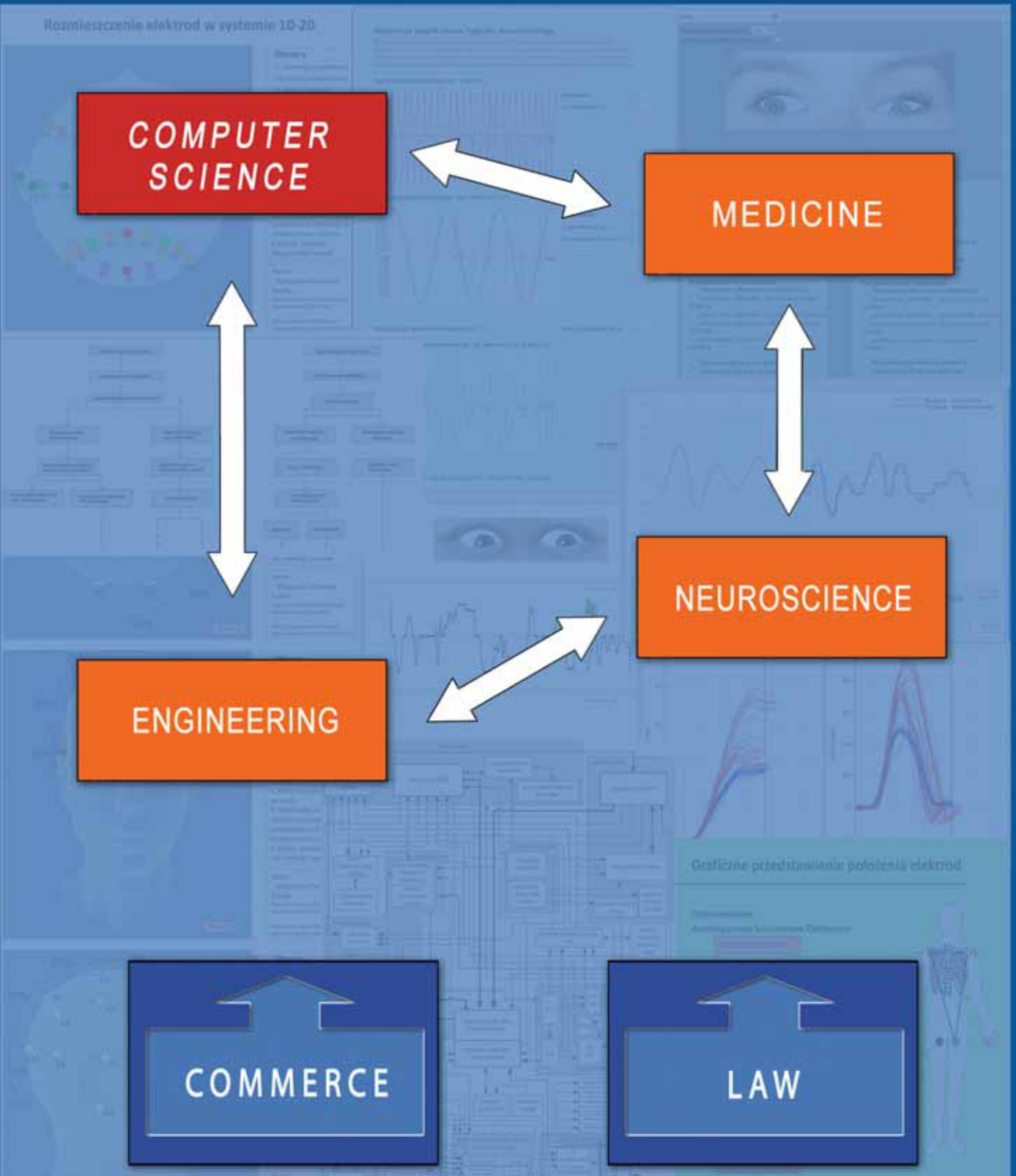
BIO-ALGORITHMS AND MED-SYSTEMS



JOURNAL EDITED BY JAGIELLONIAN UNIVERSITY - MEDICAL COLLEGE - KRAKOW – POLAND
Vol. 5, No. 10
2009
http://www.bams.cm-uj.krakow.pl

EDITORIAL BOARD

Polish Ministry of Science and Higher Education journal rating: 4.000

Sustaining institution: Ministry of Science and Higher Education

Edition: 300 copies

ISSN 1895-9091 (print version)
ISSN 1896-530X (electronic version)

http://www.bams.cm-uj.krakow.pl

Contents

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 5–8

WHY AND HOW SHOULD A VIRTUAL PATIENT BE CONSTRUCTED?

