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

Theodoros Damoulas* and Mark A. Girolami
Department of Computing Science, University of Glasgow, S. A. W. Building, G12 8QQ, UK

Received on November 13, 2007; revised on March 3, 2008; accepted on March 27, 2008
Advance Access publication March 31, 2008

ABSTRACT

Motivation: The problems of protein fold recognition and remote homology detection have recently attracted a great deal of interest as they represent challenging multi-feature multi-class problems for which modern pattern recognition methods achieve only modest levels of performance. As with many pattern recognition problems, there are multiple feature spaces or groups of attributes available, such as global characteristics like the amino-acid composition (C), predicted secondary structure (S), hydrophobicity (H), van der Waals volume (V), polarity (P), polarizability (Z), as well as attributes derived from local sequence alignment such as the Smith–Waterman scores. This raises the need for a classification method that is able to assess the contribution of these potentially heterogeneous object descriptors while utilizing such information to improve predictive performance. To that end, we offer a single multi-class kernel machine that informatively combines the available feature groups and, as is demonstrated in this article, is able to provide the state-of-the-art in performance accuracy on the fold recognition problem. Furthermore, the proposed methodology offers some insight to assessing the significance of recently introduced protein features and string kernels. The proposed methodology is well-founded within a Bayesian hierarchical framework and a variational Bayes approximation is derived which allows for efficient CPU processing times.

Results: The best performance which we report on the SCOP PDB-40D benchmark data-set is a 70% accuracy by combining all the available feature groups from global protein characteristics but also including sequence-alignment features. We offer an 8% improvement on the best reported performance that combines multi-class k-nn classifiers while at the same time reducing computational costs and assessing the predictive power of the various available features. Furthermore, we examine the performance of our methodology on the SCOP 1.53 benchmark data-set that simulates remote homology detection and examine the combination of various state-of-the-art string kernels that have recently been proposed.

Contact: theo@dcs.gla.ac.uk
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Much effort has been directed to the prediction of the 3D structures of proteins for which no experimental structures are available (Baker and Sali, 2001). Where there is sequence similarity to proteins of known structure, a comparative matching procedure is often adopted. However, where no such sequence similarity exists, the prediction problem is formidable, not least because the overall structure may be unlike that of any protein, the structure of which has been determined.

In this context, one approach, known as the taxonomic approach (Ding and Dubchak, 2001; Shen and Chou, 2006), has been to divide the problem of determining the overall 3D structure into that of determining its ‘fold’. The term ‘fold’ is used to denote a particular arrangement of a specific number of secondary structure components (usually alpha-helices and beta-strands) that is the basis of the overall structure of several different proteins which may have little or no amino acid sequence similarity. The appearances of some of these arrangements have given rise to names like ‘barrel’, ‘bundle’, ‘sandwich’ and ‘propeller’, although these tend to encompass several more specific folds e.g. the TIM beta/alpha barrel and the 5-bladed beta-propeller. Hence, protein fold prediction can be seen as a challenging multiclass recognition problem where proteins are classified into folds based on their characteristics and available measurements.

Past work on the problem of predicting protein folds has employed artificial neural networks (ANNs), support vector machines (SVMs), Bayesian networks, Hidden Markov Models and k-nn classifiers (Chou and Zhang, 1995; Dubchak et al., 1995; Jaakkola et al., 1999; Raval et al., 2002) with varying success. In Ding and Dubchak (2001) an extensive study on a publicly available data-set, consisting of 27 SCOP folds (Andreeva et al., 2004; Lo Conte et al., 2000), was conducted exploring the use of various multi-class adaptations of the well-known binary SVM classifier methodology. In that work, the best methodology for combining binary SVMs was identified for the particular problem giving an accuracy of 56%, and furthermore, via an extensive experimental procedure the most predictive protein characteristics were selected from the initial group considered. These were found to be the amino-acid composition (C), the secondary structure (S) and the hydrophobicity (H).

Recently, Shen and Chou (2006) proposed two modifications to the method of Ding and Dubchak (2001) that raised the best performance accuracy from 56 to 62.1%. Firstly, they proposed a somewhat ad hoc ensemble learning approach where
multi-class $k$-nn classifiers individually trained on each feature space (such as C or S) were later combined and secondly, they proposed the use of four additional feature groups to replace the amino-acid composition. These were pseudo-amino acid compositions (PseAA) (Chou, 2005) designed to capture sequence-order effects by using a correlation function between hydrophobicity and hydrophilicity in different intervals of the protein sequence.

