RIGA TECHNICAL UNIVERSITY

Information Technology and Management Science

ISSN 2255-9094 (online)
ISSN 2255-9086 (print)
December 2018, vol. 21, pp. 98–103
doi: 10.7250/itms-2018-0016
https://itms-journals.rtu.lv

Integrated Network Approach to Protein Function Prediction

Natalia Novoselova[1], Igar Tom[2]
[1,2]*United Institute of Informatics Problems, Minsk, Belarus*

***Abstract* – One of the main problems in functional genomics is the prediction of the unknown gene/protein functions. With the rapid increase of high-throughput technologies, the vast amount of biological data describing different aspects of cellular functioning became available and made it possible to use them as the additional information sources for function prediction and to improve their accuracy.**

In our research, we have described an approach to protein function prediction on the basis of integration of several biological datasets. Initially, each dataset is presented in the form of a graph (or network), where the nodes represent genes or their products and the edges represent physical, functional or chemical relationships between nodes. The integration process makes it possible to estimate the network importance for the prediction of a particular function taking into account the imbalance between the functional annotations, notably the disproportion between positively and negatively annotated proteins. The protein function prediction consists in applying the label propagation algorithm to the integrated biological network in order to annotate the unknown proteins or determine the new function to already known proteins. The comparative analysis of the prediction efficiency with several integration schemes shows the positive effect in terms of several performance measures.

***Keywords* – Computational biology, data mining, functional association network, binary classification.**

I. Introduction

In the past decade due to biotechnological advances, the various types of molecular data at a genome-wide scale have been produced, but despite a lot of fully sequenced genomes the function of large numbers of proteins still remains unknown. The classical way is to find homologies between a protein and other proteins in protein databases using programs such as BLAST, PSI-BLAST [1] and then predict functions based on sequence homologies. However, roughly 20 %–40 % of proteins in newly sequenced genomes do not have statistically significant sequence similarity to functionally annotated proteins. In addition, sequence similarity does not necessary imply functional equivalence and thus Blast-based annotations can be erroneous [2]. Therefore, the additional sources of information from a variety of high-throughput experiments have been extensively used for the study of protein functions. Among them are gene expression patterns, phylogenetic profiles, protein fusions and protein-protein interactions (PPI) [3]. Clustering analysis of gene expression data has been used to predict functions of unannotated proteins based on the idea that co-expressed genes are more likely to have similar functions [4]. The great popularity of using the PPI data can be explained by providing information on the biological context of protein functions. As a rule, proteins are not operating in isolation but interact with one another in order to provide cell functioning, taking part in some metabolic pathway or biological process. Therefore, it is possible to deduce functions of a protein through the functions of its interaction partners [3]. PPI functional linkage networks (graphs), where nodes represent proteins and edges represent the detected interactions, are extensively used for deriving the functions of unannotated proteins using different probabilistic and graph algorithms. Probabilistic analysis of graph neighbourhoods in a protein-protein interaction network is described in [2]. In [5], [6] the network propagation algorithm for protein function prediction is proposed. The algorithm allows obtaining functional evidence from non-neighbouring nodes in functional-linkage graphs.

However, each information source can possess the inherent noise, e.g., protein interaction databases such as MIPS [7], BioGRID [8] and STRING [9], which have assembled a large collection of putative functional links between proteins by including information provided by diverse computational and experimental screens, can produce large numbers of both false positive and false negative interactions. Additionally, each type of data describes only one part of cellular activity; therefore, it was proposed to combine the heterogeneous data sources in order to increase the coverage and the accuracy of protein function prediction [4], [10]. The need to integrate several sources of information has increased the task complexity and more computationally efficient approaches must be developed. All the existing up-to-date methods can be roughly divided into two groups: kernel methods and functional linkage network methods [10]. In the first group, for each data source the similarities between proteins are determined using the kernel similarity matrix, and different kernel integration methods are applied in order to combine heterogeneous data sources, e.g., in [11] on the basis of the integrated similarity kernel a support vector machine (SVM) was used to predict protein functions. In the second group, each data source is presented as the functional linkage graph and the network integration algorithm is applied. After that probabilistic graphical models or network-based classification algorithms are used to infer the annotations for unknown proteins [3], [4], [12], [13]. There are also some approaches, where individual classifiers are trained on each network and then the ensemble learning technique is used to combine classifiers [10], [14].

