

# Bringing folding pathways into strand pairing prediction

Jieun Jeong<sup>1,2</sup>, Piotr Berman<sup>1</sup>, and Teresa Przytycka<sup>2</sup>

<sup>1</sup> Computer Science and Engineering Department  
The Pennsylvania State University  
University Park, PA 16802 USA

<sup>2</sup> National Center for Biotechnology Information  
US National Library of Medicine, National Institutes of Health  
Bethesda, MD 20894

email: jeongji@ncbi.nlm.nih.gov, berman@cse.psu.edu,  
przytyck@ncbi.nlm.nih.gov

**Abstract.** The topology of  $\beta$ -sheets is defined by the pattern of hydrogen-bonded strand pairing. Therefore, predicting hydrogen bonded strand partners is a fundamental step towards predicting  $\beta$ -sheet topology. In this work we report a new strand pairing algorithm. Our algorithm attempts to mimic elements of the folding process. Namely, in addition to ensuring that the predicted hydrogen bonded strand pairs satisfy basic global consistency constraints, it takes into account hypothetical folding pathways. Consistently with this view, introducing hydrogen bonds between a pair of strands changes the probabilities of forming other strand pairs. We demonstrate that this approach provides an improvement over previously proposed algorithms.

## 1 Introduction

The prediction of protein structure from protein sequence is a long-held goal that would provide invaluable information regarding the function of individual proteins and the evolution of protein families. The increasing amount of sequence and structure data, made it possible to decouple the structure prediction problem from the problem of modeling of protein folding process. Indeed, a significant progress has been achieved by bioinformatics approaches such as homology modeling, threading, and assembly from fragments [16]. At the same time, the fundamental problem of how actually a protein acquires its final folded state remains a subject of controversy. Can successes/failures of computational method shade some light on this issue?

It is generally accepted that proteins fold to their global free energy minimum. Through his famous Paradox, Levinthal made an important point that a protein cannot explore all conformational states in the search of the optimal conformation and therefore a protein chain has to fold by following some directed process or a folding pathway [14]. One view that has been gathering a lot of support since at nearly three decades is the concept of hierarchical protein

## II

folding [1, 2, 6, 12, 13, 19]. Consequently, many structure prediction algorithms use hierarchical approach in which the structure is assembled in a bottom up fashion (e.g. where smaller locally folded fragments are assembled into larger folded units [4, 9, 21]).

Studies of  $\beta$ -sheets topology indicate that the way strands assemble into larger sheets may be quite complex. While about half of hydrogen bonded pairs of strands are adjacent in the sequence of strands on the chain, many are separated by a significant distance. How pairs of strands which are distant in sequence find their hydrogen bonded partners? In her classic 1977 paper, Jane Richardson proposed a set of folding rules where consecutive  $\beta$ -strands grow into larger hydrogen-bonded structures in successive steps, and blocks of strands obtained in this way coalesce, providing they are consecutive in the chain. Richardson showed, by manual inspection, that 37 known strand topologies can be constructed using these rules.

A smaller, more restricted set of folding rules was shown by Przytycka *et al.* [17] to be sufficient for 80% of fold families, while proteins in more than 90% consist of at most three substructures that can be completely folded using proposed rules.

It is tempting to hypothesize that such procedures are related to actual folding pathways. If this hypothesis is correct, such folding rules should be helpful in prediction of  $\beta$ -sheet topology in general, and the pairing of  $\beta$ -strands in particular.

The latter problem, despite of many attempts, remains unsolved. Early work by Hubbard [7] has been followed by other studies directed towards understanding and predicting  $\beta$ -sheet topology [8, 15, 20, 22, 23, 25, 26]. In a recent work, Cheng and Baldi [5] addressed the strand pairing problem using a three-stage approach. In the first stage they compute, for the input protein sequence, the scores (estimated probabilities) of residue pairs as potential partners in a  $\beta$ -strand pairing. This computation is performed by a neural network with input describing a window of size five around each residue and the information about the distance between the two residues in the protein sequence. In the second stage the above pairwise scores are used to define alignment scores for pairs of strands, and for each pair a highest scoring alignment is found with the use of dynamic programming. The alignment scores are used in the third and final stage to run a greedy selection algorithm.

