Sequence Evolution Correlates with Structural Dynamics

Ying Liu[1] and Ivet Bahar*,[1]

[1]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh

***Corresponding author:** E-mail: bahar@pitt.edu.

Associate editor: Willie Swanson

Abstract

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. Structural dynamics enable the adaptation of the protein to binding substrates and to undergo allosteric transitions, while maintaining the native fold. Key residues that mediate structural dynamics would thus be expected to be conserved or exhibit coevolutionary patterns at least. Yet, the correlation between sequence evolution and structural dynamics is yet to be established. With recent advances in efficient characterization of structural dynamics, we are now in a position to perform a systematic analysis. In the present study, a set of 34 enzymes representing various folds and functional classes is analyzed using information theory and elastic network models. Our analysis shows that the structural regions distinguished by their coevolution propensity as well as high mobility are predisposed to serve as substrate recognition sites, whereas residues acting as global hinges during collective dynamics are often supported by conserved residues. We propose a mobility scale for different types of amino acids, which tends to vary inversely with amino acid conservation. Our findings suggest the balance between physical adaptability (enabled by structure-encoded motions) and chemical specificity (conferred by correlated amino acid substitutions) underlies the selection of a relatively small set of versatile folds by proteins.

Key words: protein dynamics, sequence evolution, Gaussian network model, specificity and adaptability.

Introduction

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, or multimeric membrane proteins involved in allosteric signaling or transport, 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," for example, 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 elastic network models, 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, and allow for gate opening/closing, 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 nuclear magnetic resonance relaxation experiments (Mittermaier et al. 2003). Our examination of the collective dynamics of catalytic sites (Yang and Bahar 2005) and metal-binding sites (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; and recent studies highlight the importance of combining information on evolutionary conservation and structural dynamics in order to recognize proteins that share functional properties (Tang and Altman 2011). The prevalence of such a relationship between sequence evolution and structural dynamics 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 (supplementary table

Research article

S1, Supplementary Material online). 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 1) conserved residues have minimal fluctuations in the global modes, their high stability presumably being a prerequisite for their precise functioning, 2) increase in sequence variability is accompanied with increase in conformational mobility, this feature being most distinctive at intermediate levels of conservation/mobility typical of coevolving pairs of amino acids, 3) the coevolving residues, identified after removing effects originating from common ancestry, fall into two groups: those involved in substrate recognition and others in the neighborhood of substrate-binding sites, presumably assisting in substrate stabilization or signal transmission (the former group is distinguished by its enhanced mobility in the global modes of the enzyme), and 4) it is possible to define an intrinsic mobility scale for the 20 types of amino acids, which might be utilized for customizing protein dynamics.

Materials and Methods

Enzyme Data Set

The data set used in a previous study (Zen et al. 2008) was adopted as starting point. This data set 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. For each enzyme, the multiple sequence alignment (MSA) retrieved from the Pfam database (Finn et al. 2008) was refined using the following procedure: 1) iteratively align the primary (query) sequence from the Protein Data Bank (PDB) with each sequence in the MSA using the Smith–Waterman algorithm (Smith and Waterman 1981) and identify a "matched" sequence with the highest score, which shares at least 95% sequence identity the PDB sequence; 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 residues structurally resolved in the PDB sequence; and 3) remove the redundant sequences in the refined MSA using a threshold of 99% and eliminate the sequences that have more than 20% gaps. Supplementary table S1 (Supplementary Material online) lists the specifications of the proteins in the resulting data set.

Structural Dynamics

Gaussian network model (GNM) analysis was performed according to the well-established protocol described in our previous work (Bahar et al. 1997; Yang et al. 2006), with a cutoff distance of 7.3Å. For an enzyme of N residues, each GNM mode k is represented by an N-dimensional eigenvector, $\boldsymbol{u}^{(k)}$, and eigenvalue λ_k, describing the mode shape and frequency (squared), respectively. The ith element $[\boldsymbol{u}^{(k)}]_i$ of $\boldsymbol{u}^{(k)}$ describes the displacement of residue i along the kth mode axis. By definition, eigenvectors are normalized, that is, the plot of $[\boldsymbol{u}^{(k)}]_i^2$ as a function of residue index i also represents the normalized distribution of square displacements in mode k. The reciprocal λ_k^{-1} serves as the weight of mode k, such that the slow modes, also called soft modes, make the largest contributions to observed dynamics. The fractional contribution (or probability) of m modes to the overall dynamics is given by $w(m) = \sum_{k=1}^{m} \lambda_k^{-1} / \sum_{k=1}^{N-1} \lambda_k^{-1}$. The corresponding weighted-average mobility of residue i is defined as $<M_i>|_m = \sum_{k=1}^{m} \lambda_k^{-1} [\boldsymbol{u}^{(k)}]_i^2 // \sum_{k=1}^{m} \lambda_k^{-1}$.

The "mobility profile" driven by m modes for a given protein is obtained by plotting $<M_i>|_m$ as a function of residue index i. Minima in the mobility profile refer to sites that exhibit minimal "translational" movements in the collective motions. Note that a region may be structurally constrained (e.g., in the core of a domain) but exhibit high mobility (being embedded in a 'moving' domain), or vice versa, that is, a region that enjoys some local (e.g., rotational) flexibility may exhibit minimal mobility if it maintains its spatial position during the collective dynamics of the protein (thus acting as a hinge/anchor).

Conservation and Coevolution Patterns

The tolerance of sequence position i to mutations or amino acid substitutions is measured by the Shannon information entropy (Cover and Thomas 1991) $S(i) = -\sum_{a_i=1}^{20} P(a_i) \log P(a_i)$, where $P(a_i)$ is the probability (or fraction) of occurrence of amino acid type a at the ith column of the MSA. $S(i)$ varies in the range $0 \leq S(i) \leq \ln(20) = 3.0$, the lower and upper limits corresponding to fully conserved and fully random (equal probability of all twenty amino acid types) amino acids at the ith position. Gaps in each column are treated as uniformly distributed amino acids.

