WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Ranking the factors that contribute to protein β-sheet folding

Marc Parisien and François Major*

Department of Computer Science and Operations Research, Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, Québec, Canada

ABSTRACT

The formation of β-sheet domains in proteins involves five energetically important factors: the formation of networks of hydrogen bonds and hydrophobic faces, and the residue propensities, or preferences, to be found at the edges of the β-sheet, to adopt the extended conformation, and to make contact with other residues. These relative energy contributions define a potential energy function. Here, we show how optimizing this potential energy function reveals the formation of hydrophobic faces as the utmost factor. The potential energy function was optimized to minimize the Z-scores of the native topologies among the exhaustive sets of over 400 different β-sheets. These results corroborate with experimental data that showed the environment of a protein is an important modulator of β-sheet folding. The contact propensities were found to be the least important, which could explain the poor predictive power of β-strand alignment methods based on pairwise contact matrices.

Proteins 2007; 68:824–829.

Key words: *protein; β-sheet; topology; folding; energy; structure; prediction; hydrophobicity.*

INTRODUCTION

The goal of determining the correct, or native, three-dimensional structure of a given sequence of amino acids, known as the protein folding problem, is still not realized, and it represents one of the major challenges in molecular biology.[1] Two of the fundamental characters of the protein folding problem were established by Anfinsen more than three decades ago. First, Anfinsen demonstrated that the sequence of amino acids determines the structure.[2] Second, he made the observation that the final structure is at a minimum energy state.[3] Since these great findings, many great steps toward finding a solution to the protein folding problem have been made. However nobody came up with an algorithm to properly fold amino acid sequences in their native free-energy state structures.[4,5]

To simplify and to make the protein folding problem tractable, a divide-and-conquer approach can be taken, where the proteins are divided smaller and easier to fold domains. In particular, β-sheets have been extensively studied as they are present in ~80% of the globular proteins. β-sheets form an extensive, but simple, hydrogen bonded network, as predicted by Pauling and Corey in the early fifties.[6] As the number of available X-ray crystal structures has increased, it was realized that the β-sheets presented additional characteristics, which under statistical analysis were shown to be significant and useful in β-sheet prediction methods.

In particular, Chou and Fasman observed that a number of amino acids are more frequent in either α-helices or β-strands.[7] Then, Lifson and Sander pointed out that antiparallel β-strands have a different amino acid composition than the parallel β-strands,[8] and then observed a number of residue pairs that are more likely to be found in parallel or antiparallel β-sheets.[9] At the same time, Cohen et al. derived a set of logical rules to address the packing of α-helices on parallel β-sheets, and then discussed about the presence of hydrophobic patches, or grooves, on one side of the β-sheets.[10] Correctly folded globular proteins are soluble,[11] and solubility, which is conserved through evolution,[12] is another property that can be employed to predict β-strands from sequence data.[13]

More recently, Wouters and Curmi revisited the residue pair preferences observed by Lifson and Sander, and evaluated the likelihood of a pair to be hydrogen bonded or not in antiparallel strands.[14] However, there were discrepancies between their results and those obtained experimentally.[15] A more thorough analysis was made by Hutchinson et al. who considered antiparallel strand alignment according to the χ_1 side-chain

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: Canadian Institutes of Health Research; Grant number: CIHR MT-14604.

*Correspondence to: François Major, Institute for Research in Immunology and Cancer, Université de Montréal, PO Box 6128, Downtown station, Montréal, Québec, Canada H3C 3J7. E-mail: Francois.Major@UMontreal.CA

Received 21 September 2006; Revised 8 February 2007; Accepted 14 February 2007
Published online 21 May 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21475

