the DRC residues tend to occupy positions that are highly constrained in the global mode, whereas PhVC residues are located at relatively flexible positions. These distinctive dynamics of the two groups of residues explains the fact that the PhVCs are accommodated without altering the structure and function; whereas mutations at the DRC sites that are more buried and spatially constrained have functional consequences. Calculations repeated for the substrate-bound complex (PDB id: 2FNS

(Prabu-Jeyabalan et al., 2006)) confirmed that the slow-mode profile is insensitive to structural asymmetry and yielded the almost identical profiles for the two subunits, while the 2nd mode exhibited a stronger dependence on structural asymmetry.

Finally, we compare the global mobility profile (panel c) with the sequence position of the two clusters (**Figure 47**) reproduced in **Figure 50** to ease the visual comparison. The residues in the DRC are seen to usually lie close to global hinge regions (minima), while those in the PhVC are distributed in high mobility regions. Calculations were repeated for the 2nd and 3rd GNM modes as well. Comparison of the minima and maxima in these modes with the PhVC (red) and DRC (blue) sites along the sequence shows that PhVC modes exhibit relatively high mobilities in modes 2 and/or 3 as well, whereas the confinement of DRC residues to hinge sites is characteristic of the first (global) mode. The DRC residues located at the flap region (residues 46–54) show a high mobility in modes 2 and 3. Co-localization of MDR sites with global hinge regions thus emerges as an effective means of impacting the cooperative dynamics, and hence the function of the enzyme (Bahar et al., 1998) and on the catalysis.

6.3 CONCLUSION

In the present study, we analyzed the covariance patterns in HIV-1 protease sequences using a simple metric, MI, followed by spectral clustering. The approach proved to discriminate between two groups of correlated mutation sites, shortly referred to as DRC and PhVC. Mutations in the DRC tend to confer MDR while those in the PhVC seem to differentiate between different HIV-1 protease subtypes. We have further explored the biophysical basis of the observed differences between the two clusters of correlated sites. The two clusters were found to significantly differ

with regard to their role in the intrinsic structural dynamics of the enzyme. The DRC sites select key mechanical regions, near the global hinges that control the most cooperative motions of the enzyme; PhVC residues, on the other hand, preferentially occupy flexible regions that can easily accommodate residue substitutions.

Covariance analysis of related protein sequences is known to be problematic in many aspects (Fodor and Aldrich, 2004; Halperin et al., 2006). Many options exist to improve the basic method presented here. In the future, it may be worth considering different essential covariance measures for further analysis. Methods for assigning significant scores using the original MI scores and shuffling of the original data (Hoffman et al., 2003; Shackelford and Karplus, 2007) can also help in obtaining more meaningful results.

One major goal here was, however, to draw attention to the utility of clustering the covariance data. This step is important due to various reasons. First, although the CMA is performed in a pairwise manner (mainly due to technical and statistical reasons), it is clear that in nature larger sets of residues are expected to co-evolve to meet particular structural/functional requirements. Second, the clustering procedure is expected to help in distinguishing the real correlations from the background noise. The choice of clustering technique may also depend on the adopted CMA. When an asymmetric metric is used in step 2, a hierarchical clustering is conveniently applied (Hatley et al., 2003; Shulman et al., 2004; Chen et al., 2006). For symmetric metrics such as Pearson correlation coefficient and MI, on the other hand, a common procedure is to perform a principal component analysis (Fleishman et al., 2004). Here we utilized a relatively less detailed, but objective and theoretically robust approach. Significantly, this approach allowed us to separate the sequence covariance arising from functional pressures (e.g. MDR) from those evolutionarily selected within the examined phylogeny. Both groups of

correlations exhibit strong signals when covariance properties are quantified in terms of MI. Yet, the distinctive character of the two groups, confirmed by experiments (**Figure 48**), and rationalized by comparison with structural dynamics (**Figure 50**), supports the utility of adopting a spectral bi-clustering method for efficiently discriminating between potential correlations of fundamentally different nature/origin. It will be of interest to further explore the utility of spectral bi-clustering for differentiating between correlated mutations that reflect 'real' inter-residue interactions and those reflecting other evolutionary signals, often considered as noise for most analyses purposes (Noivirt et al., 2005).

Notably, some of the sites for potential MDR, indistinguishable in the untreated sequences (**Figure 45b**), can be detected upon rank ordering the residues via spectral clustering of MI data; furthermore, treated sequences subjected to different regimens share common DRC residues (**Figure 47**). These two observations invite attention to the intrinsic tendency of the enzyme to potentially select those effective sites to develop mutations that confer MDR, irrespective of treatment.

A challenging, yet important task, which is a natural continuation to this work, is to detect correlations between protease residues and residues of other mature/pre-mature proteins of HIV-1. A recent work demonstrates how such correlations can be detected between a protease mutation (V82A) and a mutation at the nucleocapsid-p1 cleavage site (Prabu-Jeyabalan et al., 2004). It remains to be seen if current methodology can be extended to investigating the relation between the protease and other cleavage sites as well as the correlations with other regions in HIV-1 pre-proteins, toward shedding more light on the late stages of the virus maturation.

7.0 CONCLUSION AND FUTURE WORK

In this dissertation we have performed comprehensive analyses of a collection of example proteins, combining the computational tools of both sequence and structural analysis. Our results have indicated the intrinsic correlation between sequence evolution and structural dynamics, which provide related yet complementary information about the protein functions. The current results have improved our understanding about the interplay between sequence, structure, dynamics, and function; nevertheless, it still remains to search for more quantitative descriptions to some of the fundamental questions. For example, what is the statistical significance of claiming a loop as being involved in substrate recognition if it contains highly co-evolving residues with high mobility? Or more generally, to what degree can we predict the function of the protein given sufficient sequence and structural information? The answers to these questions are critical for developing knowledge-based models aiming to predict the protein function.

As shown in Chapter 5, a common feature shared by the conservation and mobility profiles is their continuity: sequentially neighboring residues tend to exhibit similar level of conservation/mobility. Based on the co-varying conservation and mobility profiles, the residues are naturally divided into groups centered at the peaks/valleys. For allostery, this means the interactions that mediate the signal transduction may largely depend on the collective efforts of groups of aggregated residues. Most existing computational models, unfortunately, have not accounted for such information by only focusing on pairwise interactions/co-evolution between

individual residues. It may be interesting to see if the incorporation of such information can lead to significant improvement of the existing approaches.

Our current work on the Hsp70 has focused on identifying key interactions based on structure information in the "stable" states; a more challenging problem is to understand the conformational changes involved in the state transition, which may underlie multiple transient states of both domains. Indeed, the interfacial residues studied in Chapter 4 may play a key role in the early stage of the conformational transition; e.g., the PRS results can be used to investigate the driving forces that yield the initial departure of the two domains. Our collaborators are performing "soft mutations" (personal communications with Dr. Zhuravleva from Gierasch lab) to probe the local dynamical influence of mutating some of the key residues without impairing the stable conformation, and the obtained results may be used to benchmark against our computational results. In addition, the directional analysis using PRS may also be used to obtain more specific information about the mutual regulation between the allosteric sites.

More recently, increasing attention has been drawn to the C-terminus of the SBD, including the α-helical lid and an unstructured C-terminal fragment. In 2009, an anti-cancer inhibitor of human Hsp70 (PES, (Leu et al., 2009)) has been proposed to interact with the α-helical domain; moreover, a study from our collaborator has proposed the C-terminal disordered region interacts with the folding client transiently to enhance the chaperone function (Smock et al., 2011). It is expected that the emerging new experimental data may help to explain some of the observations obtained from the computational studies, thus to reveal a more complete view of the Hsp70 allosteric mechanism.

