BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 5 2008, pages 613–620
doi:10.1093/bioinformatics/btm626

Sequence analysis

Prediction of protein functional residues from sequence by probability density estimation

J. D. Fischer†, C. E. Mayer and J. Söding*,‡

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

Received on October 10, 2007; revised on November 20, 2007; accepted on December 14, 2007

Advance Access publication January 2, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The prediction of ligand-binding residues or catalytically active residues of a protein may give important hints that can guide further genetic or biochemical studies. Existing sequence-based prediction methods mostly rank residue positions by evolutionary conservation calculated from a multiple sequence alignment of homologs. A problem hampering more wide-spread application of these methods is the low per-residue precision, which at 20% sensitivity is around 35% for ligand-binding residues and 20% for catalytic residues.

Results: We combine information from the conservation at each site, its amino acid distribution, as well as its predicted secondary structure (ss) and relative solvent accessibility (rsa). First, we measure conservation by how much the amino acid distribution at each site differs from the distribution expected for the predicted ss and rsa states. Second, we include the conservation of neighboring residues in a weighted linear score by analytically optimizing the signal-to-noise ratio of the total score. Third, we use conditional probability density estimation to calculate the probability of each site to be functional given its conservation, the observed amino acid distribution, and the predicted ss and rsa states.

We have constructed two large data sets, one based on the Catalytic Site Atlas and the other on PDB SITE records, to benchmark methods for predicting functional residues. The new method FRcons predicts ligand-binding and catalytic residues with higher precision than alternative methods over the entire sensitivity range, reaching 50% and 40% precision at 20% sensitivity, respectively.

Availability: Server: http://frpred.tuebingen.mpg.de. Data sets: ftp://ftp.tuebingen.mpg.de/pub/protevo/FRpred/

Contact: soeding@lmb.uni-muenchen.de

Supplementary information: Supplementary data are available at *Bioinformatics* Online.

1 INTRODUCTION

An important aspect of the functional characterization of a protein is the determination of the residues mediating its function, such as catalytic residues, those forming the ligand-binding pocket, or residues involved in protein–protein interactions. To guide experiments, functional residues can be predicted by inference from homologous proteins whose functional sites have already been studied. Many tools and databases have been developed for this purpose (Hulo *et al.*, 2006; López *et al.*, 2007). Whenever no such information is available, functional residues can be predicted *de novo*. In the wake of the structural genomics initiative, a lot of effort has gone into developing methods for the *de novo* prediction of catalytic residues and ligand-binding sites from protein structure (reviewed by Jones and Thornton, 2004). However, structures are only available for a small fraction of proteins, which underscores the importance of being able to reliably predict functional residues based only on sequence. Also, any advance in this area is directly transferable to methods combining sequence and structural information since these sources of information have been shown to be largely complementary (Gutteridge *et al.*, 2003; Petrova and Wu, 2006; Youn *et al.*, 2007). By training their machine-learning methods on different subsets of sequence and structure-based features, these studies have identified the most important ones: residue conservation clearly tops the list, followed by amino acid type (or frequency distribution), surface geometry, and rsa (or similar measures).

The conservation of a residue is calculated from the amino acid frequency distribution in the corresponding column of a multiple sequence alignment of homologs. It is a measure for the functional or structural constraints that have acted on this position. Practically every known conservation measure has been tested for its ability to predict functional residues (Capra and Singh, 2007; Chelliah *et al.*, 2004; Madabushi *et al.*, 2002; Pei and Grishin, 2001; Pupko *et al.*, 2002; Valdar and Thornton, 2001; Wang and Samudrala, 2006; Zhang *et al.*, 2007), but there is no consensus so far as to what score works best (see review by Valdar, 2002).

A related group of methods detect residues that determine the functional subtype of proteins. Examples are positions that influence which substrate of a class of similar compounds is bound by a group of related enzymes. To detect such subtyping or tree-determining positions, these methods generally look for columns whose amino acid distributions differ strongly between the subtypes or between automatically clustered

*To whom correspondence should be addressed.
†Present address: EMBL-EBI, Hinxton, Cambridge, CB10 1SD, UK.
‡Present address: Gene Center Munich, University of Munich (LMU), 81377 Munich, Germany.

groups of homologous sequences (Casari *et al*., 1995; del Sol Mesa *et al*., 2003; Hannenhalli and Russell, 2000; Kalinina *et al*., 2004; Marttinen *et al*., 2006; Mihalek *et al*., 2004; Pei *et al*., 2006).

Motivated by the work of Youn *et al*. (2007), Petrova and Wu (2006), and Gutteridge *et al*. (2003), we aim here to use all information available from a protein's sequence to predict functional residues. We combine a new conservation score that takes into account the predicted local environment (Chelliah *et al*., 2004), predicted ss and rsa, and the profile amino acid frequencies at each position, in a simple and transparent statistical framework.

2 METHODS

2.1 Benchmark sets

Small or unevenly sampled test sets suffer from intrinsic noise and make it difficult to distinguish chance effects from systematic differences. We have therefore constructed two large and diverse benchmark sets, based on the Catalytic Site Atlas by Thornton and coworkers (Porter *et al*., 2004) and on PDB SITE records, which we name CSA and SITE. For comparison purposes, we also test all methods on a recently published, large data set by Capra and Singh (2007). The construction of the two sets is described in detail in the Supplementary Material. Briefly, for CSA we use two alternative definitions of true positive residues: catalytic (CSA-cat) and ligand-binding (CSA-ligand). Catalytic residues are defined according to the CSA, whereas ligand-binding residues are those that are in contact with a validated physiological ligand. A non-protein molecule is validated as physiological ligand by being in contact with a protein residue annotated as catalytic in the Atlas (with a 4$\mathring{A}$ distance cut-off). The SITE-ligand dataset uses the same definition of true positive residues as CSA-ligand. Here, a molecule is validated as physiological ligand by being in contact with a protein residue annotated in a PDB SITE record. Table 1 gives an overview over the benchmark sets. Note that the CSA and SITE sets are very diverse and evenly sampled, containing one member per SCOP family.

