Computing and Informatics, Vol. 29, 2010, 1221–1231

PARALLEL APPROACH FOR VISUAL CLUSTERING OF PROTEIN DATABASES

Patryk Orzechowski

Department of Automatics
Faculty of Electrical Engineering, Automatics, Computer Science and Electronics
AGH University of Science and Technology, Kraków
e-mail: patrick@agh.edu.pl

Krzysztof Boryczko

Department of Computer Science
Faculty of Electrical Engineering, Automatics, Computer Science and Electronics
AGH University of Science and Technology, Kraków
e-mail: boryczko@agh.edu.pl

Revised manuscript received 14 January 2010

Abstract. Visualization of a large-scale protein databases may help biologists in discovering similarity between sequences of different organisms. In this article we present a complex approach for visually representing relations between proteins in large scale databases. Our approach includes sequence alignment, mutual distance measurement, clustering and classification of protein sequences. We propose a visual representation method for considered as well-established Pfam 4.0 proteins database. Our objective is to visually reflect the similarity of protein sequences in three dimensional space using nonstandard approach.

Keywords: Clustering algorithms, proteins, sequence alignment, multidimensional scaling

1 INTRODUCTION

Bioinformatics has become one of the crucial grounds of science in the recent years. Decoding human DNA within Human Genome Project (1990–2003) and identifying all of approximately 20 000–25 000 genes catalyzed numerous investigations on DNA, RNA and proteins sequencing, what resulted in appearance of numerous databases.

Proteins are linear polymers of aminoacids that mediate most of the essential functions of the cell [2]. During research on decoding genomes of different organisms, it was revealed that even 45–65 % of proteins detected during the process of decoding shows considerable similarity to previously analyzed and well documented proteins. The reason is that vast majority of proteins is built upon a combinations of one or a couple of functional regions (so-called domains) [7]. Therefore, identifying protein domains permits to group proteins according to their function. Different sequence alignment techniques have been applied so far in order to designate protein families. Major accomplishments were the algorithms for creation pairwise sequence alignment [5], multiple sequence alignment [18, 14] and using some probabilistic models, such as Hidden Markov Models [7, 20].

The growing number of sequences included in databases as well as queries against the whole datasets require heavy computations and started to expand beyond the computational capabilities of individual computers. Consequently, the use of multiprocessors, PC clusters and computer grids for computing became inevitable. It is tried to improve the performance by using more powerful algorithms as well as effective parallelism algorithms [21].

Considering the high computing effort, not many attempts have been taken so far to visually reflect the whole datasets according to similarities of the proteins sequences. A large number of proteins was presented in a single 2D overview around a circle (so called doughnut view) in [19]. The more complex methods intend to visually represent the similarity of the protein sequences by using contact-maps – colored images with chromatic information encoding the chemical nature of the contacts between proteins – and to measure dissimilarity between images by applying image-processing algorithms [3, 8, 15].

In this paper we present an approach for visually representing relations between aligned protein sequences in large scale databases in the form of 3-dimensional visualization of the whole dataset.

2 APPROACH

The proposed approach comprised of the following steps:

- performing alignment of protein sequence pairs and computing their mutual distances (using protein substitution matrix BLOSUM-62 [10], Needleman-Wunsch algorithm [17] and proposed distance metric),
- determining patterns in the database using graph-based algorithm (Shared Nearest Neighbors [6]),

- reducing problem dimensionality (by Multidimensional Scaling [4]),
- visualizing the database,
- assigning evaluated clusters to the original protein families in seed alignments (with Munkres algorithm [13, 16]),
- assessing the classification efficiency by referring to original classification of proteins families.

The approach built on the previously mentioned algorithms allows to obtain effective segmentation of similar protein sequences.

Our approach is presented in Figure 1.