Cheng and Baldi reported 59% specificity and 54% sensitivity which is significantly better than what is achieved by a naive algorithm predicting that all pairs of strands that are consecutive in the sequence form hydrogen bonded partners in space. (The performance of such naive algorithm was approximated to be 42% specificity and 50% sensitivity [5]. )

The important novelty of the approach of Cheng and Baldi when compared with previous methods (*e.g.* Hubbard [7], Zhu and Braun [26] and Steward and Thornton [22]) is that the prediction of residue pairs that are partners in strand pairing is not performed independently for each pair, but instead it takes into

account a wider context; to wit, the information about 10 surrounding residues and the distance between them.

The approach of Cheng and Baldi does not employ explicitly folding rules, although some bias towards formation of hairpins (and in general contacts between close in sequence strands) is included in the (learned) scoring function. On the other hand, the third stage of their algorithm is a very simple greedy algorithm, which raises a question: Would a more elaborate approach increase the quality of prediction even further? What is more important – a better optimization method (*e.g.*, as discussed by Berman and Jeong in [3]), or a biological insight, in particular, the knowledge of folding rules?

To address these questions, we investigated two new algorithms for predicting strand partners. To make direct comparisons, we use the same scoring function as of Cheng and Baldi. The objective of the first algorithm is very similar to the approach of Cheng and Baldi, but rather than having a two-stage greedy selection heuristic, it poses the problem as integer linear programming optimization problem and solves it using ILOG CPLEX<sup>TM</sup> package.

The second approach is greedy, but it explicitly encourages two simple folding rules. This is achieved by dynamically increasing the scores of pairs of strands (as potential partners), depending on the pairs of strands predicted so far. In particular, we double the pairs of strands whose pairing is consistent with one of the rules, which are based on the pairs that are already formed. Our rules are simple and biochemically justified (as we explained later in the paper).

Both methods provided noticeable improvement over the previous approach. Importantly, a more significant improvement was obtained with the approach that promotes folding rules. This is remarkable, since in the case of integer linear program we are heuristically solving a NP-complete problem using about 100 times more time than folding rules promotion algorithm (almost entire time of the latter algorithm is consumed by the dynamic programming that computes optimal pairing/alignment for each pair of strands).

While the improvement, taken in absolute numbers, is not drastic (about 2.7% in sensitivity and ca. 1% in specificity), one has to keep in mind that the difference in sensitivity between the naive algorithm and the initial algorithm of Cheng and Baldi was only 4%. Furthermore, applying rules has domino effect as it alters the subsequent predictions and this effect was quite important; with new rules, fewer predictions are made, but the number of correct predictions increases slightly.

## 2 Methods

### 2.1 ILP formulation

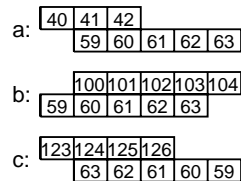
While there are many ILP methods used for protein structure prediction (*e.g.*, see [10, 11, 24]) none of them operated in our particular framework, instead, they were used in the context of all-atom model, threading etc.

We view the strand pairing problem as an optimization problem which identifies a maximum score set of strand pairs which is a subject to constraints [5] that eliminate the most obvious contradictions between the selections.

#### IV

A strand contact in this formulation is specified by listing its *upper* strand, *lower* strand, *offset* (relative shift between the two strands), and an indicator that tells whether the contact is parallel or anti-parallel. For each contact  $c$  we have a 0-1 selection variable  $x_c$ .

In particular a strand can form contacts with other strands on either of its sides but the contacts on one side cannot overlap. In Fig. 1, contacts **b** and **c** are in conflict as they overlap on residues 61 and 62. The second restriction is that if two strands are making a contact on the same side of one strand, the direction of these contacts (parallel or antiparallel) must be the same. In Fig. 1, contacts **a** and **c** are in conflict for this very reason.



