10.25100/iyc.v18i2.2154

SYSTEMS ENGINEERING

Multi-class superfamily prediction using 3D models enriched with physicochemical properties

INGENIERÍA DE SISTEMAS

Predicción de superfamilias usando modelos 3D enriquecidos con propiedades fisicoquímicas

Oscar F. Bedoya*§, Irene Tischer*

**Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle. Cali, Colombia.*
§oscar.bedoya@correounivalle.edu.co, irene.tischer@correounivalle.edu.co

(Recibido: Agosto 11 de 2015 – Aceptado: Abril 24 de 2016)

Abstract

In this paper, two new methods that address the multi-class superfamily prediction problem are presented. In the multi-class superfamily recognition problem each amino acid sequence has to be classified into one of the known structural classes (i.e., superfamilies). Most of the strategies that have been proposed to predict superfamilies are based on using the binary classifiers that detect remote homologs. The remote homology detection problem is about finding a classifier that is able to separate remote homologs from non-remote homologs. The current methods for multi-class superfamily recognition take the outputs of the binary classifier (i.e., the scores) for each SCOP superfamily in the data set and build a classification model (i.e., multi-class classifier). Unlike the current methods, which represent a protein considering the amino acids composition, in this research we use the number of times that 3D models enriched with physicochemical properties occur in both its predicted contact map and its interaction matrix. We hypothesize that including both 3D information and physicochemical properties might have an impact in the accuracy obtained during the superfamily prediction. In this paper, we present two new strategies for predicting superfamilies that use 3D models enriched with physicochemical properties, the single-MCS and the hierarchical-MCS methods, which reach an accuracy percentage of 74% and 76% on the SCOP 1.53 data set, respectively. In addition, tests on the SCOP 1.55 and the SCOP 1.61 are also presented.

Keywords: *3D enrich models, Binary classifiers, Physicochemical properties, SCOP superfamily, Superfamily prediction.*

Resumen

En este artículo se presenta dos nuevos métodos para la predicción de superfamilias. En el problema de la predicción de superfamilias cada secuencia de aminoácidos se debe clasificar en una de las clases estructurales conocidas (i.e., superfamilias). La mayoría de las estrategias que se han propuesto para predecir superfamilias se basan en usar los clasificadores binarios que detectan homólogos remotos. Detectar homólogos remotos está relacionado con encontrar un clasificador que es capaz de indicar si una proteína es, o no, un homólogo remoto de un conjuntos de proteínas. Los métodos actuales para detectar superfamilias toman las salidas de los clasificadores binarios para cada superfamilia y construyen un modelo de clasificación. A diferencia de los métodos actuales, los cuales representan a las proteínas considerando la composición de aminoácidos, nosotros usamos el número de veces que modelos 3D enriquecidos con propiedades fisicoquímicas ocurren tanto en el mapa de contacto predicho como en la matriz de interacción. Nuestra hipótesis es que al incluir los modelos 3D con las propiedades fisicoquímicas se puede tener un impacto en la exactitud obtenida durante la predicción de superfamilias. En este artículo se presenta dos nuevas estrategias para predecir superfamilias, los métodos single-MCS y hierarchical-MCS, los cuales alcanzan una exactitud del 74% y 76% en el conjunto SCOP 1.53, respectivamente. Además, se presentan otras pruebas realizadas en los conjuntos SCOP 1.55 y SCOP 1.61.

Palabras clave: *clasificadores binarios, modelos 3D enriquecidos, predicción de superfamilias, propiedades fisicoquímicas, superfamilia SCOP.*

1. Introduction

Remote homology detection focuses on the binary classification problem of discriminating between positives samples (i.e., remote homologs) and negative samples (i.e., non-remote homologs). Positive samples are proteins from a single structural class (i.e., a superfamily) and negative samples represent the rest of the proteins in the data set. For instance, in the SCOP 1.53 data set, 54 families are considered, and thus, 54 classifiers have to be built. Each classifier in the remote homology detection problem is trained to distinguish between proteins from a specific structural class (i.e., a given superfamily) and proteins from other folds. Even though each classifier is able to separate remote homologs from non-remote homologs, a more real problem is related to the multi-class superfamily recognition. The multi-class superfamily recognition problem is defined as the task of taking an amino acid sequence and predicting its corresponding superfamily. The importance about predicting the superfamily of a protein based on its primary sequence is that it allows understanding the function of a protein, which is considered a difficult task in Bioinformatics.

