Oversampling Negative Class Improves Contact Map Prediction

Grzegorz Markowski, Krzysztof Grabczewski, and Rafal Adamczak
Department of Informatics, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Grudziadzka 5, 87-100 Torun, Poland
Email: grze.markowski@gmail.com, {kgrabcze, raad}@is.umk.pl

***Abstract*—In this paper we present a contact map predictor that has been trained using unbalanced training. The training set has been built based on typical, for this problem, feature space: predicted solvent accessibilities and predicted secondary structures. To show that oversampling negative class improves prediction accuracy we have built two predictors that are based on neural networks and decision trees, respectively. The influence of the size of the non-contact class in the training set has been analyzed. We have observed that significantly better results are obtained when the size of the non-contact class is at least 4 times larger than contact class, while the optimal oversampling depends on the type of contacts and learning algorithm used. Our final predictor - PLCT – took part in CASP11 where in one of the category took 3th place. PLCT is available at http://promap.is.umk.pl/.**

***Index Terms*—neural networks, decision trees, contact maps, contact maps prediction**

I. Introduction

Resolving protein structure is one of the ultimate tasks in structural biology. Experimental methods for protein structure determination such as X-ray crystallography, NMR (Nuclear Magnetic Resonance) are expensive, time consuming and in some cases not feasible. Therefore prediction of protein tertiary structure from the primary structure remains central problem of computational biology. There are many approaches to this problem, most of which relay on intermediate sequence-based predictions of amino acids residues attributes, including secondary structure [1], [2], solvent accessibility [3], [4] or contact map [5], [6]. These intermediate predictions have been shown to play a critical role in improving the overall accuracy of protein structure prediction.

The 3D structure of a protein can be represented by a two-dimensional matrix, called distance map, which contains distances between all residues in the protein structure. Distance maps are often simplified to their binary projection, called contact map. The distances are replaced by binary values that symbolize contacts. Two residues are in contact if they are close enough (according to some arbitrary threshold) to each other. Although contact map is just an approximation of the overall topology of the protein it is very useful in many applications including protein structure and fold prediction [7].

Many approaches for the contact map prediction from the primary sequence have been proposed. They may be divided into two types: pure statistical methods and those utilizing machine learning algorithms. The most important statistical methods are based on identification of correlated mutations [8], calculation of the mutual information between columns in the multiple sequence alignment [9] or recently published sparse inverse covariance matrix [10]. Some of these methods have also been utilized in machine learning approaches.

Application of machine learning methods involve 3 distinct phases: training set preparation, learning from data (often coupled with feature selection to identify the optimal representation of the problem at hands) and validation. Finding the most appropriate features that enable accurate discrimination of contact and not-contact classes is not an easy task. To date, many feature spaces for the problem of contact map prediction have been proposed. Many of previously published methods incorporate predicted secondary structures and solvent accessibilities as part of their feature spaces. Among the other features used in contact map prediction the most commonly used are: evolutionary information represented by profiles generated using PSI-BLAST [11], sequence conservations and correlated mutations [12], [13], hydrophobicity profiles of amino acids [6], some statistical measures such as: mutual information, correlation of the profiles at specific positions or pairwise potentials [5].

Neural networks are most often selected as the machine learning algorithm for contact map prediction [14], [15]. There is a large number of architectures of neural networks used for contact map predictions. Starting from simple ones that consist of one hidden layer with from 8 to 100 nodes and output layer with one or two nodes [16], [13] through more complex architectures that are utilizing sub-networks and final cascade network that combines results from sub-networks [17], or architectures that are based on recurrent neural networks [14] up to deep architectures [18]. Other machine learning algorithms, including SVMs [5] have also been applied for contact map prediction problem.

It has been argued in [19] and recently in [20] that for binary classification problems, and contact map prediction can be cast as the binary problem, it is better to

Manuscript received February 14, 2016; revised July 7, 2016.

doi: 10.18178/ijpmbs.5.4.211-216

use balanced training set. Following this argument, most published methods aims to build a balanced training set by including a roughly equal number of vectors from each class. Since there are much more of non-contact than contact cases balancing the training set can be achieved e.g. by randomly sampling non-contact instances. In this work we argue that using a balanced training set has some drawback: a lot of information that exists in discarded vectors is not exploited.

