International Journal of Bio- Science and Bio- Technology
Vol. 1, No. 1, December, 2009

HMM based approach for classifying protein structures

Georgina Mirceva[1] and Danco Davcev[2]

Faculty of Electrical Engineering and Information Technologies
University Ss. Cyril and Methodius, Skopje, Macedonia
[1]georgina@feit.ukim.edu.mk, [2]etfdav@feit.ukim.edu.mk

Abstract

To understand the structure-to-function relationship, life sciences researchers and biologists need to retrieve similar structures from protein databases and classify them into the same protein fold. With the technology innovation the number of protein structures increases every day, so, retrieving structurally similar proteins using current structural alignment algorithms may take hours or even days. Therefore, improving the efficiency of protein structure retrieval and classification becomes an important research issue. In this paper we propose novel approach which provides faster classification (minutes) of protein structures. We build separate Hidden Markov Model (HMM) for each class. In our approach we align tertiary structures of proteins. Viterbi algorithm is used to find the most probable path to the model. We have compared our approach against an existing approach named 3D HMM, which also performs alignment of tertiary structures of proteins by using HMM. The results show that our approach is more accurate than 3D HMM.

Keywords*: Protein Data Bank (PDB), protein classification, Structural Classification of Proteins (SCOP), Hidden Markov Model (HMM), 3D HMM.*

1. Introduction

To understand the structure-to-function relationship, life sciences researchers and biologists need to retrieve similar structures from protein databases and classify them into the same protein fold. The structure of a protein molecule is the main factor which determines its chemical properties as well as its function. Therefore, the 3D representation of a residue sequence and the way this sequence folds in the 3D space are very important.

With the technology innovation and the rapid development of X-Ray crystallography methods and NMR spectrum analysis techniques, a high number of new 3D structures of protein molecules are determined. The 3D structures are stored in the world-wide repository Protein Data Bank (PDB) [1], which is the primary repository for experimentally determined 3D protein structures. The Protein Data Bank [1] is the primary repository for experimentally determined 3D protein structures. It was created in 1971 at Brookhaven National Laboratories (BNL) in the USA and contained seven macromolecule structures. These structures were created using crystallography methods. During the 1970s, the increase rate of entries was low. Since 1980, the increase rate has become dramatically high due to the rapid technological development. Nowadays, the number of the 3D molecular structure data increases rapidly, since more than 6000 new structures are stored per year in PDB. Today there are more than 61695 protein structures in this repository. In addition to the Euclidean coordinates of atoms, PDB entries contain additional information such as references, structure details, and other features. Every new structure undergoes a correctness control by using appropriate software. After its validation, the protein is given an ID and it becomes available for public use.

In order to find the function of protein molecule, life sciences researchers and biologists need to classify the protein structure. There are several sophisticated methods for classifying proteins structures.

The SCOP (Structural Classification of Proteins) protein database [2], which is held at the Laboratory of Molecular Biology of the Medical Research Council (MRC) in Cambridge, England, describes the structural and evolutionary relationships between proteins of known structure [3]. Since the existing automatic tools for the comparison of secondary structure elements cannot guarantee 100 percent success in the identification of protein structures, SCOP uses experts' experience to carry out this task. This is not a simple task considering the complexity of protein structures, which vary from single structural elements to vast multidomain complexes. SCOP has been accepted as the most relevant and the most reliable classification dataset, due to the fact that SCOP builds its classification decisions based on visual observations of the structural elements of the proteins made by human experts. Proteins are classified in a hierarchical manner that reflects their structural and evolutionary relationship. The main levels of the hierarchy are "Family" (based on the proteins' evolutionary relationships), "Superfamily" (based on some common structural characteristics), and "Fold" (based on secondary structure elements). There are four main structural classes of proteins according to the way of folding their secondary structure elements: all-a (consist of a-helices); all-b (consist of b-sheets); a/b (a-helixes and b-sheets alternating in protein structure); and a+b (a-helixes and b-sheets located in specific parts of the structure). Due to its manual classification methods, the number of proteins released in PDB database which have not yet been classified by SCOP methods drastically increases.

