Coarse-Grained Description of Protein Internal Dynamics: An Optimal Strategy for Decomposing Proteins in Rigid Subunits

R. Potestio,[†] F. Pontiggia,[†] and C. Micheletti[†‡*]

[†]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy; and [‡]Consiglio Nazionale delle Ricerche-Istituto Nazionale di Fisica della Materia, Democritos Center and Italian Institute of Technology, SISSA Unit, Trieste, Italy

ABSTRACT The possibility of accurately describing the internal dynamics of proteins, in terms of movements of a few approximately-rigid subparts, is an appealing biophysical problem with important implications for the analysis and interpretation of data from experiments or numerical simulations. The problem is tackled here by means of a novel variational approach that exploits information about equilibrium fluctuations of interresidues distances, provided, e.g., by atomistic molecular dynamics simulations or coarse-grained models. No contiguity in primary sequence or in space is enforced a priori for amino acids grouped in the same rigid unit. The identification of the rigid protein moduli, or dynamical domains, provides valuable insight into functionally oriented aspects of protein internal dynamics. To illustrate this point, we first discuss the decomposition of adenylate kinase and HIV-1 protease and then extend the investigation to several representatives of the hydrolase enzymatic class. The known catalytic site of these enzymes is found to be preferentially located close to the boundary separating the two primary dynamical subdomains.

INTRODUCTION

The biological functionality of many proteins depends on their capability to sustain large-scale conformational changes (1–12). Recent experimental advancements have provided novel insight into the complex relationship among protein structure, elasticity, and functionality, and have indicated that several enzymes possess innate modes of structural fluctuation that are functionally oriented and encoded in the overall structural architecture of the protein (13). By this it is meant that the modes are limitedly affected by the chemical or structural differences across wild-type conformers or mutants. The observed robustness of the modes of structural fluctuations arguably reflects the collective, large-scale character of the lowest-energy modes of fluctuation of proteins (3).

The collective character of these modes is here taken as the motivation to seek an optimal description of a protein internal dynamics, in terms of the relative motion of a preassigned number of approximately-rigid units.

The problem of decomposing proteins into groups of amino acids that have definite correlations in their equilibrium fluctuations has been previously addressed by a number of studies (14–19). In the approach of Hinsen and co-workers (14,15), for instance, large quasirigid blocks are built from small clusters of amino acids whose rigid-body motion is deduced from one low-energy mode of fluctuation. Clusters with similar rigid-body motions are next lumped together irrespective of their spatial separation. A related clustering strategy is employed by the DynDom web-server to decompose single-chain proteins based on the deformation vector bridging two given conformations (16). Other dynamics-based groupings of amino acids have been performed based on positive correlations of amino-acid displacements entailed by a single low-energy mode (17) or from pairwise correlation patterns in the covariance matrix itself (18).

A further interesting approach is offered by the translation-libration-screwlike (TLS) motion analysis introduced by Schomaker and Trueblood (20). In this scheme, which is used for crystallographic data refinement, the presence of approximately-rigid protein subunits may be inferred from the spatial modulation of the B-factors of the various amino acids (21). Current TLS implementations employ stochastic optimization techniques to subdivide proteins into a given number of nearly rigid blocks that are uninterrupted along the primary sequence (22).

In this study we introduce a variational scheme for the identification of nearly-rigid protein subparts, i.e., groups of amino acids that experience limited fluctuation of their pairwise distances. The method requires as input the essential dynamical spaces (23), i.e., the collective degrees of freedom that mostly account for a protein's equilibrium fluctuations (2), and the target number of domains to be identified. Based on this information, the amino acids are grouped so to minimize a phenomenological strain-energy function which ensures that the internal fluctuations of the groups are as small as possible compared to those across groups. The rigidlike character of the groups, hereafter also termed dynamical domains, is identified directly from a variational principle with no prior assumptions on the proximity in sequence or space of the grouped amino acids, nor on considerations of sign and magnitude of entries of the covariance matrix (in rigid bodies that are pivoted, the motion of specific pairs of points can even be negatively correlated). Although the number of desired subdivisions is left as input to the user, the fraction of overall internal mobility captured by the decomposition can be used to identify the appropriate number

Submitted December 9, 2008, and accepted for publication March 25, 2009.

*Correspondence: michelet@sissa.it

Editor: Nathan Andrew Baker.

0006-3495/09/06/4993/10 $2.00

doi: 10.1016/j.bpj.2009.03.051

of rigid subunits to attain an acceptable compromise between structural coarse-graining and dynamical accuracy.

We first apply the method to specific proteins of high biological interest, the internal dynamics of which had been previously studied in connection to their functionality, namely *Escherichia coli* adenylate kinase and HIV-1 protease. The method is next used, in conjunction with elastic network approaches, as a tool for investigating the relation between the position of the boundary separating the two primary dynamical domains and the location of the known catalytic site.

The fact that a substantial fraction of proteins equilibrium fluctuations can be accounted for by a very limited number of degrees of freedom (the roto-translational amplitudes for each rigid block) suggests that the dynamical-domains decomposition could be profitably used in various applicative contexts, a number of which are presented in the conclusions.

METHODS

Covariance matrix and essential dynamical spaces

A natural formulation and motivation of the method discussed here is provided in terms of the few collective coordinates that best account for the structural fluctuations in a given protein (2). These generalized coordinates, also termed essential dynamical spaces or low-energy modes, are aptly identified as the eigenvectors associated to the largest eigenvalues of the covariance matrix, C,

$$C_{ij,\mu\nu} = \left\langle \delta r_i^{\mu}\, \delta r_j^{\nu} \right\rangle, \tag{1}$$

where the brackets denote the canonical ensemble average over structural configurations and δr_i^{μ} indicates the μ^{th} Cartesian component of the vector displacement of the ith C_α from the average reference position. The covariance matrix is obtained either from atomistic molecular dynamics simulations or from exactly-solvable elastic network approaches (3,14,24,25). In the latter case, the canonical weight of a protein configuration is controlled by a model potential energy, F, that is quadratic in terms of the displacements of the amino acids from a preassigned reference structure, $F = \sum_{ij,\mu\nu} \delta r_i^{\mu} M_{ij}^{\mu\nu} \delta r_j^{\nu}$. The quadratic nature of F implies that the eigenvectors of C associated to the largest eigenvalues correspond to the lowest-energy modes of fluctuation of the system.