Although contact map prediction is the typical classification problem the accuracy measure used in many papers as well as in CASP competition is very special. The accuracy measure is focused only on contact class, the prediction of non-contact instances is ignored, besides only limited number of contacts are considered. Whereas, in the typical classification problem, the contact vs. non-contact class must be assigned to all vectors in the training and test set. The special character of accuracy measure for this problem cause that the conclusions of [20] that it is better to use balanced training set does not hold any more.

We present two methods trained on unbalanced training set that are based on neural network and decision trees. In both cases accuracy of prediction substantially improved after unbalanced training set utilization.

II. Contact Map Definition

There are several different definitions of residue-residue contacts. In this paper we follow the definition used in CASP competitions (http://predictioncenter.org/casp9/index.cgi?page=format), which states that two residues are in contact when the Euclidean distance of their respective C_β atoms (C_α- for glycines) is less than a given threshold and the sequence distance is sufficiently large. Given a protein of L residues, contact map S is evaluated as a binary symmetric L x L matrix, where L is the length of a protein sequence. An [i,j] element in matrix S is set to 1, when the distance between i-th and j-th residues is smaller than a cutoff value and 0 otherwise:

$$S_{ij} = \begin{cases} 1 if\ \Delta_{ij} < T_1, T_3 \geq |i - j \geq T_2| \\ 0 \end{cases} \quad (1)$$

where Δ_{ij} is the Euclidean distance between C_β atoms of i and j residue. We adopted 8 Å as T_1 distance threshold and 6 as T_2 - the sequence separation of two residues threshold. Because of the fact that we observed that our predictor was unable to correctly predict contacts between residues that are in sequence further then 40 residues we introduced additional threshold T_3 that limits the number of considered contacts.

In section Results we also present level of accuracy prediction that depends on the range of contacts. There, we follow definition of contact range used in CASP: S R - short range contacts $6 \leqslant |i\text{-}j| < 12$, MR - medium range contacts $12 \leqslant |i\text{-}j| < 24$, LR - long-range contacts $24 \leqslant |i\text{-}j|$, LRT - long-range contacts $24 \leqslant |i\text{-}j| \leqslant 40$.

III. Accuracy Measures

Most of the existing methods that utilize machine learning algorithms for contact map prediction balance the training set. This is usually important in typical classification problems where accuracy measures like sensitivity, specificity or Matthews correlation coefficient are used. However, in the case of protein contact prediction, typical accuracy measures are dominated by the prediction of the non-contact class. Consequently, although the prediction accuracy might be very high, according to the applied measures, its usefulness is very limited. That is why measures used to evaluate prediction do not take into account false negative predictions. Moreover only limited number of the best predicted contacts are considered. Here we are following measures that are introduced in CASP competition. The standard measure for performance evaluation of contact maps prediction is defined as:

$$Acc = \frac{TP}{TP + FP} * 100 \quad (2)$$

where TP (true positive) is the number of correctly predicted contacts, FP (false positive) is the number of non-contacts predicted as contacts. Out of all possible contacts only the best 5, L/10, L/5, L/2 predicted contacts, where L is the length of protein, are taken into account. Accuracy is evaluated for each protein in the test set separately and averaged over all tested proteins.

IV. Training and Test Sets

We trained our system using the same training set as the one used for training SABLE server (designed for secondary structure predictions) [3]. The dataset consists of 860 non-redundant proteins each of which is the representative of PFAM family. Proteins whose backbone is interrupted or the number of contacts is smaller than the number of the amino acids in the protein have been excluded. A dataset that consists of 617 proteins has been obtained.