torsion angle.[16] In this case, only 10% of the possible 210 pairs were found statistically significant, and the predictive power of such correlated pairs in strand alignment is yet to be demonstrated. The information theory framework[17] was also applied to study strand pairing, and in particular to predict native strand alignments.[18] The results were far from encouraging, yielding a mere 10% better signal than using random β-strand pairs. Steward and Thornton did not improve their results by considering the β-strand relative ordering and hydrogen bonding patterns. The latter observation suggests, to the contrary of Wouters and Curmi's claims, that hydrogen bonding does not contain relevant information about residue partnership. More sophisticated residue pairing preferences were computed by Zhu and Braun, who then built a pseudo-potential energy function to discriminate among different alternatives, and to generate distance constraints for self-correcting distance geometry calculations.[19] Unfortunately, this approach has only been tested with the pancreatic trypsin inhibitor protein, and thus needs to be applied to a larger sample of proteins. Some have claimed that β-sheet topologies could be inferred from residue pair information.[20,21] However, Mandel-Gutfreund et al. have found that residue pairs in antiparallel β-sheets are equally conserved and covaried as much as noninteracting residue pairs.[22] Furthermore, Minor and Kim[23] demonstrated that the β-sheet formation is largely accountable to the external context of the sheet, which has later been confirmed experimentally.[24,25]

As one can see, several different studies have pointed to many different factors that are suspected to contribute to the formation and stabilization of the native fold of protein β-sheets. However, no one factor by itself allows us to predict accurately the native topology of a β-sheet from sequence data. Furthermore, no studies have addressed the combinations of the proposed factors. Here, we question the relative contribution of these factors. We devised a pseudo-potential energy function composed of weighted terms for the residue preferences for the extended conformation (self), the likeliness of a residue to be found at the edges of the β-sheet (loc), the residue pairing propensities (pair), the hydrogen bonded network (hbonds), and overall hydrophobicity (hydro) of the buried β-sheet faces. We used our energy function to score exhaustive sets of computer generated and alternative β-sheet topologies. We then applied an iterative minimization procedure to address the weight space and to assign the optimal weights that predict the native topologies. In this way, we quantified the relative contributions of each term. The relative contributions we found support the experimental results of Minor and Kim, and explain the poor predictive power of residue pairing approaches. In addition, our results reveal one of the most important factors of the β-sheet phenomenon: hydrophobicity.

METHODS

Database

"PDB Select 25" is a subset of the X-ray crystallographic structures in the Protein Data Bank[26] (PDB) that share a maximum of 25% of sequence identity.[27] PDB Select 25 was compiled by the Pisces server[28] on the 27th of April 2004. All structures in the PDB Select 25 have a resolution of 3.0 Å or better. The β-sheet domains were selected by using the DSSP computer program.[29] The β-sheets that satisfied the four following criteria were selected and put in a database: (1) contain five strands or less; (2) each strand is made of at least four residues; (3) allow for more than one alternative topology; and (4) are free of β-bulges or similar β-sheet "imperfections". We rejected the β-sheets of more than five strands because of the large number of possible topologies.

Generation

We developed a computer program to generate all possible alternative topologies of a given β-sheet made of $N \geq 2$ strands. First, the $N!$ strand permutations are considered. For each permutation, the 2^N strand orientations are tested. Identical topologies are avoided along the two-fold axis of symmetry, and thus a total of $N! \times 2^{(N-2)}$ topologies are produced. For each topology, the strands slide on each other. The topologies where each residue is paired with at least one partner in an adjacent strand were kept. Consecutive strands that are separated by fifteen residues or less can only be antiparallel, since such short connecting loops does not allow for the parallel arrangement. The topological rules to avoid the pretzel motif[30] were not implemented.

Energy

The pseudo-potential energy function to evaluate the β-sheet topologies was defined as:

$$E^{\text{sheet}} = \underbrace{\omega_{\text{hydro}} E^{\text{hydro}}}_{\text{I}} + \underbrace{\omega_{\text{hbonds}} E^{\text{hbonds}}}_{\text{II}} + \sum_{i,j>1}^{\text{pairs}} \left(\underbrace{\frac{1}{2}\omega_{\text{self}} E_i^{\text{self}}}_{\text{III}} + \underbrace{\frac{1}{2}\omega_{\text{self}} E_j^{\text{self}}}_{\text{III}} + \underbrace{\omega_{\text{pair}} E_{i,j}^{\text{pair}}}_{\text{IV}} \right) + \sum_{i}^{\text{residues}} \underbrace{\omega_{\text{loc}} E_i^{\text{loc}}}_{\text{V}} \quad (1)$$

where ω_x is the weight of the x energy term, E^x. The pair (i, j) is defined as the β-sheet partners; residues that are side-by-side in the β-sheet (i.e. same registry), but are located on two different β-strands. Residues at the edge of the β-sheets have only one partner, whereas those

inside the β-sheet have two partners. Although the self-terms III are single-indexed, they are defined within the context of a pair and inform on the relative strand orientations. Even though each pair is evaluated once, an interior residue can participate in two partnerships. Consequently, the self-terms are halved at the expense of depreciating the self-contributions of the edge residues. Overall, the self-contributions of interior residues are twice that of those on the edge. The terms I–V capture the specific β-sheet factors under study.

