BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–2

A structural alignment kernel for protein structures: Supplementary information

Jian Qiu[a], Martial Hue[c*], Asa Ben-Hur[d], Jean-Philippe Vert[c], William Stafford Noble[a,b]

[a]Department of Genome Sciences, [b]Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA, [c]Ecole des Mines de Paris—ParisTech, Paris, France, [d]Department of Computer Science, Colorado State University, Colorado, USA

1 TOPS KERNEL

To compare graphs representing two structures, Borgwardt et al. define a kernel k_w that compares *walks* on the two graphs. The final kernel is a sum of all kernels for walks of different lengths.

To compare graphs representing two structures, Borgwardt et al. define a kernel k_w that compares *walks* on the two graphs. The final kernel is a sum of all the walk-kernels:

$$K(G_1, G_2) = \sum_{walk_1, walk_2} k_w(walk_1, walk_2)\,, \qquad (1)$$

where k_w assigns a value of 0 to walks of different lengths. Recall that a walk on a graph is a series of nodes $v_1, \ldots, v_k$ where $(v_i, v_{i+1}) \in E,\ i = 1, \ldots, k-1$. The walk kernel is defined between two walks of equal length: $walk_1 = u_1, \ldots, u_k : (u_i, u_{i+1}) \in E_1$, $walk_2 = v_1 \ldots, v_k : (v_i, v_{i+1}) \in E_2$ as:

$$k_w(walk_1, walk_2) = \prod_{i=1}^{k-1} k_s((u_i, u_{i+1}), (v_i, v_{i+1}))$$

where

$$k_s((u_i, u_{i+1}), (v_i, v_{i+1})) = k_n(u_i, v_i) k_e((u_i, u_{i+1}), (v_i, v_{i+1})) k_n(u_{i+1}, v_{i+1})$$

and k_n, k_e are kernels between edges and nodes, respectively. In practice, the walks need to be weighted inversely by length to ensure convergence, or a bound on walk length needs to be enforced.

Following Borgwardt et al., we consider two node and edge kernels:

- A type kernel: a 0,1 kernel where the kernel entry between two nodes/edges is 0 unless they have the same type.
- A kernel that takes into account the properties of the amino acids in a secondary structural element: average hydrophobicity (quantified by the Kyte-Doolittle index [Kyte and Doolittle, 1982]), average hydrophilicity (quantified by the Hopp-Woods index [Hopp and Woods, 1981]), fraction of hydrophobic, hydrophilic, positively charged, negatively charged, and cysteine residues. A linear kernel is applied to this vector of features.

The final node kernel is a product of the type kernel and the local amino acid property kernel. In order to compute the sum over all walks (Equation 1) we consider the product graph $G_1 \times G_2$ of graphs G_1 and G_2. The nodes and edges of the product graph are defined by: $V(G_1 \times G_2) = \{(u, v) : u \in V_1, v \in V_2\}$, $E(G_1 \times G_2) = \{((u, v), (u', v')) : (u, u') \in E_1, (v, v') \in E_2\}$. An edge $((u, v), (u', v')) \in E(G_1 \times G_2)$ is assigned a weight equal to $k_s((u, u'), (v, v'))$, and is summarized by the adjacency matrix $A_\times$. The kernel can now be expressed as:

$$K(G_1, G_2) = \sum_{i,j=1}^{|V(G_1 \times G_2)|} \left[\sum_{n=0}^{\infty} \lambda^n A_\times^n \right]_{ij} . \qquad (2)$$

The parameter $\lambda < 1$ ensures convergence of the sum. We used $\lambda = 0.5$.

Our implementation of the kernel differs from that of Borgwardt et al. in several ways. First, we define proximity of two structural elements using the C-α criterion described above, with a distance threshold of 10 Å. Borgwardt et al. likely used some other proximity measure to generate edges in their secondary structure graph. Second, in our graphs, edges between beta strands are labeled according to the type of contact they form: parallel or anti-parallel. In the edge kernel, walks along edges with incompatible orientation are not allowed. Also, in Equation 1, we sum up to the seventh power of the adjacency matrix, in order to avoid computing an inverse. The information on secondary structure content was extracted using DSSP [Kabsch and Sander, 1983].

