Supplementary Materials

1. Supplementary method descriptions

1.1 Augmenting the structural prototypes of protein fragments with local environment information

We consider a fragment comprising *l* contiguous residues as a single entity. We represent the local geometric and environmental features of fragment *a* by a coding vector of *1+2l* components,

$$\mathbf{C}^a = \{C^a_{SP}, C^a_{SS,1} C^a_{SS,2}, ..., C^a_{SS,l}, C^a_{SA,1}, C^a_{SA,2}, ..., C^a_{SA,l}\}. \quad (1)$$

The first component, C^a_{SP}, encodes the structure prototype of the fragment. In this study we consider fragments comprising 5 contiguous residues and the value of C^a_{SP} will be one of the 16 structure prototypes (protein blocks) defined based on backbone torsional angles by de Brevern et al (de Brevern, et al., 2000; Etchebest, et al., 2005).

The next *l* components, $\{C^a_{SS,1} C^a_{SS,2}, ..., C^a_{SS,l}\}$, encode the secondary structure states. For each residue, we consider three possible secondary structure states, helix, coil and strand. The $C^a_{SS,i}$ encodes the secondary structure state of the *i*th residue of the fragment as a three-dimensional vectors, with possible values (0,0,1), (0,1,0) and (1,0,0) for helix, coil and strand, respectively.

The last *l* components of the coding vector, $\{C^a_{SA,1}, C^a_{SA,2}, ..., C^a_{SA,l}\}$, encode the solvent accessibility, with $C^a_{SA,i}$ encoding the relative solvent accessibility of the side chain of the *i*th residue of the fragment in the native protein structure, also as a three dimensional vector, the sub-components of which taking real-number values within [0,1] computed as

$$C^a_{SA,i,1-3} = \begin{cases} (0, \delta - \delta * \dfrac{SA_0 - SA}{SA_0}, \delta + \delta * \dfrac{SA_0 - SA}{SA_0}) & \text{if} \quad SA \le SA_0, \\ (\delta - \delta * \dfrac{SA_1 - SA}{SA_1 - SA_0}, \delta, \delta * \dfrac{SA_1 - SA}{SA_1 - SA_0}) & \text{if} \quad SA_0 < SA < SA_1 \text{ and} \\ (\delta + \delta * \dfrac{SA - SA_1}{1 - SA_1}, \delta - \delta * \dfrac{SA - SA_1}{1 - SA_1}, 0) & \text{if} \quad SA \ge SA_1. \end{cases} \quad (2)$$

SA is the relative solvent accessible areas of residue i within [0,1], SA_0=0.09 separates the complete buried states from the intermediately buried states, and SA_1=0.36 separates the intermediately buried states from the exposed stated. The $C^a_{SA,i}$ would be (0,0,1), (0,0.5,0.5), (0.5,0.5,0) and (1,0,0) when SA takes values of 0, SA_0, SA_1, and 1, respectively.

1.2. Measuring similarity between protein fragments using the coding vectors.

Given two fragments *a* and *b* with their respective coding vectors $\mathbf{C}^a$ and $\mathbf{C}^b$, we use the

following function to score their overall similarity,

$$S(\mathbf{C}^a, \mathbf{C}^b) = \delta_{c^a_{SP} c^a_{SP}} (w_{ss} S^{ab}_{SS} + w_{SA} S^{ab}_{SA}) . \quad (3)$$

The Kronecker δ function is 1 if *a* and *b* are of the same structure prototype (PB) and 0 otherwise. S^{ab}_{SS} and S^{ab}_{SA} measure the similarity of the secondary structure components and solvent accessibility components, respectively, of *a* and *b*. Here we compute them as the Pearson correlation coefficients between the corresponding components of the coding vectors. For example,

$$S^{ab}_{SS} = \frac{\sum_{i=1}^{l}\sum_{j=1}^{3}(C^a_{SS,i,j} - \bar{C}^a_{SS}) * (C^b_{SS,i,j} - \bar{C}^b_{SS})}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{3}(C^a_{SS,i,j} - \bar{C}^a_{SS})^2} \sqrt{\sum_{i=1}^{l}\sum_{j=1}^{3}(C^b_{SS,i,j} - \bar{C}^b_{SS})^2}} . \quad (4)$$

Here $\bar{C}^a_{SS}$ and $\bar{C}^b_{SS}$ stand for averages over the three subcomponents for each position and then over the *l* positions for fragments *a* and *b*, respectively. The weighting parameters w_{SA} *and* w_{SS} in equation (3) determine respectively the relative contributions of the secondary structure components and solvent accessibility components (see below for how they have been determined).

1.3 Clustering protein fragments in the space of the coding vectors.

We apply the K-means algorithm(MacQueen, 1967) to cluster protein fragments in the space spanned by their coding vectors based on the similarity measure described above. Given a set of data points and a metric for similarity between them, the K-means algorithm partitions the set into user-specified number of clusters, so that each cluster contains data points as similar to each other as possible while the similarity between points in different clusters is minimized. This is achieved through the local minimization of a pseudo energy function which depends on both the within-cluster similarity and the between-cluster dissimilarity. There are two consequences of practical implications for such an approach. One is that the number of clusters is an input parameter for rather than a result of the clustering process, thus we can choose this parameter so that certain extra properties of the clusters are optimum. The other is that the local minimization process, usually started from random initial guesses, cannot be guaranteed to reach the global minimum of the pseudo energy function. Heuristically the clustering can be repeated with different initial guesses and the "best" solution is accepted.

Usually, the aggregation of clustered data points in the coding space is employed to measure the effectiveness of clustering. Such a metric, however, do not suit our purpose, which is to bridge the space coding structure and environmental of the fragments with the sequence space, rather than to aggregate fragments in the coding space itself. It is thus appropriate to use the similarity of the structurally and environmentally clustered fragments in the sequence space to select optimum clustering solutions.

Given the size of the dataset to be partitioned into clusters, the numbers of data points

contained in individual clusters change in inverse proportion to the total number of clusters. Changing the number of data points in a structure cluster has dual effects on our efforts to extract sequence preferences of its members. On one hand, a larger cluster size implies better statistics for the sequence preferences of the members of the cluster. On the other hand, larger clusters may also imply wider spreading of the members in both the structure and the sequence spaces and the specificity of the extracted preferences decreases. To balance between these two effects, we define a cluster size-independent measurement of sequence biases as criteria for selections of the clustering parameters.

Given *m* fragments forming a cluster in the structure and environmental coding space, we can compute their similarity in the sequence space. This similarity can be compared with the sequence similarity between randomly-selected *m* fragments from the entire data set. The probability that the randomly chosen fragments aggregates better in the sequence space than the clustered fragments is used as a measure of the sequence bias of the clustered fragments. The lower this probability, the stronger the sequence bias of the clustered fragments. As there are the same numbers of structurally clustered and randomly chosen fragments, the size-dependence of the computed similarity in the sequence space is compensated for. As the sequence biases at different positions of the fragments can be very different, we measure the sequence similarity of each position separately. Consider cluster *k* containing m_k fragments, we use the following $c_i(k)$ to measure the similarity of residues at position *i* of different members of the cluster,

$$c_i(k) = \frac{2}{m_k(m_k-1)} \sum_{\substack{a,b \in cluster\ k \\ a<b}}^{m_k} \sum \sigma(\alpha_i^a, \alpha_i^b), \tag{5}$$

in which α_i^a and α_i^b are the amino acid residue types at position i of fragments a and b, respectively. The $\sigma(\alpha_i^a, \alpha_i^b)$ is the corresponding element in the BLOSUM90 amino acid substitution matrix(Henikoff and Henikoff, 1996; Henikoff and Henikoff, 1992). To define the cluster size-independent sequence bias at position *i*, we use

$$\begin{aligned} b(k) &= \sum_{i=1}^{l} 1 - \theta[P(c_0(m_k) > c_i(k)], \\ \theta(x) &= \begin{cases} 1, x > p_0 \\ 0, x \le p0 \end{cases} \end{aligned} \tag{6}$$

in which $c_0(m)$ is the similarity score of a randomly chosen set of *m* residues as measured in equation (3), and P represents probability. Equation (6) measures effectively the number of positions at which similar amino acid residues are significantly preferred (above the confidence level $(1-p_0)\times 100\%$, where p_0 equal 0.1) in member fragments of cluster *k*. For a given partitioning of an entire dataset of fragments into *N* clusters, we compute

$$b_{total} = \frac{\sum_{k=1}^{N} m_k b(k)}{\sum_{k=1}^{N} m_k} \tag{7}$$

as a measure of the effectiveness of the partitioning for extracting sequence preferences.

As the K-means computations are expensive for larger datasets, we first used a dataset comprising all 5-residue fragments in 72 proteins to explore different coding schemes and similarity measures based on the b_{total} value. These proteins have been selected randomly from a dataset containing 482 protein chains (see below). For each explored schemes and parameters multiple (>50) K-means runs were performed and the averaged b_{total} values were considered. The ratio between the weighting factors w_{SS} and w_{SA} has been systematically changed between 10:1 to 1:10, with the additional choices of 0:1 and 1:0 considered. For each chosen w_{SS} :w_{SA} ratio the total number of k-means clusters was systematically varied from 5 to 500, Comparing the resulting b_{total} values an approximate range for the w_{SS} :w_{SA} ratio between 1:10 to 1:1 was determined. Fixing w_{SA} to 1.0 and w_{ss} was systematically changed between 0.1 to 1.0 in steps of 0.1, for each of the ratio, the dataset containing 482 proteins were clustered with varying number of k-means clusters. Again by comparing the resulting b_{total} values an optimum w_{SS} :w_{SA} ratio and the corresponding optimum number of clusters for each PB type determined. In Table S8 the averaged b_{total} values resulted from different w_{SS} :w_{SA} ratios are listed, in which w_{SS} =0.4 with w_{SA} fixed at 1.0 is the optimum. Figure S1 shows as an example how the b_{total} value changes with the number of k-means clusters (with w_{SS} =0.4 and w_{SA} fixed at 1.0) for PB type C.