The hydrophobic term (I:hydro) maximizes the total hydrophobic score of a given β-sheet face[10,23,31]:

$$H_{\max}^{\text{hydro}} = \max\left(\sum_{i}^{\text{face1}} H_i^{\text{hydro}}, \sum_{j}^{\text{face2}} H_j^{\text{hydro}}\right)$$
$$E^{\text{hydro}} = -RT \cdot \ln\left(1 + |H_{\max}^{\text{hydro}}|\right) \cdot \text{sign}\left(H_{\max}^{\text{hydro}}\right) \quad (2)$$

where H_i^{hydro} is the hydrophobicity score of the i^{th} residue, as taken from Cowan and Whittaker.[32] The product RT was fixed at 2.5 kcal/mol. The maximum over both faces of a given β-sheet topology is taken.

Consider horizontally aligned strands. We can then assign the faces sequentially and arbitrarily by assigning face 1 to the strand with the lowest residue ID. We then recursively visit the residues of the β-sheet, both horizontally and vertically, and we assign the proper face to each following residues. For instance, two consecutive residues on the same β-strand (horizontal direction) point to opposite β-sheet faces, whereas β-sheet partners (vertical direction) point to the same face.

The energy of the hydrogen bond network term (II: hbonds) is proportional to the difference between the number of hydrogen bonds in a current topology and the average number of hydrogen bonds in all possible topologies. Each topology admits two possible hydrogen bonding ladder configurations (Nb_1 and Nb_2), and we take the maximum of the two. The energy scales by RT units:

$$HB_{\max} = \max(Nb_1^{\text{hbonds}}, Nb_2^{\text{hbonds}})$$
$$E^{\text{hbonds}} = -RT \cdot (HB_{\max} - \overline{HB_{\max}}) \quad (3)$$

where Nb_i^{hbonds} is the number of hydrogen bonds in the i^{th} configuration, and $\overline{HB_{\max}}$ is the average number of maximum hydrogen bonds in all possible topologies.

It is noteworthy to mention that β-sheets can accommodate, without stress, any mixed combination of relative β-strand orientations: parallel/parallel, parallel/antiparallel, or antiparallel/antiparallel. Consider again horizontally aligned strands. The N—H and C=O dipoles of a given residue backbone can either point up or down (with an analogy to spins), and thus the two possible hydrogen-bonding ladder configurations are possible. A configuration is fully determined from the first residue, for instance the one with the lowest residue ID. Then, visiting recursively the residues in both directions once again, two consecutive residues (on the same β-strand) have their dipoles in opposite directions, or opposite spins. Vertically, if the partner is on a parallel strand, the spins are the same, whereas if the partner is on an antiparallel strand, the spins oppose. We simulate all canonical hydrogen-bonding motifs, and in particular the alternating opened (wide: spins outward) and closed (narrow: spins inward) hydrogen bond rings along an antiparallel strand pair (atomic structure details of these rings can be found in Parisien and Major[33]). Indeed, the initial choice of spin for the first residue influences the number of total hydrogen bonds in the β-sheet, and this is why we consider the maximum number of bonds of the two possible arrangements.

Switching the hydrogen-bonding configuration has an impact on the residue sidechain face, since switching from up to down (or vice-versa) backbone dipoles implies a rotation of 180° along the strand axis. Since this flipping process is applied to each residue in the new hydrogen-bonding patterns, the residues on the same face prior to the flipping stay in the same direction, and, thus, are all still part of the same face. Consequently, we can calculate terms I and II independently.