For benchmarking the various flavours of our functional residue prediction method FRcons, we use two-fold cross-validation: We divide the benchmark sets into two halves, ensuring that no SCOP superfamily (or EC family) is split between the halves. We train on the first half and test on the second and vice versa, then we pool the results.

2.2 Profile generation

Following the work of Pei and Grishin (2001), we tested three schemes to build sequence profiles from MSAs: 'unweighted', 'weighted' and 'independent counts' (see Supplementary Material). We have tested all benchmarked methods with all three profile building schemes (Fig. S1) (except Rate4Site that 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. 4B).

Table 1. Overview of the benchmark sets used in this study

	Proteins	SCOP families	Positive residues	Negative residues	Alignment diversity
CSA-cat	423	423	1536	107 463	11 ± 4
CSA-ligand			5331	103 668	
SITE-ligand	711	711	9547	142 628	11 ± 4
EC-ligand	828	348	16 166	273 718	7 ± 3

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.

2.3 Benchmarked methods

In the following we describe the scores that have been benchmarked, which includes all top-performing scores from the recent functional site prediction benchmark by Capra and Singh (2007). The sum-of-pairs measures as implemented in AL2CO were also tested but proved much inferior to the other measures and their results have therefore been omitted. For all methods in this section except Rate4Site, columns with >50% gaps have been given the minimum score, as implemented in the AL2CO method. To minimize the influence of the gap treatment, we have striven to build the alignments with uniformly high coverage (Section 2.1).

2.3.1 Normalization We investigated the effect of normalizing the conservation scores, $Z_i = (\text{Score}_i - \mu_{\text{Score}})/\sigma_{\text{Score}}$, where Score_i is a placeholder for the benchmarked scores, μ_{Score} is the average of Score_i over all positions i in the alignment and σ_{Score} is the SD over all alignment positions. The normalization considerably improves all scores except FRcons (Fig. 4B). We therefore normalized all scores except Rate4Site by default.

2.3.2 Shannon entropy The entropy for a profile column with amino acid frequencies p_{ia} is

$$\text{Entropy} = -\sum_{a=1}^{20} p_{ia} \log p_{ia} \tag{1}$$

and measures the amount of disorder in the amino acid distribution. It assumes its minimum value of 0 for a totally conserved column.

2.3.3 Relative entropy Whereas entropy treats all amino acids in the same way, relative entropy measures the deviation of the amino acid distribution p_{ia} from a background distribution f_a:

$$\text{Relative entropy} = \sum_{a=1}^{20} p_{ia} \log \frac{p_{ia}}{f_a} \tag{2}$$

As a consequence, a partially conserved column with 50% tryptophane ($f_W = 1.4\%$) will score higher than a fully conserved column with leucine ($f_L = 10\%$). For the relative entropy as well as for the Jensen Shannon divergence (see below), we use the amino acid background frequencies from the Gonnet matrix.

2.3.4 Variance Pei and Grishin (2001) propose as a conservation measure the root mean square deviation between the amino acid distribution p_{ia} and the average amino acid distribution over the whole alignment p_a, which they name *Variance*:

$$\text{Variance} = \left(\sum_{a=1}^{20} (p_{ia} - p_a)^2 \right)^{1/2} \tag{3}$$

It has the advantage over relative entropy of being less extreme in scoring deviations in frequencies of rare amino acids because the difference between frequencies instead of their ratio is used to measure deviation.

2.3.5 Jensen Shannon divergence Capra and Singh (2007) have applied the Jensen Shannon divergence (JSD) to scoring residue

conservation. As in the previous two scores, the deviation between the amino acid distributions p_{ia} and the background distribution (f_a) is measured:

$$\mathrm{JSD} = H\left(\frac{p_{ia}+f_a}{2}\right) - \frac{1}{2}H(p_{ia}) - \frac{1}{2}H(f_a). \quad (4)$$

Here, $H(\cdot)$ denotes the entropy of an amino acid distribution as defined in Equation (1). JSD can be interpreted as mutual information (Grosse *et al*., 2002): Given an amino acid drawn from either of the two distributions with a probability of 1/2, JSD is the amount of information that is gained for deciding which of the two distributions the amino acid was drawn from.

2.3.6 Rate4Site Rate4Site (Mayrose *et al*., 2004; Pupko *et al*., 2002) is a method that estimates the rates of evolution for each position in an alignment by constructing a maximum-likelihood phylogenetic tree and predicting the most likely rates of evolution with Bayesian statistics. We use Version 3.1 (slow version) with default parameters on the EC set. We could not benchmark Rate4Site on the other two sets because the fast version did not work on several alignments and the slow version was prohibitively slow.

2.4 The FRcons method

In this subsection we first introduce the basic FRcons conservation score and then explain its extensions: using amino acid background frequencies conditioned on predicted rsa and ss, including the effects of local sequence neighbors through a windowing method, and integrating this information with site-specific amino acid distributions by conditional probability density estimation.