As there is not a ready solution to solve the problem of integrating data sources in a more optimal way in order to increase the prediction accuracy and to deal with the unbalanced labels in GO functional categories we have

©2018 Natalia Novoselova, Igar Tom.
This is an open access article licensed under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0).

described and analysed the performance capabilities of a two-step approach to protein function prediction, which is promising in accounting for the label imbalance and computationally efficient due to relying on the sparseness of functional association networks. It has the advantage in comparison to simple averaging of individual networks and correlation-based network weighting method, described below. The first step consists in constructing the integrated functional network from heterogeneous data sources. It is based on integration of single functional association networks using a form of kernel-target alignment [15] between the composite network and a "target" network constructed from the function label vector. The alignment task is formulated as the task of linear regression with constraints, which allows determining the weights for each data source and simultaneously excluding non-informative ones [16]. The second step consists in assigning the functions to unannotated proteins from a single composite network using the label propagation algorithm [16], [17].

II. Methods

In our study, we consider the task of protein function prediction as a binary classification task, where the labels correspond to the presence or absence of the specific function. Each data source is initially presented in the form of functional association network, which encodes information of shared protein functions from high-throughput proteomic (or genomic) data sources (i.e., protein-protein interactions (PPI), protein domains). In this representation, a node in the network corresponds to a protein, and the weights of the edges of connected nodes correspond to their similarity computed by a specific similarity metric for a given data source. Then these individual networks are combined, through a weighted sum, into a composite network, where the weights are optimised using labels, each label corresponding to a distinct protein function. The weight of the network reflects its usefulness in predicting a given function of interest. Next, the network-based classification algorithm is applied to the composite network in order to compute the association score of a specific function label for the unannotated proteins. We have applied the label propagation algorithm [17] in order to derive the protein functions on the basis of integrated functional association network.

A. Data Sources and Pre-processing

We have made several experiments on the MouseFunc I benchmark data [18] in order to evaluate the two-step approach to protein function prediction. Ten association networks were constructed from ten data sources, including gene expression, protein annotations from Pfam [19] and InterPro [20], Protein-Protein Interactions, Phylogenetic Profiles, Disease Profiles from OMIM [21]. The data sources cover 21603 mouse genes (Table I).

We have constructed networks from each profile-based high-throughput data source using Pearson correlation coefficient (PCC).

TABLE I
Data Sources Used in the Experiments

PubMed ID	Publication (Description)	Number of networks
1558831	Zhang et al. Journal of Biology 2004	1
15075390	Su et al. PNAS 2004	1
SAGE Lib.	Tag counts	1
OPHID	Protein interactions	2
Pfam	Domain composition	1
InterPro	Domain composition	1
bioMART	Phylogenetic profile	1
Inparanoid	Phylogenetic profile	1
OMIM	Disease genes	1

For network-based data (e.g., protein interaction), we have used both binary matrix of protein interactions and matrix of distances between proteins. In order to disperse the resulting networks and increase the computation efficiency without degrading the accuracy, we have set the threshold on the number of links for each gene to K=50. Each functional network W_i has been normalised using the expression $\overline{W}_i = D_i^{-1/2} W_i D_i^{-1/2}$, where D_i is the diagonal row sum matrix of W_i. The same normalization has been applied to the integrated functional network.

To evaluate the results of gene function prediction, we have used the GO biological process (BP) function categories [22]. We have evaluated several integration schemes by using the resulting integrated networks as input to the label propagation algorithm and calculated an average area under the receiver operating characteristic (ROC) curve (AUC of ROC) and area under precision-recall curve (AUP) over all BP GO categories with 3–300 annotations using 3-fold cross-validation (CV). We have concentrated on BP categories because they comprise the majority of functions in the GO hierarchy. Four evaluation groups have been created to estimate the results, which correspond to four levels of annotations ([3 to 10], [11 to 30], [31 to 100], and [101 to 300]). Category [3–10] includes 952 GO terms, category [11–30] includes 435 GO terms, category [31–100] includes 239 GO terms and category [101–300] includes 100 GO terms.

B. Integration of Data Sources

In order to integrate the individual functional networks, we have applied the kernel-target alignment, which can be formulated in the form of linear regression task [23].