In the present work, we concentrate on the same benchmark dataset of Ding and Dubchak (2001) with the extra groups of features proposed by Shen and Chou (2006), and also including sequence-alignment$^1$ features via a 

\[ \text{pairwise} \] kernel (Liao and Noble, 2003), which essentially describes the sequence based similarity of the proteins. We offer a single multi-class kernel machine able to operate on all of these groups of features simultaneously and instructively combine them. This offers a new and efficient way of incorporating multiple feature characteristics of the proteins without an increase in the number of required classifiers. In addition, we assess the importance and predictive power of the PseAA compositions proposed by Shen and Chou (2006) together with all the other available characteristics and gain insight on the protein fold recognition problem.

Furthermore, we demonstrate the generality of our methodology in a practical setting by addressing the remote homology problem on the SCOP 1.53 data-set as previously studied and described by a large number of works, see Leslie et al. (2004); Liao and Noble (2003); Lingner and Meinicke (2006); Saigo et al. (2004) and references within, where a variety of string kernels in conjunction with a discriminative SVM methodology have been proposed. Following our approach we select four of these state-of-the-art string kernels and combine them into an overall composite kernel where the multinomial probit kernel machine operates.

Related methodologies on kernel machines and multiple kernel learning (MKL) include the work by Lanckriet et al. (2004a, b); Lewis et al. (2006a, b); Sonnenburg et al. (2006) and references within, where semidefinite programming (SDP) or semi-infinite linear programming techniques are employed in order to minimize a loss function with respect to the kernel combination. These approaches build upon the SVM methodology and formulate the kernel combination problem as a further optimization procedure. The SDP approach suffers from large requirements in memory and CPU time in the order $O(S^2N^2)$, where $S$ is the number of sources and $N$ is the number of covariates. These methods also carry the inherent drawback of SVM methodologies, namely their problematic scaling for multiclass problems as they are based on binary classifiers by nature.

An explicit multi-class classifier within the Gaussian Process methodology was introduced recently by Girolami and Zhong (2007) which enables data integration by combining the covariance functions instead of kernels. Their methodology has the drawback of employing a first-order approximation for the inverse of the covariance functions. Finally, recent work by Melvin et al. (2007) in the context of protein classification employed adaptive codes to handle the multiclass prediction problem. However their methodology is based on learning a weighting of binary classifiers which, we argue, is not an efficient strategy especially for multiple feature space problems such as the one considered in this study.

2 APPROACH

The approach adopted is based on the motivation to reduce the number of classifiers needed for such challenging multi-class recognition problems where multiple feature sets are available, while improving performance. Combining binary classifiers as in the work by Ding and Dubchak (2001) increases heavily the computational resources needed since, e.g. for the best performing all-vs-all method, we need to deploy $S \times C(C - 1)/2 = 2106$ classifiers, where $S$ is the number of feature spaces or sources (only six in their work) and $C$ the number of classes.

Furthermore, even when employing multi-class classifiers in an ensemble learning framework such as the one proposed by Shen and Chou (2006), we still need as many classifiers as there are available feature spaces. Considering the nature of the protein fold prediction problem, where the fold type of a protein can depend on a large number of protein characteristics and also noting that even in the taxonomic approach the number of fold types already approaches the 1000 boundary, it is straightforward to see the need for a methodological framework that can cope with a large number of classes and can incorporate as many as there are available feature spaces while assessing their informational content.

The proposed approach, as can be seen from Figure 1, is based on the ability to embed each object description via the kernel trick (Shawe-Taylor and Cristianini, 2004) into a kernel space (Hilbert space). This produces a similarity measure between proteins in every feature space and then, having

---

$^1$Despite the apparent low homology dataset.
a common measure, we can combine informatively these similarities onto a composite kernel space. Hence now, a single multi-class kernel machine can operate on that composite space effectively ‘disregarding’ the number of feature spaces used. Inference by Bayes theorem on our hierarchical multiclass model enables us to learn the significance of each source/feature space and their predictive power by the corresponding kernel weights \( \beta_\text{s} \), to learn the regressors and the kernel parameters without resorting to \textit{ad-hoc} ensemble learning, combination of binary classifiers or parameter tuning.