There are two classes where the first class consists of residue pairs that are in contact so have the distance below specific threshold (8 Å) and all other residue pairs belong to the second class. The number of instances of first class in our training set is 165560 and is much lower than in the second class which was 3602025. Since the training file was huge, and we had limited resources, we decided to split it into three training sets that correspond to three types of contacts considered. Both classes contact and non-contact were built only from instances that correspond to the range considered, so for SR contact range we had 47315 contact instances and 830226 non-contact, MR – 57172 versus 1314767 and LR – 61073 and 1457032.

For test sets we used the same sets that were used for SABLE server evaluation. There are 603 protein chains (with 143000 residues) with no homology to proteins included in the training that were originally grouped into 4 datasets referred to as S156 (156 structures submitted to PDB from January through March of 2002), S135 (135

structures submitted from April through June), S163 (163 structures submitted from July through September), and S149 (149 structures submitted from October through December of 2002) but we created one test set that, in the paper, is called the TEST set. The list of protein structures in the training and all control sets can be downloaded from http://sable.cchmc.org.

V. Feature Space

Features space for our predictor has been built in a similar way as the one in PROFCon [16]. For each considered residue pair (i, j) the window of 9 residues, called context window, centered around residue i and j is considered. To describe the area between residue i and j additional window of length 11, centered in the middle of residue *i* and *j*, is used.

All residues in a window are represented by two types of features: Relative Solvent Accessibility (RSA), represented by real values, and Secondary Structures (SS). There is a number of definitions (and available programs) of secondary structures that are based on the different criteria: C_α distances, dihedral angles, specific patterns of hydrogen bonds. Here, we used the DSSP program with default parameters to define SS and RSA states for each amino acid residue. We used two methods to predict secondary structures, namely SABLE and Psipred [1]. Secondary structure of each residue has been coded by confidence factors obtained from predictors. This kind of coding, compared to binary coding used by most published methods, improves prediction accuracy by 1 to 3 percent points. Somewhat similar coding has been already used by PROFCon. However, PROFcon still utilizes binary encodings of secondary structure, with the reliability index for the winner class. RSA predictions used by us have been obtained from the SABLE server and were represented by real values.

In addition to the features that represent the sliding window, we used one additional feature that represents the distance between residue *i* and *j* along sequence calculated as follows:

$$d_{ij} = \frac{|i - j|}{L} \tag{3}$$

where L is the length of the protein. This feature plays important role in short contact prediction.

The contact class for our neural network was coded by value 1 and non-contact by 0.

VI. Training and Testing Protocol

We have examined two approaches to training sample selection. In the first one (called SAMPLING1), we extracted 10000 data items marked as contacts in the training data and different counts of the non-contact examples (1, 2, …, 10 times the count of contacts). In the second approach (called SAMPLING2) we took all the contact data from the training set and selected different counts of the non-contact examples, also as the multiples of the contact sample size. Here, the diversity of the models is a result of using different negative class examples. Different proportions of the classes have facilitated a robust analysis of the influence of negative class sample on the final results.

A. Neural Network

All Neural Networks (NNs) generated by us were built and trained using SNNS [21] simulator. For each of the contact range we built a separate predictor, so finally we have three predictors. We randomly selected 3% of the data from the training set. Selected data has been used at the end of building predictor phase where we were choosing networks for committee.

The best architecture of neural networks and all learning parameters were found using 10 fold cross-validation procedure. In this procedure, the dataset is randomly partitioned into 10 equal size subsets. Subsequently 10 iterations of training and testing are performed such that within each iteration a different subset of the data is held-out for testing and the remaining 9 for training. Additionally in each fold from the training set 10% randomly selected vectors are used for validation of the neural network. When the accuracy on the validation set dropped down, the training process of the neural network was stopped. The final prediction evaluation was calculated as the sum of accuracy on test set obtained in each of the fold. 10 fold cross-validation procedure was repeated 5 times for each of the tested architectures.

Final best architecture of the neural network consists of one hidden layer with 40 nodes and the output layer with 1 node. As the learning algorithm standard RProp [22] (Resilient propagation) was used.

