BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–7

Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine Learning approach for feature weighting

Kai Ye[†], K. Anton Feenstra[‡], Jaap Heringa[‡], Adriaan P. IJzerman[†], Elena Marchiori[*‡]

[†]Division of Medical Chemistry, LACDR, Leiden University, P.O. Box 9502, 2300 RA, Leiden, The Netherlands
[‡]Dept. of Computer Science, IBIVU, Vrije Universiteit, De Boelelaan 1081A, 1081 HV, Amsterdam, The Netherlands

ABSTRACT

Motivation:

Identification of residues that account for protein function specificity is crucial, not only for understanding the nature of functional specificity, but also for protein engineering experiments aimed at switching the specificity of an enzyme, regulator or transporter. Available algorithms generally use multiple sequence alignments to identify residue positions conserved within subfamilies but divergent in between. However, many biological examples show a much subtler picture than simple intra-group conservation versus inter-group divergence.

Results: We present multi-RELIEF, a novel approach for identifying specificity residues that is based on RELIEF, a state-of-the-art Machine Learning technique for feature weighting. It estimates the expected 'local' functional specificity of residues from an alignment divided in multiple classes. Optionally, 3D structure information is exploited by increasing the weight of residues that have high-weight neighbors. Using ROC curves over a large body of experimental reference data, we show that *a)* multi-RELIEF identifies specificity residues for the seven test-sets used, *b)* incorporating structural information improves prediction for specificity of interaction with small molecules, *c)* comparison of multi-RELIEF with four other state-of-the-art algorithms indicates its robustness and best overall performance.

Availability: A web-server implementation of multi-RELIEF is available at www.ibi.vu.nl/programs/multirelief. Matlab source code of the algorithm and data sets are available on request for academic use.

1 INTRODUCTION

Many homologous protein families have a common biological function but different specificity towards substrates, ligands, effectors, proteins and other interacting molecules. All these interactions require a certain specificity. Identifying crucial residues for this specificity is a prerequisite for understanding the nature of functional specificity, for planning experiments on functional analysis or protein redesign, and for guiding point mutations aimed at switching the specificity of an enzyme, regulator or transporter.

In order to detect specificity residues, advanced computational techniques are needed, because of a great variety of functional specificities observed in nature and the vast amount of protein sequence data. A number of algorithms have been proposed in recent years for detecting specificity residues from a multiple sequence alignment (MSA) (Hannenhalli and Russell, 2000; Bickel *et al.*, 2002; Del Sol Mesa *et al.*, 2003; Kalinina *et al.*, 2004; Mihalek *et al.*, 2004; Carro *et al.*, 2006; Gu, 2006; Ye *et al.*, 2006; Feenstra *et al.*, 2007). Most algorithms employ information-entropy related scoring functions (Shenkin *et al.*, 1991) to rank residue positions according to the association with the subfamilies (for an overview see Whisstock and Lesk, 2003). While many algorithms require a predefined subdivision of the MSA into classes, some induce a grouping on the fly.

The SDPpred method (Kalinina *et al.*, 2004) uses mutual information to identify residue positions in which amino acid distributions correlate with the sub-family grouping (Mirny and Gelfand, 2002).

The Two-entropies analysis algorithm (TEA) (Ye *et al.*, 2006) creates a 2-dimensional plot of residue conservation in terms of Shannon entropy at both superfamily and subfamily level. Functional sites such as conserved or specificity residues can be distinguished easily from other residues.

The TreeDet approach introduced by Del Sol Mesa *et al.*, 2003 contains three algorithms for detecting so-called tree-determinant residues from an unpartitioned MSA. The Level Entropy (S) method first uses relative entropy to search for an optimal grouping of the alignment and then considers positions conserved within classes but different among classes as the tree-determinants. The Sequence Space Automatization (SS) method applies principle component analysis to the alignment and computes an optimal number of clusters and the residues that correspond to them. Finally, the Mutational Behavior (MB) method looks for residues whose mutational behavior resembles the phylogeny of the alignment.

The Sequence Harmony (*SH*) method (Pirovano *et al.*, 2006; Feenstra *et al.*, 2007) scores compositional overlap between two user-specified groups. The algorithm does not exploit the notion of sub-family conservation but focusses on compositional differences between the sub-families.

In this paper we introduce multi-RELIEF, a new algorithm for identifying specificity residues from a given MSA and predefined multiple classes using 'local' conservation properties. The approach is based on a state-of-the-art Machine Learning technique for feature weighting, called RELIEF, which exploits the notion of locality for estimating relevance of attributes in discriminating samples from

*to whom correspondence should be addressed: elena@cs.vu.nl.

two classes (Kononenko, 1994). In the biological context considered here, locality corresponds to *sequence space* (Landgraf *et al.*, 2001).

Multi-RELIEF estimates the expected 'local' specificity of residues, by comparing each sequence with the most similar sequence in the same class and with the most similar in opposite classes. The nearest neighbor sequences are selected based on global identity. A residue is considered relevant if it has high local specificity with respect to at least one pair of classes.

While other algorithms consider residue positions independently, multi-RELIEF considers global sequence similarity while scoring each residue. Furthermore, the method can cope with sub-family classifications derived from phylogeny, which generally are heterogeneous. Miss-classification, a general error that can arise from e.g. misannotation, will result in a close match between some opposing classes. Multi-RELIEF is able to 'recover' the innate specificity of a class, whenever one of the other classes can be contrasted to it. This alleviates the problem of downweighting the relevance of residue positions, for example, in cases where a single class is 'polluted' with a misplaced sequence.