Predicting superfamilies based on the binary classifiers that detect remote homologs has been addressed in previous works (Ding & Dubchak, 2001; Huang et al., 2003; Rangwala & Karypis, 2006; Ie et al., 2007; Leslie et al., 2007; Lin & Li, 2007; Lin, 2008). The current methods for multi-class superfamily recognition take the outputs of the binary classifier (i.e., the scores) and build a classification model (i.e., multi-class classifier) to predict the SCOP superfamily of each protein in the data set. Support vector machines (SVM) are commonly used to obtain the multi-class superfamily predictor. The difficulty when building the multi-class superfamily model is that the scores produced by the binary classifiers are not comparable, and thus, a classification model that captures the behaviour of the scores is needed. In this paper, we use binary classifiers that consider 3D models enriched with physicochemical properties. Unlike the current methods, which use binary classifiers that consider the frequencies of k-mers (i.e., k-length subsequences of amino acids), we use a protein representation related to the 3D structure of the protein. We hypothesize that our protein representation in the binary classifiers is suitable for predicting superfamilies because there is a relationship between the 3D information and the function of the protein (Yang et al., 2008).

In this paper, we propose two new multi-class superfamily predictors called single-MCS and hierarchical-MCS. The single-MCS method builds a multi-class superfamily model using a collection of binary classifiers that are based on 3D models enriched with physicochemical properties. The hierarchical-MCS method is divided into two steps. First, the SCOP class is predicted and then a multi-class superfamily predictor for each SCOP class is used. In the following section every step in the single-MCS and hierarchical-MCS methods are explained in detail. In Section 3, the results are shown considering the SCOP 1.53, SCOP 1.55, and SCOP 1.61 data sets. Finally, the conclusions are presented in Section 4.

2. Methodology

2.1 The single-MCS method

In the multi-class superfamily recognition problem, the binary classifiers that discriminate remote homologs are used. Each protein is represented as a vector (i.e., the output vector) that holds the real-valued discriminant scores obtained when the protein is submitted to the binary classifiers. This strategy is called one-vs-all approach and is based on the idea of using None-vs-the-rest classifiers to obtain an output vector of size N and then make a prediction.

The single-MCS (single Multi-Class Superfamily predictor) method is based on the following steps: (1) Splitting the training data set into five cross-validation sets; (2) training a classifier for every family using four subsamples of the partitioned training set; (3) submitting one subsample of the partitioned training set to the classifiers to obtain the discriminant scores (i.e., testing the classifiers);

(4) repeat steps 2 and 3 five times changing the subset that is taken for testing; (5) training a multi-class classifier with the scores obtained in the previous step and using the superfamily as the class label; (6) re-training the family classifiers on the full training set; (7) submitting the full test set to the classifiers and obtaining the scores; (8) submitting the scores to the trained multi-class classifier to predict the superfamily.

Figure 1 shows the process of obtaining the scores from the binary classifiers. Each binary classifier is built for distinguishing the remote homologs and non-remote homologs of a specific family. When tested, classifier reaches a performance score on a given amino acid sequence. Every score is a real value ranging from 0.0 to 1.0, where the higher the score the more probable the test protein is a remote homolog of the family. The single-MCS method is based on training a multi-class classifier with the scores of the binary classifiers, and thus, a strategy to obtain the scores is needed. Every family has a training set with positive and negative samples. Both positive and negative samples in the training set are divided in five parts. Four out of the five parts from both positive and negative samples are taken to build a binary classifier. Then, the proteins in the remaining part (i.e., the testing part) are submitted to 23 binary classifiers. Even though 54 classifiers are needed to detect the remote homologs in the SCOP 1.53 data set, we selected one classifier for each superfamily. The considered SCOP 1.53 data set is formed by 23 superfamilies, and thus, a total of 23 classifiers have to be used in the multi-class superfamily recognition problem. The classifiers considered in this research use 3D models enriched with physicochemical properties. Every protein is represented as the number of times that each model is observed in a predicted contact map and an interaction matrix as presented in Bedoya & Tischer (2015). Submitting a test sample to the whole set of binary classifiers produces a score-vector. In this research, the score-vector is a 23-length vector that holds the scores of a given amino acid sequence when is submitted to the 23 binary classifiers. The process shown in Figure 1 is repeated five times changing the subsamples that are taken as the testing data set.