APPENDIX A

SUPPLEMENTARY MATERIALS

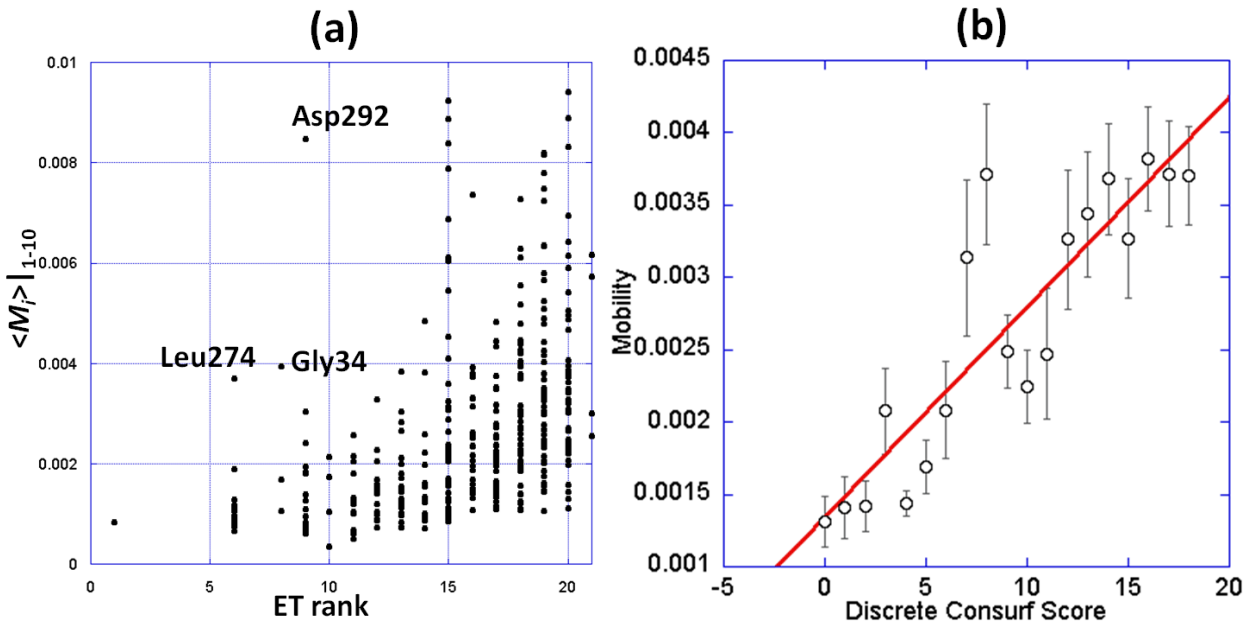

Figure A1. Comparison of residue mobilities with their evolutionary conservation properties.

(a) The mobility for each residue averaged over the first 10 GNM modes is plotted against its ET rank. Three outliers for ET rank 6, 8 and 9 are labeled, two of which (Gly34 and Asp292) are NEF-contacting residues. (b) Proportionality between the discretized ConSurf score and the average mobility (average $<M>|_{10}$) for residues with the same discrete ConSurf score. The discretization is performed by sorting all residues according to the ConSurf score, grouping every 20 consecutive residues and evaluating the mean mobility for each group. The correlation coefficient between average mobility and discrete ConSurf score is 0.88.

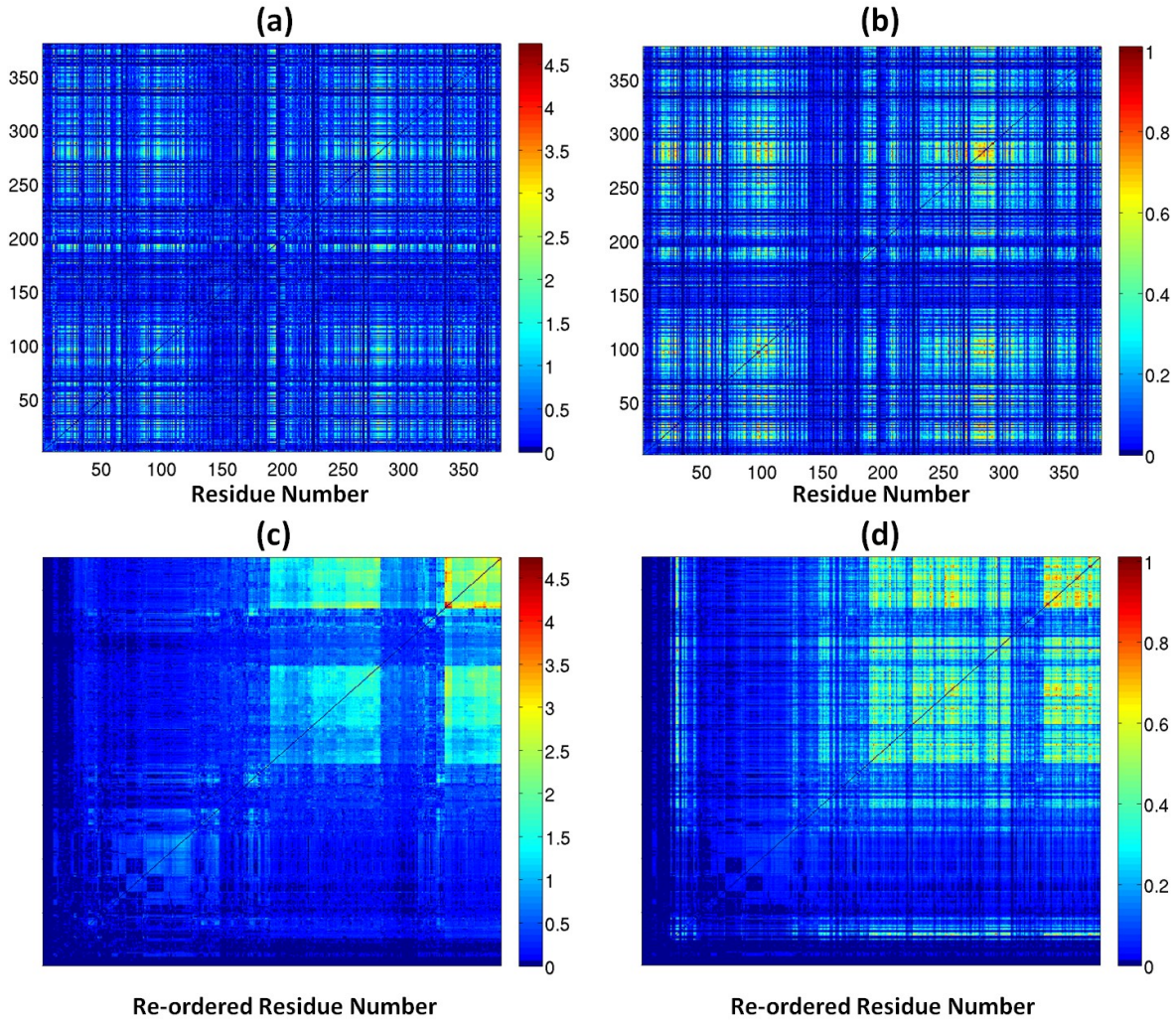

Figure A2. Comparison of correlated mutations map obtained by MI analysis and by the SCA.

The first two panels are the correlation maps calculated using (a) SCA and (b) MI. (c) The SCA correlation map after hierarchical clustering (note that the abscissa does not correspond to sequential residues anymore, but those rank-ordered according to their extent of correlated mutations. (d) The MI correlation map with residues re-ordered according to the same permutation in panel c.

As a benchmark of performance, we performed SCA calculations (SCA version 3.0) on the same MSA that were used for MI analysis, and compared the results (Figure A2).

The correlation matrices calculated in both analyses show similar patterns. Subdomain IIB exhibits the highest degree of correlation with a wide range of residues in both cases; however, we notice that the signals in the MI analysis are more distinctive, whereas the range of values in the SCA matrix is range (0-4.8 as opposed to 0-1 in the MI matrix). In the clustered maps, the high correlation in certain regions of the MI matrix has been suppressed in the SCA matrix, which may be attributed to the noise reduction step in SCA. Overall, in both matrices the majority of NEF-contact residues are identified as highly correlated with each other.

