J. Mol. Model. **1999**, *5*, 78 – 89

Journal of Molecular Modeling

FULL PAPER

Analyzing the Interplay Between Secondary and Tertiary Structure Predictions in Folding Simulations with a Genetic Algorithm

Thomas Dandekar[1] and Fuli Du[2]

[1]EMBL, Postfach 102209, D-69012 Heidelberg, Germany. Tel: +49-6221 387 372. E-mail: dandekar@embl-heidelberg.de

[2]University of Heidelberg, Germany

Received: 18 November 1998/ Accepted 16 February 1999/ Published: 1 April 1999

Abstract Three different strategies to tackle mispredictions from incorrect secondary structure prediction are analysed using 21 small proteins (22-121 amino acids; 1-6 secondary structure elements) with known three dimensional structures: (1) Testing accuracy of different secondary structure predictions and improving them by combinations, (2) correcting mispredictions exploiting protein folding simulations with a genetic algorithm and (3) applying and combining experimental data to refine predictions both for secondary structure and tertiary fold. We demonstrate that predictions from secondary structure prediction programs can be efficiently combined to reduce prediction errors from missed secondary structure elements. Further, up to two secondary structure elements (helices, strands) missed by secondary structure prediction were corrected by the genetic algorithm simulation. Finally, we show how input from experimental data is exploited to refine the predictions obtained.

Keywords Genetic algorithm, Secondary structure prediction, Tertiary structure prediction

Introduction

The following study concentrates on the interplay between secondary and tertiary structure and how to correct insufficient secondary structure prediction for fold prediction.

Many different approaches to predict secondary structure have been investigated, ranging, for instance, from amino acid properties [1] to deriving propensities from aligned structures [2] and from different stereochemical methods [3] to neural networks.[4] Nevertheless, secondary structure predictions may still contain a considerable fraction of errors.[4] In the present study we investigated whether and to what extent (1) a refined combination of several secondary structure prediction methods could correct mispredictions, (2) the mispredictions could be corrected during protein folding simulation applying a genetic algorithm or (3) by exploiting information [5] from experimental data.

The genetic algorithm used here as the tertiary fold prediction method is a robust searching algorithm which performs well on combinatorially hard problems.[6] Applying it to protein structure prediction [7] we predicted first helical proteins (RMSD to observed around 6 Ångstrœms [8]). Further, starting from sequence and secondary structure information (using secondary structure assignments according to DSSP [9]) the fold for 19 different protein topologies was successfully delineated (proteins less than 100 amino acids in length, with no more than eight secondary structure

Correspondence to: T. Dandekar

Table 1 *Fitness function criteria*

criteria	**des**[a]	**term**	**specific parameters**
constant[a]	C	weight$_C$	adjusted to 10% negative fitness in the first generation
clash[b]	cl	weight$_{cl}$• Σ overlap	weight$_{cl}$ = -500
secondary structure(ss):			
	pf[c]	weight$_{pf}$•(structural preference)	weight$_{pf}$ = +12
	co[c]	weight$_{co}$•cooperativity	weight$_{co}$ = +12
tertiary structure:			
global scatter[a]			
	gs	weight$_{gs}$•scatter	weight$_{gs}$ = -24
hydrophobic scatter[a]			
	hs	weight$_{hs}$•hydrophobic distribution	weight$_{hs}$ = -19 hydrophobic residues include Phe,Tyr, Met,Cys, Ile,Leu,Val,Trp
beta-strand criteria[d]:			
hydrogen	hyd	weight$_{hyd}$•hydrogen	weight$_{hyd}$ = + 15
bond	bond		bondcount + betapair + bondstrand + revturn + 2•bondsheet

sheetdir	sh	weight$_{sh}$•sheetdir	weight$_{sh}$ = + 6;
			within 66°, reward = +1;
			within 35°, additional reward = +6

[a] *The term "des" refers to the abbreviated designation for the criteria involved. A positive constant (C) was added, "gs" denotes the scatter of all residues, "hs" the scatter of all hydrophobic residues around the center of mass. The total fitness was the sum of all fitness terms listed using the optimised weights for each term [8,10] indicated on the right.*

[b] *C_α-C_α (closest distance 3.8 Å, pearl necklace model [28]) and any other mainchain atom overlaps (closest distance 2.47 Å [22]) were counted as clashes (cl).*

[c] *Structural preference (pf) rewards all residue conformations encoded in a bit string which agree with the secondary structure (known or predicted) used in the trial. Cooperativity (co) yields a reward for any two consecutive residues in the same dihedral conformation.*

[d] *hydrogenbonds (hyd) were counted, and specific parameters listed on the right judged whether strands or sheets were formed; suitable directions of the hydrogenbonds were rewarded (sh).*

elements, RMSD around 4.5-5.5 Ångstrœms on average).[10] Fold prediction and other applications of the genetic algorithm for protein structure analysis are also being intensely investigated by other groups [reviewed in 11-13] with the use of blind tests,[14] analysis of small helical and strand containing proteins,[15] peptide library assemblies for fold prediction [16] and simplified general models.[17]

Regarding the third step investigated here, the exploitation of experimental data for prediction refinement, fold prediction by the genetic algorithm can easily incorporate experimental information as additional fitness criteria.[18] Furthermore, theoretical estimates show [19] that the addition of experimental information should be a powerful corrective for protein topology predictions. We illustrate and apply this here to the problem of mispredicted secondary structure.

Materials and methods

Test structures

21 proteins (1IFM, 1PNH, 1PPT, 2OVO, 1MLI, 1EGL, 1HMD, 1GPT, 1EPR, 1DFN, 2CCY, 1CRO, 1CRN, 1TCG, 2CRD, 2BUS, 7PTI, 1BBI, 256B, 1ATX, 2GB1) with different topology and known three dimensional structure served as test cases to compare different secondary structure prediction programs.