The elastic network model employed here is based on the β-Gaussian model of Micheletti et al. (25), where each amino acid is represented by two centroids (for the backbone and side chain). To avoid artifacts in the dynamical domain decomposition arising from overestimation of the mobility of exposed loops/termini, the range of the pairwise interaction of centroids at distance x is prolonged beyond the default value of 7.5 Å and weighted by an exponentially decreasing term, $\exp[-(x - 7.5\ \text{Å})/2\ \text{Å}]$.

Rigid-block decomposition

The collective character of low-energy modes in proteins suggests that the internal dynamics of these biomolecules might be described in terms of the relative motion of a limited number of approximately rigid subunits. This picture holds if the distance fluctuation of amino-acid pairs within the blocks is appreciably smaller than for pairs in different blocks. In the following we shall discuss a general variational framework, apt for numerical implementation, for performing an optimal decomposition of a protein into a preassigned number of approximately-rigid groups.

A convenient criterion of optimality is given by the maximization, over the possible groupings, of the system mean-square fluctuation captured by the lowest-energy modes that is compatible with a nearly-rigid character of each group of amino acids.

For definiteness we discuss a subdivision of a protein's amino acids into Q putatively-rigid groups. We shall further indicate with $\vec{v}_l$, the l^{th} (nonzero) lowest-energy mode of the system $C\,\vec{v}_l = \lambda_l\,\vec{v}_l$, with λ_l being the associated eigenvalue. We consider the following decomposition of the mode

$$\vec{v}_l = \vec{v}_l^{\,rb} + \Delta\vec{v}_l, \tag{2}$$

where $\vec{v}_l^{\,rb}$ is the best rigid-body fit of the mode and $\Delta\vec{v}_l$ is the correction that accounts for the internal distortions within the putatively rigid blocks. Notice that the spatial contiguity of amino acids belonging to the same group is not enforced a priori. The first term of the right-hand side is entirely specified by $6Q$ parameters that correspond to the roto-translations of each rigid-body block. The modes are assumed to describe perturbative fluctuations of the reference structure (it should be borne in mind that this condition may be poorly met by highly mobile regions such as exposed loops or termini). The rigid-body motion of a group of residues consequently admits a linear parameterization in terms of the roto-translational degrees of freedom. In fact, the l^{th} modal displacement of the ith amino acid belonging, say, to the q^{th} group, is given by

$$\vec{v}_l^{\,rb}(i) = \vec{t}_l(q) + \vec{\omega}_l(q) \times \left(\vec{r}_i - \vec{r}_q^{\,cm}\right), \tag{3}$$

where $\vec{r}_q^{\,cm}$ denotes the center of mass of the q^{th} group and $\vec{r}_i$ is the position of the ith C_α. Notice that the same translation and rotation parameters ($\vec{t}_l$ and $\vec{\omega}_l$, respectively) are used for each residue in a given group. The optimal rigid-body approximation to $\vec{v}_l$ is obtained by maximizing the norm of $\vec{v}_l^{\,rb}$ over the $6Q$-dimensional parameter space. The accuracy of the fit is readily obtained by computing the fraction of the norm of the essential dynamical space captured by the rigid-body approximation:

$$\frac{\left|\vec{v}_l^{\,rb}\right|^2}{\left|\vec{v}_l\right|^2} = \left|\vec{v}_l^{\,rb}\right|^2 = 1 - \left|\Delta\vec{v}_l\right|^2. \tag{4}$$

Considering the space of the top n essential modes, the above expression generalizes to

$$f = \frac{\sum_{l=1}^{n} \lambda_l \left|\vec{v}_l^{\,rb}\right|^2}{\sum_{l=1}^{n} \lambda_l}. \tag{5}$$

Unless otherwise stated, considerations will be limited, as customary, to the $n = 10$ lowest-energy modes, which are usually sufficient to capture most of the fluctuations in a given system (alternatively n could be chosen so to capture a preassigned fraction of the overall internal fluctuations). The goal of the procedure is to identify, within the possible partitioning of amino acids into Q dynamical domains, the one yielding the largest possible value of f.

Elastic deformation energy

Because the volume of configuration space (number of distinct amino acids groupings) is large, the optimization of f in Eq. 5 requires the stochastic exploration of tens of thousands of possible amino-acid groupings. Even if optimized algorithms exist for the calculation of the rigid fits (26), the number of repeated matrix operations involved is so large that it is not computationally convenient to extremize directly the quantity f. A more effective strategy is to perform a preliminary exploration of configuration space by optimizing a simple objective function to efficiently identify candidate subdivisions over which f is finally evaluated and maximized.

The objective function that we consider is

$$F(\{\sigma\}) = \frac{1}{2}\sum_{i\neq j}\delta_{\sigma_i,\sigma_j}\sum_{l=1}^{n}\lambda_l\left[(\vec{v}_l(i)-\vec{v}_l(j))\cdot\vec{d}^{\,0}_{ij}\right]^2 + \frac{\alpha}{2}\sum_{i\neq j}(1-\delta_{\sigma_i,\sigma_j})\frac{1+\tanh\left(R_c-|\vec{d}^{\,0}_{ij}|\right)}{2}, \quad (6)$$

where $\sigma_i = 1 \cdots Q$ denotes the group to which amino acid i belongs, δ is the Krönecker delta, $\vec{d}^{\,0}_{ij}$ is the distance vector of amino acids i and j in the reference conformation, R_c is an interaction cutoff distance set equal to 7 Å, and $n = 10$. The sought optimal grouping of amino acids is the one that minimizes F, similarly to the spirit of graph clustering methods (27). For systems consisting of truly-rigid subparts, this will be analogous to maximizing f. The first term in the sum represents the cost of the average elastic energy associated to the internal deformation of the molecule. This term penalizes fluctuations in the distance of any two points belonging to the same putatively-rigid group, consistently with the definition of rigid bodies. The second term introduces a penalty, controlled by the parameter $\alpha \geq 0$, for dynamical domains consisting of regions that are disconnected in space. Upon increasing α, in fact, the term disfavors the number of pairs of neighboring amino acids (those closer than the cutoff distance $R_c = 7$ Å) that belong to different groups. The optimization of F, therefore, leads to group assignments that minimize the interface area between the groups, while not strictly enforcing the spatial compactness of the domains. The minimization of F is straightforwardly done within a simulated annealing protocol, with elementary moves corresponding to changes of the group assignment of individual amino acids. The corresponding changes of F only require the summation of N precalculated quantities, N being the number of amino acids in the protein. Proteins of 200 residues can thus be subdivided into, e.g., 10 dynamical domains in ~1 min, on present-day personal computers.