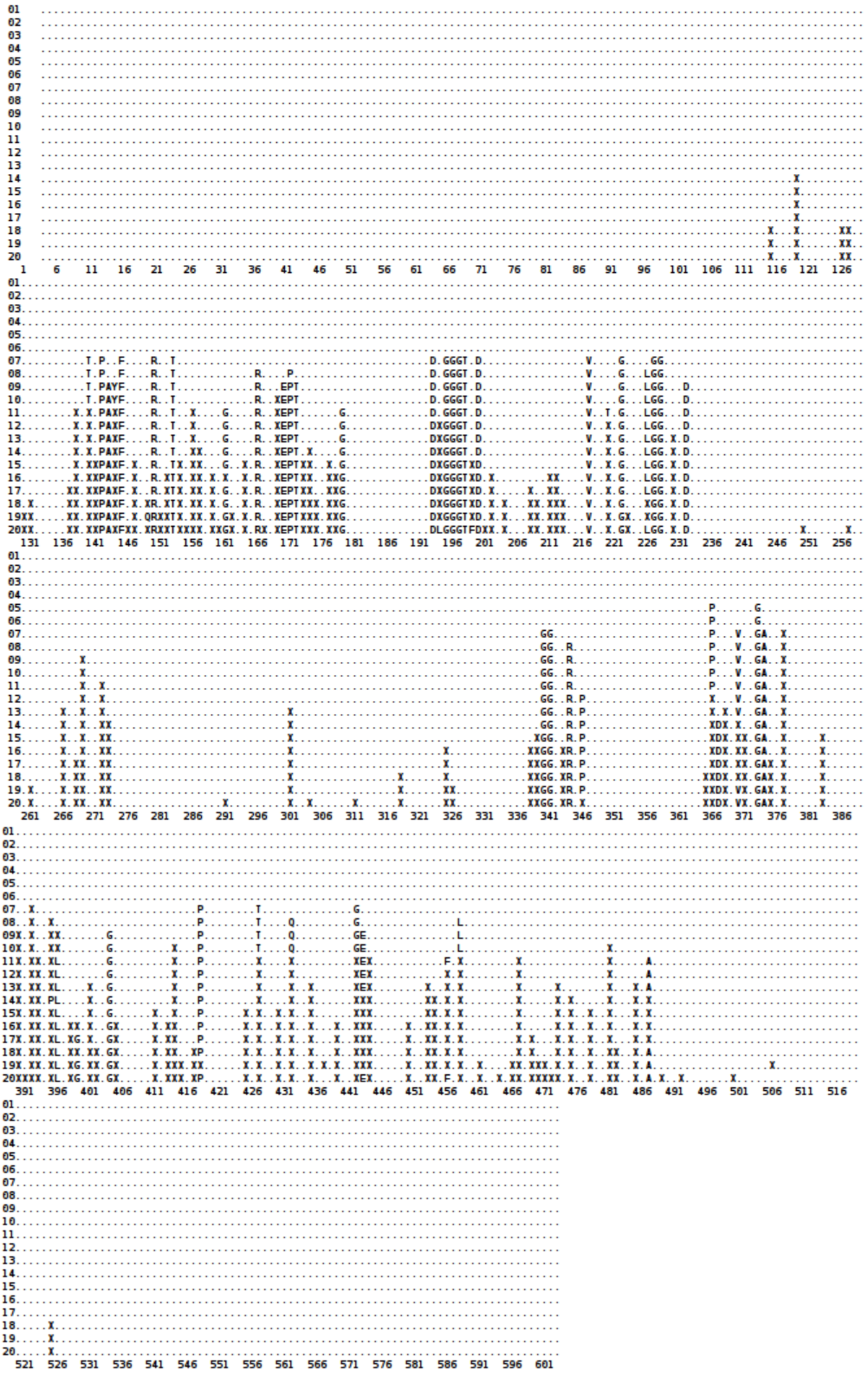

Figure A3. Evolutionary trace (Lichtarge et al., 1996) of DnaK residues 4-604.

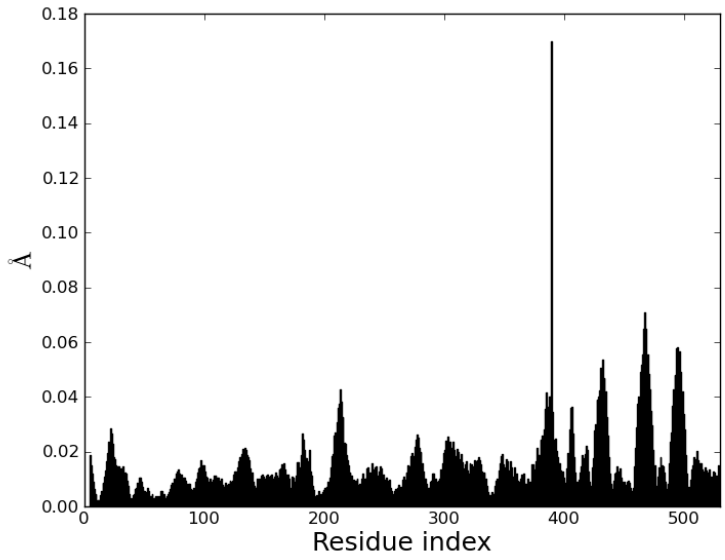

Figure A4. Displacement of residues in response to 100 pN applied at Val389 in DnaK$_{530}$.

The force constant (48 pN/Å) of the homology model DnaK$_{530}$ is calculated based on its template using the ANM web server (Eyal et al., 2006).

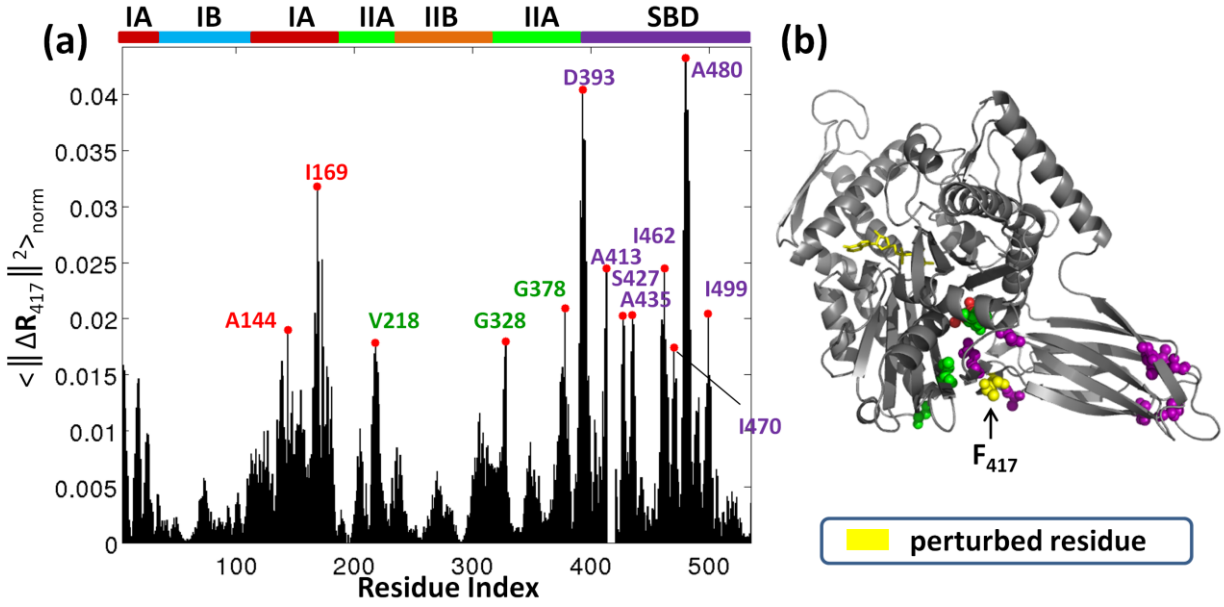

Figure A5. Responses to the perturbation at Thr417.

(a) Response profile of perturbing Thr417 ($<\|\Delta \mathbf{R}^{(417)}\|^2>_{norm}$). Peaks highlight the HS residues in the presence of this perturbation. Labels are colored according to the subdomains in which the labeled residues participate. (b)

Ribbon diagram highlighting the HS residues. The HS residues are shown in spheres representation and colored according to their subdomain location in the ATPase domain and SBD, and the ATP molecule is shown in yellow stick representation.

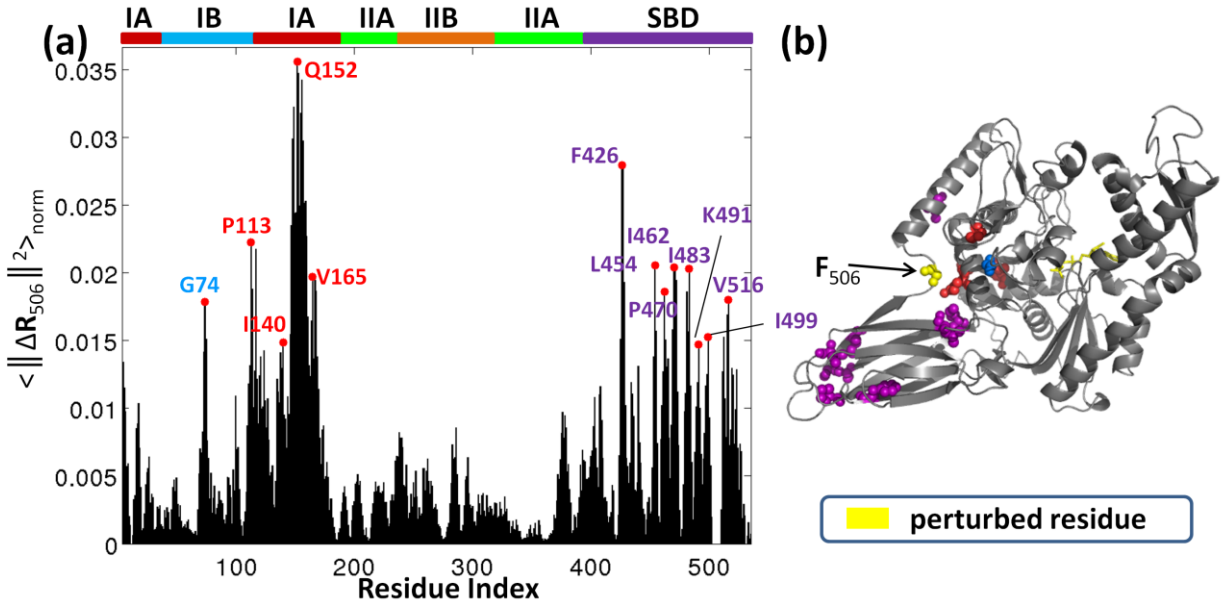

Figure A6. Responses to the perturbation at Gly506.