Using a similar notation, the coevolution propensity of the amino acids at the ith and jth positions along the sequence is given by the mutual information (MI), $I(i,j) = \sum_{a_i=1}^{21} \sum_{b_j=1}^{21} P(a_i, b_j) \log \frac{P(a_i,b_j)}{P(a_i)P(b_j)}$, where $P(a_i, b_j)$ designates the joint probability of observing amino acid types a and b and the respective sequence positions i and j (Liu et al. 2008; Dunn et al. 2008). Here gaps are treated as the 21st amino acid type. The MI profile for each sequence position i, $<I(i)>$, is calculated by taking the average $<I(i)> = \sum_{j=1}^{N} I(i,j)/N$, where the summation is performed over all $j \neq i$.

The inclusion of related sequences in the MSA may lead to an overestimation of conservation and coevolution propensities by including those amino acids that retain their identity due to common ancestry rather than selective functional requirements. This (phylogeny) effect is particularly important in the evaluation of MI values. For example, our analysis of HIV-1 protease sequences clearly showed two classes of correlated pairs, attributed to phylogeny and multidrug resistance, respectively (Liu, Eyal, and Bahar 2008), and it is important to filter out ancestral effects so as to assess functional correlations. Several methods have been proposed to reduce the effect of shared ancestry in evaluating coevolutionary patterns (Atchley

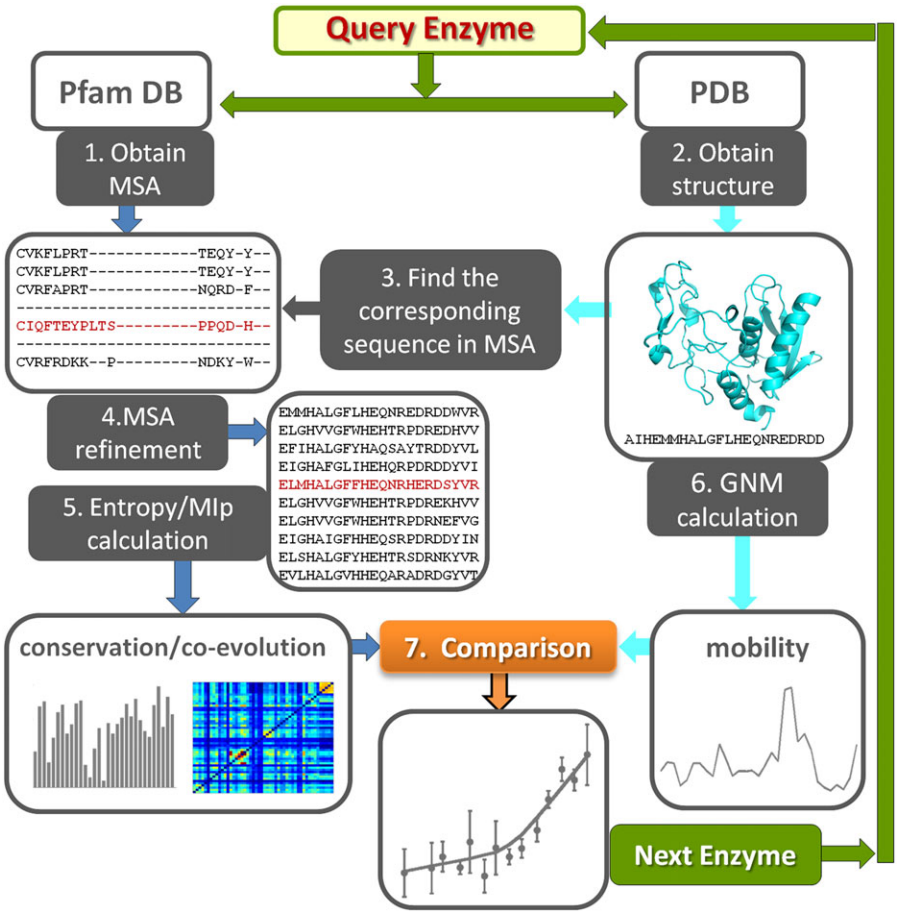

Fig. 1. Workflow of the study. For each query enzyme in the data set, we retrieve the structure from the PDB and the MSA from Pfam database. These are used as input for 1) GNM evaluation of residue mobilities (right branch) and 2) generation of conservation profile and coevolution maps (left branch), respectively. Comparison of the outputs shows 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. Statistically significant results are obtained by compiling the outputs for 34 enzymes.

et al. 2000; Tillier and Lui 2003; Dunn et al. 2008). Here we adopt the average product correction (APC) proposed by Gloor and coworkers, which proved to dramatically improve residue contact predictions based on coevolution statistics. Accordingly, the corrected MI, designated as MIp, is evaluated using the expression $MIp(i, j) = I(i, j) - APC(i, j)$, where the correction term is given by $APC(i, j) \equiv [\langle I(i)\rangle \langle I(j)\rangle]/\langle I(i, j)\rangle$. Here, $\langle I(i, j)\rangle$ is the average over all MI values evaluated for a given MSA. Subtraction of the $APC(i, j)$ lowers MI values in general, especially at residue pairs whose average MI values are generally high, allowing to distinguish the pairwise covariations devoid of the inherent variation/conservation properties of the individual residues. Although the absolute strength of the signals are weakened in general, the evaluation of $MIp(i, j)$ allows for better discrimination/visualization of the sites that undergo correlated mutations upon removal of phylogenetic effects specific to the examined protein family.

The above metrics have been normalized to introduce a uniform numerical scale for the profiles among different proteins and detect recurrent patterns. To this aim, the normalized distributions were uniformly multiplied by the number of residues so as to eliminate the dependence of the resulting mobility/conservation/coevolution profiles on the size of the proteins. Effective properties were assessed 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 evaluate the average mobility within each cluster.