Multi-RELIEF can optionally include 3D-structural information, if available. It does this by employing a new heuristic based on the assumption that a specificity residue does not evolve in isolation, but within a functional cluster in the protein structure. This means that a residue would be more likely to be a specificity residue if its neighboring residues are also specific.

To test our novel approach thoroughly, seven experimentally determined benchmark sets were considered, taken from five widely studied protein families: G protein-coupled receptors (GPCRs), the LacI family of bacterial transcription factor, the Ras-superfamily of small GTP-ases, the MIP-family of integral membrane transporters and the Smad family of transcription factors. The performance of multi-RELIEF was compared with TEA and SDPpred (both acting on multiple classes), TreeDet/MB (no class division required) and *SH* (acting on two classes). Using ROC curves we show that *a)* multi-RELIEF identifies specificity residues, *b)* incorporating structural information improves prediction for specificity of interaction with small molecules, and *c)* comparison of multi-RELIEF with other algorithms indicates its robustness and best overall performance.

2 METHODS

2.1 RELIEF

Many interesting feature weighting algorithms based on different approaches have been introduced in Machine Learning (Guyon and Elisseeff, 2003). One particular class uses a multivariate 'filter' prior to the construction of a model (the classifier) to quantify the relevance of features as to their ability to jointly discriminate between classes. RELIEF is considered one of the most successful filter multivariate feature weighting algorithms (Guyon and Elisseeff, 2003), due to its simplicity and effectiveness (Kononenko, 1994). We recently applied RELIEF for selecting specificity residues ('subtype specific functional sites') from protein sequences of the Smad receptor binding family(Marchiori *et al.*, 2006).

Given samples from two classes, RELIEF iteratively assigns weights to features based on how well they separate samples from their nearest neighbor (nnb) within the same class relative to that within the opposite class (Marchiori *et al.*, 2006). To do this, RELIEF employs a feature weight vector. At each iteration, one sequence seq is selected. The weights are updated by adding the 'difference' between seq and its nnb from the opposite class, $miss(seq)$, and subtracting the difference between seq and its nnb from the same class, $hit(seq)$. We define nnb for a sequence seq to class l as $nnb(seq) = argmin\,\{d(seq, x)\, x \in X_l\}$ where d denotes the Hamming distance between strings (e.g., $d(\texttt{ALM}, \texttt{VLM}) = 1$). The difference between two sequences $seq1 - seq2$ is a vector representing matches (`0`) and mismatches (`1`) between residues (e.g., $\texttt{ALM} - \texttt{VLM} = \texttt{100}$). This procedure is iterated over all sequences of the dataset. The computational complexity of RELIEF is $O(nr\_seq^2 \cdot nr\_positions)$.

A residue position ('site') will obtain best weight if it has maximal 'local' specificity over all triplets of a sequence, its nearest neighbor in the same, and that in the opposite class, that is, local in *sequence space*. Thus if a residue position is conserved within each class but divergent between classes, then its RELIEF weight will be high. Completely conserved positions and overall divergent positions will get zero weight, while positions that are divergent within subfamilies but conserved between subfamilies will get *negative* weight.

2.2 Multi-RELIEF

RELIEF is a two-class feature weighting algorithm. However, large protein families with a variety of specificities require algorithms acting on multiple classes. Extensions of RELIEF to handle multiple classes have been proposed (Kononenko, 1994; Robnik-Sikonja and Kononenko, 2003; Sun and Li, 2006). For instance, Kononenko, 1994 introduced RELIEF-F where the weight vector is updated by the sum of $miss(seq)$ weighted by the estimated *a priori* probabilities of the classes. Here we present a new ensemble approach based on random sub-sampling of pairs of classes. The multi-RELIEF algorithm is illustrated below in pseudo-code.

```
Multi-RELIEF
%input:  X1,..,Xm (m classes of aligned proteins)
%parameters: nr_iter, nr_sample
%output: multi_W (weights assigned to positions)
nr_positions = total number of positions;
weights = zero vector of size nr_positions;
for i=1: nr_iter
     select randomly two classes
    X = select randomly nr_sample sequences
        from each selected class
    W_i = apply RELIEF to X
end;
for s=1: nr_positions
multi_W(s) = (average across positive W_i(s)'s);
end;
return multi_W
```

In multi-RELIEF, multiple runs (nr_iter) of RELIEF are performed. At each run i, first two classes are randomly selected. Next, nr_sample sequences from each class are randomly selected. Finally, RELIEF is applied to the resulting two classes, yielding an output vector W_i. When the multiple runs are completed, the weight $multi\_W(s)$ of a position s is computed by averaging the positive weights assigned to that position by the nr_iter runs of RELIEF. That is, using $N^+ = |\{W_i(s) > 0\ \forall\ i\}|$ and $N^- = |\{W_i(s) < 0\ \forall\ i\}|$,

$$multi\_W(s) = \begin{cases} \frac{1}{N^+} \sum_i \{W_i(s) > 0\ \forall\ i\} & \text{for } N^+ > 0 \\ \frac{1}{N^-} \sum_i \{W_i(s) < 0\ \forall\ i\} & \text{for } N^+ = 0 \wedge N^- > 0 \\ 0 & \text{for } N^+ = 0 \wedge N^- = 0 \end{cases}$$

Note that in the definition of $multi\_W(s)$, only those runs where RELIEF assigned a positive weight to s are considered. In this way, $multi\_W(s)$ assigns a high score to position s only if it discriminates at least two classes. In particular, a maximum weight will be assigned if s fully discriminates two specific classes but does not differentiate (i.e. weight less than or equal to zero) any other pair of classes.

