Structural and dynamical alignment of enzymes with partial structural similarity

J. Phys.: Condens. Matter **19** (2007) 285206 (14pp) doi:10.1088/0953-8984/19/28/285206

Structural and dynamical alignment of enzymes with partial structural similarity

Vincenzo Carnevale, Francesco Pontiggia and Cristian Micheletti

International School for Advanced Studies (SISSA) and CNR-INFM Democritos, Via Beirut 2-4, 34014 Trieste, Italy

Received 2 October 2006, in final form 14 November 2006
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/285206

Abstract

Proteins and enzymes, in order to carry our their biological tasks, often need to sustain concerted displacements of a large number of amino acids. In recent years many theoretical and computational studies have pointed out how these large-scale movements, also termed slow modes or essential dynamical spaces, are mostly dictated by the overall structural organization of the protein. Several fundamental questions arise when this fact is complemented by the observation that enzymes within the same enzymatic superfamily can have remarkable conformational differences. Could their large-scale movements be similar despite the difference in structure? In this study we address this issue and present a quantitative scheme for comparing the slow modes in proteins that, though differing in sequence, length and tertiary structure, still admit a partial structural superposition. We illustrate the application of the method to two representatives of the protease enzymatic superfamily, carboxypeptidase A and pyroglutamyl peptidase.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

A comprehensive understanding of the properties of proteins requires a detailed characterization of the sequence, structure and function of the protein [1–4]. Of paramount interest to molecular biophysicists and chemists is the elucidation of the fundamental laws that interrelate these aspects. The best known example is the Anfisenian principle, asserting that, for naturally occurring proteins of moderate length, the native structure is dictated by the primary sequence [5–7]. A second important avenue is the elucidation of the relationship between structure and function [1, 8]. The present study considers several aspects of this problem and particularly addresses the possibility that similar concerted large-scale movements around the native structure, traditionally associated with functional dynamics, can be exhibited by enzymes or proteins with different structural organization [9]. In particular we shall report on a general quantitative strategy for the comparison of large-scale movements in pairs of enzymes whose structures, though differing in sequence, length and tertiary content, are still partially

0953-8984/07/285206+14$30.00 © 2007 IOP Publishing Ltd Printed in the UK

superimposable. The study complements and advances a recent one [9] in which pervasive dynamical correspondences were detected within the protease enzymatic superfamily. Here we introduce a novel method for the comparison of essential dynamical spaces and apply it to two important members of the protease superfamily [10], namely carboxypeptidase A and pyroglutamyl peptidase. Though the overall architecture of the two biomolecules is analogous according to the CATH database [11], the differences in length (208 and 307 residues, respectively) and number of secondary elements is sufficiently large to require the use of a non-subjective and quantitative scheme to detect and quantify the dynamical correspondences. The starting point of the analysis is the identification of subregions of the two proteins that can be put in structural correspondence. A statistical mechanical procedure is then used to measure the consistency of the fluctuation dynamics over the residues in structural correspondence. The case of carboxy and pyroglutamyl peptidases allows us to discuss the methodology with a high level of transparency, given their analogies in fold. The comparison of their large-scale dynamics is, however, not a mere exercise. In fact, the two enzymes rely on different catalytic chemistries and belong to different clans (clans MC and CF for carboxypeptidase A and pyroglutamyl peptidase, respectively). Hence, the possibility to establish a consistency in their large-scale fluctuation dynamics provides valuable insight into generality of the structural rearrangements that accompany the proteolysis reaction.

The material presented here is organized as follows. First we summarize the main structural and chemical features of the two enzymes and review the motivations for the search of a general and quantitative tool for comparing their large-scale movements. Next we illustrate the detailed methodology, previously outlined in [9], through which the essential dynamical spaces, putatively influencing enzymatic functionality, can be obtained and compared. In particular we discuss the three main steps of the analysis, which regard: (a) the identification of significant partial structural correspondences (if any) between the two enzymes, (b) the use of a suitable coarse-grained model to calculate the large-scale movements of the residues in structural correspondence, and (c) the measure of the dynamical accord over the matching residues. The measure of dynamical consistency introduced here generalizes others commonly employed to compare two sets of essential dynamical spaces in that it depends on the weight with which the essential spaces contribute to the overall system fluctuations. By considering the dynamical correspondence of individual residues it is found that the highest accord occurs for Tyr248 for carboxypeptidase A and Gly138 for pyroglutamyl peptidase. Notably, both residues are strictly conserved within the respective protein family and have been previously argued, on the basis of experimental evidence, to influence the binding and/or processing of the substrate [10]. This observations illustrate how valuable biochemical informations, arguably dictated by functional selection mechanisms, can be elucidated through dynamical alignments.

2. Proteases

Proteases play crucial roles in the life cycle of all organisms by affecting a wide spectrum of physiological processes such as cell growth, cell death, blood clotting, immune defense and secretion. At a molecular level they act as ‘scissors’ capable of cleaving polypeptide chains, that is other proteins and enzymes. The repertoire of known proteases covers a wide range of:

- (a) catalytic/reactive mechanisms and substrate specificities (the hydrolysis reaction leading to the cleavage of the peptide bond can involve different catalytic residues, such as Ser, Asp, Cys, Glu and Thr or even Zn metal ions) [10],
- (b) structural folds, as the approximately 2000 proteases of known structure can be assigned to as many as 13 distinct folds [12].

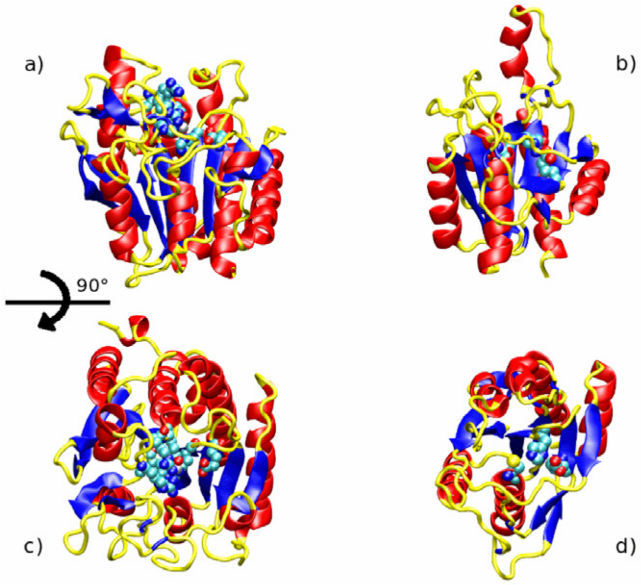

Figure 1. Cartoon representations of (a) carboxypeptidase A and (b) pyroglutamyl peptidase, PDB codes [17] 8cpa [13] and 1iof [14], respectively. The bottom panels presents a different view of the same structures. The catalytic residues are shown with atomistic detail.

