Ensemble of Diversely Trained Support Vector Machines for Protein Fold Recognition

Abdollah Dehzangi^{1,2}, Abdul Sattar^{1,2}
{a.dehzangi, and a.sattar}@griffith.edu.au

¹ Institute for Integrated and Intelligent Systems (IIIS), Griffith University,
Brisbane, Australia

² National ICT Australia (NICTA), Brisbane, Australia

Abstract. Protein Fold Recognition (PFR) is defined as assigning a given protein to a fold based on its major secondary structure. PFR is considered as an important step toward protein structure prediction and drug design. However, it still remains as an unsolved problem for biological science and bioinformatics. In this study, we explore the impact of two novel feature extraction methods namely overlapped segmented distribution and overlapped segmented autocorrelation to provide more local discriminatory information for the PFR compared to previously proposed methods found in the literature. We study the impact of our proposed feature extraction methods using 15 promising physicochemical attributes of the amino acids. Afterwards, by proposing an ensemble Support Vector Machines (SVM) which are diversely trained using features extracted from different physicochemical-based attributes, we enhance the protein fold prediction accuracy for up to 5% better than similar studies found in the literature.

Keywords: Overlapped segmented distribution, Overlapped segmented autocorrelation, Physicochemical-based features, Ensemble of different classifiers, Support vector machine

1 Introduction

Protein Fold Recognition (PFR) is considered as an important step towards protein structure prediction problem and drug design. It provides critical information of the classification of proteins based on their general major secondary structure. In pattern recognition perspective, PFR is defined as solving a multi-class classification task which its prediction performance depends on features and classification techniques being used. During the past two decades, a wide range of classification techniques such as: *Artificial Neural Network (ANN)* [1,2], Meta Classifiers [3,4], and *Support Vector Machines (SVM)* [5] have been used to solve this problem. Among the employed classifiers, ensemble of different classifiers attained the best results for PFR [6,7]. Similar to exploring the classification techniques, a wide range of features have been proposed and used

to solve this problem. Features being using for PFR can be generally categorized into three groups namely, physicochemical-based features (extracted from the physicochemical properties of the amino acids (e.g. hydrophobicity) [8–10]), sequential-based features (extracted from the alphabetic sequence of proteins (e.g. occurrence of amino acids [11])), and evolutionary-based features (extracted from the *Position Specific Scoring Matrix (PSSM)* [12]). Among these groups, physicochemical-based features are the only group that maintain their discriminatory information when the sequential similarity rate is low. Therefore, they attract tremendous attention for PFR [4, 7, 13].

To explore the impact of different physicochemical-based attributes for PFR, Gromiha and his co-workers [8] studied 49 different properties of the amino acids. they extracted 49 features based on the global density of these attributes (49-D feature vector) and used for the PFR. Despite a wide range of attributes explored in this study, they failed to properly explore the local discriminatory information of these attributes (due to use of a feature that solely describe the global density of a given attribute). To explore more local information for PFR, later studies shifted the focus to more sophisticated feature extraction methods [9, 14]. However, they merely relied on a few popular physicochemical-based attributes for feature extraction (e.g. hydrophobicity, polarity, flexibility). Furthermore, in all these studies, the whole protein sequence used as a building block of extracting local discriminatory information. Therefore, they failed to provide adequate local discriminatory information for large proteins.

In this study, we aim at enhancing the protein fold prediction accuracy by addressing the limitations highlighted earlier in the following two steps. First, to provide more local information compared to previously explained methods, we propose two feature extraction methods namely, overlapped segmented distribution and overlapped segmented autocorrelation. We explore the impact of our proposed feature extraction method for 15 most promising physicochemical-based attributes using five classification techniques that attained good results for PFR (Adaboost.M1, SVM, Random Forest, Naive Bayes, and *Ensemble of Different Classifiers (EDC)* proposed in [2]). Then, by proposing an ensemble of diversely trained SVM classifiers using features extracted from different physicochemical-based attributes, we enhance protein fold prediction accuracy up to 5% better than similar studies found in the literature.

