Protein Function Prediction Using Semantic Driven *K*-Medoids Clustering Algorithm

Ilinka Ivanoska, Kire Trivodaliev, and Slobodan Kalajdziski

***Abstract*—The proposed protein function prediction methods are mostly based on sequence or structure protein similarity and do not take into account the semantic similarity extracted from protein knowledge databases such as Gene Ontology. Many studies have shown that identification of protein complexes or functional modules can be effectively done by clustering protein interaction network (PIN). A significant number of proteins in such PIN remain uncharacterized and predicting their function remains a major challenge in system biology. In this paper we present a "semantic driven" clustering approach for protein function prediction by using both semantic similarity metrics and the whole network topology of a PIN. We apply *k*-medoids clustering combined with several semantic similarity metrics as a weight factor in the distance-clustering matrix. Protein functions are assigned based on cluster information. Results reveal improvement over standard non-semantic similarity metric.**

***Index Terms*—Protein clustering, gene ontology, semantic similarity.**

I. Introduction

Nowadays, most of the similarity-based methods for determining protein function rely on protein's sequence or structure. Unfortunately, the big drawback of these methods is that structure/sequence similarity is not directly related to the protein function, since proteins with significant structure/sequdence similarity can have different functions. Furthermore, proteins with different ancestors and no significant sequence similarity can have the same function, due to evolution.

One of the most important challenges of molecular biology is finding a method for extracting protein function and protein similarity knowledge, consisted in the great amount of protein and genome data in well-known protein databases. An important breakthrough in protein annotation is the creation of the Gene Ontology (GO) [1], the most famous bio-ontology; structured and controlled vocabulary for describing gene and protein products. The GO, structured as a directed acyclic graph (DAG), defines a set of terms used for protein annotation. The GO-annotated interacting proteins can be used as a fertile basis for performing semantic driven protein comparison. This type of comparison is called semantic similarity, and is based on the structure of the GO and the relations between its terms, focusing on the semantic similarity between the terms themselves. It is still not clear which is the best way to calculate semantic similarity considering the current bio-ontologies, but several metrics have been proposed to calculate protein semantic similarity in the context of the GO [2], [3].

A protein interaction network (PIN) consists of nodes representing proteins, and edges representing interactions between proteins. Such networks are stochastic as edges are weighted with the probability of interaction. There is more information in a PIN compared to sequence or structure alone. A network provides a global view of the context of each gene/protein. Hence, our computational function prediction is characterized by the use of a protein's interaction context within the network to predict its functions.

It has been shown that clustering PINs is an effective approach to understand the relationship between the organization of a network and its function [4]. Clustering in PIN is to group the proteins into sets (clusters) that demonstrate greater similarity among proteins in the same cluster than in different clusters. Since biological functions can be carried out by particular groups of genes and proteins, dividing networks into naturally grouped parts (clusters or communities) is an essential way to investigate some relationships between the function and topology of networks or to reveal hidden knowledge behind them.

There are many clustering techniques proposed, but usually standard clustering algorithms are the most effective. In this paper we use *k*-medoids [5] clustering algorithm for standard protein clustering without the use of PIN, and for graph clustering with the use of PIN. Semantic similarity is added to these clustering techniques as a distance metric in the distance matrix. The aim is to present our work for evaluating the semantic similarity metrics and presenting a new system for protein function prediction by the use of this clustering algorithm based on semantic similarity. For a given protein the system can determine similar proteins based on functional (semantic) similarity, and our goal is to see the impact of semantic similarity metrics on determining protein function.

In Section II we present a related work of the existing semantic similarity metrics that will be used, while Section III will give the proposed system architecture for protein function prediction based on the semantic similarity metrics and the whole network topology using the semantic driven *k*-medoids clustering algorithm. Section IV presents experimental results and a discussion of the way that semantic similarity metrics influence the prediction process. Finally, Section V concludes the paper.

II. Overview of Semantic Similarity Metrics

Several approaches are available to quantify semantic similarity between terms or annotated entities in an ontology

Manuscript received September 25, 2013; revised December 3, 2013. This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University, Macedonia.

The authors are with Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, 1000 Skopje, Macedonia (e-mail: ilinka.ivanoska@finki.ukim.mk, kire.trivodaliev@finki.ukim.mk, slobodan.kalajdziski@finki.ukim.mk).

represented as a DAG such as GO. There are essentially two types of methods for comparing terms in a graph-structured ontology: edge-based, that use the edges and their types as data source; and node-based, in which the main data sources are the nodes and their properties.