This chemical and structural diversity is so significant that, prior to the identification of common fluctuation dynamics [9], the various classes were thought to be related only by the fact that the peptide substrate in the binding cleft adopts an extended beta-conformation [12].

The two enzymes considered here, carboxyl peptidase [13] and a pyroglutamyl peptidase [14], provide an example of the differences in protease catalytic chemistry and, to a lesser extent, of the structural traits. A ribbon representation of the two enzymes is given in figure 1. According to the MEROPS classification of proteases [10], carboxypeptidase A belongs to the M14 family of metallo proteases, whose function in mammals is related to alimentary digestion. It acts by cleaving a single C-terminal amino acid (particularly aromatic ones or residues with branched side chains). Its fold is constituted by a three-layer alpha/beta/alpha-sandwich with an antiparallel beta-sheet of eight strands. The active site is on the beta-layer and is composed by a single catalytic zinc ion, tetrahedrally coordinated by two histidines (His69 and His196), a glutamate (Glu72) and the catalytic water molecule (responsible for the nucleophilic attack), Arg127 and Glu270. His69 Arg71 and Glu72, arranged in a sequence pattern well conserved among the family, are in a loop connecting one beta-strand and one of the helices; Arg127 is also in a loop connecting two helices; His196 and Glu270 are placed on adjacent strands. Another key residue, Tyr248, which is strictly conserved within the family, is located in a loop surmounting the active site; experimental evidence has indicated its important role in substrate binding and/or catalysis [15].

The second protease is a bacterial pyrrolidone carboxyl peptidase, from hyperthermophile, and belongs to the cysteine protease enzymatic class. Its molecular function consists of removing one pyroglutamate residue from the N-terminus of a peptide. As in almost all cysteine proteases, the active site is constituted by a catalytic triad, namely a glutamate, a histidine and a cysteine (the latter being the nucleophilic agent) [10]. Akin to carboxypeptidase A the pyroglutamyl peptidase is also organized as an alpha/beta/alpha-sandwich with four long parallel and two short antiparallel beta-strands surrounded by three helices on one side and two on the other [16]. However, its crystallographic quaternary organization can be very different for different organisms: monomeric for mammals and bacteria as opposed to tetrameric for archaea. The oligomeric state of the enzyme in solution is still a matter of debate. However, as the active site is contained completely within a monomer we will consider here the enzyme in its monomeric form. Interestingly, the fold is very different from that of any other cysteine peptidase, but structural similarities were detected between members of its clan and those of clans of metallopeptidases. In fact, despite the fact that the MEROPS [10] classification scheme assigns it to a different clan from the metallo carboxypeptidases the core of both enzymes presents visible analogies (see figure 1). Moreover, as for carboxypeptidase A, the active site is located on the beta-layer with the Glu79 and His166 positioned on two adjacent strands of the layer and the nucleophilic Cys142 located on a alpha-helix flanking the beta-layer.

3. Large-scale movements

We shall now discuss in detail how the native structure of the two enzymes can be exploited to gain insight into their putative functional movements. At the heart of our theoretical/computational approach is the use of simplified mesoscopic models which allow us to capture, with minimal computational expenditure, the concerted large-scale fluctuations customarily believed to be capable of steering and influencing the enzymatic functionality [8].

The viability of coarse-grained models can be argued *a priori* on the basis of extensive experimental and theoretical evidence which indicates that proteins and enzymes, in order to carry out their biological tasks, undergo conformational changes in which several amino acids are significantly displaced, in a concerted manner, from their reference (native state) positions [18]. The residues taking part in these coordinated changes may, in fact, be easily displaced by several angstroms over time spans of 1–10 ns [19]. This remarkable degree of elasticity (arguably facilitated by the presence of secondary motifs [20]), motivates and justifies the introduction of simplified models [21] to predict and describe them. In fact, since the main objective is the modelling of the mesoscopic structural modulations in a protein it is convenient to reduce the spatial degrees of freedom of the biopolymer through a coarse-graining procedure in which each amino acid is represented by a limited number of interaction centres. Following the model of [22], of which we retrace the derivation for the sake of completeness, we shall adopt a two-centroid amino acid representation, one for the mainchain, coinciding with the CA atom, and one for the sidechain. Following a geometrical rule akin to the one introduced by Park and Levitt [23], we construct the latter interaction centre as a fictitious CB centroid:

$$\vec{r}_{\mathrm{CB}}(i) = \vec{r}_{\mathrm{CA}}(i) + l\frac{2\vec{r}_{\mathrm{CA}}(i) - \vec{r}_{\mathrm{CA}}(i+1) - \vec{r}_{\mathrm{CA}}(i-1)}{|2\vec{r}_{\mathrm{CA}}(i) - \vec{r}_{\mathrm{CA}}(i+1) - \vec{r}_{\mathrm{CA}}(i-1)|}, \tag{1}$$

where $l = 3$ Å and $\vec{r}_{\mathrm{CA}}$ indicates the coordinates of the ith CA centroid. For amino acids at the beginning/end of the peptide chain(s) the construction of equation (1) is not applicable and hence the effective CB centroid is taken to coincide with the CA one.

A schematic view of the coarse graining procedure is given in figures 2(a) and (b).

The coarse-grained structural representation is instrumental for modelling in a simple and effective way how a protein fluctuates around the average (equilibrium) conformation. The latter is aptly taken to coincide with the known crystal structure of the protein, possibly energy-minimized and relaxed with atomistic force fields. The simplest (roto-translationally invariant) energy function ensuring that in thermal equilibrium the system fluctuates around the reference conformation is obtained by introducing the following quadratic penalties for displacing two

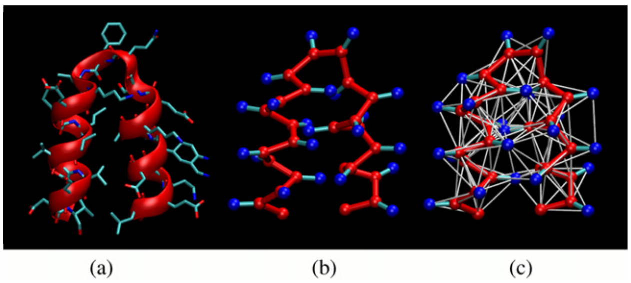

Figure 2. Pictorial representation of the coarse-graining procedure: (a) atomic representation of a two-helix bundle (backbone highlighted as a ribbon); (b) simplified structural representation in terms of the CA centroids for the backbone and the CB ones for the sidechains; (c) all pairs of non-consecutive centroids within 7.5 Å interact through an harmonic potential, schematically shown as a thin bond.

