Predicting Protein Structure using only Sequence Information

This is a preprint of an article accepted for publication in *Proteins: Structure, Function, and Genetics* copyright 1999.

Kevin Karplus*, Christian Barrett,

Melissa Cline, Mark Diekhans, Leslie Grate, Richard Hughey

17 May 1999

Jack Baskin School of Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064 USA

ABSTRACT

This paper presents results of blind predictions submitted to the CASP3 protein structure prediction experiment. We made predictions using the SAM-T98 method, an iterative hidden Markov model based method for constructing protein family profiles. The method is purely sequence based—using no structural information—and yet was able to predict structures as well as all but five of the structure-based methods in CASP3.

1 Introduction

One method of protein sequence analysis is the identification of *homologous* proteins—proteins that share a common evolutionary history and have similar overall structure and function [5]. Here we report on the use of SAM-T98 [11, 16], a newly developed hidden Markov model (HMM) [13, 9] method for recognition of homologs with low sequence similarity, and how it fared in the fold-recognition section of the CASP3 experiment.

HMMs combine the best aspects of weight matrices and local sequence alignment methods, and can be used to assign probabilities to proteins in database search [6]. Our HMM fold-recognition method differs from protein threading methods [10, 19, 14, 15] in that pairwise interactions are not modeled or used. Instead, we employ Bayesian methods [3, 2, 17] to incorporate prior information in the form of *Dirichlet mixture* densities [20] over position-specific amino acid distributions. The components of the mixture reflect different patterns of sequence conservation and can be combined with data from aligned homologs to form data-dependent estimates of amino-acid probabilities.

In the CASP3 experiments, we used the recently developed SAM-T98 remote homology detection method to compare the CASP3 targets against a database of proteins whose structures are known (Section 2). We discuss how successful this method was in finding similar structures for the targets in Section 3, and discuss the lessons learned in Section 4.

*email:karplus@cse.ucsc.edu Mailing address: Computer Engineering, UCSC, Santa Cruz, CA 95064 USA. Phone: 1-831-459-4250, Fax: 1-831-459-4829. Mail to other authors may be similarly addressed.

A prediction server using the SAM-T98 method discussed here is available on the World-Wide Web[1], as is documentation and licensing information for the SAM hidden Markov model software suite [9].

2 Methods

Since the SAM-T98 method is fully described elsewhere [11], it will only be described briefly here. The method is purely sequence-based and does not employ any structural information. The method iterates through the following steps several times (four for template library, six or seven for the target models), using the initial sequence for input in the first iteration.

1. Build an HMM from a sequence or multiple alignment, using sequence weighting and Dirichlet mixtures. The total sequence weight is chosen to get an appropriate level of generality in the resulting model.
2. Score a nonredundant sequence database with the HMM and retain as a training set those sequences that score better than some threshold value. Scoring is based on log odds, where the likelihood of a sequence having been generated from the HMM is compared to the likelihood of the sequence having been generated from some null model. For the null model we used the reverse of the HMM—the score generated using this null model is the same score one would get from scoring the reversed sequence with the original HMM. This novel null model cancels out artificially strong scores due to length and composition biases and more subtle sources such as conserved rare residues and long helices.
3. Re-estimate the HMM with these sequences, using sequence weighting and Dirichlet mixture priors [20].
4. Re-align the training set using the re-trained HMM. This multiple alignment is used an input to the first step in the next iteration.

During each round of iteration, the score threshold in step 2 is made less stringent in order to capture less similar sequences that are still, we hope, homologs. The final multiple alignment, called the SAM-T98 alignment,

[1] http://www.cse.ucsc.edu/research/compbio/HMM-apps/

is used to construct the HMM used for database search and alignment.

For CASP3, we first built a SAM-T98 HMM for every sequence in a representative set of structure templates from PDB [4] and for every target sequence. To find possible templates for a target sequence, we scored all of PDB with the target HMM, scored the target sequence with every template HMM, and summed the two scores. The structures corresponding to the best summed scores were then investigated manually. For most targets, we submitted only one structure as our prediction—usually the best-scoring one. If we had a high-scoring PDB sequence that was not in our template library, we sometimes augmented the template library with an HMM built from this PDB sequence, in order to be able compare summed scores. We ended up with about 2100 HMMs in our template library.

