Pacific Symposium on Biocomputing 5:140-151 (2000)

HOW UNIVERSAL ARE FOLD RECOGNITION PARAMETERS. A COMPREHENSIVE STUDY OF ALIGNMENT AND SCORING FUNCTION PARAMETERS INFLUENCE ON RECOGNITION OF DISTANT FOLDS.

KRZYSZTOF A. OLSZEWSKI
Molecular Simulations Inc., 9685 Scranton Road,
San Diego, CA 92121, USA
kato@msi.com

A choice of sequence-structure similarity scoring function parameters can significantly alter results of the performance of the recognition of distantly related folds. It therefore constitutes a critical part of fold recognition process. In order to increase an understanding of the influence of parameter choice, a comprehensive benchmark of very hard (SFOLD) and medium hard (SFAM) fold recognition examples has been derived from the SCOP database of protein structure families. These benchmarks have subsequently been used to optimize, validate and analyze dependence of recognition sensitivity on alignment and fold similarity score parameters for different scoring functions. Significant variation of the common parameters has been observed for different functions, leading to the conclusion that optimal parameter sets are not universal. The scope of solutions common to any pair of scoring function is relatively small, hence, using jury method for fold prediction seems not appropriate. Also, using a redundant version of fold libraries significantly increases odds of identification of distantly related fold.

1. Introduction

1.1 Twilight zone challenge

The number of available sequence entries in the public databases increased dramatically in recent years. Around 50% of the new sequences may have an already existing relative, but less than half can be assigned a putative structure or function based on their similarity to other, already well annotated proteins. Established procedures of sequence data mining are usually based on straightforward sequence similarity screening using e.g. the BLAST[1] or FASTA[2] programs. The major obstacle to the sequence similarity approach to identifying distantly related folds is dubbed the twilight zone[3] of protein similarity and is customarily located around 20-30% of the pairwise protein sequence identity computed for optimal alignment. Beyond the twilight zone of protein similarity, assignment based on sequence similarity become statistically marginal[4]. The new generation of sequence similarity based algorithms[5-8], based on creation and comparison of sequence family models rather that a single sequence, can vastly expand the reach of the sequence based

method pushing the twilight zone limits down. However, with the increasing gap between the number of available protein sequences and the number of structural or functional annotations of those sequences, there is a need for the algorithms that increase even more protein structure recognition sensitivity.

Recently, a subclass of hybrid threading algorithms[9-13] based on the sequence similarity augmented by structural information has been widely used to venture into the twilight zone limit. Methods using predicted secondary structure and predicted buried-exposed patterns matching structural information available from PDB[22] guide the sequence-structure alignment. Fischer and Eisenberg[9] have proposed a gain/penalty for secondary structure match/mismatch in addition to sequence similarity scoring, while Rice and Eisenberg[10] have used predicted secondary structure combined with a sequence-structure scoring matrix. Rost et al.[11] have used predicted secondary structure and per residue solvent accessibility to construct a local assignment profile. They reported a significant improvement of fold recognition performance when sequence similarity score was added to their scoring system. Recently, Hargbo and Elofsson[14] combined hidden Markov protein family models with the secondary structure scoring function and evaluated a number of various strategies for fold recognition. They reported the somewhat puzzling conclusion that using hidden Markov models did not lead to increased fold recognition performance, compared to a pairwise sequence scoring, which they attributed to a poor quality of HMMs they used.

The lower bound of the twilight zone of sequence homology detection has been estimated to be around 15%[15] based on the analysis of random distribution of pair sequence identities. However, the methods based on the predicted structural features often fail to identify relatives with even higher percent identity due to the large number of false positive solutions. When the reasoning based on the single score paradigm is inconclusive, a complementary approach can be used based on the analysis of *a posteriori* alignment scores[15]. Applications of the multiple scores paradigm has been shown to work well for difficult examples in CASP3 probe[16] and recently has led to the identification of a novel member of a importin family in *Drosophila* and human[17]. However, so far the *a posteriori* scores analysis is based on the human perception rather than being automated. Increased size of the benchmark containing difficult prediction examples will be used to construct an automated strategy of fold prediction.