2 Data Sets and Features

2.1 Data Sets

To investigate the performance and generality of our proposed methods, two datasets namely, EDD (extended version of the DD dataset introduced by [9]) and TG (introduced by [11]) are used in this study. We extract the EDD data set from the latest version of *Structural Classification of Proteins (SCOP)* (1.75) in the similar manner used in [5] to replace the old DD dataset which no longer used due to its inconsistency with the SCOP 1.75. This dataset consist of 3418 proteins with less than 40% sequential similarity belonging to 27 folds as it was

used in DD. The EDD dataset is mainly used to investigate the performance of our proposed method compared to similar studies found in the literature. We also used the TG benchmark extracted by [11] from the SCOP 1.73 consists of 1612 proteins with less than 25% sequential similarities belonging to 30 most populated folds in SCOP. The TG benchmark is mainly used to investigate the performance of our proposed approaches when the sequential similarity rate is low. To simulate the DD dataset condition and to be able to directly compare our results with previous studies, we divided each of our employed datasets into train and test sets (in the manner that 3/5 of the data is used in training set and 2/5 of the data is used in the testing set [9]).

2.2 Physicochemical-based Attributes

In this study, we explored the impact of our proposed approaches for 15 most promising physicochemical-based attributes. These attributes were selected by the authors from a wide range of physicochemical-based attributes explored experimentally. We studied the performances of features extracted from 115 different physicochemical-based attributes (extracted mainly from the APD database [15] and the [8] study) and selected the following 15 attributes: (1) structure derived hydrophobicity value, (2) polarity, (3) average long range contact energy, (4) average medium range contact energy, (5) mean RMS fluctuational displacement, (6) total non-bounded contact energy, (7) amino acids partition energy, (8) normalized frequency of alpha-helix, (9) normalized frequency of turns, (10) hydrophobicity scale derived from 3D data, (11) HPLC parameters to predict hydrophobicity and antigenicity, (12) average gain ratio of surrounding hydrophobicity, (13) mean fractional area loss, (14) flexibility, and (15) bulkiness. Note that to the best of our knowledge, most of the selected attributes (attributes number 3, 4, 5, 6, 7, 10, 11, 12, 13, and 14) have not been adequately (or not at all) explored for the PFR. However, the conducted comprehensive experimental study showed they are able to outperformed many popular attributes that have been widely used for PFR [1, 7, 9, 10].

3 Proposed Feature Extraction Methods

In this study, we propose two novel feature extraction methods namely overlapped segmented distribution and overlapped segmented autocorrelation. The proposed feature extraction methods are aimed at providing more local discriminatory information than previously used approaches found in the literature. These approaches are discussed in the following subsections.

3.1 Overlapped Segmented Distribution

As it was highlighted earlier, previously, global density was used as descriptor of a given physicochemical-based attributes in [8]. To calculate this feature, the amino acids in a given protein sequence ($A_1, A_2, \dots, A_L$ where L is the length of

the protein) is first replaced with their numerical values assigned to them based on a given attribute ($R_1, R_2, \dots, R_L$). Then it is calculated as follow:

$$T_{glob\_dens} = \frac{\sum_{i=1}^L R_i}{L}. \quad (1)$$

However, it could not properly explore the discriminatory information, embedded locally in a given physicochemical-based attribute. To address this limitation, we propose overlapped segmented-based distribution feature set. This feature set is calculated as follows. Beginning from the left side of a given protein, we sum the attribute values of the amino acids until reaching to $K\%$ of the total sum (which is equal to $T_{sum} = T_{glob\_dens} \times L$) as follows:

$$C_k^l \leq (T_{sum} \times K)/100. \quad (2)$$

Then, the number of summed amino acids divided by the length of the protein is returned as the distribution of the first $K\%$ of the global density. We repeat this process for $2K, 3K, \dots, nK$, calculate the $C_{2K}^l, C_{3K}^l, \dots, C_{nK}^l$ and return the given distribution-based features accordingly (where $nK = 75$). Same process is conducted from the right side. We start from right side and for $K, 2K, \dots, nK$, we calculate the $C_{2K}^r, C_{3K}^r, \dots, C_{nK}^r$ and return the corresponding distribution-based features (Figure 1). Therefore, $75/K$ features ($n=75/K$) from each side are extracted in this feature set (totally $150/K = 75/K \times 2$). We also added the global density feature to this features as a global descriptor ($150/K + 1$ features in total). In this study, the distribution factor $K = 5$ is adopted due to its better performance attained experimentally compared to other distribution factors such as $K = 10$ and $K = 25$ [9,10]. We also adopt 75 as the overlapping factor which showed better performance with respect to the number of features generated compared to other factors. Hence 31 features extracted in this feature set ($150/5 + 1 = 31$).