**Fig. 1.**

To exclude all possibilities that cannot be realized globally without excluding any arrangement that can be extremely difficult. After some experimentation we decided to exclude cycles (as did Cheng and Baldi [5]). In the data set, among all 916 protein chains and ca. 9000 strands there were only 80 cycles, of which four were very atypical (lengths 3, 4 and 7) and 76 “typical” (lengths 5, 6 and, most frequently, 8). At the same time, without prohibition of all cycles, our program was returning solutions that almost always had cycles of atypical length, hence at least 99% wrong. One can formulate a condition that would allow for cycles of  $\beta$ -strands that may occur in typical barrels, but this would complicate the program enormously with extremely small gain.

Another exclusion was removing all contacts with score below 0.06 from further consideration. This removal caused the number of predicted contacts (true and false positives in Table 1) to be roughly coincident with the number of actual contacts (true positives and false negatives).

To describe the solutions in the language of linear programming we introduced a variable  $x_c$  for every contact  $c$  and a variable  $y_{i,j}$  for every possible strand pairing. The value of  $x_c$  indicates if contact  $c$  is in the solution ( $x_c = 1$ ) or not ( $x_c = 0$ ). Similarly,  $y_{i,j} = 1$  means that strands  $i$  and  $j$  were paired, *i.e.* that we selected a contact that binds these two strands together.

Each contact  $c$  has *score*  $E(c)$  that is computed using dynamic programming (we allowed a single gap of length 1 in the alignment). Thus vector of variable values  $x$  is our solution and the inner product  $xE$  is our objective function.

To formulate our ILP we introduce two classes of 0-1 vectors:  $C_{i,j}$  such that  $C_c = 1$  if and only if contact  $c$  binds strand  $i$  with strand  $j$ , and  $\gamma(S)$  such that  $\gamma(S)_{i,j} = 1$  if and only if  $\{i, j\} \subset S$ . We also set  $\text{conflict}(c, d)$  to be true if there is a conflict between contact  $c$  and contact  $d$ .

We wish to solve the following ILP:

$$\begin{aligned}
 & \text{maximize } Ex \\
 & \text{subject to} \\
 & C_{i,j}x \leq y_{i,j} \quad \text{for } \{i, j\} \subset \{1, \dots, n\} \quad \text{contact/pairing} \\
 & x_c + x_d \leq 1 \quad \text{for } i, j \text{ s.t. } \text{conflict}(c, d) \quad \text{no-conflict} \\
 & \gamma(S)y \leq |S| - 1 \quad \text{for } S \subset \{1, \dots, n\} \quad \text{no-cycle}
 \end{aligned}$$

However, we cannot do it as the number of possible subsets used in no-cycle constraints may be too large, as the number of strands often exceeds 20, and sometimes even 40.

Instead, we start with a single no-cycle constraint with  $S = \{1, \dots, n\}$  and run a *row generation* loop: we submit ILP, we obtain a solution, and if it contains a cycle of strands we add a no-cycle constraint for its set of nodes. When the number of repetitions is too large (as it happened in ca. 15% of the cases) we give up and return the solution of a greedy program.

## 2.2 Greedy algorithm with pathway-based promotion

For each pair of strands we pre-select the best parallel and the best anti-parallel contact, and we order them according to their score. We consider candidates starting with the one with the largest score, and we never consider a candidate again.

We represent contacts with *unordered pairs* of strands, which means that we do not declare which strand is the upper one and which one is lower. Otherwise we could get the following anomaly: we greedily choose contacts for pairs (1,2) and (3,4), and decide that, say, strands 1 and 3 are upper ones. Then we cannot choose contact (1,4): if in the latter strand 1 is upper, we have conflict with (1,2), and if strand 4 is lower, we have a conflict with (3,4).

The consistency of a set of contact is verified as follows. We have the same notion of contradiction between contacts as before, but we view these contradictions as *edges* in the graph that consists of the contacts selected so far and the candidate. We require that this graph is bipartite. One can see that a simple cycle has to consist of contacts sharing a common strand, and if this cycle has an odd length, there is no way of consistently placing conflicting contacts on two sides of the common strand. Fig. 1 shows such a situation: strands (40-41-42), (100-...-104) and (123-...-126) cannot be placed on two sides of the strand (59-...-63).

