Confidence Measures for Fold Recognition

Ingolf Sommer and **Niklas von Öhsen** and **Alexander Zien** and **Ralf Zimmer** and **Thomas Lengauer**
Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, D-53754 Sankt Augustin,
Tel.: +49 2241 142360, Fax: +49 142656
{ingolf.sommer, niklas.von-oehsen, alexander.zien, ralf.zimmer, lengauer}@gmd.de

Introduction

It is a standard procedure to compare new amino acid sequences to databases of proteins that have been studied already in order to find similarities in structure and function. This comparison can be sequence–sequence or sequence–structure based. In order to compare, an alignment is performed of the target protein sequence (whose structure we are searching) with a template protein (whose structure we know). For a sequence–sequence alignment, the alignment algorithm optimizes a certain scoring function that quantifies the similarities of the amino acids at individual positions. For a sequence–structure alignment, also known as threading, usually the scoring function that is optimized is designed to capture the essence of structural similarity among proteins.

These scores are supposed to be comparable between different proteins, since we want to select the template which achieves the highest alignment score to the target protein as our candidate for the structural model of a protein. The involved scoring functions are inaccurate, however. Thus, it is very helpful if the method can augment the generated alignment and its score with a statistical significance value which captures the confidence that we can put into the generated alignment.

While important for protein alignment and threading as stand-alone tools, significance scores are even more essential if protein alignment is used in an automated cascade of tools for protein structure prediction.

In this paper we analyze the performance of several variants of the 123D protein threading method Alexandrov, Nussinov, & Zimmer and compare it to several variants of optimal sequence alignment. Where theoretically available, we analyze the statistical significance of scores. For the other methods, we propose empirical approximations to p-values and evaluate their validity.

Experimental Setup

Protein Data In order to estimate score distributions and to evaluate different confidence measures we aligned and threaded a set of amino acid sequences versus a reference set of structures. For both sequences and structures we chose proteins whose structure was determined experimentally. For our threading experiments we used two subsets representative for the Protein Data Base (PDB)[1] (Berman *et al.* 2000). The subset PDB40D is a set of all SCOP protein domains with less than 40% sequence homology as computed by the Astral-Server (Brenner, Koehl, & Levitt 2000), where SCOP is the Structural Classification of Proteins database[2] (Murzin *et al.* 1995). Subset PDB40C is the set of protein chains where at least one domain of the protein belongs to the PDB40D. The PDB40D contains 2860 protein domains, the PDB40C set contains 2232 proteins.

The pair-scores investigated in this study are computed by sequence aligning and threading the sequences of the PDB40C versus the structures of the PDB40D. This setup was chosen to resemble experiments with unknown sequences, as e.g. the Critical Assessment of Techniques for Protein Structure Prediction (CASP)(Benner, Cohen, & Gerloff 1992; Barton & Russell 1993; Defay & Cohen 1995; Jones 1997; Fischer *et al.* 1999; Fischer, Elofsson, & Rychlewski 2000).

Alignment & Threading Parameters We compare eight different methods for detecting remotely homologous protein folds, resulting from three independent binary choices: sequence alignment and threading, global and local, with and without frequency profiles

sequence alignment / threading	s/t
plain sequence / frequency profile	s/f
global / local	g/l

Optimal global sequence alignment with affine gap costs is done with a standard Gotoh algorithm (Gotoh 1982), local sequence alignment is performed using the Smith-Waterman algorithm (Smith & Waterman 1981). The alignment was performed using the Dayhoff PAM250 matrix, the algorithms were applied to plain sequences as well as to frequency profiles.

The basic version of the 123D threading tool is described in (Alexandrov, Nussinov, & Zimmer 1996). All threading experiments described in this paper are performed with an implementation that is further developed in two respects: 1. Instead of threading the target sequence itself against a template structure with its native amino acids, we can also employ frequency profiles. 2. The scoring function, originally a

[1] version of February 29th 2000

[2] release 1.50 of February 29th 2000, classifying 10650 PDB protein entries and 24186 protein domains

sum of inverse Boltzmann derived potentials, is tuned by optimally weighting the individual contributions against each other (Zien, Zimmer, & Lengauer 2000).

Fold Recognition Performance vs. Confidence Measure Methods

In the fold recognition protocol, for each target protein sequence we try to find a template protein structure, identifying a corresponding fold class. This is done by evaluating a scoring function for each target-template pair and sorting the template scores for one target. The fold of the highest scoring template is then predicted to be the fold of the target sequence. For this template a confidence score can be computed in order to estimate the validity of the prediction.