Table 2 *Comparing different standard secondary structure predictions.*

(a) performance of standard secondary structure predictions:

Testsequence 1: 1ifm	GVIDTSAVESAITDGEGDMKAIGGYIVGALVILAVAGLIYSMLRKA
observed secondary structure	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
Ptitsyn and Finkelstein [3]	aaaaaaa bbbbbbbbb aaaaaaaaaa
Mehta et al. [1]	bbbbb aaaaaa bbbbbbbbbbbbbbbbbbbb
Frishman und Argos [21]	aaaaaaa aaaaaaaaaaaaaaaa
Rost [20]	bb bbb bbbb bbbaaaaaaaaaaaaaaaa
Levin [2]	bb aaaaabb aaaa aaaaaaaaaaaaaaaaaaaaa
Testsequence 2: 1pnh	TVCNLRRCQLSCRSLGLLGKCIGVKCECVKH
observed secondary structure	aaaaaaaaa bbbbb bbbbb
Ptitsyn and Finkelstein [3]	aaaaaaa aaaaaaa bbbbbbb
Mehta et al. [1]	bbbb aaaaa aaaaaaaaaaaa
Frishman und Argos [21]	aaaaaa
Rost [20]	bbbb bbbbb
Levin [2]	aaaaaaaaaaa bbbb bbb

(b) combined prediction example:

Testsequence 3: 1ppt	GPSQPTYPGDDAPVEDLIRFYDNLQQYLNVVTRHRY		
observed secondary structure (36 residues)	aaaaaaaaaaaaaaaaaaa		
combined prediction schemes:			
	aaaaaaaaaaaaaaaaaaa	5	(33)
	aaaaaaaaaaaaaaaaaaaa	6	(34)
	aaaaaaaaaaaaaaaaaa	12	(35)
	aaaaaaaaaaaaaaaaaaa	24	(33)
	aaaaaaaaaaaaaaaaaaa	25	(33)
	aaaaaaaaaaaaaaaaaaa	40	(33)
	aaaaaaaaaaaaaaaaaaa	42	(33)
Ptitsyn and Finkelstein [3]	aaaaaaaaaaaaaaaaaaaaaaa		
Mehta et al. [1]	aaaaaaaaaaabbbbbbb		
Frishman und Argos [21]	aaaaaaaaaaaaaaaaaaa		
Rost [20]	aaaaaaaaaaaaaaaaaa		
Levin [2]	aaaaaaaaaaaaaaaaaaaa		

Predictions by Ptitsyn and Finkelstein [3], by Mehta et al. [1], by Frishman and Argos [21], by Rost [20] and by Levin [2] were compared. The top line denotes in each example the brookhaven file structure and amino acid sequence, followed by the observed secondary structure (a helix, b strand). Next there are given the secondary structure assignments according to the prediction programs (a helix, b strand). Table 2(b) has the same format as in (a), but several combination schemes are investigated (listed on the right; the first number indicates the scheme tested; the second number in brackets shows how many residues from the total of 36 residues in the example are correctly predicted).

Secondary structure prediction

Different secondary structure prediction programs were run using original executables from the authors (either available locally to EMBL by license or freely on the Web). Thus, a range of different strategies for secondary structure prediction could be compared: Pure stereochemical considerations are utilised for secondary structure prediction in the algorithm developed by Ptitsyn and Finkelstein.[3] Profile based neural networks are used by Rost.[20] Frishman and Argos [21] utilise local pairwise alignment of the sequence to be predicted with each related sequence rather than utilisation of multiple alignment. Metha et al. [1] use residue exchange weight matrices relying only on amino acid substitutions of structurally related proteins. Levin's program [2] assigns secondary structure comparing the blosum 62 similarity scores of best matching fragments in a database of known structures. The output for each secondary structure prediction pro-

Table 3a *Different combinatorial schemes for secondary structure prediction tested – First strategies*

Consensus predictions

scheme 1: Accept first helix predictions, next coil predictions from the consensus, believe strands only when there is a clear majority.

scheme 2: Accept only clear majorities, otherwise leave region coil

Combine only the best prediction programs

scheme 3: Accept predictions by PHD (best secondary structure prediction program in the comparison), but change strand regions to helices if the region is predicted helical by Alb (seems to be a good helix prediction for small proteins)

scheme 4: Accept predictions by predator, but change strand regions to helices if the region is predicted helical by Alb (good helix prediction for small proteins)

Improve a good triple combination

scheme 5: combine PHD, predator and simpa; accept all predictions where two of them agree.

scheme 6: like scheme 5 but if PHD and predator assign no helix or strand region fill in secondary structure predicted by simpa.

scheme7: like scheme 5 but if predator and simpa assign no helix or strand region fill in secondary structure predicted by PHD.

scheme 8: like scheme 5 but if PHD and simpa assign no helix or strand region fill in secondary structure predicted by predator.

Use one program alone

scheme 9: Use the prediction by Alb and assign accordingly helix or strand regions.

scheme 10: Use the prediction by sspred and assign accordingly helix or strand regions.

scheme 11: Use the prediction by predator and assign accordingly helix or strand regions.

scheme 12: Use the prediction by PHD and assign accordingly helix or strand regions.

scheme 13: Use the prediction by simpa and assign accordingly helix or strand regions.