Table 1. Weights computed by multi-RELIEF applied to a toy example.

	a	b	c	d	e
C_1	R	F	T	I	T
	R	F	T	Q	F
	R	F	T	N	V
	R	F	T	A	D
C_2	R	F	Y	S	T
	R	F	Y	F	F
	R	F	Y	D	V
	R	F	Y	V	D
C_3	R	Y	D	E	T
	R	Y	D	V	F
	R	Y	D	W	V
	R	Y	D	G	D
C_4	R	Y	H	H	T
	R	Y	H	P	F
	R	Y	H	Y	V
	R	Y	H	C	D
weights	0	1	1	0	−1

Random sampling of pairs of classes is mainly employed for efficiency reasons, while random sub-sampling of sequences is applied for handling unbalanced classes as well as for gaining efficiency. The computational complexity of multi-RELIEF is $O(nr\_iter \cdot nr\_sample^2 \cdot nr\_positions)$, while that of RELIEF-F is $O(nr\_seq^2 \cdot nr\_positions)$. Algorithms that do not consider the context (univariate scoring algorithms), such as TEA and *SH*, are generally more efficient with complexity $O(nr\_seq \cdot nr\_positions)$.

Table 1 illustrates the application of multi-RELIEF to a toy dataset. Note that positions *b* and *c* both get maximum weight. This is expected for position *c*, because it fully discriminates each pair of classes. Instead, position *b* only discriminates a subset of classes, e.g., C_1/C_3, while it does not separate the remaining pairs of classes, i.e. C_1/C_2 and C_3/C_4 . So only residue positions that, at least partly, discriminate between pairs of classes have a positive weight assigned by multi-RELIEF. This property of the algorithm is desirable, e.g., in cases where the number of subfamilies is larger than the number of amino acids, such as the GPCR benchmark (see below) that consists of 77 classes.

2.3 multi-RELIEF + 3D contacts

As an additional step in multi-RELIEF, 3D-structural information can be exploited. We use a simple heuristic based upon the notion that functional specificity generally does not evolve for a single residue but typically involves a cluster of residues in the protein structure. For each position s, we adjust the corresponding multi-RELIEF weight by adding the average weight of its 3D neighbors. Thus the score of a residue will be boosted if its neighbors have a high average score. 3D neighbors are residues that share surface with a given residue as calculated by the web server at http://ligin.weizmann.ac.il/cma/ (Sobolev *et al.*, 1999). From the list returned, residue pairs with a sequence distance of two or less are removed.

2.4 Comparison to other Algorithms

To assess the performance of multi-RELIEF versus other methods, we included the following four recent stat-of-the-art algorithms:

1. TEA: Ye *et al.*, 2006
2. SDPpred: http://bioinf.fbb.msu.ru/SDPpred/index.jsp
3. TreeDet/MB: http://www.pdg.cnb.uam.es/Servers/treedet/
4. *SH*: http://www.ibi.vu.nl/programs/seqharmwww/

To be able to use TEA automatically we used the following scoring function:

$$score(s) = Entropy(s, D) - \frac{1}{N} \sum_{C \in Classes} Entropy(s, C),$$

for each position s in a given MSA D partitioned into N $Classes$ C, and using $Entropy(s, X)$ for the entropy of s computed on dataset X.

SDPpred was applied with 10 000 shuffles for each column, and a maximum allowed percentage (70%) of gaps in a group in each column; these are the highest possible settings allowed through the web-interface of SDPpred.

TreeDet/MB was applied with the following setting, in order to obtain a ranking of the residues: Advanced run for MB method, cutoff set to 10^{-12} and percentage of High Scoring Residues set to 100%. We could not run TreeDet on the GPCR dataset because its web server accepts a maximum of 200 sequences. For this reason, we compiled a GPCR-190 reduced set (see below), to which TreeDet was applied.

SH has no adjustable parameters, except for the cut-off value that is irrelevant for the generation of the ROC curves used. Note that for a fair comparison between the methods, the tie-braking by sequential groups ('Rank') and entropy was excluded from the *SH* method. A similar mechanism could be added to the other methods in a post-processing step. *SH* was not applied to the GPCR and LacI datasets since these consist of more than two classes.

Multi-RELIEF was run using parameters $nr\_iter = 1000$ and $nr\_samples = 10$. These values were chosen based on the number of classes and their sizes, albeit no parameter tuning was applied. In general, a high value of nr_iter and a reasonably small value of $nr\_samples$ are recommended. Ties were broken by sorting residue positions with equal score in increasing sequence position.

2.5 Benchmark Studies

The performance of a method may depend on the type of protein family and functional specificity properties considered. We therefore carried out a benchmark involving seven different protein families with various associated functional specificity properties (Table 2).