The confidence can be easily computed if the score distribution is known. However, this is not the case for all alignment methods considered.

For optimal local sequence alignments of independent random sequences the scores are known to be asymptotically extreme-value or Gumbel distributed (Altschul *et al.* 1990; Karlin & Altschul 1990; Waterman & Vingron 1994; Altschul & Gish 1996; Pearson 1998; Levitt & Gerstein 1998; Mott 2000). Local alignments with sequence-profiles were also shown to follow an extreme-value distribution (Mott 2000). For optimal global alignments, whether with plain sequences or sequence profiles, neither the family of distributions nor the dependence of the expected score on the sequence lengths is known. The situation is similar for 123D threading: The local threading scores of sample sequence-structure pairs closely follow a Gumbel distribution. For global threading the distribution is unknown. Sample score distributions depend on the length of the sequences but neither resemble a Gaussian nor a Gumbel distribution.

Scoring & Confidence Functions

Let SEQ be a set of amino acid sequences of proteins and STR be a set of structures of proteins, then a scoring function

$$\begin{aligned} f_{score} : \mathrm{SEQ} \times \mathrm{STR} &\longrightarrow \mathbb{R} \\ (seq, str) &\longmapsto f_{score}(seq, str) \end{aligned}$$

is applied to each target sequence *seq* to find a related template structure $pred_{seq}$ with

$$f_{score}(seq, pred_{seq}) = \max_{str \in \mathrm{STR}_{seq}} f_{score}(seq, str)$$

where

$$\mathrm{STR}_{seq} = \{str \in \mathrm{STR} : sequence(str) \text{ not subsequence of } seq\}.$$

The fold of the structure $pred_{seq}$ is predicted to be the most plausible fold for sequence *seq*. A confidence function

$$\begin{aligned} f_{conf} : \mathrm{SEQ} \times \mathrm{STR} &\longrightarrow \mathbb{R} \\ (seq, str) &\longmapsto f_{conf}(seq, str) \end{aligned}$$

is then used to estimate the validity of the prediction as $f_{conf}(seq, pred_{seq})$. This measure tells how much trust to put into the result of the alignment or threading, i.e. how closely related we expect the target and the template protein to be.

The confidence function f_{conf} is usually calibrated over a set of sequences and structures CAL $\subset$ SEQ $\times$ STR. For example, z-scores (see section below) are computed by estimating mean and standard deviation of the distribution of scores for CAL $=$ *seq* $\times$ STR and normalizing scores according to these. Meta information, like knowledge of the relatedness of sequences and structures or of the relatedness of different structures, may also be used for calibration.

Apart from the calibration, scoring functions and confidence measure functions require the same input and output and can be used interchangeably.

Raw-Score Function (*s*) With raw scores we denote the score of an optimized threading or sequence alignment. Thus, for plain sequences, $s_{\mathrm{ssg}}(seq, str)$ is the score output of the Gotoh alignment of *seq* and *str*, while s_{ssl} denotes the score of the Smith-Waterman alignment, s_{tsl} is the score computed with local 123D threading, and s_{tsg} is the score computed with global 123D threading. Computed with frequency profiles, we denote the raw scores with $s_{*\mathrm{f}*}$ accordingly.

Z-Score Function (*z*) The set of all template proteins that the target sequence is threaded or aligned against is used to normalize the threading scores s_{***} into z-scores z_{***}.

Fitted p-values (*a*,*b*) With a known score distribution the probability that an alignment score of at least this magnitude occurs by chance can be computed. This probability is generally called *p-value*. As noted previously, the parametric form of the score distribution is known for local methods only. These scores are known to depend on the lengths of the two aligned sequences.

For gapless alignments, this dependency is shown to be either logarithmic ($E(s) = c\log(l_1 l_2)$) or linear ($E(s) = c\sqrt{l_1 l_2}$) for scoring matrices with negative or positive expectation value, respectively, with a transition phase in between (Karlin & Altschul 1990). We fit an parameterized extreme value distribution in order to estimate p-values for these scores.

Two different protocols are used for parameter fitting: In the first case (*a*), the fit is performed seperately for each sequence CAL $=$ *seq* $\times$ STR. In the other case (*b*), the data arising from all sequences are joined and the fit is carried out over all seq-struct-pairs CAL $=$ SEQ $\times$ STR.