Given $\{W_m\}_{i=m}^{M}$ functional association networks we construct the composite network $W_{ij} = \sum_{m=1}^{M} \alpha_m W_{ij}^m$, where $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_M]$ are the weights of each individual network, which determine the accuracy of protein function prediction and can be estimated by the kernel-target alignment in the form

$$\alpha = \arg\min_{\alpha} tr\left(\left(K - W\right)^{T}\left(K - W\right)\right)$$
$$s.t. \quad W = \sum_{m=1}^{M} \alpha_m W^m, \quad \alpha_m \geq 0 \tag{1}$$

where $K \in R^{n \times n}$ is the target network of functional label, computed as follows:

$$K\left(i,j\right) = \begin{cases} \left(\frac{n^-}{n}\right)^2, & if\ y_i = y_j = 1 \\ \frac{n^+ n^-}{n^2}, & if\ y_i \cdot y_j = -1 \,\&\, y_i + y_j = 0 \\ 0, & otherwise \end{cases} \tag{2}$$

where n^+, n^- are the numbers of positives and negatives in label vector $\vec{y} \in \{+1,-1\}^n$, where positive and negative genes are labelled as +1, −1, respectively. Pairs of negatively labelled genes have no influence in determining the weights. In order to exclude the negative–negative pairs of proteins from consideration, the entries in K and each network W_m that corresponds to negative pairs of genes are removed.

By minimising (1), larger weights are assigned to the networks, which consider highly similar proteins that share function of interest, and smaller weights – to networks, which consider highly similar proteins that do not share the function. Finally, networks, coherent with functional labels, get higher weights. By using the equality $trace(WK) = vec(W)^T vect(K)$, where $vec(K)$ is the vectorization operator that stacks the columns of K on top of each other we can write (1) as a non-negative unregularized linear regression problem:

$$\alpha^* = \arg\min_{\alpha}\left(\left(V_K - V_W\alpha\right)^T\left(V_K - V_W\alpha\right)\right)$$
$$s.t. \quad \alpha_m \geq 0, 1 \leq m \leq M \tag{3}$$

where $V_W = \left[vec(W_1),\ldots,vec(W_M)\right] \in R^{(n\times n)\times M}$, $V_K = vec(K)$.

We have also used the constraints in linear regression in order to increase the robustness to the inclusion of irrelevant and redundant networks. It can help dealing with the different level of importance of each individual data source for the prediction of a particular functional class. In this case, to obtain the weight vector α, we solve the following ridge regression problem:

$$\alpha^* = \arg\min_{\alpha}\left(V_K - V_W\alpha\right)^T\left(V_K - V_W\alpha\right) + J\left(\alpha\right)$$
$$s.t. \quad \alpha_m \geq 0, 1 \leq m \leq M \tag{4}$$

where $J \geq 0$ is the regularization function.

For ridge regression with the prior, the regularization function is as follows:

$$J\left(\alpha\right) = \sum_{m=1}^{M}\left(\alpha_m - \nu_m\right)^2 s_m, \tag{5}$$

where $\vec{\nu}$ is the prior weight vector and s_m is the strength of the regularization on α_m. For ridge with uniform prior, we set $\nu_m = 1$. When all the $s_m, 1 \leq m \leq M$ are set to zero, then cost function (4) is unregularized and solving for α becomes equivalent to unregularized linear regression.

Solving equations (3–4) requires at most M iterations, where each iteration involves solving a system of linear equations with M variables.

C. Network-Based Prediction Algorithm

The first approach based on the "guilt by association" principle to predict the protein function on the basis of functional association network annotates the unknown protein with the functions of its neighbours, which can lead to errors. In our experiment, we have used the network-propagation algorithm, which allows using a global topology of the entire interaction network instead of the local neighbourhood and increasing the reliability of prediction.

Let $W = \sum_{m=1}^{M} \alpha_m W_m, 1 \leq m \leq M$ be a weight matrix of the composite functional network. Non-zero elements of W correspond to the strength of association between the connected proteins; association is absent when $W_{ij} = 0$. Weights α_m represent the relevance of the m-th network for the prediction task. Among n nodes in W we have l proteins, labelled with specific function and u unlabelled proteins. The labels are used to specify the label vector $y \in \{+1,-1,k\}$ for positive, negative and unlabelled nodes. The following expression is used to specify k:

$$k = \frac{n^+ - n^-}{n}, \tag{6}$$

where n^+, n^- are positive and negative labels. The initial association values k for unknown genes in (6) help account for label unbalance, where as a rule only a small number of genes is annotated with gene function of interest.