1.2 Parameter choice significantly affects twilight zone performance

Sequence similarity based algorithms and their threading extensions are heavily influenced by parameter choice. There are roughly two types of parameters: those that influence the match of two residues in the alignment, and those that specify the penalty for introducing gaps into the alignment. For local alignments, high gap

penalty parameters lead to locally contiguous segments of alignments and the logarithmic dependence of the total score on the sequence length[18], while for low gap penalties the total alignment score depends linearly on the sequence length. An important link between these two parameter types is that the above linear-logarithm phase transition occurs for the scoring systems for which an expected global alignment score is negative[18]. A number of attempts have been made to understand the influence of the gap penalties on alignment, to design a biologically relevant parameter systems and detect biologically relevant alignments, for the cases when similarity of two sequence is apparent[19].

Vogt et al.[4] have demonstrated that the choice of gap penalties significantly alters the performance of the different sequence substitution matrices. However, they observed that after the proper optimization of parameters, there is a negligible difference between the recognition performance. Therefore, the Gonnet et al.[20] proposed matrix is used exclusively throughout the paper. Vogt et al also observed a noticeable difference between the performance of the global and local algorithm. A version of the alignment algorithm used in this paper, combines the local alignment threshold with the algorithmic incentive to increase overlap of the reference structure. One of the optimized parameters gradually introduces global features of the alignment.

The primary goal for the benchmark construction is an investigation of the limits of prediction performance for different variants of structure-enhanced sequence scores. However, the pair of benchmarks constructed in this work can be used not only for the optimization and assessment of the performance of single score based approaches but can establish a clear reference point for further algorithmic enhancements and allow the development of recognition strategies based on the multiple scores paradigm.

2. Methodology

2.1 SFOLD and SFAM benchmarks construction

Two benchmarks have been derived using the 1.37 version of the SCOP database[21]. For the purpose of benchmarks creation, all redundant structures corresponding to the same sequence have been removed, leaving total of 2175 different sequences with one representative structure. These sequences were further screened, and short sequence and those without a publicly available PDB[22] record were left out. For the SFAM level of the benchmark, a list of all pairs of sequences that belong to the same superfamily but to different families was created. There is a total of 19317 family pairs, while there are 456 designated pairs in the SFAM benchmark after one representative sequence have been chosen at random from each family. A collection

of designated pairs within a superfamily will be included in the SFAM level benchmark. Note that the designation of sequence-structure pairs (rather than requiring a match of any family representative) makes this benchmark more difficult than in reality, mimicking a lack of family information in the fold library creation and therefore estimating a lower bound of the scoring system performance. Alternatively, identifying any matching sequence from a designated family may also be perceived as an (optimistic) measure of success, hence the corresponding true positive number is also reported. For the SFOLD level benchmark, sequence pairs that correspond to the same fold class (as defined by SCOP) and to different superfamilies were taken into account.

The total number of sequences used for both benchmarks is 695. While most of sequence will form both SFOLD and SFAM pairs, some of sequences can form pair only within a single layer of the benchmark. There are 11841 superfamily pairs within a SFOLD benchmark and there are 1176 fold pairs remaining after picking a representative for a family rather than constructing all possible matches. The granularity of SFAM and SFOLD benchmarks can be roughly characterized by the cardinality of SCOP family and superfamily clusters. Table 1 reports a cardinality of family and superfamily clusters for both benchmarks and training and testing subsets of both benchmarks.

Table 1. The "granularity" of SFAM and SFOLD benchmarks. Numbers in the Table report a count of single, double, triple or bigger families (superfamilies) included in SFAM (SFOLD) benchmark, respectively. Whole benchmark, as well as, training and testing sets are reported.

