Stability of different feature selection methods for selecting protein sequence descriptors in protein solubility classification problem

Simon Kocbek[1], Gregor Stiglic[1], Igor Pernek[1], Peter Kokol[1,2]
1 Research Institute, Faculty of Health Sciences, University of Maribor, Zitna ulica 15, 2000 Maribor, Slovenia
2 Laboratory for System Design, Faculty of Electrical Engineering and Computer Science, Smetanova ulica 17, 2000 Maribor, Slovenia
simon.kocbek@uni-mb.si

Abstract

Predicting protein solubility has gained lots of intention in the recent years and several descriptors have been defined to describe proteins in these works. Therefore, different feature selection methods have been used for selecting the most important attributes. An empirical study, that aims to explain the relationship between the number of samples and stability of seven different feature selection techniques for protein datasets, is presented.

1. Introduction

Understanding protein structure and its behaviour is important research area in the bioinformatics field. Therefore, classifying proteins has gained lots of attention in recent years. Researchers try to analyse protein primary structure (sequence of different amino acids) and find answers on questions such as: prediction of structural and functional classes of proteins [1-5], prediction of secondary structure of proteins [6], protein-protein interactions prediction [7-10], subcellular location prediction [11-14] and protein solubility prediction [15-20]. Several solutions with use of machine learning techniques [21] have emerged to solve these problems: k-nearest neighbour method [22], neural network [23], decision trees [24], support vector machines [15, 16] and ensemble methods [15].

Different sequence-derived structural and physicochemical descriptors have frequently been used as the input for these methods. They range from simple descriptors, such as sequence length and molecular weight of protein, to more complex ones, such as amino acid distribution and Geary autocorrelation [27]. Number of these descriptors can be very high and they can contain redundant information. Therefore, different feature selection techniques are often applied to select the most important descriptors. Stability of these methods is very important and removing or adding learning instances should not influence the feature subset selection. Therefore, the aim of this paper is to analyse and present results for stability of different popular feature selection methods in protein sequence descriptors space.

We will focus on protein solubility prediction problem since it presents an important part of research area [15-20] and gathering soluble proteins is often a difficult but an important challenge in biophysical studies. Many proteins are insoluble when over-expressed in bacteria, therefore, targeting soluble protein is often a trial and error process with low success rate [25]. Researchers can use an alternative way to target soluble proteins. With use of machine learning algorithms they can predict which proteins have higher chance to be soluble.

Several methods have been developed to cope with protein solubility prediction problem in recent years. The first simple method was introduced by Wilkinson and Harrison in 1991 [18] and it was improved in 1999 [20]. In 2004 Goh et al. [26] used random forest algorithm and in 2006 Idicula-Thomas et al. [16] used support vector machines to predict the protein solubility. The latter was improved in 2007 when a secondary classifier which based on Naive Bayes algorithm was added [17]. All mentioned methods try to optimize the classification performance of the protein solubility problem based on features mainly derived from protein sequences. On the other hand, we try to evaluate different feature selection methods to

978-1-4244-9166-7/10/$26.00 ©2010 IEEE

demonstrate the difference between ranking of protein descriptors based on chosen feature selection method.

In the next section we will describe the dataset of soluble and insoluble proteins that we used. We will also present the features that describe those proteins. In Section 3 we will explain the feature selection and evaluation methods which we used. In Section 4 we will present the results which will be discussed in the final section.

2. Dataset description

Dividing proteins into soluble and insoluble is a hard task since there is no publicly available dataset which would unambiguously describe protein solubility property. Most of the databases, that provide the information on the solubility of proteins, often do not offer detailed information about the experimental details under which solubility was assessed. Moreover, researchers usually deal with redundant and unbalanced data when gathering soluble proteins. Therefore, several research groups tried different approaches to gather reliable protein datasets which would divide proteins into soluble and insoluble groups.

We decided to use the SOLP [15] dataset which was collected in one of the recent protein solubility prediction studies and copes well with the above problems. SOLP contains 17 408 of non redundant proteins expressed in E.coli. Its proteins were collected from three different databases: the PDB (Protein Data Bank), the SwissProt and the TargetDB database. Additionally, these proteins were merged with the proteins used in study made by Idicula-Thomas and Balaji [28]. Furthermore, the sequence redundancy was removed with a rigorous threshold, which was 25% sequence similarity. The level of 25% is considered to be necessary to sufficiently reduce the bias introduced by homologue protein sequences [29]. The SOLP database is balanced and it contains equal number of soluble and insoluble proteins.

