Design of the Comprehensive Fold Recognition Benchmark. Application to SeqFold, Training and Validation

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

Keywords: fold recognition benchmark, SeqFold

1 Introduction

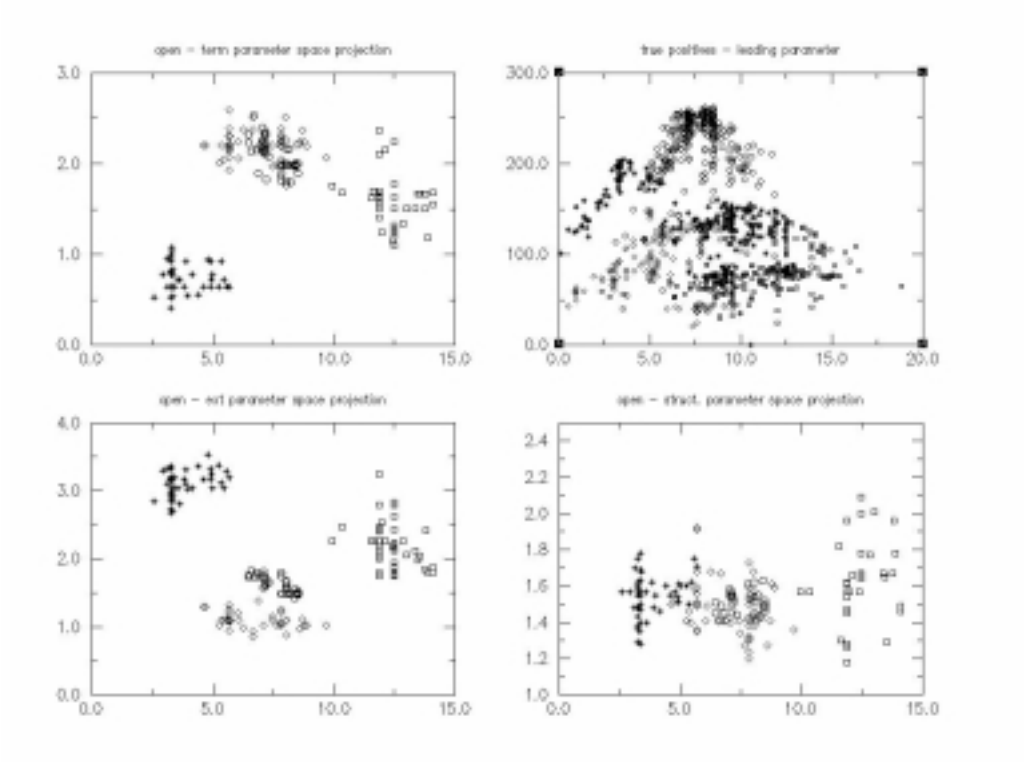
Recent exponential increase of protein sequences creates a challenge for automated annotation methods. When sequence based methods (e.g. PSIBLAST [1]) fail to identify a possible homologue (generally below 25% of protein identity i.e. within so-called twilight zone), fold recognition methods offers additional sensitivity [2,4,5,8]. However, training, validating and comparing fold recognition performance proves to be difficult. This work presents a comprehensive attempt to design a universal, two-tier, balanced fold recognition benchmark that can be used to perform all above mentioned tasks. Also, results of training and comparison of different fold recognition scoring strategies are discussed.

2 Method and Results

A benchmark has been derived using 1.37 version of SCOP [3] library with 2175 representative sequences. Two tiers with varying degree of difficulty were used by walking up the SCOP tree and discarding down in the tree hits as trivial. There are total 19317 distinctive family pairs and 111841 superfamily pairs in the benchmarks. Additional degree of difficulty was imposed by selecting specific structural pairs that must match, which reduces number of possible pairs to 456 and 1176 respectively [6]. This simulates single sequence families that are generally more difficult to predict.

Seqfold training for four possible sequence and profiles matching was performed using a self-contained subset of the benchmark in order to test for memorization effects. Four important scoring function and alignment parameters were optimized during an extensive Monte Carlo search in the parameter space. The results indicate that there is little consistency in between different scoring functions and that optimal sensitivity parameters must be designed for each scoring function variant independently. Additionally, optimal parameter set significantly increases number of near missed false negatives. This indicates that additional a posteriori scores may be used to manually analyze fold recognition results in the cases when automated assignment fails as was demonstrated for CASP3 targets [7]. The newly derived optimal parameters tested on the new Drosophila gene helped establish a novel functional link consistent with knockout phenotype [9].

Summary of SeqFold training results: a-c) parameter space projection for most sensitive sets of parameters d) parameter sensitivity correlation. Squares correspond to sequence-sequence scoring, circles to sequence profile scoring, diamonds to profile-sequence scoring and stars to profile-profile scoring functions.



References

[1] Altschul, S.F., Madden, T., Schaffer, A., Zhang J., Zhang Z., Miller W. and Lippman D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acid Res.*, 25:3389–3402, 1997.

[2] Fischer, D. and Eisenberg, D., Protein fold recognition using sequence-derived predictions, *Protein Sci.*, 5:947–955, 1996.

[3] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chotia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247(4):536–540, 1995.

[4] Olszewski, K.A., Ab initio folding. Is conformational searching enough, *Pol. J. Chem.*, 72:1667–1679, 1998.

[5] Olszewski, K.A, Yan, L., and Edwards, D.J., *Theor. Chem. Acc.*, 111:57, 1999.

[6] Olszewski, K.A., How universal are fold recognition parameters. A comprehensive study of alignment and scoring function parameters influence on recognition of distant folds, *Proc. Pacific Symp. Biocomputing 2000*, World Scientific, 143–154, 2000.

[7] Olszewski, K.A., Yan, L., Edwards, D.J., and Yeh, T., From fold recognition to homology modeling. An analysis of protein modeling challenges at different levels of prediction complexity, *Computers and Chemistry*, 2000, in press.

[8] Rost, B., Schneider, R. and Sander, C., Protein fold recognition by prediction-based threading, *J. Mol. Biol.*, 270:471–480, 1997.

[9] Pál, M., Mink, M., Komonyi, O., Deák, P., Olszewski, K.A. and Maróy, P., A novel member of β-catenin superfamily in *Drosophila* and Human, in prep.