Eur. Phys. J. B **54**, 131–136 (2006)
DOI: 10.1140/epjb/e2006-00417-1

THE EUROPEAN
PHYSICAL JOURNAL B

Vectorial representation of single- and multi-domain protein folds

F. Teichert and M. Porto[a]

Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstr. 6-8, 64289 Darmstadt, Germany

Received 13 April 2006 / Received in final form 9 October 2006
Published online 8 December 2006 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2006

Abstract. We discuss a vectorial representation applicable to both single- and multi-domain protein folds. This generalized vectorial representation is essentially identical to the previously described vectorial representation for single-domain proteins folds when applied to these, but allows for the additional consistent representation of multi-domain structures. We show that the generalized vectorial representation enables the accurate analytical prediction of site-specific amino acid distributions for both single- and multi-domain protein folds, similarly as the previously described vectorial representation does for single-domain folds.

PACS. 87.14.Ee Proteins – 87.15.Cc Folding and sequence analysis – 87.15.Aa Theory and modeling; computer simulation

1 Introduction

The question of how to properly represent protein sequences and structures has concerned researchers since the first sequences and structures have been determined. It has been realized very early that protein sequences can be represented by various profiles, the most prominent one being the amino acids' hydrophobicity [1,2], but also using other physical and chemical characteristics such as charge and secondary structure propensities. Structural information, on the other hand, can also be reduced to profiles describing structural properties of the amino acids in the fold [3], prominently secondary structure and solvent accessibility [4]. It has been shown that the hydrophobicity profile of a sequence is correlated with the solvent accessibility profile of the native structure [5], indicating that such sequence and structure profiles are interrelated [6,7].

Recently, a vectorial representation of single-domain globular protein folds has been introduced [8], which consists of a real positive number for each amino acid and is derived from the structure's contact map (note that besides such static analysis of the contact network, it might be interesting to resort to a dynamical network analysis similarly as in Ref. [9]). This vectorial representation was found to be related to the sequences attaining that fold via their hydrophobicity profile [10]. Based on this observation, an accurate analytical prediction of site-specific amino acid distributions for single-domain structures using this vectorial representation has been obtained [11–13]. Interestingly, it has also been shown that this structural profile is furthermore equivalent to the protein structure in the sense that the structure's contact matrix can be efficiently recovered from the sole knowledge of the structural profile [8], from which the full three-dimensional structure can be determined [14].

One drawback of the vectorial representation introduced in reference [8] is that multi-domain protein folds cannot be represented in the sense that, for multi-domain folds, the vectorial representation contains only information on the largest or most compact domain, but not on the smaller or less compact ones. Hence, there is the need to generalize the present vectorial representation to allow for the consistent description of both single- and multi-domain protein folds. Such a generalized vectorial representation should fulfill two central conditions: (i) the generalized vectorial representation should give the same vectorial representation as the original definition when applied to single-domain folds; and (ii) the generalized vectorial representation should have the same predictive power for both single- and multi-domain protein folds the original definition has for single-domain proteins, for instance concerning site-specific amino acid distributions. For a vectorial representation being furthermore equivalent to the protein structure, a third condition has to be fulfilled; namely (iii) that it allows to recover the full three-dimensional structure, either directly or via reconstruction of the contact matrix.

In the following, we describe a generalized vectorial representation that fulfills condition (i) in the sense that the previous and the generalized vectorial representation for the set of 404 single-domain protein folds studied in reference [11] have a mean correlation coefficient of 0.96 (so that they are not fully but almost identical) [15]. We furthermore show that condition (ii) is fulfilled, and that the generalized vectorial representation has the same predictive power for both single- and multi-domain protein folds concerning site-specific amino acid distributions the

[a] e-mail: porto@fkp.tu-darmstadt.de

original definition has for single-domain proteins. Whether the additional condition (iii) for an equivalent representation is fulfilled as well has not been ascertained yet and is subject to ongoing research, results will be discussed in a separate publication. However, the compliance of condition (iii) does not hinder immediate application of the generalized vectorial representation, for instance concerning the abovementioned prediction of site-specific amino acid distributions and their use in phylogenetics, which is the central goal of this work.

The paper is organized as follows: in Section 2, we define the generalized structural profile, analyze the resulting profiles for a large representative set of single- and multi-domain structures, and show that condition (i) is fulfilled. The predictive power of the generalized structural profile is exemplified in Section 3, where we predict the site-specific amino acid distribution for the same large representative set of single- and multi-domain structures, in very good agreement with empirically observed distributions, and hence show that condition (ii) is met. The conclusions are summarized in Section 4.