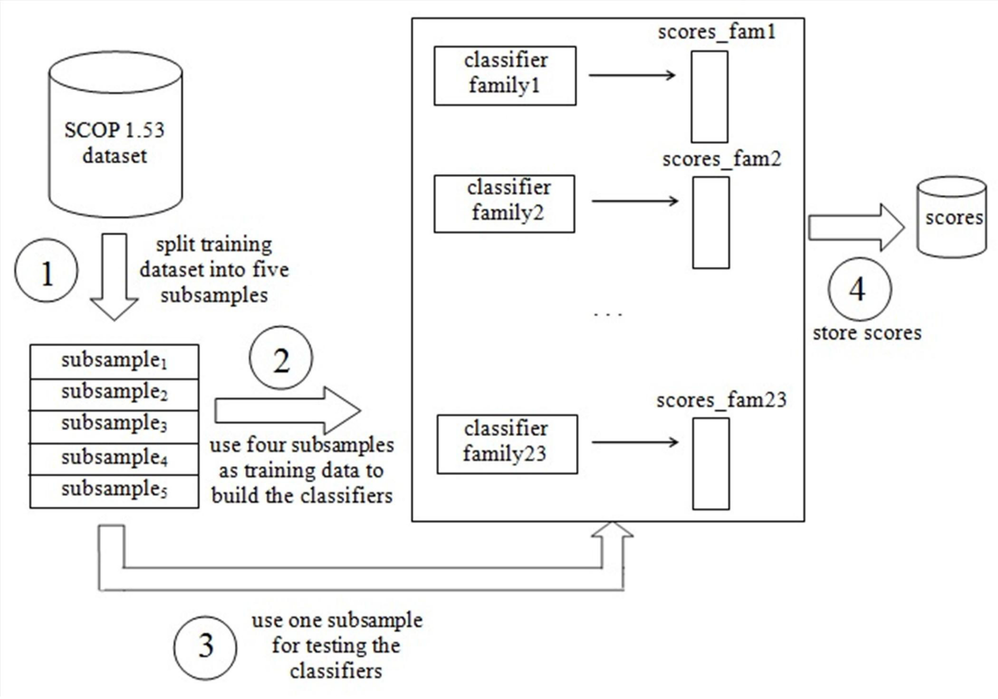

Figure 1. *Four steps included in the single-MCS method to obtain the scores from the proteins in the SCOP 1.53 data set.*

The score-vectors and their corresponding class labels are used to train a multi-class classifier. The class label is the SCOP superfamily of a given protein. Each protein is represented by the 23 values in the score-vector (i.e., holding the scores when submitting the protein to the binary classifiers) and one class label (i.e., indicating the superfamily of the protein). Support Vector Machines (SVM) along with One-vs-all classifiers have been used to build a model in the multi-class recognition problem (Leslie et al., 2007; Rangwala & Karypis, 2006). In this research, in addition to using the SVM technique we also use three classification strategies that are suitable for multi-class problems (i.e., LogitBoost, RandomSubSpace, and RandomForest). LogitBoost is a boosting classification algorithm based on additive logistic regression. Boosting strategies try to use a set of weak learners to create a single strong learner. A weak learner is a classifier that exhibits a low correlation with the true classification. RandomSubSpace is a classification technique that uses multiple trees constructed systematically by randomly selecting subsets of the components in the feature vector (i.e., the score-vector), which are represented by trees constructed in subspaces that are randomly chosen. RandomForest is a classification technique for constructing a forest of random trees. The output of the RanfomForest technique is the mode of the classifications obtained by the individual random trees.

2.2 The hierarchical-MCS method

The hierarchical-MCS method addresses the multi-class superfamily recognition problem by introducing a hierarchical model. The hierarchical-MCS method is divided in two stages. First, the SCOPclass of a given protein is predicted by using a multi-class SCOP class predictor. Then, a multiclass superfamily classifier is used to predict the superfamily. Each SCOP class has a multi-class superfamily predictor. In the first stage of the hierarchical-MCS method, a multi-class SCOP class recognition model is obtained by using the following three steps.

