What is the value added by human intervention in protein structure prediction?

Kevin Karplus*, Rachel Karchin, Christian Barrett, Spencer Tu,
Melissa Cline, Mark Diekhans, Leslie Grate, Jonathan Casper, Richard Hughey

October 10, 2001

This is a preprint of an article accepted 21 August 2001 for publication in the CASP4 special issue of *Proteins: Structure, Function, and Genetics* (46) copyright 2002.

Abstract

This paper presents results of blind predictions submitted to the CASP4 protein structure prediction experiment. We made two sets of predictions: one using the fully automated SAM-T99 server, and one using the improved SAM-T2K method with human intervention. Both methods use iterative hidden Markov model-based methods for constructing protein family profiles, using only sequence information. While the SAM-T99 method is purely sequence based, the SAM-T2K method uses the predicted secondary structure of the target sequence and the known secondary structure of the templates to improve fold-recognition and alignment.

In this paper we try to determine what aspects of the SAM-T2K method were responsible for its significantly better performance in the CASP4 experiment, in the hopes of producing a better automatic prediction server. The use of secondary structure prediction seems to be the most valuable single improvement, though the combined total of various human interventions is probably at least as important.

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

1 Introduction

For CASP4, the University of California, Santa Cruz bioinformatics group entered two related prediction methods: a fully automated server (SAM-T99) and a more experimental method (SAM-T2K) that included considerable human intervention. We were interested in seeing how much benefit we gained from the various improvements to the method and whether the human intervention offered any advantages. We were particularly concerned about the quality of human intervention, because it was done by people trained as computer scientists, not protein chemists.

In the CASP fold-recognition assessment [1, Sippl-eval], the SAM-T99 server did only adequately (45th overall and 9th of the servers), but the SAM-T2K method with human intervention did quite well (4th overall). In this paper we attempt to analyze what improvements and interventions caused this improved performance, in the hope of improving the next generation of the automatic server.

2 Methods

Both SAM-T99 and SAM-T2K are upgraded versions of the SAM-T98 method that we used in CASP3[2, 3]. Given a single target sequence, these algorithms iteratively build a HMM and multiple alignment of the sequence and its homologs. Rather than search the entire NR protein database [4], subsets to search are extracted with

WU-BLAST [5, 6].

Each iteration

- takes a multiple alignment, a sequence search set, and a score threshold;
- builds a HMM from the multiple alignment (using sequence weighting, Dirichlet mixture and transition regularizers);
- does HMM scoring of the search set;
- retrains the HMM using only sequences in the search set which score above the threshold; and
- aligns these sequences to the new HMM.

The new alignment is input to the next iteration. (The initial multiple alignment is the trivially aligned target sequence.) Four rounds of database searching and model building produce the final HMM used in fold recognition.

To predict the fold of a target sequence, we use bi-directional scoring. An HMM is built from the target sequence and scored against all sequences in a structural database. The target is also scored against a pre-built HMM structure library. The template-to-target and target-to-template scores are averaged, and fold prediction is based on the top scoring target-template pair.

SAM-T99 is a re-parameterized version of SAM-T98 [2]. SAM-T98 uses a tightly thresholded search set of close homologs (E-VALUE $\leq$ 0.00003) on iteration 1, and a loosely thresholded search set of potential homologs (E-VALUE $\leq$ 500) on subsequent iterations. In SAM-T99, more close homologs (E-VALUE $\leq$ 0.0005) and fewer potential homologs (E-VALUE $\leq$ 300) are allowed in the search sets. The HMM scoring thresholds are relaxed after each iteration in both methods, but SAM-T99 expresses the thresholds differently, using E-value rather than raw score. The threshold used on the first iteration is looser than in SAM-T98 and the threshold on the last iteration is tighter. SAM-T99 also uses different Dirichlet mixture priors and transition priors for HMM building.