Label propagation algorithm is applied to the composite network W to predict functions of the unknown proteins. Using the algorithm, the scores f (discriminative values) for each node in the network are computed using the following optimization function:

$$f = \arg\min_{f} \sum_{i}\left(f_i - y_i\right)^2 + \sum_{i}\sum_{j} w_{ij}\left(f_i - f_j\right)^2, \tag{7}$$

which consists of two terms, where the first term penalizes the differences between the discriminant values of nodes and their initial labels and the second term penalizes the differences between the discriminant values of neighbouring nodes in the network. In such a way, the labelling information propagates

through the network allowing one to determine the labels of unknown proteins not directly connected to the positive nodes. The conjugate gradient (CG) method is used to solve the system of linear equations, which presents the solution to the optimization task in (7). Due to sparseness of the composite network W a conjugate gradient method is very efficient in solving (7); potentially the runtime of CG depends only on the number of connections in W and it is possible to get very close to the exact solution with only less than few dozens of iterations.

III. Results

We have made the comparative analysis of different integration schemes, including the unregularized linear regression, ridge regression with uniform prior weight vector and a network combination with uniform weights (equal weighting), where the network weights are all set to $1/M$; M is the number of networks.

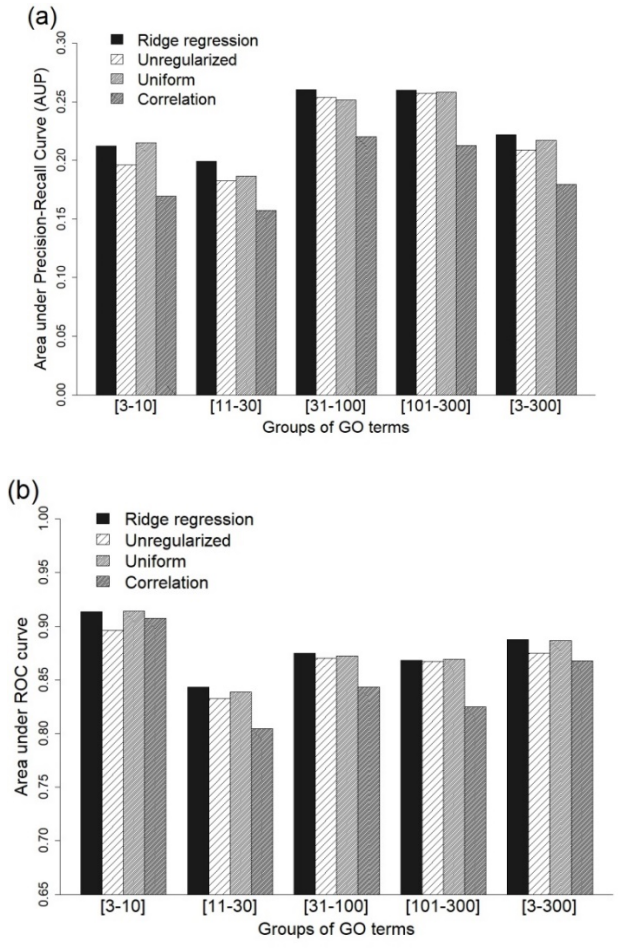

Fig. 1. Efficiency measures of protein function prediction using different network integration schemes (ridge regression with uniform priors, unregularized regression, equal weighting and correlation weighting): (a) area under precision-recall curve; (b) average area under the ROC curve. The bars indicate average performance in evaluation categories with a different number of positive annotations.