Mobility, Conservation, and Coevolution Propensities as a Function of Amino Acid Types

We extracted three subsets of amino acids distinguished by their high conservation, high coevolution, and high mobility properties, shortly designated as *C*-, *E*-, and *M*-sites, respectively. The propensity of a given amino acid (of type a) to belong to the subset of *X*-sites ($X = C, E$, or M) is given by the ratio, $P_X(a) = [N_{a,X}/N_X]/N_a/N_{total}]$. Here N_a and $N_{a,X}$ denote the numbers of amino acids of type a in the complete dataset of 34 enzymes and in the subset X respectively, $N_{total} = \sum_{a=1}^{20} N_a$ and $N_X = \sum_{a=1}^{20} N_{a,X}$. More details on the evaluation of amino acid propensities is presented in the supplementary text S1 (Supplementary Material online), and protocols for evaluating mobility

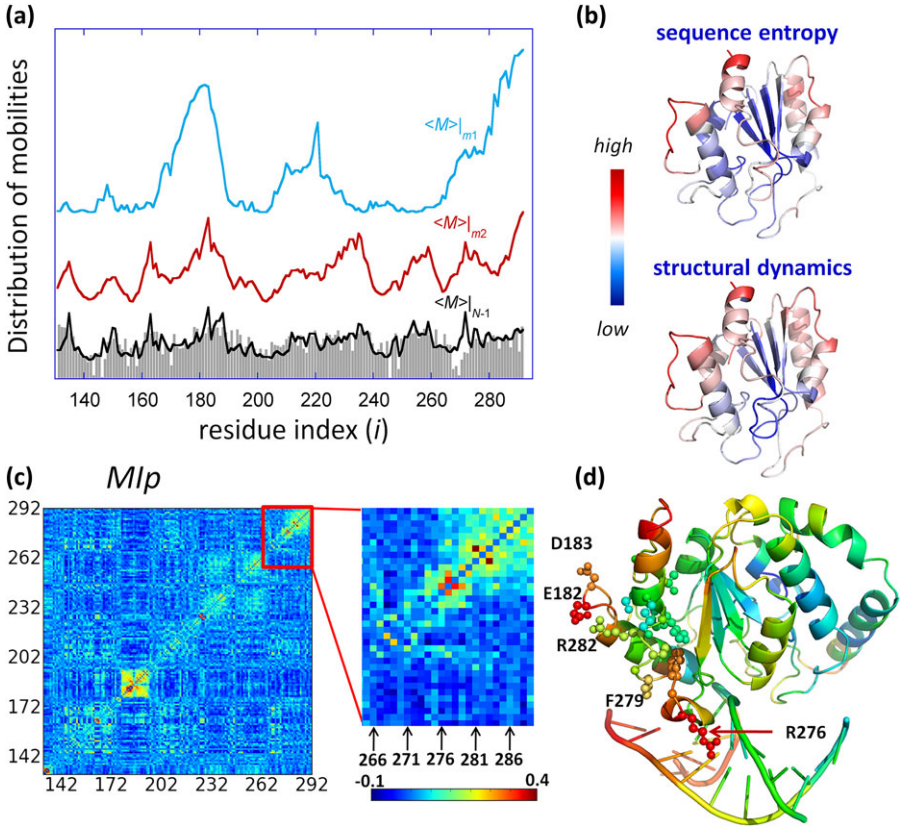

Fig. 2. An illustrative example: comparative analysis of residue conservation, conformational mobility, and coevolutionary patterns for UDG. (*a*) Mobility and conservation profiles as a function of residue index. Blue, red, and black curves represent the mobility profiles $<M_i>|_{m1}$, $<M_i>|_{m2}$, and $<M_i>|_{N-1}$ (or MSFs) computed using the GNM. The curves are shifted vertically for clarity. The bars represent the information entropy derived from 1599 Pfam sequences (supplementary table S1, Supplementary Material online). 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*) MIp map for the UDG family (see supplementary fig. S2, Supplementary Material online, for the corresponding MI map). The magnified portion refers to the DNA-binding region of UDG. Highest signals are detected at M131-I134, P163, V164, I181-F184, C281, and H283. (*d*) Location of residues distinguished by high MIp values at DNA-binding site. The diagram is color coded based on the crystallographic B factors (red/blue: most/least mobile) reported for UDG.

profiles, sequence entropy profiles, and MI maps may be found in our previous work (Liu et al. 2008, 2010).

Results and Discussion

Overview

Figure 1 illustrates the method of approach. We adopted a two-pronged analysis for each enzyme: 1) perform a GNM (Bahar et al. 1997) analysis of collective dynamics using the PDB structures and 2) analyze the residue conservation and coevolution properties using the approach described in *Materials and Methods*.

The GNM analysis yields a "mobility profile" for each enzyme. $N - 1$ GNM modes of motion contribute to the structural dynamics of an enzyme of N residues. The mobility profile based on all modes, $<M_i>|_{N-1}$, scales with the mean square fluctuations (MSFs) of residues. 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 the profiles $<M_i>|_{m1}$ and $<M_i>|_{m2}$ associated with $m1$ and $m2$ modes at the low-frequency end of the spectrum, which make fractional contributions of 0.1 and 0.4, respectively, to collective dynamics (see supplementary table S2, Supplementary Material online).

The MSAs are utilized to generate the "conservation profile" ($S(i)$ as a function of residue index i) and "coevolution maps" (both $I(i, j)$ and $MIp(i, j)$ as a function of i and j) for each protein. Comparison of mobility profiles and conservation/coevolution 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. The intermediate regime exhibits a linear increase in mobility with increasing sequence entropy. Most coevolving residues fall in this regime. The results highlight the importance of structural adaptability in sustaining the functional dynamics of the enzyme notwithstanding sequence variations that confer specificity.

An Illustrative Example

For clarity, some of the basic steps and outcomes are illustrated for a DNA repair enzyme, uracil-DNA glycosylase (UDG), in figure 2. Panel *a* displays the mobility profiles based on $m1$, $m2$, and $N - 1$ GNM modes. In UDG, $m1 = 1$, that is, the softest mode alone accounts for >10% of the dynamics (see highlighted entry in supplementary table S2, Supplementary Material online). $<M_i>|_{m1}$ shows the distribution of square displacements of residues in this softest mode. $<M_i>|_{N-1}$ 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 2*a* 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 data set. The resulting $<M_i^{\text{eff}}>|_{N-1}$ values yielded a correlation of 0.82 with sequence entropy, whereas the plot for individual residues gave a correlation of 0.52 (supplementary fig. S1, Supplementary Material online). This observation underscores the significance of consolidating the outputs with an ensemble of proteins, rather than examining single proteins where the patterns may be barely detectable.