The CATH (Class, Architecture, Topology, and Homologous superfamily) database [4], which is held at the UCL University of London, contains hierarchically classified structural elements (domains) of the proteins stored in the PDB database. The CATH system uses automatic methods for the classification of domains, as well as experts' contribution, where automatic methods fail to give reliable results. For the classification of structural elements, five main hierarchical levels are used: *Class* (is determined by the percentage of secondary structure elements and their packing); *Architecture* (describes the organization of the secondary structure elements); *Topology* (provides a complete description of the whole schema and the way the secondary structure elements are connected); *Homologous Superfamily* (structural elements that have at least 35 percent amino-acid sequence identity belong to the same Homologous Superfamily); and *Sequence* (at this last level of hierarchy, the structures of the same Homologous Superfamily are further classified according to the similarity of their amino acid sequences). CATH database is constructed by applying the Secondary Structure Alignment Program (SSAP) [5]. SSAP (Secondary Structure Alignment Program) utilizes a two-layer dynamic programming technique to align two proteins. In this way the optimal structural alignment of two proteins is determined.

The FSSP (Families of Structurally Similar Proteins) database [6] was created according to the DALI classification method [7] and is held at the European Bioinformatics Institute (EBI). It provides a sophisticated classification of protein structures. The similarity between two proteins is based on their secondary structure. The evaluation of a pair of proteins is a highly time consuming task, so the comparison between a macromolecule and all the macromolecules of the database requires days. Therefore, one representative protein for each class is defined. Every new protein is compared only to the representative protein of each class. However, for an all-to-all comparison of the representative proteins of the database, an entire day is needed.

The classification method of the DALI algorithm [7] is based on the best alignment of protein structures. The 3D coordinates of every protein are used for the creation of distance matrices that contain the distances between each pair of C_α atoms. These matrices are first decomposed into elementary formats, e.g., hexapeptidic-hexapeptidic submatrices. Similar formats make pairs and the emerging formats create new coherent pairs. Finally, a Monte Carlo procedure is used for the optimization of the similarity measure concerning the inner-molecular distances. The DALI method contains a definition of representatives, which are proteins with some special characteristics so that no two representatives have more than 25 percent amino-acid sequence identity. This method is very time-consuming due to the many different alignments performed, the optimization procedures, and the extremely high number of distances between amino acids since a protein may consist of thousands of amino acids.

SCOP, CATH, FSSP and many other sophisticated classifiers are very time consuming. SCOP method is the slowest due to manual classification from experts. CATH method is semi-manual, while FSSP is totally automated, but still is not able to follow the speed of determining novel protein structures.

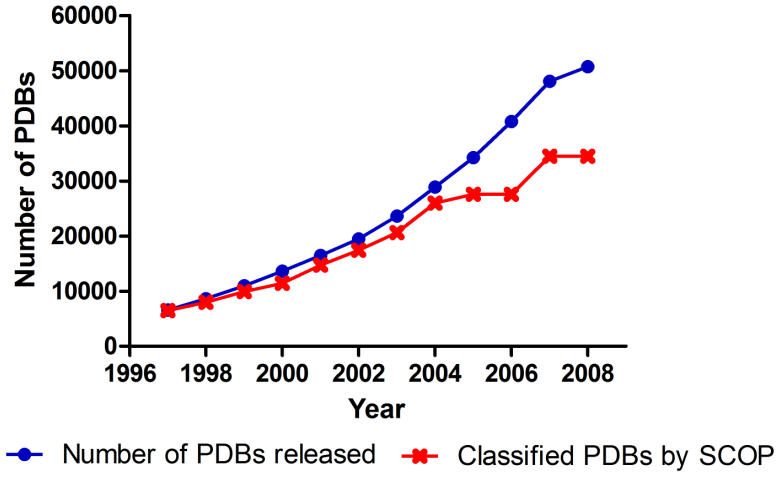

Figure 1. Number of proteins released in PDB versus number of proteins classified by SCOP

Figure 1 presents the gap of number of released proteins in PDB database [1] and number of proteins classified by SCOP. As it can be seen, the number of determined protein structures which are yet not classified by SCOP increases every day. It is due to the fact that retrieving structurally similar proteins using current structural alignment algorithms may take hours or even days to compare protein structures and return the search results Therefore, a need for fast and accurate automated methods for protein classification is obvious. There are various methods for protein classification which are trying to offer efficient and completely automated protein classification.

There are many classification algorithms that can be used for protein classification as Naive Bayesian classifier, k nearest neighbor (K-NN), decision trees, neural networks, Support vector machines (SVM), Hidden Markov Model (HMM) and so on.