Tabulated p-values (*u*) The global threading and alignment scores are known to depend on the sequence and structure length, but the nature of that dependence is not known. An approach to estimating score distributions from data is to generate tables of score-percentiles from a set of unrelated sequence-structure pairs. Later, these tables are used to look up the estimated probability of a new score to belong to this set of unrelated pairs.

The scores $s_{***}(seq, str)$ of the unrelated pairs of a calibration set are stored in tables according to lengths of sequences and structures. To estimate the p-value for a new score t with given sequence length and structure length, the previously generated table which suits that sequence and structure length is searched, and the relative frequency of scores below the threshold t is used as an estimate P_{est} for the p-value. Thus

$$u(seq, str) = P_{\text{est}}(score > s(seq, str) : \text{length}(seq), \text{length}(str)).$$

Since this leaves all examples with a score higher than the highest score seen during calibration with a p-value of zero, for later comparisons we sort all pairs with a tabulated p-value of zero according to their raw scores.

Raw Score Gaps and Z-Score Gaps (*sg,zg*) For a target sequence *seq*, the *raw score gap* is the difference $sg(seq, str) = s(seq, str) - s(seq, \text{next}(str))$ of the raw score of a template protein *str* and the next best raw score of a template protein belonging to a different fold, next(str). This gap can be computed for all but the lowest scoring fold. Thus it is always defined for the highest scoring fold and can be used as a confidence measure for fold recognition. Analogously, the *z-score gap* is the difference of the z-score of a template protein and the next best z-score of a template protein belonging to a different fold $zg(seq, str) = z(seq, str) - z(seq, \text{next}(str))$.

Results & Discussion

With our experiments we address two questions: I) Which scoring function $f_{score} \in \{s, z, a, b, u\}$ yields the best fold recognition performance? II) Which confidence measure function $f_{conf} \in \{s, z, a, b, u, sg, zg\}$ is best to evaluate the prediction produced with a given scoring function?

I) Fold Recognition Performance In order to evaluate the fold recognition performance of the different methods, we use the testset PDB40C × PDB40D as described. For the 2232 sequences in PDB40C, 1944 (87.1%) have a structure with corresponding superfamily in the PDB40D, 121 (5.4%) have a structure with corresponding fold, but none with a corresponding superfamily, and 167 (7.5%) have no structure with a corresponding fold (other than themselves) in the PDB40D, according to the SCOP classification.

When we test fold recognition on this set, the maximal performance can be 92.5%, since 7.5% of the sequences have no corresponding fold in the database. We included those sequences into the test set, in order to simulate the ”real world” situation, where a new protein does not necessarily resemble a known structure.

The fold recognition performances of the different scoring functions (raw score (s), z-score (z), estimated p-values (a, b) and tabulated p-values (u)) are listed in Table 1.

II) Comparison of Confidence Measures On the testset PDB40C × PDB40D and the fold recognition results obtained with the scoring function $f_{score} = s$, for the different parameter sets we compare the confidence measure functions $f_{conf} \in \{s, z, a, b, u, sg, zg\}$.

For each confidence measure function, the predictions $\{(seq, pred_{seq}) | seq \in \text{PDB40C}\}$ are sorted according to their confidence function scores $f_{conf}(seq, pred_{seq})$. We thus evaluate the recognition performance for the case that a fold is predicted only for those sequences which attain a confidence higher than a given threshold (i.e. those that the confidence measure declares as “reliable” predictions). For each of the methods and for each threshold t we count the number of predictions, $predicted(t) = |\{seq : f_{conf}(seq, pred_{seq}) > t\}|$ and the number of true positives $tp(t)$, false positives $fp(t)$, true negatives $tn(t)$, and false negatives $fn(t)$ and compute specificity and sensitivity accordingly.

	ssl	sfl	tsl	tfl
s	61.4	69.6	61.6	**70.7**
z	61.4	69.6	61.6	**70.7**
a	51.9	64.9	60.1	66.1
b	**63.6**	**69.7**	**65.4**	**70.7**
u	62.9	68.8	63.2	69.8

	ssg	sfg	tsg	tfg
s	57.2	66.8	60.6	**72.7**
z	57.2	66.8	60.6	**72.7**
u	**59.2**	**67.3**	**62.3**	70.2

Table 1: Fold recognition performance results on the PDB40C × PDB40D for local (above) and global (below) methods. Performances are given in percent, the maximum of each column is set in boldface.