### 3 MATERIALS AND METHODS

#### 3.1 Fold recognition

The original dataset from Ding and Dubchak (2001) (based on SCOP PDB-40D) consists of 313 proteins for training and 385 proteins for testing with \(<35\%\) sequence identity between any two proteins in the train and the test set. Furthermore, the extensions proposed by Shen and Chou (2006) exclude four proteins from the original dataset, namely proteins 2SCMC and 2GPF from the training set plus 2YHX_1 and 2YHX_2 from the test set, due to lack of sequence records.

The 27 SCOP fold types (Dubchak et al., 1995) together with the original feature spaces in Ding and Dubchak (2001), the four proposed by Shen and Chou (2006) which describe PseAA compositions estimated on different intervals of the protein sequence, and the two local alignment Smith-Waterman (SW) based feature spaces, with different scoring matrices, are described in Tables 1 and 2 in the Online Supplementary Materials (OSM hereafter).

#### 3.2 Remote homology detection (RHD)

The SCOP 1.53 benchmark data-set\(^5\) as described in Liao and Noble (2003) is employed to simulate the RHD problem. It consists of 4352 proteins belonging to one of 54 families and the positive training is performed on low-homologs while the positive testing on members of the same family. We consider four state-of-the-art string kernels, namely a local alignment (LA) kernel (Saigo et al., 2004), a mismatch (MM) kernel (Leslie et al., 2004), an oligomer kernel (Mono) (Linger and Meinicke, 2004) and a pairwise (PW) kernel (Liao and Noble, 2003), taking the best performing case from each string kernel category as a separate informational source. We follow the above past works within the kernel machine paradigm by adding a class-dependent regularization parameter to the diagonal of the kernels to improve performance on this highly imbalanced problem.

#### 3.3 Methodology

Consider now \( S \) feature spaces or sources of information; From each one we have input variables \( X_j \) as object descriptors such as strings or \( D' \) – dimensional vectors for \( s = 1, \ldots, S \) and corresponding multinomial target variables \( t_s \in \{1, \ldots, C\} \) for \( n = 1, \ldots, N \) where \( N \) is the number of observations and \( C \) the number of classes. By applying the kernel trick on the individual feature spaces created by the \( S \) sources we can define the \( N \times N \) composite kernel as

\[
K^{\text{comp}} = \sum_{s=1}^{S} \beta_s K^{\text{th}}
\]

with each element of the matrix given as

\[
K^{\text{th}}(x_i, x_j) = \sum_{s=1}^{S} \beta_s K^{\text{th}}(x_i^s, x_j^s)
\]

where \( \beta \) is an \( S \times 1 \) column vector, indicating each kernel's contribution and significance, and \( \Theta \) is an \( S \times D' \) matrix, describing the \( D' \) – dimensional kernel parameters \( \theta \), of all the base kernels \( K' \), which intuitively corresponds to the level of smoothing within each kernel. Now as we can see the mean composite kernel is a weighted summation of the base kernels, where each one describes a similarity measure between proteins based on specific features, with \( \beta_s \) as the corresponding weight for each one.

Following the standard approach for the multinomial probit by Albert and Chib (1993), we introduce auxiliary variables \( Y \in \mathbb{R}^{C \times N} \) and define the relationship between the auxiliary variable \( y_{ij} \) and the target variable \( t_i \) as

\[
t_i = \delta_{ij} \text{ if } y_{ij} > y_{ij} \forall j \neq i
\]