The MIp map in figure 2*c* displays the coevolutionary properties of UDG residue pairs. Red regions indicate the pairs that exhibit the highest MIp values, that is, the loci of most correlated mutations. The upper right portion of the map magnified in panel *c* reveals the high coevolutionary properties of residues near DNA-binding site, shown in panel *d*. Supplementary figure S2 (Supplementary Material online) shows that the corresponding MI map exhibits similar features. Comparison of MI and MIp maps shows that in general signals are weakened in the MIp map (due to complete removal of contributions potentially arising from common ancestry). However, those at the DNA-binding regions highlighted in panel *d* are maintained and become even more distinctive in the background of weakened correlations.

Our previous examination of the sequence evolution properties of Hsp70 ATPase domain in relation to its intrinsic dynamics suggested that among coevolving residues those distinguished by high mobility in the global modes serve as substrate recognition sites (Liu et al. 2010). Many coevolving residues in Hsp70 have been reported to be involved in allosteric responses (Smock et al. 2010). The role of structurally labile, sequentially correlated residues in substrate recognition was also pointed out for PDZ domains by Kosik and coworkers (Sakarya et al. 2010). E182, D183, R276-F279, and C281-H283 appear to be such residues in UDG (fig. 2*a*). Notably, as evidenced by the structure shown in figure 2*d*, the residues R276-G282 do interact with DNA, and I181-F184 distinguished by their high MIp (and MI) signals are also suggested here to be interacting with DNA due to their spatial location adjacent to the former group and also their high coevolutionary propensity.

Sequence Entropy versus Conformational Mobility for All Enzymes

We repeated the comparative analysis summarized for UDG for all 34 enzymes in our data set. The results, compiled in supplementary table S3 (Supplementary Material online), confirm that mobilities/restrictions and sequence variabilities/conservations exhibit weak but statistically significant correlations; and these correlations become apparent when the effective entropies for the complete set of enzymes are consolidated based on entropy bins of $\Delta S = 0.1$ (as opposed to plotting individual values). Figure 3*a* shows the results for all the 8,254 residues in our data set. The curves show the best fit to effective mobility $<M_i^{\text{eff}}>|_{m1}$ and MSFs (or $<M_i^{\text{eff}}>|_{N-1}$). The number distribution of residues in each entropy interval is shown by the histogram (gray bars).

Several interesting features are observed in figure 3*a*. First, the coupling between structural dynamics and sequence variability is more pronounced when the global motions driven by a few soft modes ($m1 = 1$–2; supplementary table S2, Supplementary Material online) are examined, as opposed to the resultant of all $N - 1$ modes.

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 30% of residues lie in this low-mobility/high conservation 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 fig. 3*a*) 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. GNM calculations confirm that this subgroup of residues exhibit minimal fluctuations (supplementary fig. S3, Supplementary Material online). 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 (supplementary fig. S3, Supplementary Material online).

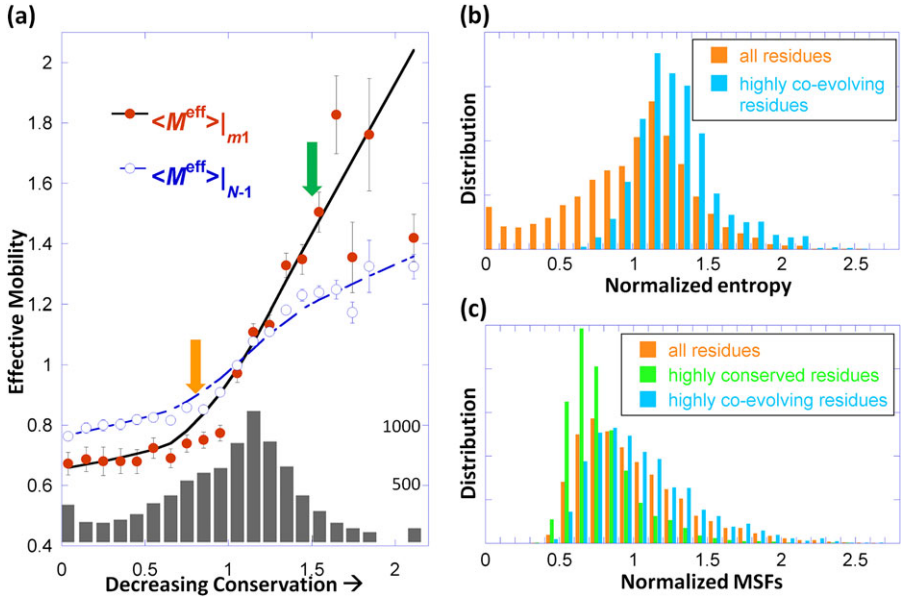

Fig. 3. Relationship between structural dynamics and sequence evolution properties. (*a*) Effective mobility as a function of sequence conservation, based on softest modes (red circles) or $N - 1$ modes (open circles) computed for all residues in the data set of 34 enzymes. The curves are the weighted least square fits to computed data, with respective correlation coefficients of 0.90 and 0.95. The number distribution of residues in different entropy intervals is shown by the gray bars (right ordinate). Entries with $S_i > 2$ are merged in the last bin. Arrows delimit distinctive mobility versus conservation regimes. (*b*) Sequence entropy distribution for all residues (orange) and a subset distinguished by their high coevolution propensities (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.