Edge-based approaches are based mainly on counting the number of edges in the graph path between two terms. The most common technique is the distance that selects either the shortest path or the average of all paths, when more than one path exists. This technique gives a metric of the distance between two terms, which can be easily converted into a similarity metric. While this approach is intuitive, terms at the same depth do not necessarily have the same specificity, and edges at the same level do not necessarily represent the same semantic distance, therefore in this paper we do not take these metrics into account.

A. Node-Based Metrics

Node-based approaches are mainly based on comparing the properties of the terms involved, which can be related to the terms themselves, their ancestors, or their descendants. The most commonly used concept in these approaches is information content (IC), which gives a measure how specific and informative a term is. The IC of a term c is defined as the negative log likelihood $-\log p(c)$, where $p(c)$ is the probability of occurrence of c in a specific knowledgebase, being normally estimated by its frequency of annotation. The use of IC is important because it is more probable (and less meaningful) that two gene products share a commonly used term than an uncommonly used term.

Four most common node-based semantic similarity metrics include Resnik's, Lin's, Jiang and Conrath's and Relevance metric. They were originally developed for the WordNet, and then applied to GO. The definition of Resnik metric [6] is the similarity between two terms as the IC of their most informative common ancestor (MICA) as the following:

$$sim_{\mathrm{Res}}(c_1, c_2) = IC(c_{MICA}) \tag{1}$$

Resnik's metric does not take into account how distant are the terms from their common ancestor. To consider that distance, Lin's metric [7] is defined as:

$$sim_{Lin}(c_1, c_2) = \frac{2 \times IC(c_{MICA})}{IC(c_1) + IC(c_2)} \tag{2}$$

Jiang and Conrath's metric [8] is based on Resnik's metric considering the IC of the two terms compared as in Lin's metric. It is defined as follows:

$$sim_{JC}(c_1, c_2) = 1 - IC(c_1) + IC(c_2) - 2 \times IC(c_{MICA}) \tag{3}$$

Another metric used is the relevance similarity metric [9], which is based on Lin's metric, but uses the probability of annotation of the MICA as a weighting factor to provide graph placement:

$$sim_{\mathrm{Rel}}(c_1, c_2) = sim_{Lin}(c_1, c_2) \times (1 - p(c_{MICA})) \tag{4}$$

B. Hybrid Metrics

Furthermore, there are hybrid methods that combine the two types of methods for semantic similarity, and they give weights to the GO nodes or edges according to their type. In this paper, for the purpose of protein function prediction, we use two semantic similarity hybrid metrics: Wang's [10] and Shortest path [11] semantic similarity metric.

In Wang's metric [10] each edge is given a weight according to the type of developed relationship. For a given term and its ancestor, a semantic contribution of the ancestor to the term is defined, as the product of all edge weights in the "best" path from *the ancestor* to *the term*, where the "best" path is the one that maximizes the product. Semantic similarity between two terms is then calculated by summing the semantic contributions of all common ancestors to each of the terms and dividing by the total semantic contribution of each term's ancestors to that term.

Shortest Path semantic similarity metric [11] uses a shortest path algorithm between terms in GO and gives weights to the terms as a reciprocal value of their IC. The distance between two terms is given by:

$$dist_{SP}(c_1, c_2) = \frac{arctan(\sum_{t_1 \in path_1} \frac{1}{IC(t_1)} + \sum_{t_2 \in path_2} \frac{1}{IC(t_2)})}{\pi / 2} \tag{5}$$

where $path_1$($path_2$) is the shortest path between term c_1(c_2) with its MICA, and t_1(t_2) are the terms on $path_1$($path_2$). Semantic similarity is defined as

$$sim_{SP}(c_1, c_2) = 1 - dist_{SP}(c_1, c_2) \tag{6}$$

III. Protein Function Prediction System Architecture

In our developed approach function prediction process is consisted of few steps: preprocessing; semantic matrix formulation; protein clustering and function prediction; and results evaluation. Fig. 1 shows the developed protein function prediction system architecture using semantic clustering algorithm. The preprocessing step [12] is made to get a highly reliably dataset. Following, semantic similarity between each protein pair is computed to formulate the dataset semantic similarity matrix. This means that we use the semantic similarity as e weighting factor while computing protein distance.

We consider two scenarios: protein function prediction using standard *k*-medoids clustering without the use of PIN, and protein function prediction using PIN graph *k*-medoids clustering. In the former scenario, each protein pair semantic similarity matrix is an input to the *k*-medoids clustering algorithm as a distance matrix.