The search for the optimal solution is carried out separately for increasing values of α. Notice that for sufficiently large α, the minimization of F eventually leads to solutions having fewer groups than Q, which are hence not considered (this is intuitively expected, as in the limit $\alpha \to \infty$ the presence of boundaries is forbidden, and a single dynamical domain is returned by the minimization of F). The solution corresponding to the largest value of f is taken as providing the best subdivision.

RESULTS

We first discuss the application of the rigid block decomposition to *E. coli* adenylate kinase (AKE) and HIV-1 protease (HIV-1 PR), two enzymes whose internal functionally oriented dynamics has been extensively studied; see, e.g., the literature (5,8,9,25,28–35) and references therein. The rigid-block decomposition is performed based on data from atomistic molecular dynamics simulations (for AKE) and from elastic network models (for HIV-1 PR). The method is next applied for a rigid-subunit decomposition of two sets of proteins. The first set consists of monomeric enzymes representing the main CATH structural classes (36) of hydrolases (class 3 according to enzyme classification, i.e., EC (37)). For these enzymes we investigate the existence of systematic biases in the location of the known catalytic site, with respect to the boundaries separating primary dynamical subdomains. We conclude the analysis by investigating the extent to which the optimal subdivision returns groups of residues that span uninterrupted stretches of the primary sequence or occupy compact regions in space. This analysis will be performed on a set of 90 protein monomers that, according to the CATH classification, consist of three-to-six structural domains each being uninterrupted along the primary sequence.

Adenylate kinase

Adenylate kinase is a phosphotransferase regulating the relative abundance of AMP, ADP, and ATP within the cell. The enzyme is composed by a central core and two domains, the ATP binding domain (Lid) and the AMP binding one, which are highly mobile. In the available ''closed'' crystallographic state (Protein Data Bank (PDB) code: 1ake), they are displaced toward the core by >7 Å, with respect to the ''open'' crystal structure (PDB code: 4ake).

Recent experiments have indicated that the enzyme is spontaneously capable of interconverting between the closed and open forms even in the absence of ligands (8,9). This points to the predisposition of AKE's internal dynamics to bridging the open/closed conformations and to the absence of large free-energy barriers separating the two reference states, consistently with indications from atomistic molecular dynamics simulations of AKE (10,29,30).

For example, in a recent computational/theoretical study carried out by some of us (30), the dynamical evolution of the free *E. coli* adenylate kinase was followed from two starting structures for as long as 100 ns. Over this extensive time-span, which is nevertheless much smaller than the experimental interconversion time, the molecule was found to populate a fair number of structurally distinct substates. Most of the structural fluctuations within and across the substates were described by very few low-energy collective modes entailing the independent motion of the Lid and AMP-bd subdomains with respect to the core (see Figs. 6 and 9 in (30)).

The rigid-block decomposition scheme here applied to AKE lends naturally to assessing if, and to what extent, the molecule's internal dynamics can be described in terms of a few parts that move as nearly-rigid units. We begin by considering the fluctuations within the substate where the 50-ns-long trajectory started from the open structure, 4ake, dwelled for ~10 ns (30). The reference structure for the substate, which is the most populated of the MD trajectory, is provided in Fig. 1 *a*, along with the representation of the lowest energy mode. The mobility the Lid and of the AMP-binding subdomains, corresponding to regions 117–164 and 30–64, respectively, is evident.

The $n = 10$ lowest energy modes within the substate were used to subdivide the enzyme into $Q = 2, 3 \ldots 10$ dynamical domains. A representation of the subdivisions into three and four groups is provided in Fig. 1, *b* and *c*. The fraction of essential dynamics motion (see Eq. 5) captured by the various subdivisions is shown in Fig. 1 *d*.

The graph indicates that a very limited number of dynamical domains is already sufficient to account for most of the essential dynamics. In fact, subdivisions into $Q = 2$, 3, and

