

Predicting protein structure using hidden Markov models

| | | | |
|---|--|---|--|
| Kevin Karplus [†] Computer Engineering U.C. Santa Cruz karplus@cse.ucsc.edu | Kimmen Sjölander Computer Science U.C. Santa Cruz kimmen@cse.ucsc.edu | Christian Barrett Computer Engineering U.C. Santa Cruz cbarrett@cse.ucsc.edu | Melissa Cline Computer Science U.C. Santa Cruz cline@cse.ucsc.edu |
| David Haussler Computer Science U.C. Santa Cruz haussler@cse.ucsc.edu | Richard Hughey Computer Engineering U.C. Santa Cruz rph@cse.ucsc.edu | Liisa Holm EBI, England holm@ebi.ac.uk | Chris Sander EBI, England sander@ebi.ac.uk |

Proteins: Structure, Function, and Genetics. Pages 134–139, Supplement 1, 1997

Abstract

We discuss how methods based on hidden Markov models performed in the fold-recognition section of the CASP2 experiment. Hidden Markov models were built for a representative set of just over one thousand structures from the Protein Data Bank (PDB). Each CASP2 target sequence was scored against this library of HMMs. In addition, an HMM was built for each of the target sequences, and all of the sequences in PDB were scored against that target model, with a good score on both methods indicating a high probability that the target sequence is homologous to the structure. The method worked well in comparison to other methods used at CASP2 for targets of moderate difficulty, where the closest structure in PDB could be aligned to the target with at least 15% residue identity.

1 Introduction

One method of protein sequence analysis is the identification of *homologous* proteins—proteins which share a common evolutionary history and have similar overall structure and function [5]. Here we report how new extensions of the hidden Markov model (HMM) methods [15, 12] for recognition of remote homologs fared in the fold-recognition section of the CASP2 experiment. We used linear HMMs trained on sets of aligned or unaligned sequences, using sequence weighting and Dirichlet mixture methods to estimate the emission probabilities for the amino acids in each state based on the training data (see Sections refsec:target and 2.3).

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 [13, 23, 16, 17] in that pairwise (residue-residue) interactions are not modeled or used. Instead, we employ Bayesian methods [4, 3, 21] to incorporate prior information in the form of *Dirichlet mixture* densities [24] over position-specific amino acid distributions, and over insertion and deletion probabilities in different structural environments (Section 2.1). The priors reflect different patterns of sequence conservation, such as invariant or hydrophobic, and can be combined with data from aligned homologs to form data-dependent parameter estimates. This differentiates our approach from that of Eisenberg and colleagues [7], which incorporates more structural and less sequence information.

In the CASP2 experiments, we developed a new sequence weighting scheme (see Section 2.3), and a method for constructing *joint models* for two sets of presumably homologous proteins (Section 2.4). We also applied a number of posthoc analysis tools to discriminate among the top potential matches (Section 2.5).

[†]To whom correspondence should be addressed. Mailing address: Computer Engineering Department, Jack Baskin School of Engineering, University of California, Santa Cruz, CA 95064 USA. Phone: 1-408-459-4250, Fax: 1-408-459-4829. Mail to other UCSC authors may be similarly addressed.

Section three discusses the success of our predictions. We used the SAM¹ HMM software suite in these experiments [12].

2 Methods

Our method for predicting the structure of a target sequence involved a two-pronged approach: constructing an HMM from the target and identified homologs, and scoring the sequences in PDB with this model; and scoring each target sequence against a library of HMMs constructed on a representative subset of PDB.

Those PDB sequences that scored high on one or (preferably) both lists of potential matches to a target were examined more closely (Section 2.5).

2.1 The HMM library

Our model library included about 1000 (now 1312) structures from PDB, the core of which was a representative set of PDB structures. For each of these structures, we constructed an HMM (a *structure model*) using the associated HSSP alignment [22] of the structure and its homologs as the initial basis. This alignment and the corresponding HMM parameters were re-estimated using standard HMM methods in combination with priors over amino acids and transition probabilities in various structural environments [14]. The transition priors allowed us to incorporate general structural information, such as the low probability of an insert in the middle of a helix. Following re-estimation, we applied sequence weighting (Section 2.3) to generalize the models for recognition of remote homologs.