*The first two authors contributed equally to this work

Table 1. Accuracy of vector kernel one-vs-one enzyme class classification. Each row lists the accuracy and balanced accuracy for SVM classifiers trained using a vector kernel. Columns labeled "Vec" are our own implementation, and columns labeled "D&D" contain data from Table 4 in [Dobson and Doig, 2005]. In the table, "Oxido" stands for "Oxidoreductase." The final column lists the optimal value of the radial basis parameter γ, selected via cross-validation.

Enzyme Classes		Acc. (%)		Bal. Acc.(%)		γ
		Vec	D&D	Vec	D&D	
Hydrolase	Isomerase	55.9	61.1	68.3	63.7	1
Hydrolase	Ligase	68.9	51.7	71.6	59.7	1
Hydrolase	Lyase	63.2	55.0	65.8	62.3	1
Hydrolase	Oxido	65.7	67.4	69.2	70.5	1
Hydrolase	Transferase	61.5	58.7	62.7	58.7	1
Isomerase	Ligase	71.8	64.8	50.0	66.4	0.01
Isomerase	Lyase	54.1	58.6	51.0	59.3	1
Isomerase	Oxido	61.5	73.8	65.2	74.0	1
Isomerase	Transferase	60.0	57.5	51.2	60.3	10
Ligase	Lyase	45.0	52.5	48.3	55.0	0.01
Ligase	Oxido	78.8	79.8	58.7	78.0	1
Ligase	Transferase	85.1	61.4	51.3	67.2	10
Lyase	Oxido	64.7	75.5	62.6	75.5	1
Lyase	Transferase	51.6	48.4	51.6	55.5	1
Oxido	Transferase	70.5	66.2	68.2	66.6	10

2 PRELIMINARY EC EXPERIMENTS

The enzyme classification benchmark was developed by Dobson and Doig [2005]. In that work, the authors perform 15 one-vs-one classifications, and report accuracy and balanced accuracy for each of the 15 tasks. To test whether our implementation matches that of Dobson and Doig, we replicated their experimental setup. For our vector kernel, three-fold cross-validations were performed. In each fold, a second five-fold cross-validation was performed within each training set to select an optimal value of $\gamma \in \{0.01, 0.1, 1, 10\}$ to apply to the test set. As in [Dobson and Doig, 2005], we use an asymmetric soft margin to account for differences in class sizes.

Table 1 compares the results of our implementation of the vector kernel with those reported in [Dobson and Doig, 2005]. The two sets of results show very little agreement. In terms of accuracy, the two vector kernel implementations differ in a non-systematic fashion, with our implementation outperforming the original implementation in 10 out of 15 cases. Conversely, for balanced accuracy, the original implementation outperforms our implementation in 10 out of 15 cases. In some cases, the difference in accuracy or balanced accuracy is quite large—up to 20%. We suspect that the observed difference in performance results from differences in the heuristic used to set the SVM soft margin parameter so as to optimize balanced accuracy. In subsequent experiments on this benchmark, our implementation of the vector kernel provides good performance.

Table 3. Features used in the vector kernel. Each protein is represented as a 55-element vector, with entries as described in the table. More details are given in the text.

1–20	Residue frequencies
21–40	Fraction of surface area, by residue type
41	Total number of residues
42	Total surface area
43	Surface-area-to-volume ratio
44	Fractal dimension
45–47	Secondary structure content
48–54	Presence of ATP, FAD, NAD, Ca, Cu, Fe, Mg
55	Number of disulfide bonds

We used the 15 one-vs-one classification tasks to select a value for the γ parameter of the radial basis kernel function. The final column in Table 1 lists the optimal $\gamma \in \{0.01, 0.1, 1, 10\}$ for each classification task. These values were selected via a single five-fold cross-validation on the entire data set. For 10 of the 15 two-class problems, $\gamma = 1$ yields the best balanced accuracy. We therefore used $\gamma = 1$ for all subsequent experiments.

3 RESULTS FROM FIVE-KERNEL SUMMATION ON GO BENCHMARK

On the SCOP benchmark, the sum of five non-MAMMOTH kernels achieves a mean ROC score of 0.929 and $\text{ROC}_{10\%}$ of 0.624. Based on a Wilcoxon signed-rank test, its ROC score is statistically better than the torsion, mismatch, contact and TOPS kernels, not significantly different from the MAMMOTH nearest neighbor classifier and vector kernel, and worse than the MAMMOTH and sum kernels. The $\text{ROC}_{10\%}$ performance is worse than the MAMMOTH, MAMMOTH NN, and sum kernel, and tied with torsion, mismatch and vector kernels. On the GO benchmark, the five-kernel sum kernel has a mean ROC score of 0.790 and $\text{ROC}_{10\%}$ of 0.350. This ROC performance is only statistically better than the torsion and TOPS kernels, and worse than both MAMMOTH methods, the sum, mismatch and vector kernels.

