BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–4

Prediction of Protein Functional Residues from Sequence by Probability Density Estimation: Supplementary material

J. D. Fischer, C. E. Mayer,and J. Söding

Department for Protein Evolution, Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany

1 BENCHMARK SETS

The difficulty in defining a benchmark set for the prediction of ligand-binding residues is to distinguish the physiologic ligands in crystal structures from the numerous molecules that are added to the buffers in the search for good crystallization conditions. Capra and Singh solve this problem by mapping the protein chains to EC numbers and accepting only ligands that are likely to be catalyzed by the corresponding reaction. Ligand-binding residues are then defined using a standard $4\mathring{A}$ distance cut-off between atoms. This EC benchmark set is one of the largest and most diverse reported to date (Table 1).

The Catalytic Site Atlas (Porter *et al.*, 2004) is a large database listing catalytic residues for enzymes of known 3D structure. To construct our CSA set, we take all 6083 entries in version 2.2.2 that are generated directly from the literature. Many of these sequences will contain more than one domain. This can be problematic since often only one of several domains in a protein chain will have its catalytic residues annotated. The other domains are then quite likely to contain unannotated catalytic residues that would wrongly be considered as false positives. In order to obtain single domains, we map the structural domains with the annotated residues to the corresponding domains from the SCOP database (Murzin *et al.*, 1995).

To avoid oversampling of particular protein families, one domain from each SCOP family is selected. Since we would like to define not only catalytic residues but also ligand-binding residues, we pick only those SCOP domains which have a ligand in contact with at least one annotated residue. Contact is defined by a $4\mathring{A}$ distance cut-off. Note that the ligands are validated, i.e. distinguished from co-crystallized buffer molecules, by being in contact with the catalytic residues. We exclude water and sulfate ions as ligands. In the few cases, where more than one ligand is in contact with the annotated residues, we chose both ligands. We are then left with 428 domain sequences.

For generating the MSAs, we run PSI-BLAST on each of the sequences with the non-redundant sequence database from NCBI filtered at 90% sequence identity by CD-HIT (Li and Godzik, 2006). Only sequences covering at least 80% of the query sequence residues are accepted into the evolving MSA, ensuring that only few columns are highly gapped. The PSI-BLAST search stops when 500 or more sequences have been found or after a maximum of ten iterations. After the PSI-BLAST search, we use `hhfilter` from the HHsearch package (Söding, 2005) to remove sequences with less than 0.25 bits score per column with the query sequence.

Table 1. Overview of the benchmark sets used in this study. The CSA set uses two definitions of true positive residues: original CSA-annotated, (CSA-cat) and ligand-binding (CSA-ligand). The diversity is measured by the average number of different amino acids per column.

	Proteins	SCOP families	Positive residues	Negative residues	Alignment diversity
CSA-cat	423	423	1,536	107,463	11 ± 4
CSA-ligand			5,331	103,668	
SITE-ligand	711	711	9,547	142,628	11 ± 4
EC-ligand	828	348	16,166	273,718	7 ± 3

This step guarantees that the sequences in the alignment are not too distantly related to the query sequence because functionally important residues need not be conserved in very distant relatives. (In a preliminary analysis on a smaller test set, we optimized the filter and found a very flat maximum for all methods at a threshold around 0.25-bits-per-column.) Finally, we eliminate alignments with only one or two sequences leaving us with 423 out of 428 alignments.

The SITE set is generated in an analogous fashion, using the PDB SITE records instead of the Catalytic Site Atlas for the validation of physiological ligands. We have constructed 726 alignments by PSI-BLAST, 711 of which have more than two sequences after the 0.25-bits-per-column filter.

The functional residues are defined in two alternative ways. First, to test the prediction of catalytic residues, the original CSA-annotated residues are defined as positives and all other residues are defined as negatives. This set is named "CSA-catalytic" (see table). Second, to test the prediction of ligand-binding residues, we define all residues in contact ($4\mathring{A}$) with the validated ligands as positives and all others as negatives. These sets are named "CSA-ligand" and "SITE-ligand". The table shows that all three data sets are fairly large, and the CSA and SITE sets are the most diverse and evenly sampled, containing just one member per SCOP family. Note that the alignment diversity is much higher in the CSA and SITE set than in the EC set.