1.4. Modeling the sequence preferences of clustered fragments

We can describe the sequence preferences of fragments in cluster *k* by the following conditional probability, $p(\alpha_1, \alpha_2 ..., \alpha_l \mid k)$. Given any reasonable database size, the number of members contained the clusters would not allow for any meaningful direct estimation of this completely jointed distribution. Thus we approximate the distribution using the single site marginal distributions and the correlations between two sites,

$$p(\alpha_1, \alpha_2 ..., \alpha_l \mid k) \propto \prod_{i=1}^{l} p(\alpha_i \mid k) * (\prod_{i=2}^{l} \prod_{j=1}^{i-1} \frac{p(\alpha_i, \alpha_j \mid k)}{p(\alpha_i \mid k) p(\alpha_j \mid k)})^{\zeta} \tag{8}$$

The conditional probabilities $p(\alpha_i \mid k)$, and $p(\alpha_i, \alpha_j \mid k)$ are estimated from the amino acid sequences of the fragments contained in the cluster, and $\zeta = 0.2$ for *l*=5. The inter-site coupling will be used only in the local structure prediction tests. In the sequence prediction tests (see below) we will treat the distributions at different sites as mutually independent (ζ =0).

The member fragments contained in a cluster can be considered as a sample of the sequence

distribution of all fragments in similar shape and local environments. The size of the sample usually does not allow for accurate estimation of the probability of rare sequences even with approximations in equation (8). We used pseudocounts to partly compensate for the effects of this insufficient sampling,

$$p(\alpha_i = \alpha \mid k) \approx \frac{n_k(\alpha_i = \alpha) + b_{c\alpha}}{N_k + B_c} \tag{9}$$

in which N_k is the size of the sample or the number of members in cluster *k*, $n_k(\alpha_i = \alpha)$ is the actual count of residue α at position i, B_c is the total number of pseudocounts, and b_{ca} is the number of pseudocounts for residue α at position *i* based on certain a prior distributions. The following choices of B_c and b_{ca} were found to perform well(Henikoff and Henikoff, 1996; Marti-Renom, et al., 2004; Sadreyev and Grishin, 2004; Sunyaev, et al., 1999),

$$\begin{aligned} b_{ca} &= B_c * \mathrm{p}(\alpha | \mathrm{SS}, SA) \\ B_c &= 5 * R_c \end{aligned} \tag{10}$$

in which R_c is the total number of different residue types observed at a given position, and *p(α|SS,SA)* is the amino acid distribution at positions with given secondary structure and solvent accessibility state, estimated directly using the entire data set.

The two-site joint probability $p(\alpha_i, \alpha_j \mid k)$ in equation (8) have also been estimated using pseudocounts, with the total number of pseudocounts $B_c = \sqrt{N_k}$ and the pseudocounts for residue pair (α_i, α_j) proportional to the corresponding joint probability in all training fragments belonging to the PB associated with cluster k.

1.5. Local protein structure predictions by a hidden Markov model.

The simplest scheme to predict local structures using the above model is to compare directly the likelihoods of different structure types (equation (8)) given the sequence of the *l* contiguous residues, ignoring the influences of surrounding residues and their preferred structures. In such a model the strong correlations between the shapes of neighboring or overlapping fragments would have been ignored. We considered several different approaches to take such correlations into account and to generate consistent predictions for contiguous fragments. The first is a hidden Markov model (HMM(Rabiner and Juang, 1986)) , in which the joint structural and environmental cluster a fragment may belong to is considered as a hidden state, and the sequence of the fragment are the observed values. The emission probabilities for different sequences are modeled by equation (8). The distribution of the initial hidden states and the transition probabilities between neighboring hidden states can be estimated directly from a database of training protein structures. Given the complete sequence of a protein (the sequence of the observed states), the standard forward-backward algorithm can be applied to compute the distribution of hidden states for any fragment. From the distribution the probability for a fragment to be of certain structure prototype

(protein block) can be computed. We compared different choices (from 1 to *l-1* residues along the sequence) for the frame shift between neighboring hidden states and they produced similar prediction results. The results from a frame shift of one residue will be reported here.

1.6. Another scoring scheme for local structure prediction by sequence –structure database matching

This scheme is based on the distribution of central fragment structure prototypes among the N top-scoring template windows. The integrated score for structure prototype SP is given by

$$\sigma_{N-top}(SP) = \frac{N_{top}(SP)}{N_{top\_total}}, \tag{11}$$

in which $N_{top}(SP)$ is the number of template windows whose central fragments are of structure prototype SP among the N_{top_total} top-scoring template windows for the given target window. The predicted structure prototype for the target fragment would be the prototype associated with the highest score defined in equation (11). The results are not very sensitive to the exact values of N_{top_total} between several hundred to one or two thousand. And the reported results have been obtained using N_{top_total}=1000.

1.7. Datasets, cross tests and control calculations

Both the probabilistic sequence preference models in equation (8) and the sequence-structure database matching method for local structure prediction depend on a set of training or template data comprised of proteins of known structures. We considered two datasets of different sizes for training and testing the methods. The small dataset contains 482 complete, non-homologous peptide chains (Table S1) with known structures selected by de Brevern et al (http://condor.ebgm.jussieu.fr/~debrevern/DOWN). The proteins contained a total number of 117377 overlapping, five-residue long fragments. The large dataset comprises 1462 complete chains (Table S2) contained in PDB-REPRDB (Noguchi and Akiyama, 2003). These 1462 proteins are not homologous with pair-wise sequence identities <20%, chain lengths > 40, structures determined by X-ray crystallography at resolutions < 2.0 Å and with R-factors < 0.3. The total number of fragments contained in these proteins is 375511. The secondary structure state of each residue in both datasets has been computed using STRIDE(Frishman and Argos, 1995) and the relative solvent accessibility computed using the Lee and Richards algorithm(Lee and Richards, 1971).

For each datasets, we performed 6-fold cross tests for both sequence preference predictions and local structure prediction. In the tests the dataset was randomly partitioned into 6 groups of protein chains. When chains in one group were used as test data, chains in the remaining five groups were used not only to derive the probabilistic sequence preference models, but also to construct the template database for sequence-structure matching. The tests covered the entire dataset after each of the 6 groups has been used as test data in turn.

To investigate the effects of augmenting the fragment shape information by solvent accessibilities, we performed the following control calculations on the large dataset. Probabilistic sequence preference models have been built separately (i) without any partitioning of fragments of each PB into clusters (i.e., one cluster for each PB type, noted as "PB only"), (ii) with partitioning based on only the PB types and secondary structure states (w_{SS}=1.0 and w_{SA}=0.0 in equation (3), noted as "PB+SS"), and (iii) with partitioning based on only the PB types and the solvent accessibility states (w_{SS}=0.0 and w_{SA}=1.0 in equation (3), noted as "PB+SA"). For (i) we consider the additional case in which only a single 5-residue long fragment is included for local structure prediction (m=0 in equation 12, noted as "1PB only". For control models (ii) and (iii) we partition each PB into the same number of clusters as used for the default case (both secondary structure and solvent accessibility states are included, noted as "PB+SS+SA" when necessary). All control calculations have been preformed using the σ_{total} scoring scheme and the 6-fold cross test scheme described above.

2.Supplementary results and discussions

2.1. Clustering by K-means.

Number of clusters for each PB type. Figure S1 shows how the b_{total} score defined in equation (7) depends on the number of K-means clusters, using PB type c as an example. Although given the number of clusters, there are significant variations in the b_{total} scores generated by different K-means runs started from different initial guesses, the overall dependence relation indicates a discernible minimum number of clusters above which the b_{total} score no longer increases with the number of clusters. Similar dependencies have been observed for other PB types. From these relations an optimum number of clusters for each PB type has been determined and listed in Table S3. In total all training fragments of different PBs form 300 clusters. While the sizes of the clusters (the numbers of fragments contained in individual clusters) are not uniform, ranging between a few tens to several thousands of fragments, most clusters contained one to several hundred fragments for the small dataset. The optimum numbers of clusters for individual PB types are obviously not proportional to the occurring frequencies of the PBs in native protein structures. In their original work de Brevern et al have reported these frequencies (de Brevern, et al., 2000; Etchebest, et al., 2005). Similar frequencies were observed in both datasets used in this work (Table S3 listed for the small dataset the number of fragments belonging to different PBs).

Variations in secondary structures and in solvent accessibility within PBs and within clusters. To some extent the number of clusters for each PB reflects the variations in the secondary structure and solvent accessibility states of different fragments contained in the same PB. Figure S2a compared the within-PB and within-cluster distributions of the three secondary structure states at individual sites. Figure S2b shows similar distributions of the three solvent accessibility states. PBs and secondary structure states are both descriptors of local shapes and thus strongly correlated. Despite of this some of the within-cluster secondary structure distributions still deviate significantly from the within-PB distributions. The within-PB distributions of the solvent

accessibility state for most sites do not indicate strong preferences of any particular state, although there are some general trends that residues contained in PBs corresponding to coil regions are more likely to be exposed than residues in regular secondary structures. Thus fragments of similar shapes do locate in different environments in native proteins. As expected, the within-cluster distributions are more homogeneous: the same sites on fragments within the same cluster are more likely to be in the same secondary structure and solvent accessibility state than sites on fragments belonging to the same PB. It is interesting to note that that for a given site, the within-cluster distributions of the secondary structure or environment states can be completely different from the within-PB distributions. For example, the within PB distributions suggest that the first site of PB type a are more likely to be exposed to solvent, while for fragments within some clusters associated with this PB type their first sites are dominantly in buried environments.