The different secondary structure prediction programs used are abbreviated as follows: Alb [3], sspred [1], predator [21], PHD [20] and simpa [2]. They are combined applying the rules listed.

gram on the complete test set of proteins with known three dimensional structure was collected and stored. Applying custom written programs in VAX Pascal that could read the prediction output from the different programs, the performance of each secondary structure prediction program was noted as well as the performance of various combinations calculated. Further, individual prediction program combinations for helical regions, strand regions and turn regions were investigated. The following combinations of secondary structure prediction programs were investigated and compared to each other: Different ways to achieve consensus predictions (scheme 1 and 2); taking one of the two best overall secondary structure prediction programs and trying to improve this further by an additional prediction for helical regions (scheme 3 and 4); combine the three best prediction programs to make conservative, careful predictions (scheme 5) and then try to improve unassigned regions (schemes 6-8); the five standard prediction programs investigated (scheme 9-13); combinations of two predictions (scheme 14-19); combinations of three predictions (scheme 20-29); combinations of three predictions, further improved in strand regions by additional prediction (scheme 30-39). Apply best helix and best strand specific prediction combinations (scheme 40). Apply only best strand and best helix prediction program (scheme 41). Extend too small strand regions from scheme 40 (scheme 42). The high probability regions derived from one individual prediction scheme predicted then nuclei of secondary structure deemed to be certain which were kept fixed during the simulation.

Protein folding simulations

Folding simulations [8,10] implemented basic protein building principles [22] as fitness criteria in a genetic algorithm. The protein main chain (C, O, N, and C_α) was modelled. Different Φ and Ψ dihedral values were taken for each residue from a set of seven possible standard conformations, representative of frequently populated regions in known tertiary structures.[23]

Side chain atoms were not represented explicitly. Instead different amino acid properties are implicitly taken into account by some of the fitness criteria (Table 1): The secondary structure propensity of amino acids for helix, strand or coil regions is taken into account by the secondary structure prediction used as starting information, as well as by the cooperative growth and preference terms for new or already present secondary structure applied in the fitness function

Table 3b *Different combinatorial schemes for secondary structure prediction tested – Systematic permutations*

Combine two programs

scheme 14: Use the prediction by Alb (conservative in predictions for helices and strands), if coil is assigned, fill in the coil regions additional helices or strands regions predicted by sspred.

scheme 15: Use the prediction by Alb (conservative in predictions for helices and strands), if coil is assigned, fill in the coil regions additional helices or strands regions predicted by simpa

scheme 16: Use the prediction by sspred, if coil is assigned, fill in the coil regions additional helices or strands regions predicted by Alb.

scheme 17: Use the prediction by sspred, if coil is assigned, fill in the coil regions additional helices or strands regions predicted by simpa.

scheme 18: Use the prediction by simpa, if coil is assigned, fill in the coil regions additional helices or strands regions predicted by sspred.

scheme 19: Use the prediction by simpa, if coil is assigned, fill in the coil regions additional helices or strands regions predicted by Alb.

C*ombine three prediction programs*

scheme 20: Where at least two of the predictions by Alb, sspred and predator agree, assign helix or strand regions.
scheme 21: Where at least two of the predictions by Alb, sspred and PHD agree, assign helix or strand regions.
scheme 22: Where at least two of the predictions by Alb, sspred and simpa agree, assign helix or strand regions.
scheme 23: Where at least two of the predictions by Alb, predator and PHD agree, assign helix or strand regions.
scheme 24: Where at least two of the predictions by Alb, predator and simpa agree, assign helix or strand regions.
scheme 25: Where at least two of the predictions by Alb, PHD and simpa agree, assign helix or strand regions.
scheme 26: Where at least two of the predictions by sspred, predator and PHD agree, assign helix or strand regions.
scheme 27: Where at least two of the predictions by sspred, predator and simpa agree, assign helix or strand regions.
scheme 28: Where at least two of the predictions by sspred, PHD and simpa agree, assign helix or strand regions.
scheme 29: Where at least two of the predictions by predator, PHD and simpa agree, assign helix or strand regions.

The different secondary structure prediction programs used are abbreviated as follows: Alb [3], sspred [1], predator [21], PHD [20] and simpa [2]. They are combined applying the rules listed.

during the genetic algorithm simulation. Furthermore, hydrophobic amino acids (Phe, Tyr, Met, Cys, Ile, Leu, Val, Trp) are packed together tighter and more central to the protein core than other amino acids.

The standard conformations (3 bits encoding one standard conformation) of all residues along the amino acid sequence were successively collected together and decoded from a long bit string (a „chromosome“). Starting from a population of random bit strings, the quality of each encoded structure was judged by a fitness function composed of rewards and penalties. The total fitness (Table 1) measured the quality of the structure encoded by an individual bit string. This is the sum of the general fitness terms (criteria 1-4) and the beta-strand fitness terms (criterion 5):

1) the total scatter of all (n) residue C_α-atoms (res), each(j) with coordinates (x,y,z) around their common centre of mass (C_m),

$$\sum_{i=1}^{n}\sqrt{\sum_{j=x,y,z}(res_{ij}-Cm_{ij})^2} \qquad (1)$$

2) distribution of hydrophobic residues only (M, I, L, V, Y, C, F and W) around the centre of mass (same centre as in 1);

3) mainchain van-der-Waals atom overlaps;

4) conformational states that agree with the secondary structure (either known or predicted) for a given subsequence and

5) the selection for the formation and direction of hydrogen bonds in beta-strands and beta-sheets and the formation of reverse turns in beta-hairpins (Table 1).

