Theor Chem Acc (1999) 101:57–61
DOI 10.1007/s002149800m66

Theoretical
Chemistry Accounts

Regular article

SeqFold – fully automated fold recognition and modeling software – evaluation and application*

Krzysztof A. Olszewski, Lisa Yan, David J. Edwards

Molecular Simulations, 9685 Scranton Road, San Diego CA 92121, USA

Received: 25 May 1998 / Accepted: 4 August 1998 / Published online: 2 November 1998

Abstract. SeqFold is a fold recognition program based on sequence-similarity detection aided by predicted secondary structure [1–3]. Critical validation and evaluation of SeqFold fold recognition performance based on the latest Critical Assessment of protein Structure Prediction (CASP2) targets has been performed. It has revealed that four out of seven CASP2 threading targets were assigned a correct fold using this method. SeqFold has also been applied to the problem of fold recognition for leptin. Mice with a defective leptin gene are extremely obese and diabetic. Leptin does not exhibit clear sequence homology to any protein with known structure. SeqFold predicts that leptin belongs to the class of short-chain four-helical cytokines. The structure of leptin, which has recently been solved by X-ray crystallography, reveals that leptin is a long-chain four-helical cytokine. The 3D model of leptin demonstrates that SeqFold alignment-based homology modeling captures essential features of the leptin structure.

Key words: Fold recognition – Threading – CASP2 – SeqFold

1 Introduction

Recent advances in whole genome scale sequencing deliver new putative protein sequences at enormous speed. Complete new genomes of small microorganisms are now being published almost every month and mammalian genomes are expected at the break of the new millennium. This wealth of information offers researchers an opportunity to describe, model, and possibly understand whole living organisms (see, for example, Ref. [4]). However, we are rapidly approaching the limit of information that may be drawn from sequence data. This limit stems from the fact that only a relatively small number of sequences can be confidently annotated with the predicted function and/or structure. The most optimistic estimates conclude that about 50–60% of all sequences in a novel genome may be annotated based on similarity to previously characterized sequences [5]; however, more confident structural annotation can be expected for only about 20% of sequences.

Homologous sequences (i.e., sequences that share a common ancestor) are likely to preserve overall structure and function, regardless of their residue-by-residue similarity. At the level of 30% or less a number of sequences appear to be homologous, but amino acid similarity is only superficial. When two sequences are optimally aligned the range of 20–30% amino acid identity, which, used as a measure of homology of the two sequences, is customarily called the twilight zone [6]. Unfortunately, a large majority of homologous sequences exhibit less than 25% amino acid identity [3], even though functional or structural similarity is preserved. Hence, the limit of information derived from sequence-based annotation is quickly exhausted. Because of the overwhelming number of new sequences it is vital that annotation methods should be automated, so that the structure/function recognition process is relatively effortless.

A protein structure is determined by its amino acid sequence alone [7]. Fortunately, the number of possible protein structures (folds) is predicted to be far less that the number of sequences [8]. It might therefore be possible to test how well a novel sequence folds into an already known protein structure. The subclass of sequence-structure space searching algorithms that attempts to solve such a question is known as a threading algorithm. Specifically, algorithms that perform a search for a given novel sequence and return the most compatible structure are called fold recognition algorithms [9]. There are a number of threading algorithms that differ by the sequence-structure scoring function and by the alignment algorithm used to optimize an alignment for a specific scoring function [10–11].

* Contribution to the Proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France

Correspondence to: K.A. Olszewski
e-mail: kato@msi.com, Tel.: +1-619-5465347,
Fax: +1-619-4580136

So far, one of the most successful attempts to solve the fold recognition problem is to explore the twilight zone of protein similarity by enhancing sequence-based similarity scores with the predicted secondary structure of the sequence under study [1–3]. In this paper we study the performance of SeqFold, our implementation of such a method using fold recognition targets from the second edition of the critical assessment of protein structure prediction (CASP2) experiment [12]. Also, we test the dependence of fold recognition prediction results on the accuracy of the secondary structure prediction and we demonstrate the application of SeqFold as the alignment provider for homology model construction using yet another example – the human hormone, leptin.

Two editions of the CASP2 experiment have already been summarized at meetings in Asilomar in 1994 and 1996. There are four categories of predictions, namely: homology modeling, threading, ab initio folding, and docking. To establish a common basis for comparison of the various techniques all participating predictors do not know the structures of their targets beforehand, which makes it a true blind prediction contest. For the purpose of validation, in this paper we have used seven targets from the threading category of CASP2. Obviously, this does not represent a blind prediction; CASP2 targets are used purely as a representative, independent test set for the validation of our tool.