3 Results and Discussion

Since we predicted on all of the targets for CASP3, we have divided them into three categories to simplify their evaluation. These categories are based on the difficulty of finding the correct structure. Those targets that had very similar sequences of known structure have been placed in the *easy targets* category, while those that had only more distantly related known structures are members of the *moderately difficult targets* category. Those targets that had little or no similarity to known structures are in the *very difficult targets* category. Table 1 shows the results for the first two categories. Except for T0085, the multiheme cytochrome, a submitted cost less than -9 was a successful prediction, though scores as strong as -27.37 would have been incorrect, had we not filtered out those predictions by hand.

3.1 Easy Targets

For the fifteen "easy" targets—T0047, T0048, T0049, T0055, T0057, T0058, T0060, T0062, T0064, T0068, T0069, T0070, T0074, T0076, T0082—our method gave unambiguous results: the correct structural template always had the best (most negative) cost (results not available yet for T0062 and T0069). This cost was always less than -28, for which we expected fewer than 1% false positives [11].

Our submitted alignments for these targets were generally the automatically produced alignments, sometimes subject to minor hand editing. Figure 1 shows our predicted alignment of T0074 to the template structure 2scpA. It shows that our alignment was quite accurate, apart from the first region which is shifted by two residues. The figure is also an example of one of

Easy Targets

		Cost		Was Correct		
Target	PDB	Predict	Top Hit	Predict	Top	CASP3
T0058	1akz	-416.33		+	+	CM
T0047	1mup	-261.04		+	+	CM
T0068	1rmg	-245.31		+	+	CM
T0069	1rtm1	-244.64		?	?	
T0060	1gifA	-226.38		+	+	CM
T0076	1almC	-155.88	-226.34	+	+	CM
T0049	3pte	-217.70		+	+	CM
T0048	1dcpA	-205.74	-209.62	+	+	CM
T0062	2pia+2cnd	-149.19		?	?	
T0055	1esl	-134.70		+	+	CM
T0064	1adr+1ois	-102.26		+	+	CM
T0082	1bol	-82.19		+	+	CM
T0057	1gd10	-47.30		+	+	CM
T0070	2omf	-44.92		+	+	CM
T0074	2scpA	-23.87	-25.75	+	+	FR/CM

Moderately Difficult Targets

		Cost		Was Correct		
Target	PDB	Predict	Top Hit	Predict	Top	CASP3
T0085	3cyr	-62.68		−	−	FR
T0053	1fvkA	-5.72	-27.37	−	−	FR
T0083	1lmb3	-20.18	-20.51	+	+	FR
T0044	1eps	-12.70		+	+	FR
T0059	2dri	-4.89	-10.47	−	−	FR/AB
T0079	1neq+1san	-9.70		+	+	FR
T0071	1hviB	-8.88		−	−	FR/AB
T0075	1oya	-4.88	-8.12	−	−	FR/AB
T0080	1t7pB+1mugA	-4.12	-7.14	−	−	FR/AB
T0043	NF	NF	-7.14	−	+	FR
T0054	NF	NF	-6.97	−	−	FR
T0077	1tif	-5.90	-6.63	−	−	FR/AB
T0067	1rhi2	-5.59	-6.04	−	−	FR/AB
T0081	3chy	0.54	-5.82	±	−	FR
T0061	1amj	-3.31	-5.76		−	FR/AB
T0046	2mcm	-5.75	-6.49	+	+	FR
T0063	1pex	-3.30	-5.65	−	−	FR

Table 1: The targets are ranked by the score of the top hit as found by the SAM-T98 method. The third column gives the sum of costs for the target model and template model (for T0043 and T0054 we predicted "new fold"). If we did not submit our best-scoring template, then its cost is also reported in column three. The symbols +/- refer to whether the prediction was correct/incorrect (see the text for discussion of T0081). "CM", "FR" and "AB" refer to the CASP3 target classifications "Comparative Modeling", "Fold Recognition" and "Ab Initio." Structures have not been released yet for T0062 and T0069, but we are fairly confident of our predictions.