Helix and strand regions predicted with high accuracy (by an optimised choice or combination of secondary structure programs; see methods for secondary structure prediction) were kept fixed in an appropriate standard conformation [23] during the simulation. The genetic algorithm kept these nuclei of secondary structure elements fixed, but operated freely on all other residues (including extension and deletion of new elements and extending or limiting the regions deemed certain). These criteria are sufficient to predict (RMSD 4-6 Ångstrœms between observed and predicted C_α-atoms) the mainchain topology of all-helical folds [8] starting from sequence and secondary structure prediction without any experimental information. Further, 19 proteins, (most with fewer than 100 amino acids) and with different topologies and secondary structural types, could be similarly predicted with sequence information and known secondary structure.[10] Parameters and suitable weights were determined empirically

Table 3c *Different combinatorial schemes for secondary structure prediction tested - Further refinement*

Predict helices by three predictions, strands by PHD

scheme 30: Where at least two of the predictions by Alb, sspred and predator agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

scheme 31: Where at least two of the predictions by Alb, sspred and PHD agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

scheme 32: Where at least two of the predictions by Alb, sspred and and simpa agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

scheme 33: Where at least two of the predictions by Alb, predator and PHD agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

scheme 34: Where at least two of the predictions by Alb, predator and simpa agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

scheme 35: Where at least two of the predictions by Alb, PHD and simpa agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

scheme 36: Where at least two of the predictions by sspred, predator and PHD agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

scheme 37: Where at least two of the predictions by sspred, predator and simpa agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

scheme 38: Where at least two of the predictions by sspred, PHD and simpa agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

scheme 39: Where at least two of the predictions by Alb, PHD and simpa agree, assign helix regions. Strand regions are assigned by PHD, over-ruling the first rule.

Refining the best combinations

scheme 40: Rule for strand prediction: Where at least two of the predictions by predator, PHD and simpa agree, assign strand regions. Rule for helix regions: Where at least two of the predictions by Alb, predator and simpa agree, assign helix regions, over-ruling the first rule.

scheme 41: Strand regions are assigned by PHD, Alb assigns helix regions, over-ruling the first rule.

scheme 42: Rule for strand prediction: Where at least two of the predictions by predator, PHD and simpa agree, assign strand regions. Rule for helix regions: Where at least two of the predictions by Alb, predator and simpa agree, assign helix regions, over-ruling the first rule. Single residues determined by these two rules to be in strand conformation and found in a coil region are extended by two further C-terminal strand residues; two strand residues found in a coil region are extended by one further C-terminal strand residue.

The different secondary structure prediction programs used are abbreviated as follows: Alb [3], sspred [1], predator [21], PHD [20] and simpa [2]. They are combined applying the rules listed.

by many simulations on these different protein topologies. Different implementations and weights for various fitness function criteria were studied to optimise the structures selected during the genetic algorithm simulation.[8,10]

High quality bit strings (after a random start) were selected preferentially as parents and mutated (1 bit per string per generation was mutated on average, choosing a random position on the string) and recombined through cross-over to yield the next parental generation of folds (probability of recombination was 0.2 per bit string per generation and occured at exactly one equivalent site chosen at one random bit on each of the parental chromosome pairs). A positive constant kept the population of prediction trials richer, since low fitness individuals may also survive (C in Table 1). Simulations were run over many generations to allow convergence (the product of population and generation equals 4 x 10^5, corresponding to a processing time for one simulation run of 20 minutes on a VAX 7620 for a 46-residue protein). At least ten simulations with different random starting populations, found to be sufficient to achieve good predictions in various test structures, were investigated for each fold prediction. The structure with the highest fitness value from the trials was taken as that predicted.

Results

Small proteins with known three dimensional structure were analyzed with respect to accuracy of secondary structure prediction and genetic algorithm based fold prediction. Two different strategies to correct mispredictions from incorrect secondary structure prediction were examined. The first investigated different combinations of secondary structure prediction programs, the second strategy applied folding simulations. Neither required experimental data.

Table 4a *The best secondary structure prediction combinations – Single residue scores* (three state prediction, helix, strand, coil) *[a]*

Scheme	correctly assigned	not assigned	wrongly assigned	net result
5	960	232	94	866
6	967	182	137	830
12	927	203	156	771
23	954	251	81	873
24	955	243	88	867
25	929	244	113	816
39	973	184	129	844
40	961	235	90	871
42	964	225	97	867

Table 4b *The best secondary structure prediction combinations – Analyzing prediction of secondary structure elements (e.g. scheme 42) [b]*

secondary structure	present	crashed	found	wrong
helices	42	4	31	4
strands	44	4	26	7

Testing accuracy of different secondary structure predictions

Table 2 shows example files of proteins with known three dimensional structure and different predictions of their secondary structure. Five standard secondary structure predictions were systematically compared on 21 proteins with different topology and known structure: The prediction algorithms by Ptitsyn and Finkelstein [3], by Metha et al. [1], by Frishman and Argos [21], by Rost [20] and by Levin [2]. As each of these has a different method to determine secondary structure (see Materials and Methods) they yield different predictions on the same protein (Table 2a).

Next we investigated whether combinations would improve the overall secondary structure prediction accuracy. An example comparing several different combinations is shown in Table 2b. We then investigated in detail different combinations of secondary structure prediction programs and decision rules to optimise the resulting combined prediction (Table 3a-3c). Note that for further improvement of these secondary structure predictions by the genetic algorithm simulations (next part of the results) more conservative estimates are valuable: They predict only the sure, central regions for the secondary structure elements. Such nucleus regions are then kept fixed during the genetic algorithm simulation (see Materials and methods). However, the genetic algorithm may extend or limit these. Further, the genetic algorithm may fill in new secondary structure elements or introduce and extend coil regions and modify or delete them again during the simulation in all other parts of the protein.