We have also compared the results with the correlation-based network method. In this method, each network weight corresponds to the kernel-target alignment score for this network:

$$\alpha_m = \frac{\vec{y}^T W_m \vec{y}}{\langle W_m, W_m \rangle} = \frac{\sum_{ij} w_{ij}^m y_i y_j}{\sum_{ij} \left(w_{ij}^2 \right)} . \quad (8)$$

Figure 1 depicts the performance of each analysed integration scheme in five categories: predicting gene GO functions, which have [3–10], [11–30], [31–100], [101–300] positive annotations and the whole set of [3–300] positive annotations. Figure 1 shows that ridge regression significantly outperforms unregularized regression and equal weighting in terms of AUP ($P=1.60\times10^{-4}$ and $P=7.70\times10^{-4}$, Wilcoxon paired signed rank test) taken all the gene GO functions together. The same tendency is for the AUC values ($P=2.50\times10^{-10}$ and $P=1.80\times10^{-5}$, Wilcoxon paired signed rank test). The weights assigned by the correlation method give the worst prediction using both AUC and AUP measures. It can be explained by inability of the method to register redundancy between the networks.

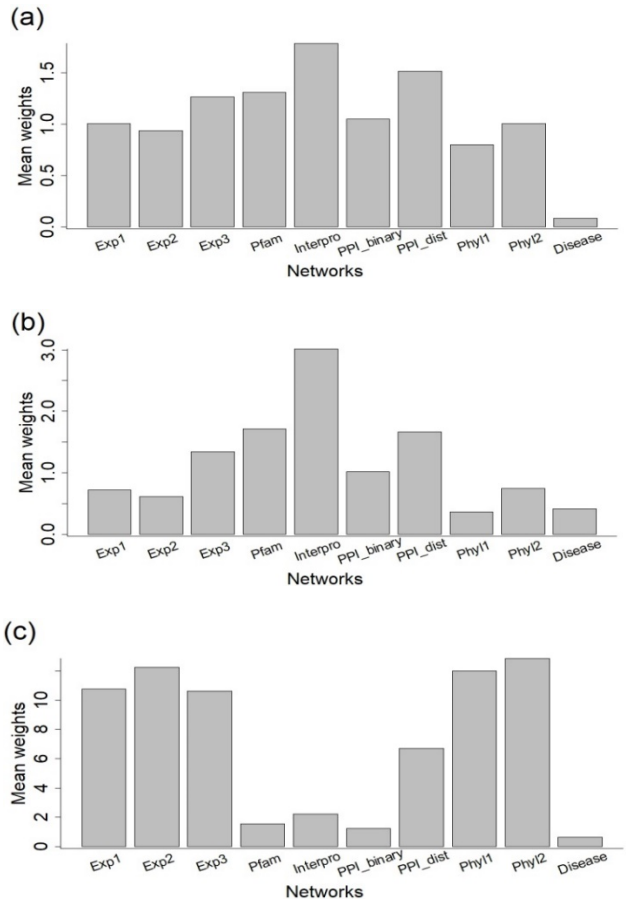

Fig. 2. Average weights assigned to each network while predicting 1726 gene functions with different integration schemes: (a) ridge regression with uniform priors, (b) unregularized regression, (c) correlation weighting method. Abbreviations: Exp1, Exp2, Exp3 – gene expression data; Pfam and Interpro – protein domains; PPI_binary and PPI_dist – protein-protein interaction data; Phyl1, Phyl2 – phylogenetic data; Disease – disease profiles.

The performance of the compared schemes in terms of AUP improves with the increase of the number of positive annotations, where the ridge regression is always on the top position in all the evaluation categories. The AUC values are higher for the category with [3–10] annotations, which is the result of a low number of positive annotations and the possible absence of them in the testing set. In all the other categories, the AUC values increase with the number of positive annotations.

We have also compared the network weights that were assigned by different integration schemes for individual networks (Fig. 2). Unregularized linear regression is the most selective and assigns non-zero weights to only several functional networks for each evaluation category. Composite network, constructed using ridge regression with uniform priors, includes more networks with positive weights. Therefore, efficiency of ridge regression and equal weighting scheme differ to a lesser extent.

The correlation weights have the opposite tendency to unregularized regression. The possible redundancy of gene expression and phylogenetic profiles is not taken into consideration. The average weights assigned to these data sources are much higher than those of linear regression.

In all the schemes, a high proportion of the weights is assigned to the networks derived from gene expression and protein-protein interaction data sources.