Now, the model response regressing on the variable \( y_{ij} \) with model parameters \( W \in \mathbb{R}^{C \times N} \) where \( y_{ij} = w_i \) is the weight with which data point \( n \) \('s votes\) for class \( c \), and assuming a standardized normal noise model as in Albert and Chib (1993) and Girolami and Rogers (2006) is given by

\[
y_{ijw} \sim \mathcal{N}(w^T C_k^{ij}, 1)
\]

where \( \mathcal{N}(m, v) \) denotes the normal distribution of \( x \) with mean \( m \) and variance \( v \), \( W \) and \( Y \) are \( C \times N \) matrices, \( w_i \) is a 1 \( \times \) \( N \) row vector and \( K^{ij} \) is a \( N \times N \) column vector from the \( nb \) column of the composite kernel \( K^{\text{comp}} \). Note that \( w_i K^{ij} \) is similar for any two data points \( n, n' \) that are similar in feature space. Hence now, the likelihood, can be expressed as the following by simply marginalizing over the auxiliary variable \( y_{ij} \) and making use of relations 1 and 2:

\[
P(t_n = i|W^c, K^{ij}) = \int P(t_n = i|y_{ij})P(y_{ij}|W^c, K^{ij})dy_{ij}
\]

\[
= \delta_{ij} \prod_{c=1}^C \mathcal{N}(y_{ij}, (w_i K^{ij} , 1)dy_{ij}
\]

where the expectation \( E \) is taken with respect to the standardized normal distribution \( p(u) = \mathcal{N}(0, 1) \). Hence, we can easily calculate the likelihood by averaging the quantity inside the expectation for a sufficient number of random samples of \( u \).

The proposed graphical model as depicted in Figure 2 is completed by considering prior distributions on the model variables and, following a hierarchical approach, hyper-prior distributions on the parameters of the model. We place a product of zero mean Gaussian distributions on the regressors \( W \sim \prod_{c=1}^C \prod_{i=1}^N \mathcal{N}(w_i, 0, \sigma_w) \) with variance \( \sigma_w \) (described by the variable \( Z \) in the graph) and a gamma distribution on each scale with hyper-hyper-parameters \( \nu, v \), reflecting our lack of prior knowledge and taking advantage of the conjugacy of these distributions.

Furthermore, we place a gamma distribution with associated hyper-hyper-parameters \( \omega, \phi \) on each kernel parameter since \( \theta_{ij} \in \mathbb{R}^+ \). In the case of the mean composite kernel, a Dirichlet distribution with parameters \( \rho \) is placed on the combinatorial weights in order to satisfy the constraints imposed on the possible values which are defined on a simplex. A further gamma distribution is placed on each \( \phi \) with associated hyper-hyper-parameters \( \mu, \lambda \). The hyper-hyper-parameters \( \Psi = \{\nu, v, \omega, \phi, \mu, \lambda\} \) can be set by type-II maximum likelihood or set to uninformative values and the hyper and first level parameters \( \Theta = \{Y, W, \beta, \rho, \Theta, Z\} \) are sampled accordingly.

It is now straightforward to see that a Gibbs sampler can be readily constructed and standard MCMC approaches (Andrieu, 2003) can be employed for Bayesian inference in our model. In this article though we offer a variational Bayes approximation in order to achieve efficient computational processing times without loss of predictive performance.

Hence, we bound the model evidence by using an ensemble of factored posteriors to approximate the joint parameter posterior distribution.

\[\text{Available from http://www.ccls.columbia.edu/compbio/svm-pairwise}\]
The approximate posterior for the variances $Z$ is an updated product of inverse-gamma distributions and the posterior mean is given in the OSM or Denison et al. (2002). Finally, the approximate posteriors for the kernel parameters $Q(\Theta)$, the combinatorial weights $Q(\beta)$ and the associated hyper-prior parameters $Q(\alpha)$ can be obtained by importance sampling (Andrieu, 2003) in a similar manner to Girolami and Rogers (2006) since no tractable analytical solution can be offered.

Having described the approximate posterior distributions of the parameters and hence obtained the posterior expectations we turn back to our original task of making class predictions $\hat{f}$ for $N_{new}$ new proteins $X^*$ that are represented by $S$ different information sources $X^{**}$ embedded into Hilbert spaces as base kernels $K^{\beta,\Theta}$ and combined into a composite test kernel $K^{**}$. The predictive distribution for a single new protein $X^*$ is given by $p(f^* = c|X^*, \mu) = \int p(f^* = c|X^*, t) \phi(t) dt$ which ends up, see OSM, as

$$p(f^* = c|X^*, \mu) = E_{\mu(\alpha)} \left[ \prod_{n=1}^{N_{test}} \phi \left( \frac{1}{\sqrt{2\pi}\sigma_n^2} \left( \mu_n - \mu_n^c \right) \right) \right]$$ (10)

where, for $N_{test}$ objects, $\mu_n^c = \tilde{y}_n^c K^{**} V_{c}^{-1} \nu_{*}$ and $V_{c} = (1 + K^{**} V_{c}^{*})$ while we have dropped the notation for the standardization of the train $K(N \times N)$ and test $K(N \times N_{new})$ kernels on $\Theta, \beta$ for clarity.