centroids, i and j from their reference positions, $\vec{r}_i^{\,0}$ and $\vec{r}_j^{\,0}$ to generic ones, $\vec{r}_i$ and $\vec{r}_j$:

$$V(\vec{r}_{ij}) = \frac{k}{2} \sum_{\mu,\nu} \frac{r^0_{ij,\mu}\, r^0_{ij,\nu}}{|\vec{r}^{\,0}_{ij}|^2}\, \delta r_{ij,\mu}\, \delta r_{ij,\nu}, \tag{2}$$

where $\vec{r}^{\,0}_{ij} \equiv (\vec{r}^{\,0}_i - \vec{r}^{\,0}_j)$ is the native distance vector of the centroids, $\delta\vec{r}_{ij}$ is the distance vector change, $\delta\vec{r}_{ij} \equiv (\vec{r}_i - \vec{r}^{\,0}_i) - (\vec{r}_j - \vec{r}^{\,0}_j)$, μ and ν run over the three Cartesian components, and k is a parameter controlling the strength of the quadratic coupling. The quadratic form of equation (2) is at the heart of the widely used elastic or Gaussian network approaches [24–28, 20, 29, 22], which typically adopt a single-centroid amino acid description. The effective free energy function introduced in [22] and used here includes, instead, pairwise contributions from all pairs of centroids, be they of the CA or CB type, whose reference distance falls within a given interaction cutoff, as pictorially illustrated in figure 2(c). Accordingly, the resulting free energy of a trial structure, Γ, takes on the form

$$\mathcal{F}(\Gamma) = 2\sum_i V(\vec{r}^{\,\mathrm{CA-CA}}_{i,i+1}) + \sum_{i<j}{}' V(\vec{r}^{\,\mathrm{CA-CA}}_{i,j}) + \sum_{i,j}{}' V(\vec{r}^{\,\mathrm{CA-CB}}_{i,j}) + \sum_{i<j}{}' V(\vec{r}^{\,\mathrm{CB-CB}}_{i,j}), \tag{3}$$

where i and j run over the residue indices, $\vec{r}^{\,\mathrm{X-Y}}_{i,j}$ denotes the distance vector of the centroids of type X and Y of residues i and j, respectively, and the prime denotes that the sum is restricted to the pairs whose reference separation is below the cutoff distance of 7.5 Å. Consistently with the spirit of elastic network models and other approaches [21], the last three terms in equation (3) have the same strength irrespective of the identity of the amino acids. The first term, on the other hand, accounts for the protein chain connectivity and has a double strength to reflect the geometrical constraints of the peptide chain.

As the positions of the CB centroids depend linearly on the coordinates of the CA ones, it is possible to recast expression (3) in the following quadratic form:

$$\mathcal{F} = \tfrac{1}{2}k \sum_{ij,\mu\nu} \delta r_{i,\mu}\, \mathcal{M}_{ij,\mu\nu}\, \delta r_{j,\nu}, \tag{4}$$

where $\delta\vec{r}_i = \vec{r}^{\,\mathrm{CA}}_i - \vec{r}^{\,0\mathrm{CA}}_i$ is the deviation of the ith CA centroid from the reference position and $\mathcal{M}$ is a symmetric matrix whose linear size is three times the number of residues in the protein. The characterization of the fluctuation dynamics of the system is conveniently made in terms of the eigenvectors and eigenvalues of $\mathcal{M}$. A transparent framework through which this can be illustrated is the overdamped Langevin scheme for the dynamical evolution of the system [30, 31]. Accordingly, the stochastic equations of motion for each amino acid subject to the thermodynamic potential of equation (4) and the surrounding medium [32, 26, 33–35] are

$$\gamma_i \dot{\delta r}_{i,\mu}(t) = -k \sum_{j,\nu} \mathcal{M}_{ij,\mu\nu} \delta r_{j,\nu}(t) + \eta_{i,\mu}(t), \tag{5}$$

where γ_i is the effective viscous friction coefficient acting on the ith particle, and $\eta_{i,\mu}(t)$ is a stochastic noise whose first and second moments satisfy the usual fluctuation–dissipation relationships [31, 30]

$$\langle \eta_{i,\mu}(t) \rangle = 0, \qquad \langle \eta_{i,\mu}(t) \eta_{j,\nu}(t') \rangle = \delta_{ij}\, \delta_{\mu\nu}\, \delta(t - t')\, 2K_B\, T\, \gamma_i, \tag{6}$$

where $K_B\, T$ is the thermal energy. From equations (5) and (6) the average correlations among the displacements of various pairs of residues can be calculated [36]. For the case in which the various viscous coefficients in equation (5) take on the same value, γ, one has

$$\left\langle \delta r_{i,\mu}(t)\, \delta r_{j,\nu}(t + \Delta t) \right\rangle = \frac{K_B T}{k} \sum_l{}' v^l_{i,\mu} v^l_{j,\nu} \frac{1}{\lambda_l}\, \mathrm{e}^{-\lambda_l \frac{k}{\gamma} \Delta t}, \tag{7}$$

where $\mathbf{v}^l$ and λ_l are, respectively, the lth eigenvector and the lth eigenvalue of the matrix $\mathcal{M}$ and the prime indicates the omission from the sum of the six eigenspaces associated to the zero eigenvalues of $\mathcal{M}$ (roto-translational degrees of freedom) [22]. Expression (7) indicates that the eigenvectors of $\mathcal{M}$ represent the independent modes of structural relaxation in the protein while the associated eigenvalues are inversely proportional to the relaxation times. It is important to notice that the eigenvectors associated to the smallest non-zero eigenvalues of $\mathcal{M}$ coincide, within the quadratic approximation, with the essential dynamical spaces of the protein. In atomistic simulation contexts, the essential dynamical spaces are identified as the dominant eigenvectors of the covariance matrix, $\mathcal{C}$, defined as

$$\mathcal{C}_{ij,\mu\nu} \equiv \left\langle \delta r_{i,\mu}\, \delta r_{j,\nu} \right\rangle, \tag{8}$$

where, assuming ergodicity, the average is intended to be equivalently taken uniformly over the simulated time or with the canonical statistical weight. By comparing equation (8) with equation (7) specialized to equal times, $\Delta t = 0$, it is possible to establish the correspondence of $\mathcal{C}$ and the pseudo-inverse of $\mathcal{M}$:

$$\mathcal{C}_{ij,\mu\nu} \equiv \left\langle x_{i,\mu} x_{j,\nu} \right\rangle = \frac{K_B T}{k} \sum_l{}' \frac{v^l_{i,\mu} v^l_{j,\nu}}{\lambda_l} = \frac{K_B T}{k} \mathcal{M}'^{-1}_{i,j,\mu,\nu}, \tag{9}$$

where the prime indicates the removal of the zero eigenspace of $\mathcal{M}$ prior to inversion. The above identity manifests the equivalence, within the quadratic free energy approximation, of the slow modes with the essential dynamical spaces. The utility of this observation in practical contexts is limited, *a priori*, by the viability of the quadratic approximation to describe the near-native free energy of proteins and enzymes [37–39]. In past years this comparison has been carried out for several different proteins and enzymes by analysing the consistency of the slowest (non-zero) modes obtained from the simplified model of equation (4) and the principal components of the covariance matrix obtained from atomistic simulations spanning several nanoseconds [22, 40, 41, 9]. In all the cases considered the consistency between the simplified and the atomistic approaches has been very significant from a statistical and also a practical point of view. We shall therefore build on these previous validations and thus establish the essential dynamical spaces of carboxypeptidase A and pyroglutamyl peptidase through the analysis of the quadratic free energy matrix of equation (4) constructed from the crystal structures of the two enzymes. The two slowest modes thus identified for each enzyme are illustrated in figure 3.

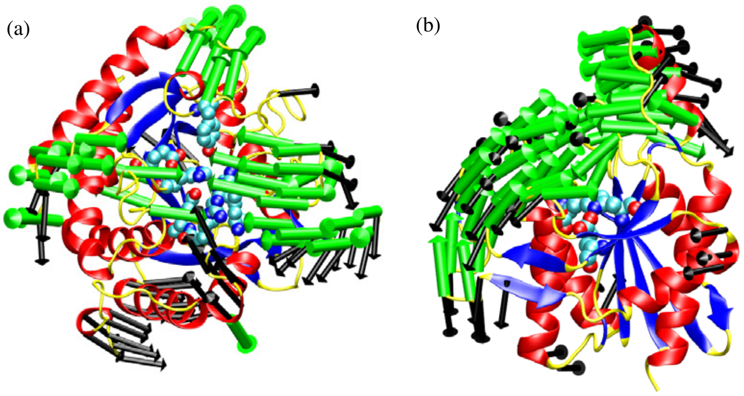

Figure 3. The displacements of the most mobile residues in the first [second] slowest mode for (a) carboxypeptidase A and (b) pyroglutamyl peptidase are shown with thick [thin] arrows of equal length.

For carboxypeptidase A, the slow modes appear mostly localized in the regions delimiting the binding site. In particular they result in a contraction/elongation of the distances between the two unstructured regions (from Ser157 to Tyr169 and from Thr274 to Phe279) and between the top of one helix (from Phe118 to Leu125) and the loop from Ile244 to Gln249. It is interesting to notice that the latter is constituted by residues involved in enzyme/substrate interactions. The overall behaviour suggests the pictorial representation of a 'breathing' motion of the binding site.

In pyroglutamyl peptidase six long loops embrace the binding pocket (three on each side), resulting in an active site cleft flanked by two lobes. Most of the movements entailed by the slowest modes are concentrated in these regions. The concerted motion of two lobes can be aptly described in terms of a bend (first mode) and shear (second mode) motion. It is interesting to note that, as for carboxypeptidase A, there is a modulation of the overall shape of the cleft. However, apart from this qualitatively similar behaviour, it is very difficult to infer, by mere visual inspection, the existence of any dynamical consistency of the two enzymes in the neighbourhood of the active region. In the next sections we shall illustrate how this question can be answered within a quantitative framework.

4. Partial structural alignment

The first step of the dynamical comparison procedure relies on the identification of amino acids that can be put in structural correspondence in the two enzymes. In fact, a one-to-one structural correspondence of a certain number of residues in the two enzymes lends naturally to measuring the consistency of the large-scale displacements of the matching residues. The interest in detecting and characterizing dynamical similarities grows with the chemical and structural diversity of the enzymes. For those considered here the conformational difference, though not impacting on the overall architecture, is significant both for the secondary content and enzyme length. This is sufficient to call for an automated scheme to single out the structural analogies which, by necessity, cannot encompass the biomolecules in their entirety. The search for the best *partial* structural match of the enzymes is aptly performed with the DALI algorithm [42]. The algorithm is based on a scoring function (formulated in terms of the matrix of pairwise residue distances of each enzyme) that quantifies the geometrical consistency of any

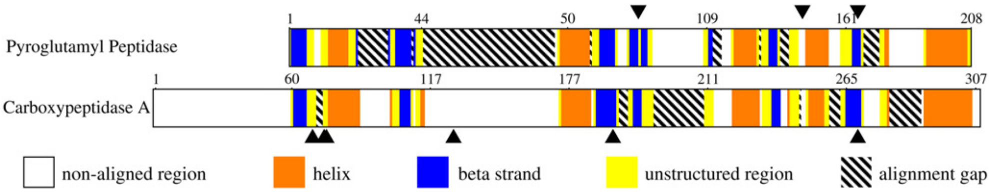

Figure 4. Primary sequence correspondence induced by the optimal DALI structural superposition of the two proteases. Each primary sequence is coloured so as to reflect the secondary structure. Blank boxes represent non-aligned regions while alignment gaps are represented with hatched boxes. The triangles mark the location of the catalytic triad for pyroglutamyl peptidase and of the five residues coordinating the Zn ion for carboxypeptidase A.

given set of corresponding residues in the two enzymes. The best partial structural alignment is identified by a stochastic optimization of the scoring function. Typical enzymatic DALI structural alignments lead to identifying several corresponding blocks of consecutive residues, each involving 10–15 amino acids. It is clear that, owing to the pervasive presence of secondary motifs in proteins, the search of partial structural matches in two proteins will almost always be successful. Discriminating between meaningful structural alignments and ‘accidental’ ones is therefore a key point of the analysis, which entails a statistical significance test. Indeed, for each DALI alignment, the optimized value of the scoring function is compared against a reference distribution of scores expected for two generic enzymes of length equal to the assigned ones. The statistical significance of the best DALI alignment is finally summarized in a Z-score which measures the number of standard deviations by which the optimal score exceeds the reference one. The better the alignment, the larger the Z-score.

For pyroglutamyl peptidase and carboxypeptidase A, their DALI Z-score was equal to 8.9, which indicates that they can be aligned with high statistical confidence. The optimal partial alignment involves 151 residues out of 208 [307] for pyroglutamyl peptidase [carboxypeptidase A]. The alignment has a simple character in that aligned blocks have the same succession in the two enzymes (and no inversion of sequence directionality occurs). This simple organization allows us to represent the alignment with the simple graphical plot of figure 4.

