Multi-RELIEF: specificity-determining residues from alignments by Machine Learning and Feature Weighting

K. Anton Feenstra[2], Kai Ye[1], Jaap Heringa[2], Adriaan P. IJzerman[1], Elena Marchiori[2]

[1] *Div. of Med. Chem., LACDR, Leiden Univ., Leiden, NL* [2] *Dept. of Comp. Sci., IBIVU, Free Univ., Amsterdam, NL*

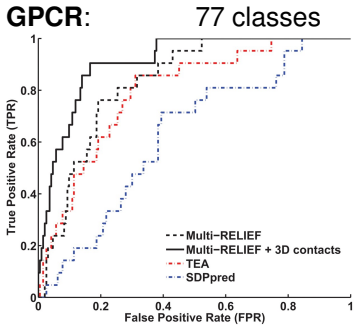
Contact: feenstra@few.vu.nl, k.ye@lacdr.leidenuniv.nl, heringa@few.vu.nl, elena@few.vu.nl, ijzerman@lacdr.leidenuniv.nl Download this poster at: www.few.vu.nl/~feenstra/posters.html

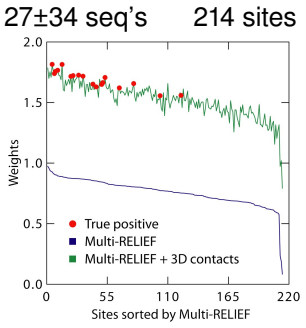
Multi-RELIEF

We introduce multi-RELIEF, a new algorithm for identifying specificity-determining residues in proteins from an alignment and predefined multiple classes. 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 two classes [1], and is complementary to previous method developed by us, Sequence Harmony [2] and Two Entropies [3]. Multi-RELIEF can be used through a web-interface at `www.ibi.vu.nl/programs/multirelief/`.

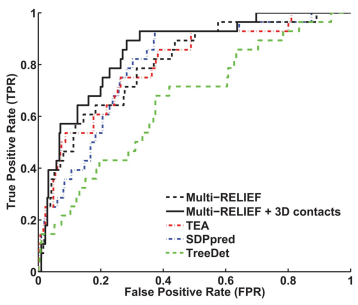
Benchmark data

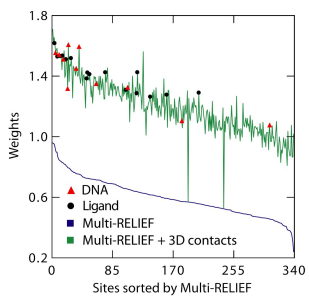
	ROC plots	weight plots	structures
GPCR:	77 classes	27±34 seq's	214 sites

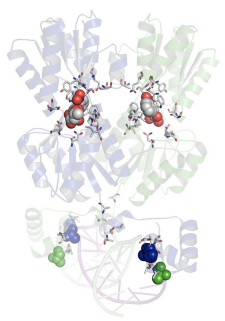




LacI:	15 classes	3.6±2.5 seq's	339 sites







Smad:	2 classes	10±2 seq's	211 sites
GPCR-190:	39 classes	5±4 seq's	214 sites
AQP/GLP:	2 classes	30±18 seq's	430 sites
Rab5/Rab6:	2 classes	5±1 seq's	163 sites
Ras/Ral:	2 classes	44±25 seq's	218 sites

Discussion & Conclusion

- **Multi-RELIEF outperforms the other methods.** The ROC plots and areas under the curve show multi-RELIEF to be superior to the other methods in five out of seven datasets. Overall best performance is obtained by Multi-RELIEF + 3D.
- **Inclusion of 3D contacts improves the predictions.** Maximum gain is seen in GPCR and LacI datasets, which have specificity sites defined around ligand-binding areas that are local in the structure. In the other datasets, specificity is related to protein-protein interactions (Smad and Ras superfamily) or (membrane) channel function (AQP/GLP).
- **Complex cases of specificity can be handled.** The GPCR family contains 77 classes; with 20 aminoacids no single site could be strictly specific for all classes. The high performance of multi-RELIEF shows its strength in identifying sites that are specific for only a subset of classes.
- **Multi-RELIEF is sensitive to the variation within classes** and captures subtle evolutionary signals from the more redundant sequences. This may explain the lower performance on the GPCR-190 dataset which was pruned by a 65% redundancy threshold.

Concluding, the use of feature weighting by multi-RELIEF incorporates conservation locally in sequence space and can capture correlations between different sites in a protein sequence. The benchmark cases shown here indicate the strength and robustness of this approach for specificity prediction in proteins.

Areas under curve | in the ROC plots, averages per method relative to multi-RELIEF, and per dataset (best in bold).

method	GPCR	GPCR 190	LacI	Rab5/ Rab6	Ras/ Ral	AQP/ GLP	Smad	Avg.
multi-RELIEF	0.83	0.78	0.80	**0.90**	**0.97**	0.83	0.97	0.000
+3D	**0.91**	0.84	**0.85**	0.86	0.91	**0.84**	0.96	**0.003**
TEA	0.80	0.89	0.80	0.79	0.86	**0.84**	0.96	-0.039
SH	–	–	–	0.86	0.95	0.75	**0.98**	-0.033
SDPpred	0.63	**0.90**	0.80	0.83	0.96	0.78	0.84	-0.058
TreeDet/MB	–	0.63	0.66	0.85	0.92	0.79	0.96	-0.073
Average	0.79	0.81	0.78	0.85	0.93	0.81	0.95	

Method

Toy example | and weights computed by multi-RELIEF

	a	*b*	*c*	*d*	*e*
C1	R	F	T	I	T
	R	F	T	Q	F
	R	F	T	N	V
	R	F	T	A	D
C2	R	F	Y	S	T
	R	F	Y	F	F
	R	F	Y	D	V
	R	F	Y	V	D
C3	R	Y	D	E	T
	R	Y	D	V	F
	R	Y	D	W	V
	R	Y	D	G	D
C4	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

Pseudo-code | of the multi-RELIEF algorithm

```
%input: X_1,..,X_m
%  (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 the
          two selected classes
    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
```

The input of multi-RELIEF is a multiple alignment of a protein family and a subdivision into groups. The groups are considered classes and the sites features. In multi-RELIEF, multiple runs of RELIEF are performed on pairs of classes. At each run *i*, a number of sequences are randomly selected from two randomly selected classes, and RELIEF is applied to the resulting two classes. From the output vectors W_i, the positive weights of each position are averaged to yield the final multi-RELIEF output vector with feature weights. 3D structural information can be used to increment multi-RELIEF weights with the average weight of all 3D neighbors (non-sequential and have shared surface area according to the 'CMA' method [4]).

References

[1] Ye, K, KA Feenstra, J Heringa, AP IJzerman, E Marchiori "*Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine Learning approach for feature weighting*", Bioinformatics, in press (2007).

[2] Feenstra, KA, W Pirovano, K Krab & J Heringa "*Sequence Harmony: Detecting Functional Specificity from Alignments*", Nucl. Acid. Res., 35: W495 2007.

[3] Ye, K, E Lameijer, M Beukers, and A IJzerman. "*A two-entropies analysis to identify functional positions in the transmembrane region of class a g proteincoupled receptors*" Proteins, 63, 1018–30 2006.

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

Leiden / Amsterdam Center for Drug Research

Leiden University / Vrije Universiteit Amsterdam

CENTRE FOR INTEGRATIVE BIOINFORMATICS VU

nbic Netherlands Bioinformatics Centre