2.4.1 The basic conservation score In devising a new conservation score, we were guided by Valdar's criteria (Valdar, 2002). Briefly, the score should (a) be continuous and bounded, (b) depend on the relative amino acid frequencies, (c) take the similarities between amino acids into account, (d) penalize gaps in the alignment column, (e) weight the sequences according to their diversity, and (f) be as simple as possible. In addition, we demand that (g) a maximally unconserved column get a score of 0, and (h) a fully conserved column get the maximum score, *independent* of the conserved amino acid.

The following score comes close to obeying these conditions:

$$\mathrm{FRcons}_{\mathrm{basic}} = \frac{\log \sum_{a=1}^{20} p_{ia}^2/f_a}{\log \sum_{a=1}^{20} p_{ia}/f_a}. \quad (5)$$

As does relative entropy and JSD, this score relates the profile amino acid distribution p_{ia} to a background distribution f_a. One can show that it attains its minimum of 0 when $p_{ia}=f_a (a=1,\ldots,20)$ and its maximum of 1 for a fully conserved column.

Equation (5) thus fulfills all criteria except (c) and (d). To make it respect (d), we penalize gaps in a straightforward way, multiplying the score by one minus the fraction of internal gaps in the alignment column. Here, an internal gap is a gap that is bordered by residues on both sides. We use Henikoff sequence weights to calculate this fraction.

2.4.2 Pseudocounts To fulfill (c), we add pseudocounts to the profile frequencies p_{ia} by the substitution matrix method (Durbin *et al*., 1998; Altschul *et al*., 1997):

$$\tilde{p}_{ia} = (1-\tau)\, p_{ia} + \tau \sum_{b=1}^{20} M(a,b)\, p_{ib}, \quad (6)$$

Here, τ quantifies how much pseudocounts are mixed into the original profile (see following paragraph). In the standard substitution matrix method, $M(a,b)$ would be the conditional probability matrix $P(a|b)$ that underlies the log-odds representation of substitution matrices: $S_{ab}=\log(P(a|b)/f_a)$. However, in that case condition (h) would be violated: For the same value of τ, a tryptophane would receive fewer pseudocounts than a serine, for instance, since a serine is much more likely to mutate than a tryptophane in the same time span. Hence the FRcons score would be higher for a column with only tryptophanes than for a column with only serines. We therefore define a matrix $M(a,b)=(1-\tau_b)\delta_{ab}+\tau_b P(a|b)$, which is a mixture of the identity matrix δ_{ab} and $P(a|b)$. We determine the mixture coefficients τ_b such that $\mathrm{FRcons}_{\mathrm{basic}}\ M(\cdot,b) = \min_{b' \in \{1,\ldots,20\}} \mathrm{FRcons}_{\mathrm{basic}} P(\cdot|b') = \text{const.}$ for all $b \in \{1,\ldots,20\}$. As substitution matrix, we chose the Gonnet matrix, but the particular choice is not critical. (The pseudocount matrix $M(a,b)$ can be obtained from the authors.)

The value of the pseudocount admixture is chosen in Equation (6) in a similar way as in PSI-BLAST (Altschul *et al*., 1997), $\tau=(\beta+1)/(N_{\mathrm{eff}}-1+\beta)$, where $\beta=5$ and N_{eff} is the average entropy over all alignment columns with <50% gaps. Thus, very diverse alignments receive few pseudocounts, whereas an alignment consisting only of a single sequence ($N_{\mathrm{aa}}=1$) gets the maximum amount of pseudocounts ($\tau=1$). The effect is that *after* the addition of pseudocounts, the profiles have approximately the same degree of diversity (or entropy) in their columns, independent of their initial diversity. In this way, conservation scores for alignments with very different alignment diversities can be compared. (We will show in Figure 4B, however, that a similar effect can also be achieved, at least in our benchmarks, by normalizing the conservation scores.) To treat all scores similarly, we normalize FRcons (Section 2.3) and name the resulting score $\mathrm{FRcons}_{\mathrm{basic}}$ in the following.

2.4.3 Trained background frequencies Instead of simply taking fixed background frequencies f_a, we can estimate the background frequencies given the predicted rsa and ss (Chelliah *et al*., 2004). We thereby assess how unusual the amino acid distribution of a profile column is compared to what would be expected for the predicted rsa and ss. This should allow us to better distinguish conserved core residues from conserved functional residues since core positions will mostly exhibit amino acid distributions that are common for their predicted low solvent accessibility.

We first construct training alignments in the same way as described in Section 2.1 for 5000 randomly chosen sequences from the nonredundant protein sequence database and predict the solvent accessibility with SABLE (Adamczak *et al*., 2004) and the secondary structure with PSIPRED (Jones, 1999). We divide the predicted rsa into 10 equally populated bins to obtain a single number $r_i \in \{0,\ldots,9\}$ for each position. Similarly, we divide the PSIPRED confidence values for helix and extended sheet states into 10 bins, obtaining h_i, e_i. For each profile column we then sum up the training profile frequencies p_{ia} for each amino acid a in the bin $(r_i,h_i,e_i) \in \{0,\ldots,9\}^3$ determined by the predicted rsa and ss states. After normalizing, we obtain a matrix containing the conditional background frequencies $f(a|r,h,e)$. These frequencies can now be used in place of the unconditioned frequencies f_a in Equation (5). (Figure S3 illustrates this procedure.)