IV. Conclusion

In the paper, we have analysed the performance capabilities of the two-step approach to protein function prediction, which is promising in accounting for the label imbalance and computationally efficient due to relying on the sparseness of functional association networks. The experiments were conducted on the MouseFunc I benchmark data and GO evaluation categories of protein annotations. Two different performance measures were applied, notably AUC and AUP. AUP measure is more appropriate to the estimation of the results of binary classification tasks with significant label unbalance, i.e., a small number of positive in comparison to negative cases.

In the first step, the integrated functional network is constructed from heterogeneous data sources. The weights for individual networks correspond to the solution of linear regression task with constraints, which is formulated on the basis of kernel-target alignment method and takes into account the known protein annotations. The second step makes the prediction of the protein functions on the basis of an integrated network using the label propagation algorithm.

Several experiments with different integration schemes have shown that the scheme on the basis of ridge regression with uniform priors has preference in comparison to the widely assumed equal weighting. The correlation weighting method was the worst in all the evaluation categories. It can be partly explained by the inability of this scheme to filter information redundancy. The application of unregularized regression to the integration of individual functional networks was restricted to the selection of only few networks with positive weights, which could explain the loss in performance in comparison to ridge regression.

The experiments have shown that the selection of integration scheme has a great influence on the accuracy of protein function prediction, and more extensive experiments with different priors for ridge regression must be conducted. The possible future research direction is the development of approach, which takes into account the hierarchical organisation of GO ontology and makes simultaneous predictions for the groups of functional terms. This direction can lead to the improvement of accuracy of protein function prediction from multiple networks.

References

[1] S. F. Altschul, et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res*., vol. 25, no. 17, pp. 3389–3402, 1997. https://doi.org/10.1093/nar/25.17.3389

[2] S. Letovsky and S. Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach," *Bioinformatics*, vol. 19 (Suppl. 1), pp. i197–i204, 2003. https://doi.org/10.1093/bioinformatics/btg1026

[3] Y. Kourmpetis, A. van Dijk, M. Bink, R. van Ham, and C. ter Braak, "Bayesian Markov Random Field Analysis for Protein Function Prediction Based on Network Data," *PLoS ONE*, vol. 5, no. 2, p. e9293, 2010. https://doi.org/10.1371/journal.pone.0009293

[4] N. Nariai, E. D. Kolaczyk, and S. Kasif, "Probabilistic Protein Function Prediction from Heterogeneous Genome-Wide Data," *PLoS ONE*, vol. 2, no. 3, p. e337, 2007. https://doi.org/10.1371/journal.pone.0000337

[5] U. Karaoz et al., "Whole genome annotation by using evidence integration in functional-linkage networks," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 9, pp. 2888–2893, 2004. https://doi.org/10.1073/pnas.0307326101

[6] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," *Journal of Computational Biology*, vol. 10, no. 6, pp. 947–960, 2004. https://doi.org/10.1089/106652703322756168

[7] P. Pagel, S. Kovac, M. Oesterheld et al., "The MIPS mammalian protein-protein interaction database," *Bioinformatics*, vol. 21, no. 6, pp. 832–834, 2005. https://doi.org/10.1093/bioinformatics/bti115

[8] C. Stark, B. J. Breitkreutz, A. Chatr-Aryamontri et al., "The BioGRID interaction database: 2011 update," *Nucleic Acids Res*., vol. 39, no. 1, pp. D698– D704, 2011. https://doi.org/10.1093/nar/gkq1116

[9] D. Szklarczyk, A. Franceschini, M. Kuhn M et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res*., vol. 39, pp. D561–568, 2011. https://doi.org/10.1093/nar/gkq973

[10] M. Re and G. Valentini, "Ensemble Based Data Fusion for Gene Function Prediction," Lecture Notes in Computer Science, pp. 448–457, 2009. https://doi.org/10.1007/978-3-642-02326-2_45

[11] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast," *Proc. of the Pacific Symposium on Biocomputing*, vol. 9, pp. 300–311, 2004. https://doi.org/10.1142/9789812704856_0029

[12] M. Frasca, A. Bertoni, and G. Valentini, "UNIPred: Unbalance-aware Network Integration and Prediction of protein functions," *Journal of Computational Biology*, vol. 22, no. 12, pp. 1057–1074, 2015. https://doi.org/10.1089/cmb.2014.0110

