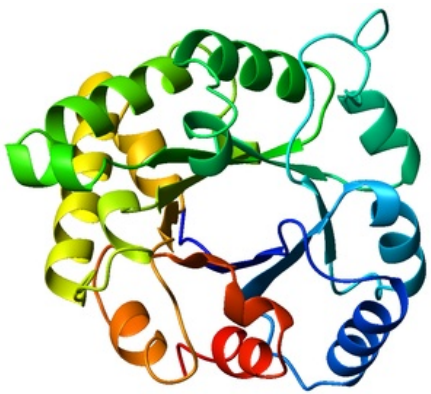
Probabilistic multi-class multi-kernel learning: On protein fold recognition and remote homology detection

Theodoros Damoulas

Supervisor: Prof. M. A. Girolami
Inference Group
Department of Computing Science
University of Glasgow

Work supported by NCR Financial Solutions Group Ltd.

- Predicting protein structural folds given protein characteristics when sequence similarity is limited ("twilight zone").



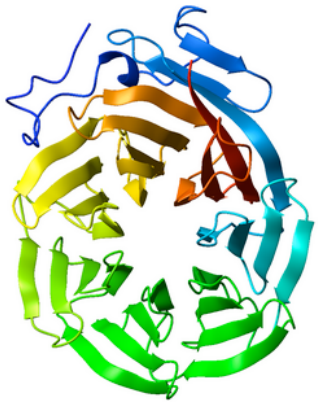


Figure: Tim-barrel 7-bladed beta-propeller

- Multi-class (≈1,000 folds) and Multi-feature (As many as 12 feature spaces have been proposed) problem.

Ding and Dubchak, Bioinformatics 2001 — Shen and Chou, Bioinformatics 2006

- Employ the PDB-40D dataset with 27 most populated SCOP folds, $\leq 35\%$ sequence similarity, 313/385 train/test split.

- 6 20-D **global** characteristics.
- ANNs and SVMs $S \times \frac{C(C-1)}{2}$=2,106 classifiers.
- Best performance **56 %**

- Replaced amino acid composition with 4 more F.Spaces: Pseudo-amino acid compositions (PseAA λ).
- Ad-hoc optimized ensemble of 9 base k-nn classifiers.
- Best performance **62.1 %**

Damoulas and Girolami, Bioinformatics 2008

- Composite Kernel regression via the Multinomial probit. Girolami and Rogers, (ICML 2005, NC 2006).
- Multi-kernel Multi-class pattern recognition machine. Best performance: **70 %.**
- Combine information from both **global** and **local** protein characteristics via the addition of *pairwise string* kernels.
- Full MCMC MH within Gibbs sampling solution and an efficient VB approximation for reduced CPU times.