(a) Response profile of perturbing Gly506 ($<\|\Delta\mathbf{R}^{(506)}\|^2>_{norm}$). Peaks highlight the HS residues in the presence of this perturbation. Labels are colored according to the subdomains in which the labeled residues participate. . (b) Ribbon diagram highlighting the HS residues. The HS residues are shown in spheres representation and colored according to their subdomain location in the ATPase domain and SBD, and the ATP molecule is shown in yellow stick representation.

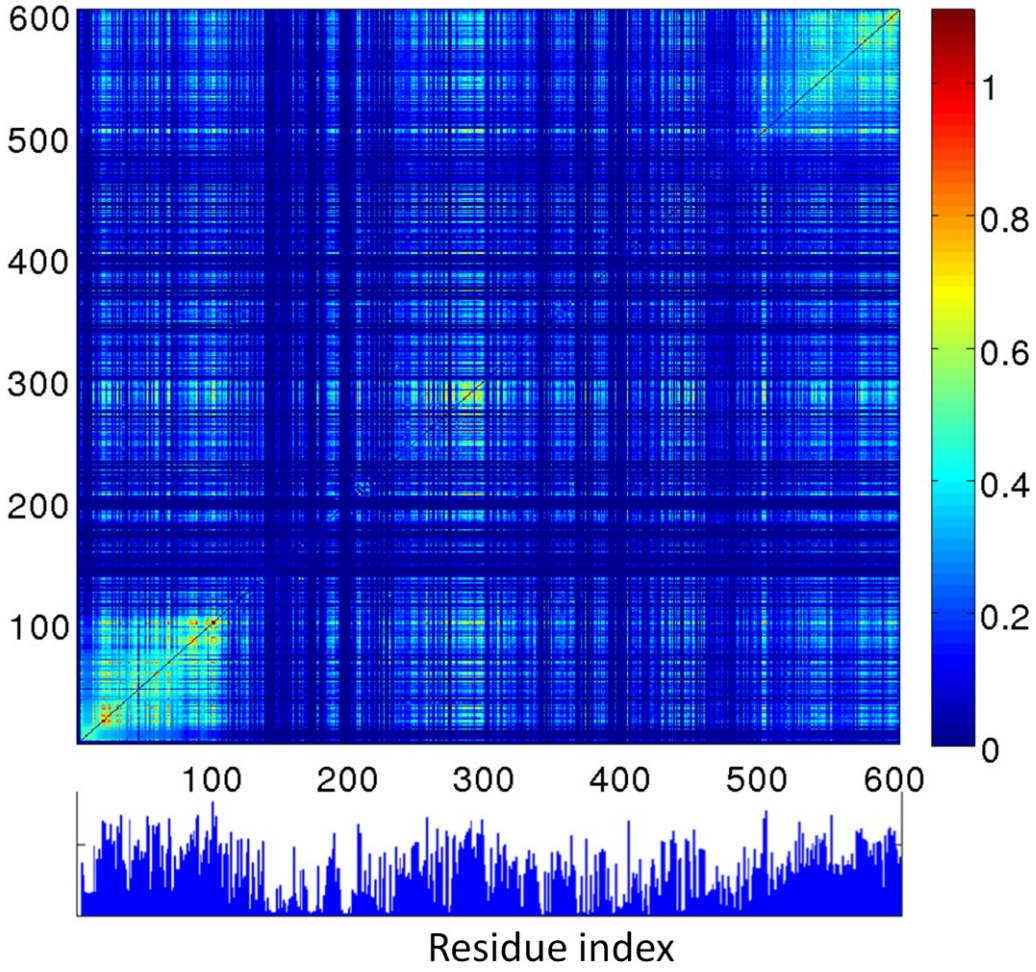

Figure A7. Mutual information map of DnaK (residues 4-604).

The blue bars below show the average MI values I(*i*) of each residue.

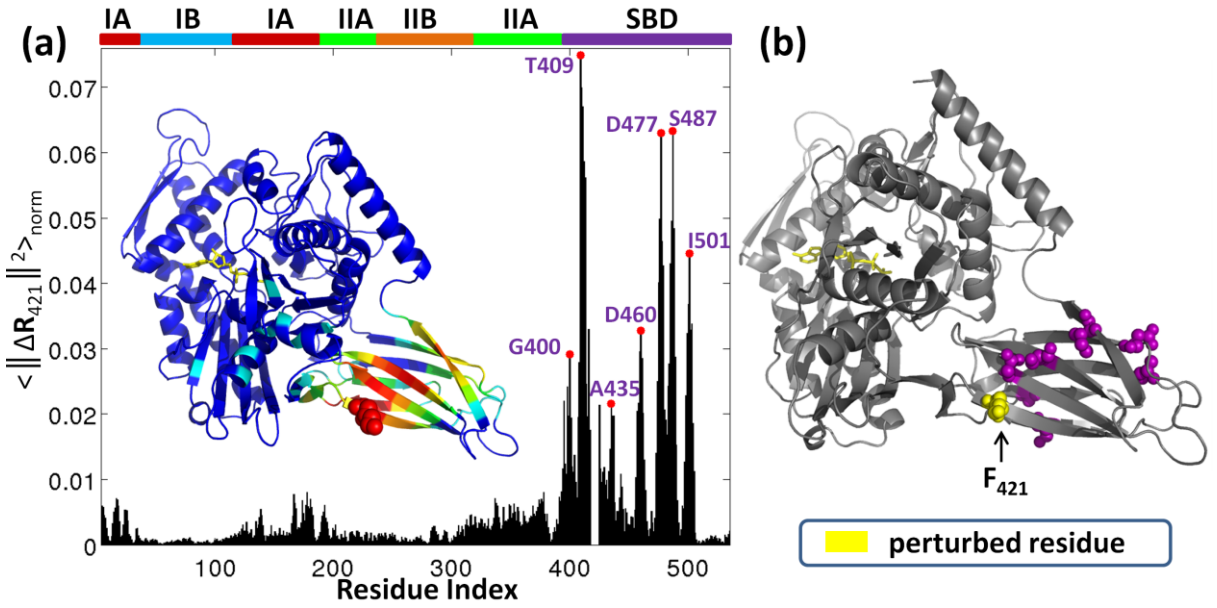

Figure A8. Responses to the perturbation at Lys421.

(a) Response profile of perturbing Lys421 ($<\|\Delta\mathbf{R}^{(421)}\|^2>_{norm}$). Peaks highlight the HS residues in the presence of this perturbation. Labels are colored according to the subdomains in which the labeled residues participate. The inset ribbon diagram is color-coded in the order of decreasing response $<\|\Delta R^{(421)}\|^2>_{norm}$, from red to blue.. (b) Ribbon diagram highlighting the HS residues. The HS residues are shown in spheres representation and colored according to their subdomain location in the ATPase domain and SBD, and the ATP molecule is shown in yellow stick representation.

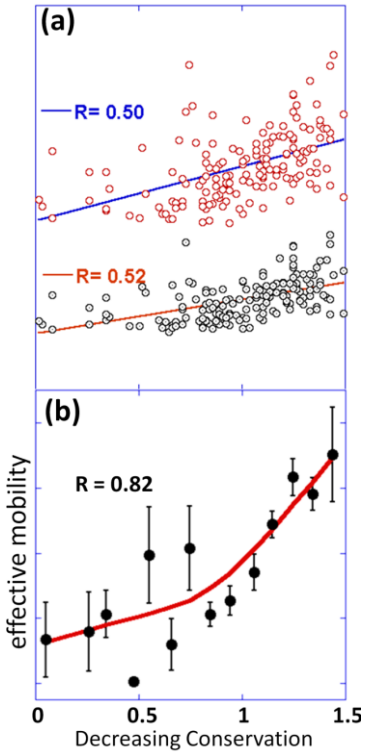

Figure A9. An illustrative example: uracil-DNA glycosylase (UDG).

(a) Scatter plot of mobility against entropy. Each point refers to a residue. Red and black dots correspond to $<M_i>|_{m2}$, and $<M_i>|_{N\text{-}1}$, respectively. The respective correlation coefficients are 0.50 and 0.52, although the correlations are statistically significant (p-values of the order of 10^{-12}) (**Table 8**). (b) Relationship between the conservation and effective mobility for UDG. The data in panel a are consolidated by evaluating the average mobilities (here MSFs) for entropy intervals of $\Delta S = 0.1$, and the corresponding grid-based mapping scheme of $<M_k^{\text{eff}}>|_{N\text{-}1}$ values are plotted for $1 \leq k \leq 15$. Best fitting curve and error bars are indicated. The resulting correlation coefficient is 0.82.