Specificity of within-PB and within-cluster sequence preferences. Intuitively, fragments within the same PB but located in different environment may have different sequence preferences, and amino acid preferences learned from PBs separated into clusters should be more specific than those learned from PBs. We calculated the respective relative entropies of the within-PB and the within-cluster amino acid distributions with the background amino acid distribution. Larger relative entropies imply larger deviations from the background distributions, or more specific and strongly biased amino acid preferences. The relative entropies computed for each PB type clustered by different criteria and the overall averaged values are listed in Table S4. For the within-PB distributions the relative entropy averaged over all sites of all PBs is 0.0051. The averaged relative entropy for distributions within fragments clustered by PB and the secondary structure states is 0.0070. The value increases to 0.0102 if the fragments have been clustered by PB and the solvent accessibility states, and 0.0105 if both secondary structure and solvent accessibility states have been considered together with PB to cluster the fragments. The increases in the relative entropy are distributed to almost all PB types, although not evenly. One exception is PB type j, for which the within-PB distributions result in the largest relative entropy (0.0179), indicating very specific sequence preferences of fragments having local shapes represented by this PB. The relative entropies for the within-cluster distributions associated with this PB type are slightly lower, probably because that for clusters of smaller size, the contributions of the background distributions through pseudo counts are larger.

2.2. Comparing different control calculations for sequence design

For comparisons we report sequence design results from the control calculations. Besides control calculations which considered PB but left out either the secondary structures (“PB+SA”) or the solvent accessibilities (“PB+SS”) or both (“PB only”), additional control calculations which only considered the secondary structure and solvent accessibility of single sites either separately (“SS” and “SA”, respectively) or jointly (“SS+SA”) without representing the containing fragments as PBs have also been performed.

Table 1 in the main text shows that the control model “PB+SA” and the default model

"PB+SS+SA" predict best the native residues. The averaged ratios between the predicted and background probabilities are above 1.3. The median values of these ratios are above 1.4, or for 50% of the sites the predicted probabilities of the native residues are more than 1.4 times of the corresponding background probabilities. These averaged and median ratios are significantly larger than those obtained using other control models (averaged values 1.05~1.17 and median values 1.07~1.15). Probably due to information redundancy between PB types and secondary structure states (both describe the local shape), the "PB+SS" model does not bring about as much improvement as the "PB+SA" model over the "PB only model". Even the "SS+SA" model predicts better sequence preferences than the "PB+SS" model, although it is not as informative as the "PB+SA" model.

If we consider the number of sites for which the native residues are predicted to be significantly preferred, joint considerations of both local shape and environment outperform models considering only the shape or environment by even larger margins. For more than 45% (28%) of the sites the predicted probabilities by the "PB+SA" or "PB+SS+SA" model are more than 1.5 (2.0) times of the corresponding background probabilities. Same percentages produced by the "PB only" or "SS" or "SA" only models are significantly lower (24-38% for the 1.5 probability ratio threshold and 8-13% for the 2.0 threshold). Interestingly, when we consider whether the native residue is predicted to be the most or among the top two or three most preferred residues, the "PB only" model is as good as the "SS+SA" model, although the models combining local shape represented as PB type with environment still perform the best.

2.3. Local structure predictions from sequences

Comparisons between different scoring schemes in sequence-structure database matching. The σ_{total} scoring scheme outperforms the σ_{N-top} scheme in terms of Q_{14}, although in terms of Q_{16} the σ_{N-top} scheme produced better results (Table S5). The σ_{total} scheme produced more balanced predictions of regular and non-regular secondary structure PBs. Comparisons of the overall SLR values for the 14 non-regular secondary structure PBs indicate that the higher Q_{16} rates of the σ_{N-top} scoring scheme are associated with over-predictions of PBs m and d which are associated with regular secondary structures, especially for the small dataset. Thus in what follows by the sequence-structure matching results we will refer to those obtained with the σ_{total} scheme, although we cannot exclude that using datasets larger than those used here the σ_{N-top} scheme could also produce balanced predictions and catch up with or even exceed the σ_{total} scheme in terms of Q_{14} (Table S5) shows that going for the small to the large dataset, the Q_{14} associated with σ_{N-top} increased from 24.6 to 31.0 without loss in the corresponding SLR),

Comparisons between sequence-structure database matching and the first-order HMM. In general, the two methods produced similar predictions. The HMM however results in lower Q_{14} for the small dataset and both lower Q_{14} and Q_{16} for the larger dataset (Table S5). For most of the

PB types the HMM model also does not perform as well as the sequence-structure database matching method and will not be discussed further.

Comparing the sequence-structure database matching strategy with the "new sequence family" method. We will first compare the "1PB only" and "PB only" control models with the "new sequence family" method of de Brevern et al, as all these models derive sequence preferences from local shapes of peptide fragments represented as PBs without considering environments of the PBs. Hence the differences may be attributed mostly to the different representations of sequence preferences (the "sequence family" concept is used only in the "new sequence family" model but not in the "PB only" model) and the different prediction strategies.

The results of the "1PB only" and "PB only" control calculations are given in Table S6 and in Table S7. Going from considering 5 residue windows each containing a single PB (the "1PB only" model) to considering 9 residue windows each containing 5 overlapping PBs (the "PB only model"), the Q_{14} increases from 32.2% for 33.9% and Q_{16} from 39.3% to 41.8%.

We applied the "new sequence family" method of de Brevern et al to each of the 482 test proteins in the small dataset. We believe the total number of test fragments (more than one hundred thousand) contained in this dataset is already large enough to obtain accurate estimates for the various prediction rates (this model has been trained using a necessarily fixed training set to achieve optimum results and dataset size dependences like that of the sequence-structure database matching method are not expected). The prediction results are summarized as overall Q_{14} and Q_{16} rates in Table 2 and as SNRs and SLRs for individual PBs in Table 3. The overall Q_{14} and Q_{16} rates are 30.7% and 43.6%, respectively. Compared with the "PB only" control model (Table S6) the Q_{14} is 3.2 percentage points lower but the Q_{16} is 1.8 percentage points higher.

Comparing the SNRs and SLRs of the "new sequence family" method (Table 3) with those of the "PB only" control model in (Table S7 in supplementary) indicate that for most PB types, including a, c, e, h, i, and p, the comparison fit the third scenario defined above, i.e., the SNR and SLR for the same PB changed in inversed directions and neither method is definitely favored. Comparisons for PB types d and m fit the first scenario in favor of the "new sequence family" method, in consistence with its higher Q_{16}. Comparisons for PB types n and o fit the first scenario in slight but definite favor of the sequence-structure database matching approach. Comparisons for PB types f, k, l fit the second scenario also in slight favor of the sequence-structure database matching approach. Interestingly, the predictions of PB types b, g, and i by the "new sequence family" method are much more sensitive than the sequence-structure database matching approach. Although the SLRs are in favor of the sequence-structure database matching methods, the comparisons for these PB types fit the second scenario in favor of the "new sequence family" method. However, for the PB types g and i, the SNRR/SLR ratios produced by the "new sequence family" model are significantly above 1.0 (1.7 for g and 3.8 for j), indicating significant over-predictions of these PBs. On the contrary, the sequence-structure matching approach under-predicts these three PB types.

We note that de Brevern et al have developed the "new sequence family" method based on a training set containing 425 chains and a test set containing 250 chains(Etchebest, et al., 2005). For this particular test set they obtained much higher Q_{14} and Q_{16} rates (37.4% and 48.7%, respectively). We have applied the sequence-structure database matching method using the same training and test sets, the resulting Q_{14} and Q_{16} rates are 36.1% and 46.6%, respectively, both significantly higher than what we obtained using 6-fold cross tests on either the small or the large dataset. This indicated that local structure prediction results obtained using the particular training and test data of reference (Etchebest, et al., 2005) may not be generalized, probably due to unwanted similarities between training and testing data. In fact de Brevern et al ((Etchebest, et al., 2005) have specifically pointed out that their inclusions of data into the training and test sets were necessarily not random, and had been adjusted so that the resulting model achieves similar prediction performances on both sets.

Effects of including secondary structures and solvent accessibility. Further clustering of local structures defined as PBs by the secondary structure states of individual residues improved the Q_{14} from 34.0% to 35.7% (Table S6). This is, however, accompanied with a decrease in Q_{16} from 42.9% to 41.7%. If comparisons in terms of individual PB types are made (Table S7), only a few PBs belong to the first (PB e) or the second (PBs b, c, and p) scenarios described earlier in favor of the "PB+SS" model. Among them, the SNR for PB type b increased significantly. For PB type m, the SNR decreased from 58.1% to 49.8% in the "PB+SS" model although the accuracy rate increased from 65.9% to 69.5%. This may be the major contributor to the decrease in Q_{16}. Thus augmenting the structure alphabet defined as PBs by secondary structures does not bring about general improvements for the prediction of local structures as PBs. This is probably because of the strong correlations between secondary structure states of individual residues and the PB type of the containing fragments.

Including solvent accessibility states have definitively positive effects on the prediction results. Compared with the "PB only" model, the Q_{14} for the "PB+SA" model increased from 34.0% to 35.2%, and Q_{16} from 42.9% to 45.0% (Table S6). The SLR rate for the 14 non-regular secondary structure states also increased from 30.5% to 32.1%. When results for individual PBs are compared (Table S7), most of them belongs to either the first (PB types c, d, e, f, i, j, k, l, m, o and p) or the second (PB types a, b and n) scenarios in favor of the "PB+SA" model, remaining results for PB types g and h belonging to the third scenario which is not in definite favor of either model if both SNR and SLR are considered. The largest improvements in SNR are for PB types b and d. For PB types a, i, k, n, o, p the SLR rates increase by more than 2 percentage points. Results from the default "PB+SS+SA" model are very close to the "PB+SA" model in terms of both the overall accuracies and performances for individual PB types, with very small increases in Q_{14}, Q_{16}, and SLR for the non-regular secondary structures. In later discussions we will focus on results of the "PB+SS+SA" model (the default model) unless stated otherwise.