2.4.4 Windowing over neighboring positions Capra and Singh showed that incorporating information about the conservation of sequentially neighboring positions improves the prediction of both catalytic and ligand-binding residues. They summed the conservation scores Z of the central position i and the neighboring positions $i+d$,

$$Z_{\mathrm{win}} = \sum_{d=-D}^{D} w_d Z_{i+d} \quad (7)$$

and empirically optimized the window length $2D+1$ and the total weight of the neighboring positions, weighting all neighboring positions the same. To improve this successful idea, we drop the restriction of constant weights for the neighbors and analytically optimize all weights

w_d independently. Technically speaking, we would like to optimize a signal-to-noise ratio, where the signal measures how much more score on average is given to the positive (i.e. functional) positions in comparison with the negatives. The noise is the standard deviation of the scores of the negative positions. Both these entities can be estimated from the training data. The signal can be written

$$\begin{aligned}\text{signal} &= \langle Z_{\text{win}}\rangle_{\text{pos}} - \langle Z_{\text{win}}\rangle_{\text{neg}} \\ &= \sum_{d=-D}^{D} w_d\big(\langle Z_d\rangle_{\text{pos}} - \langle Z_d\rangle_{\text{neg}}\big),\end{aligned} \tag{8}$$

where $\langle Z_d\rangle_{\text{pos}} = \sum_{i \text{ is pos}} Z_{i+d}/N_{\text{pos}}$ is the average conservation score Z over all positions at $+d$ residues from a positive position and $\langle Z_d\rangle_{\text{neg}} = \sum_{i \text{ is neg}} Z_{i+d}/N_{\text{neg}}$ is the average conservation scores Z over all positions at $+d$ residues from a negative position. The squared noise is

$$\text{noise}^2 = \text{var}(Z_{\text{win}}) = \sum_{d=1}^{4}\sum_{e=1}^{4} \text{cov}(Z_d, Z_e) w_d w_e, \tag{9}$$

where $\text{cov}(Z_d, Z_e) = \sum_{i \text{ is neg}} Z_{i+d}\, Z_{i+e}/N_{\text{neg}}$ is the covariance between scores at distance d and e from a negative residue. The signal-to-noise ratio can be maximized using the method of Lagrange multipliers, by maximizing the signal under the constraint of constant noise. To separate training and test data, we use two-fold cross-validation as described in Section 2.1. The optimum value for D is 2 for all data sets. For the CSA-catalytic set, we get (averaged over both halves) $(w_{-2}, \ldots, w_{+2}) = (0.10, 0.14, 0.90, 0.15, 0.10)$, for the CSA-ligand set $(w_{-2}, \ldots, w_{+2}) = (0.09, 0.12, 0.94, 0.13, 0.09)$ and similar values for SITE-ligand and EC-ligand. This is not far from the values Capra and Singh empirically optimized: $D = 3$ and $(w_{-3}, \ldots, w_{+3}) = (1/8, 1/8, 1/8, 1, 1/8, 1/8, 1/8)$. (We have scaled their weights by a factor 7/4 to show the correspondence).

2.4.5 Probability density estimation The constraint on residues to take part in a specific catalytic activity or to bind ligands certainly influences the observed frequency distribution considerably. Valines are much underrepresented at catalytic sites, whereas lysines or aspartates are highly overrepresented, for instance. The degree of under- or over-representation may also be correlated with other properties, such as rsa or conservation.

We aim to exploit this information by estimating the probability that a position i is positive (i.e. catalytic or ligand-binding), given its amino acid frequency distribution p_{ia}, its predicted rsa r_i, predicted helix and extended sheet propensities h_i, e_i, and its conservation score Z_i. We first use Bayes' theorem (Sivia, 2006; Durbin *et al.*, 1998) to calculate the *posterior probability* of finding a positive residue, given the data $(p_{ia}, r_i, h_i, e_i, Z_i)$,

$$P(i \text{ pos}|p_{ia}, r_i, h_i, e_i, Z_i) = \frac{P(p_{ia}, r_i, h_i, e_i, Z_i|i \text{ pos})}{P(p_{ia}, r_i, h_i, e_i, Z_i)}\; P(\text{pos}) \tag{10}$$

and then estimate the numerator and denominator by modeling the probabilities with the Bayesian network (Needham *et al.*, 2007) shown in Figure S2A: For the likelihood in the numerator we get

$$\begin{aligned}&P(p_{ia}, r_i, h_i, e_i, Z_i|i \text{ pos}) \approx P(p_{ia}|r_i, Z_i, i \text{ pos}) \\ &\times P(r_i|Z_i, i \text{ pos})\; P(h_i, e_i|Z_i, i \text{ pos})\;\; P(Z_i|i \text{ pos})\; P(i \text{ pos})\end{aligned} \tag{11}$$

where the first factor on the right-hand side can be approximated by

$$P(p_{ia}|r_i, Z_i, i \text{ pos}) \approx \prod_{a=1}^{20} P(a|r_i, Z_i, i \text{ pos})^{p_{ia}}, \tag{12}$$

and analogously for the denominator. We can now substitute Equation (12) into (11) and then into (10). The result is expressed

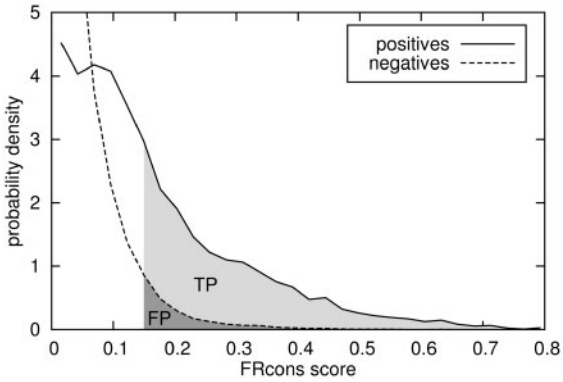

Fig. 1. FRcons score distribution for the 1536 positive and 107 463 negative residues in the CSA-ligand data set.