As can be seen in the figure, the first 60 residues of carboxypeptidase A are excluded from the alignment. The first, third, fourth, fifth, eight, and ninth beta-strands of pyroglutamyl peptidase match respectively the third, fourth, fifth, sixth, seventh and eight of carboxypeptidase A; helices 1, 2, 3, 4, 6 of pyroglutamyl peptidase are in correspondence with helices 2, 5, 6, 8, 9 of carboxypeptidase A; helix 5 of pyroglutamyl peptidase has a partial match with helix 9 of carboxypeptidase A. Also, there are non-trivial correspondences between different secondary structure elements: beta-strand 2 of pyroglutamyl peptidase matches a fragment of helices 2 and 3 of carboxypeptidase A, parts of beta-strands 3, 5, 6 and 7 of pyroglutamyl peptidase match loops in the partner structure, as well as segments of helices 2, 4, 7 and 9 of carboxypeptidase A. A view of the aligned structures is given in figure 5.

5. Dynamical alignment

We are interested in calculating the concerted displacements in the two proteins only of the aligned residues, yet we wish to take into account the dynamical influence exerted over them by the non-aligned ones. The spirit of the procedure can be conveniently illustrated in the context of the analysis of dynamical trajectories obtained from atomistic simulations. The trajectory is first obtained by integrating the Newtonian equations of motion for the atomic constituents

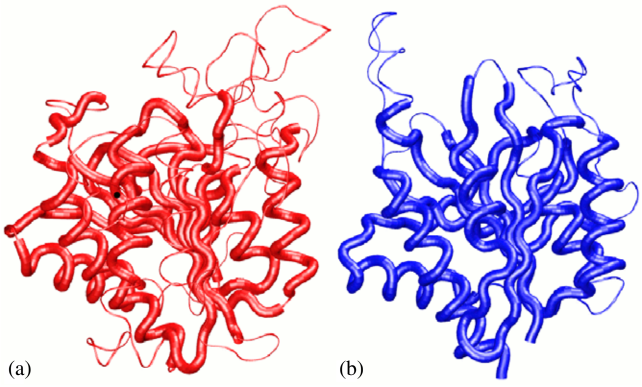

Figure 5. Backbone traces of (a) carboxypeptidase A and (b) pyroglutamyl peptidase. The matching regions are highlighted as a thick CA trace.

of all residues. To extract the essential dynamical spaces of only a subset of residues one would construct their covariance matrix by processing the recorded 'frames' and neglecting the presence of the other residues [43]. In the spirit of the coarse-grained model considered here this amounts to calculating the slow modes of the effective free energy matrix obtained by integrating over (and not omitting!) the degrees of freedom of the non-aligned amino acids. As we shall discuss, the quadratic form of the model free energy allows to carry out this integration in an exact way [35, 44, 9].

For simplicity of notation we assume that amino acids have been reindexed so that the first n residues are those taking part to the DALI alignment (i.e. corresponding residues have the same index in the two enzymes). For each enzyme we denote the displacements of this set of residues with a collective vector $\delta\vec{R} \equiv \{\delta\vec{r}_1, \delta\vec{r}_2, \ldots, \delta\vec{r}_n\}$. Analogously, the displacement of the collective set of non-matching residues is formally indicated with $\delta\vec{Q}$.

The energy function (4) of each enzyme then reads

$$\mathcal{F} = \frac{k}{2}(\delta\vec{R}^\dagger\ \delta\vec{Q}^\dagger)\begin{pmatrix}\mathbf{F}_1 & \mathbf{G}\\ \mathbf{G}^\dagger & \mathbf{F}_2\end{pmatrix}\begin{pmatrix}\delta\vec{R}\\ \delta\vec{Q}\end{pmatrix},$$

where $\mathbf{F}_1$ [$\mathbf{F}_2$] is the interaction matrix within the set of [non-] matching residues and $\mathbf{G}$ contains the pairwise couplings across the two sets. The probability of occurrence of displacements $\delta\vec{R}$ and $\delta\vec{Q}$ in thermal equilibrium is given by the Boltzmann distribution. Neglecting the normalization factor, it reads

$$P(\delta\vec{R}, \delta\vec{Q}) = \exp\left(-\frac{\mathcal{F}}{k_B T}\right) \propto \exp\left(-\frac{\delta\vec{R}^\dagger\mathbf{F}_1\,\delta\vec{R} + \delta\vec{Q}^\dagger\mathbf{F}_2\,\delta\vec{Q} + 2\delta\vec{R}^\dagger\mathbf{G}\,\delta\vec{Q}}{2k_B T/k}\right). \tag{10}$$

Since we focus only on the free energy change associated with residues in the first set, we calculate the probability distribution for this set *integrated* over all displacements of set 2, $\delta\vec{Q}$. The integration can be evaluated analytically and yields [35]

$$\tilde{P}(\delta\vec{R}) = \int \mathrm{d}\delta\vec{Q}\ P(\delta\vec{R}, \delta\vec{Q}) \propto \exp\left[-\left(\frac{\delta\vec{R}^\dagger\left(\mathbf{F}_1 - \mathbf{G}\,\mathbf{F}_2^{-1}\mathbf{G}^\dagger\right)\delta\vec{R}}{2k_B T/k}\right)\right]; \tag{11}$$

hence the effective free energy matrix that controls the displacement of residues in the first set is

$$\tilde{\mathbf{F}}_1 = \left(\mathbf{F}_1 - \mathbf{G}\,\mathbf{F}_2^{-1}\mathbf{G}^\dagger\right). \tag{12}$$

Thus, the eigenvectors associated to the smallest eigenvalues of $\tilde{\mathbf{F}}_1$ represent the integrated slow modes of the matching regions; the term ‘integrated’ is used to stress the fact that the modes depend also on the non-matching ones via the contributions $\mathbf{G}$ and $\mathbf{F}_2$.

The eigenvectors of $\tilde{\mathbf{F}}_1$, calculated separately for the two enzymes, can be directly compared component by component (we assume that the proteins are represented in the Cartesian coordinate set providing the optimal structural superposition of the DALI matching regions). To quantify the agreement of the integrated dynamics we applied the following heuristic procedure which straightforwardly leads to defining a novel measure of dynamical accord which generalizes the widely used RMSIP value [45].