2 Definition of the structural profile

The contact matrix $\mathbf{C}$ of a protein structure of N amino acids is a binary symmetric matrix of size $N \times N$, with elements $C_{ij} = 1$ if amino acids at positions i and j are in contact, and 0 otherwise [16]. Only residues separated by at least three positions along the sequence are considered in contact, so that $C_{ij} = 0$ for all i and j with $|i - j| < 3$. Two residues are considered to be in contact if any two of their heavy atoms (excluding hydrogen) are closer than 4.5 Å in space. Therefore, the contact condition depends on the size of the amino acids at positions i and j. This yields the standard definition of the contact matrix $\mathbf{C}$,

$$C_{ij} = \begin{cases} 0 & \text{for } |i-j| < 3 \\ 1 & \text{if } i \text{ and } j \text{ are in contact} \\ 0 & \text{if } i \text{ and } j \text{ are not in contact.} \end{cases} \qquad (1)$$

Based on this definition, the vectorial representation of protein folds has been defined as the eigenvector $\mathbf{c}$ of the largest eigenvalue of $\mathbf{C}$ (as $\mathbf{C}$ is a real symmetric matrix, it has N real eigenvalues) [8]. This principal eigenvector $\mathbf{c}$ maximizes the quadratic form $\sum_{ij} C_{ij}\, c_i\, c_j$ under the constraint of $\sum_i c_i^2 = \text{const}$. Since all elements of $\mathbf{C}$ are positive or zero, it follows that all components c_i have the same sign or are zero, of which the positive sign is chosen by convention. It should be noted that the principal eigenvector, in case the contact matrix represents a multi-domain protein, contains only information on the largest or most compact domain (a property that has already been used to identify structural domains [17]). For such multi-domain proteins, the principal eigenvector contains non-vanishing components only for the residues belonging to the largest or most-compact domain, and contains zero or vanishing components for the residues not belonging to it. This property limits the applicability of this vectorial representation to single-domain proteins.

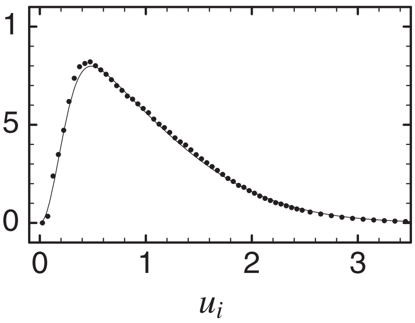

Fig. 1. Histogram $P(u_i)$ of the components u_i for the dataset NR50select. The points show the numerical data, and the line displays the fit, equation (3), with $a = 13$, $b = 0.32$, $c = 0.47$, and $d = 0.115$.

To develop a vectorial representation applicable to both single- and multi-domain protein folds, we start with a slightly modified real symmetric $N \times N$ matrix $\tilde{\mathbf{C}}$, which is defined as

$$\tilde{C}_{ij} = \begin{cases} 0 & \text{for } |i-j| < 3 \\ 1 & \text{if } i \text{ and } j \text{ are in contact} \\ \epsilon(N) & \text{if } i \text{ and } j \text{ are not in contact,} \end{cases} \qquad (2)$$

applying the same definition of contact as above. The difference between $\tilde{\mathbf{C}}$, equation (2), and $\mathbf{C}$, equation (1), is in the elements when i and j are not in contact and $|i-j| \geq 3$. We replace these 0's in equation (1) by a finite length-dependent value $\epsilon(N) = \min\{\epsilon_{\max}, \epsilon_0/[\log(N) - \epsilon_1]\}$, where we use $\epsilon_{\max} = 0.01$, $\epsilon_0 = 0.02$, and $\epsilon_1 = 2$ as parameters. They are chosen such that the value $\epsilon(N)$ is the smallest value for a given sequence length N yielding an eigenvector of the largest eigenvalue which is non-vanishing for all sites and approximately homogeneous over the domains. From the matrix $\tilde{\mathbf{C}}$, we obtain the eigenvector $\tilde{\mathbf{c}}$ of the largest eigenvalue. All components of $\tilde{c}_i$ have the same sign, which we choose to be positive. It is clear that $\tilde{\mathbf{c}}$ maximizes the quadratic form $\sum_{ij} \tilde{C}_{ij}\, \tilde{c}_i\, \tilde{c}_j$, and hence approximately maximizes $\sum_{ij} C_{ij}\, \tilde{c}_i\, \tilde{c}_j$, for $\sum_i \tilde{c}_i^2 = \text{const}$.