in terms of the odds matrices

$$\begin{aligned}&A(a, r, Z) = \frac{P(a|r, Z, \text{pos})}{P(a|r, Z)}, \qquad R(r, Z) = \frac{P(r|Z, \text{pos})}{P(r|Z)}, \\ &S(h, e, Z) = \frac{P(h, e|Z, \text{pos})}{P(h, e|Z)}, \qquad C(Z) = \frac{P(Z|\text{pos})}{P(Z)}: \\ &P(i \text{ pos}|p_{ia}, r_i, h_i, e_i, Z_i) \approx \\ &\textstyle\prod_{a=1}^{20} A(a_i, r_i, Z_i)^{p_{ia}}\; R(r_i, Z_i)\; S(h_i, e_i, Z_i)\; C(Z_i)\; P(\text{pos}).\end{aligned} \tag{13}$$

The odds matrices and the *a priori* probability of a functional residue $P(\text{pos})$ can be estimated from the training alignments (Fig. S2B). We first determine the total count matrices $C(a, r_i, h_i, e_i, Z_i)$ and $C(a, r_i, h_i, e_i, Z_i|i \text{ pos})$ by iterating over all positions i of the training alignments and adding p_{ia} to the count matrix bins $C(a, r_i, h_i, e_i, Z_i)$ for all a, and to $C(a, r_i, h_i, e_i, z_i, \text{pos})$ if position i is positive. Here, r_i, h_i, e_i, and Z_i are obtained by running SABLE (Adamczak *et al.*, 2004), PSI-PRED (Jones, 1999) and $\text{FRcons}_{\text{basic}}$, respectively, and dividing the results into 10 bins (20 bins for Z_i).

Since there are far too many bins ($20 \times 10^3 \times 20$) in the count matrix to be sufficiently populated by the $\sim 5 \times 10^4$ positions in the training alignments, we smear out the counts with functions $w_r(r') = 0.5^{-|r'-\mathrm{r}|}$, $w_Z(Z') = 0.5^{-|Z'-Z'|}$, and $w_{he}(h', e') = 0.7^{(-|\mathrm{h}'-h| = |e'-\mathrm{e}| + |h'+e'-h-e|)/2}$, similar to the way Gaussian envelopes in classical probability estimation are convoluted over the counts. We calculate a smoothed matrix with

$$\begin{aligned}&\tilde{C}(a, r, h, e, Z) \\ &= \frac{\sum_{r', h', e'=1}^{10}\sum_{Z'=1}^{20} w_r(r') w_{he}(h', e') w_Z(Z') \times C(a, r', h', e', Z')}{\sum_{r', h', e'=1}^{10}\sum_{Z'=1}^{20} w_r(r') w_{he}(h', e') w_Z(Z')}.\end{aligned} \tag{14}$$

In an analogous way, we smear out the counts of the positive residues to obtain $\tilde{C}(a, r, h, e, Z|\text{pos})$. From these smoothed matrices, the conditional probabilities in the numerator and denominator of the odds matrices in Equation (13) can be obtained by summing over the appropriate indices.

3 RESULTS AND DISCUSSION

Figure 1 shows the probability density for the FRcons score, calculated as explained in Sections 2.4.1–2.4.5, for positive (i.e. functional) and negative residues on the CSA-ligand data set. Positive residues are strongly enriched in the high-score range relative to negatives: If we set the score threshold to 1.5 (see shaded areas in Fig. 1), about 40% of positive residues

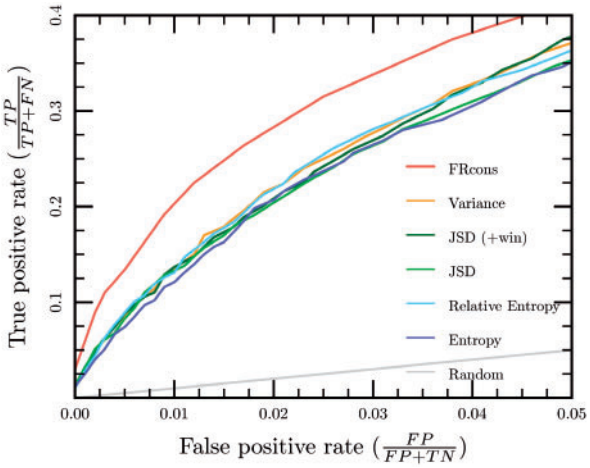

Fig. 2. ROC plot on the CSA-ligand data set.

will be predicted as true positives (light gray area), whereas only about 4% of negative residues will be predicted as false positives (dark gray area).

Let us introduce a few definitions. The positive residues above (below) the threshold score are called true positives (false negatives), and the negative residues above (below) the threshold are called false positives (true negatives). These are abbreviated TP, FN, FP, and TN, respectively. The *true positive rate* is defined as TPR = TP/(TP + FN) and the *false positive rate* as FPR = FP/(FP + TN). To compare the predictive power of different methods, a ROC plot is often drawn, tracing TPR vs. FPR while varying the threshold score from $-\infty$ to ∞. (Often, TPR and FPR are called sensitivity and 1–specificity, respectively.)