The residue preferences for the extended conformation term (III:self) accounts for the frequency of the residue type in the context of the parallel and antiparallel strands, after the observation of Lifson and Sander.[8] We represent it by:

$$E_i^{\text{self}} = -RT \cdot \ln(\varphi_i^{\text{self}}) \quad (4)$$

where φ_i^{self} is the propensity of the i^{th} amino acid to be in the parallel or antiparallel configuration. To increase the number of data per bins, we partitioned the 20 residues in the 10 classes of Li et al.[34] The propensities are such that $\varphi_{i,\text{anti}}^{\text{self}} \cdot \varphi_{i,\text{para}}^{\text{self}} = 1$ (see Table SM-I). A value of $\varphi_{i,\text{anti}}^{\text{self}} > 1$ indicates that the i^{th} residue prefers the antiparallel arrangement.

The residue pairing propensity term (IV:pair) scores the residue pairings. In Zhu and Braun,[19] our term would correspond to the ε_0 first order contact matrix. We represent the pair term as:

$$E_{i,j}^{\text{pair}} = -RT \cdot \ln(\varphi_{i,j}^{\text{pair}}) \quad (5)$$

where $\varphi_{i,j}^{\text{pair}}$ is the propensity of pair (i, j) to be in the parallel or antiparallel configuration. The propensities are shown in Table SM-II.

Finally, the location term (V:loc) deals with the preference of an amino acid to be on the edge of a β-sheet (edge strand) and to partner with only one residue, or, alternatively to be buried inside the sheet paired with two partners. We represent the term as:

$$E_i^{\text{loc}} = -RT \cdot \ln(\varphi_i^{\text{loc}}) \quad (6)$$

where φ_i^{loc} is the propensity of the i^{th} amino acid to be found at the edge or inside a β-sheet. The propensities are such that $\varphi_{i,\text{edge}}^{\text{loc}} \cdot \varphi_{i,\text{interior}}^{\text{loc}} = 1$ (see Table SM-I). A

DOI 10.1002/prot

value of $\varphi^{\mathrm{loc}}_{i,\mathrm{edge}} > 1$ indicates that the i^{th} residue prefers to be at the edge of a β-sheet. This term is different from the pair term as is does not try to evaluate the quality of the partnership.

Optimization

The Z-score has been used to rank the possible β-sheet topologies. The Z-score of the i^{th} topology of the j^{th} β-sheet of energy E^j_i is defined by:

$$Z^j_i = \frac{E^j_i - \mu_{E^j}}{\sigma_{E^j}} \tag{7}$$

where μ_{E^j} is the mean energy over all generated alternative topologies for the j^{th} β-sheet. σ_{E^j} is the standard deviation of the energy. A negative Z-score indicates a lower (better) energy than average. The optimization process minimizes the sum of the Z-scores for each native β-sheet topology in the database:

$$\min\left(\sum_{j}^{\beta-\mathrm{sheets}} z^j_{\mathrm{native}}\right) \tag{8}$$

The results are given in terms of the mean Z-score for the native topologies, $\overline{Z_{\mathrm{native}}}$, and corresponds to the sum of the Z-scores of the natives [Eq. (7)] divided by the number of β-sheets. A grid search over the vector $(\omega_{\mathrm{hydro}}, \omega_{\mathrm{hbonds}}, \omega_{\mathrm{self}}, \omega_{\mathrm{pair}})$ was used to locate the absolute minimum. Once the optimal weights ω_i were obtained, they were normalized so that $\sum_i \omega'_i = 1$ using the following transformation: $\omega'_i = \omega_i / \sum_j \omega_j$.

RESULTS AND DISCUSSION

Database

The statistics for the derivation of the various pseudo-energy values were computed from 447 β-sheets collected in the database (see Methods). These β-sheets represent only around 26% of the total number of β-sheets of five or less strands, indicating that most β-sheets contain β-bulges and other "imperfections," preventing the generation of native topologies. The β-sheets of only one possible topology were also discarded, such as those composed of two β-strands connected by a short loop, which forces the antiparallel topology. The β-sheets in the database are partitioned in 342 pure antiparallel (76.5%), 82 pure parallel (18.3%), and 23 hybrids (5.1%). The distribution of the number of strands per β-sheet in the database reflects the overall distribution of the β-sheets in the PDB Select 25[27]: 169 β-sheets of two strands; 117 of three; 109 of four; and 52 of five.