	Single	Double	Triple	4 or more
SFAM	378	130	51	187
SFOLD	414	57	33	224
SFAM (train)	50	20	0	15
SFOLD (train)	0	8	0	77
SFAM (test)	124	42	12	42
SFOLD (test)	138	21	10	49

2.2 profile vs. sequence scoring function

The first type of scoring function tested corresponded closely to the function proposed by Fischer and Eisenberg[9] and implemented as SeqFold version 1.0[23]. In this function, the straightforward sequence similarity (measured by sequence similarity log-odds matrix) of the target (query) sequence to the set of reference structures is augmented by the per residue gain/penalty for secondary structure

match/mismatch. Besides the relative weights of sequence and secondary structure contribution, scoring function parameters may include, e.g., deriving and scaling of secondary structure prediction confidence, sequence similarity balance, etc. In this work, only the secondary structure weight parameter has been allowed to vary during the training procedure. The standard measures of secondary structure prediction quality (e.g. Q3) do not correlate well with the fold prediction performance. Assigning secondary structure confidence was therefore, for the purpose of this work, avoided by using an actual secondary structure, rather than simulating prediction, in order to derive parameterization independent of the secondary structure prediction algorithm.

The new version of SeqFold[24] contains second type of sequence profiles based scoring function variants that are capable of matching target sequence with reference sequence profiles (seq-pro), target sequence profile with the reference sequences (pro-seq) and target sequence profile with reference sequences profiles (pro-pro). Two strategies of profile generation approaches using a PSI-BLAST[8] were tested: selective strategy and full (permissive) strategy. In selective strategy, a conservative e-value cutoff for sequence profile generation (10^{-7}) was used and no PSI-BLAST iterations were performed. This strategy guaranties that no false positives are included in the sequence profile, but many distantly related sequences are missing as well. For the full strategy, the default PSI-BLAST cutoff (10^{-3}) for profile inclusion and 3 iterations were performed, leading to the inclusion of many distantly related sequences but also many false positives. In this work, only the results of optimization using selective profiles are presented.

2.2 Parameters optimization

For all sequence structure scoring strategies, the Monte Carlo optimization of the fold recognition objective function has been performed in the parameter space. A few strategies of the objective function derivation have been tested. Throughout this work, the objective function is defined as the sum of all possible true positive hits for the SFAM and SFOLD benchmarks. Note the difference between the stricter benchmark definition of the true positive using designated pair and the more forgiving definition in the objective function. Since all possible true positives are counted during training, it may happen that the benchmark-designated pair may be omitted, even in case of clear hit from the different member of the same family. For different parameter combinations, an expected similarity score may oscillate around zero, switching between the linear and the logarithmic modes. The chosen true positives counting approach has been used to diminish the roughness of the optimized function and decrease the chaotic behavior near the logarithmic-linear transition point. The number of possible true positive hits in the SFAM benchmark is smaller than in the SFOLD benchmark, on the other hand the SFOLD benchmark

pairs are, in principle, more difficult to identify. No attempt has been made to normalize those effects.

In each Monte Carlo iteration, for each sequence from the sequence-structure pair from the training set, a Seqfold run with the current iteration parameters has been performed. Then, all trivial true positives have been removed from the hit list (trivial true positives are defined as the same family hits in the SFAM benchmark and same superfamily hits in the SFOLD benchmark). All the remaining true positives (all hits before the first occurrence of the true negative) are counted and included in the objective function. The basic training set contained 85 sequence-structure pairs and has been hand picked to contain the following whole fold categories: four helix bundles (001-023), Armadillo repeats (001-084), long helices oligomers (001-097), immunoglobulins (002-001), cytokines (002-028), beta-propellers (002-047), TIM-barrels (003-001), flavodoxin (003-013) and thioredoxin (003-033). Two whole SCOP classes (alpha and beta, and multi-domain) were not included in the training set to provide an additional memorization check. The complete list of sequences, family and superfamily pairs is available from the author upon request.