A binary classifier is obtained for each SCOP class. There are seven classes in the SCOP 1.53 data set, and thus, seven binary classifiers are obtained. Proteins inside a SCOP class are considered positive samples and proteins outside the class are taken as negative samples.

A multi-class classifier is trained using the seven scores of the binary classifiers. Every protein is represented by seven values (i.e., the scores that the protein obtains when it is submitted to the models that represent every SCOP class). The class labels in the multi-class SCOP class problem are the values {1,2,3,4,5,6,7}, which correspond to the SCOP classes in the data set.

A test set is submitted to the multi-class classifiers to obtain a prediction of the class. At the end of the first stage of the hierarchical-MCS method, a prediction of the SCOP class for each protein in the test set is available.

In the second stage of the hierarchical-MCS method, a multi-class superfamily classifier is built for the superfamilies in each SCOP class. The SCOP 1.53 data set includes proteins from seven SCOP classes (i.e., all alpha proteins, all beta proteins, alpha and beta proteins (a/b), alpha and beta proteins (a+b), multi-domain proteins, membrane and cell surface proteins and peptides, and small proteins). Every SCOP class has a given number of superfamilies. For instance, there are five SCOP superfamilies in the class ‘all alpha proteins’, eight SCOP superfamilies in the class ‘all beta proteins’, five SCOP superfamilies in the class ‘alpha and beta proteins (a/b)’, and five SCOP superfamilies in the class ‘small proteins’. In the hierarchical-MCS method, a multi-class superfamily classifier considers the superfamilies inside a specific class. For instance, the multi-class superfamily classifier that is built for class ‘all alpha proteins’ has only five class labels (i.e., instead of 23 as in the single-MCS method).

3. Results and discussion

In this section, the results of the experiments are shown. Tests on different SCOP versions are presented.

3.1 Accuracy measure

In this paper, the top1 accuracy is used to evaluate the multi-class superfamily predictors. For each test instance t, a given multi-class superfamily predictor outputs 23 scores. The instance t is assigned to the superfamily with the highest score yielded by the classifier. An instance is considered to be correct if its true class is among the n highest-ranked classes. The top1 value measures whether the multi-class method is able to identify the correct superfamily of the test instance.

3.2 Selecting the test set

For each superfamily, we kept one family for testing the superfamily prediction and we used the remaining families for training the multi-class superfamily predictor. For instance, the superfamily 1.4.1 has three families (i.e., 1.4.1.1, 1.4.1.2, and 1.4.1.3). We kept the family 1.4.1.1 for testing and the families 1.4.1.2 and 1.4.1.3 for training. Even though there are 54 families in the SCOP 1.53 data set, a total of 23 superfamilies can be used for testing the multi-class superfamily recognition. We selected eight out of the 23 superfamilies in the SCOP 1.53 data set for testing the multi-class superfamily recognition problem.

Table 1 shows the superfamilies that were chosen for testing. As observed, two superfamilies were selected from each SCOP class in the data set. Table 1 also shows the specific ROC score reached for each superfamily when different methods for detecting remote homologs are used. In the remote-3DI method (Bedoya & Tischer, 2015) every protein is represented using a predicted contact map and an interaction matrix. The remote-3DI method uses models with 3D information (i.e., typical 3D interactions that occur in the contact map) enriched with physicochemical properties. Table 1 shows the ROC score when the ‘Alpha helix propensity’ and the combination of indices ‘pK (-COOH)’ and ‘Atom based hydrophobic moment’ are used. The mean ROC score of the remote-3DI method with the ‘Alpha helix propensity derived from designed sequences’ is 0.953±0.060. In addition, the remote-3DI with the combination of physicochemical properties ‘pK (-COOH) + Atom-based hydrophobic moment’ reach a mean ROC score of 0.942±0.070. The specific ROC scores shown in Table 1 allow having an idea about which superfamilies can be more difficult to predict. For instance, superfamily 7.39.1 in class 7 has the lowest ROC score for an individual superfamily in most of the methods used to detect its remote homologs.