G Protein-Coupled Receptors (GPCRs) are integral cell membrane proteins involved in signal transduction. Their mediatory role makes them important drug targets (Gether *et al.*, 2002; Pierce *et al.*, 2002). We extracted the MSA of *class A* GPCRs in the transmembrane region from the latest version of the GPCRDB (Horn *et al.*, 2003, June 2006 release (10.0), www.gpcr.org/7tm), yielding a MSA of 2065 protein sequences with an average identity of 26% over all sequence pairs in the alignment. The MSA was classified into 77 subfamilies according to the recognition of endogenous ligands. An additional reduced MSA was derived by applying a redundancy limit of 65% identity, and subsequently discarding all subfamilies that had only one sequence remaining. This yielded a MSA of 190 protein sequences divided over 39 GPCR families, which was named 'GPCR-190'. Residues are deemed to be ligand binding whenever their mutation affects ligand binding in aminergic receptors, as listed in Table 2.

The LacI family is one of the largest families of bacterial transcription factors. This family was analysed by Mirny and Gelfand, 2002 using a technique based on mutual information. We used a multiple sequence alignment of 54 LacI protein sequences Mirny and Gelfand, 2002 classified into 15 families. Suckow *et al.*, 1996 mutated positions 2 to 329 of Lac repressor into 12 or 13 of the 20 natural occurring amino acids. These 4000 well-defined mutants yielded a functional classification for each position. We took the residues in group IX (DNA binding) and XI (IPTG binding) as the specificity residues. Some of these are actually conserved in the alignment and thus cannot contribute to specificity. These were subsequently excluded from the selection. The resulting 28 specificity determining residues are listed in Table 2.

The Ras superfamily of small GTP-ases is implicated in the regulation of growth, survival, differentiation and other processes in haematopoietic cells (Reuther and Der, 2000). It comprises six families, of which experimental evidence for functional sites was available from the literature for the Rab 5 versus Rab 6 subfamilies, and the Ras versus Ral families, as defined in Pirovano *et al.*, 2006. The 28 and 12 true positives, respectively, are listed in

Table 2. Properties of the datasets used for testing the algorithms.

dataset	nr of classes	avg (std) class size	max, min class size	nr of sites	site information	'true' sites
GPCR	77	26.8 (34)	189, 3	214	ligand	T94, T97, E113, G114, A117, T118, G121, L125, C167, L172, F203, V204, M207, F208, H211, Y268, A269, A272, A292, F293, K296
GPCR-190	39	4.9 (3.8)	21, 2			
LacI	15	3.6 (2.5)	12, 2	339	ligand & DNA	T5, L6, S16, Y17, Q18, R22, N25, Q26, H29, Q54, A57, S61, L73, A75, P76, I79, N125, P127, D149, S191, S193, W220, N246, Q248, Y273, D274, T276, F293
Ras/Ral	2	44.5 (24.5)	69, 20	218	protein	I24, Q25, D30, E31, D33, I36, E37, Q43, L53, M67, Q70, D92
Rab5/Rab6	2	5.0 (1)	4, 6	163	protein	K21, G22, Q23, H25, E26, F27, Q28, E29, S30, H62, A65, M67, Y69, G71, A72, Q73, E96, L97, Q98, R99, Q100, A101, S102, P103, N104, I105, V106, K162
AQP/GLP	2	30.0 (18)	48, 12	430	protein	L21, W48, V52, A65, H66, L67, V71, T137, Y138, P139, N140, P141, L159, I163, I187, G195, P196, L197, G199, F200, A201, M202
Smad	2	10.0 (2)	12, 8	211	protein	L263, Q264, T267, Q284, Q294, P295, L297, T298, S308, E309, A323, V325, M327, I341, F346, P360, Q364, R365, Y366, W368, N381, R427, T430, S460, V461, R462, C463, M466

Table 2. The MSAs of 4 Rab5 and 6 Rab6, and of 20 Ras and 69 Ral protein sequences described in Pirovano *et al.*, 2006 were used.

The Major Intrinsic Protein (MIP) family of Integral Membrane Transporters is mainly involved in facilitating the transport of both water and small neutral solutes through the cellular membrane in all domains of life. There are about six MIP subfamilies, the two major are the aquaporins (AQPs) and the glycerol-uptake facilitators (GLPs) (Zardoya and Villalba, 2001). The MSA of 12 AQP and 48 GLP protein sequences described in Pirovano *et al.*, 2006 was used. Residues with at least one atom closer than 7 Å to the bound glycerol molecules in the GLP pore channel in the crystal structure 1FX8 (Fu *et al.*, 2000), excluding those that were conserved in the training set of sequences, as defined in Pirovano *et al.*, 2006. This yields a set of 37 sites, which are listed in Table 2.

The Smad family of TGFβ-associated transcription factors plays a crucial role in the transforming growth factor-β signalling pathway and is critical for determining the specificity between alternative pathways (Feng and Derynck, 2005; Massague *et al.*, 2005). The family can be subdivided into two major classes: AR-Smads that are mainly induced by TGFβ-type receptors, and BR-Smads that are mainly induced by the BMP-type receptors. The MSA of 8 AR-Smad and 12 BR-Smad non-redundant sequences of the Smad-MH2 domain described in Pirovano *et al.*, 2006 was used. The 29 specificity determining residues as defined in Pirovano *et al.*, 2006 are listed in Table 2.