In the latter, the graph representing the PIN is weighted with the semantic similarity matrix where the weight shows the semantic protein distance (or the probability of interaction between protein pairs). The resulting semantic similarity matrix is an input to the graph *k*-medoids clustering algorithm.

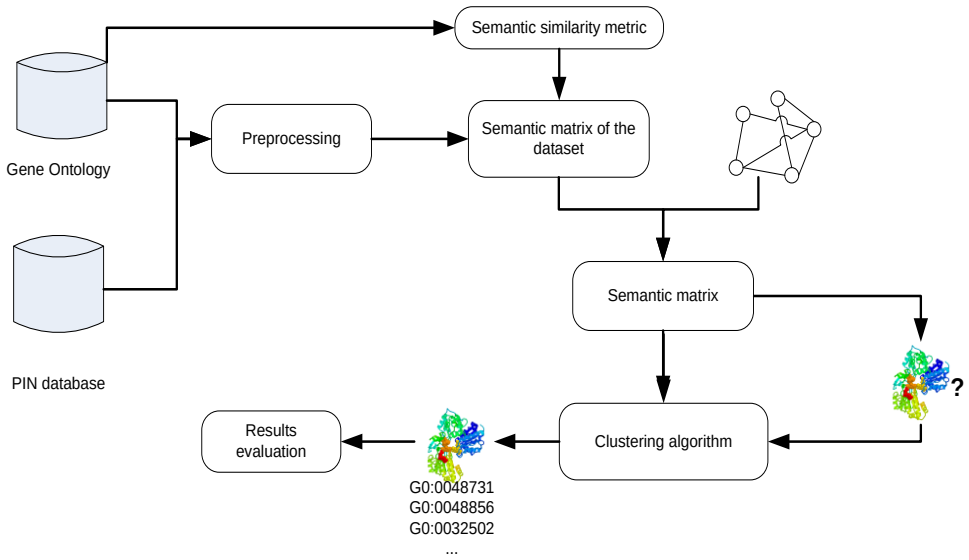


Fig. 1. Protein function prediction system architecture using semantic clustering algorithm.

After the clustering, we set up a strategy for annotating a query protein with the adequate functions according to the functions of the proteins in its cluster. Each function is ranked by its frequency of appearance as an annotation for the proteins in the cluster. This rank is calculated by (7) and it is then normalized in the range from 0 to 1.

$$f(j)_{j \in F} = \sum_{i \in K} z_{ij} \tag{7}$$

where F is the set of functions present in the cluster K, and

$$z_{ij} = \begin{cases} 1, \text{if protein } i \text{ from } K \text{ has function } j \text{ from } F \\ 0, \text{otherwise} \end{cases} \tag{8}$$

In our proposed approach the previously explained Resnik's, Lin's, Relevance, JC's, Wang's and Shortest Path metrics are used in the SS matrix, and therefore, evaluated with the semantic driven protein function prediction algorithm.

IV. Results and Discussion

For the needs of this paper the dataset and the PIN data are compiled, pre-processed and purified from a number of established datasets, like: DIP, MIPS, MINT, BIND and BioGRID. The used dataset is believed to be highly reliable and consists of 2502 proteins from the interactome of the baker's yeast has 12708 interactions between them and are annotated with a total of 888 functional labels [12]. For the purposes of evaluating the proposed methods, the largest connected component of this dataset is used, which consists of 2146 proteins.

Each protein in the dataset is streamed through the prediction process one at a time as a query protein. The query protein is considered un-annotated, that is we employ the leave-one out method. Each of the algorithms works in a fashion that ranks the "proximity" of the possible functions to the query protein. The ranks are scaled between 0 and 1 as explained in 3. The query protein is annotated with all functions that have rank above a previously determined threshold ω. For example, for ω=0, the query protein is assigned with all the function present in its cluster. We change the threshold with step 0.1 and compute numbers of true-positives (TP), true-negatives (TN), false-positives (FP) and false-negatives (FN). For a single query protein we consider the TP to be the number of correctly predicted functions, and for the whole PIN and a given value of ω, the TP number would be the total sum of all single protein TPs. To compare performance between different algorithms we use standard measures as sensitivity (9) and false positive rate (10). Experiments were performed with different number of clusters; however, we conclude that the starting number of clusters in *k*-medoids does not affect the results with different semantic similarity metrics.

$$sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$fpr = \frac{FP}{FP + TN} \tag{10}$$

For the purpose of comparing semantic similarity metrics with other non-similarity metrics we apply cosine similarity between proteins for the experiments that don't involve PIN and minimal hop distance between proteins for those that involve the PIN.