We used several sequence-derived structural and physicochemical descriptors to characterize our sequences (instances). In particular, we used the Protein Feature Server (PROFEAT) tool [30] to obtain several descriptors that have been previously often used in protein functional and structural prediction studies. PROFEAT has a web interface which allows calculation of 1497 different physicochemical values for up to 1000 proteins at once. Due to the space limitations, a short overview of the features grouped into seven groups can be seen in Table 1.They are described in details in PROFEAT's manual. Note that all these features are numerical values.

3. Methods

Our experimental study consisted of several steps which can be seen in Figure 1 and will be described in Section 3.2. First, we will make an overview of the feature selection methods that were used in the feature selection step and followed by supervised classification of proteins into soluble vs. insoluble group using 10-fold cross-validation.

3.1. Feature selection methods

Seven widely used feature selection methods that are implemented in the WEKA [31] machine learning environment were used in this study:

Table 1: An overview of the features used in the experiment.

Feature Group	Feature	No. of Descriptors	No. of Descriptor Values
Amino acid, dipeptide composition (G1)	Amino acid composition	1	20
	Dipeptide composition	1	400
Autocorrelation 1 (G2)	Normalized Moreau-Broto autocorrelation	8	240
Autocorrelation 2 (G3)	Moran autocorrelation	8	240
Autocorrelation 3 (G4)	Geary autocorrelation	8	240
Composition, transition and distribution (G5)	Composition	7	21
	Transition	7	21
	Distribution	7	105
Sequence order 1 (G6)	Sequence-order-coupling number	2	60
	Quasi-sequence-order descriptors	2	100
Sequence order 2 (G7)	Pseudo amino acid descriptors	1	50 (sequence length (SL) >= 30) 20 + SL – 1 (SL < 30)

Information Gain (IG), ReliefF (RF), Support Vector Machines Recursive Feature Elimination (SvmRfe), Gain Ratio (GR), Chi Squared (CS), One attribute rule (OR) and Symmetrical Uncertainty (SU).

The IG technique is based on measuring the decrease in entropy when a feature is selected or absent from the dataset.

The RF technique ranks the features based on their ability to distinguish between instances that are near to each other. The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes.

SvmRfe uses a linear Support Vector Machines (SVM) algorithm as the learning algorithm in the recursive selection of nested subsets of features. In the final step of each recursive cycle, all feature variables are ranked and a pre-selected number of the lowest ranked features are eliminated. In our experiments, 50% of the remaining features are removed in each cycle, since previous research has shown that feature selection performance with 50% was much faster with no significant classification performance loss.

CS is based on the chi-square test procedure which tabulates a variable into categories and computes a chi-square statistic. This goodness-of-fit test compares the observed and expected frequencies in each category to test that all categories contain the same proportion of values or test that each category contains a user-specified proportion of values

OR is an algorithm for finding association rules and it uses a simple accuracy measure. It has been shown that very simple association rules, involving just one attribute in the condition part, often work well in practice with real-world data. The idea of the OR algorithm is to find the one attribute to use to classify a novel data point that makes fewest prediction errors.

The SU algorithm evaluates a single attribute with measuring its symmetric uncertainty with respect to the class attribute. SU measures the correlation between features. Symmetrical uncertainty is defined as:

Where $IG(X_i \mid X_j)$ is the information gain between features X_i and X_j, $H(X_i)$ and $H(X_j)$ denote the entropies of X_i and X_j respectively.

3.2. Stability evaluation

Because of the computational complexity of our classification scheme we randomly selected 100 insoluble and 100 soluble proteins from the SOLP database and merged them into the SOLPmini database which contained 200 proteins. In the next step, we randomly split SOLPmini proteins into two halves SOLPminiG1 and SOLPminiG2 and on each of the group we performed each of the seven feature selection methods. For every feature selection method, we selected 25 to 1475 features in steps of 25 and in the last step we ranked all the 1497 features. In each step we calculated overlap of selected features from SOLPminiG1 and SOLPminiG2.