Generation

The total number of generated topologies, including the relative strand alignment, is upper-bounded by $N! \times 2^{(N-2)} \times N^2$ (see Fig. SM-1). The statistics on the number of generated topologies per β-sheet in the database show a minimum of 2, the first quantile is 2, the median is 4, then mean is 1251, the third quantile is 104, and the maximum is 25,730. These results indicate that, even though there are many theoretical alternative topologies, only a small subset is valid, considering that short loops imply the antiparallel arrangement.

Optimization

All factor contributions have been computed and compiled (see Table I). A grid search over the space of the free parameters, the ω_i in Eq. (1), has been applied to locate the global minimum. The free parameters were allowed to take any value in the range $[0, 10]$ using steps of 0.1. The simulations are organized by groups of terms, where N-wise contributions are associated to the combined contribution of N terms when all others are ignored (set to zero). In general, to find the global minimum, the grid search evaluates 100^N grid points ($N > 1$).

In the Table, all ω_i values have more or less the same scalar intensities, pointing to the fact that the energy terms E^i are proportional in the scoring of β-sheet topologies. The Z-score landscape near the global minimum is smooth (data not shown) and all optimal values found are within the grid, suggesting that the grid and the step sizes were appropriate. A simple affine transformation can be used to normalize the weights so that their sum equals 1: $\omega'_i = \omega_i / \sum_j \omega_j$.

The 1-wise contributions are important as they show the individual participation to the identification of the native topologies. As a control, the 1-wise simulations produced negative mean Z-scores (see Methods), indicating that each term actually participates to the identification of the native topologies. Here, since one parameter was set to one and all others to zero, the grid search was not involved. Although the most important factor is hbonds, as indicated in the upper region of the Table, the single contributions are almost of the same predictive strength.

The least potent term is loc, indicating the nature of the residues found at the edge of β-sheets, and conversely of those found in the interior of β-sheets. The loc term adds the least information about the native β-sheet topology, despite the preference of polar and charged amino acids to be found at the edge of β-sheets (see Richardson and Richardson[11]). Indeed, β-sheet interiors display a characteristic amino acid distribution (see Table SM-I and Fig. SM-2).

The results of the 2-wise contributions indicate clearly the predominance of the hydro component (over 78% of the total energy when present). The second most influent term is the hbonds, which dominates in the absence of the hydro term. The most predictive combination (mean native Z-score of −0.78) is composed of hbonds (65.9%) and pair (34.1%). The 4-wise contributions still improve

Table I
N-Wise Contributions to the Minimum Average Native Z-Scores, $min(\overline{Z_{native}})$

ω_i						$\omega_i E^i$ (%)				
hydro	hbonds	self	pair	loc	$min(\overline{Z_{native}})$	hydro	hbonds	self	pair	loc
1-Wise contribution										
1.0	*	*	*	*	−0.42	100.0	*	*	*	*
*	1.0	*	*	*	−0.58	*	100.0	*	*	*
*	*	1.0	*	*	−0.46	*	*	100.0	*	*
*	*	*	1.0	*	−0.49	*	*	*	100.0	*
*	*	*	*	1.0	−0.29	*	*	*	*	100.0
2-Wise contributions										
7.5	3.3	*	*	*	−0.71	78.6	21.4	*	*	*
9.6	*	3.8	*	*	−0.68	90.8	*	9.2	*	*
4.5	*	*	2.0	*	−0.57	89.4	*	*	10.5	*
8.0	*	*	*	1.2	−0.68	92.8	*	*	*	7.2
*	6.2	9.8	*	*	−0.66	*	60.5	39.5	*	*
*	7.3	*	8.8	*	−0.78	*	65.9	*	34.1	*
*	7.2	*	*	7.7	−0.74	*	52.8	*	*	47.2
*	*	9.8	9.4	*	−0.68	*	*	50.1	49.9	*
*	*	8.1	*	5.1	−0.70	*	*	44.0	*	56.0
*	*	*	0.5	0.5	−0.68	*	*	*	33.9	66.1
4-Wise contributions										
8.0	2.1	3.3	2.2	*	−0.86	74.6	12.1	7.9	5.4	*
9.8	1.0	3.2	*	1.8	−0.95	80.5	5.1	6.7	*	7.6
2.4	7.7	*	8.2	8.3	−0.93	17.6	34.9	*	16.0	31.5
6.8	*	3.0	1.7	2.2	−0.91	74.3	*	8.4	4.9	12.4
*	5.5	7.9	7.2	9.1	−0.95	*	28.2	16.8	15.9	39.1
5-Wise contributions										
3.8	1.0	1.6	1.0	1.6	−0.99	64.2	10.4	6.9	4.5	14.0