The alignments of the target sequence with the reference structure using the above defined scoring functions have been performed using version 2.0 of Seqfold[24]. There are three parameters that influence the sequence-structure alignment: a gap opening penalty, a gap extension penalty and a terminal gap penalty. In principle, it is possible to treat sequence and structure gaps separately. However, in this work, gap parameters for sequence and structure have been symmetrized to decrease the number of free parameters and effects of memorization. Also, the sequence terminal gap penalty has been kept zero at all time. Therefore, there was total of four modified parameters in the objective function: common gap open and gap extension penalties, terminal structure gap penalty and the weight of the secondary structure penalty.

3. Results

The outline of the Monte Carlo optimization strategy included an initial survey of the parameter space performed with a small subset of the training set. Then, a set of five starting points was chosen randomly and then the whole training set was used to perform five independent scans consisting of around 300 iterations. The majority of optimal solutions of all optimization runs did form a rough cluster in the common gaps slice of the parameter space, see Fig. 1. The notable exception is a profile to profile scoring function that exhibits two distinctive basins of suboptimal parameters.

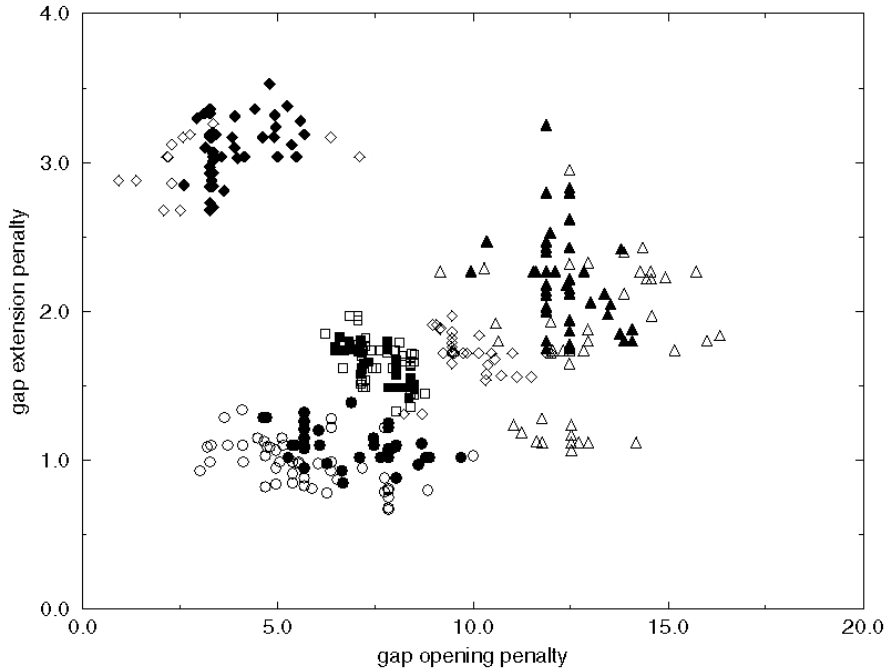
The optimal set of parameters has been chosen for each scoring function, see Table 2, corresponding to the maximum of the objective function and, if degenerated, to the highest number of SFOLD benchmark true positives. Additionally, a representative set of parameters for the suboptimal region of profile-profile scoring function and an unoptimized set of original sequence-sequence scoring parameters has been used to establish a reference point and evaluate the performance of the whole benchmark. The results for the SFOLD and the SFAM benchmark for optimal sets of parameters are reported in Table 3. Both distinctive (matchin of the distinctive pair only) and permissive (matching of the distinctive family) variants of counting of benchmark true positives are reported. Besides a number of true positives (target of the objective function optimization) false close results are reported. False close hits are defined as correct hits with less that 20 rank proceeded by at least one true negative hit. The rationale behind counting false close hits is that the high number of false close solutions is essential for a performance of post-processing of results that rely on using a posteriori alignments features. The number of false close solutions was not included in the objective function. Perhaps the most striking results of the performed optimizations is that in all cases the false close hits number did increase. In distinctive pairs case, false close count increase is roughly proportional to the number of true positive hits in both benchmarks.