2.2 Building the target model

An initial model was constructed from the target sequence only, using SAM's `modelfromalign` module. This established the length of the model to be the number of positions in the sequence, and provided a mapping between the states of the model and the residues in the sequence. This initial model was used to select homologs from a set of neighbors from the Entrez database. The model parameters were re-estimated repeatedly on the target sequence and homologs, using Dirichlet mixture densities over amino acid distributions and a variety of different transition priors [14].

Because proteins can have repeated domains, the `multidomain` module of SAM was used to select sub-sequences from the putative homolog set. For instance, some homologs to t0004 (the nucleotidyltransferase S1 motif) had 3 or 4 regions that matched the model.

The alignment of the target and homologs (with potentially several regions of alignment to some homologs) was used as the basis of an HMM. Sequence weighting was used to control the generality of the model (see Section 2.3).

For some targets (t0011,t0019, t0026, t0030), the initial set of training sequences was too small, and so a search was done of a non-redundant protein database [19] using the model, and the sequences with cost less than -8.0 nats² were considered possible homologs. The model-building procedure was repeated for this larger training set.

2.3 Weighting schemes

Almost any set of homologous proteins will contain some highly populated subfamilies and some less populated subfamilies, and a model constructed from it will favor the most highly represented sequences. To reduce training-set bias, sequence-weighting schemes assign relative weights to training sequences. The particular method used to assign the relative weights in the CASP2 contest is described in the technical report [14], but similar results would have been obtained by using a scheme such as the Henikoffs' [8] for the relative weights.

In Bayesian methods (such as our use of Dirichlet mixtures), the total weight assigned to the set of training sequences has a large impact on the posterior amino acid distributions used to estimate the model

¹ <http://www.cse.ucsc.edu/research/compbio/sam.html>

² The cost (or score) for a sequence s with respect to a model is $-\ln(P(s|\text{model})/P(s|\text{null})) + \ln|s|$, where $|s|$ is the length of s , and the null model assumes each amino acid is generated independently according to a distribution that is the geometric average of the distributions in the match states of the model, normalized to sum to 1. The probability $P(s|\text{model})$ is computed by summing over all local alignments [1, 18, 2]. Note that the more negative the cost, the better the fit to the model. The choice of -8.0 nats was rather arbitrary, based on casual observations of how well the putative homologs in HSSP scored.

parameters. Given few sequences (low total weight), there is only a faint signal for the modelled family and the posterior amino acid distributions will be close to the background frequencies of amino acids. As the number of observations increase (large total weight), the posteriors will reflect the frequencies in the data, closely modeling the training set. By adjusting the total weight, one can smoothly interpolate between the background frequencies and observed data frequencies, with intermediate weights giving very natural generalizations.

In remote homolog recognition, we need models that generalize as much as possible without losing the ability to recognize the training set. Rather than specify the total weight directly, we specify the generality of the model as the average entropy of the posterior amino acid distribution relative to the background frequencies used in the null model. For the CASP2 contest, we chose a relative entropy of 0.3 bits per alignment column for both target and structure models. By way of contrast, a PAM distance of 120 is about 1.0 bits per column and a PAM distance of 240 is about 0.5 bits per column. The total weights assigned ranged from 0.28 to 1.19 for numbers of sequences ranging from 5 to 147 [14].

2.4 Estimating joint models

If a structure and target are distant homologs, the alignment of the target homologs to the structure model may not maintain a good mutual alignment of the target homologs, and conversely for an alignment of the structure homologs to the target model. This reduces our ability to predict the correct pairwise alignment between a target and a structure.

If two sets of proteins share a common structure and evolutionary history, then we ought to be able to construct a statistical model that gives high probability to both sets, and hence better alignments. This motivated the development of two methods for estimating *joint models*, using homologs of both the target sequence and the PDB sequence together.

One method for constructing a joint model employed the method for building target models, except that the homolog set included the PDB sequence’s homologs and the thresholds were set low enough to force inclusion of them in the training set. The second method retrained an existing model using both sets of homologs, keeping the model length fixed, and assigning sequence weights to allow roughly equal weight to both groups of sequences.

The joint models were more successful at producing multiple alignments of both sets of homologs that retained the mutual alignment within each group, and provided better alignments between the regions of lower sequence identity in the two groups.