In Algorithm 1 we summarize the VB approximation in a pseudo-algorithmic fashion.

### Algorithm 1 VB Multinomial Probit Composite Kernel Regression

1. Initialize $\Xi$, sample $\Psi$, create $K^{\beta,\Theta}$ and hence $K|\beta, \Theta$
2. while Lower Bound changing do
   
3. $\mu_{-c} = \mu_{-c}^* K^{\beta,\Theta} V_{c}$
4. $\tilde{y}_n \sim N_{test}(s_n, y_{hcp})$
5. $E_{\mu(\alpha)} \left[ \prod_{n=1}^{N_{test}} \phi \left( \frac{1}{\sqrt{2\pi}\sigma_n^2} \left( \mu_n - \mu_n^c \right) \right) \right]$
6. $\tilde{y}_n = \tilde{y}_n^c K^{**} V_{c}^{-1} \nu_{*}$
7. $\tilde{y}_n^c \sim N_{test}(s_n, y_{hcp})$
8. $E_{\mu(\alpha)} \left[ \prod_{n=1}^{N_{test}} \phi \left( \frac{1}{\sqrt{2\pi}\sigma_n^2} \left( \mu_n - \mu_n^c \right) \right) \right]$
9. $\tilde{y}_n^c \sim N_{test}(s_n, y_{hcp})$
10. Update $K|\beta, \Theta$ and $V_{c}$
11. end while
12. Create composite test kernel $K^{**}$
13. end loop

### 4 RESULTS AND DISCUSSION

Reported results are averaged over 20 (fold recognition) and 10 (RHD) randomly initialized trials in order to obtain statistical measures of accuracy and precision. We monitor convergence via the lower bound to the marginal likelihood and convergence is assumed when there is <0.01% increase of the lower bound progression or when a maximum of 100 (fold recognition) and 20 (RHD) iterations have been completed. Throughout this study we have employed second-order polynomial kernels for the global characteristics and inner product kernels for the
Regarding the original features employed by Ding and Dubchak (2001), we are in agreement with their observations as further, the individual performances decrease. Although according to Shen and Chou (2006), the PseAA composition 'has the same form as the conventional amino-acid composition, but contains much more information' it seems at this stage that none of the PseAA is as predictive as the conventional amino-acid composition. Furthermore, the local characteristics (SW) surprisingly outperform every global one and SW1 achieves a higher accuracy than the best SVM-combinations proposed by Ding and Dubchak (2001). This is because although most of the proteins have <35% sequence similarity, this seems to be an adequate similarity level to achieve a good accuracy.

In Table 2 we report the effect of sequentially adding the feature spaces in the order of Ding and Dubchak (2001), extending that to the addition of the PseAA compositions and finally adding the sequence similarity based features. We compare against the best performing SVM combination methodology as reported in Ding and Dubchak (2001) and the ensemble method of Shen and Chou (2006). As we can see in all the steps the proposed method outperforms the best reported accuracies and offers the current state-of-the-art in this data-set.

The best performances can be seen in Table 3 in comparison with the best ones reported in the cited past work. We achieve an improvement over both past methods while we employ a single multiclass kernel machine without resorting to ensemble learning techniques or combining multiple binary classifiers. When we consider a weighted combination of the base kernels, with \( \sum_{i=1}^{\lambda} \beta_i = 1 \) we are able to infer the significance of the corresponding feature descriptions. In Figure 3 we plot a summary of the weights over 20 runs depicting the lower quartile, median and upper quartile values.