REFERENCES

K.M. Borgwardt, C. S. Ong, S. Schoenauer, S.V.N. Vishwanathan, A. Smola, and H-P. Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(Suppl. 1): i47–i56, 2005.

P.D. Dobson and A.J. Doig. Predicting enzyme class from protein structure without alignments. *Journal of Molecular Biology*, 345:187–199, 2005.

T. P. Hopp and K. R. Woods. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA*, 78:3824–3828, 1981.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157:105–132, 1982.

Table 2. EC benchmark results. The table lists for each kernel and each enzyme class, the mean and standard deviations of the $ROC_{10\%}$ scores for one-versus-all SVM classifiers trained using 3x5cv. The highest score in each column is indicated by boldface type.

$ROC_{10\%}$	Hydrolase	Isomerase	Ligase	Lyase	Oxidoreductase	Transferase	Enz-NEnz
MAMMOTH NN	0.05 ± 0.00	0.01 ± 0.01	0.16 ± 0.08	0.03 ± 0.01	0.07 ± 0.02	0.03 ± 0.01	0.00 ± 0.00
MAMMOTH	0.09 ± 0.01	0.03 ± 0.01	0.11 ± 0.04	0.06 ± 0.01	0.05 ± 0.00	0.09 ± 0.00	0.13 ± 0.01
vector	0.21 ± 0.01	0.09 ± 0.02	0.21 ± 0.04	0.06 ± 0.01	0.28 ± 0.02	0.10 ± 0.00	0.17 ± 0.02
torsion	0.09 ± 0.01	0.09 ± 0.01	0.06 ± 0.05	0.05 ± 0.01	0.07 ± 0.00	0.10 ± 0.03	0.13 ± 0.02
contact	0.20 ± 0.02	0.07 ± 0.04	0.20 ± 0.04	0.07 ± 0.01	0.18 ± 0.01	0.09 ± 0.02	0.20 ± 0.01
mismatch	0.20 ± 0.01	0.04 ± 0.02	0.13 ± 0.03	0.06 ± 0.02	0.17 ± 0.02	0.05 ± 0.01	0.13 ± 0.01
TOPS	0.09 ± 0.02	0.06 ± 0.02	0.02 ± 0.03	0.04 ± 0.01	0.09 ± 0.01	0.09 ± 0.01	0.15 ± 0.01
sum	0.21 ± 0.01	0.11 ± 0.03	0.23 ± 0.06	0.07 ± 0.02	0.17 ± 0.01	0.08 ± 0.00	0.18 ± 0.01

Table 5. GO benchmark results. The table lists, in each row, a GO term and the corresponding mean $ROC_{10\%}$ scores generated by the various SVM kernels and by the MAMMOTH nearest neighbor classifier. Values in boldface are the maximum among the six individual kernels. Asterisked values are the maximum among all eight methods. Rows are sorted by the difference between MAMMOTH and mismatch SVM mean ROC scores.