In SAM-T2K, a new sequence search set is produced for each iteration with a series of progressively relaxed WU-BLAST thresholds (0.01, 1.0, 10, and 400), rather than using just two search sets. The HMM score thresholds used to allow sequences into the multiple alignments have been changed from SAM-T99's E-values of (0.00001, 0.0001, 0.001, and 0.01) to (0.00001, 0.0002, 0.001, and 0.005). The fourth iteration's HMM score threshold was tightened to reduce number of false positives in the final alignment. The final alignment is computed with *posterior decoding*, rather than *Viterbi* dynamic programming [7, 8]. SAM-T2K also returned to using recode3.20comp as the Dirichlet mixture prior.

Finally, we have implemented several changes in sequence weighting. Weights are now varied for each iteration of model building to get more generalization on later iterations. To build the final HMM, both SAM-T99 and SAM-T2K use the w0.5 script, which thins the alignment to 80% residue identity, uses an entropy weighting method with the target savings set to 0.5 bits/column, and uses the recode3.20comp Dirichlet mixture prior.

Fold-recognition tests indicate that HMMs built from SAM-T2K multiple alignments are better than ones built from SAM-T99 multiple alignments (unpublished), but not by enough to account for the difference in in performance seen in CASP4.

Two-track HMMs

In SAM-T2K, fold recognition performance has improved slightly due to better multiple alignments, which results from changes in threshold parameters, regularizers, and some internal details of our basic algorithm. The most significant performance improvement between SAM-T99 and SAM-T2K involves the use of two-track HMMs. Although SAM-T99 makes a secondary structure prediction for the target sequence, this information is not used in SAM-T99's fold recognition and alignment.

SAM-T2K uses a neural net to produce a vector of three predicted probabilities (helix, sheet, and coil) for each residue in the target. These are included as emission probabilities in the match states of the target's HMM, so that each match state contains a distribution of amino acid probabilities and a distribution of three-state secondary structure probabilities. All templates in our structure library are then scored as sequence pairs (amino acid sequence and secondary structure sequence) with the two-track HMM. Our two-track HMM software was not ready until mid-summer, so we were able to apply this new method only to some of the targets.

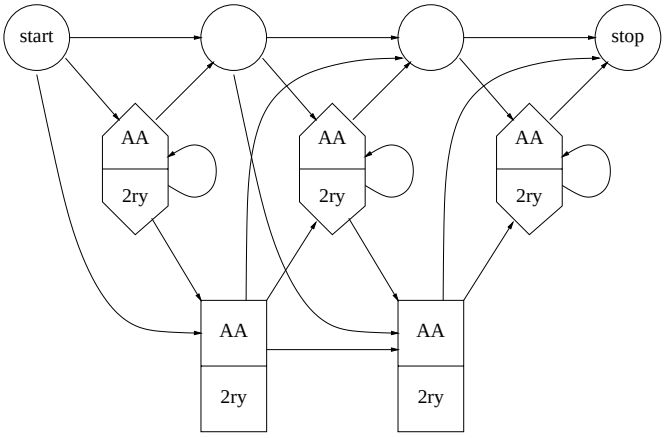

Figure 1: This picture shows graphically how the two-track HMM is organized. The only change from the profile HMMs used previously with SAM is the addition of predicted emission probabilities for secondary structure to the match states—the insertion states get background probabilities for secondary structure and the transition probabilities are identical to the amino-acid-only HMMs. A real HMM has as many match states as alignment columns in the multiple alignment (for SAM-T2K, the number of residues in the target sequence). The "AA" and "2ry" labels in the boxes refer to emission-probability tables for amino acids and secondary structure labels, respectively.

domain	SAM-T99	SAM-T2K	difference
121-2		4	4
100	0	3.5	3.5
95-2	0	3	3
110	0	3	3
118	0	1.5	1.5
101	2	3	1
116-4	1.5	2.5	1
127-2	1.5	2.5	1
87-1	0	1	1
107	0	1	1
109	0	1	1
96-1	3.5	4	0.5
87-2		0.5	0.5
95-1	0	0.5	0.5
114	0	0.5	0.5
127-1	3	3	0

Table 1: CASP4 fold-recognition targets for which either SAM-T99 or SAM-T2K received a non-zero evaluation by Manfred Sippl. The second and third columns report Sippl's evaluation, with blanks indicating that the prediction did not include the specified domain. The last column gives the difference in evaluation between the methods. Note that SAM-T2K is consistently better than SAM-T99. The domains are sorted by the difference in evaluation score.