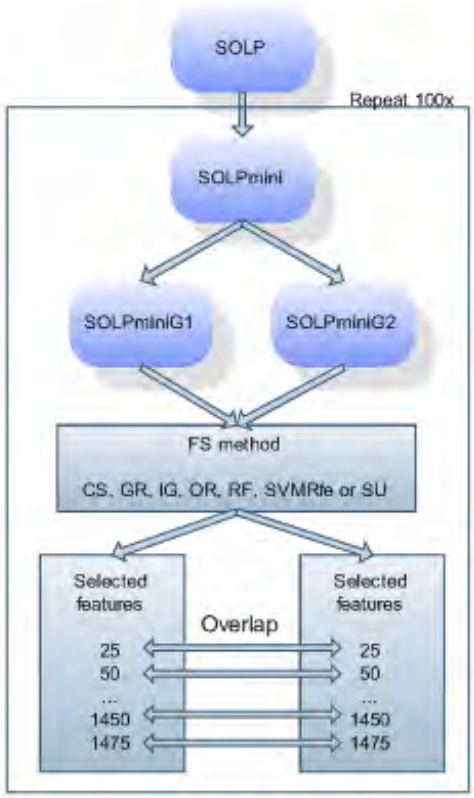

Figure 1: The experimental setup

Overlap is one of the simplest measures of similarity, where the similarity of the two lists ( ) of features is not based on the ranking of the features. The degree of similarity is calculated by simple counting of the features that are present in both the lists

and dividing them by the number of features in each list. The top-k features overlap can be defined as:

Where:

The step of splitting SOLPmini and calculating overlap of selected genes was repeated 100 times using randomized shuffling of samples. In the final step we also ranked all 1497 features with all the feature selection methods.

4. Results

In all experiments we calculated average overlap and accuracy of classification using SVM which was also used in similar protein classification studies [1-3]. These studies demonstrated the superiority of SVM over competitive classification methods. The classification results from our study do not differ from the above studies, therefore classification accuracy results are not included in this paper.

Figure 2 represents the overlap results for all seven feature selection methods. The vertical axis shows the overlap result with range from 0.0 to 1.0 and the horizontal axis shows the number of selected features.

First, we can notice that the feature selection methods form two distinctive groups. In the first group (FG1) we can find methods OR, SvmRfe and RF, while the second group (FG2) consist of CS, GR, IG and SU. Members of FG2 are representatives of univariate feature selection methods that evaluate features one by one. The first group of methods (multivariate) considers information from multiple features used in the selection process simultaneously.

Second, we can notice that overlap results grow with number of selected features which was expected. Members of FG1 have almost linear growth from the lowest overlap 0.05 (OR). These methods showed to be unstable when we choose low numbers of features. Methods in FG2 reach stable feature overlap evaluation earlier, with low selected attributes (average overlap 0.8 with 100 selected attributes).

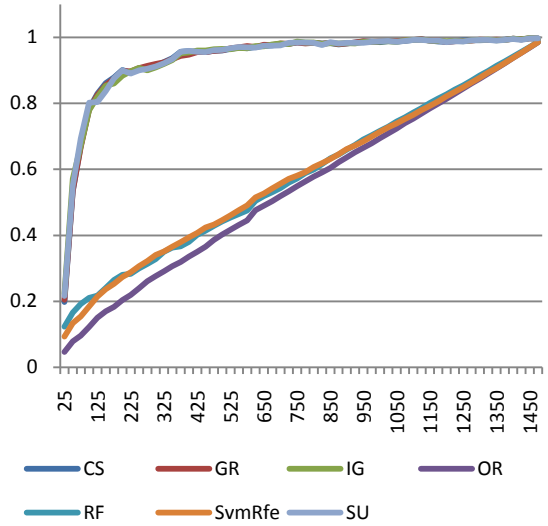

Figure 2: Average overlap for seven different feature selection methods.

Table 2 shows 10 best ranked features when performing feature selection on all 1497 features with two members of FG1 and FG2 with best average overlap score. In G1 that was the RF method, which has lowest overlap score of 0.12 and average score of 0.58. Differences between average overlaps of methods in FG2 were not so obvious and all the methods had average overlap of 0.94. However, IG had the highest minimum overlap, so we chose this method as FG2 representative.

The prefix Gi, where $0 < i < 8$, defines the group of features (defined in Table 1) that a single feature belongs to. We notice that 10 most important features ranked by both methods belong to G6 and G5.

Table 2: Top 10 ranked genes for RF and IG.