GO term		MAMMOTH	mismatch	contact	vector	random walk	torsion	sum	MAMMOTH NN
GO:0005215	MF transporter activity	**0.553***	0.306	0.339	0.358	0.156	0.225	0.514	0.253
GO:0005125	MF cytokine activity	**0.911***	0.688	0.421	0.547	0.178	0.181	0.882	0.847
GO:0016788	MF hydrolase activity, acting on ester bonds	**0.503***	0.330	0.271	0.097	0.114	0.032	0.370	0.494
GO:0016491	MF oxidoreductase activity	**0.683**	0.544	0.331	0.558	0.131	0.181	0.633	0.686*
GO:0006468	BP protein amino acid phosphorylation	**0.788**	0.679	0.383	0.268	0.229	0.173	0.738	0.818*
GO:0009058	BP biosynthesis	0.370	0.272	0.288	0.199	0.304	**0.409**	0.449*	0.315
GO:0006810	BP transport	**0.422**	0.352	0.294	0.162	0.197	0.129	0.300	0.459*
GO:0006508	BP proteolysis and peptidolysis	**0.815**	0.752	0.636	0.516	0.512	0.555	0.736	0.815*
GO:0004175	MF endopeptidase activity	**0.775**	0.713	0.643	0.365	0.532	0.511	0.803*	0.731
GO:0003700	MF transcription factor activity	**0.535***	0.493	0.382	0.419	0.325	0.237	0.529	0.478
GO:0005509	MF calcium ion binding	**0.633**	0.600	**0.633**	0.511	0.311	0.167	0.756	0.767*
GO:0017076	MF purine nucleotide binding	**0.476**	0.457	0.313	0.251	0.310	0.199	0.434	0.483*
GO:0016021	CC integral to membrane	**0.517***	0.503	0.461	0.164	0.222	0.172	0.475	0.464
GO:0006351	BP transcription, DNA-dependent	**0.471**	0.462	0.387	0.403	0.137	0.191	0.494	0.516*
GO:0007165	BP signal transduction	0.344	**0.349***	0.185	0.159	0.202	0.080	0.247	0.313
GO:0008236	MF serine-type peptidase activity	0.911	**0.922**	0.867	0.811	0.733	0.633	0.944*	0.922
GO:0004713	MF protein-tyrosine kinase activity	0.717	**0.740***	0.606	0.222	0.235	0.441	0.702	0.571
GO:0005634	CC nucleus	0.535	**0.590**	0.430	0.269	0.217	0.327	0.534	0.608*
GO:0005737	CC cytoplasm	0.456	0.511	**0.536**	0.414	0.292	0.333	0.603*	0.522
GO:0004888	MF transmembrane receptor activity	0.485	**0.577***	0.470	0.468	0.416	0.348	0.557	0.536
GO:0043234	CC protein complex	0.383	**0.494***	0.369	0.215	0.403	0.256	0.419	0.489
GO:0007596	BP blood coagulation	0.737	**0.867***	0.859	0.610	0.640	0.646	0.836	0.479
GO:0045449	BP regulation of transcription	0.310	**0.560***	0.448	0.421	0.230	0.219	0.464	0.422

Table 4. Class sizes for the Gene Ontology benchmark. For each GO term in the benchmark, the table lists the ontology from which it comes (MF = molecular function, CC = cellular compartment and BP = biological process), and the number of positive (N_+) and negative (N_-) examples associated with the term.

GO ID	Ont	GO term	N_+	N_-
GO:0005634	CC	nucleus	149	103
GO:0043234	CC	protein complex	60	181
GO:0006508	BP	proteolysis and peptidolysis	59	178
GO:0006810	BP	transport	59	178
GO:0007165	BP	signal transduction	58	175
GO:0003700	MF	transcription factor activity	57	172
GO:0006468	BP	protein amino acid phosphorylation	55	166
GO:0004175	MF	endopeptidase activity	48	145
GO:0006351	BP	transcription, DNA-dependent	43	130
GO:0005215	MF	transporter activity	40	121
GO:0005737	CC	cytoplasm	40	121
GO:0016491	MF	oxidoreductase activity	40	121
GO:0017076	MF	purine nucleotide binding	37	112
GO:0045449	BP	regulation of transcription	37	112
GO:0005125	MF	cytokine activity	36	109
GO:0007596	BP	blood coagulation	36	109
GO:0004713	MF	protein-tyrosine kinase activity	35	106
GO:0004888	MF	transmembrane receptor activity	33	100
GO:0009058	BP	biosynthesis	33	100
GO:0016021	CC	integral to membrane	32	97
GO:0016788	MF	hydrolase activity, acting on ester bonds	31	94
GO:0005509	MF	calcium ion binding	30	91
GO:0008236	MF	serine-type peptidase activity	30	91

Table 6. Summary of SCOP and GO benchmark results for various kernels. The table lists for each kernel, the mean ROC and $\mathrm{ROC}_{10\%}$ scores across all classes in the SCOP and GO benchmarks. The highest score in each column (disregarding the sum kernel) is indicated by boldface type.

Kernel	SCOP		GO	
	ROC	$\mathrm{ROC}_{10\%}$	ROC	$\mathrm{ROC}_{10\%}$
MAMMOTH	**0.997**	**0.980**	**0.877**	0.580
torsion	0.899	0.530	0.722	0.289
contact	0.883	0.499	0.816	0.459
mismatch	0.885	0.542	0.853	0.555
vector	0.907	0.569	0.801	0.366
TOPS	0.883	0.555	0.741	0.305
sum	0.985	0.901	0.874	**0.583**