Table I shows protein function prediction results based on semantic *k*-medoids without PIN. It is evident from Table I that the sensitivity for ω=0 is over 80% with all semantic similarity metrics, however, the false positive rate varies from metric to metric. The false positive rate varies between 56% with cosine distance, and 21% with Wang similarity metric for 78-81% sensitivity. This shows a significant false positive rate decrease of 62.5% in favour of semantic similarity metrics (with Wang metric) over standard non-semantic cosine metric. Furthermore, with *k*-medoids dataset clustering without PIN the highest AUC value is achieved with Wang semantic similarity metric, while the other semantic metrics also show AUC value improvement compared to cosine metric in protein function prediction.

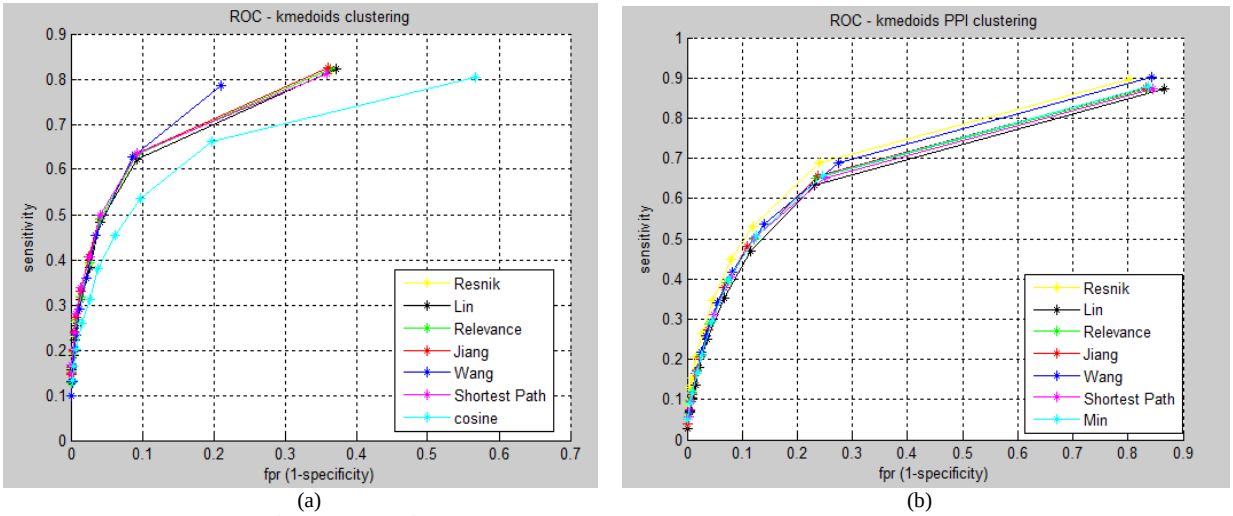
Fig. 2(a) shows the semantic similarity metrics comparison ROC curves for evaluation of annotation using semantic *k*-medoids clustering without PIN.

TABLE I: PROTEIN FUNCTION PREDICTION RESULTS BASED ON SEMANTIC K-MEDOIDS CLUSTERING WITHOUT PIN

M	ω =	0	0.1	0.2	0.3	0.5	0.7	0.9	AUC
Resnik	**sen**	0.818	0.6216	0.4869	0.3909	0.2614	0.1915	0.1273	**0.816**
	fpr	0.3633	0.0923	0.0409	0.0253	0.0076	0.0036	0.0005	
Lin	**sen**	0.8221	0.622	0.4833	0.3835	0.2551	0.1882	0.1302	**0.817**
	fpr	0.3711	0.0925	0.0407	0.0259	0.0075	0.0032	0.0005	
Rel	**sen**	0.8231	0.6322	0.4901	0.3925	0.2669	0.1958	0.1279	**0.820**
	fpr	0.3632	0.0929	0.0411	0.0259	0.0076	0.0033	0.0005	
JC	**sen**	0.8237	0.6349	0.4991	0.4059	0.2734	0.1994	0.146	**0.822**
	fpr	0.3606	0.0922	0.0409	0.0254	0.0078	0.0036	0.0005	
Wang	**sen**	0.7863	0.627	0.4529	0.3588	0.2339	0.1631	0.098	**0.831**
	fpr	0.2109	0.0855	0.035	0.0211	0.0071	0.0031	0.0012	
ShPath	**sen**	0.8126	0.6344	0.4996	0.4076	0.2815	0.2019	0.148	**0.817**
	fpr	0.3596	0.0927	0.0412	0.0262	0.0077	0.0033	0.0005	
Cosine	**sen**	0.8041	0.6629	0.5366	0.4533	0.3126	0.2017	0.1325	**0.759**
	fpr	0.5665	0.198	0.098	0.0613	0.0258	0.0072	0.0022	