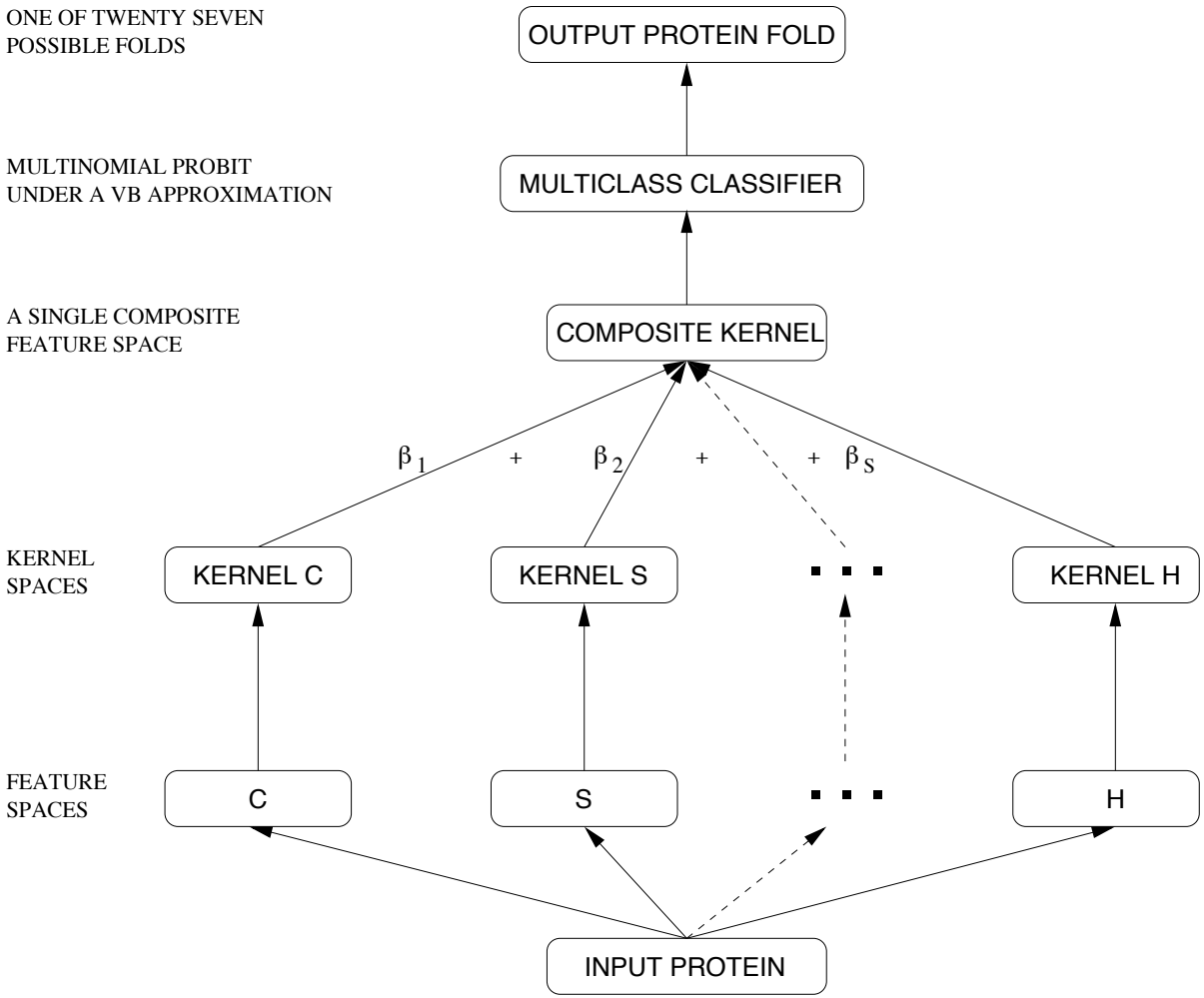


Figure: The intuition behind the approach.

- Hierarchical Bayesian model with auxiliary variables $\mathbf{Y}$.
- Classification via regression on an informatively constructed composite kernel. Inference on 3 levels, enhanced diagnostic ability and a solid Bayesian foundation.