The various combinations tested followed a systematic exploration of the possibilities to derive an accurate prediction. A first set of rules tried to achieve a consensus among all predictions. Problems arise only when their is no clear majority and we investigated the outcome of a preference for helix prediction (scheme 1) as well as the outcome of a conservative coil estimate (scheme 2). The next two schemes which test the opposite strategy, rely only on the best prediction programs, i.e. predator [21] and PHD [20] for strands and Alb [3] for helices. All the following schemes (Table 3a-c) persue intermediate strategies (see Materials and methods), trying to combine both the advantages from consensus predictions and the advantages of a program that performs particularly well on some type of secondary structure element. Before a further scheme was introduced and tested, the performance of the previous combinations was noted so

[a] A total of 1286 residues was analyzed. Predictions by Ptitsyn and Finkelstein [3], by Mehta et al. [1], by Frishman and Argos [21], by Rost[22] and by Levin [2] are compared and combined, the rules associated with each scheme are listed in Table 3. The performance and advantages of the specific schemes shown compared to all other schemes is explained in the text. The net result indicates correctly assigned residues (as strand or helical state) minus wrongly assigned residues (helix state instead of strand conformation or strand conformation instead of helix state). In all not assigned regions random coil state is assumed

[b] Prediction schemes as listed in Table 3. Mispredictions of secondary structure elements can either break ("crash") the correct secondary structure element (e.g. if a strand is predicted instead of a helix in the observed structure) or misassign ("wrong") structure in coil regions (e.g. a helix is assigned though there is a coil region in the known three dimensional structure here).

Table 4c *The best secondary structure prediction combinations. A total of 1286 residues was compared. The rules associated with each scheme are listed in Table 3. "net" indicates the total number of correctly found secondary structure elements minus those elements mispredicted.*

	Sub1[a]	Sub2[b]
the best helixpredictor is the prediction:	40	40
with the net value being:	26	22
the best betapredictor is the prediction:	12	42
with the net value being:	16	8
the best overall predictor is the prediction:	42	42
with the net value being:	38	27

Table 4d *Comparison of the best secondary structure prediction combinations with standard secondary structure prediction programs [c]*

Method	correctly assigned	not assigned	wrongly assigned	net result
Alb	786	257	243	543
sspred	738	214	334	404
predator	929	238	119	810
PHD	927	203	156	771
simpa	923	207	156	767

as to be available for further improvement. Thus the performance of the best strand predictor PHD could be outperformed by a careful combination of three secondary structure prediction programs. It turned out that the sequence of secondary structure assignment is important, i.e. whether strand assignment or helix assignment is done last, overruling in some regions the assignments made before. An extension rule for too short elements proved to be important for maximum performance (scheme 42).

A further point was also considered: The secondary structure prediction accuracy for the different combinations (Table 4) depends on how accuracy of prediction for secondary structure is defined. Several ways to measure prediction accuracy are shown and were calculated for all of the combinations tested: Per residue prediction (4a), prediction of secondary elements (4b), and different punishment weights if a secondary structure element is misassigned in coil regions (4c). For any of the criteria investigated, a suitable combination of predictions performs better than any of the standard prediction programs alone (4d; the standard prediction programs are the schemes 9-13 in Table 3a). Table 4a illustrates that all the combinations of secondary structure prediction programs shown achieve better prediction accuracy for single residues than a single secondary structure program. This is the case both for the amount of correctly predicted residues and for the net result after subtracting wrong assigned residues (the results for the standard secondary structure programs are listed in Table 4d). Several schemes are compared: The current best neural network predictor for secondary structure (scheme 12; by Rost [20]). The best combined prediction scheme maximising correctly assigned single residue states is scheme 39. However, another combined scheme (23) achieves the highest score if wrongly assigned states (e.g. predicting helix where a strand is present in the observed structure) are taken into account and subtracted from the score („net result“, Table 4a).

Good prediction schemes for whole secondary structure elements perform only slightly less well on single residue predictions and still better on this task than the uncombined prediction programs do. Compared are the two schemes that have the strand rule (scheme 24) and the helix rule (scheme 5), which are combined in the scheme 42 (also shown), as well as the somewhat weaker performance of two other related schemes (scheme 6 and scheme 25). The performance on single residue prediction for the best prediction program

[a] Every element counted equally and both types of misprediction are subtracted once.

[b] Every element counted equally, broken (see Table 4b) secondary structure elements are subtracted once, but wrong predictions are substracted twice.

[c] A total of 1286 residues was analyzed. The original prediction programs Alb by Ptitsyn and Finkelstein [3], sspred by Mehta et al. [1], predator by Frishman and Argos [21], PHD by Rost[22] and simpa by Levin [2] are compared, the rules associated with each scheme are listed in Table 3. The net result indicates correctly assigned residues (as strand or helical state) minus wrongly assigned residues (helix state instead of strand conformation or strand conformation instead of helix state). In all not assigned regions random coil state is assumed

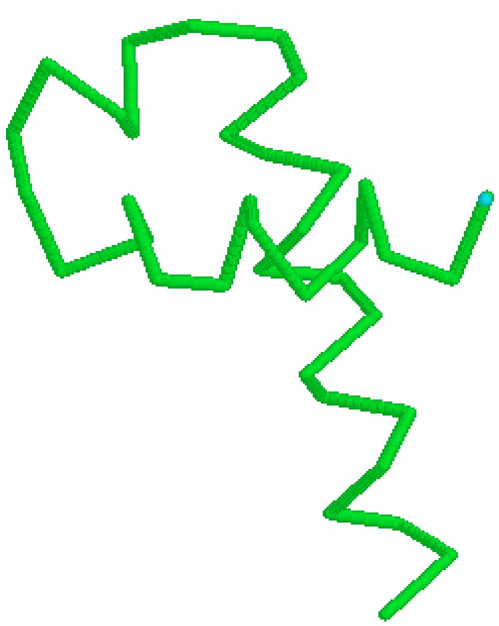

Figure 1a *Mating pheromone (1ERP); simulation result, C_α-RMSD to observed 5.0 Å*