ProCC [8], first decomposes protein structures into multiple SSE triplets. The algorithm then extracts 10 features from a SSE triplet based on the spatial relationships of SSEs such as distances and angles. R*-Tree is utilized to index 10-D feature vectors of SSE triplets. Similarly, a query protein is decomposed into multiple SSE triplets, which are searched against the R*-Tree. For each database protein, a weighted bipartite graph is generated based on the matched SSE triplets of retrieval results. A maximum weighted bipartite graph matching algorithm is used for computing an overall similarity score between the query protein and the database protein. Once the algorithm finds the top k similar database proteins, K-NN [9] and SVM [10] techniques are adopted to classify the query protein into known folds. When the classifier cannot assign a class label to the query protein with enough confidence, the algorithm employs a clustering technique to detect new protein folds. The proCC takes 9 minutes to compare a query structure with 2733 database proteins.

In [11], comparative analysis of nine different protein classification methods is performed. The profile-HMM, support vector machines (SVMs) with four different kernel functions, SVM-pair wise, SVM-Fisher, decision trees and boosted decision trees are used as classifiers.

There are many approaches, as method given in [12], for classifying protein structures which use Hidden Markov Model (HMM) for alignment of secondary structures. Alexandrov and Gerstein [13] have introduced the HMM for classifying protein tertiary structures. In [14], it is shown that HMM approach based on tertiary structure is more accurate (10% higher classification accuracy) than the approach based on secondary structure. This is due to the fact that tertiary structure cares much more information than the secondary structure.

Several works [15], [16], [17] apply a consensus strategy to classify the protein domains or folds for newly-discovered proteins by intersecting multiple classification results from classical structural alignment algorithms such as DALI [7], MAMMOTH [18], Combinatorial Extension (CE) [19] and VAST [20]. These consensus approaches yield higher classification accuracies than each individual method. However, a combination of structural alignment algorithms is computationally expensive.

In this paper we propose novel approach for classifying protein 3D structures based on HMMs which consider the tertiary structure of protein molecules. The evaluation of our classification approach is made according to the SCOP hierarchy. Additionally we have compared our approach against an existing approach named 3D HMM [13].

The paper is organized as follows: our approach is given in section 2; section 3 gives some experimental results; while section 4 concludes the paper and gives some future work directions.

2. Our HMM based approach

In this paper we propose novel approach for classifying protein molecules. Our approach uses the well known Hidden Markov Model for building profile for tertiary structure for corresponding class.

2.1. Hidden Markov Model (HMM)

Hidden Markov Models (HMMs) [21] are statistical models which are generally applicable to time series or linear sequences. They have been widely used in speech recognition applications [22], and have been introduced to bioinformatics in the late 80's [23]. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics.

Initialize: $P_{START}(0)=1;\ P_{s_1}(0)=0$ for all s_1

Recursion: for $i=1$ to $length\ (query)$

$$P_{s_2}(i)=e_{s_2}(x_i)\max_{\forall s_1}\{P_{s_1}(i-1)t_{s_1s_2}\}$$

$$backward\_ptr_i=\operatorname{argmax}_{\forall s_1}\{P_{s_1}(i-1)t_{s_1s_2}\}$$

Termination:

$$score(x,Path)=\max_{\forall s_1}\{P_{s_1}(length(query))t_{s_1START}\}$$

$$Path_{\ length(query)}=\operatorname{argmax}_{\forall s_1}\{P_{s_1}(length(query))t_{s_1START}\}$$

Backtracking: for $i=length\ (query)$ to 1

$$Path_{\ i-1}=backward\_ptr_i Path_{\ i}$$

s_1, s_2 are hidden states;
START is start state;
END is end state;
Path is the most probable path; *backward_ptr* is the backward pointer;
x_i= the i-th symbol in the emission sequence; t_{ab} is the transition probability from state a to b;
$e_a(b)$ is the emission probability of symbol b in state a;
$P_k(i)$ =the most probable path ending in state k with observation i

Figure 2. Viterbi algorithm

A HMM can be visualised as a finite state machine. Finite state machines move through a series of states and produce some kind of output, either when the machine has reached a particular state or when it is moving from state to state.