Figure 2 shows a ROC plot for six conservation methods applied to the CSA-ligand set. Before discussing the results in the next subsection, a few remarks about the graphical presentation of the data are in place. It looks at first glance as though 40% of the positive residues should be predictable with fairly high confidence. However, one must bear in mind that ligand-binding residues make up only about 5% of the total number of residues (Table 1), hence at a true positive rate of 40% and a false positive rate of 4%, the ratio of TP to FP is about $0.4 \times 0.05{:}0.04 = 1{:}2$, corresponding to a *precision* TP/(FP + TP) of only 33% and a *false discovery rate* FP/(FP + TP) of 67%. This exemplifies the importance of carefully interpreting TPR–FPR plots if positives and negatives are highly unbalanced (Davis and Goadrich, 2006). First, the FPR gives only indirect information about the false discovery rate, which is the more relevant measure for practical purposes. Second, the names 'false positive rate' and also '1–specificity' carry the risk of being misunderstood to be synonymous with 'error rate'. We have, therefore, chosen to present the benchmark results as *precision versus sensitivity* plots, where precision can be interpreted as 1–error rate and sensitivity is synonymous with TPR and recall. To allow easier comparison with previous studies, we include TPR versus FPR versions of all precision–sensitivity plots in the Suppplementary Materials (Figs. S4–S6).

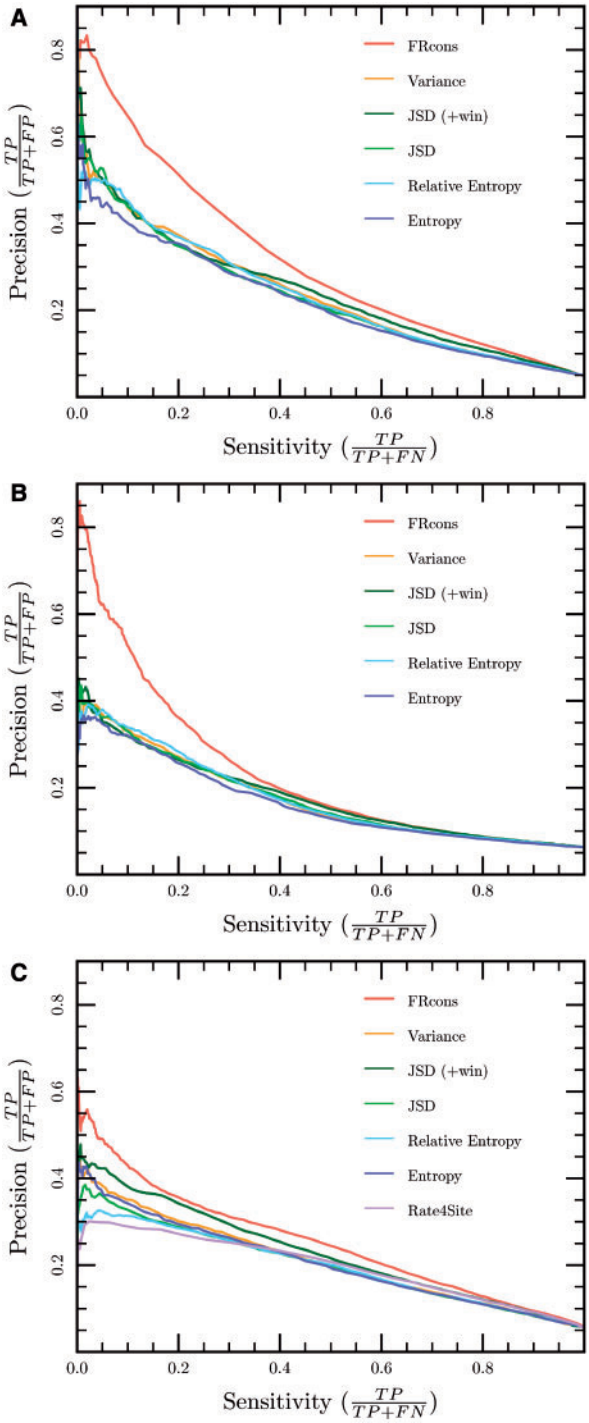

Fig. 3. Prediction of ligand-binding residues on three sets: (**A**) CSA, (**B**) SITE and (**C**) EC. (Note: sensitivity is the same as recall.)

3.1 Prediction of ligand-binding residues

Precision versus sensitivity for the CSA-ligand benchmark set is shown in Figure 3A. First, we note that the precision of all methods is below 30% at 50% sensitivity, a fairly sobering result that suggests plenty of room for improvement.

Second, the differences in performance between entropy, relative entropy, variance, and JSD are minor compared to the improvement over these methods by FRcons. The absolute improvement by FRcons is particularly pronounced at high precisions. In relative terms, FRcons precision shows between 10% and 40% higher precision than the other methods over most of the sensitivity range. The windowing (Section 2.4.4) is able to improve the JSD score, but only at precisions below 0.3.

The results on the SITE-ligand benchmark set (Fig. 3B) yield similar results. The improvement of FRcons over the other methods is more pronounced for lower sensitivities, but vanishes above 50% sensitivity. We surmise that the much more heterogenuous quality of the PDB SITE annotations compared with the manually curated, literature-based CSA annotations is responsible for the smaller differences in the right half of the plot. The larger differences toward low sensitivities might be explained by a higher coverage of ligand-binding sites in the SITE annotated structures. A higher coverage would lead to fewer falsely assigned negatives and to a higher achievable maximum precision.