In typical classification problem some threshold of activation value must be defined below or above which the tested vector is assigned to one or the other class. In contact map prediction we made ranking of all tested vectors, which belong to one protein, based on the value of activation function of output neuron. From the ordered vectors only k requested best vectors are selected as contacts. Activation value of the output node for contact vector depends on the number of non-contact instances that are "similar" to the contact one. The more non-contact instances the smaller excitation is expected, that is why adding more vectors to the non-contact class improves the ordering. But from the point of view of typical classification measures, where all, not only limited number of testing vectors are considered, this methodology will lower accuracy, because some activation threshold for classification must be defined. The distribution of the activation values strongly depends on the type of the protein. For some of the proteins activation for all tested vectors might be very low and in such cases if the threshold is defined all tested vectors (proteins) are treated as non-contacts, whereas looking at ranking we are still able to select the required number of contacts having a reasonable accuracy at the same time.

To improve final prediction, for each of the contact range, instead of single neural network we built a committee of 15 networks. To select the predictors to be included in the committee, we chose networks obtained in

cross-validation procedure with the lowest correlation between each other.

B. Decision Trees

Decision Trees (DTs) belong to the most frequently used models of machine learning. Unfortunately, single DTs may be insufficiently accurate and black-box systems seem more attractive because of more accurate predictions. Therefore, many efforts have been undertaken to learn a number of DTs that can act together providing high accuracy. Such committees of decision trees have been successfully applied to various types of data. One of the successful techniques is called bagging DTs [23]. It creates a bunch of DTs by performing a number of independent processes consisting of drawing a data sample on the basis of the training data and learning a DT. The standard way to determine the training samples is to draw independently n data items (with possible repetitions) from the training data containing n items. As a result, the expected probability of that a given item belongs to a sample (called bootstrap sample) is close to 0.632 [24]. In this way, each sample provides different information to the learner and the final trees can be diverse (most DT induction methods are deterministic, so there is no diversity without diverse training data samples). The diversity of the set of DTs is the basis of the generalization capabilities of the bagged DT committees. To maximize the diversity, the DT models are not pruned (pruning processes cut some DT splits to avoid overfitting the training data, but this is not desirable when bagging DTs, as the generalization abilities come from averaging decisions of diverse models). As a side effect, resigning from pruning speeds the learning process up.

In the case of searching for contacts in proteins, where the goal is to correctly predict a number of contacts, not to obtain high overall mean of classification accuracy, it is important to predict properly just a number of top items in the ranking of the most probable contacts. To get correct rankings, it is important to accurately predict the probabilities of a contact or another measure that can put the most probable contacts at the top of the ranking. Probability estimates based on single DTs are usually based on proportions of class representatives within the tree node of interest. Such estimates are not very accurate and in the case of bagging, where no DT pruning is applied, they are binary unless some corrections like Laplace correction or m-estimates are taken into account. Nevertheless, in sufficiently large committees, even the binary outputs average to quite successful estimates.

VII. Results

In application of the DT bagging committees and NN committees to the task of contact prediction, we have collected large amounts of results for our two types of sampling and various parameter values. Because results for SAMPLING1 in case of NNs are much lower then for SAMPLING2, for DT there is almost no difference, we focused here on SAMPLING2. The main aspect of our analysis is depicted in Fig. 1. The figure illustrates prediction accuracies of DT committees consisting of 101 DT models and NN committees consisting of 15 networks built on complete sets of contact items and various sizes of the non-contact samples. The three groups of lines (columns of the figure) present results for different contact ranges. Each group contains four lines corresponding to the four levels of contact prediction (L/2, L/5, L/10 and best 5). Each line presents the trend of the Acc measure with respect to the balance ratio between contact and non-contact samples within the training data set. In case of NNs number of non-contact samples is higher than in DT. The main reason is that variety of trees comes from sampling, so it is not possible to use all the data, whereas in NNs comes from learning process. The maximal number of non-contact instances in DTs is 10 times higher than contact vectors, whereas in NNs we used all available data for each of the contact type, so for each of the type we have different.