The minimum weight values, ω_i, as well as the relative contribution of individual terms, $\omega_i E^i$, to the total energy are shown.
Entries marked with a star (*) indicate that the corresponding weights, ω_i, were fixed to zero to remove the corresponding energy term from the total energy.

the predictive power of the function, since the mean native *Z*-score decreases further.

The 5-wise contributions, which include the five components tested in these studies, correspond to the application of Eq. (1). To avoid redundancy among the relative evaluated weight combinations, we fixed one parameter, ω_{hbond}, to 1, and thus the grid search was applied over the other four dimensions. The global minimum was found at $(\omega_{hydro}, \omega_{hbonds}, \omega_{self}, \omega_{pair}, \omega_{o},) = (3.8, 1.0, 1.6, 1.0, 1.6)$, giving a minimum mean native *Z*-score of −0.99; a better score than any individual term or any partial mixture of them. The mean native *Z*-score is also above the sum of each individual term ($-0.99 > -0.42 + -0.58 + -0.46 + -0.49 + -0.29$), indicating coupling between the energy terms. Consequently, Eq. (1) does not represent a functional basis for the scoring of β-sheets. Despite the coupling, all five terms are needed to capture β-sheet features that are not yet fully described by the other terms. It is interesting to compare the interplay of the hydro and loc terms; the extra 14% gain of the loc term (compare the first entry of the 4-wise contribution section against the 5-wise) seems to be almost at the entire expense of the hydro term, and yet the experiment demonstrates their noninterchangeability. Surprisingly, the pair term is not a determinant factor at all, showing a mere 4.5% of the total 5-wise energy weight, after a second place in the 1-wise section.

As can be seen from the Table, the hydro term is responsible for almost 65% of the total energy, which is more than four-fold the impact of any other term and still more than the combination of all other terms added, suggesting that the formation of the hydrophobic face is by far the most important factor of the β-sheet phenomenon. In other words, the formation of the β-sheet arranges the hydrophobic residues to be buried inside a protein and to be shielded by the hydrogen bond network. Consequently, the hydrophobic face of a β-sheet is subject to the context, which, as noted by Minor and Kim,[23] as well as by others,[24,25] has a capital impact on β-sheet propensities. The poor predictions of the relative β-strand alignments[18,19,22] can be explained by this hydrophobic phenomenon, suggested by our results, since the residue pairing propensities is being shadowed by it.

CONCLUSIONS

Understanding how β-sheets fold is key to the solution of the protein folding problem. β-sheets abound in protein structures, they are present in ∼80% of the available X-ray crystallographic structures. β-sheets bring in contact residues that are distant in the peptide chains, providing valuable spatial constraints in the context of molecular modeling and structure prediction. We evaluated

DOI 10.1002/prot

the relative contributions of several factors that characterize protein β-sheets: the formation of networks of hydrogen bonds and hydrophobic faces, and the residue propensities to adopt the extended conformation, to be positioned at the edge of β-sheets, and to make contacts with other residues. Each factor was represented by a term in a potential energy function. An optimization of the energy function has allowed us to compare their contributions in the identification of native β-sheet topologies. Interestingly, the optimization results revealed that the most important factor is the construction of a hydrophobic face. This observation is consistent with the finding of Minor and Kim[23] about the influence of the amino acid context on the final fold of β-sheets, which essentially form a buried hydrophobic shield adapted to the surrounding environment. The prime dominance of the hydrophobic factor over the contact propensities can also explain the problems to predict β-strand alignment using residue contact statistics alone. Now that the relative contributions to the protein β-sheet phenomenon have been determined, we intend to improve our protein structure building tools and make better structure predictions.

ACKNOWLEDGMENTS

We are indebted to the reviewers for pointing out the possible preference of amino acids for the solvated edges of β-sheets. FM is a CIHR Investigator and a member of the Centre Robert-Cedergren.