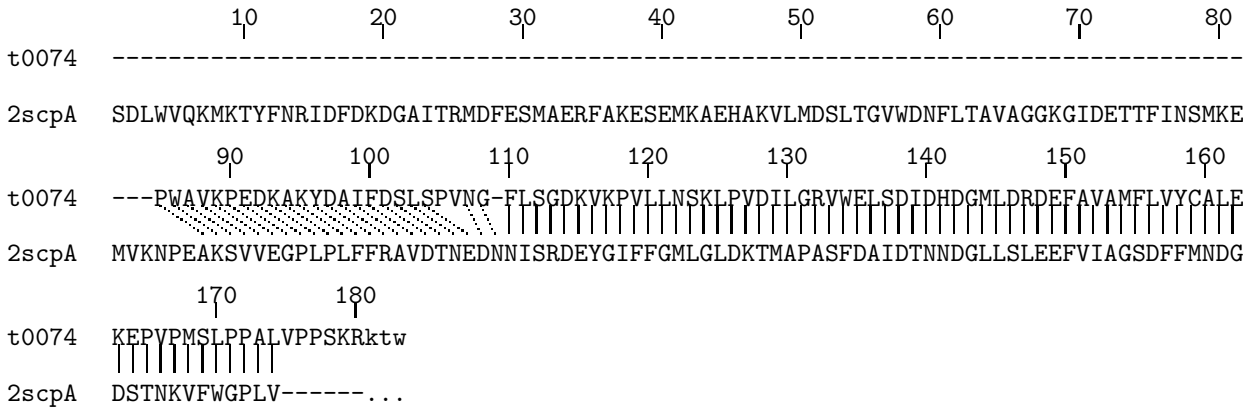

Figure 1: This figure presents the structural alignment of T0074 to 2scpA found by the Yale structural aligner [7] as aligned upper-case letters. Superimposed on the alignment is our predicted alignment, as lines connecting the residues we aligned. Slanted lines between uppercase letters indicates a shift of the predicted alignment relative to the correct structural alignment. This hand-alignment is slightly better than the automatic one we started from, which had the same correctly aligned residues, but the N-terminus of T0074 was more misaligned.

the few cases where hand-editing improved on the automatic alignment. In this case, the automatic alignment had shifted the 21 residues PWAVKPEDKAYKYDAIFDSLS of T0074 7 residues toward the N-terminal region of 2scpA, while the hand alignment shifted them 4 residues toward the C-terminus.

Overall, our alignments for the easy targets were usually among the best alignments submitted to CASP3, even though we used no structure information in generating them.

Our 3D prediction for target T0076 was quite poor. We aligned T0076 to 1almC (a theoretical model) because our top hit, 2mysC, had a probable mistracing. We thought that 1almC corrected this mistracing, but since it did not, our 3D prediction was poor even though the sequence alignment was accurate. We would have done better to use the second-highest-scoring template (1wdcC), which has an accurate 3D structure.

3.2 Moderately Difficult Targets

There are sixteen "moderately difficult" targets in this category: T0043, T0044, T0046, T0053, T0054, T0059, T0063, T0067, T0071, T0075, T0077, T0079, T0080, T0081, T0083, T0085. We correctly predicted similar structures for five targets: T0044, T0046, T0079, T0081, T0083, all but T0081 of which used the top hit. Most of the other structures in this category had costs too close to zero to yield much confidence in our predictions.

Because of the low similarity between the targets and templates, even the "correct" predictions had alignments that were accurate only for portions of the target sequence. We used local alignment to find the folds, but global alignment to provide the submitted alignment. The global alignments generally aligned more residue pairs than the correct structural alignments, but if we has submitted the local alignments, we would have missed many of the residue pairs that were correctly predicted. Because RMS deviation is very sensitive to overprediction, our RMS scores for the entire alignment look poor, even though we often have a well-predicted core alignment. Determining which parts of an alignment are worth predicting and which should be removed remains a difficult problem for us.