3 Results and Discussion

One of the first analyses we did was to compare the official assessment of our results to those of the other groups ranked near the top. We wanted to know whether our methods were stronger at alignment or at fold-recognition (using the other top groups as a comparison base to factor out the difficulty of different targets).

For most of the fold-recognition predictions, our SAM-T2K alignments were similar in quality to Baker's predictions with Rosetta and Rychlewski's with Bioinfo, though there were particular targets on which one or another of the methods did significantly better. Where Rosetta did better overall was in making somewhat reasonable predictions for new folds, where our methods, predictably, had no success. Both Murzin and Sternberg had higher quality alignments than SAM-T2K, but did not have quite as many successful fold predictions.

We submitted two sets of predictions to CASP4: a fully automated set using the SAM-T99 server, and a set involving human intervention (SAM-T2K). The SAM-T2K predictions were good [1, Sippl-eval]—much better than the SAM-T99 predictions. Table 1 lists all the fold-recognition domains that got a non-zero score for either technique in Sippl's evaluations, sorted by the difference in score.

From the table, we can clearly see that SAM-T2K consistently outperforms SAM-T99, and the rest of this section will try to analyze, for each domain, what we did differently in SAM-T2K to get the superior performance.

Domain 121-2

SAM-T2K's better performance on domain 2 of ketose reductase/sorbitol dehydrogenase (T0121) is not surprising. The protocol we used with the automatic server simply reported the best hit, which was for domain 1. No attempt was made to consider the residues not included in that prediction. For SAM-T2K, we created a new target consisting of just the C-terminus of T0121 (past residue 240). With this target, both the standard HMM and the 2-track HMM found 1b9mA as the best template. Furthermore the conserved residues in the 2-track local alignment occurred in regular stripes across the beta strands of the barrels, increasing our confidence in the quality of the alignment.

Domain 100

Target T0100 (pectin methylesterase) was a beta helix (SCOP superfamily 2.75.1 [9]), which SAM-T99 had as the eighth-ranked protein (third-ranked fold). This eighth ranked protein for SAM-T99 was pectate lyase, which we decided to examine more closely, since it binds a similar substrate as the target protein. We also noted that the CAFASP servers were generally in agreement that the domain was 2.75.1 [10, CAFASP2], and this agreed with our secondary structure prediction.

We spent a lot of effort attempting to align T0100 with various beta helices, but were only moderately successful.

At the time we had to submit our predictions for T0100, the two-track HMM scoring was not working yet, but later scoring showed that the right superfamily would have been at the top for two-track HMMs. Since this is the only target discussed in this paper for which the two-track HMMs were unavailable by the target expiration, and since we could get the right fold either by matching the function or by using 2-track HMMs, we have given credit for both in Table 2.

Domains 95-1 and 95-2

SAM-T99 was seriously handicapped for the alpha(E)-catenin fragment (T0095), since the template used by SAM-T2K was not available in PDB by the deadline for the CAFASP submission. The difference in results for this target does not reflect a difference in the methods, but a difference in the submission deadlines. SAM-T2K got a good alignment for the second domain, but not for the first domain.

Domain 110

The ribosome binding factor A (T0110) got no hits with template models in our T99 library and only one weak hit with the SAM-T99 target model, which turned out to be incorrect. We rejected this hit manually, since it was a zinc-binding protein, but none of the cysteines that coordinated the zinc were present in the target. The CAFASP servers did not offer any obvious consensus targets. We generated a few more candidates by considering the top few hits from the 2-track target HMM. We also noted that the protein functioned as a cold-shock protein, so we included a number of templates from SCOP superfamily 2.38.4 (OB-fold, nucleic-acid-binding proteins) [9].