Table 1. *Superfamilies chosen for testing and their corresponding number of proteins and accuracy values when using different physicochemical properties.*

Superfamily	Number of proteins	ROC score with remote-3DI (alpha helix)	ROC score with remote 3DI (pK (-COOH) + Atom-based hydrophobic moment)
1.27.1	6	0.996	0.995
1.4.1	23	0.984	0.989
2.1.1	31	0.989	0.991
2.56.1	8	0.996	1.000
3.1.8	10	0.992	0.996
3.42.1	10	0.996	0.997
7.3.6	26	0.884	0.864
7.39.1	14	0.877	0.797

Table 2. *Top1 accuracy achieved by the single-MCS method when using 15 combinations of physicochemical properties and six different classification techniques.*

Binary classifier	Logit-	Random-	Random-	SVM	SVM	SVM
	Boost	Subspace	Forest	(c=10)	(c=20)	(c=100)
remote-3DI (alpha helix)	0.62	0.58	0.65	0.70	0.73	0.74
remote-3DI (Hydropathy)	0.49	0.46	0.54	0.62	0.59	0.50
remote-3DI (pK (-COOH))	0.52	0.45	0.47	0.48	0.47	0.56
remote-3DI (alpha+atom)	0.46	0.37	0.46	0.61	0.60	0.62
remote-3DI (alpha+pK)	0.54	0.41	0.46	0.49	0.50	0.56
remote-3DI (alpha+C)	0.41	0.40	0.52	0.36	0.36	0.33
remote-3DI (pK+atom)	0.43	0.43	0.50	0.44	0.44	0.44
remote-3DI (pK+C)	0.52	0.46	0.46	0.53	0.49	0.48
remote-3DI (atom+C)	0.40	0.38	0.40	0.41	0.44	0.46
remote-3DI (hydropathy+pK)	0.43	0.40	0.44	0.41	0.42	0.45
remote-3DI (alpha+pK+atom)	0.46	0.45	0.39	0.53	0.50	0.54
remote-3DI (alpha+pK+C)	0.47	0.49	0.46	0.57	0.55	0.55
remote-3DI (alpha+atom+C)	0.57	0.51	0.53	0.51	0.50	0.47
remote-3DI (pK+atom+C)	0.56	0.55	0.56	0.55	0.60	0.61
remote-3DI (alpha+pK+atom+C)	0.43	0.50	0.62	0.60	0.59	0.57

3.3 Evaluating the single-MCS method

Table 2 shows the top1 values when different physicochemical properties for the binary classifiers and different multi-class superfamily classifiers are used. As observed, SVMs are tested using three different values for the misclassification cost (i.e., c=10, 20, 100). According to Rangwala & Karypis (2006), the results of the multi-class classification process depend on the value C, which is the misclassification cost that determines the trade-off between the generalization capability of the model being learned and maximizing the margin. An optimization of the value C has to be done. The optimization process helps preventing under-fitting and over-fitting the data during the training process. The mean top1 at superfamily level of the single-MCS method is 0.50±0.11. The highest top1 value is 0.74, which is achieved when binary classifiers in the remote-3DI method with 'Alpha helix propensity derived from designed sequences' and SVM c=100 as multi-class classifier are used. Table 3 shows the mean ROC score for each physicochemical property and the corresponding mean top$_1$ accuracy. As expected, the physicochemical properties that have high ROC scores tend to reach high top$_1$ values. For instance, the mean ROC score of the remote-3DI method with the 'Alpha helix propensity' is highest in the experiments (i.e., 0.953). The top$_1$ accuracy reached when the 'Alpha helix propensity' is used to predict superfamilies is also the highest in the experiments. In addition, the lowest top$_1$ value is 0.40, which is reached by the combination 'alpha+C'. The remote-3DI using 'alpha+C' also exhibits one of the lowest mean ROC scores. One of the most important results that the Table 3 suggests is that there is clear a relationship between the ROC score of the binary classifiers and the top$_1$ accuracy of the multi-class superfamily predictors. There are some physicochemical properties that exhibit higher ROC

Table 3. *mean ROC scores and their corresponding mean top1 accuracy in the single-MCS method when using 15 combinations of physicochemical properties*