Connected components of the above graph correspond to rigid parts of the chain, and they can be mapped onto a grid in such a way that strands form rows and paired partners are adjacent in common columns. Such a layout allows to form a conservative estimate of the minimal length of coils that join the strands in the components. If such a coil is actually shorter, we disallow the candidate.

As before, we disallow a candidate if it would create a cycle.

Up to this point, the algorithm does not differ from that of Cheng and Baldi in a significant way. (Their notion of consistency as exhibited by their program is a bit different than the one described in the paper, but in the evaluation it was indistinguishable).

However, we have this new element: after selecting a *consecutive* contact, say between strands  $i$  and  $i + 1$ , we double the score of contacts between strand pairs  $(i, i + 2)$ ,  $(i - 1, i + 1)$ ,  $(i - 1, i + 2)$  and change their position within the ordering to reflect that.

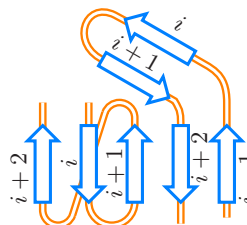


Fig. 2.

This rule is explicitly promoting a folding pathway. It is actually a part of a more general rule, but it restricts it to the cases of the smallest separation between strands and thus the most reliable scores.

There are biophysical reasons for which the probability of hydrogen bonding between strands  $i$  and  $i+2$  (Fig. 2) is increased under assumption that  $i$  is already hydrogen bonded. Namely, strand  $i+2$  would stabilize the conformation already acquired by strands  $i$  and  $i+1$ . The higher probability of bonding between strands  $i-1$  and  $i+2$  upon hydrogen bonding between  $i$  and  $i+1$  is in turn justified by the loss of entropy of subchain separating strands  $i-1$  and  $i+2$  resulted from the hairpin formation. This rule can be extended to strands  $i-2$  and  $i+3$  but with the current scoring schema it had no effect on the results (see Discussion section).

### 3 Results

We used the data set of Cheng and Baldi (see [5], page 176) that consists of 916 protein chains that contain up to 45  $\beta$ -strands.

We also used the output of their program that given a sequence of amino acids (residues) returns (a) a sequence of secondary structure identifications ( $\alpha$ -helix,  $\beta$ -strand, coil) and (b) for every pair of residues classified as  $\beta$ -strand it provides a pseudo-probability that these two residues face each other in a pairing of two  $\beta$ -strands. To evaluate the result we used their file of DSSP identifications of correct secondary structure identifications and correct pairing of  $\beta$ -strand residues.