Fig. 1. Segmented distribution-based feature extraction method

In this method, we calculated the distribution factor from both side of a protein sequence to bring the emphasize to the sides of proteins. To highlight the impact of the middle part of a protein sequence as well, we adopt overlapping style. In this manner, the impact of a give attributes with respect to each side of the protein sequence are represented.

3.2 Overlapped Segmented Autocorrelation

The overlapped segmented-based distribution which was introduced earlier, is mainly based on the density and distribution properties. In this section, we propose overlapped segmented-based autocorrelation which is based on the autocorrelation property. Autocorrelation of the amino acids have been widely used as an effective feature which reveals important information of how amino acids are ordered in the protein sequence [14]. However, previous approaches (even the most sophisticated ones (e.g. pseudo amino acid composition [14])) failed to properly explore the potential of this method [7].

Therefore, in this study, we propose the concept of segmented autocorrelation aiming to address this limitation. These feature set is extracted in the following manner. We first segments the protein sequence as it used in the overlapping segmentation-based distribution method (where segmentation factor $K = 10$ and overlapping factor 70 is adopted due to better performance attained experimentally). Then we calculate the autocorrelation with distance factor $D_F = 10$ (as it was shown the most effective value for this parameter in [16]) cumulatively. In the similar manner to overlapped segmented-based distribution method, we calculate these features starting from each side (left and right). We therefore calculate $T_F = 7 \times 10$ features from each side (totally $140 = T_F \times 2$ from both sides). We also added the autocorrelation calculated using the whole protein sequence as the global descriptor of this method (totally $150 = 140 + 10$ features). The autocorrelation in each segment is equal to:

$$A\_Ci, a = \frac{1}{(L \times (a/100) - i)} \sum_{j=m}^n S_j \times S_{j+i}, (i = 1, \dots, 10 \& a = 10, \dots, 70), \quad (3)$$

where a is the segmentation factor, m and n are respectively the begin and the end of a segment, and S_j is the attribute value for each amino acid.

4 Classification Techniques

In this study, we use four different classifiers namely, AdaBoost.M1, Random Forest, Naive Bayes, and SVM to evaluate the performance of the explored physicochemical-based attributes with respect to the proposed feature extraction methods. These classifiers are selected based on their performances attained in the previous studies for the PFR [2–5, 16]. These classifiers are briefly introduced as follows.

Naive Bayes: As the most popular Bayesian-based classifier is based on the naive assumption of independency of the studied features for a given task. Despite its simplicity, it attained promising results for this task [2]. Naive Bayes is also able to provide important information of the correlation between the features being used (better performance of Naive Bayes is the prove for low level of correlation between features while poor performance will support the high rate of correlation between them) [2].

AdaBoost.M1: Is introduced by [17] and is considered as the best-of-the-shelf meta classifier. Adaboost.M1 also attained promising results for the PFR [2]. It uses a classifier called base learner sequentially in K iteration and adopt the weight of the misclassified samples in each iteration to improve the performance. It builds its final structure by combining the output of each step for a given sample using majority voting as its final decision. In this study, Adaboost.M1 implemented in WEKA is used [18]. The C4.5 decision tree is used as its base learner and the number of its iterations is set to 100 ($K=100$) (as it is shown as the best parameter for this algorithm for the PFR [19]).

Random Forest: Is introduced by [20] based on the concept of bagging. It randomly select K subset of features and train K different classifiers (base learners) independency and then, combine their results using majority voting as its final decision. Despite its simplicity, it was shown as an effective classifier for different tasks as well as the PFR [3]. In this study, for the Random Forest (implemented in WEKA) $K = 100$ and random tree based on the gain ratio is used as its base learner.