Table2. Optimal parameters encountered during Monte Carlo scan of the parameter space. (i) unoptimized set of original parameters, (ii) representative set of suboptimal solutions for the pro-pro scoring

	seq-seq (i)	seq-seq	seq-pro	pro-seq	pro-pro (ii)	pro-pro
gap open	10.8	11.8	8	7.8	8.7	3.3
gap extend	0.6	2.3	1.5	1	1.3	3.2
gap ref. Terminal	0.6	1.7	2	2.2	0.9	0.64
sec. str. Weigth	1	1.6	1.43	1.4	1.6	1.6

It is also apparent from the Table 3 that the unoptimized set of parameters for sequence-sequence scoring performs fairly poorly, when compared to all other optimized scoring functions. Since the training set was composed excluding two SCOP classes: alpha and beta and multidomain, it is possible to treat the subset of SFOLD and SFAM benchmarks as a separate test benchmark. There are 208 sequences in this benchmark and 658 of distinctive superfamily pairs in the SFOLD part of the test benchmark and 194 family pairs in the SFAM part of the test benchmark. The fold library used still contains the whole set of SCOP sequences, therefore the relative number of true negatives is larger for the test benchmark than for the complete benchmark. The results with optimized parameters for test subset of both benchmarks are presented in Table 4.

Fig. 1. Clustering of the best (filled symbols) and suboptimal solutions (open symbols) for all scoring functions: seq-seq (triangle up), seq-pro (square), pro-seq (circle) and pro-pro (diamond).



With the exception of the sequence to profile scoring strategy, there is a significant improvement in the performance of the test benchmark. Overall, the increase in the test set performance is well correlated with the training set performance, even though both sets consists of completely different classes of protein folds. Therefore, possible memorization effects of the training set are less likely.

Both SFOLD and SFAM benchmark are fairly difficult for all sequence-structure scoring strategies, therefore, the question of mutual consistency of different scoring strategies arises. In Table 5, overlaps of the true positive pairs identified for optimal set of parameters for different scoring strategies are reported. It is apparent that sequence-sequence and profile-sequence pair overlap is relatively high, as well as sequence-profile and profile-profile overlap leading to the conclusion that the quality of the reference fold library definition distinguishes methods to a larger extent than the definition of the target sequence. Still, the overlap of any two

Table 3. The total number of true positives and close false negatives for complete benchmarks for the optimized (Table 2.) set of parameters for all scoring function variants. (i) unoptimized set of parameters for seq-seq scoring, (ii) representative suboptimal parameter set for pro-pro scoring. dist refers to a distinctive pairs count and all to all possible true positive hits within a family.

	seq-seq (i) SFOLD dist	seq-seq (i) SFOLD all	seq-seq SFOLD dist	seq-seq SFOLD all	seq-pro SFOLD dist	Seq-pro SFOLD all	pro-seq SFOLD dist	pro-seq SFOLD all	pro-pro (ii) SFOLD dist	pro-pro (ii) SFOLD all	pro-pro SFOLD dist	pro-pro SFOLD all
true positive	16	31	40	165	34	288	61	175	39	202	43	249
false close	86	178	169	281	80	176	187	303	152	252	127	231
false negative	2216	2126	2143	2021	2238	2132	2104	1978	2161	2048	2182	2066
true pos %	0.7	1.2	1.7	6.4	1.4	11.1	2.6	6.7	1.7	7.8	1.8	9.6
false close %	3.7	6.9	7.2	10.8	3.4	6.8	8.0	11.7	6.5	9.7	5.4	8.9
false neg. %	95.6	91.9	91.1	82.8	95.2	82.1	89.5	81.6	91.9	82.5	92.8	81.5