Our T0085 prediction (2cthA) was an incorrect multiheme cytochrome, despite the high score. Matching three heme-binding sites provided a strong similarity signal, even though the overall fold was different. The correct multi-heme cytochrome was in our top 6 hits (out of about 2000 templates).

For T0053, we were misled by our post-scoring sequence analysis. We considered the correct template 1ak1 for T0053 (our 10th highest-scoring template), because it scored well in one of our template libraries and was also a chelatase. We correctly rejected our top hit (1djxB), because it did not cluster well the known metal-binding residues in T0053. We chose 1fvkA (our 8th highest scoring template), because it clustered the residues well. Unfortunately, we did not analyze the clustering on 1ak1. The target HMM (which gave the erroneous high score to 1djxB) was poor because there were only three short matches to the target found in the non-redundant protein database by the SAM-T98 method (other than the target

itself), so the HMM had to generalize from very little data. In such cases, it may be advantageous to put more weight on the template HMM scores, but we did not attempt this.

For T0071, we were again misled by our post-scoring sequence analysis. We looked at, but rejected, some correct templates (1euu and 1dlhA) for the first domain because of low scores and unconvincing alignments. We wanted to find an SH3 domain, because the C-terminus of EPS15 binds to T0071 and is known to bind to an SH3 domain [18, 1]. This hint from the literature was used to decide between a small number of folds, all of which had fairly weak scores with our method.

We were surprised that 2mcm turned out to be a correct prediction for T0046, because the similarity to immunoglobulins was weak and most immunoglobulins are quite similar to each other. Our alignment turned out to be terrible, as can be seen in Figure 2

The known active site residues for T0081 clustered well when the target sequence was aligned to 3chy. This was our rationale for choosing 3chy as our prediction. It turns out that 3chy has a similar alpha-beta-alpha structure, but threaded in a different order than T0081. If we do a circular permutation of the chain, we can get a much better superposition of the structures—unfortunately, our method did not predict this circular permutation, but an incorrect alignment. Even constructing an HMM for the chimeric sequence 3chy followed by 3chy does not allow our methods to find the permuted alignment. We would have done better to predict 1rvv1, which scored better than 3chy, and had the correct threading order. We had rejected 1rvv1, because our alignments for it did not cluster the aspartic acid residues, which we had expected.

During the early part of CASP3, we predicted "new fold" for targets that produced only weak scores to template structures. For this reason we predicted that T0043 would be a new fold, even though our top hit turned out to be the correct fold.

There were a number of targets for which the correct structural template was in our list of top 10–20 hits, but we were not able to pick it out. These targets, with the rank (out of approximately 2000) of the correct hit in parentheses, are T0043(1), T0053(16), T0054(9), T0059(16), T0063(10), T0067(16), T0071(6), T0085(6). We hope that small improvements to the method, as well as the increase in the number of homologs in the databases will allow the method to discriminate better in future.

The low similarity between targets and structures in this category reduced alignment quality considerably compared to the alignments for the easier targets in Section 3.1. We almost always hand edited our automatically generated alignments for these targets. In general, though, hand alignment did not provide much improvement, and we would have done about as well with considerably less effort had we submitted our automatic alignments. For example, we show one of the better hand alignments in Figure 3, but the automatic alignment it was based on did not include the incorrect alignment at the C terminus.

3.3 Very Difficult Targets

Almost all of the remaining "very difficult" targets can be characterized as targets that represented new folds or that bore similarity only to fragments of solved protein structures. For all of the targets in this category, our method indicated with high likelihood that there was no similar structure.