Support Vector Machine: Is considered as the most promising classification technique which outperformed other individual classifiers for the PFR [5,16]. SVM aims as finding the *Maximal Marginal Hyperplane (MMH)* to minimizing the classification error. To find the appropriate support vector, it transforms the input data to different dimension based on the concept of kernel function. In this study, We employ SVM using *Sequential Minimal Optimization (SMO)* as a kind of polynomial kernel (implemented in WEKA) which its kernel degree is set to one ($p = 1$).

5 Results and Discussion

In the first step, we construct a feature vector based on each attribute explored in this study with respect to the proposed feature extraction methods. Therefore, for a given attribute a feature vector consists of two feature groups namely overlapped segmented distribution feature group (31 features) and overlapped segmented autocorrelation (150 features) is constructed. We also added the composition of the amino acids feature group (20 features) as well as the length of the protein sequence feature (1 feature) to these feature vectors as important sources of sequential-based features [1, 19]. Composition of the amino acid feature group consists of the percentage of occurrence of amino acids divided by the length of protein sequence which used in [9] and later works as an effective feature group that provide important sequential-based information. The length of the amino acids also showed as an effective features for the PFR [2,4]. Therefore, based on each attribute being explored in this study, a feature vector consist of 202 features ($31 + 150 + 20 + 1$) constructed to explore their local and global discriminatory information in detail. For the rest of this study, each feature group will be referred as *comb_numb* where *numb* is the number assigned to the corresponding physicochemical-based attribute in section 2.2.

We then apply the employed classifiers in this study (Naive Bayes, AdaBoost.M1, SVM, and Naive Bayes) to the constructed feather vectors and report the results in Table 1. We also study the performance of the *Ensemble of Different Classifier (EDC)* which we introduced in our previous work [2] to the extracted feature vectors to compare its results with the results achieved in previous studies. We first reproduce the results achieved in [10] using EDC (219 features) and achieve up to 48.8% and 41.1% prediction accuracies respectively for the EDD and TG datasets. We also reproduce the results using EDC for the features introduced in [9] (126 features) and 69D feature vector (49D in addition to 20 features for the composition feature group) and achieve to 47.6% and 36.6% for EDD and 40.7% and 33.0% for TG respectively.

Table 1. Results achieved (in percentage %) by using AdaBoost.M1 (Ada), Random Forest (RF), Naive Bayes (NA), SVM, and EDC for 15 feature vectors extracted from the explored attributes in this study (for both EDD and TG datasets).

Datasets	EDD					TG				
	Ada	Na	RF	SVM	EDC	Ada	Na	RF	SVM	EDC
Comb_1	45.8	24.9	40.6	50.1	50.3	37.4	22.2	33.1	37.7	39.3
Comb_2	45.4	28.8	38.5	49.7	51.1	35.7	24.0	31.7	40.9	39.3
Comb_3	45.8	26.3	41.2	49.6	49.5	38.6	22.4	34.9	41.3	40.7
Comb_4	43.9	25.6	39.3	47.9	46.7	35.3	24.5	34.4	37.9	37.9
Comb_5	42.9	24.4	40.3	50.5	50.0	35.8	17.5	31.3	39.8	39.1
Comb_6	45.1	19.5	39.9	51.3	50.5	39.1	17.5	34.7	42.4	43.9
Comb_7	42.6	23.0	38.9	49.2	49.0	35.5	17.5	30.6	39.4	39.0
Comb_8	45.8	29.0	40.6	48.1	48.6	36.0	23.0	36.3	39.1	40.7
Comb_9	43.0	23.3	39.8	48.1	47.1	35.8	17.8	32.8	40.2	39.3
Comb_10	45.6	26.9	38.3	52.4	52.0	36.3	22.1	33.0	41.0	42.7
Comb_11	43.4	25.0	39.4	50.6	50.1	33.9	20.7	32.0	42.1	40.4
Comb_12	43.3	24.0	39.2	49.7	49.7	38.5	20.8	35.2	40.4	40.5
Comb_13	43.9	23.2	38.6	52.4	52.8	37.7	18.9	34.2	40.5	41.6
Comb_14	43.6	24.2	38.3	50.7	50.2	36.8	20.0	31.7	40.4	42.3
Comb_15	42.9	19.1	38.4	48.1	48.1	35.3	17.7	34.2	40.5	39.1