Residues Distinguished by Coevolutionary Properties Usually Exhibit Intermediate Levels of Mobility

We evaluated the MI and MIp maps for all the enzymes in our data set and identified the residues that yielded the strongest coevolution signals. Figure 3*b* and c show the respective conservation and mobility distributions (cyan bars) evaluated for the residues that yielded the top 20% $<I(i)>$ values (1,639 of them), referred to as highly coevolving residues. Panel *b* compares their sequence entropy distribution to that of the entire residue set (orange). Notably, a large majority (82%) of highly coevolving residues fall in the intermediate entropy regime identified above. And the distributions in figure 3*c* show that these residues tend to enjoy larger mobilities compared with 'all' residues. Calculations repeated with MIp values confirmed the same trend, with a slight shift of the overall distribution toward higher entropy (and higher mobility) regime, consistent with the elimination of a number of residue pairs that appear to covary due to their common ancestry (fig. S4). 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 with all residues.

Substrate Recognition Is Assisted by Coevolving Residue Pairs that Enjoy Enhanced Global Mobility

Coevolution 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, 2010; Xu et al. 2009; 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 4 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 the supplementary figures S5 and S6 and table S4 (Supplementary Material online). In all cases, the "highest" peaks in the global mode include residues distinguished by their high coevolution propensities (indicated by squares on the global mobility curves), and these residues are noted to assist in substrate recognition. Figure 4 shows that in procathepsin the residues distinguished by their strong MIp values are mainly clustered in the occluding loop N113-T125 that is involved in substrate recognition (Illy et al. 1997) and inhibitor binding (Renko et al. 2010). The high coevolution propensity of this loop is apparent even by examining the MI map (supplementary fig. S7, Supplementary Material online). Figure 4*b* shows the pronounced mobility of this loop in the global mode of motion of the enzyme.

It is worth noting that apart from the residues involved in substrate binding, those located near active sites (e.g., catalytic or signal transduction sites) may also exhibit coevolutionary 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) presents such sites (S65,

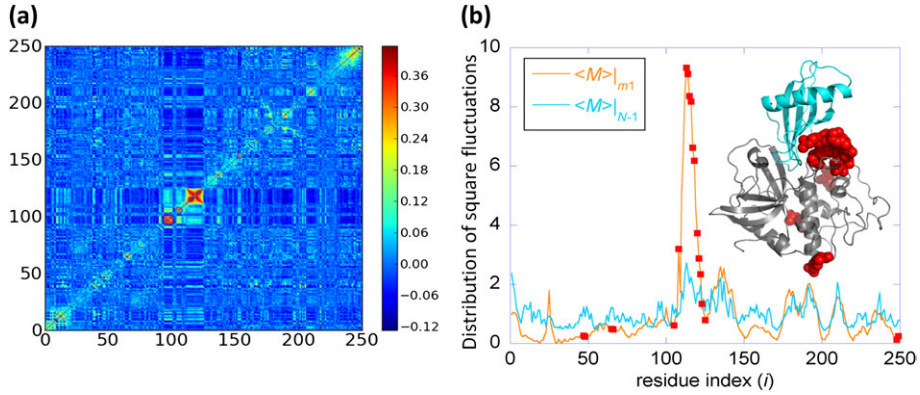

FIG. 4. Sequence coevolution and high mobility properties at the ligand recognition site of procathepsin B catalytic domain. (*a*) MI map, highlighting (in red) the coevolving amino acid pairs. Residues corresponding to the top 0.05% MIp values (N47, A48, S65, M66, I105, C108, N113-P118, T120-G123, T125, A248, and G249) are indicated by squares on the $<M_i^{\text{eff}}>|_{m1}$ curve, color. They are shown by spheres in the ribbon diagram for the complex formed with stefin A (cyan). (*b*) Global mobility profile (orange) and MSF distribution of residues (cyan) for procathepsin B. The residues distinguished in panel *a* by their coevolutionary propensities are shown by red spheres in the ribbon diagram of the protein (gray/red). Note the close neighborhood of this region to the binding site of the substrate stefin A (cyan).

C67, and G68), in close spatial proximity of substrate-binding site; the restricted mobility of these residues in the global mode suggests a signal transduction role.

Mobility Scale of Amino Acids and Contrasting Mobility and Conservation Propensities

We developed an automated procedure for identifying the sites distinguished by their high mobility in the global modes, shortly referred to as *M*-sites (see Materials and Methods, and supplementary text S1, Supplementary Material online), and evaluated the propensity P_M of different types of amino acids to take part in these sites. Figure 5 displays the resulting distribution (orange bars), obtained after normalizing the results with respect to the frequency of occurrence of different types of residues in our data set. Note that $P_M =$ if the probabilistic participation in *M*-sites is not different from that expected from a priori frequency (natural occurrence) of amino acids; $P_M > 1$ refers to amino acids that undergo relatively large displacements (based on backbone fluctuations); and $P_M < 1$, to those restricted. Calculations repeated with $m2$ and $N - 1$ modes yielded similar propensities, as shown by the respective dark orange and red bars.

Figure 5*a* 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, least conserved. The high level of conservation of histidines may be attributed to their unique multidirectional proton transfer capability, which also makes them the most common amino acid at active sites (Betts and Russell 2007). Their lowest P_M value compared with 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 5*a* 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.

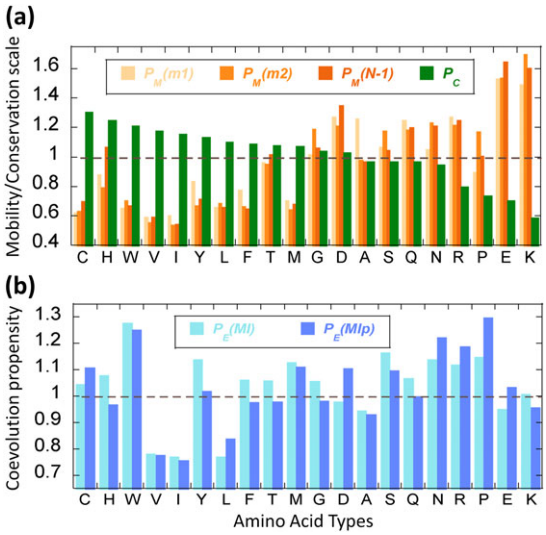

FIG. 5. Mobility, conservation, and coevolution propensities of amino acids. (*a*) Distributions of amino acids within the subsets composed of highly conserved (*C*-) (green bars) and highly mobile (*M*-) sites (light-to-dark orange bars, based on $m1$, $m2$, or $N - 1$ modes, as labeled). The bars represent the propensities with respect to those expected a priori based on the frequency of occurrence of the particular amino acid types in the data set. (*b*) Coevolution propensities of amino acids based on MI (light blue) and MIp (dark blue) values, as labeled. Amino acid types (shown by one-letter codes) are listed in the order of decreasing entropy in both panels.