In the following, we focus on the dataset NR50 [18], which is a subset of the structures available in the Protein Databank (PDB) [19], where structures of similar sequences have been clustered and ranked. Of this set, we use all structures of rank 1, yielding a representative set of known structures. We exclude from this set of rank 1 structures only those that are clearly not globular, which is verified by enforcing that the number of contacts per residue N_c/N is $N_c/N \geq 2.63 + 5.85N^{-1/3}$, where the factor $N^{-1/3}$ comes from the surface to volume ratio and coefficients are chosen such that essentially all globular structures are correctly identified and non-globular structures are excluded [20]. The reason for this exclusion is that, in most cases, non-globular structures do not attain a well-defined folded structure and are hence already problematic in the contact matrix representation. We remain with 7195 single- and multi-domain globular structures of lengths between 25 and 1491 amino acids, which we refer to in the following as dataset NR50select. For each structure in the set, we construct the matrix $\tilde{\mathbf{C}}$ according to equation (2), from which we obtain the principal eigenvector $\tilde{\mathbf{c}}$. We normalize the vector components as $u_i \equiv \tilde{c}_i/\langle\tilde{c}\rangle$, where $\langle\tilde{c}\rangle = N^{-1}\sum_i \tilde{c}_i$ indicates the average over the components of the given structure, so that $N^{-1}\sum_i u_i = 1$ for each structure, independent of the sequence length N. The histogram of the components for NR50select is shown in Figure 1. This distribution $P(u_i)$ is very well fitted by a functional form

$$P(u_i) = (1 - \exp[-a\,u_i^2])\,\exp[-b\,(u_i + c)^2 + d] \qquad (3)$$

with parameters $a = 13$, $b = 0.32$, $c = 0.47$, and $d = 0.115$.

As the distribution of components of the previous vectorial representation for single-domain proteins has an exponential shape, we transform the components u_i to components v_i which shall obey an exponential distribution. We decompose this transformation into two transformations, of which the first transformation $\mathcal{F}$ transforms the components u_i to the uniform distribution in the unit interval,

$$\begin{aligned}\mathcal{F}(u_i) &= \int_0^{u_i} \left(1 - \exp\left[-ax^2\right]\right)\,\exp\left[-b\,(x + c)^2 + d\right]\,\mathrm{d}x \\ &= \frac{\sqrt{\pi}}{2}\exp(d)\left(\frac{\operatorname{erf}\left(\sqrt{b}(u_i + c)\right) - \operatorname{erf}\left(\sqrt{b}c\right)}{\sqrt{b}}\right. \\ &\quad\left. - \exp\left[-\frac{abc^2}{a+b}\right]\frac{\operatorname{erf}\left(\frac{au_i + b(u_i+c)}{\sqrt{a+b}}\right) - \operatorname{erf}\left(\frac{bc}{\sqrt{a+b}}\right)}{\sqrt{a+b}}\right) \qquad (4)\end{aligned}$$

with erf being the error function. With the fitted parameters (see Fig. 1), one gets with three digits accuracy (reflecting the accuracy of the fitted parameters)

$$\begin{aligned}\mathcal{F}(u_i) = &-0.503 + 1.76\,\operatorname{erf}(0.266 + 0.566\,u_i) \\ &- 0.254\,\operatorname{erf}(0.0412 + 3.65\,u_i). \qquad (5)\end{aligned}$$

Afterwards, we apply a second (inverse) transformation $\mathcal{G}^{-1}$ from the uniform distribution in the unit interval to the intended form

$$P(v_i) = \begin{cases}\alpha\,\exp[-v_i/\lambda] + \beta & \text{for } v_i \le v_{\max} \\ 0 & \text{for } v_i > v_{\max}.\end{cases} \qquad (6)$$

The transformation $\mathcal{G}$ from the distribution $P(v_i)$, equation (6), to the uniform distribution in the unit interval is given by

$$\mathcal{G}(v_i) = \begin{cases}\displaystyle\int_0^{v_i} \alpha\,\exp[-x/\lambda] + \beta\,\mathrm{d}x = \alpha\,\lambda\,(1 - \exp[-v_i/\lambda]) \\ \qquad\qquad\qquad\qquad + \beta\,v_i \quad \text{for } v_i \le v_{\max} \\ \alpha\,\lambda\,(1 - \exp[-v_{\max}/\lambda]) + \beta\,v_{\max} \quad \text{for } v_i > v_{\max}.\end{cases} \qquad (7)$$