We introduce infinitesimal harmonic couplings between corresponding residues in the two proteins. More precisely, we consider the following effective free energy function for two coupled proteins, A and B (optimally rototranslated):

$$\frac{k}{2}\,(\delta\vec{R}_A^\dagger\;\delta\vec{R}_B^\dagger)\begin{pmatrix}\tilde{\mathbf{F}}_A & \epsilon\,\mathbf{1}\\ \epsilon\,\mathbf{1} & \tilde{\mathbf{F}}_B\end{pmatrix}\begin{pmatrix}\delta\vec{R}_A\\ \delta\vec{R}_B\end{pmatrix}, \qquad (13)$$

where ϵ indicates the strength of the harmonic coupling between corresponding residues and $\mathbf{1}$ indicates the identity matrix. We shall be concerned with the limit $\epsilon \to 0$, where the harmonic coupling becomes a weak perturbation. The coupling of equation (13) has an abstract character in that it introduces isotropic interactions between corresponding residues without reference to any particular notion of space proximity of residues in the two proteins. Due to its minimalistic nature, the free energy function of equation (13) does not have the full spatial invariance properties of for example equation (4), though it is still invariant for rotations of both proteins.

The introduction of the coupling between the matching residues facilitates the detection of similar large-scale fluctuations of corresponding residues in the two proteins. The information on the extent to which these correlations exist is aptly conveyed by the covariance matrix obtained by inverting the effective matrix, see equation (9). To leading order in ϵ the covariance matrix is given by

$$\begin{pmatrix}\tilde{\mathbf{F}}_A^{-1} & -\epsilon\tilde{\mathbf{F}}_A^{-1}\tilde{\mathbf{F}}_B^{-1}\\ -\epsilon\tilde{\mathbf{F}}_A^{-1}\tilde{\mathbf{F}}_B^{-1} & \tilde{\mathbf{F}}_B^{-1}\end{pmatrix}.$$

According to this expression, the degree of dynamical correlation of the displacements of pairs of corresponding residues is provided by the diagonal terms of the submatrix, $(-\epsilon\tilde{\mathcal{F}}_A^{-1}\tilde{\mathcal{F}}_B^{-1})$. To turn this observation into a practical procedure it is convenient to express $\tilde{\mathcal{F}}_A$ and $\tilde{\mathcal{F}}_B$ in terms of their eigenvalues and eigenvectors. Indicating with $\vec{v}_i$ and $\vec{w}_i$ the ith eigenvector of the first protein and second protein, respectively (with associated eigenvalues λ_i and μ_i), we have

$$\tilde{\mathbf{F}}_A^{-1} = \sum_l{}' \lambda_l^{-1}\vec{v}_l^\dagger\vec{v}_l \qquad \text{and} \qquad \tilde{\mathbf{F}}_B^{-1} = \sum_l{}' \mu_l^{-1}\vec{w}_l^\dagger\vec{w}_l.$$

Hence, the sum of the diagonal terms of $-\epsilon\tilde{\mathbf{F}}_A^{-1}\tilde{\mathbf{F}}_B^{-1}$ is equal to $\epsilon\sum_{l,m}\lambda_l^{-1}\,\mu_m^{-1}|\vec{v}^l\cdot\vec{w}^m|^2$. In case of perfect correspondence of both sets of ranked eigenvectors and eigenvalues, the previous quantity attains its maximum value, that is $\sum_l \lambda_l^{-1}\,\mu_l^{-1}$.

This observation allows us to introduce a novel normalized measure for the agreement between two dynamical spaces, that we shall term the root weighted square inner product, RWSIP,

$$\mathrm{RWSIP} = \sqrt{\frac{\sum_{l,m}\frac{1}{\lambda_l}\frac{1}{\mu_m}|\vec{v}_l\cdot\vec{w}_m|^2}{\sum_l\frac{1}{\lambda_l\mu_l}}}. \qquad (14)$$

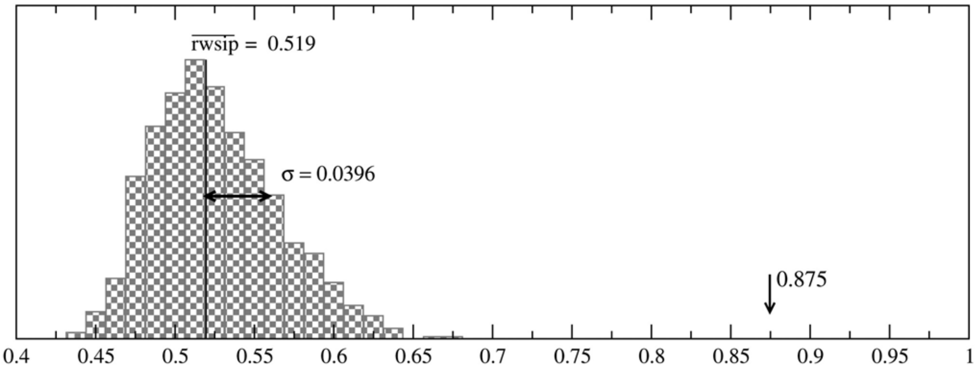

Figure 6. Normalized probability distribution of the RWSIP values computed on the 1000 randomly generated alignments. The arrow indicates the RWSIP of the optimal DALI alignment.

By comparison with the familiar root mean square inner product (RMSIP) expression,

$$\text{RMSIP} = \sqrt{\sum_{l,m=1,\ldots,10} |\vec{v}_l \cdot \vec{w}_m|^2/10}, \tag{15}$$

it can be observed that the RWSIP includes information about the eigenvalues of the effective energy matrix (i.e. the inverse eigenvalues of the covariance matrix). Hence it provides a more stringent and comprehensive account of the degree of accord of two dynamical spaces, and avoids the introduction of a subjective limit to the number of essential eigenvectors to keep. For the two enzymes under consideration the RWSIP of the DALI aligned region was equal to 0.875. To assess the statistical significance of this value we have compared it with a control distribution. The term of comparison is given by the distribution of RWSIP values resulting by randomly choosing the residues in set 1, that is for arbitrary choices of the blocks of corresponding residues in the two structures. Accordingly, we stochastically generated 1000 ‘decoy’ sets of matching residues involving the same number of amino acids (151) as the optimal DALI alignment of carboxypeptidase A and pyroglutamyl peptidase. Also the typical size of DALI matching blocks (10–15 residues) is respected in the control alignments. For each stochastic alignment we carried out the dynamical integration described above numerically, and hence obtained the corresponding RWSIP value from equation (14). By processing the results of the 1000 decoy alignments we calculated the average value and dispersion of the control RWSIP distribution, $\langle\text{RWSIP}\rangle$ and ΔRWSIP. These quantities were used to define the *dynamical Z-score*: $(\text{RWSIP}_{\text{DALI}} - \langle\text{RWSIP}\rangle)/\Delta\text{RWSIP}$. In analogy to the structural Z-score, it provides a measure of how unlikely it is that the RWSIP of the DALI matching regions could have arisen by chance. The value obtained for RWSIP in the DALI alignment was accordingly found to yield a dynamical Z-score of 8.98. The control distribution of RWSIP values is shown in figure 6.