	seq-seq (i) SFAM dist	seq-seq (i) SFAM all	seq-seq SFAM dist	seq-seq SFAM all	seq-pro SFAM dist	seq-pro SFAM all	pro-seq SFAM dist	pro-seq SFAM all	pro-pro (ii) SFAM dist	pro-pro (ii) SFAM all	pro-pro SFAM dist	pro-pro SFAM all
true positive	98	345	134	461	31	202	158	471	77	292	61	245
false close	81	156	110	190	58	106	107	178	120	168	103	151
false negative	733	649	668	593	823	770	647	587	715	660	748	692
true pos %	10.7	27.7	14.7	37.1	3.4	16.2	17.3	37.9	8.4	23.5	6.7	19.7
false close %	8.9	12.5	12.1	15.3	6.4	8.5	11.7	14.3	13.2	13.5	11.3	12.1
false neg. %	80.4	59.7	73.2	47.7	90.2	75.2	70.9	47.8	78.4	63.0	82.0	68.2

Table 4. The number of true positives and close false negatives for the optimized (Table 2.) set of parameters for all scoring function variants. (i) unoptimized set of parameters, (ii) representative suboptimal parameter set. dist refers to a distinctive pairs counting and all to counting all possible hits within a family.

	seq-seq (i)	seq-seq	seq-pro	pro-seq	pro-pro (ii)	pro-pro
SFOLD dist						
true positive	0	13	6	21	15	14
false close	9	57	17	66	52	44
true pos. %	0.0	2.0	0.9	3.2	2.3	2.1
false close %	1.4	8.7	2.6	10.0	7.9	6.7
SFOLD all						
true positive	1	16	6	26	16	15
false close	17	72	22	80	63	54
true pos. %	0.2	2.4	0.9	3.9	2.4	2.3
false close %	2.6	10.9	3.3	12.1	9.5	8.2
SFAM dist						
true positive	15	28	6	35	14	11
false close	20	26	7	22	30	25
true pos. %	7.7	14.4	3.1	18.0	7.2	5.7
false close %	10.3	13.4	3.6	11.3	15.5	12.9
SFAM all						
true positive	50	75	25	87	45	35
false close	26	40	14	32	37	34
true pos. %	20.0	30.0	10.0	34.8	18.0	14.0
false close %	10.4	16.0	5.6	12.8	14.8	13.6

methods rarely exceeds 50% and that leads to the conclusion that there is no apparent optimal strategy in the twilight zone limit.

A lack of significant overlap between alternative scoring strategies suggests also that certain fold pair combinations are more amenable for certain scoring system than for others. This may explain an apparent lack of performance increase reported by Hargbo and Elofsson[14] upon inclusion of family information in the scoring function. Similarly, in a recent CASP3 experiment for fold recognition targets, we used exclusively seq-seq scoring function and noted a significant difference between performance for fold and superfamily targets[15].

Table 5. SFOLD true positives overlap between different scoring strategies. Upper triangle includes all true pairs, lower triangle includes only designated pairs.

	seq-seq	seq-pro	pro-seq	pro-pro
seq-seq		47 (16-28)	78 (45-47)	61 (30-37)
seq-pro	6 (15-18)		29 (10-17)	116 (40-57)
pro-seq	28 (46-70)	6 (10-18)		62 (31-35)
pro-pro	11 (28)	17 (44-50)	19 (31-49)	

Further extensions of theoretical limits of recognition performance were explored by an attempt to optimize parameters within an extended SCOP fold class. We tested immunoglobulin, TIM barrels and ferrodoxin fold classes consisting of many distant superfamilies and families. Parameters optimization within a single fold, led to an apparent significant increase of recognition performance (in some cases double). However, fold derived parameters are significantly different from the parameters derived using a more diverse training set. Although fold derived parameters are of little use in practice, the above results demonstrate that the training set derived parameters are optimal only when we lack *a priori* knowledge about the protein sequence.