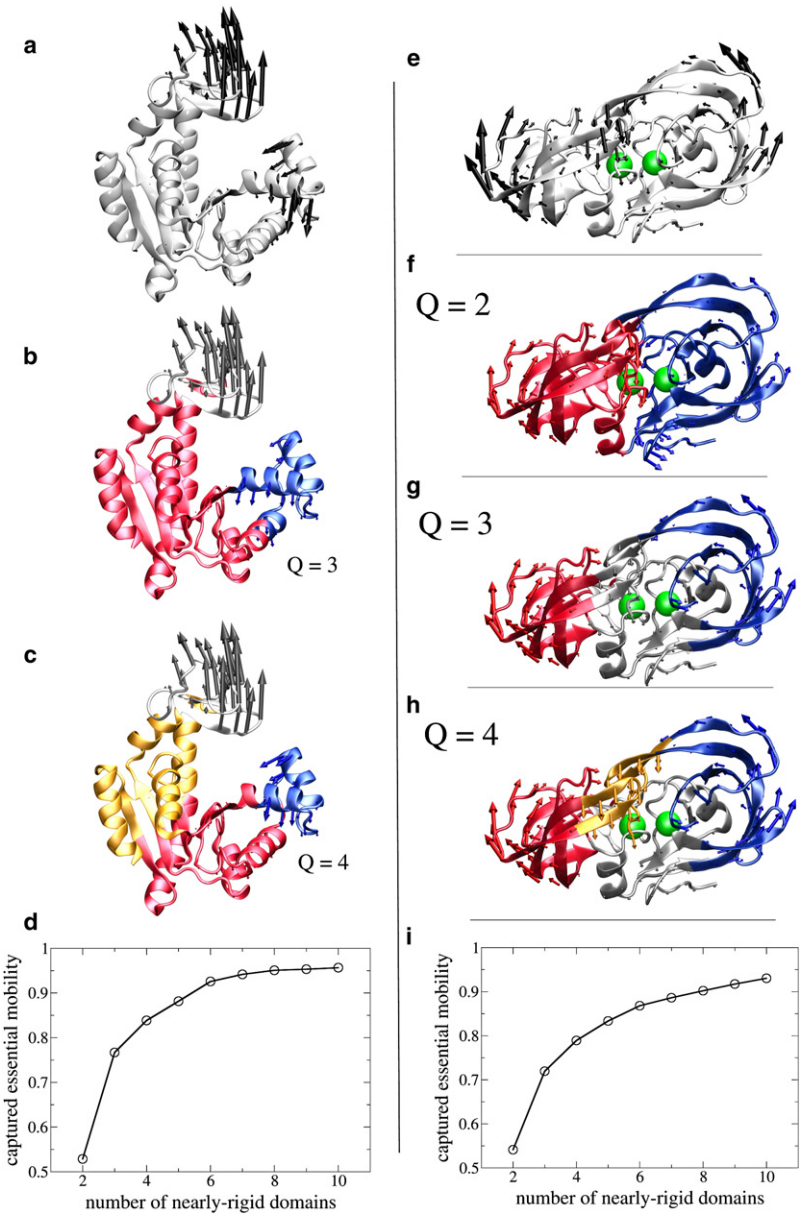

FIGURE 1 (*a*) First essential mode of *E. coli* adenylate kinase. Subdivisions of the enzyme in $Q = 3$ and $Q = 4$ rigid subunits, identified by different colors, are shown in panels *b* and *c*. The decomposition was performed taking into account the 10 lowest-energy modes. For clarity, only the rigid-body approximation to the first mode is shown. The fraction of essential dynamical motion (see Eq. 5) captured by the subdivision into $Q = 2 \ldots 10$ rigid domains is shown in panel *d*. Panels *e–i* show analogous results for HIV-1 protease. The two catalytic residues (Asp^{25} and Asp^{124}) are highlighted in green in the three-dimensional structure.

4 blocks capture as much as 52%, 77%, and 83% of the fluctuations entailed by the $n = 10$ essential modes (which account for 80% of the overall mobility).

The subdivision for $Q = 2$ identifies region 122–156 as an approximately rigid, but highly mobile, unit. The region overlaps well with the Lid indicated before. The less mobile AMP-binding domain is identified as a distinct unit when using $Q = 3$. In fact, for $Q = 3$, the regions corresponding to the two mobile nearly-rigid subdomains are 122–158 and 32–59, and are compatible with the customary tripartite subdomain division of AKE. If the entire 50-ns-long trajectory is used rather than the most populated substate, it is found that the boundary of the AMP-binding domain is virtually unaltered (sequence interval 32–60). The larger configurational space spanned by the more mobile Lid domain instead reflects into an extension of the both the left and right subdomain boundaries by ~10 residues, thus covering the interval 112–167.

In summary, for $Q = 3$, the three units cover five sequence intervals: one for each of the two mobile domains and three for the nearly-fixed core. It is interesting to compare this dynamics-based subdivision with the one provided by the TLS analysis of crystallographic data (22), which enforces the sequence continuity of each rigid block. The TLS decomposition of 4ake into five intervals (as many as those found with $Q = 3$) returns the following segments: 1–27, 78–116, and 171–214, identifiable with the core, and 28–77 and 117–170, compatible with the AMP-bd and Lid subdomains, respectively. With the exception of one of the AMP-bd/core boundaries, the TLS subdivisions and those of our analysis

of the full MD trajectory are mismatched by only approximately five residues. They are hence generally consistent, despite the differences not only in method but also for the nature of the input data (crystallographic B-factors for TLS and MD data for our method). An important distinction between the two results is, however, that the three segments constituting the core regions are encompassed in a single rigid unit by this variational method, while the others are treated as independent ones within the TLS scheme.

The optimal subdivision was compared also with the one returned by the DynDom server (16), which requires the input of two structures representing the conformational variability of the molecule of interest. Accordingly, from the set of MD-sampled conformers we selected the pair with the largest root mean-square deviation. DynDom returned a subdivision in two domains, the smallest corresponding to the Lid (sequence interval 110–169) plus a small loop (residues 6–12) and the other to the core plus the AMP-binding domain. Interestingly, this latter subdomain is recognized as a separate dynamical domain if the open and closed crystallographic conformers of AKE (1ake, 4ake) are used as input structures.

We conclude by commenting on the subdivision in $Q = 4$ dynamical units of the most populated MD substate. With respect to the $Q = 3$ case, the boundaries of the two mobile domains are only slightly adjusted to 35–60 and 118–159, respectively. However, a new domain, comprising several sequence segments 7–25, 108–117, 160–174, and 195–214, is identified at the interface between the core and the Lid. This group of hinge residues have consistently been shown, by independent methods (28,30), to be subject to a significative strain during the free enzyme dynamical evolution.

HIV-1 protease

As a further example of the dynamics-based decomposition we consider the HIV-1 protease dimer complexed with a peptide substrate. The internal dynamics of the enzyme has a distinctive collective character that has been extensively investigated in past years by means of both atomistic MD simulations as well as coarse-grained models (see (5,7,33–35) and references therein). The enzyme flaps, in fact, carry out particular large-scale movements, resulting in detectable anticorrelated displacements of the flap's tip, which contacts the bound (inhibiting) peptide and the distal region of the flaps, where several mutations causing drug resistance are located (5).

To illustrate the applicability of the method in the absence of data from atomistic simulations, we obtained the essential dynamical spaces from the βGM elastic network model of Micheletti et al. (25), modified as described in the Methods section.

The complex shown in Fig. 1 *e*, which corresponds to the equilibrium structure of the MD study of Piana et al. (5), was again subdivided from 2 to 10 domains (see Fig. 1, *f–h*). The fraction of internal dynamics captured by the various decompositions is shown in Fig. 1 *i*. The curve has a slightly slower increasing trend compared to AKE (Fig. 1 *d*). In fact, when $Q = 3$, 4 domains are used, ~72% and 79%, respectively, of the essential dynamical fluctuations is captured for the HIV-1 protease/substrate complex.

The subdivisions into $Q = 2$, 3, 4 approximately-rigid units are represented in Fig. 1, *f–h*. As for AKE, the units are compactly organized in space but do not cover a single stretch of the primary sequence. The sequence-disconnected nature of the domains does not lead simply to a detailed comparison with the TLS decomposition. We shall therefore restrict ourselves to considering the primary hinge-points, represented by amino acids 20, 35, 57, and 70, which emerge from the precalculated subdivision offered by the TLS webserver of the HIV-1 PR monomer (PDB structure 1t3r) in five-to-seven intervals. The first three hinges fall within three amino acids (along the primary sequence) from boundaries indentified for the optimal subdivision in two primary domains, $Q = 2$ (see the Supporting Material), suggestive of good consistency.

As a further comparison we considered the precalculated subdivision of the HIV-1 PR monomer offered by the DynDom server (based on structures 1aid and 1hsg). The returned subdivision consisted of two domains, the smaller one comprising segments 32–60 and 75–77, and broadly corresponding to the monomer flap. Though it should be borne in mind that the subdivision might depend on the fact that only one monomer is considered (multimers are not accepted by the DynDom server), the identified modular nature of the flaps is compatible with salient aspects of the TLS and variational decomposition.