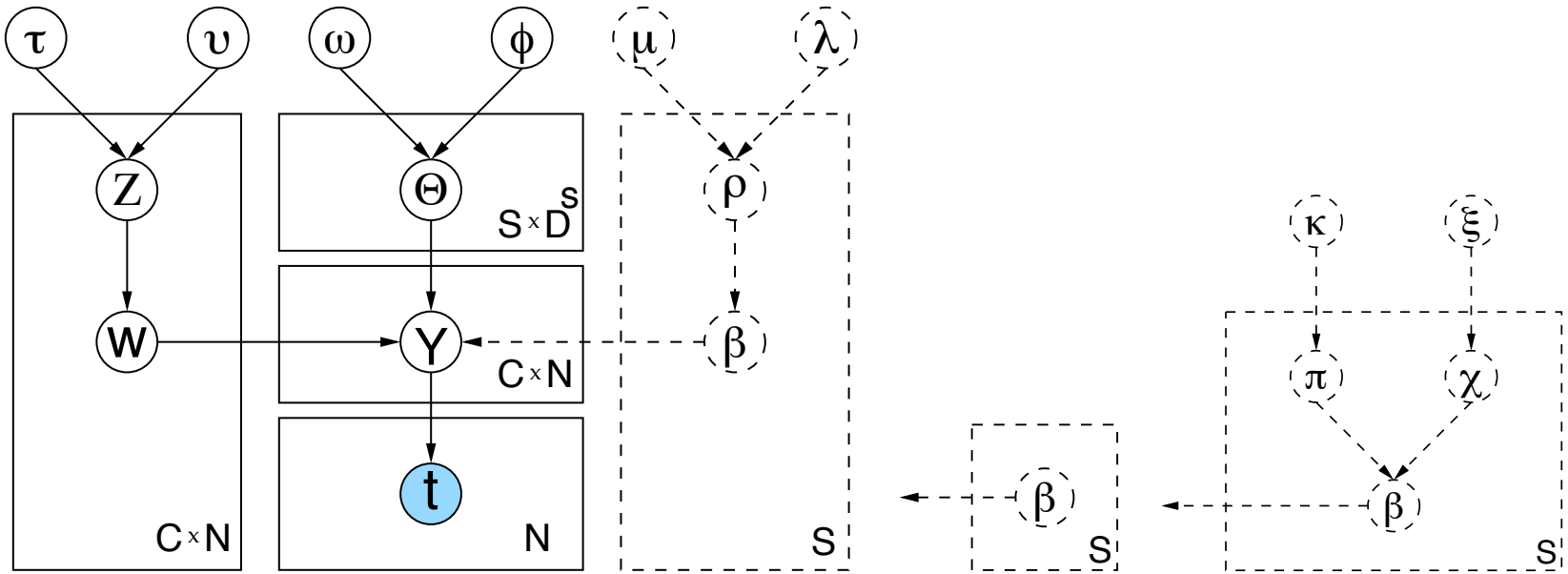


Figure: Plates diagram of the model with extensions.

Table: Average individual F.S percentage accuracy.

Feature Space	VBKC	Ding and Dubchak
Amino Acid Composition (C)	$\mathbf{51.2 \pm 0.5}$	44.9
Predicted Secondary Structure (S)	38.1 ± 0.3	35.6
Hydrophobicity (H)	32.5 ± 0.4	36.5
Polarity (P)	32.2 ± 0.3	32.9
van der Waals volume (V)	32.8 ± 0.3	35
Polarizability (Z)	33.2 ± 0.4	32.9
PseAA $\lambda = 1$ (λ_1)	41.5 ± 0.5	—
PseAA $\lambda = 4$ (λ_4)	41.5 ± 0.4	—
PseAA $\lambda = 14$ (λ_{14})	38 ± 0.2	—
PseAA $\lambda = 30$ (λ_{30})	32 ± 0.2	—
SW with BLOSUM62 (SW_1)	$\mathbf{59.8 \pm 1.9}$	—
SW with PAM50 (SW_2)	49 ± 0.7	—

- Amino Acid Composition (C) individually more predictive than any PseAA. String kernels outperform the rest.
- In general agreement with Ding and Dubchak.

Table: Effect of F.S combination. % Accuracy reported.

Feature Spaces	VBKC	Ding and Dubchak (AvA SVM)
C	51.2 ± 0.5	44.9
CS	55.7 ± 0.5	52.1
CSH	57.7 ± 0.6	56.0
CSHP	57.9 ± 0.9	56.5
CSHPV	58.1 ± 0.8	55.5
CSHPVZ	58.6 ± 1.1	53.9
CSHPVZλ_1	60 ± 0.8	—
CSHPVZ$\lambda_1\lambda_4$	60.8 ± 1.1	—
CSHPVZ$\lambda_1\lambda_4\lambda_{14}$	61.5 ± 1.2	—
CSHPVZ$\lambda_1\lambda_4\lambda_{14}\lambda_{30}$	62.2 ± 1.3	—
CSHPVZ$\lambda_1\lambda_4\lambda_{14}\lambda_{30}SW_1$	66.4 ± 0.8	—
CSHPVZ$\lambda_1\lambda_4\lambda_{14}\lambda_{30}SW_1SW_2$	$\mathbf{68.1 \pm 1.2}$	—
		Shen and Chou
SHPVZ$\lambda_1\lambda_4\lambda_{14}\lambda_{30}$	61.0 ± 1.4	62.1

- An apparent 2% increase in predictive performance by the inclusion of PseAA. Significant improvement via the addition of the string kernels

Table: Best single run performances (% Accuracy).

Feature Spaces	Ding and Dubchak	Shen and Chou	VBKC
CSHP	56.5	—	59.3
SHPVZ$\lambda_1\lambda_4\lambda_{14}\lambda_{30}$	—	62.1	63.5
CSHPVZ$\lambda_1\lambda_4\lambda_{14}\lambda_{30}$	—	—	63.9
CSHPVZ$\lambda_1\lambda_4\lambda_{14}\lambda_{30}SW_1SW_2$	—	—	**70**
No. of Classifiers	2,106	9	1

- State-of-the-art performance with a single classifier.

Table: Mean CPU times (sec) $\pm\sigma$ over 20 runs for the VBKC.

CSHPVZ	2,243 $\pm$485
SHPVZ$\lambda_1\lambda_4\lambda_{14}\lambda_{30}$	2,844$\pm$644
CSHPVZ$\lambda_1\lambda_4\lambda_{14}\lambda_{30}$	2,713$\pm$453

- Reported times for ANNs and SVMs are a minimum of 20 times higher (12 hours).