Comparing the "PB+SS+SA" sequence-structure database matching model with the new

sequence family method. When the results of our default model are compared with the "new sequence family" model, Q_{14} increased from 30.7% to 35.6% and Q_{16} from 43.6% to 45.3% (Table 2). The SLR for the 14 non-regular secondary structure states also increased from 29.6% to 32.3%. When individual PBs are compared (Table 3), five PB types (c, d, f, i, n and o) can be assigned to the first scenario in favor of our default model. The largest improvements are in the SNRs for PB types c (from 28.2% to 33.6%), d (from 42.7% to 48.5%) and i (from 33.5% to 41.0%) at no cost of the respective SLRs. The improvements for f and n are mainly in the SLRs (from 31.3% to 35.3% for f and 31.0% to 35.6% for n) accompanied also by slight increases in the respective SNRs. For six PB types (a, e, h, k, l, p) the results can be assigned to the second scenario also in favor of our default model. Especially for PB types e, k and p there are significant increases in the SNRs (more than 10 percentage points) accompanied with relatively small drops in the respective SLR rates. Result for no PB type can be assigned to the first scenario in favor of the "new sequence family" model. However, augmenting the PB by solvent accessible states does not improve enough the sensitivity rates obtained with the sequence-structure database matching method for PB types b, g, and j to make them comparable to predictions made by the "new sequence family" method. The under-predictions of these PB types remain in contrast to the over-predictions of them by the "new sequence family" method.

2.4. Implications on local sequence to structure relations

The results reported here confirmed that there exist strong correlations between local structure/environment and sequence, although such correlations should be understood in a complex, probabilistic, and environment dependent multiple sequence types-to-multiple local structure prototypes mapping sense rather than with a simple, deterministic, environment-free one sequence type-to-one local structure type mapping picture.

The local conformation combined with the local environment put stronger constraints on the sequence than either element alone. On average close to 20% of the native residues correspond to residues predicted to be most preferred by the local shape and environment of the containing fragments. And the ratio for strongly preferred (with a predicted probability 1.5 times of the background probability) native residues is 45%.

In the other direction, protein local structures depend strongly on the local sequence. On average 45% of the five-residue fragments contained in globular proteins can be predicted from sequence to have the most preferred local structure type among the 16 possible PB types, and such fragments cover 75% of all residues. The ratio of fragments adopting not necessarily the most but strongly preferred local structures are even higher, c.a. 62% and 72% respectively with local structure types among the predicted top two and top three most preferred types.

The success rates of deriving sequences from local structures and of the reversed predictions vary greatly among different proteins. Figures S3 and S4(in supplementary)shows the distributions of the per-protein success rates among the 1462 proteins for the "prediction" of native sequences from PB and environment types and for the prediction of PB types from sequences. In different

proteins, the ratio of native residues predictable from local structures and environments varied between a few percent to more than 30%, and the ratio of fragments with local structures uniquely predictable from local sequences varied between 20% to 70%. Do these variations reflect different weights of local sequence-structure dependences in the global sequence-structure relationships of different proteins, or do they simply reflect biases in the computational models (that is, the local sequence-structure dependences captured by the model apply better to some proteins than others)?

To explore this issue, we divided the 1462 protein chains into three groups according to the per-protein Q_{16} rates: the top 30% protein chains with the highest Q_{16} rates, the bottom 30% protein chains with the lowest Q_{16} rates, and the remaining 40% protein chains with Q_{16} rates in between. The local structure prediction accuracy-Δ_{12} relations were recomputed for each protein group. Figure 1 shows that different protein groups generated exactly the same relation between prediction accuracy and Δ_{12}. This strongly suggests that the local sequence-structure relations captured by our model apply equally to different protein groups. The protein-independence of the prediction accuracy- Δ_{12} relation also strongly suggests that Δ_{12} may be associated with some physical meaning. A larger Δ_{12} may implicate a wider "free energy gap" between the most preferred local structure state and other possible state, thus stronger local structure preferences of the respective sequence segment. This in turn implicates a larger probability for the segment to adopt the preferred local structure when integrated into the native structure of a complete peptide chain. The lower success rates for some proteins are probably due to that they contained more local sequence segments which do not have strong preferences for unique local structures, possibly because these proteins rely less on such preferences but more on longer range interactions to fold into native three dimensional structures. Should this be the case, the accuracy of deterministic, unique local structure predictions would be intrinsically limited, and predictions ranking different local structures probabilistically would be more natural.

3. Supplementary Tables

Table S1. PDB codes of the 482 protein chains forming the small dataset**.**

153l 1a12A 1a1x 1a2pA 1a2zA 1a3aA 1a3h 1a44 1a4iB 1a4uA 1a76 1a7uA 1a8e 1a8h 1a8l 1a8p 1aew 1af7 1afwB 1agjA 1ah7 1ahc 1ai3 1ai9A 1aj8A 1ajz 1ako 1amm 1amp 1aocA 1apyA 1apyB 1aq0A 1aqb 1arb 1aru 1atlA 1axn 1ayoA 1ayx 1b00A 1b1cA 1b2pA 1b4kA 1b5eA 1b5qB 1b71A 1b8aA 1b8pA 1b94A 1b9hA 1bbpA 1bd8 1bea 1bf2 1bf6B 1bfd 1bgf 1bgvA 1bhe 1bhtA 1bj7 1bjwA 1bkpB 1bkzA 1bn8A 1bolA 1bqk 1bslB 1bsmA 1btl 1btn 1bu7A 1buoA 1bxaA 1byfB 19gsA 1byqA 1bz0A 1bzyA 1c02A 1c1fA 1c1kA 1c2pA 1c3kA 1c3qA 1c44A 1c8kA 1c8uA 1ca1 1cczA 1celA 1cem 1cewI 1cfb 1chd 1chmA 1cjcA 1cmbA 1cnzA 1cozA 1cp2A 1cpn 1cq3A 1cqxA 1crzA 1cs6A 1css 1cv8 1cvrA 1cz1A 1czfA 1cznA 1cztA 1d0bA 1d0qA 1d2oA 1d2vA 1d2vC 1d3sA 1d4oA 1d6oA 1d8wA 1d9cA 1dbfA 1dciA 1dd3A 1ddvA 1dfx 1dgwA 1dhn 1dixA 1dj0A 1dk0A 1dk8A 1dlwA 1dmhA 1dmr 1doi 1dorA 1dowA 1dozA 1dp4A 1dqeA 1dqgA 1dqiA 1dqtA 1dqzA 1ds0A 1dsbA 1dts 1dugA 1dupA 1dusA 1duwA 1dxeA 1dxy 1dysA 1dytA 1dz3A 1dzfA 1e0cA 1e15A 1e19A 1e29A 1e2uA 1e3aA 1e3uB 1e5mA 1e6oL 1e6qM 1e6uA 1e87A 1ecsA 1edg 1edqA 1edt 1ee8A 1eejA 1eg9A 1eg9B 1eguA 1ej2A 1ejbA 1ejdA 1ejjA 1ek0A 1ekgA 1el4A 1el6A 1emvB 1eo6B 1eo9A 1eo9B 1eokA 1ep0A 1eq6A 1erzA 1esgB 1esl 1eswA 1eu3A 1eu8A 1euaA 1euhA 1eur 1evxA 1ew0A 1ew4A 1ew6A 1ex2A 1extA 1ey0A 1eyhA 1eyqA 1eyvB 1ez3A 1f00I 1f08A 1f0kA 1f2dA 1f2tB 1f2uA 1f32A 1f39A 1f5mB 1f5vA 1f5wA 1f6kA 1f7sA 1f8mA 1f9zA 1fc3A 1fc9A 1fd7D 1fgyA 1fi2A 1fit 1fj2A 1fkmA 1fl2A 1flmA 1flp 1fn9A 1fnc 1fp2A 1fs7A 1ft5A 1ftrA 1fua 1fupA 1fus 1fvaB 1fzqA 1g0sB 1g12A 1g13A 1g1bA 1g1kA 1g291 1g3qA 1g5tA 1g6sA 1g72A 1g73A 1g73B 1g73D 1g8lA 1gakA 1gcuA 1gd0A 1gd1O 1gefA 1gg6B 1ggxA 1gia 1gnd 1gof 1gp1A 1gpeA 1gpr 1h2rL 1h2rS 1hcz 1he7A 1hf8A 1hfc 1hhsA 1hruA 1hsbA 1htrB 1i0dA 1i0rB 1i39A 1i6pA 1iab 1iakA 1iakB 1iazA 1icjA 1ido 1igs 1ihgA 1io7A 1jbc 1jfrA 1jkmB 1kdj 1koe 1kpf 1lam 1lenC 1lib 1lki 1ltsA 1mba 1mgtA 1mkaA 1mla 1mml 1mugA 1muyA 1nah 1nar 1nbaA 1nbcA 1nkr 1nlr 1nox 1npk 1nseA 1nsf 1nsj 1nsyA 1nwpA 1nzyA 1obwA 1onrA 1pamA 1pbn 1pbv 1pbwA 1pdo 1phnA 1php 1pmi 1pnkB 1poa 1ppn 1pprM 1prn 1puc 1pud 1qazA 1qb8A 1qccA 1qcxA 1qd9A 1qgiA 1qh4A 1qh5A 1qhqA 1qhvA 1qi7A 1qjdA 1qk8A 1qksA 1qnrA 1qnxA 1qqjA 1qsaA 1qstA 1qtoA 1qtsA 1qu1F 1regX 1rhs 1rl6A 1rmg 1rom 1rpjA 1rro 1sacA 1seiA 1sftA 1skf 1sll 1smlA 1sra 1srvA 1stmA 1sur 1svb 1svy 1tca 1tf4A 1tfe 1thfD 1thm 1thv 1tib 1tkiA 1tl2A 1tpfA 1trkA 1ttqB 1udh 1uok 1uroA 1vcaA 1vfrA 1vid 1vls 1vpnB 1vsd 1vsrA 1wab 1wdcC 1wgtA 1whi 1xer 1xgsA 1xib 1xnb 1xsoA 1xwl 1yacA 1yge 1zin 1zpdA 256bA 2abh 2abk 2baa 2bbkH 2bbkL 2cba 2cpl 2ctb 2e2c 2end 2fcbA 2gdm 2hrvA 2hvm 2i1b 2lisA 2mcm 2mnr 2nacA 2pgd 2pia 2pii 2plc 2por 2pth 2rn2 2scpA 2sil 2spcA 2tgi 2tlxA 2tnfA 3chy 3cyr 3daaA 3grs 3hsc 3lzm 3mbp 3pah 3pte 3sdhA 3stdA 3thiA 3vub 3wrp 4fgf 4pgaA 4uagA 5nll 7nn9

Table S2. PDB codes of the 1462 protein chains forming the large dataset.