Hidden Markov models, which are extensions of Markov chains, have a finite set of states ($a_1,\ldots,a_n$), including a begin state and an end state. The HMM generates a protein sequence by emitting symbols as it progresses through a series of states. Each state has probabilities associated with it:

- the transition probability T_{ij} that a state a_i will transit to another state a_j, and
- the emission probability $E(x|j)$ that a state a_j will emit a particular symbol x.

Any sequence can be represented by a path through the model. This path follows the Markov assumption, that is, the choice of the next state is only dependent on the choice of the current state (first order Markov Model). However, the state sequence is not known; it is hidden. A HMM is generated for each class.

To obtain the probability that a query belongs to the corresponding class, the query sequence is compared to the HMM by aligning it to the model. The most probable path taken to generate the sequence similar to the query gives the similarity score. It is calculated by multiplying the emission and transition probabilities along the path [24]. The most likely path through the model can be computed

with the Viterbi [25] or forward algorithm [26]. In this paper we have used the Viterbi algorithm. The most probable sequence is determined recursively by backtracking, see Figure 2.

2.2. Efficient representation of protein structures

In our approach we consider the arrangement of protein structure in 3D space. According to our previous analysis [27], by taking into account only the C_α atoms of the protein which form the protein backbone, we can get higher accuracy. The main idea of our approach is to model the folding of protein backbone around its centre of mass by using HMM. In this way we align tertiary structures of protein molecules.

Proteins have distinct number of C_α atoms. So, we have to find a unique way to represent all proteins with sequences with same length. In this approach, we interpolate the backbone of the protein with fixed number of points, which are equidistant along the backbone. According to our previous work [27], in this way by uniformly interpolation of protein backbone we can efficiently extract the most relevant features from the protein structure. Different number of approximation points can be used. In this paper we interpolate the backbone with 64 approximation points, which is sufficient for extracting the most relevant features of protein tertiary structure [27]. In this way, we can present each protein structure with same number of approximation points.

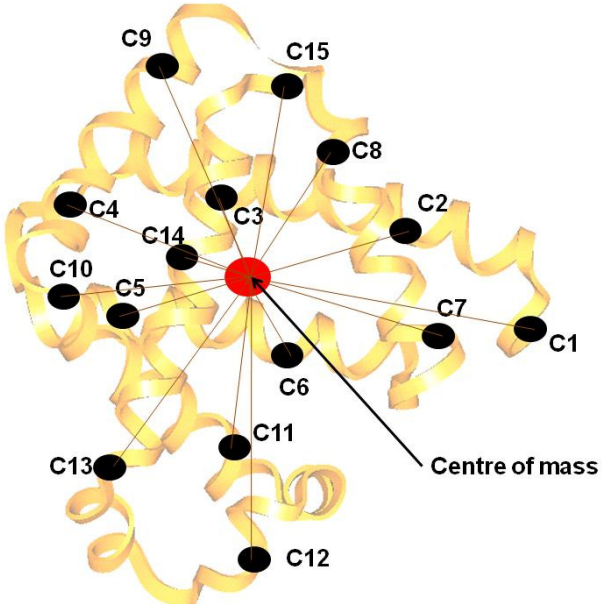

Figure 3. Backbone interpolation

After backbone interpolation, we calculate the Euclidean distances from these points to the centre of mass, as shown on Figure 3. In this way the folding of protein backbone around the centre of mass is considered.

Additionally, distances are quantized in order to obtain discrete values of symbols that can be emitted in each state of the HMM. Different type of quantization can be

used. In order to model the hydrophobic effect [28], we have used uniformly quantization. Experimentally we determined that 20 quantization levels are enough to efficiently present the protein backbone. In this way, by quantizing the distances from approximated points to the centre of mass, our approach models the folding of protein backbone into concentric spheres, as shown on Figure 4.

Figure 4. Our HMM approach

3. Experimental results

We have implemented a system for protein classification based on the HMM approach described above. Our ground truth data contains 6979 randomly selected protein chains from SCOP 1.73 database [2] from 150 domains. 90% of the data set serves as the training data and the other 10% serves as the testing data. We will examine the classification accuracy of our approach according to the SCOP hierarchy.

Table 1. Experimental results of our approach by using 64 approximation points

Q (number of states)	Classification accuracy (%)	Classification time for all test proteins (sec)
16	92.35	420
20	92.51	420
30	90.88	450

In this research we approximated the backbone with 64 approximation points which are sufficient for describing the most relevant features of proteins [27]. First, we examined the influence of number of states (Q) on classification accuracy, see Table 1. As it can be seen, by using 20 states a highest precision is achieved. By using 30 HMM states classification time increases, while classification accuracy is getting worse.