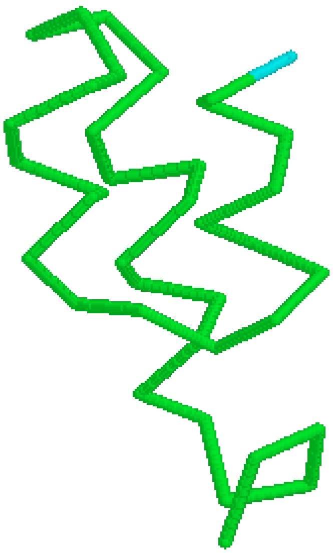

Figure 1b *Mating pheromone (1ERP); experimentally determined structure; the mainchain backbone is shown in green and the first N-terminal residue indicated in cyan*

to predict helical secondary structure elements is also shown (scheme 40).

The performance of the best scheme to predict secondary structure elements (if both strands and helices are considered) is shown in detail in Table 4b (scheme 42).

Furthermore, Table 4c gives the optimal results and schemes for several combinations of prediction programs looking only at the strands or helices predicted or the overall performance. The choice of the optimal strategy to predict secondary structure elements depends in addition on how strong mispredictions (predicting the wrong secondary structure or predicting secondary structure in loop regions) are punished. This proved to be useful in identifying prediction combinations optimal as input for protein folding simulations.

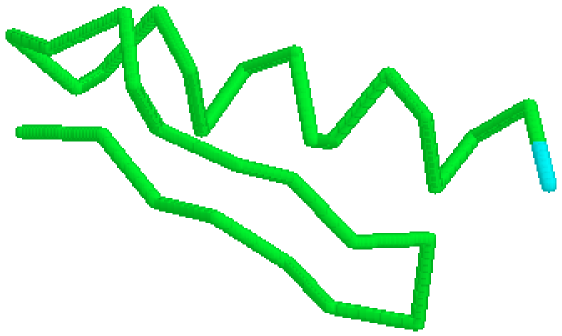

Figure 2a *Scorpion toxin (1PNH); simulation result, C_α-RMSD to observed 5.7 Å*

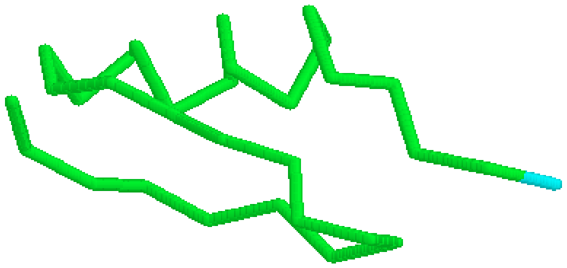

Figure 2b *Scorpion toxin (1PNH); experimentally determined structure; the mainchain backbone is shown in green and the first N-terminal residue indicated in cyan*

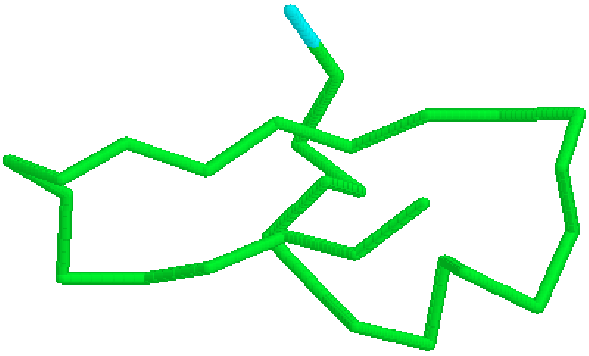

Figure 3a *Defensin (1DFN); simulation result, C_α-RMSD to observed 5.1 Å*

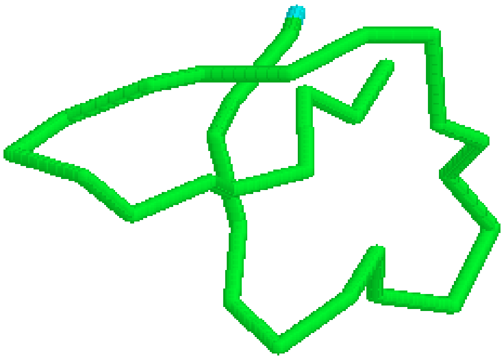

Figure 3b *Defensin (1DFN); experimentally determined structure; the mainchain backbone is shown in green and the first N-terminal residue indicated in cyan*

Correcting mispredictions exploiting the tertiary folding simulations and the genetic algorithm.

The next set of experiments investigated the ability of the genetic algorithm folding simulations to correct mispredicted secondary structure remaining after the optimisation trials on the secondary structure predictions. The best scheme (42) for prediction of secondary structure elements was used in these trials. Comparing the 21 different protein topologies, the genetic algorithm improved the accuracy of the secondary structure prediction during the simulations in 17 of the 21 cases. Conditions where secondary structure elements (helices, strands) were completely overlooked by the secondary structure prediction became increasingly challenging the more secondary structure elements had been missed. Here also topology predictions could be achieved. Note, however, that the proteins or domains to be predicted are not big (22-121 amino acids) and have no more than six secondary structure elements, which is advantageous in the sampling of the conformational space.

Three examples are shown, simulation result and experimentally determined structure are viewed from the same perspective. In each case a batch of ten simulations is compared and the fittest fold obtained is studied.