2.6 Evaluation of the Algorithms' Performance

The Receiver-operator characteristic (ROC) curve is used for testing the performance of an algorithm for separating true and false positives (Swets, 1988; Provost and Kohavi, 1998). Here known functional specificity residues are considered true positives. The remaining residues are considered true negatives. We use the scoring (weight) values as threshold for generating the ROC curve. For each weight value v the set of residues with weight higher than or equal to v is considered: the true positive percentage is reported on the y-axis (sensitivity, or coverage), and the false positive percentage (1−specificity, or error) on the x-axis. The ROC curve thus describes the goodness of a method in giving higher ranking to the *given* functionally important residues.

3 RESULTS

From the ROC curves in Figure 1A, D, the results of our multi-RELIEF method appear superior to the other methods over all the datasets. The addition of 3D contacts yields a clear improvement for the GPCRs and LacI family, as is shown in the 'weights' plots in Figure 1B, E. This is more evident for the GPCRs, for reasons explained in the next section.

The Smad, Ras/Ral and Rab5/Rab6 datasets contain two classes, which are rather balanced. In this case nearly all algorithms achieve high performance, but some variations are still observable (see Figure 2, suppl. inf.). In general, the distributions of true positives with respect to the computed scoring weights are similar for Smad, GPCR and Ras/Ral, and for Rab5/Rab6 to somewhat lesser extent. The true positives occur in the upper part of the curve, that is, they satisfy the multi-RELIEF condition of being locally specific.

For the LacI and MIP datasets the situation is slightly different. Here, the majority of true positives also occurs on the upper part of the curve, but some are retained at the central or lower parts of the curve. Clearly, some of the LacI true positives do not conform to the model of local specificity exploited by multi-RELIEF. Upon detailed examination, these sites turn out to be largely conserved, a-like positions, as discussed further below.

The overall performance of the methods can be captured by the area under the curve (AUC) in the ROC plots, as listed in Table 3. Here we observe that in four of the datasets, multi-RELIEF or multi-RELIEF + 3D contacts is the best-scoring algorithm. In two others, they are not far below the best. A notable exception is the GPCR-190 reduced set, for reasons that are explained below. Importantly, the other methods are top-scoring only in at most one single dataset. The average scores over all datasets, in the last column of Table 3, also shows that multi-RELIEF and multi-RELIEF + 3D contacts are the top-scoring methods, with a consistent but modest lead for multi-RELIEF + 3D contacts.

4 DISCUSSION

4.1 Evolving Specificity Residues

It is well accepted in sequence analysis that conserved residues are likely to be functionally important. Indeed, many early approaches select functional sites by simply picking the most conserved positions in a given MSA. Since vast amounts of sequence data have become available, sequence comparison between paralogous and orthologous proteins is performed routinely in order to identify specificity residues that account for differences between functional subgroups. Most state-of-the-art approaches for functional specificity detection require a MSA with predefined functional classes. They then forward MSA positions conserved within each group but different between groups as functionally