We have additionally compared our approach against an existing approach named 3D HMM [13]. We have used HMMs with Q=20 states. In this analysis, we have used dataset of proteins from globins and IgV (V set from immunoglobulin superfamily) families, as in [13]. We have randomly chosen one training protein from each domain, while other proteins serve

as test data. Namely, test set consists of 754 proteins from globins and 1326 proteins from IgV family. Analysis showed that our approach is more accurate than existing 3D HMM approach [13], see Table 2. Namely, our approach achieves classification accuracy higher for 1.5% for globins and 1.7% for IgV family.

Table 2. Comparison of our approach against 3D HMM

Approach	Classification accuracy for globins family (%)	Classification accuracy for IgV family (%)
Our approach	99.7	98.3
3D HMM	98.2	96.6

Classifiers such as our which are based on HMM can be used for classification at lower levels, but aren't suitable at upper levels of the SCOP hierarchy. Namely, HMM builds profiles for all classes, so if we use this classifier at upper levels we want to model a profile for proteins which are dissimilar. So, if we want to classify proteins at upper levels we have to use other classifier. However, this approach can be incorporated into a hybrid hierarchical classifier, where this approach can be used at family and lower levels of the SCOP hierarchy.

4. Conclusion

In this paper we proposed novel approach for classifying protein molecules by using Hidden Markov Model. We build separate Hidden Markov Model for each class. In our approach we align tertiary structures of proteins.

We have used part of the SCOP 1.73 database for evaluation of the proposed approach. Analysis showed that our approach achieves high precision. Additionally we have compared our HMM approach against an existing 3D HMM approach [13]. The results showed that our approach is more accurate than 3D HMM. Namely, our approach achieves classification accuracy higher for 1.5% for globins and 1.7% for IgV family.

Our future work is concentrated on investigating other protein classifiers in order to obtain higher precision. Also, we want to make a hybrid hierarchical classifier, so HMM can be used at family and lower levels of the SCOP hierarchy, while other corresponding classifiers can be used at upper levels.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The Protein Data Bank", Nucleic Acids Research, vol. 28, 2000, pp. 235-242.

[2] SCOP (Structural Classification of Proteins) Database, http://scop.mrc-lmb.cam.ac.uk/scop/.

[3] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia, "Scop: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures", J. Mol. Biol. 247, 1995, pp. 536-540.

[4] C.A. Orengo, A.D. Michie, D.T. Jones, M.B. Swindells, and J.M. Thornton, "CATH--A Hierarchic Classification of Protein Domain Structures", Structure, vol. 5, no. 8, 1997, pp. 1093-1108.

[5] W.R. Taylor, and C.A. Orengo, "Protein structure alignment", J. Mol. Biol., vol. 208, 1989, pp. 1-22.

[6] L. Holm, and C. Sander, "The FSSP Database: Fold Classification Based on Structure- Structure Alignment of Proteins", Nucleic Acids Research, vol. 24, 1996, pp. 206-210.

[7] L. Holm, and C. Sander, "Protein structure comparison by alignment of distance matrices", J. Mol. Biol., vol. 233, 1993, pp. 123-138.

[8] Y.J. Kim, and J.M. Patel, "A framework for protein structure classification and identification of novel protein structures", BMC Bioinformatics, vol. 7, no. 456, 2006.

[9] T. Hastie, and R. Tibshirani, "Discriminant adaptive nearest neighbor classification", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 6, 1996, pp. 607-616.

[10] C. Cortes, and V. Vapnik, "Support vector networks", Machine Learning, vol. 20, 1995, pp. 273-297.

[11] P. Khati, "Comparative analysis of protein classification methods", Master Thesis, University of Nebraska, Lincoln, 2004.

[12] T. Plötz, and G.A. Fink, "Pattern recognition methods for advanced stochastic protein sequence analysis using HMMs", Pattern Recognition, vol. 39, 2006, pp. 2267-2280.

[13] V. Alexandrov, and M. Gerstein, "Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures", BMC Bioinformatics, vol. 5, no. 2, 2004.

[14] M. Fujita, H. Toh, and M. Kanehisa, "Protein sequence-structure alignment using 3D-HMM", Fourth International Workshop on Bioinformatics and Systems Biology (IBSB '04), Kyoto, Japan, 2004.