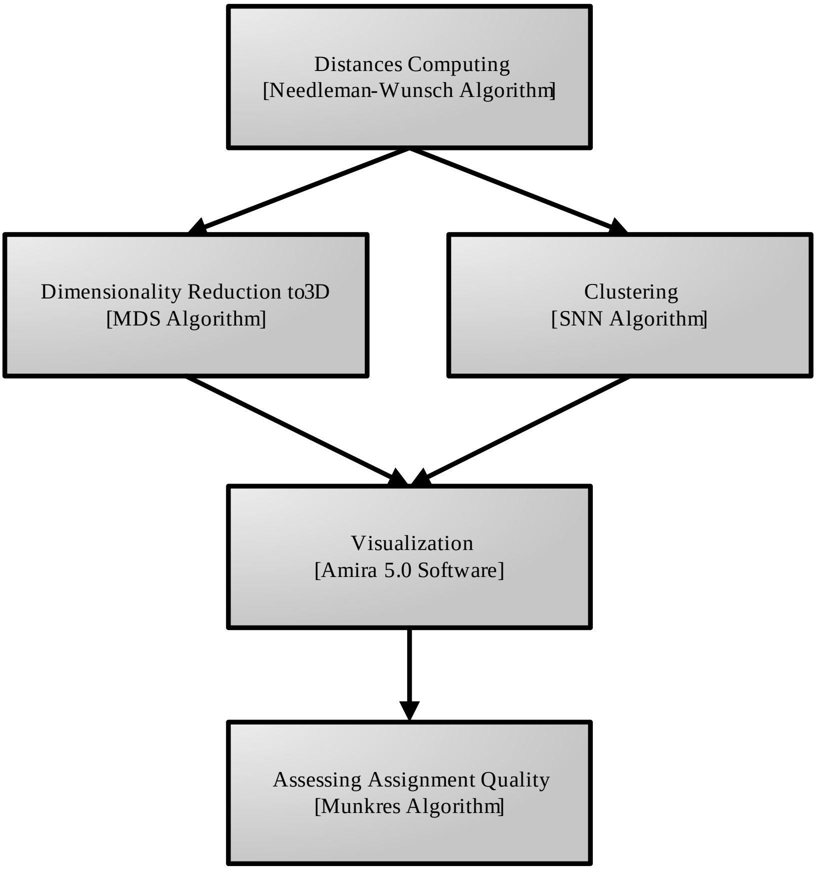


Fig. 1. Algorithmic approach

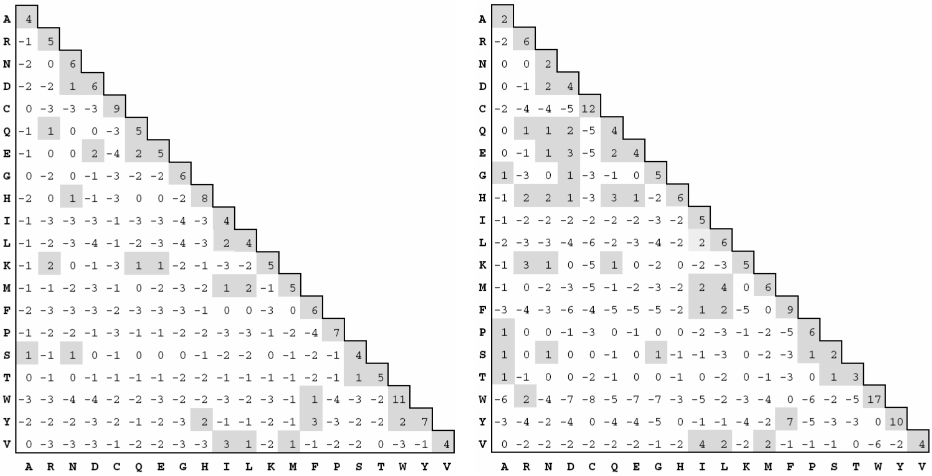
3 APPLIED ALGORITHMS AND TOOLS

In this section we present a short description of algorithmic approach used in our work. The briefly presented set of algorithms constitutes a complex tool for visually representing, interpreting and clustering of large scale datasets.

3.1 Distance Calculation

One of the fastest and most widely used tool for performing search in protein databases is a program performing pairwise sequence alignments called Basic Local Alignment Search Tool (BLAST) [1, 12]. The BLAST algorithm strategy bases on using scoring matrices to compare sequences against the entire DNA or protein sequence database in order to find best matches of the most similar sequences or subsequences.

Our distance calculation algorithm is based on the BLAST concept of pairwise alignment with exploitation of similarity matrices. In order to compute distances between proteins, all proteins sequences are initially aligned against each other in pairs. The optimal mutual alignment is obtained by Needleman-Wunsch algorithm [17] with commonly used similarity matrices: PAM250 and Blosum62 (see Figure 2) that are determining the similarity of amino acids.