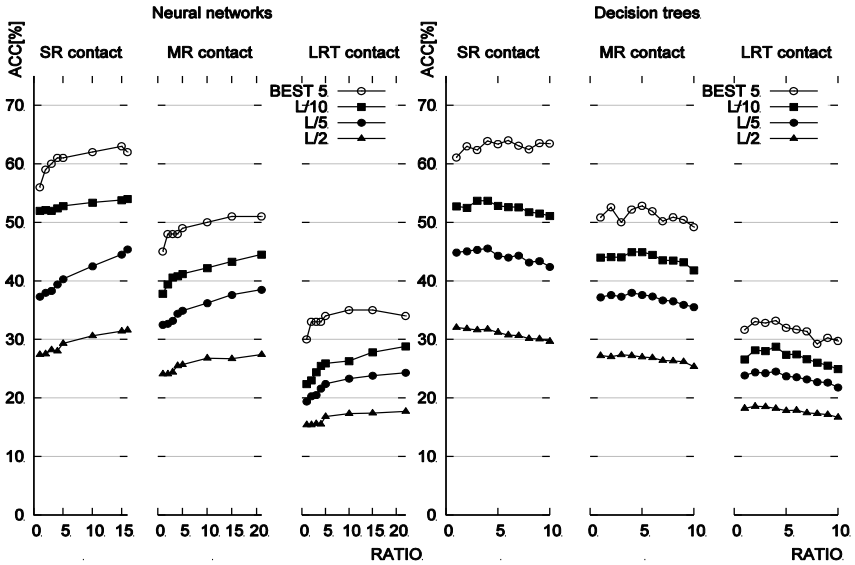

Figure 1. Trends of changes in prediction accuracy (Acc measure) in relation to the balance factor of the training data set. Left panel represents results for NNs based predictor, right for DTs.

For both completely different predictors and almost all contact types we observe improvements when unbalanced training set is used. This is particularly pronounced when Best 5 limitation of contacts is considered. The more contacts is analyzed, the closer we are to binary classification task (where accuracy measure takes into account prediction of all tested vectors), the lower improvements are observed. This is very consistent to the fact that for binary classification task it is better to use balanced training set [20]. For both predictors we observe also important differences, for NNs there is almost constant improvement of *Acc* measure with growing number of non-contact instances, whereas in DTs usually after ratio 5 *Acc* decreases. This varieties may come for completely different nature of the presented methods, decision trees are built without any regularization (trees where not cut) whereas in neural networks number of adaptive parameters is limited by the architecture. Nevertheless the best *Acc* measures for both types predictors in all types of contacts are very similar, moreover predicted contacts are almost identical, so combining both predictors does lead to very small further improvements.

A. Results on CASP 11

Comparing different methods for contact map predictions is challenging. Results that are published are already obsolete, in the meantime newer versions of sequence databases became available, which may lead to better SS prediction and consequently better contact map prediction accuracy. Moreover usually different training sets and testing sets are used, so some very similar proteins in the testing set might be already incorporated in training set of the other method. So, the best way of comparison with the other methods is to use an independent evaluation, such as the one employed in CASP competition.

Our PLCT method took part in the newest CASP competition CASP11. According to the automatic evaluation available at http://www.predictioncenter.org/casp11/rr_summary_results.cgi PLCT for the medium (see Table I) and medium+long contact ranges (*L*/10 first contacts considered) took 3rd place among 21 sequence based methods that took part in CASP11.

TABLE I. RANKING OF CONTACT MAP PREDICTION METHODS OBTAINED FROM HTTP://PREDICTIONCENTER.ORG/CASP11/RR_SUMMARY_RESULTS.CGI FOR PARAMETERS: CONTACT RANGE – MEDIUM, LIST SIZE L/10. THE ORDER OF PRESENTED METHODS IS MADE ACCORDING TO THE LAST COLUMN (Z-SCOREACC+Z-SCOREXD)