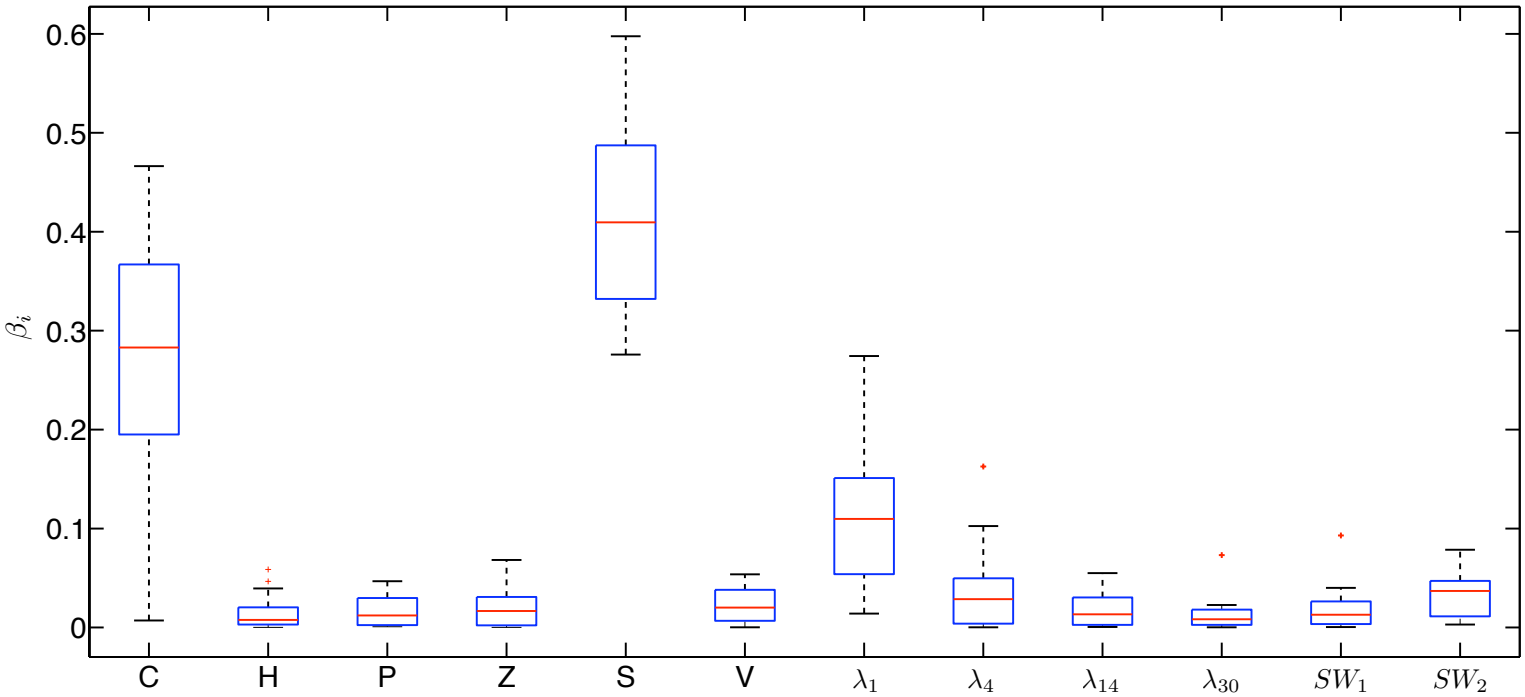
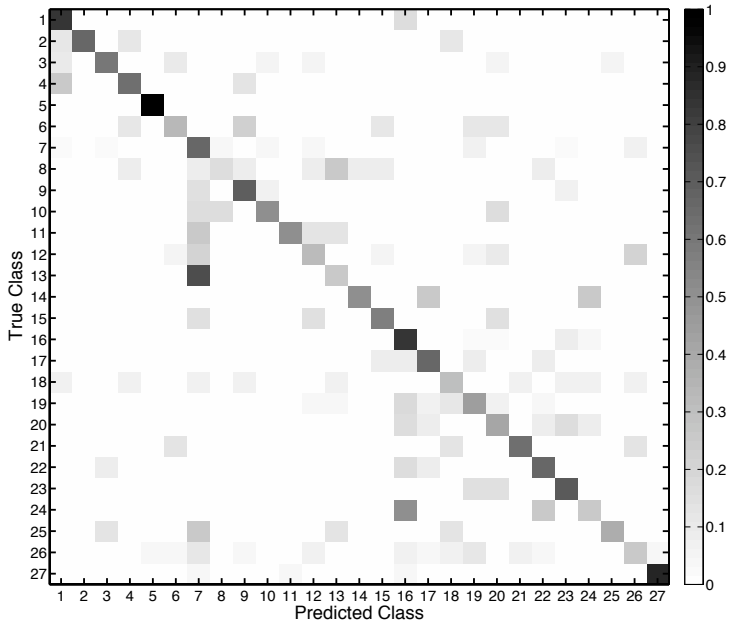


Figure: Inferring the statistical strength via the weighted combination.