A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Fig. 2. PAM-250 and Blosum-62 matrices

The Needleman-Wunsch alignment and calculation of its cost was performed according to:

$$m_{ij} = \max\{M_{i-1,j-1} + s(a_i b_j); \max_{x>=1}(M_{i-x,j} - g_x); \max_{y>=1}(M_{i,j-y} - g_x)\}, \tag{1}$$

where:

- M_{ij} – result in Needleman-Wunsch matrix of aligning first i amino acids in sequence a against first j in sequence b,
- $s(a_i b_j)$ – similarity between i^{th} amino acid in sequence a and j^{th} in b,
- g_x – gap penalty for sequence a,
- g_y – gap penalty for sequence b.

The distance metric d was proposed as invariant to the length of proteins:

$$d(a,b) = \left(1 - \frac{align(a,b)}{align(a,a)}\right)\left(1 - \frac{align(a,b)}{align(b,b)}\right). \tag{2}$$

3.2 SNN Algorithm

Our work is based on the concept of the Shared Nearest Neighbors algorithm presented in [6]. The algorithm discovers clusters of different shapes in noisy data and comprises of the following steps:

1. Compute the similarity matrix (containing the nearest points).
2. Keep only the k most similar neighbors (k nearest neighbors of the similarity graph remain).
3. Construct the SNN-graph (Jarvis-Patrick algorithm [11]). Similarity threshold is applied, components are connected to obtain the clusters.
4. Find the SNN density of each point. Using a user specified parameter, *Eps*, find the number of points that have an SNN similarity of *Eps* or greater to the point. This is the SNN density of the point.
5. Find the core points. Using a user specified parameter *MinPts* find the core points, i.e. all points that have an SNN density greater than *MinPts*
6. Form clusters from the core points. If two core points are within a radius *Eps* of each other, then they are placed in the same cluster.
7. Discard all noise points. All non-core points that are not within a radius of *Eps* of a core point are discarded.
8. Assign all non-noise, non-core points to clusters. Assign points to the nearest core point.

3.3 MDS Algorithm

Multidimensional Scaling Algorithm (MDS) [4] aims at reducing the dimensionality of a large set of features to a smaller one by creating a special mapping reflecting the distances between the original and the reduced set. Configuration of points in reduced space is being generated in which differences between positions in original space and reduced one determine formation of forces. Each step of the algorithm aims at minimizing non-linear stress function, which causes the points to change their location in the reduced 3D space.

3.4 Munkres Algorithm

Initially, the assignment between original Pfam families and the clusters computed by SNN-algorithm remains unknown. Therefore an algorithm needs to assign the original Pfam families in the dataset to the protein clusters. To compute the assignment, we have been using Munkres algorithm also known as Hungarian Algorithm [16], which solves in polynomial time the Assignment Problem (AP).

4 RESULTS

For testing purposes we used Pfam 4.0 seed alignments proteins database, which was reduced to families containing at least 25 members [7]. The total size of database was 12 977 sequences grouped in 280 protein families. The full Pfam 4.0 protein database with seed alignments containing 27 650 sequences grouped in 1 467 families was used as the reference one. The histogram of the families was presented in Table 1.

Proteins	Family ID	Family description
589	PF00560	Leucine Rich Repeat
200	PF00096	Zinc finger, C2H2 type
182	PF00270	DEAD/DEAH box helicase
165	PF00969	Class II histocompatibility antigen, beta domain
131	PF01011	PQQ enzyme repeat
131	PF00904	Involucrin repeat
122	PF00098	Zinc knuckle
110	PF00818	Ice nucleation protein repeat
109	PF00041	Fibronectin type III domain
108	PF00073	Picornavirus capsid protein

Table 1. Histogram of 10 largest families in Pfam 4.0 dataset

All algorithms were implemented in C++ programming language using OpenMP v2.0 standard for parallelism. Computations were performed on SGI Altix 3700 machine, running 128 1.5 GHz Intel Itanium2 processors. Amira 5.0 Software was used for visualization.