For the parameter sets ssl, sfl, tsl, tfl, ssg, sfg, tsg, and tfg, for each confidence measure function $f_{conf} \in \{s, z, a, b, u, sg, zg\}$, specificity-sensitivity plots are shown in Figures 1 and 2.

Specificity is plotted on the x-axis, sensitivity on the y-axis. By construction, the lines for all methods join at the endpoints. The specificity in the upper left endpoint corresponds to the fold-recognition rate, as listed in Table 1. These endpoints thus show the performance of the underlying scoring function. The second criterion for interpreting the goodness of a confidence measure is the form of the corresponding curve; the closer the curve gets to the upper right corner, the better.

While the tendency for the fold-recognition is clearly visible – threading performs better than sequence alignment, and methods with frequency profiles perform better than methods without –, the interpretation of the different confidence measures is more subtle: For all the local methods, the tabulated p-values (u) perform nearly as well as the p-values (b) calculated according to the Karlin-Altschul model with parameters estimated on the same data basis. We therefore conjecture that these tabulated p-values are reasonable estimates for the significances of global alignments, where no theoretical model is available.

We further notice that the p-values (b, u), tabulated and estimated over all pairs, compared to the p-values (a), with parameters fit on a per sequence basis, have a significant advantage for the plain-sequence cases (*s*), but not for frequency profile cases (*f*). In the latter case, to the best of our understanding, this is due to the lack to fit an additional, unknown parameter. This might be the entropy of the frequency profile, or the number of homologous sequences that the profile was generated from. For the per sequence fit (a), this parameter remains constant over all data and can thus be compensated for by the fitting to the remaining parameters.

For each parameter set ([st][sf][lg]), the curves can roughly be clustered into two groups, better and worse ac-

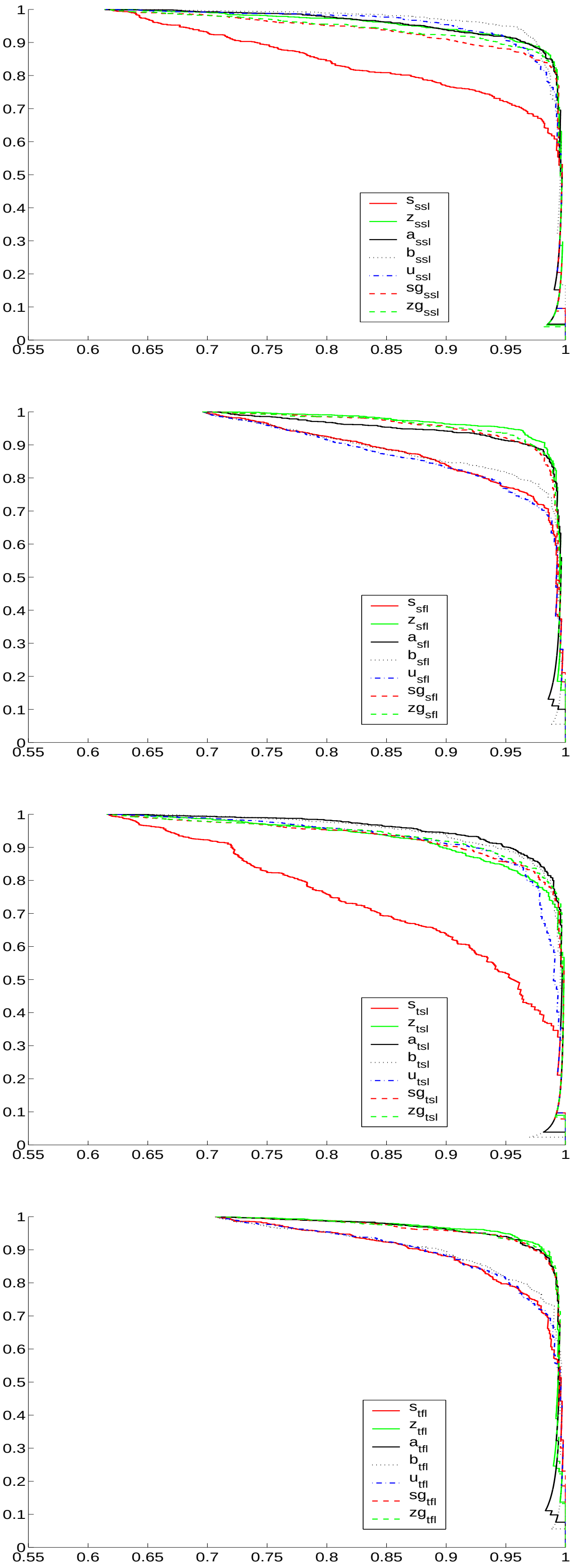

Figure 1: Specificity-Sensitivity plots for the local methods from top to bottom: ssl, sfl, tsl, tfl