From this definition, the transformation $\mathcal{G}^{-1}$ from the uniform distribution in the unit interval to the desired distribution of the components v_i, equation (6), follows by inversion. Even though $\mathcal{G}^{-1}$ can be obtained from equation (7) in closed form, we do not show the very lengthy expression here. Instead, we directly proceed to the form of $\mathcal{G}^{-1}$ that one obtains for the numerical values $\alpha = 0.733$, $\beta = 0.283$, and $\lambda = 0.4$, with three digits accuracy,

$$\mathcal{G}^{-1}(y) = -1.04 + 3.53y + 0.4\ \operatorname{prodlog}(34.5\ \exp[-8.83y]), \qquad (8)$$

where prodlog is the product log or Lambert's W function, so that prodlog(z) is the principal solution for w of $z = w\,\exp(w)$.

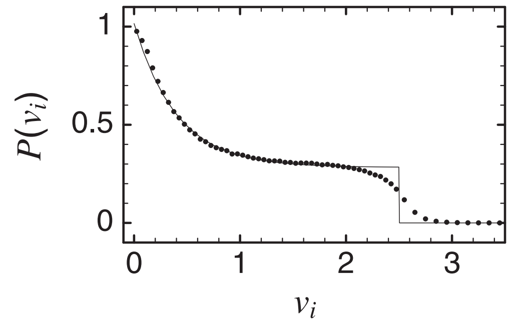

Fig. 2. Histogram $P(v_i)$ of the components v_i for the dataset NR50select. The points show the numerical data, and the line the desired functional form equation (6) with $\alpha = 0.733$, $\beta = 0.283$, and $\lambda = 0.4$.

The final value v_i of the structural profile at residue i is then obtained as $v_i \equiv \tilde{v}_i/\langle\tilde{v}\rangle$ with $\tilde{v}_i \equiv \mathcal{G}^{-1}(\mathcal{F}(\tilde{c}_i/\langle\tilde{c}\rangle))$, where $\tilde{c}_i$ is the ith component of the eigenvector of the largest eigenvalue of $\tilde{\mathbf{C}}$, and $\langle\tilde{c}\rangle$ and $\langle\tilde{v}\rangle$ are the averages of the $\tilde{c}_i$ and $\tilde{v}_i$, respectively, over the given structure. Hence, one has $N^{-1}\sum_i v_i = 1$ for each structure, independent of the sequence length N. The numerically obtained distribution $P(v_i)$ is shown in Figure 2, which follows very well the desired form equation (6), with the exception of the smoothened edge at $v_{\max}$, which is due to enforcing the normalization $N^{-1}\sum_i v_i = 1$.

The correlation coefficient between the original definition of the structural profile (based on diagonalizing the contact matrix in Eq. (1)) and the new definition $\mathbf{v}$ is 0.96 for the set of 404 single-domain globular structures studied in reference [11], so that both definitions are essentially identical for single-domain globular folds [15]. Hence, the generalized structural profile fulfills condition (i) imposed in the Introduction.

3 Properties of the structural profile

In the following, we apply the generalized structural profile defined above to derive site-specific amino acid distributions for the single- and multi-domain structures contained in the dataset NR50select. This shall exemplify that the generalized structural profile has the same predictive power for single- and multi-domain folds the previous structural profile has for single-domain folds. We follow below the discussion in reference [11].

To derive site-specific amino acid distributions, we first define the hydrophobicity profile $\mathbf{h} = \mathbf{h}(\mathbf{A})$ of a given sequence $\mathbf{A}$, where the profile components h_i are given by $h_i = h(A_i)$, and $h(a)$ is the so-called interactivity of amino acid a [10]. The 20 elements $h(a)$, one for each amino acid, are given by the components of the eigenvector for the in absolute value largest eigenvalue $\epsilon_1 < 0$ of the contact interaction matrix $\mathbf{U}$ derived in reference [21]. The values $h(a)$ are strongly correlated with experimentally observed hydrophobicities [10], for instance with the octanol scale derived by Fauchere and Pliska [22]. This is a general property of contact interaction matrices [23,24], as hydrophobicity is one of the major contributions to protein energetics. The parameters $h(a)$, however, should not be interpreted as hydrophobicities in the strict biochemical sense, since they also take into account other kinds of interactions. For instance, aromatic amino acids have very large $h(a)$, in part due to the strength of the interactions between aromatic rings.