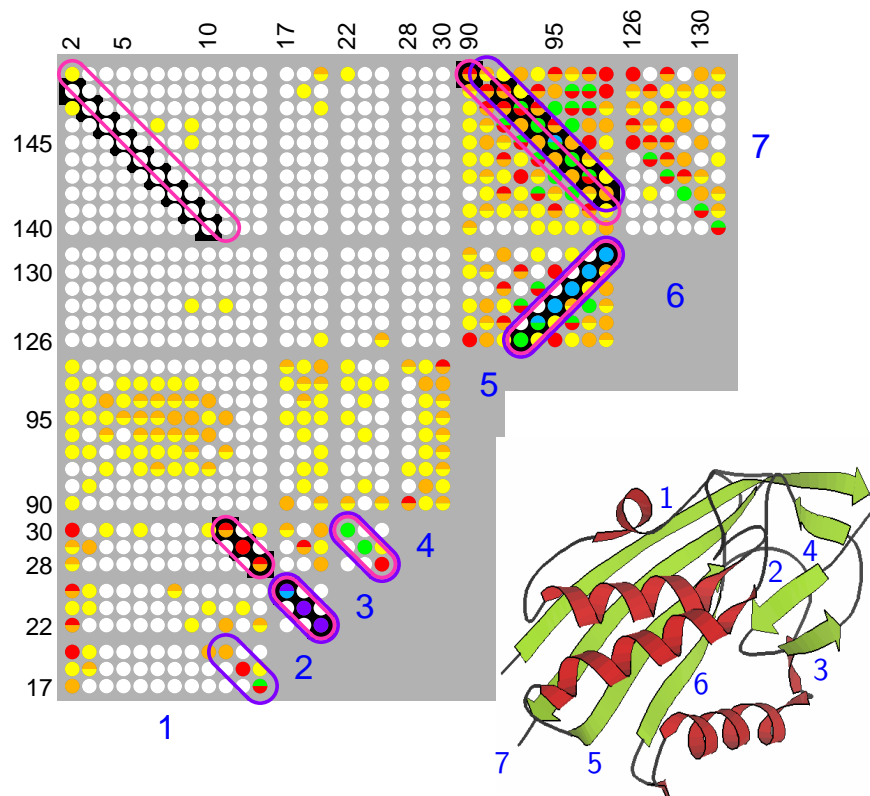
We defined the population of possible answers in two ways: pairs of  $\beta$ -strands as identified by PREDICT\_BETA\_FASTA.SH and as identified by DSSP. Given a pair of predicted (true) strands, we defined the pairing to be true (correctly predicted) if for at least one residue of one strand there was a residue in the other strand that was in a contact described by DSSP (predicted by the evaluated program). These two definitions yielded different numbers, but they registered roughly the same differences between various programs, so our conclusions do not seem to depend on this somewhat arbitrary definition.

We compare three programs: the three-stage program of Cheng and Baldi, ILP optimizer and our greedy algorithm with pathway based promotion. The differences in the quality of predictions are very consistent when we use various measures. We use  $T$  and  $F$  to indicate the number of true and false predictions and  $\oplus$  and  $\ominus$  to indicate positive and negative predictions. To evaluate the set of prediction, we use the correlation coefficient, as well as selectivity/specificity pairs.

$$\text{Correlation coefficient} = \frac{T^{\oplus}T^{\ominus} - F^{\oplus}F^{\ominus}}{\sqrt{(T^{\oplus} + F^{\oplus})(T^{\oplus} + F^{\ominus})(T^{\ominus} + F^{\oplus})(T^{\ominus} + F^{\ominus})}}$$

$$Spe = \frac{T^{\oplus}}{T^{\oplus} + F^{\oplus}} \quad Sel = T \frac{\oplus}{T^{\oplus} + F^{\ominus}}$$

The correlation coefficient was 0.555 for Cheng and Baldi's, 0.567 for ILP optimizer and 0.577 for the greedy with pathway based promotion.



**Fig. 3.** Example how promotion may have good secondary effects. We show here the table of pairwise scores for 2C-Methyl-D-erythritol-2,4-cyclodiphosphate Synthase (PDB id: 1iv1, chain a). The entries in the table are color-coded, purple codes interval  $2/3$  to 1, and each subsequent code (purple-blue, blue, blue-green etc.) codes an interval decreased by  $2/3$  factor (and white for the remaining values down to zero). Black background codes the true contacts, purple ovals are the contacts found by Cheng & Baldi, and the pink ovals are the contacts found by our version of greedy. After contact 2-3 was selected, contact 1-4 (between strand 1 and strand 7) was promoted over 1-2; once we got contacts 1-4-3-2, contact 1-2 was blocked by no-cycle rule; moreover 1-5 was blocked by 5-6 and 5-7, thus 1-7 became the best available contact for 1 — as well as for 7.

#### 4 Discussion and conclusions

We considered two methods of predicting  $\beta$ -sheet pairing partners using the machine learned scores for inter-residue contacts from [5]. In the first method, we computed optimal set of pairs by solving an instance of integer linear program.