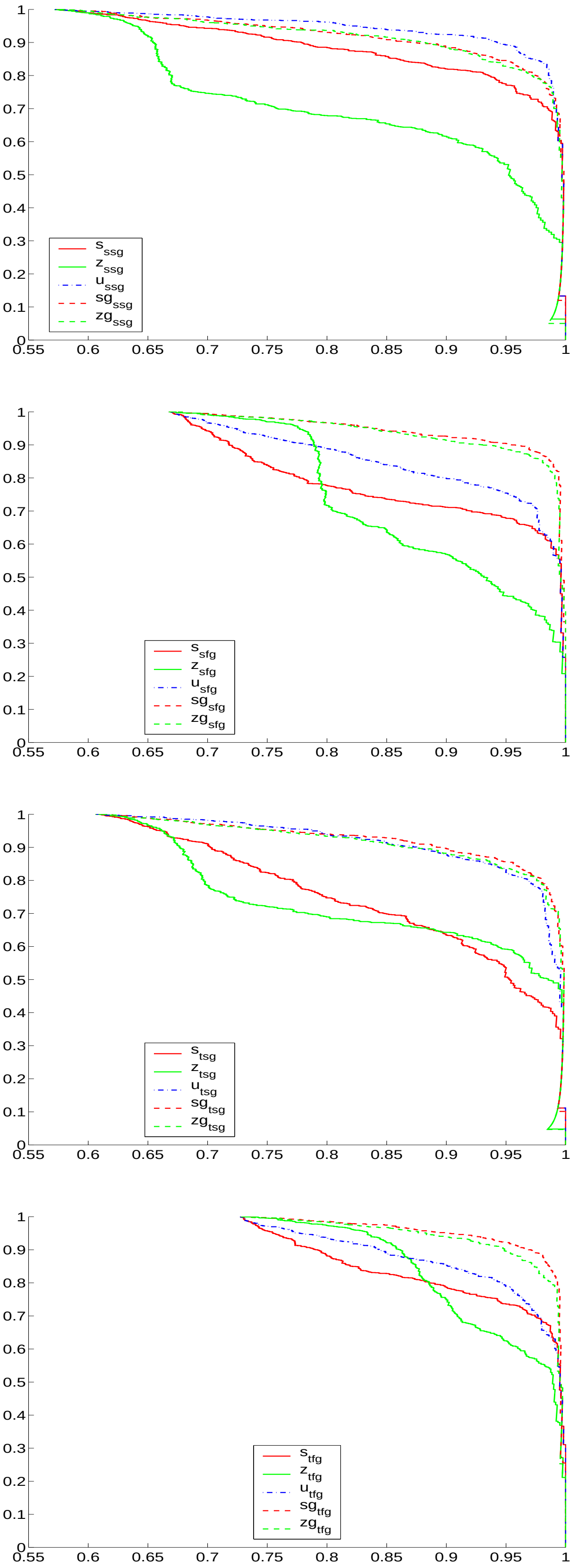

Figure 2: Specificity-Sensitivity plots for the global methods from top to bottom ssg, sfg, tsg, tfg

cording to their performance. For the local parameter sets (**l), z-scores (z), score-gaps (sg), z-score gaps (zg) and p-values fit per sequence (a) perform better than the raw scores (s). As mentioned above, the p-values, tabulated (u) and estimated (b) for all pairs, perform competitively for the plain sequence methods (*s*) only. For the global methods (**g), score-gaps (sg), z-score gaps (zg) and tabulated p-values (u) perform significantly better than raw scores (s) and z-scores (z); Again, the p-values (u) suffer in the frequency profile cases (*fg). Clearly, the z-scores are not adequate for global methods (**g). Contrary to local methods, global methods can produce negative scores. These scores can achieve a high z-score being much better than average for still very negative scores (we found examples for this in our data).