- The original amino-acid composition (C) and the secondary structure (S) are shown to be most significant.

- A further look on mis-classifications reveals...



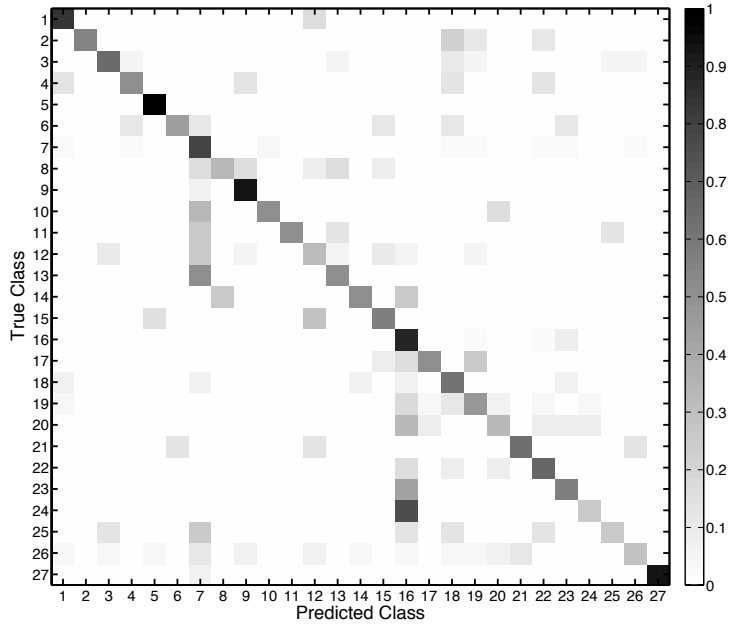


Figure: Left: Without PseAA Right: With pseudo-ones

- The patterns persist and error increases from 27% to 32%.
- Pseudo-amino acids carry *non-complementary* information.

- Global and local protein characteristics.

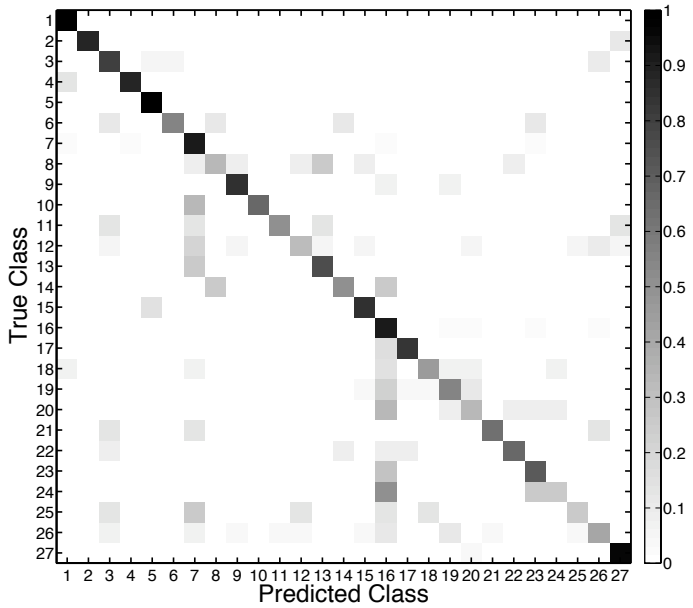


Figure: With the inclusion of the pairwise string kernels.

- Improvement on misclassification patterns via complementary information

Remote Homology Detection

- SCOP 1.53 database
- Predicting proteins at the super-family level. 54 binary imbalanced problems on a total of 4,352 proteins.
- Imbalance is treated via either the inclusion of an *ad-hoc* diagonal term, or through an alternative Bayesian prior specification.