Physicochemical properties	mean ROC score using remote-3DI	mean top$_1$ accuracy
Alpha-helix propensity	0.953±0,060	0.67
Hydropathy index	0.945±0,067	0.53
pK (-COOH)	0.950±0,069	0.49
Alpha-helix propensity, Atom-based hydrophobic moment	0.950±0.066	0.52
Alpha-helix propensity, pK (-COOH)	0.947±0.071	0.49
Alpha-helix propensity, Relative preference value at C’	0.947±0.071	0.40
pK (-COOH), Atom-based hydrophobic moment	0.942±0.070	0.45
pK (-COOH), Relative preference value at C’	0.945±0.071	0.49
Atom-based hydrophobic moment, Relative preference value at C’	0.944±0.071	0.42
Hydropathy index, pK (-COOH)	0.943±0.052	0.43
Alpha-helix propensity, pK (-COOH), Atom-based hydrophobic moment	0.944±0.069	0.48
Alpha-helix propensity, pK (-COOH), Relative preference value at C’	0.944±0.076	0.52
Alpha-helix propensity, Atom-based hydrophobic moment, Relative preference value at C’	0.944±0.072	0.52
pK (-COOH), Atom-based hydrophobic moment, Relative preference value at C’	0.945±0.070	0.57
Alpha-helix propensity, pK (-COOH), Atom-based hydrophobic moment, Relative preference value at C’	0.941±0.069	0.55

scores than others. The results suggest that the higher the ROC score of a physicochemical property used in a binary classifier, the higher the top$_1$ accuracy. Another result that the test set suggests is about the classification technique that is more suitable for predicting superfamilies in the single-MCS method. As observed in Table 2, even though we are using six strategies to obtain a multi-class superfamily classifier, the Support vector machines with c=100 is the technique that reaches the highest top1 accuracy in eight out of 15 physicochemical properties. The test set suggests that SVMs with a misclassification cost of 100 is a classifier that is suitable for the single-MCS method.

3.4 Evaluating the hierarchical-MCS method

There are seven classes in the SCOP 1.53 data set. SCOP classes 'all alpha proteins', 'all beta proteins', 'alpha and beta proteins (a/b)', 'alpha and beta proteins (a+b)', 'Multi-domain proteins', 'Membrane and cell surface proteins and peptides', and 'Small proteins' have 804, 950, 694, 737, 54, 121, and 992 positive samples, respectively. The negative samples for a given SCOP class are the proteins outside that class. For instance, because there are 4352 proteins in the SCOP 1.53 data set and 804 proteins in the class ‘all alpha proteins’, a total of 4352-804=3548 negative samples are available for that SCOP class. In the hierarchical-MCS method, we kept all the proteins of some superfamilies for testing and we trained the classifiers with the remaining proteins. For instance, the superfamily 1.27.1 has 18 proteins in the SCOP 1.53 data set. We kept the 18 proteins in the superfamily 1.27.1 for testing the multi-class SCOP class predictor and we trained the binary classifiers and the multi-class classifier for class 1 with 786 positives samples (i.e., 804-18) and 3548 negative samples. We repeated the same methodology for each superfamily in our test set (i.e., the eight superfamilies 1.27.1, 1.4.1, 2.1.1, 2.56.1, 3.1.8, 3.42.1, 7.3.6, and 7.39.1).

In the hierarchical-MCS method, a binary classifier is obtained for each SCOP class. We select the classifier with the highest ROC score using 5-fold cross validations on the training data set. We chose between the same classification techniques (BayesNet, NaiveBayes, NaiveBayes Multinomial, Multilayer perceptron, Hyper pipes, VFI, LMT) that were considered for detecting remote homologs in Bedoya and Tischer (2015). Then, a multi-class SCOP class predictor is obtained by using the same steps that we used in the single-MCS method but considering SCOP classes instead of SCOP superfamilies. In the second stage of the hierarchical-MCS method the multi-class superfamily classifiers are built using the same binary classifiers that were considered in the single-MCS method (i.e., the binary classifiers for the 23 superfamilies). Every protein in the data set is represented by the scores of the 23 binary classifiers and the class label. However, each multi-class superfamily classifier in the hierarchical-MCS method has fewer class labels than in the single-MCS method, and thus, the prediction might be improved. For instance, eight class labels are used in the SCOP class 'all beta proteins' (i.e., class labels 6, 7, 8, 9, 10, 11, 12, and 13) and five class labels are used in the SCOP class 'alpha and beta proteins (a/b)' (i.e., class labels 14, 15, 16, 17, and 18).