2 PROFILE GENERATION

Following the work of Pei and Grishin (2001), we tested three schemes to build sequence profiles from MSAs: "unweighted",

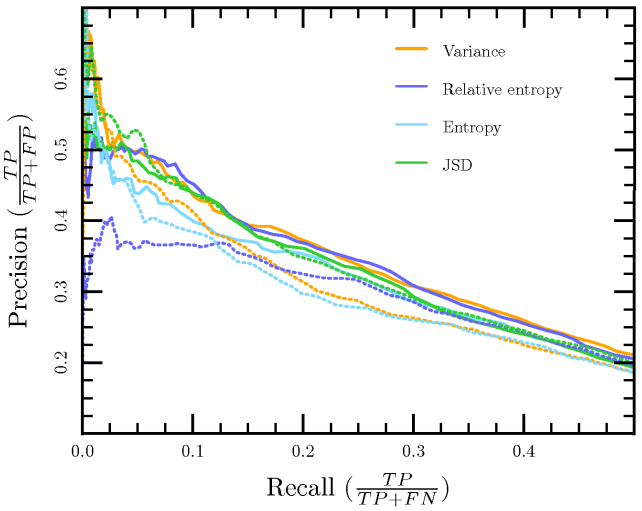

Fig. 1. ROC plot comparison of two profile building schemes on the CSA-ligand benchmark data. The solid traces show the conservation score for profiles built with independent counts, the broken traces refer to the Henikoff weighted scheme. The unweighteed case is not shown since performance is worse than for the other schemes in all methods.

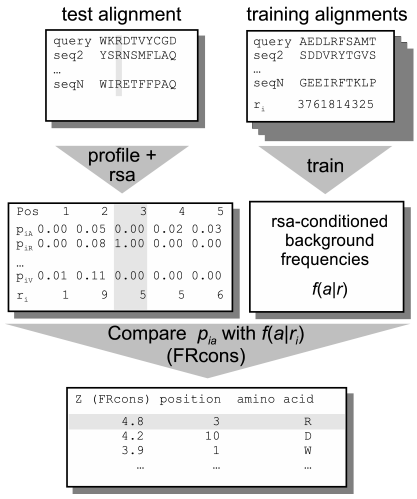

Fig. 2. Procedure for predicting the conservation with FRcons using trained amino acid background frequencies conditioned on the predicted relative solvent accessibility (rsa) state ($r_i \in \{0, \ldots, 9\}$). The conditioning on predicted secondary structure (ss) is omitted for simplicity.

"weighted" and "independent counts". In the unweighted scheme, no sequence weights are used to calculate the amino acid frequencies, whereas in the weighted scheme, Henikoff sequence weights are employed (Durbin *et al.*, 1998; Henikoff and Henikoff, 1994). In the independent count scheme (Sunyaev *et al.*, 1999; Pei and Grishin, 2001), the frequency p_{ia} of amino acid a at position i in the MSA is proportional to the effective number of sequences $N_{\text{eff}}(i,a)$ in the sub-aligment $\text{MSA}(i,a)$ composed of those sequences that have amino acid a at position i. $N_{\text{eff}}(i,a)$ is derived from the average number $N_{\text{aa}}(i,a)$ of different amino acids per column in $\text{MSA}(i,a)$: $N_{\text{eff}}(i,a) = \log(1 - N_{\text{aa}}(i,a)/20)/\log(1 - 1/20)$. Normalization of $N_{\text{eff}}(i,a)$ over the 20 amino acids yields the profile frequencies p_{ia}. In comparison to the sequence weighting scheme, the independent count scheme weights amino acids that are rare in a particular MSA column higher, effectively pushing the amino acid composition more towards an equi-frequency distribution.

We have tested all benchmarked methods with all three profile building schemes (see Fig. S1) (except Rate4Site which takes alignments as input) and picked the best scheme for each method. All methods except Jensen-Shannon Divergence performed best with independent counts. The latter was slightly better with the Henikoff-weighted scheme, which was also employed in the original work (Capra and Singh, 2007).

Except for the FRcons method, no pseudocounts are added to the profiles because our tests have shown that pseudocounts do not improve the performance of the methods once the scores are normalized (Fig. 4 middle).