The inspection of the optimal subdivisions in Fig. 1, *f–h*, prompts two considerations. The motion of the flaps is largely consistent with a coordinated rotatory movement around the central fulcrum regions. It is evident from the $Q = 3$, 4 cases that within the nearly-rigid parts comprised by the flaps, the points at the two extremes are displaced in opposite directions. On the one hand, this feature illustrates that the motion of nearly-rigid units in proteins can be sufficiently general to allow for the presence of anticorrelated motion within its constituents parts. On the other hand, the analysis supports the qualitative description of the flap motion first given by Piana et al. (5), based on the visual inspection of the first essential mode of a multi-nanosecond MD simulation (see Fig. 6 *b* in (5)). Indeed, if the first mode only is used for decomposing the proteins in to $Q = 2$ blocks, it is found that each entire flap is identified as a nearly-rigid unit (see the Supporting Material).

The second observation regards the location of the catalytic site of HIV-1 protease with respect to the "primary dynamical boundaries." By the latter, we mean the boundaries separating the most prominent rigidlike regions in a protein (i.e., when using $Q = 2$ or $Q = 3$). By inspecting Fig. 1, *f–h*, it is seen that the highlighted catalytic aspartic dyad, which has low mobility, straddles the rigid-domains interface for $Q = 2$ and is close to one or more domain boundaries for $Q = 3$ and 4.

This suggests that the proximity of the catalytic amino acids to the primary dynamical boundary is instrumental for accompanying the limited mobility of the aspartic dyad with a functionally-oriented modulation of the bound peptide substrate (38).

Catalytic site location

It is interesting to note that modulations of the active site analogous to HIV-1 protease have been found for several other proteolytic enzymes differing by catalytic chemistry and structural architecture (31,32,39). The existence of such common large-scale movements is taken as the starting point for examining what relationship, if any, exists between the location of the cleavage sites and the proximity of primary dynamical boundaries for other enzymes belonging to hydrolases. Is the cleavage site commonly located near primary dynamical boundary, as for HIV-1 protease?

The question appears particularly appealing also in view of recent considerations made by del Sol et al. (11) that functional sites in proteins with allosteric behavior are preferentially located at the boundary between regions that are modular in terms of contacting amino acids.

The question will be formulated within a rather comprehensive framework, where proteolytic enzymes are considered along with other enzymes members of class 3 of EC. The enzymes were taken from the list of 76 representatives of the main EC and CATH groups singled out in Zen et al. (39). The list, restricted for simplicity to monomeric enzymes (following the indication in annotated UNIPROT (40) entries), is reported in Table 1 along with the EC and CATH code and with the indication of the amino acids constituting the catalytic site.

TABLE 1 Monomeric members of the EC class 3 enzymes (hydrolases)

PDB	Length	Catalytic site
4p2p	124	H48, D99
1ako	268	N7, D151, N153, D229, H259
1vas	137	T2, R22, Q23, R26
2fmb	104	D25
1bol	222	H46, E105, H109
1k2a	136	H15, H129
1de3	150	H137, E96, H50
1kab	136	R35, R87
3eng	213	D10, D121
2f47	175	E11, D20
2ayh	214	E105, E109
4skn	223	N145
1avp	204	H54, E71, C122
1qjj	200	E93
1lqy	184	E154

Enzymes were taken from the representative list of Zen et al. (39), which covers the main CATH groups. To avoid excessive dispersion in length, only enzymes with 100–270 amino acids were considered. The amino acids constituting the catalytic site were taken from the catalytic site atlas (48) when literature evidence was available, otherwise they were obtained by intersecting the catalytic site atlas and Uniprot data.

For each entry, the boundary between the two primary dynamical domains was identified from the $Q = 2$ subdivision. To measure the separation of a catalytic residue from the primary dynamical boundary, we considered the distance of its C_α from the nearest C_α belonging to the other dynamical domain. The normalized distribution of these boundary distances is shown with a thick line in Fig. 2 along with the reference distribution (*dashed line*) of the boundary distance of every amino acid in the 15 proteins. The two distributions present appreciable differences as the catalytic residues are preferentially closer to the interface than other amino acids in the proteins.

The overall indication of catalytic-site/boundary proximity in hydrolases conveyed by Fig. 2 was complemented by a case-by-case analysis of the 15 enzymes in Table 1. This detailed investigation is necessary in view of the fact that the cumulated data in Fig. 2 reflect properties of a group of enzymes with a certain heterogeneity in length, structural architecture, and number of catalytic sites.

The $Q = 2$ subdivisions of the 15 enzymes are provided as Supporting Material and consistently reveal the good proximity of the cleavage sites with the boundaries between the dynamical domains. Here we limit the discussion to three enzymes whose dynamical role in the functional cleavage of peptides or nucleic acids has been previously considered (41–45), namely: exonuclease III (PDB 1ako); human adenovirus proteinase (PDB 1avp); and endo-1,3-1,4-β-D-glucan 4-glucanohydrolase (PDB 2ayh). Their $Q = 2$ subdivision is represented in Fig. 3, *a–c*.

Exonuclease III and adenovirus proteinase bind DNA in double- and single-stranded forms, respectively. In Zen et al. (39), a dynamics-based connection between them was established, which is particularly interesting as they are not evolutionarily related and are characterized by two different architectures, 4-Layer Sandwich (CATH: 3.60.10.10) and 3-Layer ($\alpha\beta\alpha$) Sandwich (CATH: 3.40.395.10), respectively. In both cases the catalytic residues are found to be located at the primary boundary. As visible in Fig. 3, the low-energy

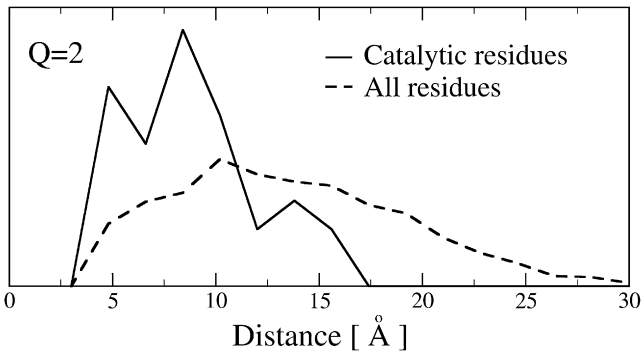

FIGURE 2 Distribution of amino acid distances from the boundary separating the two primary dynamical subdomains. The dashed line indicates the distribution of boundary distances for all 2690 amino acids in the data set of Table 1, while the thick line gives the distribution only for the 34 catalytic amino acids. Both distributions are normalized.