The reasoning to define a hydrophobicity profile $\mathbf{h}(\mathbf{A})$ for a sequence $\mathbf{A}$ becomes clear when looking on the free energy of a sequence $\mathbf{A}$ in a structure given by contact matrix $\mathbf{C}$, which in pair-contact-approximation is given by $\sum_{i<j} C_{ij}\, U(A_i, A_j)$. This free energy can be well approximated by $\epsilon_1 \sum_{i<j} C_{ij}\, h(A_i)\, h(A_j)$, since the components $U(a,b)$ and $h(a)\, h(b)$ are strongly correlated, displaying a correlation coefficient of 0.83 for the matrix $\mathbf{U}$ derived in reference [21]. As, on the one hand, wild-type sequences have low free energy in their native configuration so that $\sum_{i<j} C_{ij}\, h(A_i)\, h(A_j)$ is large (note that $\epsilon_1 < 0$) and, on the other hand, $\mathbf{v}$ is an approximation for the vector that maximizes the quadratic form $\sum_{ij} C_{ij}\, v_i\, v_j$ for $\sum_i v_i^2 = \text{const}$, it is not surprising that the hydrophobicity profile $\mathbf{h}$ is correlated with the structural profile $\mathbf{v}$. The mean correlation coefficient for the dataset NR50select is 0.44, which is similar to the value obtained for the previous vectorial representation when analyzing single-domain folds only [11,15]. Note that despite the correlation coefficient being small, it is significant, as the probability to observe such correlation by chance is 10^{-13} for a structure of 250 residues, to mention one example.

For a given structure, one can define an optimal hydrophobicity profile $\mathbf{h}_{\text{opt}}$, for fixed mean and mean square of the hydrophobicity profile $\langle h \rangle$ and $\langle h^2 \rangle$, being the averages of $h(A_i)$ and $h^2(A_i)$, respectively, over the given structure. These values need to be fixed to constrain the energy gap [10,11]. Based on the discussion above, it is clear that this optimal hydrophobicity profile $\mathbf{h}_{\text{opt}}$ is strongly correlated with the structural profile $\mathbf{v}$. The optimal sequence displaying the hydrophobicity profile $\mathbf{h}_{\text{opt}}$ is very unlikely to be realized during evolution. However, all viable sequences have to be sufficiently stable, which implies that their hydrophobicity profile has to have a large correlation coefficient with the optimal hydrophobicity profile $\mathbf{h}_{\text{opt}}$. Thus, the hydrophobicity profiles of protein sequences are expected to move around the optimal hydrophobicity profile in the cause of evolution, so that the evolutionary average $\big[\mathbf{h}\big]_{\text{evol}}$ of the hydrophobicity profile almost coincides with the optimal hydrophobicity profile $\mathbf{h}_{\text{opt}}$, which is in fact observed in simulations of protein evolution [10]. Consequently, the evolutionary average $\big[\mathbf{h}\big]_{\text{evol}}$ is, in turn, strongly correlated with the structural profile $\mathbf{v}$. We follow the approach of reference [11] and assume, for a given fold, a correlation coefficient of 1 between $\big[\mathbf{h}\big]_{\text{evol}}$ and $\mathbf{v}$, yielding

$$\big[h_i\big]_{\text{evol}} \equiv \sum_{\{a\}} \pi_i(a)\, h(a) = A\,(v_i - 1) + B, \qquad (9)$$

where $\pi_i(a)$ is the probability to observe amino acid a at site i, the sum over $\{a\}$ is taken over all amino acids, and

$$A = \sqrt{\frac{\big\langle \big[h\big]^2_{\text{evol}} \big\rangle - \big\langle \big[h\big]_{\text{evol}} \big\rangle^2}{\big\langle v^2 \big\rangle - 1}} \text{ and } B = \big\langle \big[h\big]_{\text{evol}} \big\rangle. \qquad (10)$$

Equations (9, 10) involve two different kinds of averages: The angular brackets $\big\langle f \big\rangle$ again denote the average over the N positions of the protein, whereas square brackets denote position-specific evolutionary averages, $\big[f_i\big]_{\text{evol}} = \sum_{\{a\}} \pi_i(a)\, f(a)$, and consequently $\big\langle \big[f\big]_{\text{evol}} \big\rangle = N^{-1} \sum_i \sum_{\{a\}} \pi_i(a)\, f(a)$. As A and B are given by the mean and mean square of the given profiles' components via equation (10), the above ansatz does not contain any free parameter.