153l 1a12A 1a1iA 1a1x 1a34A 1a3aA 1a4iB 1a6m 1a73A 1a76 1a8d 1a8l 1a8s 1adeA 1af7 1agi 1ah7 1aj2 1ajsA
1ak0 1ako 1al3 1amm 1amx 1aocA 1apyB 1arb 1b0yA 1b2pA 1b3aA 1b5pA 1b5qB 1b8oA 1b94A 1b9hA 1bd0A
1bea 1behA 1bf2 1bgf 1bgvA 1bhe 1bif 1bm9A 1bn6A 1bn8A 1brfA 1bslB 1byrA 1c02A 1c0pA 1c1dA 1c1kA
1c1yB 1c30B 1c39A 1c52 1c5kA 1c7cA 1c7nA 1c7sA 1c8kA 1c8uA 1c96A 1c9oA 1cc8A 1ccwB 1cczA 1cdcA
1cdy 1cfb 1chd 1chmA 1cjcA 1cl8A 1cmbA 1cmlA 1cpn 1cq3A 1cqxA 1cruB 1cs6A 1csh 1csn 1cuoA 1cv8
1cvrA 1cz9A 1czfA 1czpA 1czqA 1cztA 1d02B 1d0cA 1d0qA 1d2oA 1d2vA 1d2zD 1d4oA 1d4tA 1d9cA 1dad
1dbfA 1dciA 1dd3A 1dgwA 1dj0A 1dk0A 1dk8A 1dljA 1dmhA 1dmr 1dozA 1dp4A 1dpgA 1dqaA 1dqgA 1dqzA
1ds1A 1dvoA 1dzfA 1dzkA 1e19A 1e29A 1e2wA 1e3uB 1e59A 1eaoA 1eb6A 1ecfB 1edg 1edt 1ee8A 1eejA
1eerB 1eexA 1eexB 1eg3A 1ekjG 1el5A 1el6A 1elkA 1eokA 1epfB 1es5A 1esgB 1esjA 1eu3A 1euvA 1evhA
1ew0A 1ew4A 1ex2A 1ex7A 1exrA 1extA 1ey4A 1eyqA 1ezgA 1ezjA 1ezwA 1f08A 1f0kA 1f1mA 1f1uA 1f20A
1f24A 1f2dA 1f2tB 1f2uA 1f39B 1f3uA 1f46B 1f5mB 1f60B 1f74A 1f86A 1fbqB 1fc9A 1fczA 1fil 1fiuA 1fk5A
1fl0A 1fl2A 1flmA 1fn9A 1fnf 1fp2A 1fp3A 1fr2B 1fs7A 1fsoA 1ft5A 1ftrA 1furA 1fxlA 1fxoB 1fy7A 1g0sB
1g12A 1g1tA 1g2qA 1g38A 1g5tA 1g61A 1g66A 1g6gA 1g6sA 1g6xA 1g73A 1g85A 1g8aA 1g8eA 1g8kA
1g8kB 1g8lA 1g97A 1g9gA 1g9zA 1ga6A 1gakA 1gk2A 1gk9A 1gl4A 1glg 1gmuC 1gnlA 1gnuA 1gnyA 1gof
1gotB 1gp0A 1gpeA 1gpiA 1gpr 1gpuA 1gq8A 1gqiA 1gqnA 1gqvA 1gqyB 1gs5A 1gs9A 1gsa 1gtzA 1gu2A
1gu7A 1gudA 1guiA 1guqA 1gv4A 1gvp 1gwmA 1gwyA 1gx0A 1gxjB 1gxmB 1gxqA 1gxuA 1gxyA 1gy6A
1h0aA 1h16A 1h1yA 1h2cA 1h2wA 1h4gB 1h4pA 1h4yA 1h6fB 1h6hA 1h6lA 1h6tA 1h7eA 1h7wD 1h8eH
1h8pA 1h8xA 1h9sB 1hbnB 1hbnC 1hbzA 1hd2A 1hdoA 1he1A 1he7A 1hf8A 1hh8A 1hq0A 1hqkA 1hs6A
1ht6A 1htwA 1hufA 1hw5A 1hyoB 1hz4A 1i0dA 1i0rB 1i1dD 1i1nA 1i2aA 1i39A 1i4mA 1i4uA 1i5gA 1i5rA
1i6mA 1i7nA 1i8dA 1i8oA 1i9gA 1ia9B 1iab 1iakA 1iapA 1ib2A 1ibyA 1ic6A 1iccA 1idpA 1ifc 1ifgA 1ifrA
1ig0B 1ig3A 1ihrB 1ijhA 1ijyA 1iktA 1inlC 1io0A 1io1A 1iow 1ipbA 1iq4A 1iq6B 1iqzA 1ituA 1itvA 1iu8A
1iv3A 1iv9A 1ivuA 1iw0A 1iwmA 1ixbA 1iz6C 1izcA 1j09A 1j0pA 1j1bB 1j1nA 1j1tA 1j23A 1j2rA 1j30A
1j33A 1j34B 1j3wB 1j8qA 1j8uA 1j9jA 1ja1A 1jakA 1jayA 1jb3A 1jb7A 1jb7B 1jbc 1jcdA 1jdw 1jetA 1jf2A
1jf8A 1jflA 1jfrA 1jfuB 1jfxA 1jg9A 1jh6A 1jhfA 1jhjA 1jhsA 1ji1A 1jidA 1jixA 1jkeC 1jkxA 1jl1A 1jnrA 1jnrB
1josA 1jovA 1jpeA 1jr2B 1jr8A 1jsrA 1ju2A 1jubA 1juvA 1jx6A 1jyeA 1jyhA 1jyoE 1k04A 1k0iA 1k0mB
1k1eA 1k2eA 1k32A 1k3yA 1k4gA 1k4iA 1k7cA 1k7hA 1k7iA 1k94A 1kafD 1kblA 1kbqA 1kcmA 1kdgB 1keiA
1kg2A 1khiA 1kjlA 1kl1A 1km4A 1kncA 1knlA 1koe 1kolA 1kpf 1kphB 1kq3A 1kqfB 1kqfC 1kqpA 1krhA
1ks8A 1kt6A 1ku0A 1kufA 1kwgA 1kyfA 1kzkB 1kzqA 1l2tA 1l3jA 1l6rA 1l6wA 1l6xA 1l7aA 1l8fA 1l9lA
1l9xA 1lam 1lc0A 1lgpA 1lh0A 1lhuA 1lj9A 1lk2A 1lk2B 1lk5A 1lki 1lkoA 1llfA 1ln4A 1lniB 1lo7A 1loeB
1lovA 1lplA 1lq9A 1lqbC 1lqvB 1lr5A 1lriA 1lslA 1luaA 1lwbA 1lwdA 1lxkA 1lxzA 1ly2A 1lyqB 1lyvA 1lzlA
1m15A 1m1hA 1m1nA 1m1nB 1m2aB 1m2kA 1m2tB 1m3kA 1m3uA 1m4iA 1m4jA 1m55A 1m5wA 1m70A
1m9xC 1m9zA 1mbmA 1mbyA 1mdl 1mf7A 1mgtA 1mho 1mixA 1mk4A 1mkaA 1mkkA 1ml4A 1mla 1mml
1mopA 1mqdA 1mqvA 1msc 1msk 1mtpA 1muwA 1mw7A 1mxrA 1n08B 1n0qA 1n0sA 1n13B 1n13E 1n2sA
1n2zA 1n5uA 1n62A 1n62B 1n67A 1n7fA 1n7sC 1n8fA 1n8kA 1na0A 1nar 1nawA 1nbcA 1nbuA 1nc5A 1ne9A
1nepA 1ng2A 1ng6A 1nijA 1nkgA 1nkiA 1nkr 1nlnA 1nm8A 1nofA 1nogA 1nox 1np3A 1np6B 1nrzA 1nsf
1nsxB 1ntvA 1ntyA 1nv0A 1nvmB 1nvmG 1nwaA 1nwzA 1nxjA 1nxmA 1nycA 1nykA 1nytA 1o08A 1o0sA
1o0wB 1o26B 1o4yA 1o5uA 1o6sB 1o7iA 1o7nB 1o98A 1o9iA 1oa8D 1oaoC 1obnA 1ockA 1ocyA 1oflA 1ofzA
1oh0B 1oh4A 1oi6B 1oioA 1oizA 1okoA 1okqA 1olrA 1omrA 1on3E 1onrA 1oo0A 1ooeA 1oohA 1oqqA 1oqvA
1orb 1orsC 1orvA 1osyB 1otkB 1owlA 1oxjA 1oz2A 1oz9A 1p1jA 1p1xA 1p36A 1p4kA 1p4oB 1p6oB 1p71A
1p99A 1pbyA 1pbyB 1pcfB 1pdo 1pi1A 1pii 1pjcA 1pk6A 1pkhA 1pl3A 1pm4A 1pmi 1pnf 1po5A 1poc 1pot
1pp0B 1pt7A 1pvmB 1pwbA 1pwmA 1px4A 1pxrA 1pxzA 1pzxB 1q08A 1q0nA 1q0qA 1q0rA 1q16A 1q1cA
1q4gA 1q5zA 1q6zA 1q7fB 1q7lA 1q7lB 1q7zA 1q8fA 1q92A 1q98A 1qazA 1qb0A 1qb5D 1qb7A 1qcxA 1qd1B
1qd9A 1qdvA 1qftA 1qgiA 1qh5A 1qhdA 1qhoA 1qipB 1qj8A 1qjcA 1qksA 1ql0A 1qlmA 1qmgA 1qmyA 1qnrA