Albeit computationally much simpler to handle than tabulated p-values, we find that the score gaps (sg, zg) perform highly competitive as confidence measures for all methods proposed.

Conclusion We evaluated the performance of different fold recognition methods for a large dataset. We find that threading with frequency profiles performs best according to our measures. For the data set analyzed here, global threading performs better than local. Further we analyzed several confidence measures in order to estimate the validity of a prediction made with one of the above fold recognition methods. We find that score gaps and z-score gaps perform competitively to p-values. From the confidence measures presented, we can empirically estimate the probability of a prediction being correct or incorrect. This estimate becomes essential if protein threading is used in an automated cascade of tools for protein structure prediction.

References

Alexandrov, N.; Nussinov, R.; and Zimmer, R. 1996. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In *Proc. Pacific Symposium on Biocomputing*, 53–69.

Altschul, S., and Gish, W. 1996. Local Alignment Statistics. *Methods Enzymol.* 266:460–480.

Altschul, S.; Gish, W.; Miller, W.; Myers, E.; and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Altschul, S.; Madden, T.; Schäffer, A.; Zhang, J.; Zhang, Z.; Miller, W.; and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17):3389–3402.

Barton, G. J., and Russell, R. B. 1993. Protein structure prediction. *Nature* 361:505–506.

Benner, S. A.; Cohen, M. A.; and Gerloff, D. 1992. Correct structure prediction? *Nature* 359:781.

Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; and Bourne, P. 2000. The protein data bank. *Nucleic Acids Research* 28:235–242.

Brenner, S. E.; Koehl, P.; and Levitt, M. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Research* 28:254–256.

Defay, T., and Cohen, F. E. 1995. Evaluation of current techniques for ab initio protein structure prediction. *Proteins* 23:431–445.

Fischer, D.; Barret, C.; Bryson, K.; Elofsson, A.; Godzik, A.; Jones, D.; Karplus, K. J.; Kelley, L. A.; MacCallum, R. M.; Pawowski, K.; Rost, B.; Rychlewski, L.; and Sternberg, M. 1999. CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins* Suppl 3:209–217.

Fischer, D.; Elofsson, A.; and Rychlewski, L. 2000. The 2000 Olympic Games of protein structure prediction; fully automated programs are being evaluated vis-a-vis human teams in the protein structure prediction experiment CAFASP2. *Protein Eng* 13:667–670.

Gotoh, O. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology* 162:705–708.

Gribskov, M.; McLachlan, A. D.; and Eisenberg, D. 1987. Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the USA* 84(13):4355–4358.

Jones, D. T. 1997. Progress in protein structure prediction. *Curr Opin Struct Biol* 7:377–387.

Kallberg, Y., and Persson, B. 1999. KIND – a non-redundant protein database. *Bioinformatics* 15(3):260–261.

Karlin, S., and Altschul, S. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the USA* 87(3):2264–2268.

Krogh, A., and Mitchison, G. 1995. Maximum Entropy Weighting of Aligned Sequences of Protein or DNA. In Rawlings, C.; Clark, D.; Altman, R.; Hunter, L.; Lengauer, T.; and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligen t Systems for Molecular Biology*, 215–221. Menlo Park, California 94025: AAAI Press.

Levitt, M., and Gerstein, M. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences of the USA* 95(11):5913–5920.

Mott, R. 2000. Accurate Formula for P-Values of Gapped Local Sequence and Profile Alignments. *Journal of Molecular Biology* 300(3):649–659.

Murzin, A. G.; Brenner, S. E.; Hubbard, T.; and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* 247:536–540.

Park, J.; Karplus, K.; Barrett, C.; Hughey, R.; Haussler, D.; Hubbard, T.; and Chothia, C. 1998. Sequence Comparisons Using Multiple Sequences Detect Twice as Many Remote Homologues as Pairwise Methods. *Journal of Molecular Biology* 284(4).

Pearson, W. R. 1998. Empirical Statistical Estimates for Sequence Similarity Searches. *Journal of Molecular Biology* 276(1):71–84.

Smith, T., and Waterman, M. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147:195–197.

Waterman, M., and Vingron, M. 1994. Rapid and accurate estimates of statistical significance for sequence base searches. *Proceedings of the National Academy of Sciences of the USA* 91(11):4625–4628.

Zien, A.; Zimmer, R.; and Lengauer, T. 2000. A Simple Iterative Approach to Parameter Optimization. *Journal of Computational Biology* 7(3,4):483–501.