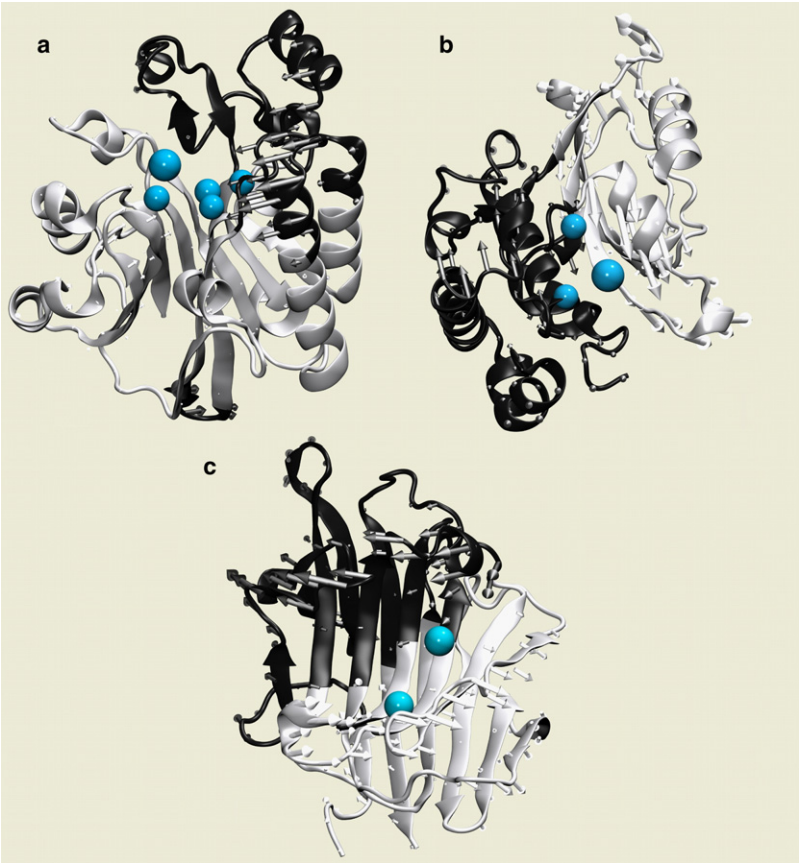

FIGURE 3 Subdivision into $Q = 2$ dynamical domains (represented in different colors) of exonuclease III (*a*), human adenovirus proteinase (*b*), and endo-1,3-1,4-β-D-glucan 4-glucanohydrolase (*c*). The decomposition was performed taking into account the 10 lowest-energy modes. For clarity, only the rigid-body approximation to the first mode is shown. Catalytic residues are shown as spheres.

modes have a common character as they entail an outward/inward concerted movement between the two blocks in the surroundings of the catalytic sites, with the latter at the center. The analysis carried out on endo-1,3-1,4-β-D-glucan 4-glucanohydrolase also shows that the two catalytic residues of the enzyme are located in proximity to the interface between the two primary dynamical domains, both surrounded by loops forming a groove that can arguably accommodate the corresponding ligand.

The consistent indication of Fig. 3 is that the catalytic site is located close to the primary interface. This fact appears particularly interesting when considering how the primary boundary is modulated by the lowest-energy modes of the domains (which are compatible with the opening/closing of the catalytic cleft (41,44,45)). By comparison to noninterfacing amino acids, it is found that interface residues cover a fairly large range of values both for overall mobility and for the degree of distortion of the local structural environment (see the Supporting Material). Interestingly, the catalytic site is accommodated at, or close to, an interface subregion having both low mobility and low-structural deformation. While these properties are consistent with the expected rigidity of the catalytic region, it is interesting that they can be realized in proximity of the primary dynamical boundary, where appreciable elastic strain can be built up due to the relative motion of the dynamical domains.

Approximately-rigid units: connectedness in sequence and space

We conclude the analysis with a systematic evaluation of the extent to which the subdivision of a protein into a limited number of approximately-rigid units results in dynamical domains that are compact in space and/or cover uninterrupted regions of the primary sequence. The interest in this question is twofold. On the one hand, it can provide indications on the viability, for computational efficiency, of enforcing a priori the proximity in sequence or space of the amino acids belonging to the same group. On the other hand, it can shed some light on the existence of consistent modular organizations of proteins at the level of sequence, structure, and dynamics.

An interesting general context where these questions can be posed is provided by multidomain proteins. We considered a data set of 90 protein monomers, with overall sequence identity $<90\%$ and constituted by three-to-six CATH domains each consisting of a single sequence interval (see the Supporting Material). Each protein was subdivided into a number of dynamical domains equal to the number of CATH domains. To analyze robust aspects of the sequence-integrity of the dynamical domains, the rigid-units subdivisions were postprocessed to eliminate domain fragments covering excessively short sequence intervals. Specifically, fragments smaller than 1/20th of the protein length (and in any case no longer

than 10 amino acids) were removed. The amino acids in these fragments are reassigned to the nearest flanking unit. The resulting dynamical domains subdivisions (along the primary sequence) were compared with the ones provided by CATH.

It was found that only for 30 proteins out of 90, did the number of domain boundaries along the sequence match. Therefore, in two-thirds of the cases, the dynamical domain subdivisions gathered regions that were disconnected along the primary sequence, at variance with structural subdivisions employed in domain identifications. Notably, for the corresponding 30 cases, the dynamical subdivisions were very well consistent with the CATH ones. In fact, out of the 91 boundaries occurring in the 30 proteins, as many as 71 occurred at a separation of <10 residues of the CATH ones. By commonly-employed criteria (46) this reflects a very strong agreement of the subdivisions. It is worth noting that also for the 60 nonmatching proteins, most of the CATH subdivisions fall within 10 residues from the dynamical ones, which are, however, more numerous.

For all the 90 proteins we checked the extent to which the non-postprocessed dynamical domains, despite possibly comprising segments that are not contiguous in sequence, occupy compact regions in space. The compactness of a domain was ascertained by measuring the diameter of the graph given by the contact map of the residues (with a contact cutoff distance of 7.5 Å). A finite value of the diameter, which measures the minimum number of graph edges that need to be traversed for connecting any two nodes in the graph, indicates the spatial compactness of the domain. It was found that <5 dynamical domains out of 308 comprised disconnected, though nearby, regions. This provides an a posteriori indication of the fact that rigidlike units comprise amino acids that occupy spatially connected regions.