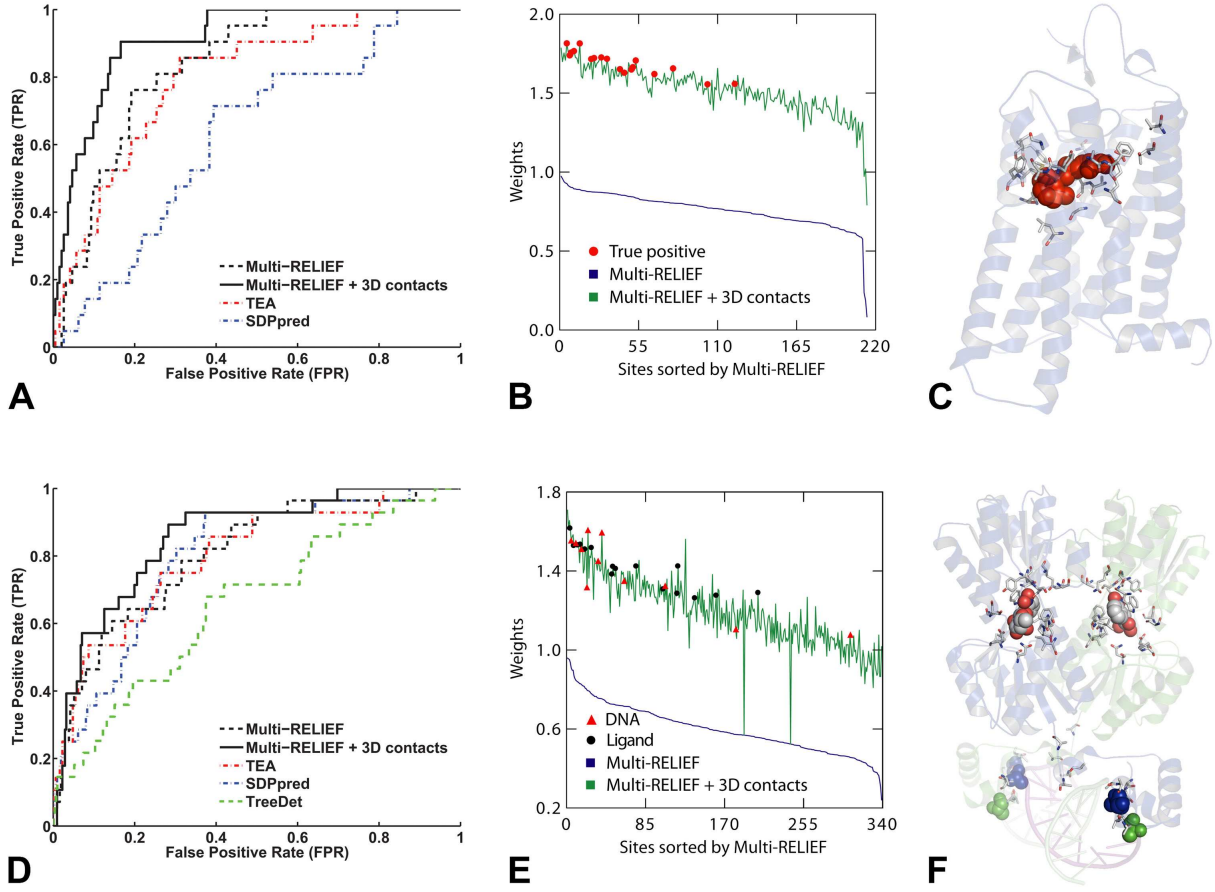

Fig. 1. Results for all methods on the GPCR (**A**-**C**) and LacI (**D**-**F**) datasets. In columns are the ROC curves (**A**, **D**); the weights assigned by multi-RELIEF without (blue) and with (green) 3D contacts, and true positives highlighted with symbols (**B**, **E**); and the respective protein structures with true positive residues in sticks, and ligands in space filling balls (red for GPCR, **C**, and atom colours for LacI, **F**). Note that TreeDet could not be applied to the GPCR dataset (**A**) due to its size ($>$ 200 sequences). For LacI the residues S21 and A27 mentioned in the text are highlighted as blue spheres (**F**).

Table 3. Area under curve for the ROC plots of the six methods and seven datasets, and average scores relative to multi-RELIEF over the common datasets (best scores in bold). Average scores per dataset are also given.

method	GPCR	GPCR 190	LacI	Rab5/ Rab6	Ras/ Ral	AQP/ GLP	Smad	Rel. Avg.
multi-RELIEF	.83	.78	.80	**.90**	**.97**	.83	.97	0
+ 3D	**.91**	.84	**.85**	.86	.91	**.84**	.96	**+.003**
TEA	.80	.89	.80	.79	.86	**.84**	.96	−.039
SH	—	—	—	.86	.95	.75	**.98**	−.033
SDPpred	.63	**.90**	.80	.83	.96	.78	.84	−.058
TreeDet/MB	—	.63	.66	.85	.92	.79	.96	−.073
Average	.79	.81	.78	.85	.93	.81	.95	

specific. However, different degrees of specificity may be relevant. For example, position c in the Table 1 toy example provides a perfect explanation of such subdivision in classes. Although position b is insufficient for differentiating all four classes, it does provide some information about the difference between $C1$, $C2$ and $C3$, $C4$.

The specificity residues considered in this study include c-like positions, that are fully class-specific, but also partially class-specific b-like positions are present, especially in the GPCR case study. The following evolutionary scenario can explain this observation. After proteins 'learn' how to fold correctly in order to perform their main function, they can start evolving new functional sites in order to interact with other components such as small molecules, DNA, RNA or another protein. Such a process can be conducted in a stepwise fashion, first by establishing general interaction anchor points (conserved, a-like positions), next by evolving to more selective recognition sites (specific, b- and c-like). For example, if proteins want to interact with DNA, they first evolve some positively charged residues in a certain region of the protein just to attract the negatively charged phosphoric acid group(s) of DNA. They then can evolve b-like positions to selectively bind to a specific category of DNA and finally, they can obtain c-like positions to achieve specific recognition of a particular DNA fragment.

Functional specificity sites can therefore contain different types of specificity positions. The proportions of *a*-, *b*- and *c*-like positions (see Table 1) may vary within different protein families. In our benchmark studies, for the GPCRs, Smad and LacI datasets, we defined all residues at the specificity interaction interface according to the experimental evidence and excluded *a*-like. Such definition is straightforward but results in *c*- and *b*-like positions being taken as "true" positives. If a family contains a high percentage of *c*-like positions, methods focussing on intra-group conservation will all perform well, while a more varying performance is likely with larger proportions of *b*-like positions.

4.2 Using 3D contacts

Although multi-RELIEF attains similar or better performance than its counterparts considered here, we have demonstrated that specificity detection can be further enhanced by taking 3D-contact information into account (Figure 1A, D). In this scenario, the score of a residue position will be boosted if its neighboring residues score high, introducing a bias towards spatially clustered residues. Depending on the ligand being a small molecule or a larger protein or DNA structure, employing contact information may affect predictions differently. If the ligand is a small moiety, the specificity residues form a small, compact cavity, such that application of 3D contacts improves prediction. On the other hand, interaction interfaces to large protein or DNA ligands will generally be larger and more planar, often leading to relatively few isolated interface residues providing specificity recognition. This renders 3D contacts potentially less beneficial for datasets associated with proteins interacting with larger ligands.