2.5 Posthoc analysis tools

Automated methods identified potential matches in our structure library and produced the respective pairwise alignments. Each potential match was then inspected for the quality of the alignment, similarity in biological function, and consistency with other structural assessments. We checked PhD secondary structure predictions [20] for consistency with our structure predictions. The alignments were inspected with Leslie Grate’s SAE, a graphical tool combining RASMOL and an alignment viewer.³ This allowed us to see if insertions and deletions occurred in reasonable regions of the structure and whether the resulting protein structures were compact and contiguous. Alignments were further examined with Liisa Holm’s solvation analysis software [9, 10], which built 3D models given the target-structure alignment, assessing whether a protein-like hydrophobic core was formed. This solvation analysis is very sensitive to the correctness of the input alignment.

3 Results

3.1 Fold-recognition results

Table 1 shows how our HMM models scored on the eight targets for which we submitted predictions and received feedback. For consistency since our methods and library evolved significantly over the course of the summer, all scores are reported with the method and library from the end of the summer.

³SAE is a prototype tool that we do not intend to release, but a successor tool, DINAMO, will be released soon—see <http://tito.ucsc.edu/dinamo> for more information.

| target | structure | bet cost | target | | library | | DALI | | predicted |
|--------|--------------------|-------------|---------------|------|---------------|-------|------|-----|-----------|
| | | | rank /7991 | cost | rank /1312 | score | %ID | %ID | |
| t02 | 1psdA | 0.0 | -3.8 | 221 | -1.5 | 278 | 1.00 | 6 | |
| t02 | 1wsyB [‡] | 1.0 | -38.6 | 1 | -17.0 | 3 | 1.00 | 20 | 24 |
| t04 | 1csp [‡] | 0.6 | -7.1 | 1 | -0.9 | 223 | 1.00 | 21 | 29,30 |
| t04 | 1mjc [‡] | 0.4 | -5.4 | 7 | -2.6 | 24 | 1.00 | 23 | 22,24 |
| t11 | 1grl [†] | 1.0 | -2.0 | 1656 | -2.7 | 55 | 0.00 | | 25 |
| t11 | 3gapB [†] | 0.5 | -3.5 | 197 | -6.0 | 3 | 0.00 | | 21 |
| t11 | 1frpA [†] | 0.5 | -3.1 | 505 | -5.9 | 4 | 0.00 | | 19 |
| t12 | 1mdyA [†] | 0.5 | -0.6 | 3173 | -6.0 | 2 | 0.00 | | 20 |
| t12 | 1pht [‡] | 0.5 | -2.1 | 302 | -2.1 | 85 | 0.00 | | 16 |
| t12 | 1atr | 0.0 | -0.7 | 2829 | -0.7 | 432 | 0.22 | | |
| t12 | 1gerA | 0.0 | -1.9 | 372 | | | 0.28 | | |
| t20 | 1arv [‡] | 1.0 | -7.5 | 2 | -1.4 | 319 | 0.00 | | 18,24 |
| t20 | 7aatA | 0.0 | -1.8 | 2903 | -5.0 | 4 | 0.43 | 5 | (18) |
| t20 | 1scuB | 0.0 | -2.7 | 1130 | -3.1 | 42 | 0.55 | | (11) |
| t20 | 2dln | 0.0 | -5.6 | 24 | -0.6 | 705 | 0.82 | | (17) |
| t20 | 1minA | 0.0 | -4.9 | 67 | -2.7 | 64 | 1.00 | | (19) |
| t30 | 2hwf1 [†] | 0.1 | -1.2 | 805 | -0.4 | 446 | 0.00 | | 19 |
| t30 | 1hsbA [†] | 0.1 | -0.3 | 4254 | -2.9 | 7 | 0.00 | | 32 |
| t30 | NONE [†] | 0.8 | | | | | 1.00 | | |
| t31 | 1fonA | 0.0 | -10.8 | 1 | | | 1.00 | | |
| t31 | 1hcgA | 0.0 | -1.9 | 990 | -12.3 | 6 | 1.00 | 14 | |
| t31 | 4ptp | 0.0 | -1.8 | 1085 | -14.9 | 3 | 1.00 | 15 | |
| t31 | 1elt [†] | 0.2 | -8.6 | 4 | -11.3 | 7 | 1.00 | 14 | 16 |
| t31 | 1mctA [†] | 0.2 | -2.1 | 816 | -15.0 | 1 | 1.00 | 16 | 21 |
| t31 | 1try [‡] | 0.6 | -10.5 | 3 | -13.1 | 5 | 1.00 | 18 | 18 |
| t38 | 1exg [‡] | 1.0 | -3.0 | 200 | -0.9 | 217 | 0.57 | 15 | 11,17,28 |
| t38 | 1lpbB | 0.0 | -5.2 | 1 | -0.7 | 268 | 0.85 | 8 | |
| t38 | 1celA | 0.0 | -0.8 | 2897 | -2.4 | 22 | 1.00 | 10 | |
| t38 | 1bglA | 0.0 | -4.3 | 75 | -1.0 | 181 | 1.00 | 10 | |
| t38 | 2ayh | 0.0 | -0.8 | 3082 | -0.3 | 705 | 1.00 | 10 | (21,31) |