Finally, despite the differences in sequence integrity of the dynamical and CATH subdivisions, we performed a test to quantify the overlap between these two decompositions. We found that the mutual one-to-one overlap of the CATH domains and rigid units was, on average, 80%, which underscores a nontrivial, albeit not perfect, consistency between the two subdivision criteria. The degree of overlap, specialized for the two principal CATH codes (class and architectures) is provided as Supporting Material.

CONCLUSIONS

In consideration of the collective nature of the lowest-energy modes of structural fluctuations in proteins, it appears natural to investigate the extent to which the internal dynamics of these biomolecules can be described in terms of a few groups of amino acids, each moving, under thermal agitation, as an approximately-rigid unit.

The method presented and applied here introduces a variational scheme, for optimally grouping proteins' amino acids into a preassigned number of approximately-rigid units. The search in configuration space (possible amino acid groupings) is guided by the maximization of the fraction of the proteins' equilibrium fluctuations captured after the suppression of the internal fluctuations within the putative rigid units. The method is computationally efficient as proteins of 200 residues can be subdivided in, e.g., 10 dynamical domains in ~1 min, on present-day personal computers. The method presents the following appealing features. First, its variational formulation allows the straightforward control of the viability of the rigid-body approximation for each unit. Secondly, no a priori assumption is made on the fact that a unit should comprise an uninterrupted stretch of the primary sequence, nor should occupy a compact (not disconnected) region of space. Finally, the subdivision is not guided by considerations on the high degree of correlation of equilibrium displacements of pairs of amino acids belonging to the same group, as this criterion is not necessarily respected for generic movements of a rigid body.

The subdivision into rigid units was performed and discussed in a number of contexts of biophysical interest. We first used the method to analyze and describe the internal dynamics of adenylate kinase using data obtained from atomistic molecular dynamics simulations. For this specific enzyme, the subdivision into as few as three units is already sufficient to account for 77% of the fluctuation entailed by the top-10 low energy modes. In addition, these rigid units and the bridging hinge regions are consistent with previous studies (28,30) and in general agreement with the partitioning obtained through uncorrelated methodologies, such as TLS crystallographic data analysis (22). This suggests that the method can possibly be used to identify primary hinge regions in proteins (e.g., through the comparison of the location of boundaries separating dynamical domains upon increasing the number of subdivisions, Q).

A further case study was constituted by the dimeric HIV-1 protease with a bound peptide substrate. The complex was subdivided using the essential dynamical spaces identified via an elastic network model. The resulting subdivisions reflected the large-scale mechanical couplings that are known to exist between the distal regions of the flaps, which are capable of modulating the proximity of the peptide with the cleavage site. Interestingly, the catalytic aspartic dyad was found to be located at the boundary separating the most prominent rigid units. This localization appears instrumental for accompanying the limited mobility of the catalytic site (which has to be preserved in the correct catalytic geometry) with specific concerted movements of the flaps, compatible with the modulation of the substrate in a β-extended configuration (38).

The computational effectiveness of the decomposition strategy, in conjunction with elastic network models, suggests the potential applicability of the method (in which general features of proteins' functionally oriented elasticity are sought, in terms of the collective motion of a few units). As an example of this avenue, and motivated by the findings of HIV-1 protease, we investigated the relationship between the relative location of primary boundaries among dynamical

domains and known catalytic sites for several monomeric representatives of the hydrolases (class 3 of the enzyme classification). Despite differences in structural organization and nature of the bound substrate, the catalytic site of these enzymes is found to be preferentially located at a particular subregion (experiencing both low mobility and small structural deformation) of the primary interface.

Consistently with HIV-1 protease, the motion of the rigid units delimiting the active region are found to be generally compatible with functionally oriented movements leading to the binding or processing of the substrate.

These applications indicate that the variational method may be profitably used to not only gain insight into the modular organization of structure and functionally oriented dynamics of individual proteins, but also as a comparative tool to highlight common dynamical features in protein families and superfamilies (39).

The method could also be used to identify order parameters apt for capturing the relative displacements and correlations of the rigid domains. The order parameters could be used in MD contexts where all the atomic degrees of freedom are retained, either to analyze a posteriori MD trajectories or to add a controlled bias for aiding the exploration of conformational space. The latter strategy might be of help in protein/protein docking schemes (47) through the generation of conformers of partner biomolecules.

The program implementing the decomposition is freely available, upon request, for academic use. We are currently working on making the tool available also as a Web service named PiSQRD (after protein structure quasi-rigid domain decomposition).

SUPPORTING MATERIAL

Additional details on the algorithm, 10 figures, and three tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(09)00781-4.

We thank Fabio Caccioli, Giorgio Colombo, Alejandro Giorgetti, and Konrad Hinsen for valuable discussions.

We acknowledge financial support from the Italian Ministry for Education (grant No. PRIN-2006025255).