We proceed by assuming that the site-specific distributions $\pi_i(a)$ are the distributions of maximum entropy compatible with the above conditions, so that the only constraint is given by the average, $\sum_{\{a\}} \pi_i(a)\, h(a) = \big[h_i\big]_{\text{evol}}$. The solution to this problem is well-known and is given by exponential or Boltzmann distributions,

$$\pi_i(a) = \frac{\exp[-\beta_i\, h(a)]}{\sum_{\{a'\}} \exp[-\beta_i\, h(a')]}, \qquad (11)$$

with the constraint, equation (9),

$$\sum_{\{a\}} \exp[-\beta_i\, h(a)]\, [h(a) - A\,(v_i - 1) - B] = 0. \qquad (12)$$

Note that equation (11) can be generalized by including weights $w(a)$ for each amino acids, $\pi_i(a) \propto w(a) \exp[-\beta_i\, h(a)]$, which reflect organization of the genetic code, mutational bias, etc., as discussed in detail in references [12,13]. However, as the focus here is the generalized structural profile, we remain with equation (11), following the discussion of reference [11].

The Boltzmann parameter β_i (which can be both negative and positive) can be calculated in an implicit form by rewriting equation (12) as

$$v_i = 1 + A^{-1} \left[\frac{\sum_{\{a\}} h(a)\, \exp[-\beta_i\, h(a)]}{\sum_{\{a\}} \exp[-\beta_i\, h(a)]} - B \right], \qquad (13)$$

giving v_i as a function of β_i, without any free parameter. Hence, for a given fold defined by the structural profile $\mathbf{v}$, the Boltzmann parameter β_i can be obtained for each

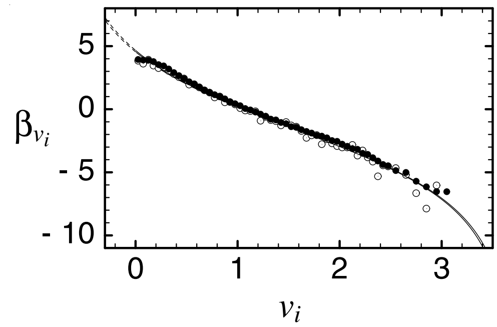

Fig. 3. Amino acid probability distributions, displaying the Boltzmann parameter β_{v_i} as a function of the component v_i of the vectorial representation. Shown are the numerically obtained values for the dataset NR50select (full circles) and for comparison the numerically obtained values for a dataset of 404 single-domain globular folds using the original definition of the structural profile (open circles, replotted from Ref. [11] using $v_i = c_i/\langle c\rangle$). The lines shown are the prediction for the dataset NR50select and for the dataset of reference [11] (the dashed part indicates the 'forbidden' regime $v_i < 0$), which essentially coincide.

site i for given mean and mean square of the hydrophobicity, $\langle h\rangle$ and $\langle h^2\rangle$. The latter two quantities are not determined by the structure, but by the mutation and selection processes (for a detailed discussion of this issue see reference [13]). In the following, we simply obtain $\langle h\rangle$ and $\langle h^2\rangle$ from the analyzed sequences.

To analyze a whole set of structures, it is convenient to perform a structural alignment in the sense that one bins together similar values of v_i. This is possible as the Boltzmann parameter β_i is solely given by the value v_i for fixed mean and mean square hydrophobicity. Hence, we rewrite equation (13) as an equation for β_{v_i} for each whole bin characterized by v_i,

$$v_i = 1 + \tilde{A}^{-1}\left[\frac{\sum_{\{a\}} h(a)\,\exp[-\beta_{v_i}\,h(a)]}{\sum_{\{a\}}\exp[-\beta_{v_i}\,h(a)]} - \tilde{B}\right], \qquad (14)$$

with parameters $\tilde{A}$ and $\tilde{B}$ now given by

$$\tilde{A} = \sqrt{\frac{\langle [h]_{\text{set}}^2\rangle_{v_i} - \langle [h]_{\text{set}}\rangle_{v_i}^2}{[\langle v^2\rangle]_{\text{set}} - 1}} \text{ and } \tilde{B} = \langle [h]_{\text{set}}\rangle_{v_i}. \quad (15)$$

The square brackets $[h]_{\text{set}}$ then denote, instead of the evolutionary average over a protein family, the average over all positions with fixed v_i, even belonging to different structures, whereas angular brackets, $\langle [h]_{\text{set}}\rangle_{v_i}$, denote the average over all values of v_i, weighted by the number of entries in the bin. The denominator $[\langle v^2\rangle]_{\text{set}}$ indicates the quantity $\langle v^2\rangle$, obtained for each structure individually, averaged over the whole set of structures. As $\tilde{A}$ and $\tilde{B}$ are given by equation (15), which in turn can be calculated for a given set of sequence/structure pairs, the above ansatz does not contain any free parameter.