REFERENCES

1. Kolata G. Trying to crack the second half of the genetic code. Science 1986;233(4768):1037–1039.
2. Anfinsen CB, Haber E, Sela M, White FHJ. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA 1961;47:1309–1314.
3. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181(96):223–230.
4. Vila JA, Ripoll DR, Scheraga HA. Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. Proc Natl Acad Sci USA 2003;100(25):14812–14816.
5. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2003; 53(Suppl 6):524–533.
6. Pauling L, Corey RB. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. Proc Natl Acad Sci USA 1951;37(11):729–740.
7. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 1974;13(2):211–222.
8. Lifson S, Sander C. Antiparallel and parallel β-strands differ in amino acid residue preferences. Nature 1979;282(5734):109–111.
9. Lifson S, Sander C. Specific recognition in the tertiary structure of β-sheets of proteins. J Mol Biol 1980;139(4):627–639.
10. Cohen FE, Sternberg MJ, Taylor WR. Analysis and prediction of protein β-sheet structures by a combinatorial approach. Nature 1980;285(5764):378–382.
11. Richardson JS, Richardson DC. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci USA 2002;99(5):2754–2759.
12. Mandel-Gutfreund Y, Gregoret LM. On the significance of alternating patterns of polar and non-polar residues in β-strands. J Mol Biol 2002;323(3):453–461.
13. Siepen JA, Radford SE, Westhead DR. β-edge strands in protein structure prediction and aggregation. Protein Sci 2003;12(10):2348–2359.
14. Wouters MA, Curmi PM. An analysis of side chain interactions and pair correlations within antiparallel β-sheets: The differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. Proteins 1995;22(2):119–131.
15. Smith CK, Regan L. Guidelines for protein design: the energetics of β-sheet side chain interactions. Science 1995;270(5238):980–982.
16. Hutchinson EG, Sessions RB, Thornton JM, Woolfson DN. Determinants of strand register in antiparallel β-sheets of proteins. Protein Sci 1998;7(11):2287–2300.
17. Shannon CE, Weaver W. The mathematical theory of communication. Urbana, IL: University of Illinois Press; 1949.
18. Steward RE, Thornton JM. Prediction of strand pairing in antiparallel and parallel β-sheets using information theory. Proteins 2002;48(2):178–191.
19. Zhu H, Braun W. Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of β-sheet formation in proteins. Protein Sci 1999;8(2):326–342.
20. Hubbard TJ, Park J. Fold recognition and ab initio structure predictions using hidden Markov models and β-strand pair potentials. Proteins 1995;23(3):398–402.
21. Klepeis JL, Floudas CA. Prediction of β-sheet topology and disulfide bridges in polypeptides. J Comput Chem 2003;24(2):191–208.
22. Mandel-Gutfreund Y, Zaremba SM, Gregoret LM. Contributions of residue pairing to β-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel β-strands. J Mol Biol 2001;305(5):1145–1159.
23. Minor DL, Jr., Kim PS. Context is a major determinant of β-sheet propensity. Nature 1994;371(6494):264–267.
24. Zaremba SM, Gregoret LM. Context-dependence of amino acid residue pairing in antiparallel β-sheets. J Mol Biol 1999;291(2):463–479.
25. He MM, Wood ZA, Baase WA, Xiao H, Matthews BW. Alanine-scanning mutagenesis of the β-sheet region of phage T4 lysozyme suggests that tertiary context has a dominant effect on β-sheet formation. Protein Sci 2004;13(10):2716–2724.
26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–242.
27. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1(3):409–417.
28. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19(12):1589–1591.
29. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–2637.
30. Richardson JS. β-Sheet topology and the relatedness of proteins. Nature 1977;268(5620):495–500.
31. Richardson JS, Richardson DC. Prediction of protein structure and the principles of protein conformation. In: Fasman GD, editor. Principles and patterns of protein conformations. New York: Plenum Press; 1989. pp 1–98.
32. Cowan R, Whittaker RG. Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography. Pept Res 1990;3(2):75–80.
33. Parisien M, Major F. A new catalog of protein β-sheets. Proteins 2005;61(3):545–558.
34. Li T, Fan K, Wang J, Wang W. Reduction of protein sequence complexity by residue grouping. Protein Eng 2003;16(5):323–330.