Table 1: Scoring of our predicted sequences (marked with [†] for UCSC-only predictions and [‡] for joint UCSC-EBI predictions) and some of the lowest-cost sequences which DALI [11] considered to have similar structure. Structures are listed in increasing order of similarity to the known structure.

The bets for t11 add up to more than 1.0 because we submitted predictions for two separate domains.

Ranks and scores are based on an October 1996 version of PDB and the April 1997 version of our HMM library. Ranks are somewhat inflated by redundancy (e.g., there are 5 sequences identical to 1csp in the PDB database, so the rank of 7 for 1mjc would be 3 in a non-redundant database).

DALI scores are rescaled so that $Z \leq 2$ becomes 0 and $Z \geq 6$ becomes 1, as reported by the CASP2 assessors. The % residue identity for DALI alignments is also as reported by the assessors. For target t2, we have had to use VAST scores and residue identity, as the DALI results were never sent to us. Missing DALI % figures were not reported by the assessors.

Residue identities in parentheses are for alignments we considered but did not submit. Multiple residue identities are for multiple competing alignments that we submitted.

Analyzing where our HMM methods succeed and where they fail shows that prediction success is correlated with sequences being only moderately divergent—having a pairwise residue identity of at least 15%. Note that higher residue identities are sometimes reported for incorrect fold predictions. In most cases these are from hand-edited alignments that increased residue identity at the cost of greatly increasing the number of gaps. Residue identity alone is a very poor measure of similarity in a gapped alignment.

HMM scores of remote homologs show some discrimination capability above this point, and posthoc analysis is sufficient to differentiate the true homologies from the pool of candidate structures, given reasonably

| target | structure | | Alignment Length | Residues Aligned Correctly | Avg. Shift | Avg. RMSD | SC%id | %ID | Alignment Specificity | Alignment Sensitivity |
|--------|-----------|------|------------------|----------------------------|------------|-----------|-------|-------|-----------------------|-----------------------|
| t02 | 1wsyB | VAST | 245 | 117 | 1.316 | 5.15 | 20.07 | 24.08 | 47.76 | 51.09 |
| t04 | 1csp | VAST | 63 | 34.80 | 0.338 | 3.52 | 24.53 | 29.37 | 55.24 | 65.66 |
| t04 | 1mjc | DALI | 62 | 39.14 | 0.471 | 3.64 | 22.58 | 23.20 | 62.62 | 63.23 |
| t27 | 2pec | DALI | 319 | 99 | 3.938 | 14.40 | 21.93 | 24.14 | 31.03 | 36.80 |
| t28 | 1celA | DALI | 359 | 342 | 0.205 | 2.37 | 48.74 | 49.30 | 95.26 | 95.80 |
| t31 | 1elt | DALI | 200 | 111 | 2.427 | 8.73 | 13.90 | 15.50 | 55.50 | 59.36 |
| t31 | 1mctA | DALI | 195 | 105 | 2.437 | 8.77 | 15.85 | 20.51 | 53.85 | 57.38 |
| t31 | 1try | DALI | 198 | 101 | 1.624 | 7.45 | 17.74 | 18.69 | 51.01 | 54.30 |

Table 2: This table compares our alignments of the targets to the structural alignments produced by VAST or DALI. *Alignment length* refers to the total number of residues aligned, including loop regions. *Residues Aligned Correctly* describes the number of positions in which the alignment was correct, as compared to the structural alignments. *Avg. RMSD* and *Avg. Shift* refer to the average RMS deviation and shift, as computed by the assessors. *SC%ID* describes the percent residue identity for each structural alignments, and *%ID* describes the percent residue identity of our alignment. *Alignment Specificity* and *Alignment Sensitivity* refer to the number of correctly aligned residues as a fraction of the number aligned in the prediction and the number aligned in the structural alignment, respectively.

accurate alignments. However, when pairwise residue identity drops below this level, HMM scores are weaker and less informative, resulting in a large pool of poorly aligned candidate structures which our posthoc analysis tools cannot differentiate amongst.