In the first example (Figure 1; 1ERP, mating pheromone) the genetic algorithm started with a secondary structure prediction where only the N-terminal helix was correctly given and the following two helices were completely missed. The folding simulation corrected for these missing two elements by filling in helical regions, however, some topological error is still left.

The next example (Figure 2; 1PNH, scorpion toxin) shows a topology close to the experimentally observed. However, two C-terminal strands were too small and shortened by the secondary structure prediction scheme and had to be extended correctly during the simulation.

In the third example (Figure 3; 1DFN, defensin) the N-terminal strand has been completely missed by secondary structure prediction and the following two strands were only partly given. Nevertheless, the algorithm achieved a reasonable topology prediction during the simulation.

There are from 1-6 secondary structure elements (most often 4 elements) present in the 21 protein structures tested (22-121 amino acids in length). Our results indicated that a loss of one secondary structure element leads to slightly higher RMSD (for values see figures), but is otherwise well tolerated in the topology prediction. A second missed element only changes the topology in some cases. However, loss of more elements in the starting information for the algorithm leads to increasing topological error. This parameter is thus more important than the correct placement of secondary structure elements in the genetic algorithm simulations (several misplacements can be corrected during the simulation). Good starting data for the genetic algorithm simulations are provided by combining secondary structure predictions in such a way that only few secondary structures are completely missed, a high number of correctly predicted residue states is less important.

Applying experimental data

A third step in the refinement of protein fold predictions applies and combines experimental data to test and refine predictions both for secondary structure and tertiary folds. Additional information on secondary structure, distance and structural constraints can be exploited to test, correct and improve fold predictions.[19] This is currently used in several application examples of our method. Our fold prediction method has the advantage that such information can easily be incorporated as further fitness criteria for the genetic algorithm. Correcting constraints [18,24,25] such as domain boundaries, distance constraints from different experimental data, including S-S bonds or antibody epitopes, but also experimental data on secondary structure elements are sometimes available and may help to achieve tertiary fold predictions which can be supported by experiment. An example shown is a domain in the large subunit of RNA polymerase II from *Drosophila* (Figure 4). A more detailed picture of the domain is desirable because the domain and its interactions are involved in a number of steps catalyzed by the large subunit of RNA polymerase. We are studying the interaction of an antibody with this domain. Residues known from experimental data to interact with the antibody (single chain antibody scFv215, [26]) are shown in cyan. The antibody epitope as calculated by this simulation is accessible to the antibody in accordance with the experimental data. However, this model is currently being developed and refined incorporating feed-back and data from experiment to minimise errors from secondary and tertiary structure prediction. Both secondary structure prediction and folding simulation have not yet assigned a specific secondary structure to this region (depicted as loop region in the model). Further experiments are currently investigating the secondary structure in this region in more detail. This includes epitope peptide scanning [27] to test and refine the prediction for this binding region. The result of the peptide scanning experiment will be used to refine the secondary structure prediction and start the next round of genetic algorithm simulations.

Discussion

Secondary structure prediction has only a limited accuracy, the difficulties in considering long range interactions being one of the major challenges.[4] The present study examines three different strategies to minimise errors caused by secondary structure prediction in the resulting protein fold prediction:

(1) We show that combining different secondary structure programs for the prediction of secondary structure can outperform the results from any of the single methods. The results are quantified for different measures of secondary structure accuracy.

(2) New secondary structure elements are also created and all elements present are optimised in length during folding simulations with the genetic algorithm. The genetic algorithm is able to improve the secondary structure prediction during the folding simulation in 17 out of the 21 protein cases studied.

(3) Experimental information is shown to be important for final refinement in application examples.

Ad (1): The accuracy of different secondary structure predictions and their combinations was examined in detail. On their own, in this comparison neural network-based methods (Rost's program PHD,[20]) and Frishman and Argos [21] achieved best results (3 state prediction). However, we note that the programs by Levin [2], Metha et al.[1] and the old

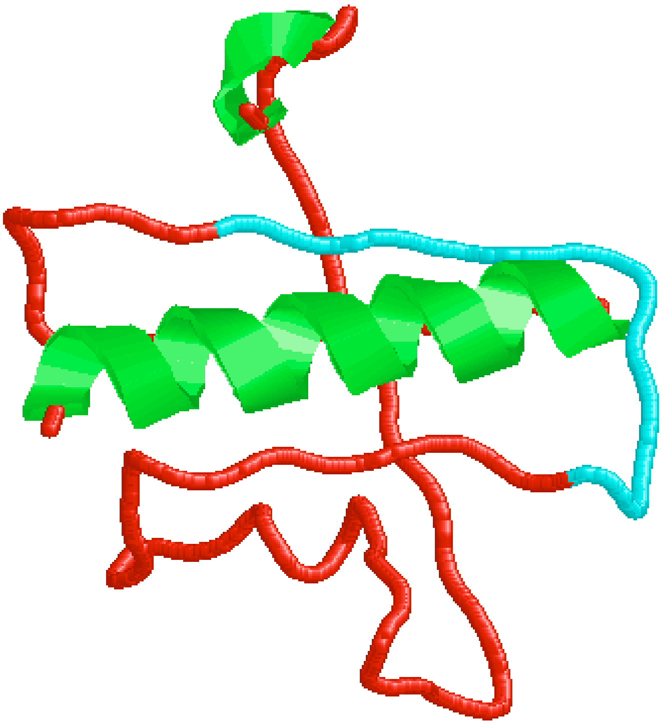

Figure 4 *Domain from the large subunit of RNA polymerase II. Overall topology prediction shown in red, helices are shown in green, antibody epitope is shown in cyan. The next refinement step will be the incorporation of additional experimental data studying the secondary structure around the antibody epitope*

stereochemical model by Ptitsyn and Finkelstein [3] perform not much worse. Different secondary structure elements are overlooked by different programs. As these approaches are independent strategies and concepts, we reasoned that a consensus approach might be even better and tested systematically different combinations (Table 3).