Table 4 shows the top$_1$ accuracy at superfamily level for 15 different combinations of physicochemical properties and different multi-class classification techniques. Even though we predicted the SCOP class for all the proteins in some superfamilies, we only took proteins in one family for testing the superfamily prediction accuracy of the hierarchical-MCS method. The other families inside a given superfamily were used for training as it was explained in the single-MCS method. The top$_1$ values at superfamily level were calculated using only the proteins in the same test set that were used in the single-MCS method. The mean top1 accuracy of the hierarchical-MCS method at superfamily level is 0.70±0.16. The highest top$_1$ accuracy is 0.76, which is achieved when the combination of physicochemical properties 'alpha+pK+atom' in the multi-class superfamily classifier and the RandomSubSpace technique are used. In the hierarchical-MCS method, the RandomSubSpace strategy reached the highest top$_1$ accuracy in 11 out of the 15 combinations of physicochemical properties.

Table 4. *Top$_1$ accuracy achieved by the hierarchical-MCS method when using 15 combinations of physicochemical properties and six different classification techniques.*

Binary classifier	Logit-	Random	Rando	SVM	SVM	SVM
	Boost	Subspace	Forest	(c=10)	(c=20)	(c=100)
remote-3DI (alpha helix)	0.71	0.72	0.73	0.71	0.71	0.71
remote-3DI (Hydropathy)	0.73	0.71	0.67	0.68	0.68	0.68
remote-3DI (pK (-COOH))	0.73	0.71	0.70	0.70	0.70	0.70
remote-3DI (alpha+atom)	0.66	0.67	0.66	0.66	0.66	0.66
remote-3DI (alpha+pK)	0.71	0.71	0.71	0.71	0.71	0.71
remote-3DI (alpha+C)	0.70	0.72	0.63	0.67	0.67	0.67
remote-3DI (pK+atom)	0.70	0.72	0.67	0.64	0.66	0.66
remote-3DI (pK+C)	0.71	0.71	0.64	0.70	0.69	0.71
remote-3DI (atom+C)	0.68	0.72	0.68	0.70	0.70	0.70
remote-3DI (hydropathy+pK)	0.70	0.69	0.67	0.69	0.69	0.69
remote-3DI (alpha+pK+atom)	0.74	0.76	0.74	0.74	0.74	0.74
remote-3DI (alpha+pK+C)	0.71	0.71	0.70	0.71	0.71	0.71
remote-3DI (alpha+atom+C)	0.71	0.71	0.69	0.71	0.71	0.71
remote-3DI (pK+atom+C)	0.69	0.69	0.67	0.68	0.68	0.68
remote-3DI (alpha+pK+atom+C)	0.71	0.67	0.70	0.69	0.71	0.71

The main difference between the single-MCS and the hierarchical-MCS methods is the fact of predicting the SCOP class and then using a multi-class superfamily predictor. In the single-MCS method, only one multi-class superfamily is trained using 23 class labels (i.e., the superfamilies). In the hierarchical-MCS method, a classifier is trained for the superfamilies that are inside each SCOP class. There are five superfamilies in class 1, eight superfamilies in class 2, five superfamilies in class 3, and five superfamilies in class 7. Once the SCOP class is predicted for a given protein, we predict its superfamily considering only the superfamilies inside the predicted SCOP class. Having a multi-class superfamily classifier inside each SCOP class allowed us having fewer class labels, which actually improved the prediction. As observed, the hierarchical-MCS method achieves higher top1 values than the single-MCS method.