To obtain the numerical approximation for the distributions $\pi_{v_i}(a)$ from the dataset NR50select, we count the number of occurrences of each amino acids as a function of v_i, where we use a bin size of 0.05 for $v_i \leq 2.5$ and a bin size of 0.1 for $v_i > 2.5$. Then, for each bin of v_i, we fit the observed distributions $\pi_{v_i}(a)$ with an exponential function of the hydrophobicity parameters, $\pi_{v_i}(a) \propto \exp[-\beta_{v_i}\,h(a)]$. The values of β_{v_i} obtained by this fit are plotted in Figure 3 and compared with the analytical prediction given by equations (14, 15), yielding a very good agreement without any adjustable parameter. The data shown in reference [11], based on the analysis of single-domain folds using the original definition of the structural profile, is displayed for comparison. As it can be seen in Figure 3, the data obtained by the analysis of the dataset NR50select using the generalized structural profile is in very good agreement with the data obtained in reference [11]. Furthermore, the analytical prediction, equations (14, 15), for both datasets essentially coincide and agree very well with the numerical data. Hence, the generalized structural profile allows for the accurate prediction of site-specific amino acid distributions for both single- and multi-domain folds, similarly as the previous structural profile does for single-domain folds.

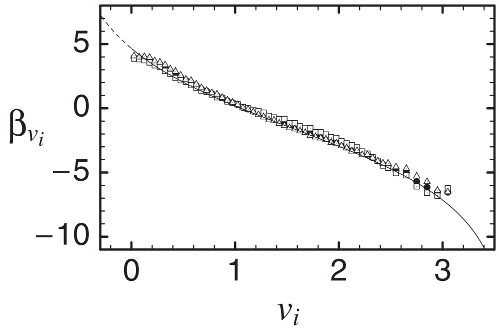

Fig. 4. Amino acid probability distributions, displaying the Boltzmann parameter β_{v_i} as a function of the component v_i of the vectorial representation. Shown are the numerically obtained values for the full dataset NR50select (full circles), as well as separately for the two disjoint subsets containing only single-domain (open squares) and only multi-domain folds (open triangles). The line shown is the prediction for the full dataset NR50select (the dashed part indicates the 'forbidden' regime $v_i < 0$).

To verify whether there is a difference between single- and multi-domain folds concerning the quality of prediction, we split the set of 7195 structures of the dataset NR50select into two disjoint subsets, one containing 4330 single-domain folds and one containing 2865 multi-domain folds. The discrimination is done using the Protein Domain Parser (PDP) [25]. The above analysis is repeated for each of the two subsets separately, yielding the values of β_{v_i} individually for single- and multi-domain folds in the dataset NR50select. The results are shown in Figure 4 and compared with the results of the whole dataset NR50select. There is a very good agreement between single- and multi-domain folds, indicating that the underlying generalized structural profile describes single- and multi-domain folds in an analogous manner. These two comparisons, one with the previous results obtained using the previous definition of the structural profile (Fig. 3), and one between single- and multi-domain folds (Fig. 4), exemplify that the generalized structural profile has the same predictive power for both single- and multi-domain folds the original definition has for single-domain folds, and hence fulfills condition (ii) imposed in the Introduction.

The prediction in equations (14, 15) can be further improved by considering the effects of mutation and selection in more detail as done in reference [13] for single-domain structures using the original definition of the structural profile. Such treatment allows to derive site-specific amino acid distributions taking into account a given mutational model. This generalized ansatz including both mutation and selection can be solved on a mean-field level, so that one obtains a protein evolution model with independent sites that reproduces site-specific amino acid distributions [13]. The generalized structural profile defined here is directly applicable within this more generalized context.