Several decision rules for combining predictions by secondary structure prediction programs can be shown to improve results in comparison to the secondary structure programs on their own, at least on the test set investigated (21 different topologies from proteins with known crystal structure and 22-121 amino acids in length). An optimal input for the genetic algorithm is a more conservative prediction (fewer residues are assigned) which makes few wrong predictions on secondary structure elements (scheme 42) as further secondary structure is filled in by the genetic algorithm and extended or shortened during the simulation. However, in other applications, a maximum number of correctly assigned residues may be important and corresponding optimal strategies are given (Table 4a). Additional investigations will analyse the different rules and combinations described here for their performance on larger sets of protein structures.

Ad (2): Another set of experiments examines to what extent secondary structure is optimised, and by this corrected through the combination of all the fitness criteria used during the genetic algorithm simulation. An improvement was seen in 17 of the 21 protein structures tested. Three examples with different topologies and challenges are illustrated.

This will be studied further. Additional criteria to judge the simulation outcome will be developed to identify where the simulation by the genetic algorithm failed to correct mistakes from secondary structure prediction and leads to a wrong topology prediction. Further research will analyse additional prediction approaches for the interplay between secondary and tertiary structure including turn prediction accuracy and overall fitness of predicted structures according to different criteria.

Ad (3): In application examples, feed-back from experiment is important for further refinement. The topology of the antibody binding regions within a domain predicted by the genetic algorithm is a simple case in point.

Outlook: The interplay between secondary structure (mainly local interactions) and tertiary fold (mainly global interactions) is a fascinating challenge. Considering the set of proteins we studied, we show that further improvement of secondary structure prediction can efficiently be achieved either by a suitable combination of different secondary structure prediction strategies („local“) or by further refinement applying the genetic algorithm („global“). For more and larger protein structures as well as new application examples, exploitation of experimental data and refined structure criteria such as additional packing rules will be applied and further developed.

Acknowledgement We thank Enrique Morett, Louise Kelly, Sean O’ Donoghue and Richard Copley for critical reading of the manuscript and stylistic corrections.

References

1. Mehta, P. K.; Heringa, J.; Argos, P. *Protein Science* **1995**, *4*, 2517.
2. Levin, J. M. *Protein Eng.* **1997**, *10*, 771.
3. Ptitsyn, O. B.; Finkelstein, A. V. *Biopolymers* **1983,** *22*, 15.
4. Rost, B. *Proteins* **1997**, suppl *1*:192.
5. Aszodi, A.; Gradwell, M. J.; Taylor, W. R. *J. Mol. Biol.* **1995**, *251*, 308.
6. Goldberg, D. *Genetic algorithms in search, optimization and machine learning*. Massachusetts, 1989.
7. Dandekar, T.; Argos, P. *Protein Eng.* **1992**, *5*, 637.
8. Dandekar, T.; Argos, P. *J. Mol. Biol.* **1994**, *236*, 844.
9. Kabsch, W.; Sander,C. *Biopolymers* **1983**, *22*, 2577.
10. Dandekar, T.; Argos, P. *J. Mol. Biol.* **1996**, *256*, 645.
11. Clark, D. E.; Westhead, D. R. *J. Comp.-Aided Mol. Design* **1996**, *10*, 337.
12. Pedersen, J. T.; Moult, J. *Curr. Op. Struc. Biol.* **1996**, *6*, 227.
13. Dandekar, T.; Koenig, R. *Biochimica et Biophysica Acta* **1997**, *1340*, 1.
14. Pedersen, J. T.; Moult, *J. Proteins* **1995**, *23*, 454.
15. Sun, S.; Thomas P. D.; Dill, K. A. *Protein Eng.* **1995**, *8*, 769.
16. Bowie, J. U.; Eisenberg, D. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 4436.
17. Unger, R.; Moult, J. *J. Mol. Biol.* **1993**, *231*, 75.
18. Dandekar, T.; Argos, P. *Protein Engineering* **1997**, *10*, 877.
19. Sibbald, P. R. *J. Theor. Biol.* **1995**, *173*, 361.
20. Rost, B. *Methods Enzymol.* **1996**, *266*, 525.
21. Frishman, D.; Argos, P. *Proteins* **1997**, *27*, 329.
22. Schulz, G. E.; Schirmer, H. R. *Principles of protein structure*. New York, 1979.
23. Rooman, M. J.; Kocher, J.-P. A.; Wodak, S. J. *J. Mol. Biol.* **1991**, *221*, 961.
24. Dandekar, T.; Leippe, M. *Folding and Design* **1997**, *2*, 47.
25. Saxena, P.; Whang, I.; Voziyanov, Y.; Harkey, C.; Argos, P.; Jayaram, M.; Dandekar,T. *Biochimica et Biophysica Acta* **1997**, *1340*, 187.
26. Kontermann, R. E.; Liu, Z.; Schulze, R. A.; Sommer, K. A.; Queitsch, I.; DŸbel, S.; Kipriyanov, S. M.; Breitling, F.; Bautz, E. K. F. *Biol. Chem. Hoppe-Seyler* **1995**, *376*, 473.
27. Kramer, A.; Schneider-Mergener, J. Meth. *Mol. Biol.* **1998**, *87*, 25.
28. Karasawa, T.; Tabuchi, K.; Fumoto, M.; Yasukawa, T. *Comp. Appl. Bio. Sci.* **1993**, *9*, 243.