Rank	ReliefF (FG1)	Information Gain (FG2)
1	G6_QsoDesc_QSOD2	G6_QsoDesc_QSOD1
2	G6_QsoDesc_QSOD1	G5_Distribution_Polarity
3	G6_QsoDesc_QSOD2	G6_QsoDesc_QSOD2
4	G5_Distribution_Polarity	G5_Distribution_Hydrophobicity
5	G6_SocNum_SOCN2(24)	G5_Distribution_Normalized
6	G6_SocNum_SOCN2(22)	G5_Distribution_Polarizability
7	G5_Distribution_Hydrophobicity(3)	G6_SocNum_SOCN2(21)
8	G6_SocNum_SOCN2(26)	G6_SocNum_SOCN2(2)
9	G6_SocNum_SOCN2(18)	G6_SocNum_SOCN2(28)
10	G6_SocNum_SOCN2(7)	G6_QsoDesc_QSOD2

G6 consists of quasi-sequence-order descriptors which were proposed by K.C.Chou, et.al [10]. They are derived from the physicochemical distance matrix between each pair of the 20 amino acids. The physicochemical properties computed include hydrophobicity, polarity, and side-chain volume.

Descriptors from G5 have been developed by Dubchak, et.al [4] and used by several research groups. Features are computed by the following procedure: the protein sequence (amino acids) is transformed into sequences of certain structural or physicochemical properties/attributes of residues. Twenty amino acids are grouped into three groups for each of the seven different amino acid attributes representing the main clusters of the amino acid indices as described in [32]. Therefore, for each attribute, every amino acid is replaced by the index 1, 2, or 3 according to one of the three groups to which it belongs. The ranges of these numerical values and the amino acids belonging to each group are shown in [30]. In the next step, three descriptors: composition, transition and distribution, are computed for a given attribute to describe the global percent composition of each of the three groups in a protein, the percent frequencies with which the attribute changes its index along the entire length of the protein, and the distribution pattern of the attribute along the sequence, respectively.

5. Conclusion

In the paper we presented a stability assessment of popular feature selection methods for protein sequence descriptors. Results of seven different methods were analyzed based on their overlap scores. The methods can be grouped into two different groups: the univariate and the multivariate methods. The experiment indicated that the univariate methods outperformed the multivariate ones in the stability context. The only exception is OR which shows multivariate behavior. Performances of the methods in the univariate group are comparable and there are no significant differences in the overlap score between them. They reach good stability score with lower number of attributes compared to the multivariate methods. This indicates that researchers should use the univariate methods rather than multivariate ones if they want stable and robust feature selection methods when selecting low number of protein descriptors in protein solubility prediction problem. They reach good stability scores at around 100 selected features while the multivariate methods need more than 1000 features to reach the same stability score.

Moreover, we showed that members of two groups from the PROFEAT server contain top ranked features: the compesition-transition-distribution group and the quasi-sequence-order descriptors group. Further research should be done in this area.

Our method experimental study has few limitations. The biggest problem is the computational complexity of the experiment. The SOLP database contains over 17 000 protein sequences and due to limitations we had to pick only 200 instances. It would be ideal if we used all the instances or if we repeated the step, where we select the instances, few times.

Another problem is the limited feature space. We used the PROFEAT tool to calculate the attributes which has limited number of sequence features. In the future, other sequence derived features that have been proven to be important in previous works should be added to the experiment.

6. References

[1] R. Karchin, K. Karplus, and D. Haussler, “Classifying G-protein coupled receptors with support vector machines”, *Bioinformatics*, 2002, 18:147-159.

[2] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen X, and Y. Z. Chen, “SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence”, *Nucl Acid Research*, 2003, 31:3692-3697.

[3] C. Z. Cai, L. Y. Han, Z. L. Ji, Y. Z. Chen, “Enzyme family classification by support vector machines”, *Proteins*, 2004, pp. 55:66-76.

[4] I. Dubchak, I. Muchnick, C. Mayor, I. Dralyuk, S. H. Kim, “Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification”. *Proteins,* 1999, 35:401-407.

[5] L. Y. Han, C. Z. Cai, S. L. Lo, M. C. Chung, Y. Z. Chen, “Prediction of RNA binding proteins from primary sequence by a support vector machine approach”, *RNA*, 2004, 10:355-368.

[6] J. J. Ward, L. J. McGuffin, B. F. Buxton, D. T. Jones, “Secondary structure prediction with support vector machines”, *Bioinformatics,* 2003, 19(13):1650–1655.

[7] J. R. Bock, D. A. Gough, “Predicting protein - protein interactions from primary structure”, *Bioinformatics,* 2001, 17:455-460.