As an independent example of SeqFold fold recognition performance, SeqFold has been tested on the human form of the leptin hormone. Leptin is a product of the obese (OB) gene. OB gene defects in rodents lead to extreme obesity, hypherphagia, diabetes, and infertility. In humans, OB gene defects are linked to early-onset obesity and inhibited sexual development (see, for example, Ref. [13]). Sequence-similarity methods fail to indicate any links between the leptin family and any other sequence. The recent structure solution by X-ray [14] reveals that it possesses a four-helix structure which corresponds to the family of long-chain cytokines. Hence, it is a perfect target for testing the fold recognition method.

2 Theory

Sequence-similarity score between target (query) sequence and the reference structure is computed in SeqFold using sequence-similarity matrices such as Gonnet [15] or Blosum62 [16]. These matrices represent the relative, observed chances of mutation of amino acid i to amino acid j with respect to the estimated random chance of such mutation, i.e.,

$$s_{ij} = \log \frac{\text{observed}_{ij}}{\text{expected}_{ij}}$$

In the twilight zone, sequence-similarity searches are not always selective enough to distinguish between a true homologue of the target sequence and a number of false positive hits. It has been demonstrated that inclusion of the predicted secondary structure in the scoring function increases the odds of identification of a correct homologue missed by the sequence-only method [1]. The source of the improvement is mainly through the increased selectivity of the scoring function [17]. The modified sequence-structure similarity score for the target amino acid i and the reference structure position j has the following form:

$$st_{ij} = w^{\text{seq}} \cdot s_{ij} + w^{\text{str}} \cdot c_i \cdot t_{ij},$$

where s_{ij} and t_{ij} are sequence- and structure-similarity matrices, respectively. s_{ij} is defined as above and t_{ij} equals 1 if the secondary structure at position i is a helix or a strand and equals to the secondary structure at position j. Otherwise t_{ij} equals 0. c_i values range from 0 to 1 and represent confidence in the secondary structure prediction of the ith position of the target sequence, and depend on the secondary-structure prediction algorithm. In the case of Chou-Fasman, GOR, and DSC predictions, c_{i} is equal to the probability of the most likely secondary structure at position i. In the case of PHD, the prediction reliability index has been used. w^{seq} and w^{str} are relative weights of sequence and structure contribution to the total score. Default values for w^{seq} and w^{str} are 1.0 and 0.6, respectively, in all our calculations.

Alignment of the annotated target sequence with every structure from the reference fold library is performed using the global-local algorithm. The global-local algorithm introduces an additional gap penalty for the terminal gaps in addition to the standard opening and extension penalties. Gaps in the target sequence are treated separately from gaps in the reference structure. In particular, only terminal gaps in the reference structure are penalized and the alignment is local in the target sequence. This has the effect of squeezing a target sequence onto the whole length of the reference structure. It has been shown that global-local alignment increases fold recognition odds even without using predicted secondary structure annotations [1].

3 Results

Sequences of seven CASP2 targets that correspond to the folds from the SeqFold database have been retrieved from the CASP2 homepage [18]. These sequences have no clear sequence homology to the structures in our database of protein folds. For the purpose of comparison, we used the PHD secondary structure prediction server [19]. The results of the secondary structure prediction were passed to the SeqFold program using the InsightII interface [20]. The default parameter set was used with no attempt to change or optimize parameters. As can be seen from the cumulative results presented in Table 1, four out of seven targets have been correctly assigned to the corresponding structure. Three targets were relatively easy to identify [21]. The S1 motif of PNPase (T4) exhibits a sequence similarity of almost 30% to 1mjc (when a gapless alignment is used) and the homology model based on the SeqFold alignment exhibits a 4.6 Å RMS deviation from the experimental structure for all alpha carbons (see Fig. 1). Exfoliative toxin A exhibits a 26% identity to the trypsin-like serine protease (1elc). The homology model based on the SeqFold alignment has a sharp decrease in quality following residue 180 of the target (see Fig. 2). The first

Table 1. SeqFold results for CASP2 threading targets

CASP2 target	Protein description	Closest fold (CASP2 jurors)	SeqFold rank 1 hit	Comments
T02	Threonine Deaminase	1psd_A	1gp1_A	False positive – difficult domain only part – second domain is homologous to tryptophan synthase
T04	S1 motif of PNPase	1csp	1mjc	Same family
T14	3-dehydroquinase	1nal_1	1wsy_A	Same fold
T20	Ferrochelatase	8abp	1ipd	False positive, closest true positive ranks as 4
T22	L-fucose isomerase	1tca	1trk_A	False positive – no true positive in first 10 hits – difficult target
T31	Exfoliative toxin	3est	1elc	Same family
T38	Cellulose binding domain CBND1	1byh	1xyn	Same fold