As we can observe, the amino-acid composition and the secondary structure are judged as more important, followed by

---

### Table 1. Average individual F.S percentage accuracy

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>VBKC</th>
<th>Ding and Dubchak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid composition (C)</td>
<td>51.2 ± 0.5</td>
<td>44.9</td>
</tr>
<tr>
<td>Predicted secondary structure (S)</td>
<td>38.1 ± 0.3</td>
<td>35.6</td>
</tr>
<tr>
<td>Hydrophobicity (H)</td>
<td>32.3 ± 0.4</td>
<td>36.5</td>
</tr>
<tr>
<td>Polarity (P)</td>
<td>32.2 ± 0.3</td>
<td>32.9</td>
</tr>
<tr>
<td>van der Waals volume (V)</td>
<td>32.8 ± 0.3</td>
<td>35</td>
</tr>
<tr>
<td>Polarizability (Z)</td>
<td>33.2 ± 0.4</td>
<td>32.9</td>
</tr>
<tr>
<td>PseAA λ = 1 (λ₁)</td>
<td>41.5 ± 0.5</td>
<td>–</td>
</tr>
<tr>
<td>PseAA λ = 4 (λ₄)</td>
<td>41.5 ± 0.4</td>
<td>–</td>
</tr>
<tr>
<td>PseAA λ = 14 (λ₁₄)</td>
<td>38 ± 0.2</td>
<td>–</td>
</tr>
<tr>
<td>PseAA λ = 30 (λ₃₀)</td>
<td>32 ± 0.2</td>
<td>–</td>
</tr>
<tr>
<td>SW with BLOSUM62 (SW₁)</td>
<td>59.8 ± 1.9</td>
<td>–</td>
</tr>
<tr>
<td>SW with PAM50 (SW₂)</td>
<td>49 ± 0.7</td>
<td>–</td>
</tr>
</tbody>
</table>

---

### Table 2. Effect of F.S combination (% accuracy reported)

<table>
<thead>
<tr>
<th>Feature spaces</th>
<th>VBKC</th>
<th>Ding and Dubchak</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>51.2 ± 0.5</td>
<td>44.9</td>
</tr>
<tr>
<td>CS</td>
<td>55.7 ± 0.5</td>
<td>52.1</td>
</tr>
<tr>
<td>CSHP</td>
<td>57.7 ± 0.6</td>
<td>56.0</td>
</tr>
<tr>
<td>CSHPV</td>
<td>57.9 ± 0.9</td>
<td>56.5</td>
</tr>
<tr>
<td>CSHPVZi₅</td>
<td>58.1 ± 0.8</td>
<td>55.5</td>
</tr>
<tr>
<td>CSHPVZi₅λ₄</td>
<td>58.6 ± 1.1</td>
<td>53.9</td>
</tr>
<tr>
<td>CSHPVZi₅λ₄λ₁₄</td>
<td>60.0 ± 0.8</td>
<td>–</td>
</tr>
<tr>
<td>CSHPVZi₅λ₄λ₁₄λ₃₀</td>
<td>60.8 ± 1.1</td>
<td>–</td>
</tr>
<tr>
<td>CSHPVZi₅λ₄λ₁₄λ₃₀ (SW₁)</td>
<td>61.5 ± 1.2</td>
<td>–</td>
</tr>
<tr>
<td>CSHPVZi₅λ₄λ₁₄λ₃₀ (SW₁, SW₂)</td>
<td>62.2 ± 1.3</td>
<td>–</td>
</tr>
<tr>
<td>CSHPVZi₅λ₄λ₁₄λ₃₀ (SW₁, SW₂)</td>
<td>66.4 ± 0.8</td>
<td>–</td>
</tr>
<tr>
<td>CSHPVZi₅λ₄λ₁₄λ₃₀ (SW₁, SW₂)</td>
<td>68.1 ± 1.2</td>
<td>–</td>
</tr>
</tbody>
</table>

---

### Table 3. Best single run performances (% accuracy)

<table>
<thead>
<tr>
<th>Feature Spaces</th>
<th>Ding and Dubchak</th>
<th>Shen and Chou</th>
<th>VBKC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSHP</td>
<td>56.5</td>
<td>–</td>
<td>59.3</td>
</tr>
<tr>
<td>SHPVZλ₄λ₁₄λ₃₀</td>
<td>–</td>
<td>62.1</td>
<td>63.5</td>
</tr>
<tr>
<td>CSHPVZλ₄λ₁₄λ₃₀</td>
<td>–</td>
<td>–</td>
<td>63.9</td>
</tr>
<tr>
<td>CSHPVZλ₄λ₁₄λ₃₀SW₁, SW₂</td>
<td>–</td>
<td>–</td>
<td>70</td>
</tr>
</tbody>
</table>