[8] J. R. Bock, D. A. Gough, “Whole-proteome interaction mining”, *Bioinformatics*, 2003, 19:125-134

[9] S. L. Lo, C. Z. Cai, Y. Z. Chen, M. C. Chung, "Effect of training datasets on support vector machine prediction of protein-protein interactions", *Proteomics*, 2005, 5:876-884.

[10] K. C. Chou, Y. D. Cai, "Predicting protein-protein interactions from sequences in a hybridization space", *J Proteome Res*, 2006, 5:316-322.

[11] K. C. Chou, Y. D. Cai, "Prediction of protein subcellular locations by GO-FunD-PseAA predictor", *Biochem Biophys Res Commun*, 2004, 320:1236-1239.

[12] K. C. Chou, H. B. Shen, "Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization". *Biochem Biophys Res Commun*, 2006, 347:150-157.

[13] K. C. Chou, H. B. Shen, "Large-scale plant protein subcellular location prediction", *J Cell Biochem*,2006, 100(3):665-678.

[14] J. Guo, Y. Lin, "TSSub: eukaryotic protein subcellular localization by extracting features from profiles", *Bioinformatics*, 2006, 22(14):1784-1785.

[15] C. N. Magnan, A. Randall, P. Baldi, "SOLpro: accurate sequence-based prediction of protein solubility", *Bioinformatics*, 2009, 25(17):2200-2207.

[16] S. Idicula-Thomas, A. J. Kulkarni, B. D. Kulkarni, V. K. Jayaraman, P. V. Balaji, "A support vector machine-based method for predicting thepropensity of a protein to be soluble or to form inclusionbody on overexpression in Escherichia coli", *Bioinformatics*, 2006, 22(3):278-284.

[17] P. Smailowski, A. J. Martin-Galiano, A. Mikolajka, T. Girchick, T., A. Holak, D. Frishman, "Protein solubility: sequence based prediction andexperimental verification", *Bioinformatics*, 2007, 23(19):2536-2542.

[18] D. L. Wilkinson, R. G. Harrison, "Predicting the solubility of recombinant proteins in Escherichia coli", *Nature Biotechnology*, 1991, 9:443-448.

[19] C.S. Goh, N. Lan, S. M. Douglas, B. Wu, N. Echols, A. Smith, D. Milburn, G. T. Montelione, H. Zhao, M. Gerstein, "Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis", *J. Mol. Biol*, 2004, 336:115-30.

[20] G. D. Davis, C. Elisee, D. M. Newham, R. G. Harrison, "New fusion protein systems designed to give soluble expression in Escherichia coli", *Biotechnol. Bioeng.*, 1999, 65:382–388

[21] E. Alpaydin, "Introduction to Machine Learning", *The MIT Press*, October 2004.

[22] J. Sim, S. Y. Kim, J. Lee, "Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method", *Bioinformatics*, 2005, 21(12):2844-2849

[23] C. Igel, J. Gebert, T. Wiebringhaus, "Protein Fold Class Prediction using Neural Networks with Tailored Early-Stopping", *Currents in Computational Molecular Biology*, The Seventh Annual International Conference on Research in Computational Molecular Biology, 2004.

[24] M. Singh, P. K. Wadhwa, S. Kaur, "Predicting Protein Function using Decision Tree", World Academy of Science, Engineering and Technology, 2008, 29:350-353

[25] S. M. Singh, A. K. Panda, "Solubilization and refolding of bacterial inclusion body proteins", *J. Biosci. Bioeng.*, 2005, 99(4):303-310

[26] C. S. Goh, , N. Lan, S. M. Douglas, B. Wu, N. Echols, A. Smith, D. Milburn, G. T. Montelione, H. Zhao, M. Gerstein, "Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis", *J. Mol. Biol.*, 2004, 336:115-130.

[27] S. Li, L. Xi, C. Wang, J. Li, B. Lei, H. Liu, X. Yao, "A novel method for protein-ligand binding affinity prediction and the related descriptors exploration", *Journal of Computational Chemistry*, 2009, 30(6):900-909.

[28] S. Idicula-Thomas, P. V. Balaji, "Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation", *Protein engineering, design & selection: PEDS*, 2005, 18(4):175-180

[29] B. Rost, "Twilight zone of protein sequence alignments". *Protein Eng.*, 1999, 12:85–94.

[30] Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, Y. Z. Chen, "PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence", *Nucleic Acids Research*, 2006, 34:32-37.

[31] I. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition*, Morgan Kaufmann, 2005.

[32] K. Tomii, M. Kanehisa, "Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins", *Protein engineering, design & selection: PEDS*, 1996, 9:27-36.