[13] W. Noble and A. Ben-Hur, "Integrating information for protein function prediction," in *Bioinformatics-From Genomes to Therapies*, Vol. 3, T. Lenguaer, Ed. WILEY-VCH Verlag GmbH & Co., 2007, pp. 1297–1314. https://doi.org/10.1002/9783527619368.ch35

[14] N. Cesa-Bianchi, M. Re, and G. Valentini, "Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference," *Machine Learning*, vol. 88, no. 1–2, pp. 209–241, 2012. https://doi.org/10.1007/s10994-011-5271-6

[15] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Proc. of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, December 03–08, 2001, Vancouver, British Columbia, Canada. [Online]. Available: https://papers.nips.cc/paper/1946-on-kernel-target-alignment. [Accessed: Oct. 5, 2018].

[16] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris, ”Genemania: A real-time multiple association network integration algorithm for predicting gene function,” *Genome Biology*, vol. 9, Suppl. 1:S4, 2008. https://doi.org/10.1186/gb-2008-9-s1-s4

[17] X. Zhu, Z. Ghahramani, and J. Lafferty, “Semi-supervised learning using Gaussian fields and harmonic functions,” *Proc. of the Twentieth International Conference on Machine Learning*, pp. 912–919, August 21– 24, 2003, Washington, DC, USA.

[18] L. Pena-Castillo, T. Murat, C. L. Myers, H. Lee, T. Joshi, C. Zhang, and et al., “A critical assessment of Mus musculus gene function prediction using integrated genomic evidence,” *Genome Biology*, vol. 9, Suppl. 1:S2, 2008. https://doi.org/10.1186/gb-2008-9-s1-s2

[19] Pfam 32.0 (September 2018, 17929 entries). [Online]. Available: https://pfam.xfam.org/. [Accessed: Oct. 5, 2018].

[20] InterPro: protein sequence analysis & classification. [Online]. Available: https://www.ebi.ac.uk/interpro/. [Accessed: Oct. 5, 2018].

[21] Online Mendelian Inheritance in Man (OMIM). [Online]. Available: https://www.ncbi.nlm.nih.gov/omim/. [Accessed: Oct. 5, 2018].

[22] Gene Ontology Consortium. [Online]. Available: http://geneontology.org/. [Accessed: Oct. 5, 2018].

[23] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, “On kernel target alignment,” in *Studies in Fuzziness and Soft Computing*, vol. 194, Innovations in Machine Learning, Springer-Verlag, 2006, pp. 205–256. https://doi.org/10.1007/3-540-33486-6_8

Natalia Novoselova. In 1987, she graduated with honour from the Belarusian State University. In 2008, she defended the Doctoral Thesis. In 2008, she received an academic title of Senior Researcher. Her major field of research is intelligent methods for data pre-processing, exploratory analysis, classification and knowledge extraction from the biological and medical data.

She is a Leading Researcher of the Bioinformatics Laboratory at the United Institute of Informatics Problems, National Academy of Sciences of Belarus. She has more than 50 scientific publications, including two monographs. Research interests include exploratory analysis of biomedical data, classification and prediction of the disease subtype using the gene expression data, finding the combinations of biomarkers for efficient disease treatment using different data mining methods, multicategory ROC analysis for the identification of the biological risk factors, etc.

Contact data: United Institute of Informatics Problems, Surganova 6, Minsk, 220012, Belarus.

E-mail: novos65@gmail.com

Igar Tom. In 1977, he graduated from Minsk Radio Engineering Institute, specialty “Automated Control Systems”. In 1986, he defended the Doctoral Thesis. His major field of research includes methods for data mining, decision support systems.

From 1986 to 1999, he has worked as an Associate Professor at the Institute of Technical Cybernetics; since 1999, he has been the Head of the Laboratory of Bioinformatics at the United Institute of Informatics Problems. He has more than 185 scientific and methodological publications, including 2 monographs, 3 educational and methodical works, more than 70 full-text papers in journals and collections of articles, including the impact factor.

Research interests include design and development of models and knowledge discovery techniques in molecular-genetic, epidemiological, clinical and laboratory data; creation on their basis of information-analytical systems for healthcare and other industries.

Contact data: United Institute of Informatics Problems, Surganova 6, Minsk, 220012, Belarus.

E-mail: tom@newman.bas-net.by