---

### Figure 3. Combinatorial weights when all the feature spaces are employed.

---

### 4.1 Fold recognition

First we examine the performance from individual feature spaces to gain an overall understanding of their predictive abilities. This however does not draw the complete picture as complementary information, may be shared across sources achieving low performances. In Table 1 we present the mean percentage accuracy with std. from our method (VBKC) together with the best ones reported by Ding and Dubchak (2001) on the original dataset.

Regarding the original features employed by Ding and Dubchak (2001), we are in agreement with their observations as the best performing feature space, seems to be the amino-acid composition (C). The \( \lambda = 1 \) and \( \lambda = 4 \) PseAA achieve the second best global individual performance and as the ‘step’ \( \lambda \) increases further, the individual performances decrease. Although according to Shen and Chou (2006), the PseAA composition ‘has the same form as the conventional amino-acid composition, but contains much more information’ it seems at this stage that none of the PseAA is as predictive as the conventional amino-acid composition. Furthermore, the local characteristics (SW) surprisingly outperform every global one and SW1 achieves a higher accuracy than the best SVM-combinations proposed by Ding and Dubchak (2001). This is because although most of the proteins have <35% sequence similarity, this seems to be an adequate similarity level to achieve a good accuracy.

In Table 2 we report the effect of sequentially adding the feature spaces in the order of Ding and Dubchak (2001), extending that to the addition of the PseAA compositions and finally adding the sequence similarity based features. We compare against the best performing SVM combination methodology as reported in Ding and Dubchak (2001) and the ensemble method of Shen and Chou (2006). As we can see in all the steps the proposed method outperforms the best reported accuracies and offers the current state-of-the-art in this data-set.

The best performances can be seen in Table 3 in comparison with the best ones reported in the cited past work. We achieve an improvement over both past methods while we employ a single multiclass kernel machine without resorting to ensemble learning techniques or combining multiple binary classifiers. When we consider a weighted combination of the base kernels, with \( \sum_{i=1}^{\lambda} \beta_i = 1 \) \( \beta_i \geq 0 \) we are able to infer the significance of the corresponding feature descriptions. In Figure 3 we plot a summary of the weights over 20 runs depicting the lower quartile, median and upper quartile values.

As we can observe, the amino-acid composition and the secondary structure are judged as more important, followed by...
the PseAA $\lambda = 1$. However, it is worth noting that by taking out
the amino-acid composition we have only a small loss in
performance as we have seen in Table 2. These two observations
suggest that the original amino-acid (C) and the pseudo-
ones ($\lambda$) carry redundant information. Furthermore, despite
the individual accuracies of the SW features, they are not
heavily weighted. This is because they depend solely on the
sequence similarity between proteins and their quality of
discriminative information is strongly related to which end of
the $0–35\%$ sequence similarity the two proteins will belong.
In reality, for the real ‘twilight-zone’ of low-homology proteins
the $0–35\%$ similarity such features have little effect by
definition.

In Figure 4 the confusion matrix for a single run is depicted.
The values on the matrix are normalized according to
$R_{ij} = P_j / N_i$ where $N_i$ is the total number of proteins belonging
in class $i$ and $P_j$ is the number of these $N_i$ proteins that were
predicted to belong to class $j$. For example when all of the
proteins in class $c$ were predicted correctly, then $R_{cc} = 1$ and
$R_{ij} = 0 \forall j \in \{1, C\}$.

First, it is worth noting that there are two areas where
consistent misclassification occurs. The first one is when
proteins of class 10–13 (conA-like barrel, SH3-like barrel,
OB, beta-trefoil) are classified as Class 7 (fold: immunoglobulin
like), and the second one is when proteins of class 19–20 and 24
(Rossmann fold, P-loop, periplasmic binding protein-like) are
classified as Class 16 (fold: TIM-barrel). Noting that folds 7 and
16 are represented by the top two largest numbers in the
training set (30 and 29 proteins, respectively), this seems to
imply that these classes are over-represented in comparison
with other folds (mean size of 10 proteins) and that features
such as (pseudo- or not) amino-acid composition and
secondary structure offer little discriminative power on the
distinction problem in these two areas.