1qnxA 1qoyA 1qq4A 1qqfA 1qqmA 1qr0A 1qs1A 1qsaA 1qtnA 1qtwA 1qu1F 1qveB 1qw9A 1qwnA 1qwzA 1qxmA 1qxrA 1qz4A 1qz9A 1r0vA 1r1hA 1r1mA 1r1tB 1r29A 1r3sA 1r45A 1r5zA 1r6dA 1r6jA 1r6xA 1r7jA 1r89A 1r8nA 1r9dA 1r9lA 1ra0A 1ra9 1rfs 1rg8A 1ri6A 1rifA 1rjdC 1rkd 1rkiA 1rkuA 1rl6A 1rmg 1ro0A 1roaA 1rocA 1rp0A 1rqpA 1rtqA 1rttA 1ru4A 1rv9A 1rw7A 1rwyA 1rwzA 1ry9A 1ryaA 1ryiA 1ryoA 1s0pA 1s1dA 1s3cA 1s3eB 1s55A 1s68A 1s6aA 1s7iA 1s95B 1s9rA 1sa3A 1sacA 1sauA 1sdwA 1se0A 1seiA 1sf9A 1sfsA 1sg4C 1sh8A 1sixA 1sjwA 1sk7A 1skz 1sqwA 1sr4A 1sr4C 1sra 1su8A 1sulB 1sur 1svb 1svdA 1svmC 1svpA 1svsA 1sw5A 1sxrA 1szhA 1t06A 1t0bH 1t0fA 1t0iA 1t0pB 1t2dA 1t3cA 1t3mA 1t4bA 1t5oA 1t61D 1t6cA 1t6nA 1t7rA 1ta3A 1ta8A 1tbfA 1tca 1te5A 1ten 1tf1A 1tfe 1tg7A 1th7A 1thzA 1ti6B 1tjlA 1tjyA 1tkeA 1tl2A 1tl9A 1toaA 1tr0A 1tt8A 1tu1A 1tu9A 1tuaA 1tuhA 1tuvA 1tv4A 1tvfA 1twbA 1twiA 1tx4A 1txgA 1txlA 1ty9B 1tyjA 1tzcA 1tzvA 1tzyA 1tzyB 1tzyG 1u07A 1u11B 1u1iA 1u1qA 1u4bA 1u55A 1u5dA 1u5hA 1u5pA 1u5uA 1u8vA 1ua4A 1uaiA 1uasA 1ub4C 1ucdA 1ucrB 1uebA 1uehB 1uekA 1uf5A 1ug6A 1ugqA 1ugxA 1ui0A 1uj8A 1ujcA 1ukuA 1ulkA 1um0A 1umhA 1umzA 1unnD 1unqA 1uok 1uowA 1uoyA 1uq5A 1ursA 1us6B 1uscA 1usgA 1uslC 1uv4A 1uvjA 1uwcA 1uwfA 1uwkB 1uxzA 1uy2A 1uylA 1v00A 1v0eA 1v2xA 1v30A 1v33A 1v3eA 1v4pA 1v58A 1v5dA 1v5vA 1v6pA 1v6sA 1v70A 1v71A 1v73A 1v74A 1v77A 1v7bA 1v7lA 1v7zA 1v8cA 1v8eA 1v8hA 1v96B 1v9fA 1vbkA 1vcaA 1vclA 1vemA 1vf8A 1vfjA 1vflA 1vfqA 1vh5A 1vhh 1vl9A 1vls 1vpsB 1vsrA 1vyrA 1vziA 1w0dA 1w0nA 1w0pA 1w1hC 1w2dA 1w4sA 1w4xA 1w5fA 1w5mA 1w5rA 1w66A 1w96C 1wab 1wbhB 1wc3A 1wckA 1wcwA 1wd3A 1wdcC 1wdjA 1wdpA 1wdyA 1wer 1wf3A 1wg8A 1whi 1wkuB 1wkvB 1wleB 1wlgA 1wmwA 1wmxB 1wn2A 1wn5A 1wnyA 1woqA 1wp4C 1wpbA 1wpnA 1wq3A 1wqaA 1wqwB 1wrrA 1wrvB 1wteA 1wtjB 1wu6A 1wu9B 1wubA 1wurA 1wvfA 1ww7A 1wwzA 1wxcA 1wz8A 1wzcB 1x0pF 1x0tA 1x1nA 1x1oA 1x2jA 1x38A 1x54A 1x6iB 1x7yB 1x91A 1xawA 1xb9A 1xc1A 1xeoA 1xffA 1xfkA 1xg4A 1xkrA 1xo0A 1xovA 1xp2A 1xq6A 1xqhA 1xrjB 1xs0C 1xszA 1xtaA 1xttA 1xubA 1xuuA 1xw3A 1xwwA 1xx1A 1y0pA 1y1pA 1y1tA 1y2tA 1y43B 1y4mA 1y50A 1y60D 1y65A 1y6iA 1y6vA 1y7bA 1y7pB 1y8aA 1y93A 1y9gA 1y9wA 1y9zB 1yacA 1yarA 1yb0B 1yb6A 1ybqA 1ydgE 1ydyA 1yfqA 1ygaA 1yge 1yiiA 1yiyA 1ykdA 1ymiA 1ymqA 1yn3A 1yn9B 1ynhB 1ynsA 1ypqB 1ypyA 1yqsA 1yqzA 1yrkA 1yroB 1yt3A 1ytbA 1ytqA 1yu0A 1yukB 1yumA 1yvxA 1yw5A 1ywmA 1yxyB 1yzxA 1z2wA 1z4rA 1z6nA 1z9nA 1zaiA 1zavA 1zb1A 1zc3B 1zcjA 1zd8A 1zdyA 1zhsA 1zhxA 1zi8A 1zjcA 1zjyA 1zkkA 1zkrA 1zl0B 1zo4B 1zodA 1zowA 1zpdA 1zpsB 1zs4C 1zsqA 1zvaA 1zvtB 1zvzA 1zx0C 1zxiC 1zxxA 1zz1A 1zzgA 2a14A 2a1kA 2a21A 2a26A 2a38A 2a38B 2a50B 2a6zA 2a7bA 2a7lA 2a9dB 2ab0A 2abh 2acfD 2acvA 2ad6A 2aebA 2aenB 2aeuA 2aexA 2ag4A 2ahfA 2airB 2ajgA 2akaA 2al1A 2amdA 2aorA 2apcA 2aqjA 2ar1A 2arzA 2asbA 2asdA 2askA 2au3A 2au7A 2avdA 2avkA 2avwA 2awgA 2axqA 2axwA 2aydA 2az4B 2b0jA 2b0pA 2b0tA 2b2hA 2b3fA 2b3sA 2b4hA 2b4lA 2b4pB 2b5hA 2b7kB 2b82A 2b8tA 2b97A 2ba2A 2ba9A 2bb6A 2bemA 2bfwA 2bgxA 2bibA 2bivA 2bj0A 2bjfA 2bjiA 2bjkA 2bjqA 2bkxA 2bm5A 2bmoA 2bmwA 2bnmA 2bo4A 2bo9B 2boqA 2bs2B 2bs2C 2bsjA 2burB 2bwrA 2bxxA 2c0hA 2c1dB 2c1lA 2c1vA 2c2uA 2c31B 2c3nA 2c42A 2c4nA 2c5gA 2c6qB 2c6zA 2c78A 2c7pA 2cb5B 2cbp 2cbzA 2ccaA 2cfeA 2cihA 2ciqA 2ciwA 2cjlA 2ck3C 2cl2A 2cl5A 2cmkA 2cn3B 2cqsA 2ctc 2cu5A 2cveA 2cwkA 2cwlA 2cx5A 2cxcA 2cxnA 2cygA 2cyjA 2cz1B 2czcA 2czvC 2d0oA 2d16A 2d29A 2d39C 2d4pA 2d4xA 2d5bA 2d5wA 2d81A 2d8dB 2dbbB 2dbpA 2dc1A 2dc4A 2ddrC 2de3B 2de6B 2dekA 2dg1B 2dgkA 2djfA 2dkoB 2dm9B 2dp9A 2dpoA 2dq6A 2dr3D 2dsjA 2dstA 2e2oA 2end 2erfA 2erl 2ervA 2et1A 2etxB 2eutA 2ewhA 2ex2A 2f01B 2f0cB 2f1fA 2f23A 2f2bA 2f2hA 2f3yA 2f4mA 2f5gA 2f5vA 2f6eA 2f6gA 2f6lA 2f6uA 2f9iD 2fa1B 2faoA 2fb5A 2fbaA 2fbqA 2fbyA 2fcbA 2fcwA 2fd5A 2fd6A 2fdiA 2fe8A 2ffuA 2fhfA 2fhzA 2fi1A 2fipA 2fj8A 2fjrA 2fkoA 2fl4A 2flhB 2fmaA 2fmpA 2fnjA 2fp7A 2fq3A 2fq6B 2fqtA 2fqxA 2fr0A 2ft0A 2fujA 2fukA 2fwfA 2fy6A 2fygA 2fzsB 2g19A 2g2sA 2g2sB 2g2uB 2g7eA 2g8oB 2gagA 2gagC 2gagD 2gbaA 2gbjA 2gc4D 2gdgA 2ge7B 2ghtA 2giaB 2ginA 2gjlA 2gjuA 2gkeA 2gmsA 2gmyA 2gn4A 2gq0B 2gqtA 2grrA 2gsoA 2gtdA 2gudB 2gwmA 2gz4A 2h6fA 2h6fB 2h6nB 2h88A 2h88C 2h8gA 2h8zA 2h9aB 2halA 2hbvB 2hd9A 2hekA 2hf9A 2hjvA 2hrvA 2hxmA 2hxtA 2hy5A