Method	Mean ROC	Mean ROC50	Mean mRFP
VBKC	**0.924**	**0.567**	**0.0661**
SVM (SW)	0.896	0.464	0.0837
SVM (LA)	**0.925**	**0.649**	**0.0541**
SVM (MM)	0.872	0.400	0.0837
SVM (Mono)	**0.919**	**0.508**	**0.0664**

- Combining a Mismatch, Monomer, Local Alignment and Pairwise String kernel.

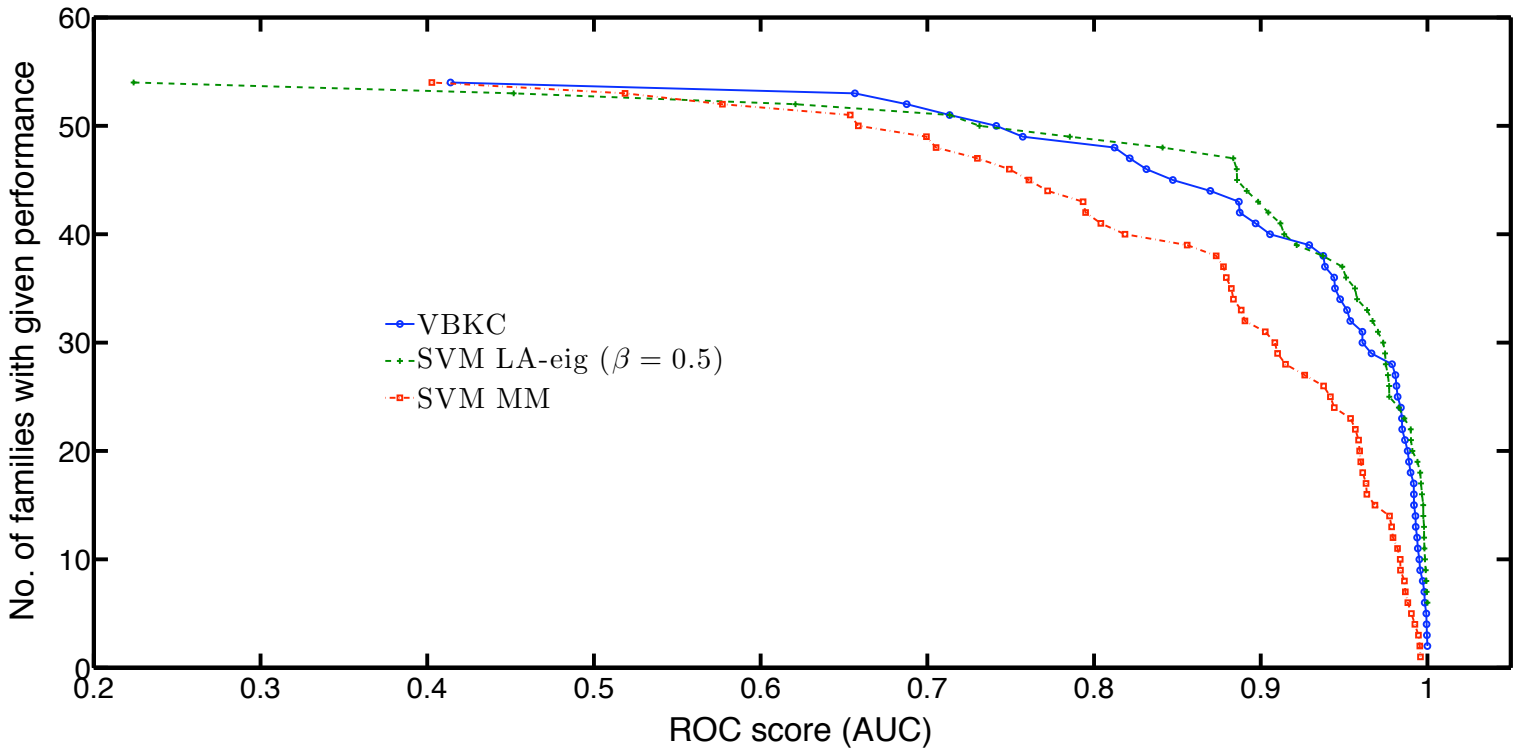


Figure: Comparison between SVM state-of-the-art string kernels and combination VBKC.

Conclusions

- The proposed Bayesian methodology has been shown to provide state-of-the-art predictive performance on two important bioinformatics problems.
- Combining local and global protein characteristics.
- Principled way of combining feature spaces without ad-hoc modifications.
- Probabilistic and multi-class kernel machine.
- The work has been accepted for publication in the Bioinformatics journal.

Thank you

theo@dcs.gla.ac.uk