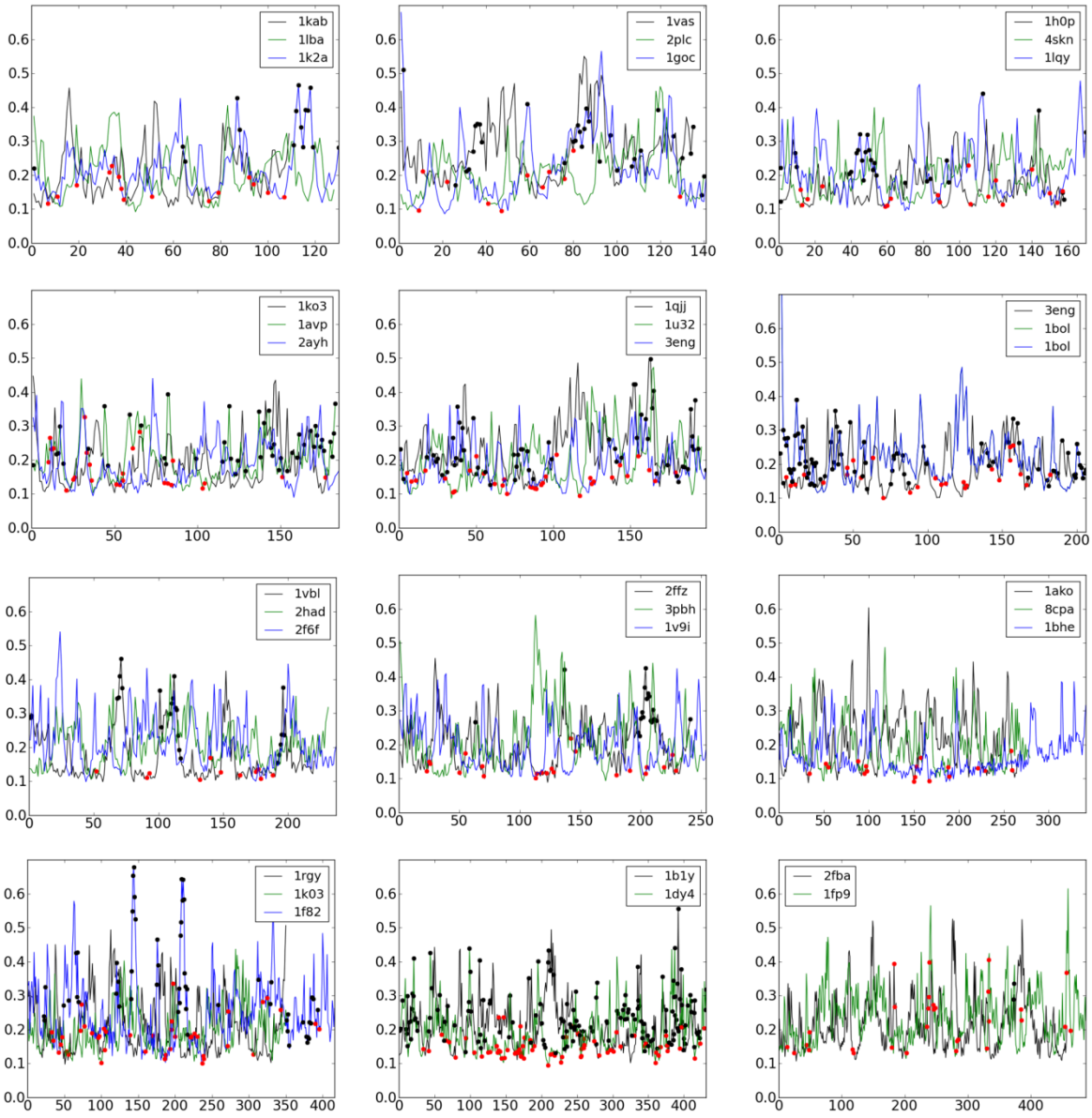

Figure A10. The location of highly conserved (red dots, $S < 0.1$) and most variable residues (black dots, $S > 1.6$) on the MSF.

The curves refer to all 34 enzymes studied here (**Table 6**). Enzymes are grouped based on their size.

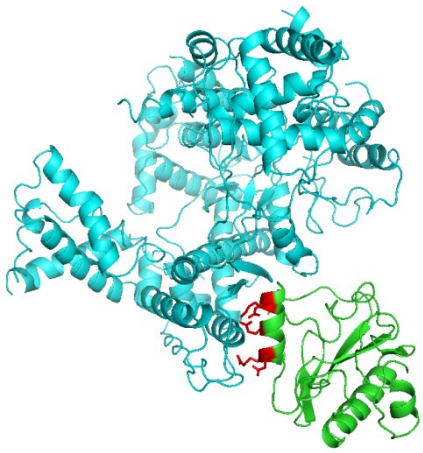

Figure A11. Interaction of T7 lysozyme with T7 RNA polymerase.

Highlighted T7 lysozyme residues (in red) are Arg30, Glu31, Gln34, Lys37 and Glu38. Among them Gln34 is distinguished by their high co-evolutionary properties (shown by squares in **Figure 42**, panel b). The structure has been generated using the PDB file 1ARO (Jeruzalmi and Steitz, 1998).

BIBLIOGRAPHY

Alberti, S., Esser, C., and Hohfeld, J. 2003. BAG-1--a nucleotide exchange factor of Hsc70 with multiple cellular functions. *Cell Stress.Chaperones.* **8**:225-231.

Ali, M. M., Roe, S. M., Vaughan, C. K., Meyer, P., Panaretou, B., Piper, P. W., Prodromou, C., and Pearl, L. H. 2006. Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature* **440**:1013-1017.

Andreasson, C., Fiaux, J., Rampelt, H., Mayer, M. P., and Bukau, B. 2008. Hsp110 is a nucleotide-activated exchange factor for Hsp70. *J.Biol.Chem.* **283**:8877-8884.

Armon, A., Graur, D., and Ben-Tal, N. 2001. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J.Mol.Biol.* **307**:447-463.

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., and Dress, A. W. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol.Biol.Evol.* **17**:164-178.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys.J.* **80**:505-515.

Atilgan, C., and Atilgan, A. R. 2009. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput.Biol.* **5**:e1000544.

Atilgan, C., Gerek, Z. N., Ozkan, S. B., and Atilgan, A. R. 2010. Manipulation of conformational change in proteins by single-residue perturbations. *Biophys.J.* **99**:933-943.

Bahar, I., Atilgan, A. R., Demirel, M. C., and Erman, B. 1998. Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Physical Review Letters* **80**:2733-2736.

Bahar, I., Atilgan, A. R., and Erman, B. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold.Des* **2**:173-181.

Bahar, I., Chennubhotla, C., and Tobi, D. 2007. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr.Opin.Struct.Biol.* **17**:633-640.

Bahar, I., Lezon, T. R., Yang, L. W., and Eyal, E. 2010. Global dynamics of proteins: bridging between structure and function. *Annu.Rev.Biophys.* **39**:23-42.

Bahar, I., and Rader, A. J. 2005. Coarse-grained normal mode analysis in structural biology. *Curr.Opin.Struct.Biol.* **15**:586-592.

Bakan, A., and Bahar, I. 2009. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc.Natl.Acad.Sci.U.S.A* **106**:14349-14354.

Barabasi, A. L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* **286**:509-512.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**:235-242.

Bertelsen, E. B., Chang, L., Gestwicki, J. E., and Zuiderweg, E. R. 2009. Solution conformation of wild-type E. coli Hsp70 (DnaK) chaperone complexed with ADP and substrate. *Proc.Natl.Acad.Sci.U.S.A* **106**:8471-8476.

Betts, M. J., and R. B. Russell. 2007. Amino-Acid Properties and Consequences of Substitutions. Pages 311-342 *in* Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data (M. R. Barnes, Ed.). John Wiley & Sons, Ltd, Chichester.

Bhabha, G., Lee, J., Ekiert, D. C., Gam, J., Wilson, I. A., Dyson, H. J., Benkovic, S. J., and Wright, P. E. 2011. A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science* **332**:234-238.

Bhattacharya, A., Kurochkin, A. V., Yip, G. N., Zhang, Y., Bertelsen, E. B., and Zuiderweg, E. R. 2009. Allostery in Hsp70 chaperones is transduced by subdomain rotations. *J.Mol.Biol.* **388**:475-490.

Bork, P., Sander, C., and Valencia, A. 1992. An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc.Natl.Acad.Sci.U.S.A* **89**:7290-7294.

Brehmer, D., Rudiger, S., Gassler, C. S., Klostermeier, D., Packschies, L., Reinstein, J., Mayer, M. P., and Bukau, B. 2001. Tuning of chaperone activity of Hsp70 proteins by modulation of nucleotide exchange. *Nat.Struct.Biol.* **8**:427-432.

Brik, A., and Wong, C. H. 2003. HIV-1 protease: mechanism and drug discovery. *Org.Biomol.Chem.* **1**:5-14.