The heuristic approach followed here to derive the RWSIP measure leads also to a transparent criterion for isolating the individual contributions of corresponding amino acids in the protein. In fact, the weighted inner product is WSIP $\propto \sum_i q_i$, where

$$q_i = \sum_{l,m} \frac{1}{\lambda_l} \frac{1}{\mu_m} \vec{v}_l^{\,i} \cdot \vec{w}_m^{\,i} (\vec{v}_l \cdot \vec{w}_m) \Bigg/ \sum_l \frac{1}{\lambda_l \mu_l} \, . \tag{16}$$

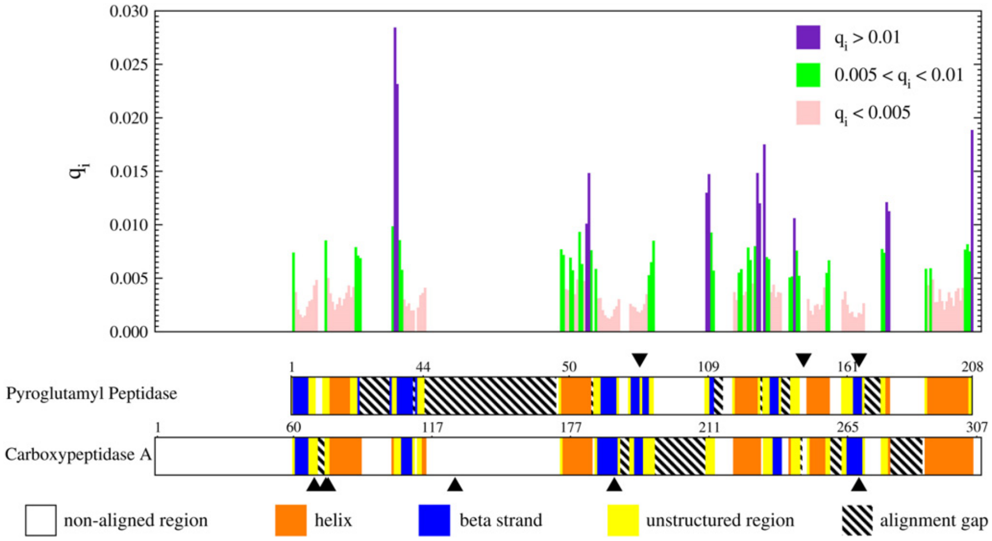

Figure 7. The q_i profile over DALI aligned residues (upper panel) is shown along with the primary alignment induced by the optimal DALI superposition. Different colours are used to partition residues according to the indicated ranges of q_i values.

Hence the q_i provides a qualitative account of the dynamical importance of the residues contributing to the dynamical accord. For the two enzymes under consideration, the non-normalized profile of q_i is illustrated in figure 7. The high inhomogeneity of the profile appropriately complements the information about the structural alignment where all aligned residues are treated on equal footing. Interestingly, the top peaks of the q_i profile involve some of the most mobile residues as well as residues close to the catalytic site. Among the latter set we mention particularly Tyr238 in carboxypeptidase A, whose role has been proven to be crucial for the substrate binding and/or processing [15], and the matching Gly138 of pyroglutamyl peptidase, which is in a crucial position (it surmounts the catalytic cysteine) and contributes to enzyme/substrate interactions [10]. In particular, this Gly is in the exact position of the Glycine of the carboxyanion hole in serine proteases, and this could be suggestive of the fact that the Gly could be directly involved in the catalytic reaction.

The one-dimensional information contained in the q_i profile is finally complemented by figure 8 which illustrates, in Cartesian space, the top two eigenvectors of $\tilde{\mathcal{F}}$ for each enzyme. For graphical simplicity only the top 50 most mobile residues are shown, and their displacement is represented with arrows of equal length. In each of the two enzymes it is possible to identify two halves of the enzymes which undergo rotational fluctuations in opposite directions. The resulting shear motion may consequently produce a mechanical stress of the bound substrate. In the same figure we have highlighted with van der Waals spheres the seven residues (for each structure) where the highest peaks of q_i are observed.

6. Conclusions

We have presented a quantitative scheme through which the essential dynamical spaces of proteins with different fold can be compared. The method relies on the detection of a partial structural correspondences between two biomolecules. An elastic network model is then used to calculate in an efficient way the large-scale concerted movements of the system in

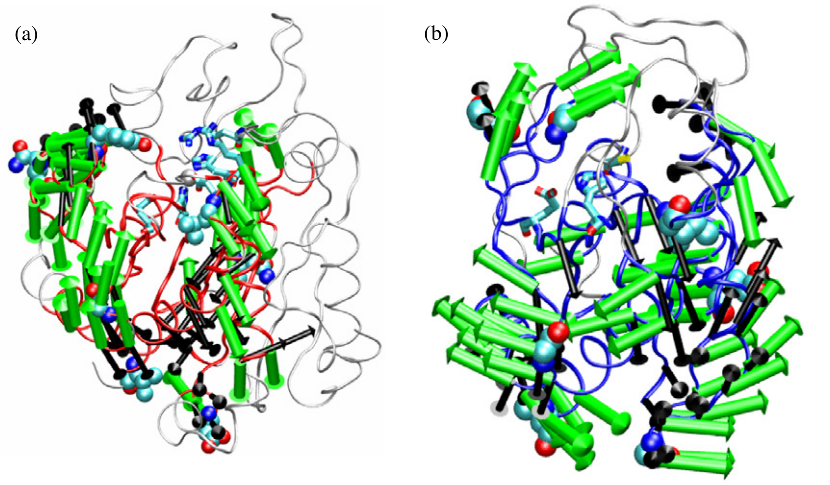

Figure 8. The most mobile residues in the first [second] slowest mode for the *aligned residues* of (a) carboxypeptidase A and (b) pyroglutamyl peptidase are shown with green thick [black thin] arrows. The structurally aligned regions are shown with a thick CA trace.

the frame of reference of the structurally corresponding residues. The consistency of the resulting dynamical spaces is finally measured by means of a novel quantitative criterion that takes into account all slow modes in the system. The dynamical measure introduced leads straightforwardly to identifying the extent to which different residues (all in structural correspondence) contribute to the overall dynamical consistency.