Algorithm calculating mutual distances between proteins was implemented in parallel version using OpenMP standard. Efficiency of processing for randomly generated sequences was presented in Figure 3.

The preliminary results show that the efficiency of algorithm increases with the number of sequences included. For relatively small number of processors efficiency reaches its maximal value.

We have managed to achieve 90.7 % conformance rate in family assignment with the original Pfam dataset (12 977 sequences, 297 families) for the reduced dataset with the following SNN-algorithm parameters: $K = 30$, $MinPts = 18$, $Eps = 12$ (see Figure 4) which is considerably good result compared to original families classifica-

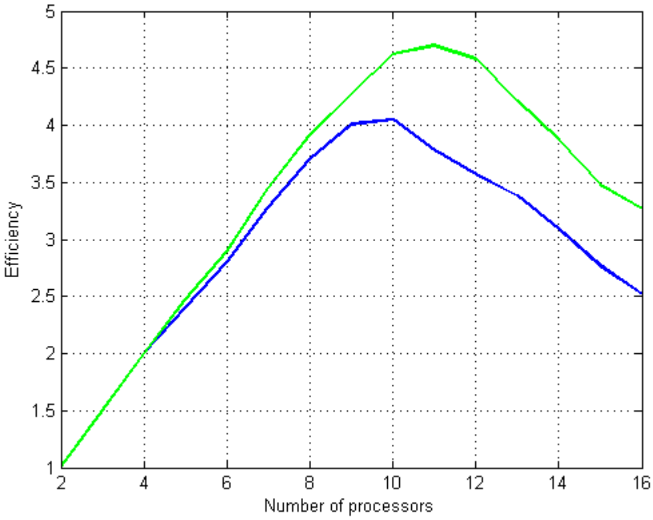


Fig. 3. Efficiency of processing 1 000 (blue) and 3 000 (green) random sequences of 50–500 residues

tion generated by multiple sequence alignments and Hidden Markov Models in the database.

In order to verify the propriety of the approach, we have analyzed the results of clustering of 30 largest original Pfam families in the reduced dataset. The results were as follows: the subset covered 3 488 sequences, out of which 930 were classified as noise which was 26.7 % of the analyzed subset. Five original Pfam families (PF00560, PF00270, PF00969, PF01011, PF00073) covering 516 sequences (14.8 %) were divided by algorithm into more than one cluster. It is noteworthy that every cluster of those families covered sequences purely from the one family. The rest, 2 012 sequences (58.5 %), were assigned correctly.

The analysis of purity of clusters was performed as well. Out of 30 largest designated clusters by SNN algorithm (covering 2 473 sequences), only one cluster covered sequences originating from more than one family: 40 sequences from Insulin/IGF/Relaxin family (PF00049) and 26 sequences from Activin types I and II receptor domain (PF01064), what is 2.2 %.

For the full dataset, the conformance felt down to 85.4 % with the parameters: $K = 15$, $MinPts = 4$, $Eps = 4$. The result is still considered to be very promising, considering the fact that the full dataset covered over twice as much sequences which were divided in nearly 5-times more families as in the reduced dataset (see Figure 5).

Similar biological reference analysis was performed for the 10 largest families and clusters of the full dataset. They covered 1 847 sequences, out of which 297 were classified as noise (16.1 %). Only one family out of 10 most numerous remained

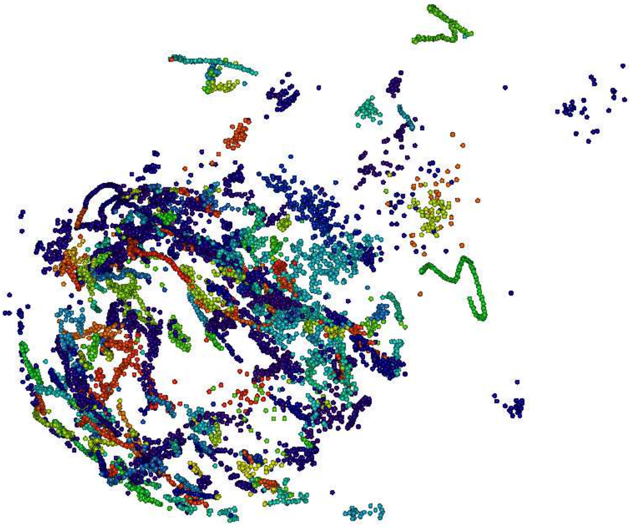


Fig. 4. Visualization of reduced Pfam 4.0 dataset – the result of clustering of data obtained by MDS with SNN algorithm. Proteins properly assigned. Total conformance rate was 90.7 %.