As it is shown in Table 1. By using EDC to the Comb_13 feature vector (mean fractional area loss attribute which to the best of our knowledge, have not been used for the PFR), we achieve to 52.8% prediction accuracy, up to 4% better than previously reported results in similar studies for the EDD dataset. We also achieve up to 43.9% prediction accuracy using EDC to the Comb_6 feature vector (total non-bounded contact energy attribute which to the best of our knowledge, have not been adequately explored for the PFR), over 2.8% better than previously reported results in similar studies for the TG dataset. Our results emphasize on the effectiveness of our proposed feature extraction methods as well as the importance of physicochemical-based attributes explored in this study against the popular feature extraction methods and attributes have been used for PFR. As it is shown in Table 1, the enhancement for the EDD and TG datasets are achieved using different feature vectors. Hence, we also propose a fusion of diversely trained ensemble of SVM to explore the potentials of the explored attributes in conjunction with each other as well as defining a system that generally perform well for both EDD and TG benchmarks.

To achieve this goal, we propose an ensemble of four diversely trained SVM classifiers. SVM is selected due its it better performance compared to the other employed classifiers in this study as well as its promising performance for the PFR. We trained these four SVM classifiers with four feature vectors extracted from four physicochemical-based attributes that attained the best results mainly for TG dataset among the explored attributes in this study (which raises the idea of exploring more optimal approach for this task in future studies.). We fuse these classifiers and use EDC (trained on Comb_10 which attained persistent results for both EDD and TG datasets) as a tie breaker in our proposed system. The voting system works in the following manner. For the case that majority of SVM classifiers, classify a given sample to a same fold, this fold will also be chosen as the decision and consequently output fold without consideration of the output of the EDC. In case that not a single fold would have the majority votes, the output of the EDC will be chosen as the output of the system (which occur when two of the SVM classifiers vote for a same fold and other two SVM classifiers for another fold or all four SVM classifiers vote for different folds). The architecture of our proposed *Ensemble of Diversely Trained SVM Classifiers (EDTSVM)* is shown in Figure 2.

Fig. 2. The overall architecture of the EDTSVM.

By applying the EDTSVM to the EDD and TG datasets, we achieve up to 53.8% and 43.5% prediction accuracies, up to 5% and 2.4% better than the best results reported in similar studies found in the literature. We also achieved up to 17.2% and 10.5% better prediction accuracy than using 69D for the EDD and TG datasets, respectively. These highlights the impact of the proposed feature extraction methods in this study to reveal significant discriminatory information based on an individual attribute rather than using a naive feature extraction method for a wide range of physicochemical-based attributes. The comparison of the results achieved in this study, compared to the similar studies found in the literature for the TG and EDD benchmarks is shown in Table 2.

5.1 Conclusion and Future Works

In this study, we explored the impact of 15 physicochemical-based attributes using two novel feature extraction methods namely, overlapping segmented-based distribution and overlapping segmented-based autocorrelation. We then, constructed a feature vector consisting of combination of features extracted using

Table 2. The best results (in percentage %) achieved in this study compared to the best results found in the literature for the EDD and TG benchmarks respectively.

Study	Attributes (Number of features)	Method	EDD (Results)	TG (Results)
[9]	Features proposed in [9] (126)	SVM	46.3	38.5
[19]	Features proposed in [9] (126)	Ada	44.7	36.4
[3]	Features proposed in [9] (126)	RF	42.9	37.1
[2]	Features proposed in [9] (126)	EDC	47.6	40.7
[2]	Features proposed in [10] (219)	SVM	47.3	40.1
[19]	Features proposed in [10] (219)	Ada	45.3	37.2
[3]	Features proposed in [10] (219)	RF	43.9	38.1
[2]	Features proposed in [10] (219)	EDC	48.8	41.1
[8]	69D (49+20)	SVM	36.6	33.0
This study	Comb.6 (202)	SVM	51.3	42.4
This study	Comb.10 (202)	SVM	52.4	41.0
This study	Comb.11 (202)	SVM	50.6	42.1
This study	Comb.6 (202)	EDC	50.6	43.9
This study	Comb.10 (202)	EDC	52.0	42.7
This study	Comb.13 (202)	EDC	52.8	41.6
This study	Fused (202 for each classifier)	EDTSVM	53.8	43.5