Bukau, B., and Horwich, A. L. 1998. The Hsp70 and Hsp60 chaperone machines. *Cell* **92**:351-366.

Burley, S. K., and Petsko, G. A. 1986. Amino-aromatic interactions in proteins. *FEBS Lett.* **203**:139-143.

Cecconi, F., Micheletti, C., Carloni, P., and Maritan, A. 2001. Molecular dynamics studies on HIV-1 protease drug resistance and folding pathways. *Proteins* **43**:365-372.

Changeux, J. P., and Edelstein, S. J. 1998. Allosteric receptors after 30 years. *Neuron* **21**:959-980.

Changeux, J. P., and Edelstein, S. J. 2005. Allosteric mechanisms of signal transduction. *Science* **308**:1424-1428.

Chen, Y., Reilly, K., and Chang, Y. 2006. Evolutionarily conserved allosteric network in the Cys loop family of ligand-gated ion channels revealed by statistical covariance analyses. *J.Biol.Chem.* **281**:18184-18192.

Chennubhotla, C., and Bahar, I. 2006. Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. *Molecular Systems Biology*.

Chennubhotla, C., Yang, Z., and Bahar, I. 2008. Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. *Mol.Biosyst.* **4**:287-292.

Chung, F. 1997. Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92). American Mathematical Society.

Cormen, T., C. Leiserson, R. Rivest, and C. Stein. 2001. Introduction to Algorithms. The MIT Press.

Cotton, F. A., Hazen, E. E., Jr., and Legg, M. J. 1979. Staphylococcal nuclease: proposed mechanism of action based on structure of enzyme-thymidine 3',5'-bisphosphate-calcium ion complex at 1.5-A resolution. *Proc Natl.Acad.Sci.U.S.A* **76**:2551-2555.

Cover, T., and J. Thomas. 1991. Elements of Information Theory. Wiley-Interscience, New York.

Craig, E. A., Huang, P., Aron, R., and Andrew, A. 2006. The diverse roles of J-proteins, the obligate Hsp70 co-chaperone. *Rev.Physiol Biochem.Pharmacol.* **156**:1-21.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res.* **14**:1188-1190.

Crowley, P. B., and Golovin, A. 2005. Cation-pi interactions in protein-protein interfaces. *Proteins* **59**:231-239.

Csermely, P., Palotai, R., and Nussinov, R. 2010. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem.Sci.* **35**:539-546.

Cuellar, J., Martin-Benito, J., Scheres, S. H., Sousa, R., Moro, F., Lopez-Vinas, E., Gomez-Puertas, P., Muga, A., Carrascosa, J. L., and Valpuesta, J. M. 2008. The structure of

CCT-Hsc70 NBD suggests a mechanism for Hsp70 delivery of substrates to the chaperonin. *Nat.Struct.Mol.Biol.* **15**:858-864.

Davis, J. E., Voisine, C., and Craig, E. A. 1999. Intragenic suppressors of Hsp70 mutants: interplay between the ATPase- and peptide-binding domains. *Proc Natl.Acad.Sci.U.S.A* **96**:9269-9276.

del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. 2006. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular Systems Biology*.

del Sol, A., and O'Meara, P. 2005. Small-world network approach to identify key residues in protein-protein interaction. *Proteins* **58**:672-682.

Dobbins, S. E., Lesk, V. I., and Sternberg, M. J. 2008. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc.Natl.Acad.Sci.U.S.A* **105**:10390-10395.

Doruker, P., Atilgan, A. R., and Bahar, I. 2000. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins* **40**:512-524.

Dunn, S. D., Wahl, L. M., and Gloor, G. B. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**:333-340.

Dutta, A., and Bahar, I. 2010. Metal-binding sites are designed to achieve optimal mechanical and signaling properties. *Structure* **18**:1140-1148.

Eyal, E., Pietrokovski, S., and Bahar, I. 2007. Rapid assessment of correlated amino acids from pair-to-pair (P2P) substitution matrices. *Bioinformatics* **23**:1837-1839.

Eyal, E., Yang, L. W., and Bahar, I. 2006. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics* **22**:2619-2627.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. 2008. The Pfam protein families database. *Nucleic Acids Res.* **36**:D281-D288.

Fischer, D., Bachar, O., Nussinov, R., and Wolfson, H. 1992. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J.Biomol.Struct.Dyn.* **9**:769-789.

Fitch, W. M., and Margoliash, E. 1967. Construction of phylogenetic trees. *Science* **155**:279-284.

Flaherty, K. M., Luca-Flaherty, C., and McKay, D. B. 1990. Three-dimensional structure of the ATPase fragment of a 70K heat-shock cognate protein. *Nature* **346**:623-628.

Flaherty, K. M., Wilbanks, S. M., Luca-Flaherty, C., and McKay, D. B. 1994. Structural basis of the 70-kilodalton heat shock cognate protein ATP hydrolytic activity. II. Structure of the active site with ADP or ATP bound to wild type and mutant ATPase fragment. *J.Biol.Chem.* **269**:12899-12907.

Fleishman, S. J., Yifrach, O., and Ben-Tal, N. 2004. An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. *J.Mol.Biol.* **340**:307-318.

Fodor, A. A., and Aldrich, R. W. 2004. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56**:211-221.

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19**:163-164.

Gonzales, M. J., Machekano, R. N., and Shafer, R. W. 2001. Human immunodeficiency virus type 1 reverse-transcriptase and protease subtypes: classification, amino acid mutation patterns, and prevalence in a northern California clinic-based population. *J.Infect.Dis.* **184**:998-1006.

Gotoh, O. 1992. Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *J.Biol Chem* **267**:83-90.

Greene, L. H., and Higman, V. A. 2003. Uncovering network systems within protein structures. *J.Mol.Biol.* **334**:781-791.

Gunasekaran, K., Ma, B., and Nussinov, R. 2004. Is allostery an intrinsic property of all dynamic proteins? *Proteins* **57**:433-443.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. 2009. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**:774-786.

Halperin, I., Wolfson, H., and Nussinov, R. 2006. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins* **63**:832-845.

Harrison, C. J., Hayer-Hartl, M., Di, L. M., Hartl, F., and Kuriyan, J. 1997. Crystal structure of the nucleotide exchange factor GrpE bound to the ATPase domain of the molecular chaperone DnaK. *Science* **276**:431-435.

Hartl, F. U., and Hayer-Hartl, M. 2002. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**:1852-1858.

Hartl, F. U., and Hayer-Hartl, M. 2009. Converging concepts of protein folding in vitro and in vivo. *Nat.Struct.Mol.Biol.* **16**:574-581.

Hatley, M. E., Lockless, S. W., Gibson, S. K., Gilman, A. G., and Ranganathan, R. 2003. Allosteric determinants in guanine nucleotide-binding proteins. *Proc.Natl.Acad.Sci.U.S.A* **100**:14445-14450.

Henzler-Wildman, K. A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M. A., Petsko, G. A., Karplus, M., Hubner, C. G., and Kern, D. 2007. Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450**:838-844.

Hertogs, K., Bloor, S., Kemp, S. D., Van den, E. C., Alcorn, T. M., Pauwels, R., Van, H. M., Staszewski, S., Miller, V., and Larder, B. A. 2000. Phenotypic and genotypic analysis of clinical HIV-1 isolates reveals extensive protease inhibitor cross-resistance: a survey of over 6000 samples. *AIDS* **14**:1203-1210.

Hoffman, N. G., Schiffer, C. A., and Swanstrom, R. 2003. Covariation of amino acid positions in HIV-1 protease. *Virology* **314**:536-548.

Hornak, V., Okur, A., Rizzo, R. C., and Simmerling, C. 2006. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc.Natl.Acad.Sci.U.S.A* **103**:915-920.

Hu, Z., Ma, B., Wolfson, H., and Nussinov, R. 2000. Conservation of polar residues as hot spots at protein interfaces. *Proteins* **39**:331-342.

Ikeguchi, M., Ueno, J., Sato, M., and Kidera, A. 2005. Protein structural change upon ligand binding: linear response theory. *Phys.Rev.Lett.* **94**:078102.

Illy, C., Quraishi, O., Wang, J., Purisima, E., Vernet, T., and Mort, J. S. 1997. Role of the occluding loop in cathepsin B activity. *J.Biol.Chem.* **272**:1197-1202.

Innis, C. A., Shi, J., and Blundell, T. L. 2000. Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng* **13**:839-847.

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J., McGarvey, P., and Gasteiger, E. 2009. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10**:136.

James, L. C., Roversi, P., and Tawfik, D. S. 2003. Antibody multispecificity mediated by conformational diversity. *Science* **299**:1362-1367.

Jeruzalmi, D., and Steitz, T. A. 1998. Structure of T7 RNA polymerase complexed to the transcriptional inhibitor T7 lysozyme. *EMBO J.* **17**:4101-4113.