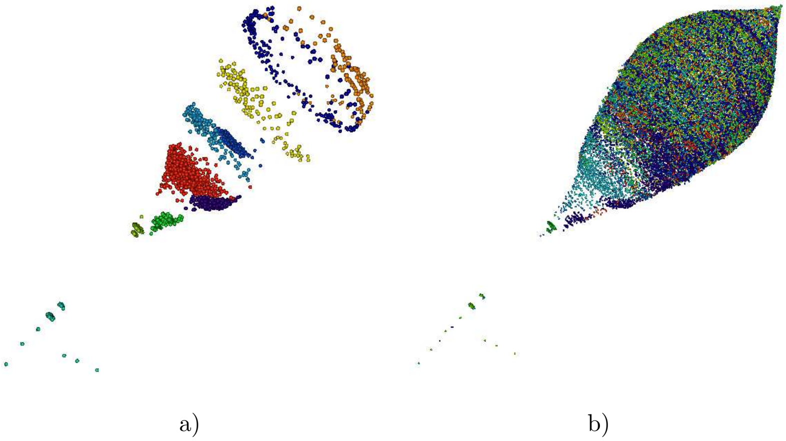


Fig. 5. Visualization of full Pfam 4.0 dataset – the result of clustering of data obtained by MDS with SNN algorithm. a) Ten largest clusters b) all proteins

undivided: PF00041. The rest got divided into two or more seperate clusters. Each of the clusters retained its purity. Ten largest clusters found by algorithm covered 1 341 sequences. Only one of the clusters mixed two families: 100 sequences from Fibronectin type III domain (PF00041) and 17 sequences of PKD domain (PF00801) – totally 8.7 % of the whole. Those families were assigned to one cluster only (undivided between different clusters).

5 CONCLUSIONS AND FUTURE WORK

We proved that technique may be applied to represent visual similarity between the sequences by projecting the distances into three-dimensional space. The proposed approach includes computer-intensive methods:

- Needleman-Wunsch Algorithm – $O(n^2)$,
- SNN – $O(n^2)$,
- MDS – $O(n^2)$,
- Hungarian Algorithm – $O(n^3)$.

Execution time was shortened by parallel computation using OpenMP paradigm.

Complexity reduction may be achieved in MDS by storing M nearest and N futhest neighbors or by using histogram of distances or reducing computational time, and parallel versions of each algorithm are being implemented. We are expecting to improve quality of clustering process by applying mutual nearest neighborhood concept [9] to clustering of core points in SNN-algorithm. The algorithms need further improvement to be able to cover the latest Pfam – 23.0 release (July 2008, 10 340 families, 3 925 943 sequences).

Acknowledgements

The authors would like to express our sincere appreciation to Professor Jacek Kitowski for his valuable remarks and to the Academic Computer Centre CYFRONET AGH for providing computing power for our computations.

REFERENCES

[1] Altschul, S.—Gish, W.—Miller, E.—Myers, E.—Lipman, D.: A Basic Local Alignment Search Tool. J. Mol. Biol., Vol. 215, 1990, pp. 403–410.

[2] Bourne, E. P.—Weissig, H.: Structural Bioinformatics. John Wiley and Sons, Inc., New York 2003.

[3] de Melo, R. C.—Lopes, C. E.—Fernandes Jr., F. A.—da Silveira, C. H.—Santoro, M. M.—Carceroni, R. L.–Meira Jr., W.—Arajo Ade, A.: A Contact Map Matching Approach to Protein Structure Similarity Analysis. Genet. Mol. Res., Vol. 5, 2006, No. 2, pp. 284–308.

[4] Dzwinel, W.—Błasiak, J.: Method of Particles in Visual Clustering of Multidimensional and Large Data Sets. Future Generation Computers Systems, Vol. 15, 1999, pp. 365–379.