Coevolution Propensity of Amino Acids

The coevolution propensities, P_E, of amino acids are presented in figure 5*b*. The propensities were based on the strongest signals observed in the MI (light blue) and MIp maps (dark blue). The two scales exhibit similar properties, suggesting that the composition of the subset of residues that yields the strongest coevolution signals (*E*-sites; see supplementary text S1, Supplementary Material online) is relatively insensitive to the choice of the metric, MI or MIp. For ease of comparison, the amino acids along the abscissa are listed in the same order as panel *a*.

Comparison of the histograms in figure 5*b* with those in panel *a* shows that the coevolutionary propensities of amino acids are practically independent of their conservation or mobility scales. Proline exhibits the highest coevolution propensity (based on MIp values). Trp is also distinguished by its high 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 coevolutionary tendencies are Cys and Tyr, which, similarly to Trp, are usually conserved (see panel *a*) and presumably involved in specific interactions 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 coevolvability and high mobility. Ser, Asn, and Arg (a charged but versatile residue that has both hydrophobic and polar moieties) lie in this group. Their combined coevolution propensity and conformational mobility suggests that these particular amino acids are suitably recruited by proteins at substrate recognition sites being at the same time specific and flexible enough to mediate substrate selectivity.

Determinants of Sequence Conservation

The present study shows that amino acids constrained in the collective motions (especially in the global modes) of enzymes tend to be conserved. However, sequence conservation may be attributed to various sources. The observed correlation may not exclusively arise from dynamic requirements but could be a manifestation of other functional or stability requirements, which in turn impose constraints on the dynamics, and on sequence identity. For example, catalytic residues are usually conserved, and so are metal-binding residues, due to their unique biochemical activities that are achieved only under well-defined coordination geometries, hence the need to stabilize specific poses/orientations of side chains at substrate/ligand-binding site. Our earlier studies demonstrated that these regions exhibit minimal motions in the global modes, consistent with these requirements (Yang and Bahar 2005; Dutta and Bahar 2010).

In the same way, one might think that the most mobile regions are usually solvent exposed, and those buried, forming the hydrophobic core, tend to be highly stable/immobile (and conserved). GNM mobilities are by definition (via the Kirchhoff connectivity matrix that describes the native contact topology) dependent on local packing density (Bahar et al. 1997; Haliloglu et al. 1997). Comparison of the relative area of solvent accessibility (RASA) (Fraczkiewicz and Braun 1998) and GNM mobilities for all amino acids in our data set yielded, for example, a correlation coefficient of 0.54. Not surprisingly RASA values also yielded a correlation of 0.43 (see supplementary table S5, Supplementary Material online) with conservation levels (Shannon entropies).

It may therefore be hard, if not impossible, to ascribe the observed conservation behavior to a "single" determinant, such as structural constraints (core, secondary structure or underlying local/specific interactions), mechanical/dynamic role (e.g., a global hinge), specific functional role (e.g., catalysis, ligand coordination), especially when interactions, structure, and dynamics are themselves interdependent. However, the present analysis permits us to make a first assessment of the potential relevance of observed correlations to collective dynamics upon examination mobility profiles.

As an example, let us consider the procathepsin B catalytic domain. As shown in figure 4*a*, the occluding loop 113–117 is distinguished by its remarkably high coevolution propensity. This loop is highly exposed (with RASA values of 94.1%, 65.9%, 90.8%, 51.6%, and 46.8% for the respective residues). Its high mobility in the global dynamics of the enzyme is manifested by the peak in the $<M>|_{m1}$ profile (fig. 4*b*). On the other hand, other solvent-exposed residues (e.g., 134–137, shown by the green spheres in supplementary fig. S8, panel *b*, Supplementary Material online) have even higher RASA values (90%, 100%, 25.8%, and 100%) but exhibit significantly lower mobility. Interestingly, the former group modulates substrate binding and the latter does not. The dynamic character of the former group presumably confers optimal substrate-binding ability. Not surprisingly, the same residues 113–117 are distinguished as coevolving residues, residues 134–137 are not. This example illustrates a case where global dynamics, rather than solvent exposure, correlates with evolutionary behavior.

Toward a more systematic examination of the origin of observed correlations, we focused on the *C*-sites (subset of most conserved 1,700 residues, ~20% of the complete set of 8,254 residues, with $S(i) < 0.65$; see fig. 3*a*), and we examined their solvent exposure and global mobility. To this aim, we first extracted two subsets of residues of the same size: 1) dynamically restrained residues usually serving as hinge centers in the global modes, designated as *H*-sites, and 2) those exhibiting the lowest solvent exposure, that is, buried residues or *B*-sites. The members of these respective subsets are found by rank ordering the normalized mobilities and normalized RASA values, starting from lowest values and simply selecting the top-ranking 20% in each list. Next we examined the overlap between the three subsets of conserved (*C*), dynamically restrained (*H*), and buried (*B*)