The application of the methodology was illustrated for two representatives of the protease enzymatic superfamily, carboxypeptidase A and pyroglutamyl peptidase. Considering these enzymes is particularly instructive as, besides having different sequence, length and secondary content, they also rely on a different catalytic chemistry. The first feature emerging from the partial structural alignment is that the 151 residues taking part to the optimal superposition are distributed in a region within ∼30 Å of the active site cleft. The slowest modes calculated over this region (which also account for the dynamical influence of the non-aligned residues) turn out to have a dynamical accord with a very high statistical significance. The largest contribution to the 'dynamical alignment' arises from two patches of residues that appear to be capable of modulating the cleft of the active site. Remarkably, the highest dynamical accord close to the active site is observed for one particular residue that is known to be crucial for substrate binding and/or processing in both enzymes. As previously elucidated for other members of the protease enzymatic superfamily [40, 9], the findings provide a strong indication of how the biological selective pressure for efficient cleavage of peptide substrate that may have promoted not only similar structural architectures in the neighbourhood of the active site, but also consistent concerted movements that putatively accompany and facilitate the substrate recognition or cleavage.

Acknowledgments

We are indebted to Paolo Carloni and Andrea Zen for valuable discussions. We acknowledge support from FIRB 2003 (grant RBNE03B8KK) from the Italian Ministry for Education and from Regione Friuli Venezia Giulia (grant Biocheck, 200501977001).

References

[1] Lesk A M 2004 *Introduction to Protein Science: Architecture, Function and Genomics* (Oxford: Oxford University Press)

[2] Fersht A R 1999 *Structure and Mechanism in Proteinscience: a Guide to Enzyme Catalysis and Protein Folding* (New York: Freeman)

[3] Branden C and Tooze J 1991 *Introduction to Protein Structure* (New York: Garland)

[4] Creighton T 1993 *Proteins, Structure and Molecular Properties* 2nd edn (New York: Freeman)

[5] Anfinsen C 1973 *Science* **181** 223–30

[6] Levinthal C 1969 *Mossbauer Spectroscopy in Biological Systems* (Urbana: University of Illinois Press)

[7] Dill K A and Chan H S 1997 *Nat. Struct. Biol.* **4** 10–9

[8] Alexandrov V, Lehnert U, Echols N, Milburn D, Engelman D and Gerstein M 2005 *Protein Sci.* **14** 633–43

[9] Carnevale V, Raugei S, Micheletti C and Carloni P 2006 *J. Am. Chem. Soc.* **128** 9766–72

[10] Barrett A J, Rawlings N D and Woessner J F (ed) 2004 *Handbook of Proteolytic Enzymes* 2nd edn (Amsterdam: Elsevier)

[11] Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J and Orengo C 2005 *Nucleic Acids Res.* **33** D247–51

[12] Tyndall J D, Nall T and Fairlie D P 2005 *Chem. Rev.* **105** 973–99

[13] Kim H and Lipscomb W N 1991 *Biochemistry* **30** 8171–80

[14] Tanaka H, Chinami M, Mizushima T, Ogasahara K, Ota M, Tsukihara T and Yutani K 2001 *J. Biochem. (Tokyo)* **130** 107–18

[15] Cho J H, Kim D H, Lee K J and Choi K Y 2001 *Biochemistry* **40** 10197–203

[16] Singleton M, Isupov M and Littlechild J 1999 *Acta Crystallogr.* D **55** 702–3

[17] Bernstein F C, Koetzle T F, Williams G J, Meyer E E, Brice M D, Rodgers J R, Kennard O, Shimanouchi T and Tasumi M 1977 *J. Mol. Biol.* **112** 535–42

[18] Gunasekaran K, Ma B and Nussinov R 2004 *Proteins Struct. Funct. Bioinf.* **57** 433–43

[19] Garcia-Viloca M, Gao J, Karplus M and Truhlar D G 2004 *Science* **303** 186–95

[20] Micheletti C, Lattanzi G L and Maritan A 2002 *J. Mol. Biol.* **321** 909–21

[21] Tirion M M 1996 *Phys. Rev. Lett.* **77** 1905–8

[22] Micheletti C, Carloni P and Maritan A 2004 *Proteins Struct. Funct. Bioinf.* **55** 635–45

[23] Park B and Levitt M 1996 *Proteins* **258** 367–92

[24] Tirion M M and ben Avraham D 1993 *J. Mol. Biol.* **230** 186–95

[25] Bahar I, Atilgan A R and Erman B 1997 *Folding Des.* **2** 173–81

[26] Hinsen K 1998 *Proteins* **33** 417–29

[27] Micheletti C, Banavar J R and Maritan A 2001 *Phys. Rev. Lett.* **87** 088102

[28] Atilgan A R, Durell S R, Jernigan R L, Demirel M C, Keskin O and Bahar I 2001 *Biophys. J.* **80** 505–15

[29] Delarue M and Sanejouand Y H 2002 *J. Mol. Biol.* **320** 1011–24

[30] Howard J 2001 *Mechanics of Motor Proteins and the Cytoskeleton* (Sunderland, MA: Sinauer Associates)

[31] Doi M 1996 *Introduction to Polymer Physics* 1st edn (Oxford: Clarendon)

[32] Brooks B and Karplus M 1985 *Proc. Natl Acad. Sci. USA* **82** 4995–9

[33] McCammon J A, Gelin B R, Karplus M and Wolynes P G 1976 *Nature* **262** 325–6

[34] Swaminathan S, Ichiye T, vanGusteren W and Karplus M 1982 *Biochemistry* **21** 5230–41

[35] Hinsen K, Petrescu A J, Dellerue S, Bellisent-Funel M C and Kneller G 2000 *Chem. Phys.* **261** 25–37

[36] Chandrasekhar S 1943 *Rev. Mod. Phys.* **15** 1–89

[37] Janezic D, Venable R M and Brooks B R 1995 *J. Comput. Chem.* **16** 1554–6

[38] Noguti T and Go N 1982 *Nature* **296** 776–8

[39] Pontiggia F, Colombo G, Micheletti C and Orland H 2007 *Phys. Rev. Lett.* **98** 048102

[40] Cascella M, Micheletti C, Rothlisberger U and Carloni P 2005 *J. Am. Chem. Soc.* **127** 3734–42

[41] Neri M, Cascella M and Micheletti C 2005 *J. Phys.: Condens. Matter* **17** 1581–93

[42] Holm L and Sander C 1996 *Science* **273** 595–603

[43] Pang A, Arinaminpathy Y, Sansom M S P and Biggi P C 2005 *Proteins* **61** 809

[44] Ming D and Wall M E 2005 *Phys. Rev. Lett.* **95** 198103

[45] Amadei A, Ceruso M A and Nola A Di 1999 *Proteins* **36** 419–24