[5] Eidhammer, I.—Jonassen, I.—Taylor, W. R.: Structure Comparison and Structure Patterns. Journal of Computational Biology, Vol. 7, 2000, pp. 685–716.

[6] Ertoz, L.—Steinbach, M.—Kumar, V.: Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA (USA) 2003.

[7] Finn, R. D.—Tate, J.—Mistry, J.—Coggill, P. C.—Sammut, J. S.—Hotz, H. R.—Ceric, G.—Forslund, K.—Eddy, S. R.—Sonnhammer, E. L.—Bateman, A.: The Pfam Protein Families Database. Nucleic Acids Research, Database Issue, Vol. 36, 2008, pp. 281–288.

[8] Fernandes, F. Jr.—Lopes, C. E. R.—de Melo, R. C.—Santoro, M. M.—Carceroni, R. L.—Meira, W. Jr.—Araujo, A. A.—Silveira, C. H.: An Imagematching Approach to Protein Similarity Analysis. Proceedings of 17th Brazilian Symposium on Computer Graphics and Image Processing, October 2004.

[9] Chidananda Gowda, K.—Krishna, G.: Agglomerative Clustering Using the Concept of Mutual Nearest Neighbourhood. Pattern Recognition, Vol. 10, 1978, No. 2, pp. 105–112.

[10] Henikoff, S.—Henikoff, J. G.: Amino Acid Substitution Matrices from Protein Blocks. Proc. Natl. Acad. Sci. USA, Vol. 89, 1992, pp. 10915–10919.

[11] Jarvis, R. A.—Patrick, E. A.: Clustering Using a Similarity Measure Based on Shared Near Neighbors. IEEE Transactions on Computers, Vol. C-22, 1973, No. 11.

[12] Korf, I.—Yandell, M.—Bedell, J.: BLAST. OReilly 2003.

[13] Kuhn, H. W.: The Hungarian Method for the Assignment Problem. Naval Research Logistic Quarterly, Vol. 2, 1955, pp. 83–97.

[14] Leibowitz, N.—Nussinov, R.—Wolfson, H. J.: MUSTA – A General, Efficient, Automated Method for Multiple Structure Alignment and Detection of Common Motifs: Application to Proteins. J. Comput. Biol., Vol. 8, 2001, No. 2, pp. 93–121.

[15] Lisewski, A. M.—Lichtarge, O.: Rapid Detection of Similarity in Protein Structure and Function Through Contact Metric Distances. Nucleic Acids Research, Vol. 34, 2006, e152.

[16] Munkres, J.: Algorithms for the Assignment and Transportation Problems. Journal of the Society of Industrial and Applied Mathematics, Vol. 5, 1957, No. 1, pp. 32–38.

[17] Needleman, S. B.—Wunsch, C. D.: A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. J. Mol. Biol., Vol. 48, 1970, No. 3, pp. 443–53.

[18] Notredame, C.: Recent Progress in Multiple Sequence Alignment: A Survey. Pharmacogenomics, Vol. 3, 2002, No. 1, pp. 131–144.

[19] Pinelle, D.—Gutwin, C.: Overview and Detail for Wide Data with Doughnut Views. Technical Report HCI-TR-2001-02, Computer Science Department, University of Saskatchewan, 2001.

[20] Sonnhammer, E. L. L.—Eddy, S. R.—Durbin, R.: Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignment. Proteins, Vol. 28, 1977, pp. 405–420.

[21] Zomaya, A. Y. ed.: Parallel Computing for Bioinformatics and Computational Biology. Wiley Series on Parallel and Distributed Computing, Wiley-Interscience 2005.

Patryk Orzechowski obtained his M.Ṡc. degree in Automatics and Robotics in 2006 and a M.Ṡc. degree in Computer Science in 2008 from AGH University of Science and Technology, Krakow, Poland, where he works as a teaching assistant. His research interests are in the areas of bioinformatics, artificial intelligence and image processing. He is also interested in feature extraction and clustering methods for computer simulation.

Krzysztof Boryczko obtained his Ph. D. degree in Computer Science in 1996 from Department of Computer Science, AGH University of Science and Technology, Krakow, Poland where he is now an Associate Professor. His research interests focus on large scale simulations with particle methods. He is also interested in feature extraction, clustering methods for analysis of simulation data and scientific visualization.