The fact that increasing the sum of scores improves the predictions suggests that these scores are indeed related to the energy of contacts. On the other hand, giving preference according to our rule leads to lower sums of scores and yet it improves the specificity significantly without decreasing the sensitivity. This

separation	true positive	false negative	false positive	true negative	specificity	sensitivity	corr. coef.
greedy — Cheng & Baldi's version							
<b>ALL</b>	<b>5032</b>	<b>3140</b>	<b>3370</b>	<b>61563</b>	<b>0.599</b>	<b>0.616</b>	<b>0.557</b>
0	3748	363	2136	3577	0.637	0.912	0.541
1	521	485	484	7418	0.518	0.518	0.457
2	407	523	355	6710	0.534	0.438	0.423
3	169	359	161	6412	0.512	0.320	0.368
4	100	276	89	5788	0.529	0.266	0.348
5	38	241	58	5130	0.396	0.136	0.209
6	29	195	32	4482	0.475	0.129	0.230
7	11	157	10	3891	0.524	0.065	0.175
8+	9	541	45	18155	0.167	0.016	0.044
ILP optimizer							
<b>ALL</b>	<b>5092</b>	<b>3080</b>	<b>3253</b>	<b>61603</b>	<b>0.610</b>	<b>0.623</b>	<b>0.568</b>
0	3781	330	2084	3621	0.645	0.920	0.558
1	538	468	552	7342	0.494	0.535	0.449
2	427	503	317	6741	0.574	0.459	0.457
3	167	361	119	6447	0.584	0.316	0.398
4	94	282	72	5798	0.566	0.250	0.352
5	36	243	39	5143	0.480	0.129	0.230
6	30	194	10	4498	0.750	0.134	0.306
7	14	154	13	3883	0.519	0.083	0.196
8+	5	545	47	18130	0.096	0.009	0.021
greedy — with pathway-based promotion							
<b>ALL</b>	<b>5089</b>	<b>3083</b>	<b>3035</b>	<b>61821</b>	<b>0.626</b>	<b>0.623</b>	<b>0.577</b>
0	3715	396	1733	3972	0.682	0.904	0.596
1	594	412	619	7275	0.490	0.590	0.473
2	472	458	385	6673	0.551	0.508	0.469
3	142	386	122	6444	0.538	0.269	0.347
4	81	295	68	5802	0.544	0.215	0.318
5	37	242	41	5141	0.474	0.133	0.231
6	30	194	10	4498	0.750	0.134	0.306
7	11	157	14	3882	0.440	0.065	0.158
8+	7	543	43	18134	0.140	0.013	0.034

**Table 1.** Comparison of results of three tested algorithms on a set of 916 protein chains. Note that the discriminating power of the potential function quickly decreases as the separation grows and the statistical quality measures are largely determined by contacts separated by up to three other strands.

suggests that a local assembly may remain stable even when it is inconsistent with a conformational state that has the minimal energy. However, for contacts separated for more than 3 strands the reliability of Cheng and Baldi's scores seems to decrease rather quickly and more complete versions of our rules do not lead to further improvements.

In the future, more complete set of rules based on work of Richardson [18] and/or Przytycka [17] should be added. However, more complete rules are also more ambiguous — the number of possible successive steps in the folding process goes up and we need to rely on pairwise predictors more, while in the same time their reliability goes down.

Improving pairwise prediction of contacts for more separated pairs of strands seems to be a necessary challenge before a qualitative improvement in *ab initio* prediction methods of the tertiary structure of proteins. In the same time, this task cannot be separated from the search of the best methods of using such predictors.

An important implication of this work is the demonstration that a simple algorithm that takes into account folding rules works better than heavy duty integer linear programming. This suggests that the future line of research should be developing a folding-rule depending scoring function that would allow to explore a richer set of folding rules.

## Acknowledgments

The authors thank George D. Rose (JHU), Bonnie Berger (MIT) and Arthur M. Lesk (PSU) for an insightful discussions. We also thank Jailing Cheng for help in using their program. This work was supported in part by the intramural research program, National Institutes of Health, National Library of Medicine.