The results on the EC-ligand benchmark set (Fig. 3C) show a much weaker improvement of FRcons over the other methods, but the former still performs best over the entire sensitivity range. Entropy is slightly better than JSD and relative entropy, although the opposite is true for the other two sets. We suspect that the main cause for the differences between the EC set and the other two benchmark sets is the presence of a fair amount of alignments in the EC-benchmark set with very low diversity (Table 1). Alternatively, some non-physiological ligands might have been used to define positive residues. This would explain the much smaller maximum precision reached (55% instead of 85%). However, this hypothesis cannot explain the observation that only some of the methods have a decreased performance on the EC set in comparison with the CSA set (FRcons, relative entropy, entropy) whereas others have similar (variance) or improved performance (JSD, JSD+win).

The relatively weak predictive performance of Rate4Site might be due to the fact that Rate4Site ranks conserved residues from the core higher on average than the other methods, degrading its performance. This disappointing result is surprising, however, since the study by Capra and Singh (2007) showed Rate4Site to be slightly better than the other tested methods. While Rate4Site scores were calculated in the same way, the differences are probably related to the calculation of JSD, relative entropy, and entropy: (a) In our study these scores are normalized, whereas they were not in the other study, and (b) we do not add pseudocounts, whereas constant pseudocounts of 10^{-6} were employed in the other study.

3.2 Prediction of catalytic residues

The results of predicting catalytic residues on the CSA catalytic benchmark set are shown in Figure 4A. In comparison with the CSA-ligand graph (Fig. 3A), the precision is lower for all tested methods. This is not surprising since many conserved ligand-binding sites will have high scores and will become high-scoring false positives in this benchmark set. However, FRcons manages at least to some extent to distinguish ligand-binding

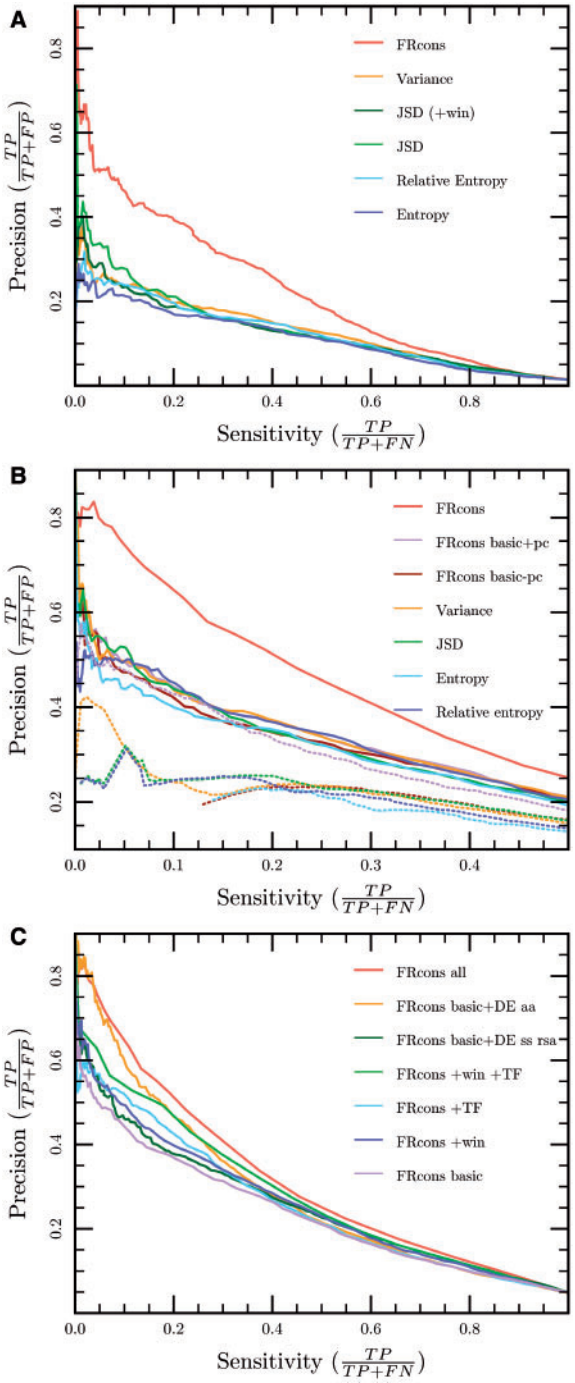

Fig. 4. (**A**) Precision versus sensitivity for the prediction of catalytic residues. (**B**) Normalization improves the performance of all scores except FRcons, as shown here for the CSA-ligand set. Unnormalized scores are plotted as dotted lines, normalized scores as solid lines. Note that the x-axis scale and the data set are different from A. (**C**) Effect of the different FRcons components, shown here for the CSA-ligand set. FRcons basic: basic score as described in Sections 2.4.1–2.4.2; TF: trained background frequencies (Section 2.4.3); win: windowing over neighboring positions (Section 2.4.4);DE ss rsa: density estimation using predicted ss and rsa (Section 2.4.5); DE aa: density estimation using amino acid distribution (Section 2.4.5); FRcons all: full FRcons score (Sections 2.4.1–2.4.5).

residues from catalytic ones, because its improvement over the other methods is larger than the CSA-ligand set. The likely reason is that FRcons uses information other than conservation, e.g. the amino acid composition of sites.

3.3 Score normalization

Normalization relative to all positions in the query protein (Section 2.3) greatly improves the performance of all scores (Fig. 4B), except FRcons with pseudocounts (see mauve FRcons basic + pc trace). Without normalization, alignments containing only a few very similar sequences look highly conserved at all positions compared to highly diverse alignments. After normalization, the conservation score of each particular position is related to the conservation at all other positions. Hence, a very narrowly sampled alignment will not have many more highly scoring positions than a diverse alignment once the normalized score is used. A similar effect can be achieved by adding diversity-dependent pseudocounts, which has been done for FRcons, explaining why FRcons performance does not critically depend on the normalization step.