Jiang, J., Maes, E. G., Taylor, A. B., Wang, L., Hinck, A. P., Lafer, E. M., and Sousa, R. 2007. Structural basis of J cochaperone binding and regulation of Hsp70. *Mol.Cell* **28**:422-433.

Jones, S., and Thornton, J. M. 1997. Analysis of protein-protein interaction sites using surface patches. *J.Mol.Biol.* **272**:121-132.

Kabani, M. 2009. Structural and functional diversity among eukaryotic Hsp70 nucleotide exchange factors. *Protein Pept.Lett.* **16**:623-660.

Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F., and Holmes, K. C. 1990. Atomic structure of the actin:DNase I complex. *Nature* **347**:37-44.

Kantor, R., Katzenstein, D. A., Efron, B., Carvalho, A. P., Wynhoven, B., Cane, P., Clarke, J., Sirivichayakul, S., Soares, M. A., Snoeck, J., Pillay, C., Rudich, H., Rodrigues, R., Holguin, A., Ariyoshi, K., Bouzas, M. B., Cahn, P., Sugiura, W., Soriano, V., Brigido, L. F., Grossman, Z., Morris, L., Vandamme, A. M., Tanuri, A., Phanuphak, P., Weber, J. N., Pillay, D., Harrigan, P. R., Camacho, R., Schapiro, J. M., and Shafer, R. W. 2005. Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration. *PLoS Med.* **2**:e112.

Kimura, E. 2001. Model studies for molecular recognition of carbonic anhydrase and carboxypeptidase. *Acc.Chem Res.* **34**:171-179.

Korber, B., and Myers, G. 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. *AIDS Res.Hum.Retroviruses* **8**:1549-1560.

Koshland, D. E., Jr., Nemethy, G., and Filmer, D. 1966. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* **5**:365-385.

Kovbasyuk, L., and Kramer, R. 2004. Allosteric supramolecular receptors and catalysts. *Chem.Rev.* **104**:3161-3187.

Kozal, M. 2004. Cross-resistance patterns among HIV protease inhibitors. *AIDS Patient Care STDS* **18**:199-208.

Lafont, V., Armstrong, A. A., Ohtaka, H., Kiso, Y., Mario, A. L., and Freire, E. 2007. Compensating enthalpic and entropic changes hinder binding affinity optimization. *Chem.Biol.Drug Des* **69**:413-422.

Lange, O. F., Lakomek, N. A., Fares, C., Schroder, G. F., Walter, K. F., Becker, S., Meiler, J., Grubmuller, H., Griesinger, C., and de Groot, B. L. 2008. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320**:1471-1475.

Leu, J. I., Pimkina, J., Frank, A., Murphy, M. E., and George, D. L. 2009. A small molecule inhibitor of inducible heat shock protein 70. *Mol.Cell* **36**:15-27.

Li, H., Helling, R., Tang, C., and Wingreen, N. 1996. Emergence of preferred structures in a simple model of protein folding. *Science* **273**:666-669.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J.Mol.Biol.* **257**:342-358.

Lichtarge, O., and Sowa, M. E. 2002. Evolutionary predictions of binding surfaces and interactions. *Curr.Opin.Struct.Biol.* **12**:21-27.

Liebscher, M., and Roujeinikova, A. 2009. Allosteric coupling between the lid and interdomain linker in DnaK revealed by inhibitor binding studies. *J.Bacteriol.* **191**:1456-1462.

Liu, Q. L., and Hendrickson, W. A. 2007. Insights into Hsp70 chaperone activity from a crystal structure of the yeast Hsp110 Sse1. *Cell* **131**:106-120.

Liu, Y., and Bahar, I. 2010. Toward understanding allosteric signaling mechanisms in the ATPase domain of molecular chaperones. *Pac.Symp.Biocomput.*269-280.

Liu, Y., and Bahar, I. 2011. Sequence evolution correlates with structural dynamics. *Proc Natl.Acad.Sci.U.S.A*(submitted).

Liu, Y., Eyal, E., and Bahar, I. 2008. Analysis of correlated mutations in HIV-1 protease using spectral clustering. *Bioinformatics* **24**:1243-1250.

Liu, Y., Gierasch, L. M., and Bahar, I. 2010. Role of Hsp70 ATPase Domain Intrinsic Dynamics and Sequence Evolution in Enabling its Functional Interactions with NEFs. *PLoS Comput.Biol.* **6**:e1000931.

Livingstone, C. D., and Barton, G. J. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput.Appl.Biosci.* **9**:745-756.

Lockless, S. W., and Ranganathan, R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**:295-299.

Lovell, S. C., and Robertson, D. L. 2010. An integrated view of molecular coevolution in protein-protein interactions. *Mol.Biol.Evol.* **27**:2567-2575.

Luque, I., and Freire, E. 2000. Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins* **Suppl 4**:63-71.

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc.Natl.Acad.Sci.U.S.A* **100**:5772-5777.

McLellan, C. A., Raynes, D. A., and Guerriero, V. 2003. HspBP1, an Hsp70 cochaperone, has two structural domains and is capable of altering the conformation of the Hsp70 ATPase domain. *J.Biol.Chem.* **278**:19017-19022.

Mintseris, J., and Weng, Z. 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc.Natl.Acad.Sci.U.S.A* **102**:10930-10935.

Mirny, L. A., and Shakhnovich, E. I. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J.Mol.Biol.* **291**:177-196.

Mittag, T., Kay, L. E., and Forman-Kay, J. D. 2010. Protein dynamics and conformational disorder in molecular recognition. *J.Mol.Recognit.* **23**:105-116.

Mittermaier, A., Davidson, A. R., and Kay, L. E. 2003. Correlation between 2H NMR side-chain order parameters and sequence conservation in globular proteins. *J.Am.Chem.Soc.* **125**:9004-9005.

Monod, J., Wyman, J., and hangeux, J. P. 1965. On the nature of allosteric transitions: a plausible model. *J.Mol.Biol.* **12**:88-118.

Montgomery, D. L., Morimoto, R. I., and Gierasch, L. M. 1999. Mutations in the substrate binding domain of the Escherichia coli 70 kDa molecular chaperone, DnaK, which alter substrate affinity or interdomain coupling. *J.Mol.Biol.* **286**:915-932.

Moro, F., Fernandez, V., and Muga, A. 2003. Interdomain interaction through helices A and B of DnaK peptide binding domain. *Febs Letters* **533**:119-123.

Nair, S. K., Calderone, T. L., Christianson, D. W., and Fierke, C. A. 1991. Altering the mouth of a hydrophobic pocket. Structure and kinetics of human carbonic anhydrase II mutants at residue Val-121. *J.Biol Chem* **266**:17320-17325.

Noivirt, O., Eisenstein, M., and Horovitz, A. 2005. Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng.* **18**:247-253.

Nowak, M. A., Boerlijst, M. C., Cooke, J., and Smith, J. M. 1997. Evolution of genetic redundancy. *Nature* **388**:167-171.

O'Brien, M. C., Flaherty, K. M., and McKay, D. B. 1996. Lysine 71 of the chaperone protein Hsc70 Is essential for ATP hydrolysis. *J.Biol.Chem.* **271**:15874-15878.

Ohtaka, H., Schon, A., and Freire, E. 2003. Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations. *Biochemistry* **42**:13659-13666.

Olmea, O., Rost, B., and Valencia, A. 1999. Effective use of sequence correlation and conservation in fold recognition. *J.Mol.Biol.* **293**:1221-1239.

Pellecchia, M., Montgomery, D. L., Stevens, S. Y., Vander Kooi, C. W., Feng, H. P., Gierasch, L. M., and Zuiderweg, E. R. 2000. Structural insights into substrate binding by the molecular chaperone DnaK. *Nat.Struct.Biol.* **7**:298-303.

Perryman, A. L., Lin, J. H., and McCammon, J. A. 2004. HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: possible contributions to drug resistance and a potential new target site for drugs. *Protein Sci.* **13**:1108-1123.

Podobnik, M., Kuhelj, R., Turk, V., and Turk, D. 1997. Crystal structure of the wild-type human procathepsin B at 2.5 A resolution reveals the native active site of a papain-like cysteine protease zymogen. *J.Mol.Biol.* **271**:774-788.

Polier, S., Dragovic, Z., Hartl, F. U., and Bracher, A. 2008. Structural basis for the cooperation of Hsp70 and Hsp110 chaperones in protein folding. *Cell* **133**:1068-1079.

Prabu-Jeyabalan, M., Nalivaika, E. A., King, N. M., and Schiffer, C. A. 2004. Structural basis for coevolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease. *J.Virol.* **78**:12446-12454.

Prabu-Jeyabalan, M., Nalivaika, E. A., Romano, K., and Schiffer, C. A. 2006. Mechanism of substrate recognition by drug-resistant human immunodeficiency virus type 1 protease variants revealed by a novel structural intermediate. *J.Virol.* **80**:3607-3616.