Furthermore, besides the proteins in the fifth class (fold:
4-helical cytokines) that are correctly classified as expected
by previous observations by Ding and Dubchak (2001), now the
first class (fold: globin-like) is also achieving a $100\%$ accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean ROC</th>
<th>Mean ROC50</th>
<th>Mean mRFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBKC</td>
<td>0.924</td>
<td>0.567</td>
<td>0.0661</td>
</tr>
<tr>
<td>SVM (SW)</td>
<td>0.896</td>
<td>0.464</td>
<td>0.0837</td>
</tr>
<tr>
<td>SVM (LA)</td>
<td>0.925</td>
<td>0.649</td>
<td>0.0541</td>
</tr>
<tr>
<td>SVM (MM)</td>
<td>0.872</td>
<td>0.400</td>
<td>0.0837</td>
</tr>
<tr>
<td>SVM (Mono)</td>
<td>0.919</td>
<td>0.508</td>
<td>0.0664</td>
</tr>
</tbody>
</table>

Table 4. ROC, ROC50 and median RFP scores

Fig. 4. Confusion matrix with each element normalized to $R_{ij}$.

4.2 Remote homology detection

As a generalization of the proposed methodology to other
related problem-domains, we consider the simulated remote
homology problem (RHD) as described in the works of Liao
and Noble (2003); Lingner and Meinicke (2006); Leslie et al.
(2004); Saigo et al. (2004). The results from the combination
of the string kernels are depicted in Table 4 together with the best
previously reported results within the SVM methodology. We
achieve a state-of-the-art performance via the combination of
the kernels and match the overall best performing SVM method
outperforming other string kernels. In Figure 5 the number of
families that achieve certain ROC scores is depicted in
comparison with some of the best performing methods reported
in the literature.

Furthermore, by employing the weighted combination, we
infer the contribution of each string kernel and as it can be seen
from Figure 6 the Monomer (Mono) and the LA kernel are
weighted most heavily as expected from Table 4 and previously reported results.

5 CONCLUSION

In this article we offer a single probabilistic multi-class multi-kernel machine that is able to operate simultaneously in multiple feature spaces via a kernel combination methodology. Furthermore, we illustrate the capabilities of our method in a well-benchmarked dataset by Ding and Dubchak (2001) in which recent studies (Shen and Chou, 2006) have improved predictive performance by the introduction of additional pseudo-amino acid composition feature spaces. We show that the additional feature spaces although overall improve performance by a factor of 1–2% in reality carry non-complementary information with the original amino-acid composition. The need for such information to tackle the two misclassification patterns can also be seen in the work of (Shahbaba and Neal, 2007) where even when the problem is treated as a hierarchical classification with parent classes, the performance is not improving beyond 61.4%

Furthermore, our methodology offers a significant reduction in computational resources as it is based on a single classifier operating over a composite space which retains the dimensionality \( (N \times N) \) of any of the individual contributing feature spaces \( (N \times N) \). This, in contrast with the past work of employing thousands of binary classifiers or an ensemble of individually trained classifiers is a significant improvement. We provide, the state-of-the-art on the problem under consideration with a best performance of 70% accuracy without resorting to \( ad-hoc \) approaches but employing a solid Bayesian formalism which enables us to infer the informative content of the feature spaces.

Finally, we extend our approach to the remote homology problem and demonstrate the generality of our approach in a practical setting by achieving a state-of-the-art performance via a combination of string kernels.

ACKNOWLEDGEMENTS

The authors would like to acknowledge insightful discussions with Dr David Leader and Dr Rainer Breitling, and helpful suggestions by the anonymous reviewers. The first author would like to acknowledge the support received from the Advanced Technology & Research Group within the NCR Financial Solutions Group Ltd company and especially the help and support of Dr Gary Ross and Dr Chao He.

Funding: NCR Financial Solutions Group Ltd. Scholarship to T.D.; EPSRC Advanced Research Fellowship (EP/E052029/1) to M.A.G.

Conflict of Interest: none declared.

REFERENCES