		AVG Acc	$Z\text{-score}_{Acc}$	AVG Xd	$Z\text{-score}_{Xd}$	$Z\text{-score}_{Acc}+Z\text{score}_{Xd}$
1.	RBO_Aleph [25]	50.27	0.95	17.84	0.86	1.81
2.	CONSIP2 [10]	51.55	0.87	18.67	0.82	1.69
3.	PLCT	47.05	0.69	17.35	0.71	1.4
4.	RaptorX-Contact [26]	46.49	0.53	17.84	0.75	1.38
5.	UCI-IGB-CMpro [18]	45.26	0.67	17.13	0.69	1.36
6.	FALCON_Contact [27]	44.83	0.61	16.48	0.63	1.24
7.	MULTICOM-NOVEL [28]	44.98	0.58	16.8	0.63	1.2
8.	MULTICOM-CLUSTER [28]	45.78	0.62	17.04	0.55	1.16

As we mentioned the best results for out PLCT method were obtained for medium and medium+long contact ranges and L/10 best first contacts. Also for other categories our method behaves very well, for: medium and medium+long and Top5 best contacts PLCT took 4th position, for short contact ranges 4th and 5th position. The worst results were obtained for long contacts (9th position).

VIII. CONCLUSION

Balancing the training set for machine learning algorithms is a well known technique in machine learning field. In this paper, we are claiming the statement that for contact map prediction it is better to use unbalanced training set. We are showing that this statement is true only because the accuracy measure used in contact map prediction is not the one used in typical classification problems. This measure is taking into account prediction of limited number of contact class instances. In the paper we presented that the improvement in accuracy is observed for different kinds of machine learning algorithms: neural networks and decision trees. Our methods have been compared with methods, that took part in CASP 11 competition. Although these methods applied many different features in their feature space including features used by us, our results in comparison with other sequence based methods belong to the best ones.

ACKNOWLEDGMENT

We would like to thank Jarek Meller for his valuable comments and suggestions.

REFERENCES

[1] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, pp. 195-202, 1999.
[2] R. Adamczak, A. Porollo, and J. Meller, "Combining prediction of secondary structure and solvent accessibility in proteins," *Proteins Struct. Funct. Genet.*, vol. 59, pp. 467-475, 2005.
[3] R. Adamczak, A. Porollo, and J. Meller, "Accurate prediction of solvent accessibility using neural networks-based regression," *Proteins Struct. Funct. Genet.*, vol. 56, pp. 753-767, 2004.
[4] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, "Prediction of coordination number and relative solvent accessibility in proteins," *Proteins*, vol. 47, pp. 142-153, 2002.