3 FRCONS METHOD

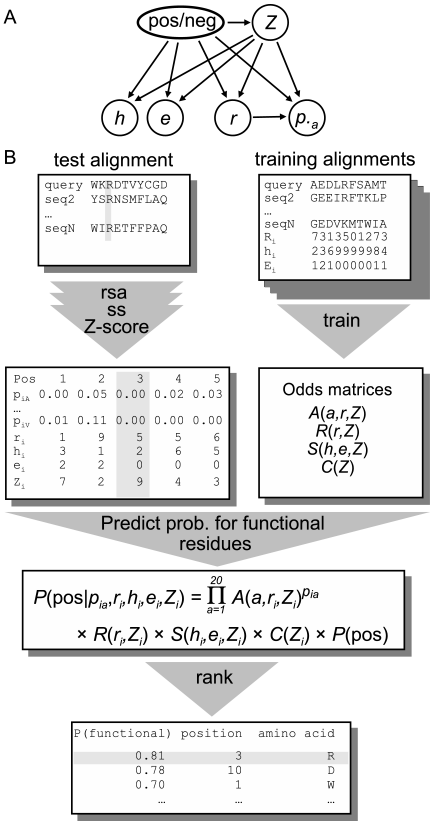

Fig. 3. (A) Bayesian network for modeling the likelihood in eq. (11). (B) Predicting the probability for a functional residue by density estimation. In the training step, the matrices A, R, S, and C are estimated from the training alignments with predicted rsa, ss, and FRcons score. In the prediction step, the conditional probability for each query residue to be functional is estimated, given its amino acid distribution, predicted rsa, ss, and FRcons values.

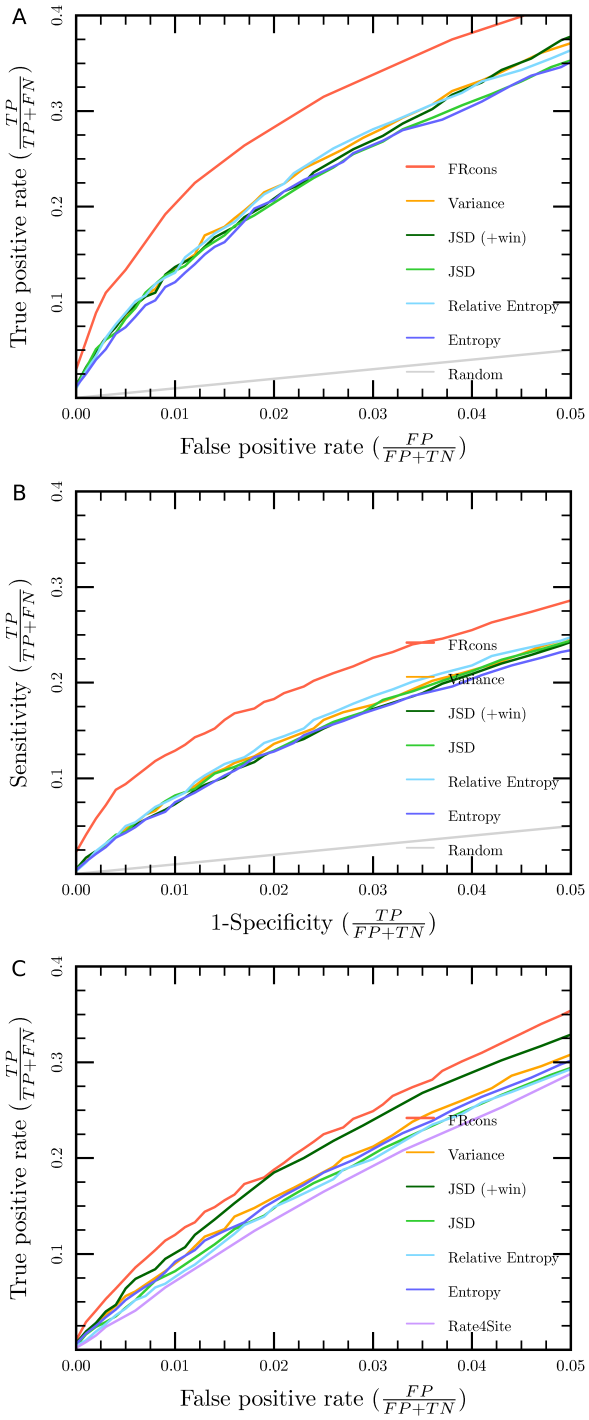

Fig. 4. True positive rate vs. false positive rate for the prediction of ligand-binding residues on three sets: (A) CSA, (B) SITE and (C) EC.