For example, for targets t0002, t0004, and t0031, the predicted structures were excellent structural matches, and the top-scoring match was correct in the target model. All three of these targets had pairwise alignments to their closest structural match of at least 16%.

For target t0002, like most groups we misunderstood the rather cryptic comment about partial homology and only predicted the domain homologous to 1wsyB for which simple sequence methods already provided an adequate prediction of homology. We had made some attempts to predict the other domain, but we did not come up with a prediction sufficiently believable to be submitted. We hope that future contests label targets more clearly when partial prediction is desired.

For t0031, the best match in Table 1 with the target model is for 1fonA, which was not in PDB at the time the initial searches were done. The predicted folds 1try and 1elt were the best scoring sequences at that time.

For t0020 and t0038, the closest structural matches that scored well in both directions were 1minA and 1bglA, respectively, but we failed to identify them for the contest.

For t0020, 1minA scored within the top 70 for both target and library models and was considered along with several other good structural matches, but we concentrated on the incorrect 1arv because of a perceived need for an iron-binding site.

For t0038, 1bglA with 9% residue identity ranked 75 with the target model and 181 with the library models and was too low ranked to be considered in our analysis. We did consider one correct structure for t0038 (2ayh), based not on its scores but on its function (we looked at all glucanases and cellulases with structures in PDB), but we rejected it because of a too strict interpretation of solvation scores. While weak, our 1exg prediction did have a somewhat similar fold, and, interestingly, was also predicted by several other groups. We wonder if there is a distant evolutionary relationship between 1exg and t0038.

Targets t0011 and t0030 were determined to be novel structures by DALI and VAST. We predicted t0030 essentially correctly, placing 80% of our “bet” on NONE, indicating that we felt that we had found no similar structure in PDB. However, we did not do this with t0011, despite its fairly weak scores.

Target t0012 had only very weak structural homologs; our prediction, 1mydA, a helix-turn-helix, aligned well to a similar secondary structure in t0012, but did not yield a useful global structural match.

3.2 Quality of alignments

We submitted alignments for ten fold-recognition targets and three comparative modeling targets (count-

ing t0002 as a comparative modeling target). Results are available for eight of those targets, and Table 3.2 summarizes our alignments for the three comparative modeling targets (t0002, t0027, and t0028) and the two fold-recognition targets for which we identified a correct fold (t0004 and t0031).

Our method searched for global, rather than local, alignments between a target and a structure. While the average shift in the alignments is generally quite low, this resulted in alignments with a higher RMS distance compared to other groups. Loop regions have a high degrees of divergence, and so identifying these regions and removing them from the alignment would improve the evaluation of our alignments—this improvement is one we hope to have implemented by the next CASP contest.

Perhaps the most striking difference between the RMS measure and the other measures of correctness for the alignment is for our t0031 predictions. The number of exactly correct residues is high and the average shift is quite low, but the RMS deviation is suprisingly high. This results from a single segment (residues 163-183) which is badly misaligned (see Figure 1. The segment should be aligned to residues 149-168 of 1mctA which includes the edge strand of a conserved beta sheet. Instead the edge strand was skipped and the segment was aligned to a loop and helix on the surface. This misalignment should have been detected before we submitted the prediction, since it results in a large distance between the predicted positions of residues 162 and 163, but we failed to notice the problem.

Our alignment for t0027 and 2pec was reasonable in the beta sheets of the core, but we included alignments for the rather variable surface helices which turned out to be rather different in the two structures. Trimming our global alignment to remove the surface elements would have considerably improved the statistics for the prediction.

It is interesting that we got better alignments for t0004 and t0031, which were classified as a fold-recognition target, than for t0027, which was classified as a comparative modeling target. Perhaps in future CASP contests, the targets should not be pre-classified, but all targets should be made available for all prediction types. The assessment for each type of prediction can then focus on the targets that show a difference between the predictors. Existing servers for sequence-based alignment can be used as baseline comparisons to see whether the more sophisticated methods provide better results on the easy targets.