4.3 Benchmark Performance

4.3.1 GPCRs On the GPCRs dataset, all algorithms except SDPpred perform well. The GPCR ligand binding site is illustrated by retinal, the endogenous ligand of bovine rhodopsin in Figure 1C. Multi-Relief outperforms the other methods substantially, and the use of structure information (multi-Relief + 3D contacts) further improves its performance. There are two factors that contribute to these observations. First, there are 77 subfamilies in the GPCRs dataset which cannot be uniquely differentiated by a single position using the 20 natural amino acids. Thus, in the absence of absolute *c*-like positions, *b*-like positions are the best alternative. This gives multi-Relief an obvious advantage in identifying *b*-like positions. Second, the *class A* GPCRs evolved to recognize small molecules so that the specificity site is relatively compact and concentrated in a small region of the protein compared to other protein families that recognize DNA, RNA or protein (Figure 1C). This also explains the relatively large performance increase, compared to the other datasets, of multi-Relief when 3D contacts are used for boosting results for the GPCR dataset.

For the reduced GPCR-190 set, average AUC of all methods is similar to that of the full GPCR set, see Table 3 (also Figure 2, suppl. inf.). Intriguingly, only multi-RELIEF performs similarly over both datasets, while all other methods perform differently. Multi-RELIEF + 3D contacts gives a much smaller improvement over multi-RELIEF than in the full GPCR set, but more strikingly, the performance of TEA and SDPpred are higher. An explanation can be found in the 65% redundancy threshold applied. This retains diversity within a subfamily, i.e. the most divergent members, but multi-RELIEF relies on differences between *nearest neighbors*, which could be entirely different in the reduced set. Even the 3D information apparently cannot overcome this. TEA and SDPpred, on the other hand, put more emphasis on entropy to measure the overall differences between the subfamily, which may be more pronounced in the reduced set.

4.3.2 LacI Results on the LacI dataset highlight the difference between specificity related binding to small molecules compared to binding DNA. LacI transcription factors bind to particular DNA fragments to prevent transcription of downstream genes. After recognition of ligands specific for each subfamily, they change conformation so that RNA polymerase is no longer blocked from binding to DNA. This leads to high expression of the encoded proteins. As illustrated in Figure 1E multi-RELIEF generally assigns higher weights to residues that recognize the small molecule than those binding to DNA. Moreover, application of the 3D contacts option boosts the weights of these residues.

Figure 1F shows the structure of the transcription factor of LacI. Among the small molecule binding sites, the position of R196 has low weight, even after being boosted by means of the 3D contacts step. When looking at its residue composition (data not shown), we can see that R196 is a *b*-like position, since amino acid R occurs in 47 out of a total of 54 protein sequences.

For this dataset, the 3D contacts information does not notably improve the detection of the DNA-binding residues, but, importantly, the prediction quality also does not suffer from the 3D contacts. The limited added value may be due to the fact that the DNA binding site is much bigger and more extended than the binding site for small molecules. Thus, interaction between protein and DNA may include several relatively isolated and spatially separated locations. For example, residue S16 interacts with DNA and indeed is assigned a high weight since it contributes to specific recognition. However, neighboring residues do not interact with the DNA and have low weight, so the score of S16 becomes worse after application of the 3D contacts step.

In addition, we identified a specific region of the protein where two residues, S21 and A27 are close to each other and have high weights before and after application of 3D contacts. Although these two residues were not characterized as DNA binding by Suckow *et al.*, 1996, they are located within 5 Å distance to the DNA.

4.3.3 Ras The two datasets from the Ras family are based upon mutation experiments, three regions of about 10 residues each for Rab5 vs. Rab6, and 12 point mutations for Ras vs. Ral and Rab. They show best performance for multi-RELIEF and worst for TEA, while other methods perform very similar and only slightly below multi-RELIEF, see Table 3 (and Figure 2, suppl. inf. for more details). Overall performance of all methods is lower for Rab5/Rab6 than for Ras/Ral (Table 3).

Although specificity in the Ras superfamily is related to recognition of various small-molecule and protein targets, multi-RELIEF is well able to recognize these sites. However, due to the presence of multiple interacting sites (see Figure 2, suppl. inf.), addition of 3D contacts information does not lead to a gain in detection of specificity residues.

4.3.4 MIP The MIP dataset is based on a structural definition of functional residues: those close to the ligand in the crystal structure. Overall performance of all methods is relatively low (see Table 3. Multi-RELIEF + 3D contacts and TEA together are

the best-scoring methods. Importantly, multi-RELIEF and multi-RELIEF + 3D contacts show the steepest initial slope in the ROC curves (Figure 2, suppl. inf.), which is relevant for experimental planning if only top-scoring sites are to be examined.

4.3.5 Smad The Smad dataset is a special benchmark because the true positive residues have been verified directly by site-directed mutagenesis experiments. It is different from the other datasets in that it contains two classes. The known functional specificity sites are a mix of *b*/*c*-like positions, i.e., specific and conserved in each class, and *d*-like positions, that are specific but not conserved within the classes.

The performance of all methods on the Smads is remarkably good, compared to the other datasets, see Table 3 and Figure 2, suppl. inf. (note the difference in scale of the FPR axis). This is likely due to the comprehensive experimental coverage of true functional Smad sites, reducing the proportion of false negatives and increasing overall performance by all methods. The 3D-contact step results in a slightly decreased performance of multi-Relief. This may be due to the fact that three different functional interactions are involved, each involving distinct interaction interfaces on the Smad protein surface.