3.5 Evaluating the proposed methods on more recent SCOP versions

The SCOP 1.53 data set is considered a gold standard in remote homology detection. Therefore, the 3D models enriched with physicochemical properties used in this research were obtained from that SCOP version in order to use the same data set that has been used in previous works. Since more recent versions of the SCOP data set have been released (i.e., SCOP 1.75 in 2009 and SCOP 2.04 in 2014), new collections of models should be obtained to include these versions in the experiments. However, in addition to the SCOP 1.53 data set, we also used the SCOP 1.55 and the SCOP 1.61 data sets during the experiments because these SCOP versions are close to the SCOP 1.53 data set, and thus, the models are still suitable. The SCOP 1.55 data set has 3527 proteins divided in 51 families. The SCOP 1.61 data set has 10569 proteins divided in 206 families. There are some families in common between the SCOP 1.53 and the SCOP 1.55 data sets, such as the families in the superfamilies 2.1.1 and 3.2.1. Some other families were not included in the SCOP 1.53 but they appear in the SCOP 1.55 data set, such as the families 3.66.1.10 and 6.2.1.2. In this section, we retrained the multi-class superfamily predictors using the proteins in some superfamilies of the SCOP 1.55 (i.e., 3.1.8) and the SCOP 1.61 (i.e., 1.4.1, 3.1.8, and 3.2.1) data sets. Including all the superfamilies in the SCOP 1.55 and the SCOP 1.61 data sets has to be done by building a binary classifier for each superfamily and then retraining the multi-class superfamily classifiers. Therefore, we built a new binary classifier and retrained the multi-class predictor only for the superfamilies 3.1.8 in the SCOP 1.55 and superfamilies 1.4.1, 3.1.8, and 3.2.1 in the SCOP 1.61 data set. The mean top1 accuracy on the SCOP 1.55 using the single-MCS (i.e., with SVM c=100 and ‘alpha helix propensity’) and the hierarchical-MCS (i.e., with RandomSubSpace and 'alpha+pK+atom') methods increased to 0.75 and 0.88, respectively. The mean top1 on the SCOP 1.61 data set using the single-MCS and the hierarchical-MCS methods increased to 0.65 and 0.81, respectively.

4. Conclusions

In this paper, two multi-class superfamily predictors were presented. The mean top1 accuracy of the single-MCS and the hierarchical-MCS methods on the SCOP 1.53 data set are 0.50±0.11 and 0.70±0.16, respectively. The mean top1 value of the single-MCS and the hierarchical-MCS methods increased on both the SCOP 1.55 and the SCOP 1.61 data sets. The main result achieved in this research is that using a hierarchical strategy allows increasing the accuracy of the superfamily prediction. When the SCOP class of a protein is initially predicted, the prediction of the superfamily becomes an easier task because fewer class labels are considered in the multi-class superfamily recognition.

We also found that the more accurate the binary classifiers, the higher the top1 accuracy in the superfamily prediction. A higher ROC score of a binary classifier means that the positive samples (i.e., the remote homologs) get higher scores than the negatives samples. When the binary classifier of a given superfamily S has a high ROC score, it means that proteins in that superfamily are getting higher scores than the proteins that do not belong to S. The results suggest that having binary classifiers with high ROC scores allows identifying the correct superfamilies in the test set. Testing on different combinations of physicochemical properties might help to improve the accuracy of the multi-class superfamily prediction. We found that more accurate binary classifiers actually help to increase the top1 accuracy, and thus, working on improving the ROC scores of the binary classifiers should be done.

5. References

Bedoya, O. & Tischer, I. (2015). Remote homology detection of proteins using 3D models enriched with physicochemical properties. *Ingeniería y Competitividad* 17 (1), 75-84.

Ding, C., Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 17 (4), 349-358.

Huang, C., Lin, C. & Pal, N. (2003). Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification. *IEEE Transactions on Nanobioscience* 2 (4), 221-232.

Le, E., Weston, J., Noble, W. & Leslie, C. (2007). Multi-class protein fold recognition using adaptive codes. *Journal of Machine Learning Research* 8 (1), 1557-158.

Leslie, C., Melvin, I., Ie, E., Kuang, R., Weston, J. & Stafford, W. (2007). SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. *BMC Bioinformatics* 22 (8), Suppl 4:S2.

Lin, H., Li, Q. (2007). Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. *Biochemical and Biophysical Research Communications* 354 (2), 548–551.

Lin, H. (2008). The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *Journal of Theoretical Biology* 252 (2), 350-356.

Rangwala, H. & Karypis, G. (2006). Building multiclass classifiers for remote homology detection and fold recognition. *BMC Bioinformatics* 7 (1), 455-467.