4. Conclusions

Two fold recognition benchmarks with varied degrees of difficulty were created. An extensive optimization of different scoring systems for distant protein fold recognition based on the two benchmarks has been performed. Optimal parameters for each scoring strategy have been derived and validated using large and diverse training and test subsets of derived SFOLD and SFAM benchmarks. New parameters not only increase significantly the performance of single score based methods (increased number of true positives) but are suitable for the derivation of an automated multiple, *a posteriori* scores based method. Note, however, that since the actual rather that predicted secondary structure was used, the results presents an upper limit of secondary structure based prediction performance. Also, no attempt was made to assess quality of resulting alignments.

For optimal parameter choice, there is little consistency between different scoring systems. Alignment parameters are not universal and have to be derived for each strategy independently. Additionally, every set of derived parameters can be significantly improved for use within a superfamily or fold class. This leads to an important question, rather than a conclusion: How can the results obtained for two different scoring systems or parameter choices be compared if consistency is not to be expected.

A significant difference between distinctive and permissive counting demonstrated that the redundant fold library should be used to increase the odds of identification of distantly related folds. Also, it suggests that strategies based on intermediate sequences are particularly suitable for distant fold recognition.

Acknowledgments

The author gratefully acknowledges Silicon Graphics assistance in providing a necessary computer power for completing this work.

References

1. S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, *J. Mol. Biol*. **215**, 403 (1990)
2. W.R. Pearson, *Proc. Natl. Acad. Sci USA* **85**, 2444 (1988)
3. R. Doolittle, *Science*, **214**, 149 (1981)
4. G. Vogt T. Etzold and P. Argos, *J.Mol.Biol*., **249**, 816 91995)
5. A. Krogh, M. Brown, I.S. Mian, K. Sjolander and D. Haussler, *J. Mol. Biol*. **235**, 1501 (1994)
6. S.R. Eddy, G. Mitchison and R. Durbin, *J. Comput. Biol.* **2**, 9 (1995)
7. S.R. Eddy, *Curr. Op. Struct, Biol.* **6**, 361 (1996)
8. S.F. Altschul, T. Madden, A Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lippman, *Nucleic Acid Res*. **25**, 3389 (1997)
9. D. Fischer and D. Eisenberg, *Protein Sci.*, **5**, 947 (1996)
10. D. Rice and D. Eisenberg, *J. Mol. Biol*. **267**, 1026 (1997)
11. Rost B., Schneider R. and Sander C. (1997) J. Mol. Biol. **270**, 471
12. L. Rychlewski, B. Zhang and A. Godzik, *Folding&Design* **3**, 229 (1998)
13. K.A. Olszewski, L. Yan, D.J. Edwards, *Theor.Chem.Acc*. **111**, 57 (1999)
14. J. Hargbo and A. Elofsson, *Proteins Struct. Func. Genet*. **36**, 68 (1999)
15. K.A. Olszewski, "Using alignment derived posterior scores increases sensitivity of fold prediction", in preparation
16. K.A. Olszewski, Yan L. Computers & Chemistry, "From fold recognition to homology modeling. An analysis of protein modeling challenges at different levels of prediction complexity", submitted
17. M. Pál, M. Mink, O. Komonyi, P. Deák, K.A. Olszewski and P. Maróy, "A novel member of β-catenin superfamily in *Drosophila* and Human", in preparation
18. M.S. Waterman, L. Gordon and R. Arratia, Proc. Natl. Acad. Sci. USA, 84, 1239 (1987)
19. see e.g. M. Vingron and M.S. Waterman, *J. Mol. Biol*., **235**, 1 (1994)
20. G.H. Gonnet, M.A. Cohen and S.A. Benner, *Science*, **256**, 1433 (1992)
21. A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chotia, *J. Mol. Biol*. **247**, 536 (1995)
22. F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.E. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol*. **112**, 535 (1977)
23. Seqfold (1998) SeqFold release 1.0, Molecular Simulations Inc.
24. Seqfold (1999) SeqFold release 2.0, Molecular Simulations Inc.
25. L. Yan, Z-Y. Zhu, A. Badretdinov, K.A. Olszewski, D. Kitson and M. Pear, in preparation