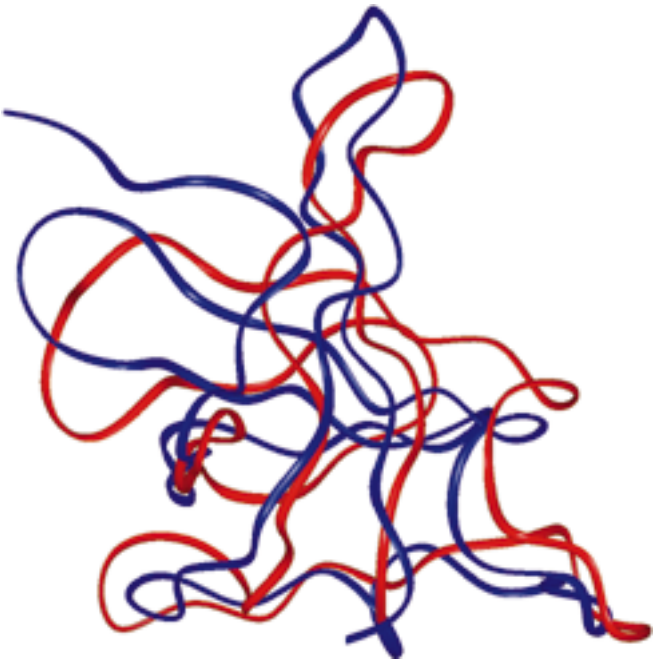

Fig. 1. Comparison of the experimental structure (*blue*) and the prediction based model (*red*) of the S1 motif of polyribonucleotide nucleotidyltransferase

part of the model deviates around 4.8 Å from the experimental structure, whereas the remainder is almost random. 3-Dehydroquinase (target 14) is also an easy target that SeqFold recognizes as a TIM barrel fold (1wsy_A) from the family of tryptophan biosynthesis enzymes. The experimental structure of that target has not yet been published so assessment of the homology model is not possible. The cellulose binding domain (target 38) is a relatively difficult target [21]. The percentage of identical residues with the best hit (1 xyn) is only 20%. That prediction is nevertheless suggestive, since there is an independent high-scoring member of the same family of folds (1scs) present in the hit list. Ferrochelatase (T20) is another difficult target with a strong false positive that prevents identification of the correct corresponding hits of 1sbp and 2gbp – these are periplasmic binding-like folds. The second domains of threonine deaminase and L-fucose isomerase are very difficult targets and sequences with the correct corresponding folds are not observed in the first ten hits. All other CASP2 threading targets correspond to folds that are not present in our fold library – they were also identified as novel folds by CASP2 jurors. We did not attempt to test the null hypothesis for these remaining CASP2 targets.

The mouse leptin sequence has been used to test the dependence of SeqFold results on the secondary structure prediction quality. The leptin sequence does not exhibit clear homology to any protein in our database, i.e, the sequence search with $w^{\mathrm{str}} = 0$ failed to identify any sequence match. Four secondary structure prediction methods have been tested, namely: Chou-Fasman, GOR II, DSC [22], and PHD [23]. In the case of the GOR II and Chou-Fasman methods the SeqFold results are inconclusive and it is not possible to identify a correct fold. The quality of the helix prediction by these methods is very low – 31% and 36%, respectively. In the case of the PHD prediction, granulocyte-macrophage colony stimulating factor GM-CSF (1gmf_A) stands out as a fold (four helical cytokine-like) prediction and the result is substantiated by other high-scoring members of the same superfamily, namely granulocyte colony stimulating factor G-CSF (1bgc) and interleukin-4 (1rcb). Similar results has been obtained by Madej et al. [24]. The leptin structure has recently been solved experimentally and shown to belong to the superfamily of long-chain four-helical cytokines [14]. Hence, although the overall fold has been correctly assigned, the GM-CSF superfamily is slightly different. The result achieved using PHD is probably due to the high quality of helix state predictions – 74% with no false positives. Using DSC made the prediction less apparent. Results were not clear; however, GM-CSF did rank second and a strong false positive – Mengo virus coat protein (2mve) was ruled out due to the very short coverage of the reference structure by the alignment. This result is surprising since the overall helix prediction accuracy is only 24%, the first two helices are completely missed, and only helix C and part of helix D are correctly rendered. A possible explanation for this effect is that the standard Q3 measure of the secondary structure prediction quality is inadequate in the context of sequence-structure compatibility score. Secondary structure information is critical for maintaining a correct path in the regions where sequence similarity is ambiguous, but is not essential in regions where sequence similarity is high.