2hy5B 2i3gA 2i47A 2i49A 2i4lB 2i53A 2i56C 2i74B 2iavA 2ic7A 2iccA 2icyB 2if6B 2iimA 2ijqA 2incB 2iprA 2iu1A 2iu5A 2iuhA 2ivfA 2ivfB 2iwaA 2ixsA 2izxB 2j1nA 2j27A 2j2jA 2j45B 2j6aA 2jbaA 2kauC 2kinA 2kinB 2lisA 2mcm 2nacA 2nllB 2nnuA 2nqoC 2nqoD 2nvhA 2nw8A 2nx4A 2nxeA 2o3tA 2o4vA 2o6dA 2o6sA 2o8lA 2pgd 2pia 2por 2pspA 2ptd 2pth 2sak 2spcA 2sqcA 2tgi 2tpsB 2wea 3ezmA 3grs 3prn 3pviA 3sil 3tmkA 3tss 4lzt 4ubpA 4ubpB 7ahlB 7fd1A 8a3hA

Table S3. Number of clusters for each PB and the number of fragments of each PB type in the dataset containing 482 protein chains.

PB	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	total
Number of clusters	36	14	24	52	32	22	11	12	25	8	14	12	13	8	5	12	300
Number of Fragments	4544	5189	9501	22349	2893	7885	1347	2808	2154	982	6391	6345	35204	2401	3264	4111	117,377

Table S4. Relative entropies or the Kullback-Leibler divergences of the within-PB and within-cluster amino acid distributions relative to the background distribution. Fragments contained in each PB are either not partitioned further (within PB) or have been clustered by considering the secondary structure states (PB+SS), or the solvent accessibility states (PB+SA), or both (PB+SS+SA).

PB type	Within PB	PB+SS	PB+SA	PB+SS+SA
a	0.0132	0.0143	0.0148	0.0153
b	0.0027	0.0062	0.0073	0.0076
c	0.0032	0.0064	0.0088	0.0088
d	0.0021	0.0054	0.0088	0.0093
e	0.0122	0.0129	0.0129	0.0143
f	0.0045	0.0072	0.0091	0.0091
g	0.0087	0.0082	0.0106	0.0104
h	0.0123	0.0124	0.0136	0.0141
i	0.0136	0.0134	0.0131	0.0139
j	0.0179	0.0152	0.0154	0.0160
k	0.0070	0.0087	0.0107	0.0109
l	0.0058	0.0076	0.0099	0.0101
m	0.0026	0.0038	0.0095	0.0096
n	0.0151	0.0161	0.0170	0.0184
o	0.0129	0.0143	0.0153	0.0157
p	0.0090	0.0105	0.0127	0.0128
averaged	0.0051	0.0070	0.0102	0.0105

Table S5. Overall prediction rates of sequence-structure database matching using the σ_{N-top} scoring scheme and the HMM model.

Method	sequence-structure database matching using σ_{N-top}		HMM	
	small l dataset	**large dataset**	**small dataset**	**large dataset**
Q_{14} [a]	24.6 (36.0)	31.0 (36.0)	31.5 (29.9)	32.3 (30.9)
TOP2 Q_{14} [b]	45.3	50.2	-	-
TOP3 Q_{14} [b]	58.6	62.9	-	
Q_{16} [a]	46.7	48.0	43.9	44.7
TOP2 Q_{16} [b]	63.6	64.7	-	-
TOP3 Q_{16} [b]	73.3	74.2	-	-

[a] See text for definitions of Q_{14} and Q_{16} rates. Data in parentheses correspond to selectivity rate for the 14 non-regular secondary structure PBs. The small dataset contains 482 and the large dataset contains 1462 selected protein chains. Results have been obtained using 6-fold cross testing. [b] TOP2 and TOP3 rates have been computed by considering any of the two or three PB types with the highest σ_{N-top} scores as acceptable predictions.

Table S6. Overall prediction rates of different control models in 6-fold cross tests using the large dataset.

Accepted predictions	1 PB only		PB only		PB+SS		PB+SA	
	Q_{14}[a]	Q_{16}	Q_{14}	Q_{16}	Q_{14}	Q_{16}	Q_{14}	Q_{16}
Top1	32.1(29.0)	40.7	34.0(30.5)	42.9	35.7(30.1)	41.7	35.2(32.1)	45.0
Top2	50.0	57.2	51.6	59.5	52.1	58.8	52.0	61.5
Top3	61.8	67.8	63.2	69.8	63.1	69.4	63.1	71.5

[a] Data in parentheses are selectivity rates for the 14 non-regular secondary structure PBs

Table S7 The sensitivity rates (SNR) and the selectivity rates (SLR) of different control models in 6-fold cross tests using the large dataset.

PB	PB+SA		PB+SS		PB only		1PB only	
	SNR	SLR	SNR	SLR	SNR	SLR	SNR	SLR
a	62.3	36.0	62.8	34.1	63.6	33.3	63.0	32.5
b	11.8	23.2	10.6	22.8	5.0	27.0	3.7	24.5
c	33.8	30.3	33.7	28.8	31.0	29.4	29.2	28.2
d	49.1	49.0	44.9	47.2	42.8	48.4	39.6	45.3
e	40.4	26.2	42.5	25.0	39.7	24.8	37.4	23.3
f	29.8	34.7	31.1	32.7	28.7	34.2	26.2	31.8
g	12.6	16.9	9.3	18.9	10.4	18.8	9.3	17.7
h	39.6	25.8	38.5	25.0	41.3	23.6	39.6	22.5
i	40.2	23.3	36.6	22.0	40.5	20.6	38.4	19.4
j	10.4	19.3	10.8	18.3	8.6	19.8	6.7	17.4
k	35.4	37.1	37.5	34.0	34.7	35.1	31.7	32.8
l	30.3	37.8	30.7	33.4	29.5	35.6	27.2	33.4
m	59.1	70.7	49.8	69.5	58.1	65.9	56.0	62.2
n	48.6	35.7	48.5	32.4	49.9	31.2	48.4	29.5
o	47.5	38.0	49.9	34.4	47.8	35.0	45.9	33.3
p	39.5	32.8	42.4	29.1	39.6	30.9	38.0	29.7

Table S8. The averaged b_{total} values at different w_{SS}:w_{SA} ratios. The line corresponding to the optimum ratio has been highlighted.

w_{SS}	w_{SA}	b_{total}
0.00	1.00	0.221
0.10	1.00	0.234
0.20	1.00	0.234
0.30	1.00	0.233
0.40	**1.00**	**0.245**
0.50	1.00	0.240
0.60	1.00	0.233
0.70	1.00	0.223
0.80	1.00	0.219
0.90	1.00	0.214
1.00	1.00	0.209
2.00	1.00	0.212
3.00	1.00	0.190
4.00	1.00	0.208
5.00	1.00	0.213
6.00	1.00	0.192
7.00	1.00	0.168
8.00	1.00	0.178
9.00	1.00	0.174
10.00	1.00	0.165
1.00	0.00	0.080

Table S9. The averaged logarithm of the probabilities of native residue types predicted using equation (1) in the main text over the respective background probabilities computed using different values for the parameter ε in the same equation**.** Results for ε from 0.01 to 0.49 are given. The line corresponding to the optimum of ε has been highlighted.

ε	averaged logarithm
0.01	0.0207
0.03	0.0595
0.05	0.0948
0.07	0.1265
0.09	0.1548
0.11	0.1798
0.13	0.2015
0.15	0.2201
0.17	0.2357
0.19	0.2484
0.21	0.2584
0.23	0.2657
0.25	0.2704
0.27	0.2728
0.29	**0.2730**
0.31	0.2709
0.33	0.2669
0.35	0.2609
0.37	0.2531
0.39	0.2436
0.41	0.2325
0.43	0.2198
0.45	0.2057
0.47	0.1902
0.49	0.1734