3.4 Secondary Structure Prediction using SAM-T98

We also used the SAM-T98 multiple alignments for the target sequences as inputs to a neural net for secondary structure prediction. This turned out to be one of the best two secondary structure predictors at CASP3, although we did not use the secondary structure predictions in our fold predictions. When we had a correct fold prediction, deducing the secondary structure from the predicted alignment provided more accurate secondary structure prediction than the neural net, but the neural net was more accurate when we had an incorrect fold prediction. We plan to combine this secondary structure predictor with two others we are building, and put them all up on the World Wide Web in the next six months.

4 Conclusion

We have discussed the HMM-based SAM-T98 method for remote homology detection and how it was applied to protein structure prediction in the CASP3 experiment.

Many of the fold-recognition methods do considerably better on multi-domain proteins when given the domain boundaries, but when we tested our method after CASP3 on the true domains, we gained no benefit from having that extra information [12]. We suspect that our use of local alignment and sum-of-all-paths scoring makes our method rather insensitive to the inclusion of extra domains, so there is little gain from excising them.

Perhaps the biggest lesson learned is that we do not know enough about proteins to adjust SAM-T98 alignments manually. We would have been better off trusting the programs even when they seemed wrong. Protein experts with more knowledge of the proteins would most likely be able to adjust the alignments better than we

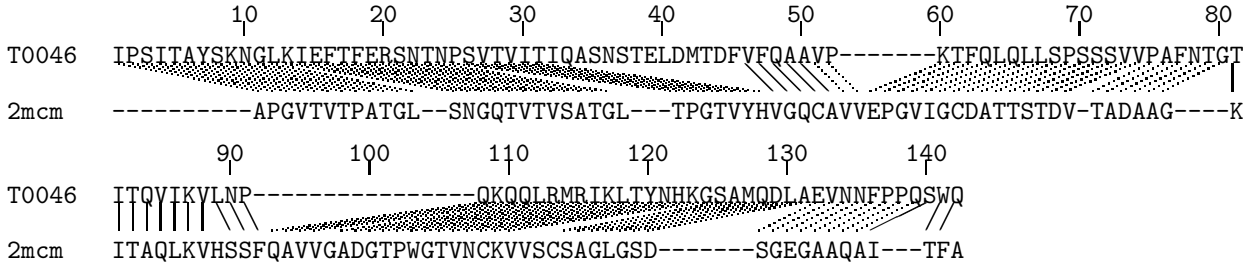

Figure 2: This figure shows the correct structural superposition of T0046 and 2mcm as found by the Yale structural aligner [7]. Lines connect the residues we predicted to be aligned. This alignment is one of our worst alignments for a correct fold. The problem here was in our hand-editing, as the automatic alignment we started from was much better.

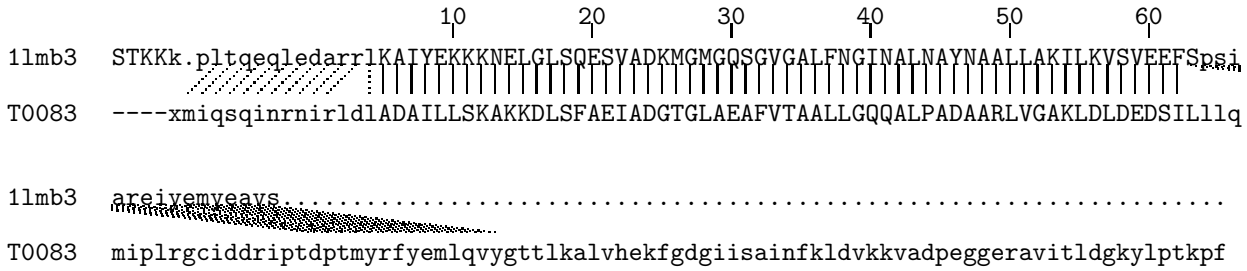

Figure 3: This figure presents the correct structural superposition of T0083 and 1lmb3 as found by the VAST structural aligner [8]. (The Yale structural aligner provides a somewhat longer alignment, but which agrees on the common part.) The lines connect the residues we predicted to be aligned. This alignment is among the best alignments we submitted for the moderately difficult targets, but automatic alignment was slightly better than this submission.

could. We do still get value from human interaction, but mainly in including functional information or information about known binding sites, rather than in adjusting alignments.