REFERENCES

1. Frauenfelder, H., H. Sligar, and P. Wolynes. 1991. The energy landscape and motions of proteins. *Science*. 254:1598–1603.
2. Garcia, A. 1992. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett*. 68:2696–2699.
3. Tirion, M. M. 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett*. 77:1905–1908.
4. Falke, J. J. 2002. Enzymology. A moving story. *Science*. 295:1480–1481.
5. Piana, S., P. Carloni, and M. Parrinello. 2002. Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease. *J. Mol. Biol*. 319:567–583.
6. Garcia-Viloca, M., J. Gao, M. Karplus, and D. G. Truhlar. 2004. How enzymes work: analysis by modern rate theory and computer simulations. *Science*. 303:186–195.
7. Perryman, A. L., J.-H. Lin, and J. A. McCammon. 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.
8. Wolf-Watz, M., V. Thai, K. Henzler-Wildman, G. Hadjipavlou, E. Z. Eisenmesser, et al. 2004. Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat. Struct. Mol. Biol*. 11:945–949.
9. Hanson, J. A., K. Duderstadt, L. P. Watkins, S. Bhattacharyya, J. Brokaw, et al. 2007. Illuminating the mechanistic roles of enzyme conformational dynamics. *Proc. Natl. Acad. Sci. USA*. 104:18055–18060.
10. Arora, K., and C. L. Brooks. 2007. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc. Natl. Acad. Sci. USA*. 104:18496–18501.
11. del Sol, A., M. J. Arauzo-Bravo, D. Amoros, and R. Nussinov. 2007. Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. *Genome Biol*. 8:R92. 10.1186/gb-2007-8-5-r92.
12. Bradley, M. J., P. T. Chivers, and N. A. Baker. 2008. Molecular dynamics simulation of the *Escherichia coli* NikR protein: equilibrium conformational fluctuations reveal interdomain allosteric communication pathways. *J. Mol. Biol*. 378:1155–1173.
13. Kumar, S., B. Ma, C. J. Tsai, N. Sinha, and R. Nussinov. 2000. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci*. 9:10–19.
14. Hinsen, K. 1998. Analysis of domain motions by approximate normal mode calculations. *Proteins*. 33:417–429.
15. Hinsen, K., A. Thomas, and M. J. Field. 1999. Analysis of domain motion in large proteins. *Proteins*. 34:369–382.
16. Hayward, S., A. Kitao, and H. J. C. Berendsen. 1997. Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins Struct. Funct. Genet*. 27:425–437.
17. Kundu, S., D. C. Sorensen, and J. G. N. Phillips. 2004. Automatic domain decomposition of proteins by a Gaussian network model. *Proteins*. 57:725–733.
18. Yesylevskyy, S. O., V. N. Kharkyanen, and A. P. Demchenko. 2006. Hierarchical clustering of correlation patterns: new method of domain identification in proteins. *Biophys. Chem*. 119:84–93.
19. Golhlke, H., and M. F. Thorpe. 2006. A natural coarse graining for simulating large biomolecular motion. *Biophys. J*. 91:2115–2120.
20. Schomaker, V., and K. N. Trueblood. 1968. On rigid-body motion of molecules in crystals. *Acta Crystallogr. B*. 24:63–76.
21. Chaudhry, C., A. L. Horwich, A. T. Brunger, and P. D. Adams. 2004. Exploring the structural dynamics of the *E. coli* chaperonin GroEL using translation-libration-screw crystallographic refinement of intermediate states. *J. Mol. Biol*. 342:229–245.
22. Painter, J., and E. A. Merritt. 2006. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. D Biol. Crystallogr*. 62:439–450.
23. Amadei, A., A. B. M. Linssen, and H. J. C. Berendsen. 1993. Essential dynamics of proteins. *Proteins*. 17:412–425.
24. Bahar, I., B. Erman, T. Haliloglu, and R. L. Jernigan. 1997. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. *Biochemistry*. 36:13512–13523.
25. Micheletti, C., P. Carloni, and A. Maritan. 2004. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. *Proteins*. 55:635–645.
26. Hinsen, K., and G. R. Kneller. 2000. Projection methods for the analysis of complex motions in macromolecules. *J. Mol. Sim*. 23:275–292.
27. Blatt, M., S. Wiseman, and E. Domany. 1996. Superparamagnetic clustering of data. *Phys. Rev. Lett*. 76:3251–3254.
28. Maragakis, P., and M. Karplus. 2005. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol*. 352:807–822.

29. Lou, H., and R. I. Cukier. 2006. Molecular dynamics of apo-adenylate kinase: a principal component analysis. *J. Phys. Chem. B*. 110:12796–12808.

30. Pontiggia, F., A. Zen, and C. Micheletti. 2008. Small- and large-scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics. *Biophys. J*. 95:5901–5912.

31. Cascella, M., C. Micheletti, U. Rothlisberger, and P. Carloni. 2005. Evolutionarily conserved functional mechanics across pepsin-like and retroviral aspartic proteases. *J. Am. Chem. Soc*. 127:3734–3742.

32. Carnevale, V., S. Raugei, C. Micheletti, and P. Carloni. 2006. Convergent dynamics in the protease enzymatic superfamily. *J. Am. Chem. Soc*. 128:9766–9772.

33. Verkhivker, G., G. Tiana, C. Camilloni, D. Provasi, and R. A. Broglia. 2008. Atomistic simulations of the HIV-1 protease folding inhibition. *Biophys. J*. 95:550–562.

34. Ishima, R., and J. M. Louis. 2008. A diverse view of protein dynamics from NMR studies of HIV-1 protease flaps. *Proteins*. 70:1408–1415.

35. Ding, F., M. Layten, and C. Simmerling. 2008. Solution structure of HIV-1 protease flaps probed by comparison of molecular dynamics simulation ensembles and EPR experiments. *J. Am. Chem. Soc*. 130:7184–7185.

36. Pearl, F., A. Todd, I. Sillitoe, M. Dibley, O. Redfern, et al. 2005. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res*. 33:D247–D251.

37. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). 2009. Enzyme Nomenclature. http://www.chem.qmul.ac.uk/iubmb/enzyme/.

38. Tyndall, J. D., T. Nall, and D. P. Fairlie. 2005. Proteases universally recognize beta strands in their active sites. *Chem. Rev*. 105:973–999.

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

40. The Uniprot Consortium. 2008. The Universal Protein Resource (UniProt). *Nucleic Acids Res*. 36:D190–D195.

41. Mol, C., C.-F. Kuo, M. Thayer, R. Cunningham, and J. Tainer. 1995. Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature*. 374:381–386.

42. Stockwell, G., and J. Thornton. 2006. Conformational diversity of ligands bound to proteins. *J. Mol. Biol*. 356:928–944.

43. Roujeinikova, A., S. Sedelnikova, G. de Boer, A. Stuitje, A. Slabasi, et al. 1999. Inhibitor binding studies on enoyl reductase reveal conformational changes related to substrate recognition. *J. Biol. Chem*. 274:30811–30817.

44. Pidugu, L., M. Kapoor, N. Surolia, A. Surolia, and K. Saguna. 2004. Structural basis for the variation in triclosan affinity to enoyl reductases. *J. Mol. Biol*. 343:147–155.

45. Gupta, S., W. Mangel, W. McGrath, J. Perek, D. Lee, et al. 2004. DNA binding provides a molecular strap activating the adenovirus proteinase. *Mol. Cell. Proteomics*. 3:950–959.

46. Redfern, O. C., A. Harrison, T. Dallman, F. M. Pearl, and C. A. Orengo, 2007. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput. Biol*. 3:e232. 10.1371/journal.pbio.0020232.

47. Dominguez, C., R. Boelens, and A. M. Bonvin. 2003. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc*. 125:1731–1737.

48. Porter, C. T., G. J. Bartlett, and J. M. Thornton. 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*. 32:D129–D133.