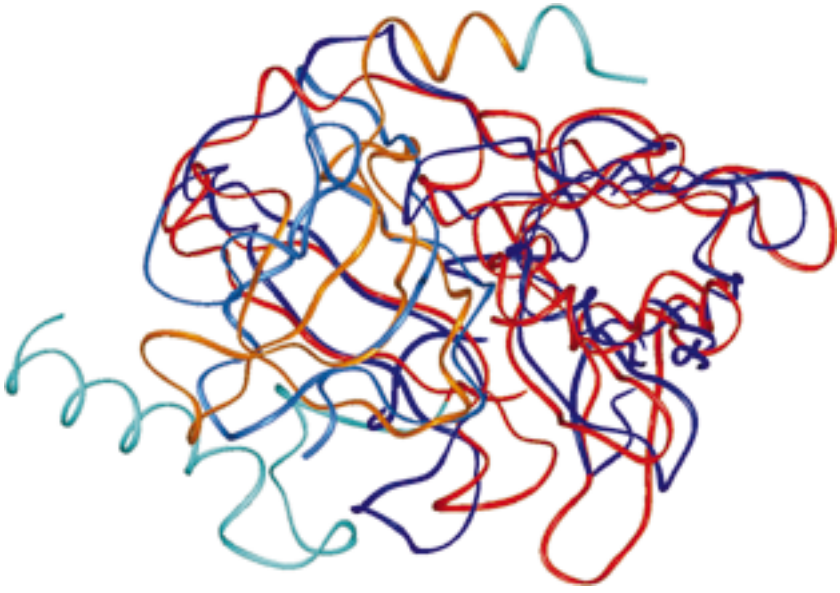

Fig. 2. Comparison of the experimental (*red*) and model (*blue*) structures of exfoliative toxin A. *Orange* and *light blue* correspond to the low-quality model and the *cyan* colored structure indicates the region which lacks a template

Alignments based on PHD and DSC predictions do not differ; thus, homology models of mouse leptin would have been the same. The X-ray structure of leptin has not yet been published; therefore only qualitative assessment of the model is possible. All four helices that form the fold are present in the model, however, one additional helix is inherited from the template and the E helix from the X-ray structure is missing from the model. The lengths of the helices in the template and in the structure are reported in Table 2. Note that model helices A and D have pronounced kinks as expected from the structure. The positions of the invariant residues observed in the leptin family in each helix are also reported in Table 2. There are five clusters of highly conserved sequence segments located in helices A, B, C, D and in the loop between the A and B helices. With the exception of the last segment that is shifted towards the N terminus of the D helix, all others are correctly positioned. Disulfide bond-forming Cys96 and Cys146 are close in the model and the only buried aromatic side chain in the leptin structure, Tyr61, is also buried in the model, even though the Tyr61 environment is not correctly reproduced (see Fig. 3).

Table 2. Obese gene mouse model evaluation statistics. Names of helices are according to Ref. [14]. Helix length column report length of model helix and length of the helix in the X-ray structure. Conserved helical residues report number of residues conserved in the leptin family that are predicted to be in the helix

Helix	Helix length	Missing N terminal residues	Missing C terminal residues	Conserved helical residues
A	16/23	5	2	8/8
B	10/17	3	4	6/6
C	9/24	10	5	8/10
D	14/23	5	4	6/11
E	10	N/A	N/A	N/A