PAWEŁ SPÓLNIK

Chair of Medical Biochemistry, Jagiellonian University – Collegium Medicum, Kopernika 7, 31-034 Krakow, POLAND

Introduction

Graduating with even the highest grade does not necessarily guarantee good preparation for the profession of a doctor. Evidence of this fact may be found in the apparent failure of students of later years and young doctors, even those who are studying or who graduated from the world's top medical universities, to successfully carry out one of the daily clinical activities, namely, the ward round (1, 2). The skills required for this activity include not only the ability to perform differential diagnosis, treatment and the ability to perform minor surgeries, but also the ability to effectively and precisely communicate with patients, and, last but not least, the ability and readiness to work in teams. As many as 30% of patients experience fear connected with medical appointments, and do not understand the language doctors use to communicate with them (3). The average doctor's visit in the case of a ward patient is approximately 5 min long. Only 44% of patients undergo cursory physical examination. One particular evaluation of ward round quality revealed that convalescents after myocardial infarction received the appropriate medication (as defined by the established guidelines) in fewer than half of all cases (4). These data suggest that there are flaws within the system of medical education. This problem has recently been attracting a lot of attention. Consequently, some new solutions are being incorporated in order to better prepare young doctors for medical practice : problem-based learning, integration of clinical disciplines and basic sciences (horizontal and vertical), as well as the application of computer systems in education (e-learning). The creation of databases of virtual patients (VP) is also among these solutions (5-8). Government-supported institutions and projects aimed specifically at improving the situation in the teaching of medical sciences have been developed (AMEE, Association for Medical Education in Europe; COMET, Consortium on Medical Education and Technology; CHEC, Canadian Healthcare Education Commons; eViP, Electronic Virtual Patients; WebSP, Web-based Simulation of Patients).

E-learning and electronic virtual patients

The development of the e-learning technology makes it possible to more effectively prepare future doctors for work in hospitals. It should be emphasized, however, that the direct doctor-patient interaction is irreplaceable in medical didactics. The computer technology facilitates the teaching of differential diagnosis, which forms the essence of a doctor's work. Arguably, the biggest advantage offered by VPs is that there are no clinical consequences of wrong decisions made by students. Other undeniable advantages include easy accessibility, a unified grading system, the attractive presentation of the problems, and the possibility of using previously prepared comments. Medical universities in Europe, supported by EU grants, are developing a program for improving virtual patients (eViP) (9). Developments in computer science are making it possible to efficiently manage the created databases. The elaboration of simple and extended diagnostic models should only be carried out in the frame of close cooperation with doctors.

Development of VP "from scratch"

The elaboration of an original virtual patient from scratch is not an easy task due to the necessity of exchanging data between existing models, and their improvement and adaptation to local conditions. However, elaboration from scratch does carry some advantages with it. For example, it makes it possible to create a patient in the reality immediate to the student. A clinician may prepare a case ex post based on experience and archival data, or formulate it concurrently, while the diagnostic-therapeutic process is still taking place. The reality of clinical work generally favours the first approach.

Among the virtual patients we have elaborated, there is one case which makes students particularly aware of the frequent problem of upper gastrointestinal bleeding (to be anatomically precise, bleeding from the proximal segment of the digestive tract to the ligament of Treitz). Around 50-100/100,000 people are afflicted with this problem each year. Clinicians often need to decide between a dozen or so possible diagnoses. This stems from the fact that each case is different. The discussed diagnostic problem may be a self-containing state which may not be detected clinically, or a rapid process leading to hypovolemic shock and death (10, 11).

Authoring of virtual patients

Let us introduce a 40-year old patient who was brought into admissions by an ambulance. He complained of pain in the epigastrium and vomiting. A gastroscopic examination revealed a non-bleeding prepyloric ulcus, and the CLO-Test was positive. A final gastroscopic examination, performed after the eradication of *Helicobacter pylori*, indicated significant improvement. The patient was discharged in good general condition. It should not be too hard for the student to solve this case, since at the outset of the task they receive vital hints to help them conduct the differential diagnosis. However, the picture may be distorted by other clinical data, such as those normally available to the doctor in the initial stage of the diagnosis, i.e. data concerning basic life parameters, ECG results, and X-ray images of the chest, in both planes. The initially presented data correspond to those available to the doctor that admitted the patient first. The medical history taken by the doctor working in the emergency service did not suffice to put forward a correct working hypothesis. As part of the standard procedure in the hospital he was admitted to, an ECG examination was carried out, and the nurse recorded the patient's life parameters. It is with such data and the information from the physical examination and the medical history that the student is confronted with. One element, the teaching which is of critical importance for the work of a doctor, especially in ward/night duty conditions, is the ability to carry out efficient differential diagnosis, which should first be focused on either ruling out or confirming potentially fatal conditions. If the student focuses on electrocardiographic symptoms that indicate myocardial ischemia, the diagnosis and the treatment may be pursued in the wrong direction, thus putting the patient in danger. The hints from the physical examination and the medical history (the type of pain experienced, taking of medication from the NSAID (Non-steroidal Anti-inflammatory Drug) group, as well as the presence or absence of melena provide data with which the correct diagnosis may be made with high probability. Self-evaluation questions concerning, for example, the molecular and pathophysiological mechanisms of ulcer formation during medication with NSAIDs are somewhat more difficult. Commentaries by experts outline the molecular mechanism of the acetylation of serine in the cyclooxygenase active site. To complete the exercise, the student must apply their knowledge in biochemistry, pharmacology, physiology, and pathophysiology. In the further stage of the diagnostic procedure, questions pertaining to the mechanism of drugs from the PPI (proton pump inhibitors) group should be addressed. This is where knowledge strictly connected with basic subjects is again required. In the presented case, we also emphasize the accurate explanation of the changes observed through laboratory tests, e.g. the results of an arterial blood gas analysis and the mechanism through which the level of urea in the serum of a patient with gastrointestinal bleeding is elevated. The biological mechanisms which are exploited in the application of the non-invasive ureatic test are also explained. In addition, the student is expected to know the different treatment options for *Helicobacter pylori*, and to know how to act if the initial treatment fails. In another case that is being elaborated, we try to introduce a lot of data that combine clinical practice with basic subjects. For example, when discussing a patient with mononucleosis, we pay attention to the mechanism of the virus's interaction with human cells containing antigen CD 21 (which explains its predilection towards the cells of the immune system). When interpreting the elevated level of transaminases, we also present their role and examples of physiological reactions.