We ended up with 31 templates to consider. We ranked these according according to the Viterbi scores of alignment by various HMMs (target, template, template HMM from FSSP alignment) with various alignment options, then examined the alignments manually. Several of the top-scoring ones were rejected because they aligned to a beta sheet but omitted an interior strand of the sheet. We also rejected a few possibilities for very poor secondary structure matches. We did not examine all the alignments produced, not bothering with the ones that scored very poorly.

We ended up with two possibilities: pieces of 1egaA and 1lehA, which were the 4th and 1st hit with the 2-track HMM. We rejected the second hit (1mml) because the predicted strands were in different domains, rather than being in the same sheet, and the third hit (1cf9A) because the predicted structure was not compact.

The alignments with 1egaA and 1lehA overlapped: 1egaA had secondary structure elements helix-strand-strand-helix-strand, with the strands part of an untwisted beta sheet, and 1lehA had elements strand-strand-helix-strand-helix with a similar untwisted beta sheet. The strand-strand-helix-strand sections were very similar in both structures. Since the pieces of 1lehA and 1egaA did not conflict, we tried combining the two predictions using Undertaker, our experimental fragment-packing program to make a helix-strand-strand-helix-strand-helix prediction. It turned out that this combination did not improve

our prediction, as the final helix that 1lehA added is not resolved in the solved structure, and the prediction based on 1egaA alone was closer to the solved structure than our combined prediction was.

The overall prediction was good, though the first helix was misaligned by two turns. For this target, the hand-selection of compact beta structures was crucial, but the 2-track HMMs greatly reduced the number of alignments we had to consider (from 31 down to 4).

Domain 118

Our prediction for Endodeoxyribonuclease I (T0118) was not an exceptionally good prediction, but the target was difficult (only four groups did better than SAM-T2K).

Since 5 catalytically essential residues had been identified for Endonuclease I: E20, E35, D55, E65, and D74, and the last 9 residues were known to be essential for DNA binding [11], we tried improving our HMM by using a two-sequence alignment as our initial seed. One sequence was the target sequence, the other had Xs is all positions except the 14 key residues.

The other sequences in the multiple alignment generated by SAM-T2K using this seed (the *expanded* alignment) were the same as in the SAM-T2K alignment from just the target (the *original* alignment), but the extra sequence with the active site residues did increase (slightly) the conservation expected in those columns. The secondary structure prediction with the expanded alignment provided more reasonable helix and strand lengths than the one from the original alignment (and turned out to be somewhat more accurate).

The best alignments with the 2-track HMM built from the expanded alignment were to 1opr (SCOP superfamily 3.56.1) and 1avqA (3.47.1). The 1opr alignment was missing an interior strand of a beta sheet, but the 1avqA alignment could be plausibly extended to a full-length alignment and, with a little editing, could cluster 4 of the 5 active site residues—this is the alignment we submitted. The best of the CASP4 predictions used 1pviA as a template, which is in the same SCOP superfamily as 1avqA, but a different family—we had 2pviA (the same sequence as 1pviA) as our fourth best hit, but did not have time to examine it manually.

The addition of the key residues to the multiple alignment moved 1avqA from superfamily 3.47.1 from the 5th position with the 2-track HMMs to the second position. The amino-acid-only HMM also benefitted from the addition of the key residues, moving sequence 2pviA (the correct superfamily) from very far down the list to second position. Increasing the weight on the key residues decreases the E-value for 1avqA, but even for fairly high weights on the key residues, the incorrect match to 1opr scores slightly better.

Domain 101

The pectate lyase (T0101) was easily recognized as a beta-helix, and both SAM-T99 and SAM-T2K used 1dbgA as the template (as did most of the servers in CAFASP). The better evaluation for SAM-T2K results entirely from better alignment. The alignment submitted is basically the local alignment using the 2-track target HMM, but some hand-editing was done to try to improve the conservation pattern along the highly conserved turns in the beta helix. The hand editing moved few residues and those only by a few positions—the unedited alignment would have scored almost as well.