The magnitude of the improvement is likely to be smaller in practice than in the CSA-ligand benchmark set, which would mean the results of all methods except FRcons are slightly optimistic. The reason is that the normalization helps more when the fraction of positives varies little from protein to protein. Due to the construction of the CSA-ligand benchmark set, we expect this variation to indeed be the lower than it will be in practice.

3.4 Contribution by various FRcons components

To gain insight into the source of improvement of FRcons over the other methods, we have tested several versions of FRcons by including the components described in Sections 2.4.1–2.4.5 in various combinations (Fig. 4C). The basic FRcons score including pseudocounts (mauve) performs similarly to the other tested scores (Fig. 4B). Windowing (blue) and trained frequencies (cyan) both improve the performance by a small amount, which adds up when combined (green). Density estimation using ss and rsa but without using trained frequencies or windowing gives only a marginal improvement (dark green). The largest contribution by far stems from density estimation using amino acid frequencies (orange). When adding the other components (red trace), there is a clear improvement at intermediate and high sensitivities. However, the effects are unfortunately far from additive (compare the green and orange trace with red). To make further progress, it will be important to understand the underlying reasons.

4 CONCLUSION

We have developed a sequence-based method for the prediction of catalytic- or ligand-binding residues that combines a new conservation score with two further ingredients: the estimation of background frequencies conditioned on predicted ss and rsa and a probability density estimation technique to integrate the information from conservation, predicted ss and rsa, and amino acid frequencies. Furthermore, we show how to analytically optimize the weights for the inclusion of neighboring residues (Capra and Singh, 2007).

The proposed method considerably improves on existing, conservation-based methods for the prediction of both ligand-binding and catalytic residues. The largest contribution comes from combining the conservation score with the amino acid frequencies by probability density estimation. This method was chosen for its simplicity and transparency, but it would make sense to compare its performance to that of other techniques such as Support Vector Machines. An advantage could be that one would not have to specify conditional dependencies in order to artificially reduce the complexity of the problem.

In a comparison to approaches using sequence and structure information, we attain similar performance for catalytic residue prediction: At 18.5% precision, our method has a sensitivity of ~50%, close to the 51.1% cited by Youn *et al.* (2007) for fold-level training. At 14% precision we get 57% sensitivity, again close to the 56% obtained by Gutteridge *et al.* (2003) without spatial clustering. However, such a comparison can only give a rough indication due to the different data sets used. Our set, for instance, contains 1.5% positives whereas the CATRES set contains only 1.1%.

The design of two large benchmark sets for the prediction of ligand-binding residues, based on the CSA and on the PDB SITE annotations, proved crucial for the development and testing of our new method, as no such large data set was around when the work was begun. Together with the new data set by Capra and Singh (2007) and the (not CAS) (Porter *et al.*, 2004), they will hopefully assist others in making progress in the prediction of functional residues.

ACKNOWLEDGEMENTS

Many thanks to John Capra and Mona Singh for making the EC benchmark set available and to Jimin Pei for the AL2CO source code. We thank Andreas Biegert, Michael Remmert, and Oliver Kohlbacher for helpful suggestions and Jasmina Ponjavic for preliminary work on the prediction of subtyping residues. We are especially grateful to Andrei Lupas for initiating this work and giving ample support and advice. Financing by the Max-Planck-Society is gratefully acknowledged.

Conflict of Interest: none declared.

REFERENCES

Adamczak,R. *et al.* (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, **56**, 753–767.

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

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

Casari,G. *et al.* (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

Chelliah,V. *et al.* (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.*, **342**, 1487–1504.

Davis,J. and Goadrich,M. (2006) The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, **148**, 233–240.

del Sol Mesa,A. *et al.* (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.

Durbin,R. *et al*. (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.
Grosse,I. *et al*. (2002) Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys. Rev. E*, **65**, 041905.
Gutteridge,A. *et al*. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, **330**, 719–734.
Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.
Hulo,N. *et al*. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, 227–230.
Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
Jones,S. and Thornton,J.M. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, **8**, 3–7.
Kalinina,O.V. *et al*. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.*, **32**, 424–428.
López,G. *et al*. (2007) Firestar–prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.*, **35**, 573–577.
Madabushi,S. *et al*. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.
Marttinen,P. *et al*. (2006) Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics*, **22**, 2466–2474.
Mayrose,I. *et al*. (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.
Mihalek,I. *et al*. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.
Needham,C.J. *et al*. (2007) A primer on learning in Bayesian networks for computational biology. *PLoS Comput. Biol.*, **3**, e139.
Pei,J. *et al*. (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, **22**, 164–171.
Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.
Petrova,N.V. and Wu,C.H. (2006) Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, **7**, 312.
Porter,C.T. *et al*. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, 129–133.
Pupko,T. *et al*. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18** (Suppl 1), 71–77.
Sivia,D.S. (2006) *Data Analysis. A Bayesian tutorial*. Oxford University Press, Oxford.
Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.
Valdar,W.S. and Thornton,J.M. (2001) Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–124.
Wang,K. and Samudrala,R. (2006) Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, **7**, 385.
Youn,E. *et al*. (2007) Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci.*, **16**, 216–226.
Zhang,S.W. *et al*. (2007) Estimating residue evolutionary conservation by introducing von Neumann entropy and a novel gap-treating approach. *Amino Acids*, DOI 10.1007/s00726-007-0586-0.