our feature extraction methods as well as composition of the amino acids feature group and the length of protein sequence feature. Then by using several classifiers that attained good results for the PFR such as, Random Forest, AdaBoost.M1, Naive Bayes, SVM, and EDC (proposed in [2]), we studied the effectiveness of our proposed approaches. Achieved results showed the impact of our proposed feature extraction methods with respect to the attributes being used compared to the similar studies found in the literature. Finally, By proposing an ensemble of diversely trained SVM classifiers (EDTSVM) applied to the feature vectors extracted with respect to the physicochemical-based attributes that have not been adequately explored for PFR, we achieved up to 5% prediction accuracy compared to the similar studies found in the literature.

For our future works, we aim to explore the impact of evolutionary-based information in conjunction with physicochemical-based information for PFR. We also aim to explore the impact of weighted ensemble of different classifiers based on the classifier as well as features being used.

References

1. Ghanty, P., Pal, N.R.: Prediction of protein folds: Extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. *NanoBioscience, IEEE Transactions on* **8**(1) (2009) 100–110
2. Dehzangi, A., Phon-Amnuaisuk, S., Ng, K.H., Mohandesi, E.: Protein fold prediction problem using ensemble of classifiers. In: *Proceedings of the 16th International Conference on Neural Information Processing: Part II. ICONIP '09* (2009) 503–511
3. Dehzangi, A., Phon-Amnuaisuk, S., Dehzangi, O.: Using random forest for protein fold prediction problem: An empirical study. *Journal of Information Science and Engineering* **26**(6) (2010) 1941–1956
4. Chen, K., Kurgan, L.A.: Pfrs: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics* **23**(21) (2007) 2843–2850

5. Yang, J.Y., Chen, X.: Improving taxonomy-based protein fold recognition by using global and local features. *Proteins: Structure, Function, and Bioinformatics* **79**(7) (2011) 2053–2064
6. Dehzangi, A., Karamizadeh, S.: Solving protein fold prediction problem using fusion of heterogeneous classifiers. *INFORMATION, An International Interdisciplinary Journal* **14**(11) (2011) 3611–3622
7. Yang, T., Kecman, V., Cao, L., Zhang, C., Huang, J.Z.: Margin-based ensemble classifier for protein fold recognition. *Expert Systems with Applications* **38** (2011) 12348–12355
8. Gromiha, M.M., Oobatake, M., Sarai, A.: Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophysical Chemistry* **82** (1999) 51–67
9. Ding, C., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17** (2001) 349–358
10. Dehzangi, A., Phon-Amnuaisuk, S.: Fold prediction problem: The application of new physical and physicochemical- based features. *Protein and Peptide Letters* **18**(2) (2011) 174–185
11. Taguchi, Y.H., Gromiha, M.M.: Application of amino acid occurrence for discriminating different folding types of globular proteins. *BMC Bioinformatics* **8**(1) (2007) 404
12. Kurgan, L.A., Cios, K.J., Chen, K.: Scpred: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC Bioinformatics* **9** (2008) 226
13. Kavousi, K., Moshiri, B., Sadeghi, M., Araabi, B.N., Moosavi-Movahedi, A.A.: A protein fold classifier formed by fusing different modes of pseudo amino acid composition via pssm. *Computational Biology and Chemistry* **35**(1) (2011) 1–9
14. Shen, H.B., Chou, K.C.: Ensemble classifier for protein fold pattern recognition. *Bioinformatics* **22** (2006) 1717–1722
15. Mathura, V.S., Kolippakkam, D.: Apdbase: Amino acid physico-chemical properties database. *Bioinformation* **12**(1) (2005) 2–4
16. Dong, Q., Zhou, S., Guan, G.: A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics* **25**(20) (2009) 2655–2662
17. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: *International Conference on Machine Learning*. (1996) 148–156
18. Witten, I., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*. 2 edn. Morgan Kaufmann, San Francisco (2005)
19. Krishnaraj, Y., Reddy, C.K.: Boosting methods for protein fold recognition: An empirical comparison. In: *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*. (2008) 393–396
20. Breiman, L.: Bagging predictors. *Machine Learning* **24** (1996) 123–140