TABLE II: PROTEIN FUNCTION PREDICTION RESULTS BASED ON SEMANTIC K-MEDOIDS CLUSTERING WITH PIN

M	ω =	0	0.1	0.2	0.3	0.5	0.7	0.9	AUC
Resnik	**sen**	0.8958	0.69	0.5316	0.4488	0.2648	0.1543	0.0928	**0.750**
	fpr	0.8014	0.2388	0.1202	0.0794	0.0285	0.0093	0.0017	
Lin	**sen**	0.8717	0.6322	0.4707	0.3517	0.1801	0.1025	0.0274	**0.701**
	fpr	0.8654	0.2308	0.1143	0.0676	0.0221	0.0091	0.0011	
Rel	**sen**	0.8739	0.6494	0.4974	0.3911	0.2131	0.118	0.0668	**0.717**
	fpr	0.8339	0.2338	0.1205	0.0717	0.025	0.0088	0.0027	
JC	**sen**	0.8747	0.656	0.4805	0.3798	0.2071	0.1111	0.0399	**0.719**
	fpr	0.8319	0.2372	0.1091	0.0671	0.0242	0.0089	0.0013	
Wang	**sen**	0.9018	0.6871	0.5354	0.4163	0.26	0.1657	0.073	**0.732**
	fpr	0.8424	0.2756	0.1393	0.0812	0.035	0.0168	0.0044	
ShPath	**sen**	0.8733	0.6518	0.5008	0.4042	0.2085	0.1167	0.0576	**0.712**
	fpr	0.8454	0.2502	0.1216	0.0796	0.0252	0.0097	0.0027	
Min	**sen**	0.88	0.6573	0.5062	0.3978	0.2095	0.1151	0.0515	**0.719**
	fpr	0.8371	0.2452	0.1251	0.0758	0.025	0.0088	0.0018	



(a) (b)

Fig. 2. ROC curves for evaluation of annotation using semantic k-medoids clustering (a) without PIN, and (b) with PIN.

Table II shows protein function prediction results for sensitivity and false positive rate, based on semantic *k*-medoids with PIN. In this case, only Resnik's semantic similarity metric gives an improvement in the AUC value, which indicates that the simplest node-based metric outperform the prediction. Also, this shows that SS metrics have a bigger impact on protein function prediction based on *k*-medoids clustering of the dataset without the use of PIN, compared to protein graph *k*-medoids clustering with the use of PIN.

V. CONCLUSION

A new method for protein function prediction using semantic similarity metrics with and without protein interaction networks was presented. It is based on a *k*-medoids clustering algorithm by using a semantic similarity metric as a weight factor in the distance-clustering matrix. The influence of 6 different semantic similarity metrics was evaluated with the protein data. Experiments revealed that the prediction accuracy of the method based on protein clustering without PIN with the use of hybrid semantic similarity metrics outperforms the use of other standard non-semantic similarity metrics. However, with PIN data and graph clustering, results show that semantic similarity metrics do not influence overall protein function prediction, but only give a small decrease in the false positive rate. High sensitivity gives a high false positive rate, therefore, the decrease of the false positive rate is a significant result. These results show the future need of detecting true protein interactions in PINs with the help of other SS metrics.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University, Macedonia.