[5] J. Cheng and P. Baldi, "Improved residue contact prediction using support vector machines and a large feature set," *BMC Bioinformatics*, vol. 8, no. 2, 2007.
[6] A. Vullo, I. Walsh, and G. Pollastri, "A two-stage approach for improved prediction of residue contact maps," *BMC Bioinformatics*, vol. 7, no. 7, 2006.
[7] A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick, "Ab initio folding of proteins using restraints derived from evolutionary information," *Proteins*, pp. 177-185, 1999.
[8] H. Ashkenazy and Y. Kliger, "Reducing phylogenetic bias in correlated mutation analysis," *Protein Eng. Des. Sel.*, vol. 23, pp. 321-326, 2010.
[9] U. Göbel, C. Sander, R. Schneider, and A. Valencia, "Correlated mutations and residue contacts in proteins," *Proteins*, vol. 18, pp. 309-317, 1994.
[10] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, "PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments," *Bioinformatics*, vol. 28, pp. 184-190, 2012.
[11] S. F. Altschul, *et al.*, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389-3402, 1997.
[12] G. Pollastri and P. Baldi, "Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners," *Bioinformatics*, vol. 18, pp. S62-S70, 2002.
[13] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio, "Prediction of contact maps with neural networks and correlated mutations," *Protein Eng.*, vol. 14, pp. 835-843, 2001.
[14] I. N. Shindyalov, N. A. Kolchanov, and C. Sander, "Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?" *Protein Eng.*, vol. 7, pp. 349-358, 1994.
[15] N. Hamilton, K. Burrage, M. A. Ragan, and T. Huber, "Protein contact prediction using patterns of correlation," *Proteins*, vol. 56, 679-684, 2004.
[16] M. Punta and B. Rost, "PROFcon: Novel prediction of long-range contacts," *Bioinformatics*, vol. 21, pp. 2960-2968, 2005.
[17] W. Ding, J. Xie, D. Dai, H. Zhang, H. Xie, and W. Zhang, "CNNcon: Improved protein contact maps prediction using cascaded neural networks," *PLoS One*, vol. 8, p. e61533, 2013.
[18] P. D. Lena, K. Nagata, and P. Baldi, "Deep architectures for protein contact map prediction," *Bioinformatics*, vol. 28, pp. 2449-2457, 2012.
[19] G. Weiss and F. Provost, "The effect of class distribution on classifier learning: An empirical study," Technical report ML-TR-44, Department of Computer Science, Rutgers University, 2001.
[20] Q. Wei and R. L. Dunbrack Jr, "The role of balanced training and testing data sets for binary classifiers in bioinformatics," *PLoS One*, vol. 8, p. e67863, 2013.
[21] A. Zell, *et al.* The SNNS users manual version 4.1. [Online]. Available: http://www-ra.informatik.uni-tuebingen.de/SNNS
[22] M. Riedmiller and H. Braun, "RPROP- A fast adaptive learning algorithm," in *Proc. International Symposium on Computer and Information Sciences*, 1992.
[23] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123-140, 1996.
[24] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *J. Am. Stat. Assoc.*, vol. 78, pp. 316-331, 1983.
[25] M. Schneider and O. Brock, "Combining physicochemical and evolutionary information for protein contact prediction," *PLoS One*, vol. 9, no. 10, Oct. 2014.
[26] Z. Wang and J. Xu, "Predicting protein contact map using evolutionary and physical constraints by integer programming," *Bioinformatics*, vol. 29, no. 13, pp. i266–i273, Jul. 2013.
[27] S. C. Li, D. Bu, J. Xu, and M. Li, "Fragment-HMM: A new approach to protein structure prediction," *Protein Sci.*, vol. 17, no. 11, pp. 1925-1934, 2008.
[28] Z. Wang, J. Eickholt, and J. Cheng, "MULTICOM: A multi-level combination approach to protein structure prediction and its assessments in CASP8," *Bioinformatics*, vol. 26, no. 7, pp. 882-888, Apr. 2010.

Grzegorz Markowski is a PhD student working under the supervision of dr hab. Rafał Adamczak. He received his Master's degree in physics from the Nicolaus Copernicus University in Torun and also degree in informatics from the Adam Mickiewicz University in Poznan. His research interests include applying machine learning tools in bioinformatics. In addition, to pursuing his PhD, Grzegorz is also working in Information&Communication Technology Centre.

Krzysztof Grąbczewski received his PhD degree in 2003 from Systems Research Institute, Polish Academy of Sciences and habilitation (D.Sc.) in 2014 from Institute of Computer Science, Polish Academy of Sciences. He is working as assistant professor at Department of Informatics, Nicolaus Copernicus University, Toruń, Poland. His scientific interests include broad spectrum of computational intelligence algorithms and applications, especially all the aspects of advanced meta-learning. He has published a book "Meta-learning in decision tree induction" and over 50 reviewed papers including book chapters, journal articles and peer-reviewed conference papers. His data mining skills have been confirmed by the 3rd place in NIPS 2003 Feature Extraction Challenge, and the 1st place in ICAISC 2006 Handwritten Digit Recognition Contest at The Eighth International Conference on Artificial Intelligence and Soft Computing.

Rafal Adamczak received his PhD degree in 2001. In the same year he started post-doc position at Children's Hospital Medical Center in Cincinnati in USA, where he was working in bioinformatics field. Currently he is working as assistant professor at Department of Informatics, Nicolaus Copernicus University, Toruń, Poland. His scientific interests include artificial neural networks, machine learning and bioinformatics. He is author and co-author of over 50 reviewed papers.