4. Supplementary Figures

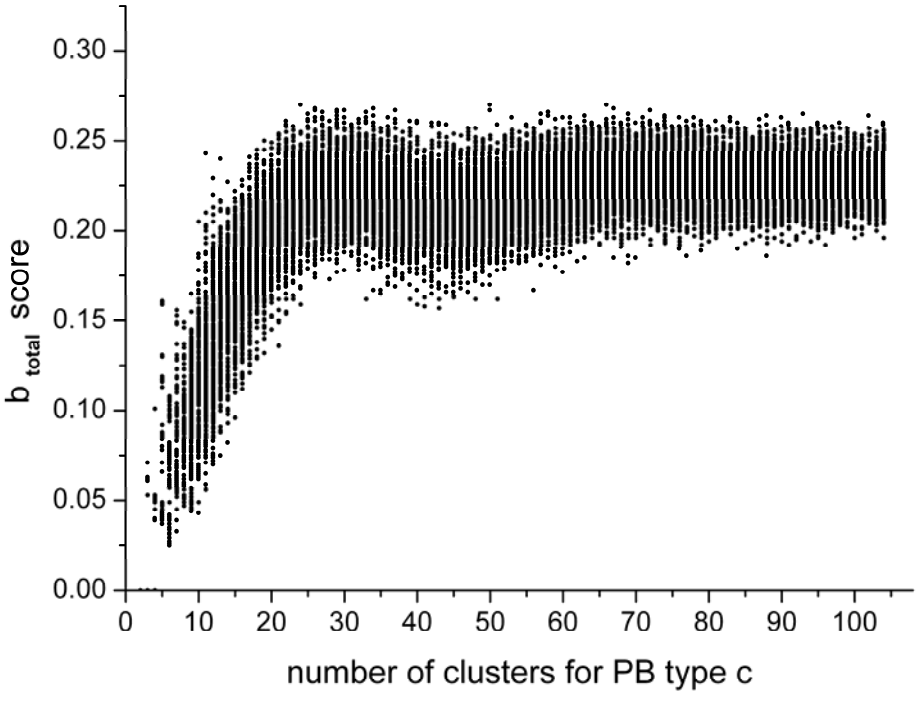
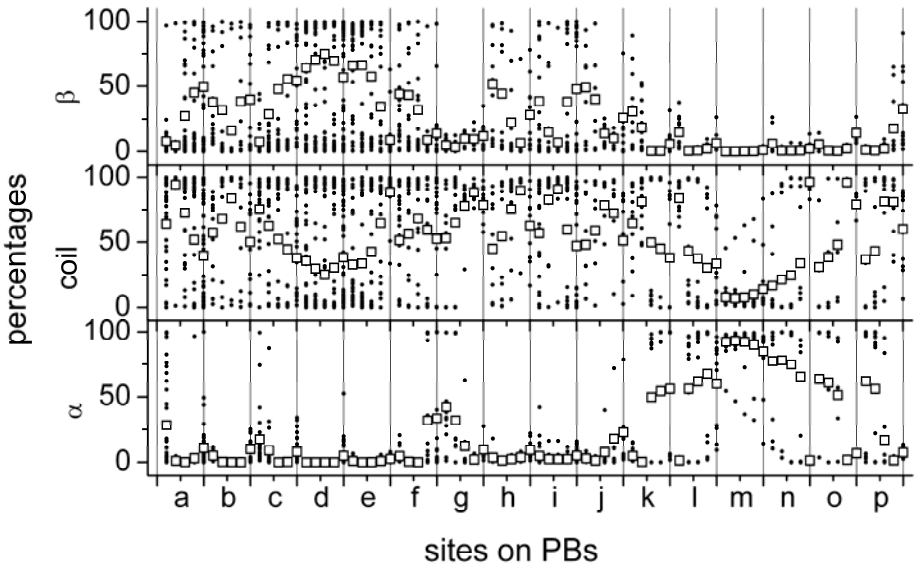
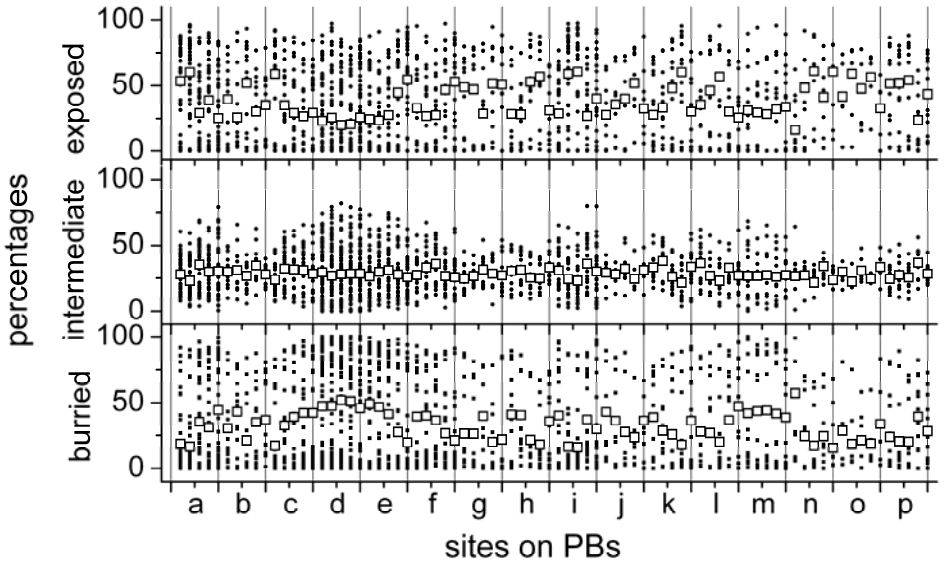


Figure S1. The b_{total} scores (see text) versus the number of clusters in the K-means clustering of fragments of PB type *c* contained in the small dataset. Each point corresponds to one K-means run starting from a random initial guess with given number of clusters.



(a)



(b)

Figure S2. The within-PB and within-cluster distributions of the (a) three secondary structure, and (b) solvent accessibility states computed using the large dataset. The within-PB (open squares) and within-cluster (dots) percentages (the y axis) of sites (the x axis) in different states are shown. Each PB contains five sites and clusters belong to the same PB are shown with the same x coordinates

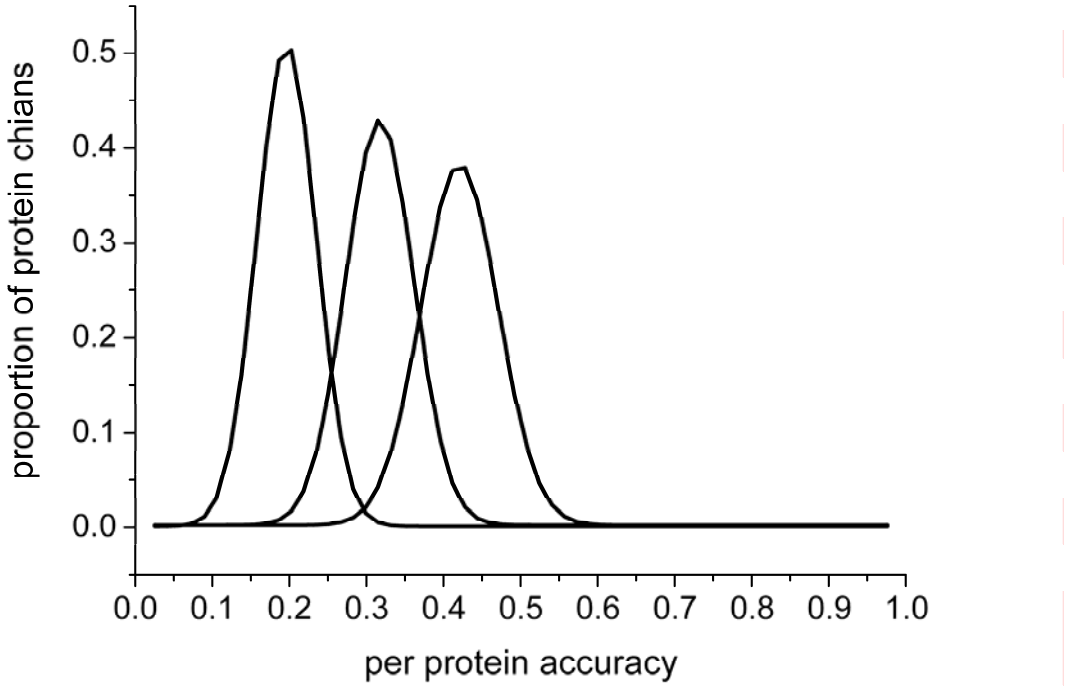


Figure S3. Distributions of the per-protein success rates for the 6-fold pseudo sequence design cross test using the large dataset. Lines from left to right correspond respectively to distributions of rates computed considering any of the one, two and three amino acid types with the highest predicted probabilities as acceptable

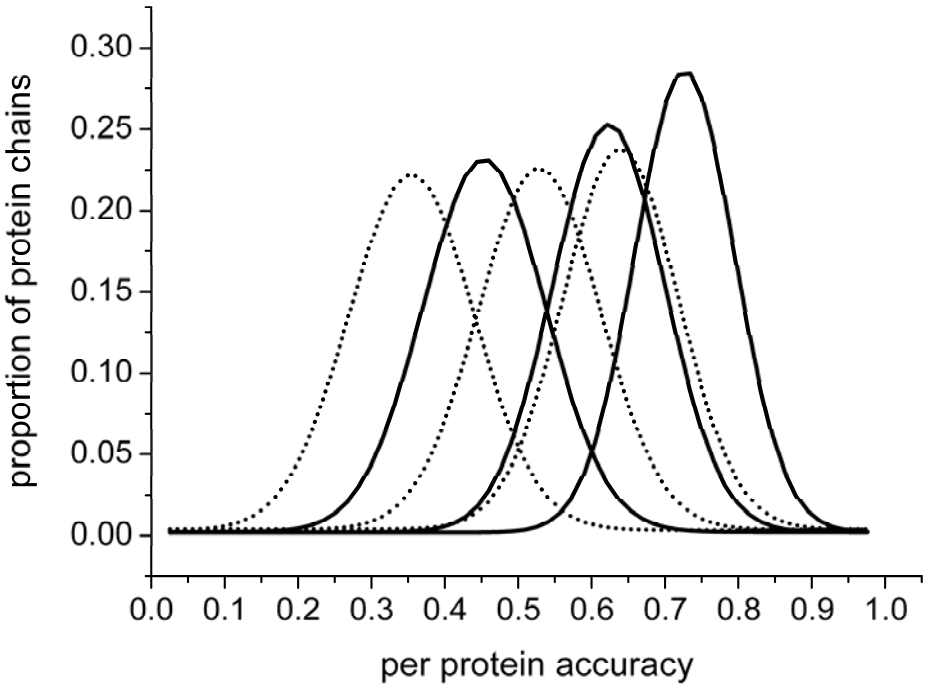


Figure S4. Distributions of the per protein success rates for the 6-fold local structure prediction cross test using the large dataset. Q_{14}: dotted lines, Q_{16}: solid lines. Lines from left to right correspond respectively to distributions of rates computed considering any of the one, two and three top-scoring PB types as acceptable predictions.