[15] T. Can, O. Camoglu, A.K. Singh, and Y.F. Wang, "Automated protein classification using consensus decision", Third International IEEE Computer Society Computational Systems Bioinformatics Conference, Stanford, 2004, pp. 224-35.

[16] S. Cheek, Y. Qi, S.S. Krishna, L.N. Kinch, and N.V. Grishin, "Scopmap: Automated assignment of protein structures to evolutionary superfamilies", BMC Bioinformatics, vol. 5, no. 1, 2004.

[17] O. Camoglu, T. Can, A.K. Singh, and Y.F. Wang, "Decision tree based information integration for automated protein classification", Journal of Bioinformatics and Computational Biology (JBCB), vol. 3, no. 3, 2005, pp. 717-742.

[18] A.R. Ortiz, C.E. Strauss, and O. Olmea, "Mammoth (matching molecular models obtained from theory): An automated method for model comparison", Protein Science, vol. 11, 2002, pp. 2606-2621.

[19] H.N. Shindyalov, and P.E. Bourne, "Protein structure alignment by incremental combinatorial extension (ce) of the optimal path", Protein Engineering, vol. 9, 1998, pp. 739-747.

[20] J.F. Gibrat, T. Madej, and S.H. Bryant, "Surprising similarities in structure comparison", Current Opinion in Structural Biology, vol. 6, no. 3, 1996, pp. 377-385.

[21] Y. Ephraim, and N. Merhav, "Hidden Markov processes", IEEE Transactions on Information Theory, vol. 48, 2002, pp. 1518-1569.

[22] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of IEEE, vol. 77, no. 2, 1989, pp. 257-285.

[23] G.A. Churchill, "Stochastic models for heterogeneous DNA sequences", Bulletin of Mathematical Biology, vol. 51, no. 1, 1998, pp. 79-94.

[24] R. Karchin, "Hidden Markov Models and Protein Sequence Analysis", Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB), 1999.

[25] A.J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", IEEE Transactions on Information Theory, vol. 13, no. 2, 1967, pp. 260-269.

[26] R. Durbin, S. Edy, A. Krogh, and G. Mitchison, "Biological sequence analysis: Probabilistic models of proteins and nucleic acids", Cambridge University Press, 1998.

[27] G. Mirceva, S. Kalajdziski, K. Trivodaliev, and D. Davcev, "Comparative Analysis of three efficient approaches for retrieving protein 3D structures", 4-th Cairo International Biomedical Engineering Conference 2008 (CIBEC '08), Cairo, Egypt, 2008, pp. 1-4.

[28] M.J. Betts, and R.B. Russell, "Amino acid properties and consequences of subsitutions", Bioinformatics for Geneticists, M.R. Barnes, I.C. Gray eds., Wiley, 2003.

Authors

GEORGINA MIRCEVA obtained B.Sc. degree from the Faculty of Electrical Engineering and Information Technologies at the University "Ss. Cyril & Methodius", Skopje, Macedonia in 2007. In 2009 she received her M.Sc. degree in Computer Science from the same university. She is working on her Ph.D. thesis in the field of bioinformatics.

She is currently working as teaching and research assistant at the Computer Science and Informatics Department at the Faculty of Electrical Engineering and Information Technologies.

Her research interests include bioinformatics, multimedia databases and systems, spatiotemporal data and 3D objects, intelligent information systems, machine learning, data mining, artificial intelligence etc.

DANCO DAVCEV obtained the degree of Engineer in Electronics and Computer Science from the Faculty of Electrical Engineering at the University of Belgrade (YU) in 1972. He received his "Doctor – Ingenieur" degree in Computer Science from the University of ORSAY (Paris, France) in 1975 and Ph.D. degree in Informatics from the University of Belgrade (YU) in 1981.

He is currently a professor and head of Computer Science and Informatics Department. He is also head of EU Open and Distance Learning Center at the University "Ss. Cyril & Methodius", Skopje (Macedonia). He has more than 250 research papers presented on International Conferences or published in International Journals in Computer Science and Informatics.

His research interests include multimedia databases and systems, spatiotemporal data and 3D objects, intelligent information systems, distance learning systems, distributed and multi-agent systems, wireless sensor systems and system biology. He is a member of ACM and a Senior Member of IEEE.