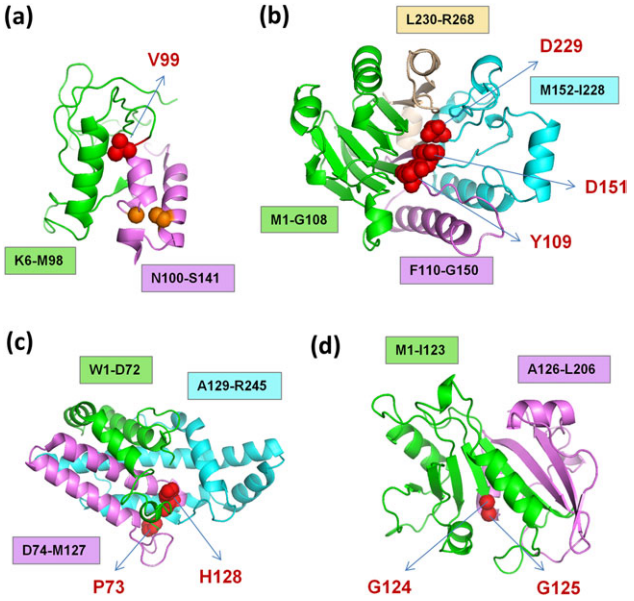

FIG. 6. Conserved sites distinguished by minimal fluctuations in global modes, despite moderate-to-high exposure to solvent. The figure illustrates four cases: (*a*) Staphyloccocal nuclease (PDB id: 1kab), (*b*) exonuclease III (PDB id: 1ako), (*c*) phospholipase C (PDB id: 2ffz), and (*d*) dehydrofolate reductase (PDB id: 3cd2). The labeled residues displayed in red space-filling representations simultaneously belong to the *C*- and *H*-subsets (of highly conserved and dynamically restrained residues) but not to the *B*-subset (of most buried residues). The identities of these residues and substructures whose collective dynamics they delimit are indicated by the labels (color coded after the substructures). The orange, space-filling residues in panel *a* illustrate a pair of residues that are highly conserved and buried (but globally moving as part of the violet substructure).

residues. The *C*-subset contained 706 *B*-sites and 531 *H*-sites (258 of which shared with the *B*-sites), and 721 'other' residues. Because of the insensitivity of effective mobilities to sequence entropies in the regime $S(i) < 0.8$ (fig. 3*a*), we did not expect a strong overlap between the *C*- and *H*-subsets. Yet, we still detect enrichments of 2.1 and 1.6, respectively, for the *B*- and *H*-sites in the *C*-subset, compared with random expectations. This statistical analysis thus shows that the majority of conserved residues are highly constrained, either structurally (*B*-sites) or dynamically (*H*-sites), and the two subsets of *B*- and *H*-residues are also interdependent.

Of interest were the 273 *H*-residues whose evolutionary conservation cannot be ascribed to their extent of burial but to their potentially hinge-bending role in the collective dynamics encoded by the architecture. Figure 6 illustrates a few such cases in four enzymes. The highlighted residues therein serve as 'anchors' or hinge centers for modulating the concerted motion of substructures (indicated by different colors; see caption). We note that the highly conserved *H*-sites tend to be located at the loop regions or at the termini of secondary structural elements, in contrast to buried residues that usually belong to secondary structural elements that make tertiary contacts (e.g., G107 and A132, orange space-filling representation, in panel *a*).

Conclusion

Several recent studies have highlighted the significance of collective dynamics in achieving biological functions or enabling biochemical activities. Yet, in previous studies, emphasis has been usually on the evolutionary pressure originating from "structure stabilization" requirements. For example, a designable protein has been viewed as one that can sustain many substitutions while maintaining its structure (Li et al. 1996; Leelananda et al. 2011). 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). The present systematic analysis, driven by the need to unravel the correlation, if any, between sequence conservation and intrinsic dynamics shows that regions severely constrained in global modes also tend to retain/conserve their amino acid identity; conversely, the most mobile regions are subject to the largest sequence variations.

These observations raise questions with regard to the origin, or causality, of the observed correlations. It is not clear whether sequence variations (that are not necessarily functional) are allowed because of the intrinsic mobility of the structure at those particular regions, or whether, on the contrary, sequence variations (or coevolution) arise from adaptability requirement at certain regions (e.g., ubiquitous recognition sites), which, in turn, selectively stabilize the particular folds that confer suitable flexibility at those functional sites. A large number of sequence variations at highly flexible regions are presumably neutral. However, some are accompanied by compensating mutations; and those coevolving pairs (or even clusters) of amino acids at regions distinguished by their uniquely high mobilities in the global modes are noted here to be involved in substrate recognition, suggesting that their behavior is driven by functional requirements. Evidently, not all observed sequence correlations are the result of functional interactions. Some may simply originate from shared ancestry. To eliminate such effects in the detection of amino acid coevolutions, we adopted the MIp method introduced by Dunn et al. (2008) and widely used in the literature (see, e.g., Buslje et al. [2009] or Dutheil [2012] for a recent extensive study).

The predisposition of highly mobile and coevolving residues to serve as substrate recognition sites, also noted 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 achieved by residues conserved within families or subfamilies. However, prior to binding, the first step is recognition, and mobility/coevolution of the recognition sites appears to be a design principle to accommodate the geometry and chemistry of the substrate (Lovell and Robertson 2010; Mittag et al. 2010). Our analysis reveals which amino acids have high coevolution propensities along with enhanced mobilities to

satisfactorily fulfill these requirements. Arg, Met, and polar residues are distinguished in this respect as versatile mediators of interactions with specific substrates. We also noted that there is another, somewhat less prominent, group of coevolving amino acids, which appears to be assisting conserved residues in either binding the substrate or coordinating cooperative responses, and this group has, in contrast to the former group, relatively suppressed mobilities in the global modes.

The reaction at the active site of an enzyme usually requires high precision; 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). The observed conservation may thus be determined by functional (e.g., catalytic) or structural (e.g., stability) requirements, rather than structural dynamics. However, systematic examination of the structural and dynamic properties of conserved residues shows that it is possible to identify sites whose conservation appears to be exclusively associated with dynamic (e.g., global hinge) role, among other conserved sites.

The evolutionary versus 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). Preorganization of conserved residues with restricted mobility has been suggested to help in stabilizing the bound conformer with minimal entropic penalty (Yogurtcu et al. 2008), whereas 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.