Domain 116-4

For the MutS DNA mismatch repair protein (T0116), we had a strong match for SCOP domain 3.31.1 using template models (particularly 2reb). We did not find any strong matches with the amino-acid-only target model, but the 2-track target HMM had hits to domain 3.31.1 in the top four, making it quite promising. Other than the 2reb match (a strand-helix alternation making a 5- or 6-strand sheet), none of the top-scoring alignments looked particularly good—they tended to have large gaps both in the sequence and in three dimensions.

With 811 residues in the target, it seemed clear that this was a multi-domain protein. Our match to 2reb started at about residue 576, and we had a weak match to 1bkdS up to about residue 470, so we tried a three-way overlapping split: 1-500, 400-600, 500-765. The domain split made by the assessors is 1-128, 129-249, 250-542, 543-765.

For the 500-765 subtarget, we had strong hits for the 2reb template model, and moderate strength ones with the target model. Other sequences from the same SCOP domain, 1b0uA and 1cr1A, also scored well either with

the template model or the 2-track target model. We hand-edited the 1b0uA global 2-track alignment to shrink the gaps where there were substantial deletions. This alignment placed the domain boundary in the right place (within 1 or 2 residues), and was apparently a good alignment (within the top 4 predictions for this domain).

Although we tried doing further splitting of sequence (1-300, 201-500), we were not successful in identifying the other domains of T0116. We submitted an alignment to 1tgoA, which might have been ok for the second domain, but our alignment was about 80 residues off, rendering it useless.

The 2-track alignment was important for our getting a good alignment for domain 116-4 and domain splitting somewhat less important, as we managed to recognize the right fold (with a poorer alignment) without domain splitting.

For some reason, we overlooked the submitters' comment that the C-terminus was homologous to ABC transporter ATPase (1b0uA)—using that hint could have saved us some time in choosing the right template from the superfamily.

Domain 127-2

For the magnesium chelatase (T0127), we found many strong hits with the target models for SCOP superfamily 3.31.1 (as did essentially all the automatic servers in the CAFASP experiment). The 2-track HMM also scored sequences from this superfamily well. One template, 1do0A, consistently scored the best with target and template models and the two-track target model. Although some hand editing was done (mainly moving a few residues from the second domain to the first domain), the alignment was essentially that provided by the 2-track HMM alignment. Using an amino-acid-only HMM did not give as good an alignment of domain 127-2 as the 2-track HMM.

It turned out that our submitted alignment was the best alignment predicted for this domain. Note that no attempt was made to split the target into two domains and predict them separately, though hand inspection of the predicted structure showed a clear domain boundary.

Our alignment for the first domain was good (comparable to several other good predictions), but not as good as that submitted by SBFold [12].

Domains 87-1 and 87-2

Although the SAM-T2K predictions were mediocre for PPase (T0087) (seven groups did better on the first domain, and at least 30 groups did better on the second domain), they were better than the SAM-T99 predictions. The main advantage came from splitting the target into two domains (1-180 and 181-310)—a domain split that was only a little different from that done after the structure was known (1-194, 195-310). The domain split was done based on the match to the Pfam DHH domain [13].

Scores with the two-track HMMs were weak and were not used for selecting templates. The alignment of Domain 1 to 8abp was obtained by hand-editing a 2-track global alignment. The alignment of Domain 2 to 1be1 was obtained by extensive hand editing of the global target model alignment (not using a 2-track HMM).

Domain 107

The prediction for family 9 carbohydrate-binding module (T0107) was poor for SAM-T2K—what slight advantage it had over SAM-T99 is probably attributable to the 2-track HMM.

Domain 109

The prediction for oligoribonuclease (T0109) was poor for SAM-T2K. We selected the template based on function and on the predictions of other servers in the CAFASP experiment. The functional matching was just for ribonucleases and DNA polymerases (the most common functions for the Swissprot matches in the SAM-T2K multiple alignment). We got the right fold, but the alignment was poor, perhaps because of inaccuracies in the secondary structure prediction.