4 Conclusions

We introduce a generalized structural profile applicable to both single- and multi-domain proteins. This generalized structural profile is essentially identical to the original definition for single-domain proteins when applied to these (correlation coefficient of 0.96 for the set of 404 proteins discussed in Ref. [11]). Furthermore, this generalized structural profile has the same predictive power for single- and multi-domain protein folds the original definition has for single-domain proteins. This is exemplified by predicting the site-specific amino acid distributions of the dataset NR50select, which is a representative set of known single- and multi-domain structures. Such site-specific amino acid distributions are very helpful for many applications, for instance in phylogenetics, as they allow to derive a protein evolution model with independent sites that reproduces site-specific amino acid distributions [13]. To facilitate such applications, a web site has been set up [26].

Inspiring discussions with Ugo Bastolla, H. Eduardo Roman, and Michele Vendruscolo, as well as generous financial support by the Deutsche Forschungsgemeinschaft (PO 1025/1-1) are gratefully acknowledged.

References

1. J. Kyte, R.F. Doolittle, J. Mol. Biol. **157**, 105 (1982)
2. R.M. Sweet, D. Eisenberg, J. Mol. Biol. **171**, 479 (1983)
3. J.U. Bowie, R. Lüthy, D. Eisenberg, Science **253**, 164 (1991)
4. B. Rost, C. Sander, Proteins **23**, 295 (1995)
5. J.U. Bowie, N.D. Clarke, C.O. Pabo, R.T. Sauer, Proteins **7**, 257 (1990)
6. M. Wilmanns, D. Eisenberg, Proc. Natl. Acad. Sci. USA **90**, 1379 (1993)
7. E.S. Huang, S. Subbiah, M. Levitt, J. Mol. Biol. **252**, 709 (1995)
8. M. Porto, U. Bastolla, H.E. Roman, M. Vendruscolo, Phys. Rev. Lett. **92**, 218101 (2004)
9. K.A. Eriksen, I. Simonsen, S. Maslov, K. Sneppen, Phys. Rev. Lett. **90**, 148701 (2003)
10. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, Proteins **58**, 22 (2005)
11. M. Porto, H.E. Roman, M. Vendruscolo, U. Bastolla, Mol. Biol. Evol. **22**, 630 (2005); Mol. Biol. Evol. **22**, 1156 (2005)
12. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, Gene **247**, 219 (2005)
13. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, BMC Evol. Biol. **6**, 43 (2006)
14. M. Vendruscolo, E. Kussell, E. Domany, Fold. & Des. **2**, 295 (1997)
15. In reference [11], the condition of being single-domain was imposed by requiring that the relative variance $(\sum_i [c_i - \langle c \rangle]^2)/(N\langle c \rangle^2) = (1 - N\langle c \rangle^2)/(N\langle c \rangle^2)$ of the components of $\mathbf{c}$ was smaller than 1.5. Note that this condition does not strictly exclude multi-domain structures if the components c_i are relatively homogenous over the domains, which is for instance true when the domains are well connected within the structure's heavy atoms contact matrix, equation (1). Such multi-domain structures behave like single-domain structures as far as the original definition of the structural profile is concerned. According to the Protein Domain Parser (PDP) [25], the set of 404 folds used in reference [11] contains 45 such multi-domain structures
16. In the following, we use bold face symbols such as $\mathbf{c}$ and $\mathbf{C}$ to indicate vectors and matrices, whereas c_i and C_{ij} indicate individual components
17. L. Holm, C. Sander, Proteins **19**, 256 (1994)
18. The list of structures of the dataset NR50 can be obtained from the Protein Databank (PDB) site at `ftp://ftp.rcsb.org/pub/pdb/derived_data/NR/`. We use the list of October 4, 2005 and consider from each cluster the structure of rank 1
19. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Research **28**, 235 (2000)
20. Note that we apply a somewhat less stringent exclusion rule than in references [10–13]
21. U. Bastolla, J. Farwer, E.W. Knapp, M. Vendruscolo, Proteins **44**, 79 (2001)
22. J.L. Fauchere, V. Pliska, Eur. J. Med. Chem. **18**, 369 (1983)
23. G. Casari, M.J. Sippl, J. Mol. Biol. **224**, 725 (1992)
24. H. Li, C. Tang, N.S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997)
25. N. Alexandrov, I. Shindyalov, Bioinformatics **19**, 429 (2003)
26. The structural profile, as defined in this paper, can be downloaded from the SLOTH homepage `http://www.fkp.tu-darmstadt.de/sloth/` for each structure of the PDB