3.3 Conclusions

Fold recognition and alignment by HMMs shows considerable promise, but there were too few targets with homologs of known structure to draw a definitive conclusion. Our method seems to be effective in cases where the residue identity between the target and the sequence of known structure is in the 15–25% range, which brings us some distance into the “twilight zone,” but we have no evidence yet that it will be effective in harder cases.

Our methods were developed very hastily while the contest was in progress, and we had no time to validate the methods before making predictions. We are now building a new library, developing new methods, and putting together a test suite for the methods. These improvements include refinement of the SAM software suite, better methods for building target models, a more complete library, better methods for building joint models, better post-processing to identify problems in proposed alignments, methods to adjust alignments to remove or realign the unreliable parts, and extensive testing of the methods to determine which are most useful.

We hope to have some tested methods available in time for the CASP3 contest, and we will be putting up at least the automatic part of these methods on our web site <http://www.cse.ucsc.edu/research/compbio>.

Since HMMs do not use pairwise contacts, they are more computationally efficient than threading models. Their minimal dependency on structure information also allows them to be used to search for remote homologs of protein families that contain no sequence with known structure. It may turn out that other techniques, which make real use of structural information, may be able to do better at finding veryd distant homologs, but we feel that the HMM methods can still be improved enough to remain competitive with more expensive methods.

Acknowledgments

This work reflects the contributions of many, including Leslie Grate, Chris Tarnas, Ole Winther, Rachel Karchin, Marc Hansen, Mark Diekhans, Rey Rivera, Tony Fink, and Lydia Gregoret.

This work was supported in part by NSF grants CDA-9115268, IRI-9123692, and BIR-9408579; DOE grant 94-12-048216; ONR grant N00014-91-J-1162; NIH grant GM17129; NFS and GAANN graduate fellowships; and the UCSC Division of Natural Sciences.

References

- [1] S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *JMB*, 219:555–565, 1991.
- [2] C. Barrett, R. Hughey, and K. Karplus. Scoring hidden Markov models. *CABIOS*, 13(2):191–199, 1997.
- [3] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.
- [4] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley and Sons, first edition, 1994.
- [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] D. Fischer and D. Eisenberg. Protein fold recognition using sequence-derived predictions. *Protein Sci.*, 5(5):947–955, May 1996.
- [8] S. Henikoff and J. G. Henikoff. Position-based sequence weights. *JMB*, 243(4):574–578, Nov. 1994.
- [9] L. Holm and C. Sander. Evaluation of protein models by atomic solvation preference. *JMB*, 225:93–105, May 1992.
- [10] L. Holm and C. Sander. Fast and simple monte-carlo algorithm for side chain optimization in proteins: Application to model building by homology. *Proteins: structure, function and genetics*, 14:213–223, Oct 1992.
- [11] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *JMB*, 233(1):123–138, 5 Sept 1993.
- [12] R. Hughey and A. Krogh. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS*, 12(2):95–107, 1996.
- [13] D. Jones and J. Thornton. Protein fold recognition. *J. Comput. Aided Mol. Des.*, 7:439–456, 1993.
- [14] K. Karplus, Kimmen Sjölander, C. Barrett, M. Cline, D. Haussler, R. Hughey, L. Holm, and C. Sander. Predicting protein structure using hidden Markov models, the CASP2 contest. Technical Report UCSC-CRL-97-13, University of California, Santa Cruz, Computer Science, UC Santa Cruz, CA 95064, 1997.
- [15] 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.
- [16] R. Lathrop, L. Ljubomir, R. Nambudripad, J. White, L. L. Conte, B. Bryant, and T. Smith. Threading through the levinthal paradox. *Nature*, To appear.
- [17] 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.
- [18] A. Milosavljević and J. Jurka. Discovering simple DNA sequences by the algorithmic similarity method. *CABIOS*, 9(4):407–411, 1993.
- [19] NRP (Non-Redundant Protein) Database. Distributed on the Internet via anonymous FTP from ftp.ncifcrf.gov, under the auspices of the National Cancer Institute’s Frederick Biomedical Supercomputing Center.
- [20] B. Rost. Phd: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.*, 266:525–39, 1996.
- [21] T. J. Santner and D. E. Duffy. *The Statistical Analysis of Discrete Data*. Springer Verlag, New York, 1989.
- [22] R. Schneider and C. Sander. The HSSP database of protein structure-sequence alignments. *NAR*, 24(1):201–205, 1 Jan 1996.
- [23] M. Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, 5:229–235, 1995.
- [24] 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, 1996.