## References

1. Robert L. Baldwin and George D. Rose. Is protein folding hierarchic? I. II. Folding intermediates and transition states. *Trends in Biochemical Sciences*, 24(2):77–83, 1999.
2. Robert L. Baldwin and George D. Rose. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends in Biochemical Sciences*, 134(3):26–33, 1999.
3. Piotr Berman and Jieun Jeong. Consistent sets of secondary structures in proteins. <http://www.cse.psu.edu/~jijeong>.
4. Christopher Bystroff and David Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *Journal of Molecular Biology*, 281(3):565–577, 1998.
5. Jianlin Cheng and Pierre Baldi. Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, 21(suppl 1):i75–84, 2005.
6. Gordon M. Crippen. The tree structural organization of proteins. *Journal of Molecular Biology*, 126:315–332, 1978.
7. Tim J Hubbard and J Park. Fold recognition and *ab initio* structure predictions using hidden markov models and  $\beta$ -strand pair potentials. *Proteins: Structure, Function, and Genetics*, 23(3):398–402, 1995.

8. E. G. Huthinson, R. B. Sessions, J. M. Thornton, and D. N. Woolfson. Determinants of strand register in antiparallel  $\beta$ -sheets of proteins. *Protein Sci*, 7(11):2287–2300, 1998.
9. Yuval Inbar, Hadar Benyamini, Ruth Nussinov, and Haim J. Wolfson. Protein structure prediction via combinatorial assembly of sub-structural units. *Bioinformatics*, 19(suppl 1):i158–168, 2003.
10. Carleton L. Kingford, Bernard Chazelle, and Mona Singh. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, 21(7):1028–1036, 2004.
11. J. L. Klepeis and C. A. Floudas. Astro-fold: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical Journal*, 85:2119–2146, October 2003.
12. Andriy Kryshtafovych, Ceslovas Venclovas, Krzysztof Fidelis, and John Moult. Protein folding: From the Levinthal paradox to structure prediction. *Journal of Molecular Biology*, 293(2):283–293, 1999.
13. Arthur M. Lesk and George D. Rose. Folding Units in Globular Proteins. *PNAS*, 78(7):4304–4308, 1981.
14. C Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, 65:44, 1968.
15. Matthew Menke, Jonathan King, Bonnie Berger, and Lenore Cowen. Wrap-and-pack: A new paradigm for beta structural motif recognition with application to recognizing beta trefoils. *Journal of Computational Biology*, 12(6):777–795, 2005.
16. John Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15(3):285–289, 2005.
17. Teresa M. Przytycka, Rajgopal Srinivasan, and George D. Rose. Recursive domains in proteins. *Protein Sci*, 11(2):409–417, 2002.
18. Jane S. Richardson. beta-Sheet topology and the relatedness of proteins. *Nature*, 268(5620):495–500, 1977.
19. George D. Rose. Hierarchic organization of domains in globular proteins. *Journal of Molecular Biology*, 134(3):447–470, 1979.
20. Ingo Ruczinski, Charles Kooperberg, Richard Bonneau, and David Baker. Distributions of beta sheets in proteins with application to structure prediction. *Proteins: Structure, Function, and Genetics*, 48(1):85–97, 2002.
21. Rajgopal Srinivasan and George D. Rose. LINUS: A hierarchic procedure to predict the fold of a protein. *Proteins: Structure, Function, and Genetics*, 22(2):81–99, 1995.
22. Robert E. Steward and Janet M. Thornton. Prediction of strand pairing in antiparallel and parallel  $\beta$ -sheets using information theory. *Proteins: Structure, Function, and Genetics*, 48(2):178–191, 2002.
23. DN Woolfson, PA Evans, EG Hutchinson, and JM Thornton. On the conformation of proteins: The handedness of the connection between parallel  $\beta$ -strands. *J Mol Biol*, 110:269–283, 1977.
24. Jinbo Xu, M. Li, D. Kim, and Y. Xu. Raptor: Optimal protein threading by linear programming. *Journal of Bioinformatics and Computational Biology*, 1(1):85–117, 2003.
25. Chao Zhang and Sung-Hou Kim. The anatomy of protein [beta]-sheet topology. *Journal of Molecular Biology*, 299(4):1075–1089, 2002.
26. H Zhu and W Braun. Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins. *Protein Sci*, 8(2):326–342, 1999.