Domain 96-1

The first domain of FadR (T0096) was the easiest of the targets in the fold-recognition category, with most groups that predicted for it getting at least the right fold. We got hits for 1lea and 2cgpA with both target and template models. Both of these are winged-helix DNA-binding domains (SCOP 1.4.4). Since the target is a DNA-binding protein, this seemed like a good hit functionally, though

we did not really need this functional information to make a confident prediction.

Because we had no good hits to the second domain and the submitters gave us the information that the acyl-coenzyme-A-binding domain had no structurally similar sequences in PDB, we did not examine that domain further, but looked only at the first domain.

With the two-track target HMM, the top-scoring six sequences were all from SCOP superfamily 1.4.4, giving us more confidence than we had with the amino-acid-only HMMs. We submitted the third-best scoring alignment (to 1qbjA), because it had fewer insertions and deletions than the higher-scoring ones, but we edited it to move the insertion away from the DNA, putting the insertion in the same place as the automatic alignment for 1bi0. We chose 1qbjA as the template, because the insertion is smaller in 1qbjA than in 1bi0. Murzin produced a better alignment to 2dtr, which is the same sequence as 1bi0 for this domain. It is not clear how we could have chosen 1bi0 over 1qbjA with the information we had.

Domain 114

Neither the SAM-T99 nor the SAM-T2K methods found any homologs for T0114 in building the multiple alignments, so the target models and the secondary structure predictions were made from a single sequence.

Although secondary structure prediction from a single sequence is inaccurate, the template we selected (1hoe) was chosen because it was a 6-strand structure, not because it scored particularly well.

We correctly predicted that the structure would be a beta sandwich, and we got a few of our predicted strands to align to real strands, but the topology of our predicted sandwich was not correct, so the overall prediction was incorrect even for SAM-T2K. The difference in evaluation between the two probably amounted to bonus points for getting a beta sandwich of roughly the right size with SAM-T2K.

Summary of differences between SAM-T99 and SAM-T2K

In Table 2 we have attempted to tabulate what made the SAM-T2K alignments better than the SAM-T99 ones. The single most important difference is the use of two-track HMMs in SAM-T2K, though splitting into domains and the use of functional information were also helpful. Overall, the the 2-track HMMs and the combined manual interventions were about equally valuable.

4 Conclusion

The hand intervention was successful in improving the protein fold recognition, but was very labor intensive. Many of the tasks done in the hand intervention are automatable, and we intend to put as many as we can into the next automatic server (SAM-T01).

Many of the improvements in SAM-T2K are automatable. The most valuable of them, the 2-track HMMs, are already programmed—we just have to tune parameters and provide some calibration of the E-values. We are planning to do some automatic checking of function based on keyword matches, perhaps using the methods of SAWTED [14]. Our template library is now updated frequently (about every two weeks—more frequently during CASP season), so additions to PDB are quickly incorporated.

Domain splitting remains a more challenging problem. We do not expect to do fully automatic domain prediction, but we may automate resubmission of the remainder of a target when one domain is strongly matched.

We frequently observe several templates for the same fold or superfamily scoring well in our predictions, but we have not used this clustering of results in automatic prediction. We are looking into using the product-of-p-values method [15] for combining information from multiple templates.

We do not expect to automate the screening that we do by looking at the predicted three-dimensional structure, though the removal of non-compact predictions or predictions that skip interior strands of a beta sheet would be useful. We have also not committed to automating the use of key-residue information, since the information is rarely available in a machine-readable form.