Finally, as suggested in a recent study (Poelwijk et al. 2007), approaches that explicitly incorporate structure, function, and fitness are likely to bring a new perspective to molecular evolution research, beyond the insights gained from comparative analyses of sequence variations. The present study is another step toward that direction, which takes advantage of the advances made in recent years in structure-based assessment of collective dynamics and accessible paths of structural changes, using coarse-grained network models for proteins' architectures.

Supplementary Material

Supplementary tables S1–S5, figures S1–S8, and text S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Support from National Institutes of Health (5R01LM007994-07) is gratefully acknowledged by I.B.

References

Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol.* 17:164–178.

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

Bahar I, Lezon TR, Yang LW, Eyal E. 2010. Global dynamics of proteins: bridging between structure and function. *Annu Rev Biophys.* 39:23–42.

Bakan A, 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 USA.* 106:14349–14354.

Betts MJ, Russell RB. 2007. Amino-acid properties and consequences of substitutions. In: Barnes MR, editor. Bioinformatics for geneticists: a bioinformatics primer for the analysis of genetic data. Chichester (UK): John Wiley & Sons Ltd. p. 311–342.

Bhabha G, Lee J, Ekiert DC, Gam J, Wilson IA, Dyson HJ, Benkovic SJ, Wright PE. 2011. A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science* 332:234–238.

Buslje CM, Santos J, Delfino JM, Nielsen M. 2009. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics* 25:1125–1131.

Cover T, Thomas J. 1991. Elements of information theory. New York: Wiley-Interscience.

Crowley PB, Golovin A. 2005. Cation-pi interactions in protein-protein interfaces. *Proteins* 59:231–239.

Dobbins SE, Lesk VI, Sternberg MJ. 2008. Insights into protein flexibility: the relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci USA.* 105:10390–10395.

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

Dutheil JY. 2012. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief Bioinform.* 13:228–243.

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

Finn RD, Tate J, Mistry J, et al. (11 co-authors). 2008. The Pfam protein families database. *Nucleic Acids Res.* 36:D281–D288.

Fraczkiewicz R, Braun W. 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J Comput Chem.* 19:319–333.

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.

Haliloglu T, Bahar I, Erman B. 1997. Gaussian dynamics of folded proteins. *Phys Rev Lett.* 79:3090–3093.
Illy C, Quraishi O, Wang J, Purisima E, Vernet T, Mort JS. 1997. Role of the occluding loop in cathepsin B activity. *J Biol Chem.* 272:1197–1202.
James LC, Roversi P, Tawfik DS. 2003. Antibody multispecificity mediated by conformational diversity. *Science* 299:1362–1367.
Jones S, Thornton JM. 1997. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol.* 272:121–132.
Leelananda SP, Towfic F, Jernigan RL, Kloczkowski A. 2011. Exploration of the relationship between topology and designability of conformations. *J Chem Phys.* 134:235101.
Li H, Helling R, Tang C, Wingreen N. 1996. Emergence of preferred structures in a simple model of protein folding. *Science* 273:666–669.
Liu Y, Eyal E, Bahar I. 2008. Analysis of correlated mutations in HIV-1 protease using spectral clustering. *Bioinformatics* 24:1243–1250.
Liu Y, Gierasch LM, 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.
Lovell SC, Robertson DL. 2010. An integrated view of molecular coevolution in protein-protein interactions. *Mol Biol Evol.* 27:2567–2575.
Luque I, Freire E. 2000. Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins Suppl.* 4:63–71.
Mintseris J, Weng Z. 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci USA.* 102:10930–10935.
Mirny LA, Shakhnovich EI. 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 LE, Forman-Kay JD. 2010. Protein dynamics and conformational disorder in molecular recognition. *J Mol Recogn.* 23:105–116.
Mittermaier A, Davidson AR, Kay LE. 2003. Correlation between 2H NMR side-chain order parameters and sequence conservation in globular proteins. *J Am Chem Soc.* 125:9004–9005.
Podobnik M, Kuhelj R, Turk V, 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.
Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ. 2007. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* 445:383–386.
Renko M, Pozgan U, Majera D, 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.
Sacquin-Mora S, Lavery R. 2006. Investigating the local flexibility of functional residues in hemoproteins. *Biophys J.* 90:2706–2717.
Sakarya O, Conaco C, Egecioglu O, Solla SA, Oakley TH, Kosik KS. 2010. Evolutionary expansion and specialization of the PDZ domains. *Mol Biol Evol.* 27:1058–1069.
Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol.* 147:195–197.
Smock RG, Rivoire O, Russ WP, Swain JF, Leibler S, Ranganathan R, Gierasch LM. 2010. An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol Syst Biol.* 6:414.
Tama F, Brooks CL. 2006. Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu Rev Biophys Biomol Struct.* 35:115–133.
Tang GW, Altman RB. 2011. Remote thioredoxin recognition using evolutionary conservation and structural dynamics. *Structure* 19:461–470.
Tillier ER, Lui TW. 2003. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 19:750–755.
Worth CL, Gong S, Blundell TL. 2009. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol.* 10:709–720.
Wright PE, Dyson HJ. 2009. Linking folding and binding. *Curr Opin Struct Biol.* 19:31–38.
Xu F, Du P, Shen H, Hu H, Wu Q, Xie J, Yu L. 2009. Correlated mutation analysis on the catalytic domains of serine/threonine protein kinases. *PLoS One* 4:e5913.
Yang LW, Bahar I. 2005. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure* 13:893–904.
Yang LW, Rader AJ, Liu X, Jursa CJ, Chen SC, Karimi HA, Bahar I. 2006. oGNM: online computation of structural dynamics using the Gaussian network model. *Nucleic Acids Res.* 34:W24–W31.
Yang Z, Majek P, Bahar I. 2009. Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. *PLoS Comput. Biol.* 5:e1000360.
Yogurtcu ON, Erdemli SB, Nussinov R, Turkay M, 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 AM, Micheletti C. 2008. Correspondences between low-energy modes in enzymes: dynamics-based alignment of enzymatic functional families. *Protein Sci.* 17:918–929.