Rader, A. J., C. Chennubhotla, L. W. Yang, and I. Bahar. 2006. The Gaussian Network Model: theory and applications. Pages 41-63 *in* Normal Mode Analysis - theory and applications to biological and chemical systems (Q. Cui, and I. Bahar, Eds.). Chapman & Hall/CRC.

Rebek, J. 1990. On the structure of histidine and its role in enzyme active sites. *Struct.Chem.* **1**:129-131.

Renko, M., Pozgan, U., Majera, D., and Turk, D. 2010. Stefin A displaces the occluding loop of cathepsin B only by as much as required to bind to the active site cleft. *FEBS J.* **277**:4338-4345.

Rhee, S. Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **31**:298-303.

Sacquin-Mora, S., and Lavery, R. 2006. Investigating the local flexibility of functional residues in hemoproteins. *Biophys.J.* **90**:2706-2717.

Sakarya, O., Conaco, C., Egecioglu, O., Solla, S. A., Oakley, T. H., and Kosik, K. S. 2010. Evolutionary expansion and specialization of the PDZ domains. *Mol.Biol.Evol.* **27**:1058-1069.

Sali, A., and Blundell, T. L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J.Mol.Biol.* **234**:779-815.

Schmeing, T. M., Huang, K. S., Strobel, S. A., and Steitz, T. A. 2005. An induced-fit mechanism to promote peptide bond formation and exclude hydrolysis of peptidyl-tRNA. *Nature* **438**:520-524.

Schneider, T. D., and Stephens, R. M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**:6097-6100.

Schuermann JP, Jiang J, Cuellar J, Llorca O, Wang L, Gimenez LE, Jin S, Taylor AB, Demeler B, Morano KA, Hart PJ, Valpuesta JM, Lafer EM, and Sousa R. 2008. Structure of the Hsp110:Hsc70 Nucleotide Exchange Machine. *Molecular Cell* **31**:232-243.

Shackelford, G., and Karplus, K. 2007. Contact prediction using mutual information and neural nets. *Proteins* **69 Suppl 8**:159-164.

Shafer, R. W. 2002. Genotypic testing for human immunodeficiency virus type 1 drug resistance. *Clin.Microbiol.Rev.* **15**:247-277.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell system technical journal* **27**.

Shi, J., and Malik, J. 2000. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**:888-905.

Shomura, Y., Dragovic, Z., Chang, H. C., Tzvetkov, N., Young, J. C., Brodsky, J. L., Guerriero, V., Hartl, F. U., and Bracher, A. 2005. Regulation of Hsp70 function by HspBP1: structural analysis reveals an alternate mechanism for Hsp70 nucleotide exchange. *Mol.Cell* **17**:367-379.

Shulman, A. I., Larson, C., Mangelsdorf, D. J., and Ranganathan, R. 2004. Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* **116**:417-429.

Smith, T. F., and Waterman, M. S. 1981. Identification of common molecular subsequences. *J.Mol.Biol.* **147**:195-197.

Smock, R. G., Blackburn, M. E., and Gierasch, L. M. 2011. Conserved, Disordered C Terminus of DnaK Enhances Cellular Survival upon Stress and DnaK in Vitro Chaperone Activity. *J.Biol.Chem.* **286**:31821-31829.

Smock, R. G., and Gierasch, L. M. 2009. Sending signals dynamically. *Science* **324**:198-203.

Smock, R. G., Rivoire, O., Russ, W. P., Swain, J. F., Leibler, S., Ranganathan, R., and Gierasch, L. M. 2010. An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol.Syst.Biol.* **6**:414.

Sondermann, H., Scheufler, C., Schneider, C., Hohfeld, J., Hartl, F. U., and Moarefi, I. 2001. Structure of a Bag/Hsc70 complex: convergent functional evolution of Hsp70 nucleotide exchange factors. *Science* **291**:1553-1557.

Sriram, M., Osipiuk, J., Freeman, B., Morimoto, R., and Joachimiak, A. 1997. Human Hsp70 molecular chaperone binds two calcium ions within the ATPase domain. *Structure* **5**:403-414.

Swain, J. F., Dinler, G., Sivendran, R., Montgomery, D. L., Stotz, M., and Gierasch, L. M. 2007. Hsp70 chaperone ligands control domain association via an allosteric mechanism mediated by the interdomain linker. *Mol.Cell* **26**:27-39.

Swain, J. F., and Gierasch, L. M. 2006. The changing landscape of protein allostery. *Curr.Opin.Struct.Biol.* **16**:102-108.

Tama, F., and Brooks, C. L. 2006. Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu.Rev.Biophys.Biomol.Struct.* **35**:115-133.

Tama, F., and Sanejouand, Y. H. 2001. Conformational change of proteins arising from normal mode calculations. *Protein Eng* **14**:1-6.

Thanki, N., Rao, J. K., Foundling, S. I., Howe, W. J., Moon, J. B., Hui, J. O., Tomasselli, A. G., Heinrikson, R. L., Thaisrivongs, S., and Wlodawer, A. 1992. Crystal structure of a complex of HIV-1 protease with a dihydroxyethylene-containing inhibitor: comparisons with molecular modeling. *Protein Sci.* **1**:1061-1072.

Tobi, D., and Bahar, I. 2005. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc.Natl.Acad.Sci.U.S.A* **102**:18908-18913.

Tokuriki, N., and Tawfik, D. S. 2009. Protein dynamism and evolvability. *Science* **324**:203-207.

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. 2008. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu.Rev.Biophys.* **37**:215-246.

Vogel, M., Bukau, B., and Mayer, M. P. 2006a. Allosteric regulation of Hsp70 chaperones by a proline switch. *Mol.Cell* **21**:359-367.

Vogel, M., Mayer, M. P., and Bukau, B. 2006b. Allosteric regulation of Hsp70 chaperones involves a conserved interdomain linker. *J.Biol.Chem.* **281**:38705-38711.

von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing* **17**:395-416.

Wahba, G. 1990. Spline models for observational data, 59 ed. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Watts, D. J., and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature* **393**:440-442.

Webb, E. C. 1992. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. Academic Press.

Wilbanks, S. M., and McKay, D. B. 1995. How potassium affects the activity of the molecular chaperone Hsc70. II. Potassium binds specifically in the ATPase active site. *J.Biol.Chem.* **270**:2251-2257.

Worth, C. L., Gong, S., and Blundell, T. L. 2009. Structural and functional constraints in the evolution of protein families. *Nat.Rev.Mol.Cell Biol.* **10**:709-720.

Wright, P. E., and Dyson, H. J. 2009. Linking folding and binding. *Curr.Opin.Struct.Biol.* **19**:31-38.

Wu, T. D., Schiffer, C. A., Gonzales, M. J., Taylor, J., Kantor, R., Chou, S., Israelski, D., Zolopa, A. R., Fessel, W. J., and Shafer, R. W. 2003. Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *J.Virol.* **77**:4836-4847.

Xu, F., Du, P., Shen, H., Hu, H., Wu, Q., Xie, J., and Yu, L. 2009. Correlated mutation analysis on the catalytic domains of serine/threonine protein kinases. *PLoS.One.* **4**:e5913.

Yang, L. W., and Bahar, I. 2005. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure* **13**:893-904.

Yang, L. W., Rader, A. J., Liu, X., Jursa, C. J., Chen, S. C., Karimi, H. A., and Bahar, I. 2006. oGNM: online computation of structural dynamics using the Gaussian Network Model. *Nucleic Acids Res.* **34**:W24-W31.

Yang, Z., Majek, P., and Bahar, I. 2009. Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. *PLoS Comput.Biol.* **5**:e1000360.

Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M., Kavraki, L., and Lichtarge, O. 2003. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J.Mol.Biol.* **326**:255-261.

Yogurtcu, O. N., Erdemli, S. B., Nussinov, R., Turkay, M., and Keskin, O. 2008. Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations. *Biophys.J.* **94**:3475-3485.

Zen, A., Carnevale, V., Lesk, A. M., and Micheletti, C. 2008. Correspondences between low-energy modes in enzymes: dynamics-based alignment of enzymatic functional families. *Protein Sci.* **17**:918-929.

Zhu, X. T., Zhao, X., Burkholder, W. F., Gragerov, A., Ogata, C. M., Gottesman, M. E., and Hendrickson, W. A. 1996. Structural analysis of substrate binding by the molecular chaperone DnaK. *Science* **272**:1606-1614.

Zhuravleva, A., and Gierasch, L. M. 2011. Allosteric signal transmission in the nucleotide-binding domain of 70-kDa heat shock protein (Hsp70) molecular chaperones. *Proc.Natl.Acad.Sci.U.S.A* **108**:6987-6992.

Zoete, V., Michielin, O., and Karplus, M. 2002. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J.Mol.Biol.* **315**:21-52.