Acknowledgments

This work was supported in part by NSF grants DBI-9808007 and EIA-9905322, DOE grant DE-FG0395-

domain	2-track HMM	domains	function	library	interior strand	active site	other
121-2		x					
100	x		x				
95				x			
110	x				x		x
118					x	x	
101	x						
116-4	x	x					
127-2	x						
87	x	x					
107	x						
109			x				
96-1	x						
114							x
total	8	3	2	1	2	1	2

Table 2: For each domain, the "x" marks the reason(s) we believe the SAM-T2K alignment with human intervention was superior to the SAM-T99 automatic one. We have not attempted a finer quantization of how much improvement each technique made, though the contribution varies from target to target. Although many targets involved some hand-editing of the alignment, we have not attempted to assess how much benefit was obtained from this editing.
For T100, we have given credit to the 2-track HMMs, which would have given the correct fold, had they been available at the time.
The single most important change is the addition of 2-track HMMs using the secondary structure predictions, though the combined effect of splitting into domains, looking at the function of the protein, using a newer template library, checking for missing interior strands of beta sheets, and using active-site information may be at least as important.

99ER62849, and a National Physical Sciences Consortium graduate fellowship. We are grateful to David Haussler and Anders Krogh for starting the hidden Markov model and Dirichlet mixture work at UCSC, as these approaches were instrumental to our success.

References

[1] Manfred J. Sippl, Peter Lackner, Francisco S. Domingues, Andreas Prlić, Rainer Maik, Antonina Andreeva, and Markus Wiederstein. Assessment of the CASP4 fold recognition category. *Proteins: Structure, Function, and Genetics*, 2001.

[2] Kevin Karplus, Christian Barrett, and Richard Hughey. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.

[3] Kevin Karplus, Christian Barrett, Melissa Cline, Mark Diekhans, Leslie Grate, and Richard Hughey. Predicting protein structure using only sequence information. *Proteins: Structure, Function, and Genetics*, Supplement 3(1):121–125, 1999.

[4] NR (All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF Database) Distributed on the Internet via anonymous FTP from `ftp://ftp.ncbi.nlm.nih.gov/blast/db`. Information on NR is available at `http://www.ncbi.nlm.nih.gov/BLAST/blast_databases.html`.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. *JMB*, 215:403–410, 1990.

[6] WU-BLAST WWW archives. `http://blast.wustl.edu/`.

[7] I. Holmes and R. Durbin. Dynamic programming alignment accuracy. *Jour. Comp. Biol.*, 5(3):493–504, 1998.

[8] Richard Hughey, Kevin Karplus, and Anders Krogh. SAM: Sequence alignment and modeling software system, version 3. Technical Report UCSC-CRL-99-11,

University of California, Santa Cruz, Computer Engineering, UC Santa Cruz, CA 95064, October 1999. Available from `http://www.cse.ucsc.edu/research/compbio/sam.html`.

[9] T. Hubbard, A. Murzin, S. Brenner, and C. Chothia. SCOP: a structural classification of proteins database. *NAR*, 25(1):236–9, January 1997.

[10] Dani Fisher? CAFASP2? *Proteins: Structure, Function, and Genetics*, 2001.

[11] M. Janine Parkinson, J. Richard G. Pöhler, and David M. J. Lilley. Catalytic and binding mutants of the junction-resolving enzyme endonuclease I of bacteriophage T7: role of acidic residues. *NAR*, 27(2):682–689, 1999.

[12] K. K. Koretke, R. B. Russell, R. R. Copley, and A. N. Lupas. Fold recognition using sequence and secondary structure information. *Proteins: Structure, Function, and Genetics*, Supplement 3(1):141–8, 1999.

[13] E.L.L Sonnhammer, S.R. Eddy, and R. Durbin. Pfam: A comprehensive database of protein families based on seed alignments. *Proteins: Structure, Function, and Genetics*, 28:405–420, 1997.

[14] Robert M. MacCallum, Lawrence A. Kelley, and Michael J. E. Sternberg. SAWTED: Structure assignment with text description-enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics*, 16(2):125–129, February 2000.

[15] Timothy L. Bailey and William N. Grundy. Classifying proteins by family using the product of correlated p-values. In *Int. Conf. Computational Molecular Biology (RECOMB99)*, pages 10–14. ACM Press, April 11-14 1999.