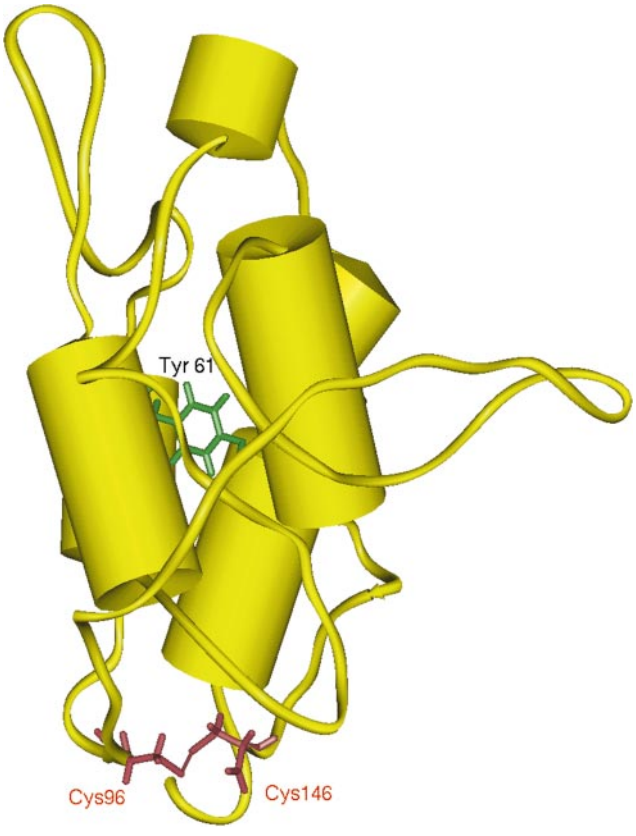

Fig. 3. The mouse leptin model displays a four helix bundle architecture with a correctly positioned Tyr61. The Cys96-Cys146 pair is also correctly reproduced by the model

4 Conclusions

Application of the SeqFold algorithm to leptin and seven CASP2 targets has demonstrated that the use of predicted secondary structure annotations substantially

increases the odds of identifying a homologous sequence for these molecules. However, the performance of the fold recognition algorithm is substantially affected by the quality of the secondary structure prediction algorithm. This can be seen by comparing the performance of SeqFold using the different predictions from Chou-Fasman, GOR II, DSC, and PHD for the leptin search.

When examining the homology models that have been constructed using a PDB template of the correctly recognized fold it has been observed that, whilst the overall quality of the model may not be high, the models generally preserve the main features of the experimental structures. In the case of the leptin model this means that the majority of the functionally important residues were arranged in a similar 3D orientation as that observed in the X-ray crystal structure. This has implications in the area of functional genomics, where recognition of a molecule's fold alone may not indicate function (e.g., many proteins possess a TIM barrel fold but have different functions). However, in many cases, a low-resolution homology model with the functionally important residues in the correct orientation would be enough to give some clues about a protein's function; this could then be verified by targeted mutagenesis experiments.

Acknowledgements. This work was presented at the European conference on Computational Chemistry and The Living World: From Sequence to Function.

References

1. Fischer D, Eisenberg D (1996) Protein Sci 5: 947
2. Russel RB, Copley RR, Barton GJ (1996) J Mol Biol 259: 349
3. Rost B, Schneider R, Sander C (1997) J Mol Biol 270: 471
4. Gilbert W (1991) Nature 349: 99
5. Das S, Yu L, Gaitatzes C, Rogers R, Freeman J, Bienkowska J, Adams RM, Smith TF, Lindelien J (1997) Nature 385: 29
6. Vogt G, Etzold T, Argos P (1995) J Mol Biol 249: 816
7. Anfinsen CB (1973) Science 181: 223
8. Chothia C (1992) Nature 387: 543
9. Olszewski KA (1998) Pol J Chem 72: 1667–1679
10. Jones DT, Thornton J (1996) Curr Op Struct Biol 6: 210
11. Jones DT (1997) Curr Op Struct Biol 7: 377
12. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT (1997) Proteins Struct Funct Genet [Suppl 1]: 2
13. O'Rahilly S (1998) Nature 392: 330
14. Zhang F, Basinski MB, Beals JM, Briggs SL, Churgay LM, Clawson DK, DiMarchi RD, Furman TC, Hale JE, Hsiung HM, Schoner BE, Smith DP, Zhang XY, Wery JP, Schevitz RW (1997) Nature 387: 206
15. Gonnet GH, Cohen MA, Benner SA (1992) Science 256: 1433
16. Henikoff S, Henikoff JG (1992) Proc Natl Acad Sci USA 89: 10912
17. Olszewski KA (unpublished results)
18. CASP2 http://moult.carb.nist.gov/casp2/
19. PHD http://www.embl-heidelberg.de/predictoprotein/
20. InsightII, 97.2 release. Molecular Simulations Inc., http://www.msi.com
21. Marchler-Bauer A, Levitt M, Bryant SH (1997) Proteins Struct Funct Genet [Suppl. 1]: 83
22. King RD, Sternberg MJ (1996) Protein Sci 5: 2298
23. Rost B, Sander C (1994) Protein Struct Funct Genet 19: 55
24. Madej T, Boguski MS, Bryant SH (1995) FEBS Lett 373: 13