5 CONCLUSION

In this study, we proposed a novel multi-Relief algorithm for identifying specificity-related functional sites. We provided an option for boosting prediction quality using structural information, if available,for specificity of interaction with small molecules. We tested the performance of multi-RELIEF and other recent algorithms on seven different experimental benchmark cases. The results demonstrate robustness and best overall performance of multi-RELIEF over a wide variety of biological cases.

ACKNOWLEDGEMENTS

We thank our student Bernhard Kammleitner for implementing the multi-RELIEF web server. We also thank Leonid A. Mirny and Mikhail S. Gelfand for making their LacI dataset available for this study. This work was financially supported by the NWO-Bioinformatics Breakthrough Project (050-71-047), The Netherlands.

REFERENCES

Bickel, P., Kechris, K., Spector, P., Wedemayer, G., and Glazer, A. (2002). Finding important sites in protein sequences. *Proc. Natl Acad. Sci. USA*, **99**, 14764–71.

Carro, A., Tress, M., de Juan, D., Pazos, F., Lopez-Romero, P., Del Sol, A., Valencia, A., and Rojas, A. (2006). Treedet: a web server to explore sequence space. *Nucleic Acids Res.*, **35**(Web Server Issue), 99.

Del Sol Mesa, A., Pazos, F., and Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J Mol Biol*, **326**(4), 1289–1302.

Feenstra, K., Pirovano, W., Krab, K., and Heringa, J. (2007). Sequence harmony: detecting functional specificity from alignments. *Nucleic Acids Res.*, **35**(web server issue), W495–W498.

Feng, X. and Derynck, R. (2005). Specificity and versatility in TGF-beta signaling through Smads. *Annu Rev Cell Dev Biol*, **21**, 659–93.

Fu, D., Libson, A., Miercke, L., Weitzman, C., Nollert, P., Krucinski, J., and Stroud, R. (2000). Structure of a glycerol-conducting channel and the basis for its selectivity. *Science*, **290**(5491), 481–6.

Gether, U., Asmar, F., Meinild, A., and Rasmussen, S. (2002). Structural basis for activation of g-protein-coupled receptors. *Pharmacol Toxicol*, **91**, 304–312.

Gu, X. (2006). A simple statistical method for estimating type-ii (cluster-specific) functional divergence of protein sequence. *Mol Biol Evol.*, **23**, 1937–45.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

Hannenhalli, S. and Russell, R. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol*, **303**(1), 61–76.

Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F., and Vriend, G. (2003). Gpcrdb information system for g protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.

Kalinina, O., Novichkov, P., Mironov, A., Gelfand, M., and Rakhmaninova, A. (2004). SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res*, **32**(Web Server issue), W424–8.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. In Springer, editor, *European Conference on Machine Learning*, volume LNCS 784, pages 171–182.

Landgraf, R., Xenarios, I., and Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteinsr. *J. Mol. Biol.*, **307**, 1487–1502.

Marchiori, E., Pirovano, W., Heringa, J., and Feenstra, K. (2006). A feature selection algorithm for detecting subtype specific functional sites from protein sequences for smad receptor binding. In *Fifth International Conference on Machine Learning and Applications*, pages 168–173. IEEE.

Massague, J., Seoane, J., and Wotton, D. (2005). Smad transcription factors. *Genes Dev*, **19**(23), 2783–810.

Mihalek, I., Res, I., and Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol*, **336**(5), 1265–1282.

Mirny, L. and Gelfand, M. (2002). Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol*, **321**(1), 7–20.

Pierce, K., Premont, R., and Lefkowitz, R. (2002). Seven-transmembrane receptors. *Nat Rev Mol Cell Biol*, **3**, 639–650.

Pirovano, W., Feenstra, K., and Heringa, J. (2006). Sequence comparison by sequence harmony identifies subtype specific functional sites. *Nucleic Acids Res.*, **34**, 6540–48.

Provost, F. and Kohavi, R. (1998). Guest editors' introduction: On applied research in machine learning. *Machine Learning*, **30**, 127–132.

Reuther, G. and Der, C. (2000). The Ras branch of small GTPases: Ras family members don't fall far from the tree. *Curr Opin Cell Biol*, **12**(2), 157–65.

Robnik-Sikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, **53**(1-2), 23–69.

Shenkin, P., Erman, B., and Mastrandrea, L. (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins*, **11**(4), 297–313.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E., and Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.

Suckow, J., Markiewicz, P., Kleina, L., Miller, J., Kisters-Woike, B., and Müller-Hill, B. (1996). Genetic studies of the lac repressor. xv: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J Mol. Biol.*, **261**(4), 509–23.

Sun, Y. and Li, J. (2006). Iterative relief for feature weighting. In *International Conference on Machine Learning*, pages 913–920. ACM.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Whisstock, J. and Lesk, A. (2003). Prediction of protein function from protein sequence and structure. *Quart Rev Biophys*, **36**(3), 307–340.

Ye, K., Lameijer, E., Beukers, M., and IJzerman, A. (2006). A two-entropies analysis to identify functional positions in the transmembrane region of class a g protein-coupled receptors. *Proteins*, **63**, 1018–30.

Zardoya, R. and Villalba, S. (2001). A phylogenetic framework for the aquaporin family in eukaryotes. *J Mol Evol*, **52**(5), 391–404.