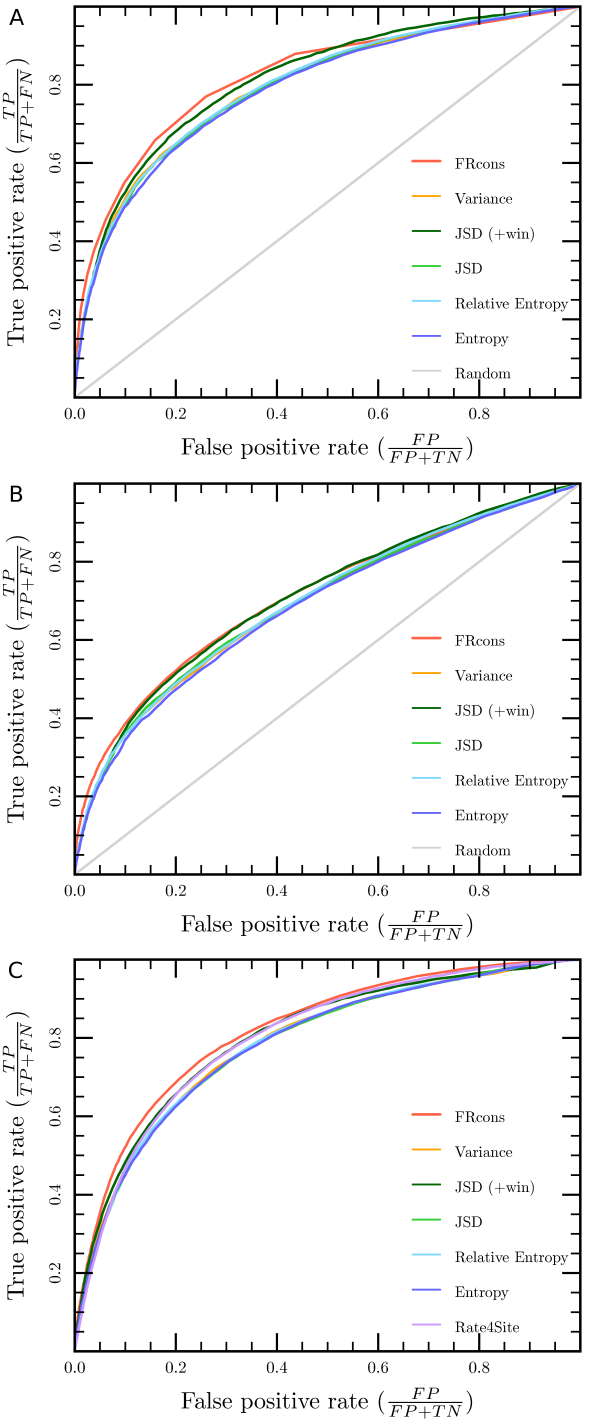

Fig. 5. Same as previous figure with false positive rate up to 1.0: (A) CSA, (B) SITE and (C) EC.

REFERENCES

Capra, J. A. and Singh, M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.

Henikoff, S. and Henikoff, J. G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Li, W. and Godzik, A. (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

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. *J. Mol. Biol.*, **247**, 536–540.

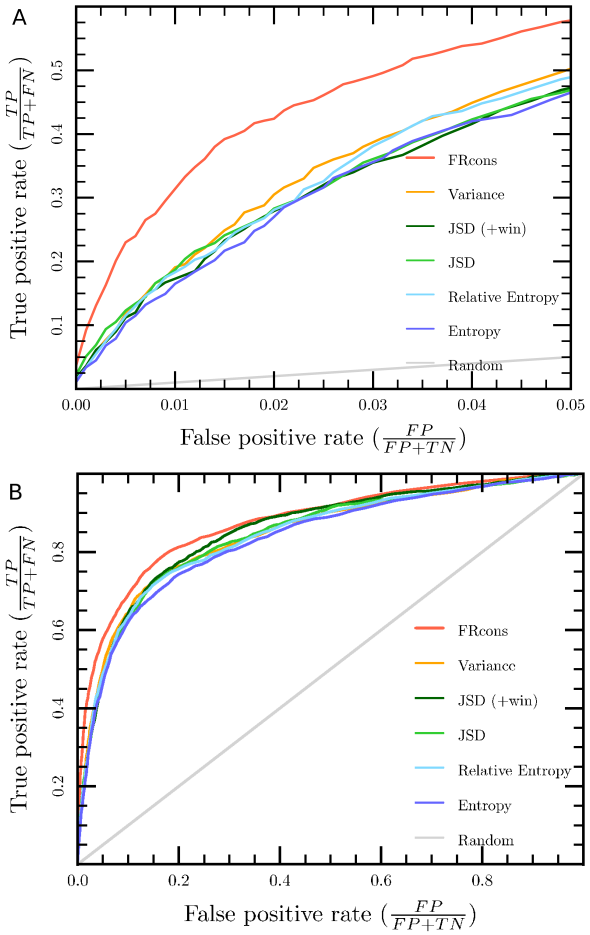

Fig. 6. (A) True positive rate vs. false positive rate for the prediction of catalytic residues. (B) Same as A but for false positive range up to 1.0.

Pei, J. and Grishin, N. V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.

Porter, C. T., Bartlett, G. J. and Thornton, J. M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, 129–133.

Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G. and Kuznetsov, E. N. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, **12**, 387–394.