Improving the Detection of Protein Remote Homologues Using Shannon Entropy Information

E. Capriotti(1,3) , P. Fariselli (2), I. Rossi (3,4), and R. Casadio(2)

(1) Department of Physics, University of Biologna, via Irnerio, Bologna, Italy
emidio@biocomp.unibo.it

(2) Department of Biology, University of Bologna, via Selmi, Bologna, Italy

(3) CIRB, University of Bologna, via Irnerio, Bologna Italy

(4) BioDec S.r.l., via Fanin, Bologna, Italy

Keywords. Fold Recognition, Shannon Entropy Information, Sequence Alignment

We analyze the quality of the alignment generated by the profile-profile alignment comparison algorithm known as BASIC [1] and compare the results with those obtained with a structural alignment code. By this we compute that a Shannon entropy value > 0.5 gives a sequence to sequence alignment of the target/template couple comparable to that obtained with the structural alignment performed with CE.
In our fold recognition/threading code *Tangram,* the BASIC profile-profile alignment is implemented as follows:

1. The composition profiles P_A and P_B for the target and template are generated by multiple alignment of the sequences obtained from a three-iteration PSI-BLAST [2] search on the Non-Redundant database (the inclusion threshold is $E=10^{-3}$).
2. the dot matrix (D) for the profile comparison of two protein sequences
$D= P^T_A\, S\, P_B$, (*with* S=BLOSUM62 [3] substitution matrix) is computed using linear algebra routines.
3. the D matrix is searched for high-scoring alignment by means local Smith-Waterman dynamic programming algorithm [4].

*T*he test set used for the evaluation is composed by 185 template/target couples of PDB structures that share the same SCOP label, but have less than 30% sequence identity
When the top-scoring alignments for each target protein in the test set is considered, our BASIC implementation detects the full SCOP label for 125 couples (68%) and generates 114 (62%) alignments with a MaxSub [5] score >=1.
Interestingly, it is found that nearly all of the high-quality alignments share a common feature: the average Shannon entropy for the profile sections aligned together is greater than 0.5 for both the template and the target.
If only the top scoring alignments for which this condition holds are considered, a subset of 119 alignments is selected, and for 116 of them (97%) the full SCOP label can be assigned to the target (see Fig. 1), while 108 (91%) gets a nonzero MaxSub score (see Fig. 2), with an average score of 4.6 MaxSub on the subset
On the same 119 couples, the structural alignment program CE [6] computes a nonzero MaxSub score for 116 of them, with an average of 5.7 points.
These results indicate that the Shannon entropy value can be used to discriminate a subset of sequence profile-profile alignments of quality comparable to that obtained by means of a structural alignment program[7].

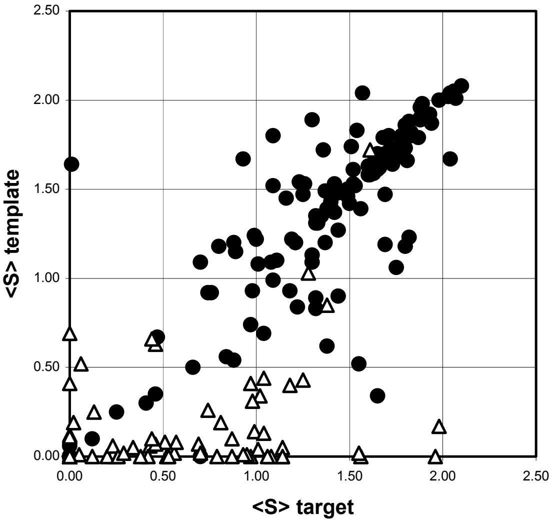

Fig. 1. Template versus target average Shannon entropy. Filled circles represent the correct SCOP fold labels, while white triangles are erroneous assignments.

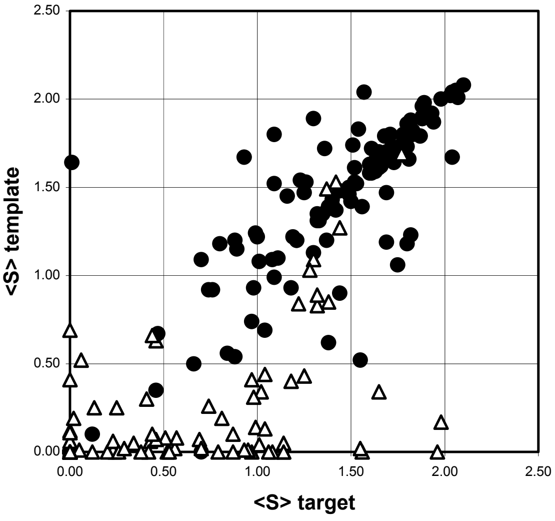

Fig. 2 Template versus target average Shannon entropy. Filled circles and white triangles represent alignments on which MaxSub score is positive or zero, respectively.

[1] Rychlewski J. et al. (1998) Fold and function predictions for Mycoplasma genitalium proteins. *Fold. Des.* ***3,*** 229-238

[2] Altschul S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* ***25*** *(17),* 3389-3402

[3] Henikoff, S. et al. (1998). Superior performance in protein homology detection with the BLOCKS database server. *Nucleic Acids Res.* ***26,*** 309-312.

[4] Smith T. S. and Waterman M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* ***147,*** 145-147

[5] Siew N. et al (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* ***16(9)*** 776-785.

[6] Shindyalov I. N. and Bourne P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path *Prot. Eng.* ***11(9)*** 739-747

[7] Capriotti E, Fariselli P, Rossi I, Casadio R. (2004)A Shannon entropy-based filter detects high- quality profile-profile alignments in searches for remote homologues. *Proteins*, **54(2)** 351-360.