REFERENCES

[1] The gene ontology consortium: gene ontology documentation. (March 2013). [Online]. Available: http://www.geneontology.org/U
[2] C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. O. Falcão, and F. M. Couto, "Metrics for GO based protein semantic similarity: a systematic evaluation," *BMC Bioinformatic*, vol. 9, 2008.
[3] C. Pesquita, "Improving semantic similarity for proteins based on the gene ontology," Master thesis, University of Lisbon, Portugal, 2007.
[4] S. Brohée and J. V. Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, 2006.
[5] T. Witsenburg and H. Blockeel, "K-means based approaches to clustering nodes in annotated graphs," *ISMIS 2011*, pp. 346-357, 2011.
[6] P. Resnik, "Using information content to evaluate semantic similarity," in *Proc. the IJCAI05*, 1995, pp. 448 – 453.
[7] D. Lin, "An information-theoretic definition of similarity," in *Proc. the 15th Int. Conf. on Machine Learning*, 1998.
[8] J. Jiang and D. W. Conrath, "Semantic similarity based on corpus and lexical taxonomy," in *Proc. 10th Int. Conf. Coling*, 1997.
[9] A. Schlicker, F. Domingues, J. Rahnenfuhrer, and T. Lengauer, "A new measure for functional similarity of gene products based on Gene Ontology," *BMC Bioinformatics* 2006, pp. 7-302, 2006.
[10] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. F. Chen, "A new method to measure the semantic similarity of GO term," *Bioinformatics*, vol. 23, no. 10, pp. 1274-1281, 2007.
[11] Y. Shen, S. Zhang, H. S. Wong, and L. Zhang, "A new method for measuring the semantic similarity on Gene Ontology," in *Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2010, pp. 533–538.
[12] K. Trivodaliev, I. Chingovska, S. Kalajdziski, and D. Davcev, "Protein function prediction based on neighborhood profiles," in *Proc. the ICT Innovations*, Springer-Verlaang, Macedonia, 2009.

Ilinka Ivanoska was born in 1986 in Skopje, Macedonia. She graduated at the Faculty of Electrical Engineering and Information Technologies in Skopje, Macedonia in 2009. Afterwards, she defended her master thesis at the Faculty of Computer science and Engineering in Skopje, Macedonia in the field of Content-based search and retrieval in 2013. During her studies, she was continually awarded for her achievements. As a student in 2005 she was awarded with the "13 November" Skopje award for achieving significant results in her education.

She had her first working experience in the IT company "Nextsense" in Skopje, Macedonia. She started working for the Institute for Computer science and Informatics at the Faculty of Electrical Engineering and Information Technologies in 2009. Her current work in teaching includes courses such as Algorithms and data structures, Object-oriented systems, Object-oriented analysis and design, Information systems and Intelligent information systems. Areas of her scientific research interest are artificial intelligence, bioinformatics, data mining, machine learning, and databases.

Kire Trivodaliev was born in 1982 in Strumica, Macedonia. He graduated at the Faculty of Electrical engineering in Skopje, Macedonia in 2006. Afterwards, he obtained his Master's degree at the same faculty in 2008. During his studies, he was continually awarded for his achievements, and as the best student in the generation he was awarded with an "Engineering ring" award by the Engineering institution of Macedonia.

From 2005 he worked as a demonstrator at the Computer Science Institute - Faculty of Electrical Engineering and Information Technologies in Skopje. The titles Junior assistant and Assistant were acquired in 2007 and 2010, respectively. As an assistant he was involved in maintaining teaching subjects in structures and databases, data mining, artificial intelligence, data analysis algorithms, logic and functional programming and bio-cybernetics. He is an author of more than 15 scientific papers published in international and domestic conferences and journals. His research interest is focused in: bioinformatics, computational biology, artificial intelligence, data mining, and databases.

Slobodan Kalajdziski was born in 1976 in Ohrid, Macedonia. He graduated in 2000 and he obtained his master's degree in Electrical Engineering in Skopje, Macedonia in 2004. His PhD. title is obtained at the Faculty of Electrical Engineering and Information Technologies in Skopje, Macedonia in 2008.

His first work experience is done at Neocom AD, Skopje, Macedonia in 2000, where he worked as a software architect, a project manager and a programmer. Starting from 2000 he worked as a demonstrator at the Computer Science Institute - Faculty of Electrical Engineering and Information Technologies in Skopje. The titles Junior assistant and Assistant were acquired in 2003 and 2005, respectively. From 2008 he was elected as an Assistant professor. As an assistant he was involved in maintaining teaching subjects in programming languages, structures and databases, structured programming, object oriented programming, object oriented systems, databases, object oriented analysis and design and information systems. As assistant professor he held teaching subjects in object oriented systems, database 2, object oriented analysis and design and information systems. He is a coordinator of the Master curricula Content based retrieval; Intelligent Information Systems and Bioinformatics at the Faculty of Computer Science and Engineering in Skopje, Macedonia. Also he holds classes on the part of the subjects covered by these study programs. He is an author of more than thirty scientific papers published in international and domestic conferences and journals and participated in 8 research projects. The fields of his research interest include bioinformatics, systems biology, and knowledge discovery in data, data mining, data warehouses, databases, analysis and design of information systems