The costs provided by SAM-T98 are a strong, but not perfect, indicator of the correctness of the predictions. Using a calibration of our method from known structures [11], many of the targets had such weak similarities that we had little confidence in those predictions.

We believe SAM-T98 has taken sequence-only methods about as far as they will go. For many of the "moderately difficult" targets we did not select the correct structure even though it was in the top 20 hits, so even a small amount of additional information should improve the method significantly. We are currently investigating a few ways to include structure information: building our template library HMMs from structural multiple alignments (rather than single sequences), using information from the structure of the template to trim alignments, and using sequence-structure compatibility measures to evaluate alignments.

Acknowledgments

This work was supported in part by NSF grants DBI-9408579 and DBI-9808007; DOE grant DE-FG0395-ER62112; and a GAANN graduate fellowship. We are grateful to David Haussler for starting the hidden Markov model and Dirichlet mixture work at UCSC, as these approaches were instrumental to our success.

References

[1] A. Benmerah, B. Begue, A. Dautry-Varsat, and N. Cerf-Bensussan. The ear of alpha-adaptin interacts with the COOH-terminal domain of the Eps 15 protein. *J Biol Chem*, 271(20):12111–6, May 17 1996.

[2] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[3] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley and Sons, first edition, 1994.

[4] F. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi,

and M. Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *JMB*, 112:535–542, 1977.

[5] R. F. Doolittle. *Of URFs and ORFs: A primer on how to analyze derived amino acid sequences*. University Science Books, Mill Valley, California, 1986.

[6] S. Eddy. Hidden Markov models. *Curr. Opin. Struct. Biol.*, 6(3):361–365, 1996.

[7] M. Gerstein and M. Levitt. Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.*, 7:445–456, 1998.

[8] J. Gilbrat, T. Madej, and S. Bryant. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, 6:377–85, 1996.

[9] R. Hughey and A. Krogh. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS*, 12(2):95–107, 1996. Information on obtaining SAM is available at `http://www.cse.ucsc.edu/research/compbio/sam.html`.

[10] D. Jones and J. Thornton. Protein fold recognition. *J. Comput. Aided Mol. Des.*, 7:439–456, 1993.

[11] K. Karplus, C. Barrett, and R. Hughey. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.

[12] L. A. Kelley, R. M. MacCallum, M. Sternberg, K. Karplus, D. Fischer, A. Elofsson, A. Godzik, L. Rychlewski, K. Pawłowski(A), D. Jones, K. Bryson, and B. Rost. CAFASP-1: Critical assessment of fully automated structure prediction methods. *Proteins: Structure, Function, and Genetics*, to appear, 1999.

[13] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *JMB*, 235:1501–1531, Feb. 1994.

[14] R. Lathrop, L. Ljubomir, R. Nambudripad, J. White, L. L. Conte, B. Bryant, and T. Smith. Threading through the levinthal paradox. *Nature*, To appear.

[15] C. Lemer, M. Rooman, and S. Wodak. Protein structure prediction by threading methods: evaluation of current techniques. *Proteins: Structure, Function, and Genetics*, 23:337–355, 1995.

[16] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods. *JMB*, 284(4):1201–1210, 1998. Paper available at `http://www.mrc-lmb.cam.ac.uk/genomes/jong/assess_paper/assess_paperNov.html`.

[17] T. J. Santner and D. E. Duffy. *The Statistical Analysis of Discrete Data*. Springer Verlag, New York, 1989.

[18] C. Schumacher, B. Knudsen, T. Ohuchi, P. Di Fiore, R. Glassman, and H. Hanafusa. The SH3 domain of Crk binds specifically to a conserved proline-rich motif in Eps15 and Eps15R. *J Biol Chem*, 270(25):15341–7, Jun 23 1995.

[19] M. Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, 5:229–235, 1995.

[20] K. Sjölander, K. Karplus, M. P. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler. Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *CABIOS*, 12(4):327–345, Aug. 1996.