Discussion and conclusion

Commentaries concerning basic subjects, which facilitate the understanding and practical application of traditionally conceived preclinical subjects in a doctor's work, require special preparation. It is easier to consider these dependencies while taking a test, whereas in actual clinical conditions there is often no time for this, and effective action is expected immediately. According to some estimates, only 15% of the knowledge acquired during preclinical education is put into practice in a noticeable way (12). That is far too little. Well-constructed VP

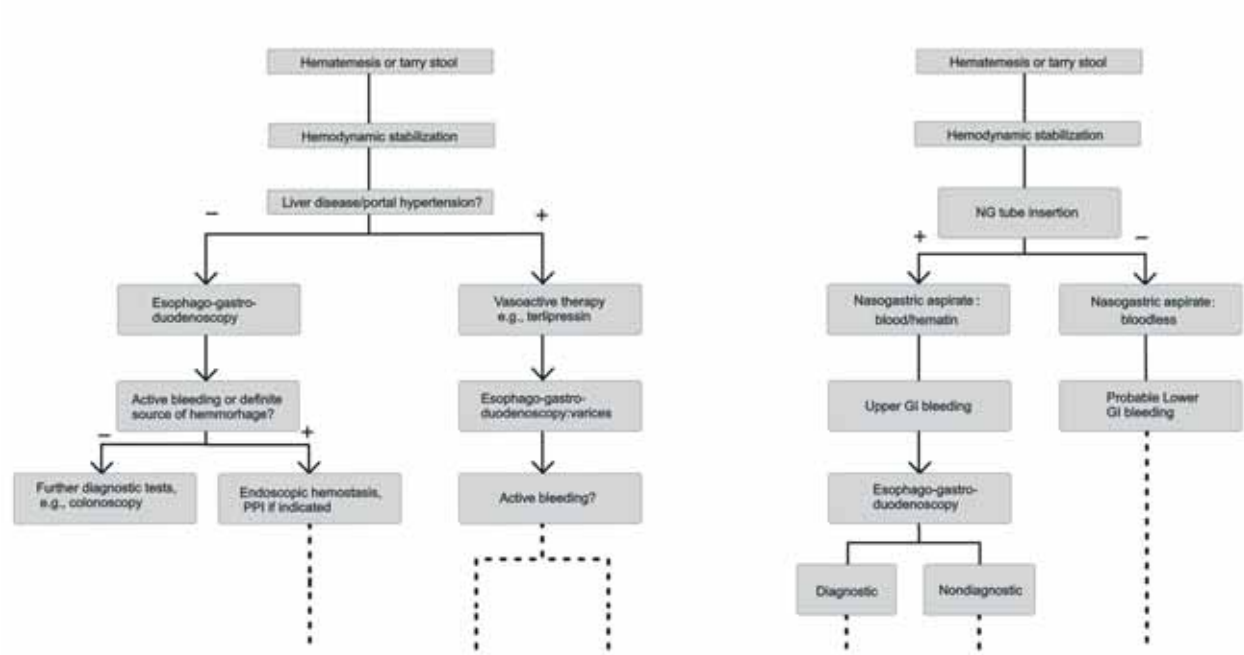


Fig. 1. Different algorithms for the management of acute upper gastrointestinal bleeding. (NG, nasogastric; GI, gastrointestinal; PPI, proton pump inhibitor).

models and the adjustment of the teaching strategy provide an opportunity to change the misguided opinion students sometimes express, namely that basic subjects are completely irrelevant to professional work. A complex presentation of each problem – in the form of a clinical case discussed by expert in various fields (in practice – the teaching of basic subjects based on clinical cases as well as presenting the basic issues again while discussing the cases during clinical practice) – appears to be one of the ways in which the level of education may be improved (5, 6, 8).

This trend towards integration in teaching may be introduced horizontally (a joint presentation of the case by teachers of preclinical subjects) and vertically (in the form of cooperation between teachers of basic subjects and clinicians) (13). It seems like a good idea to have specialists in basic subjects present the background of molecular abnormalities in a series of lectures given by clinicians.

The problems faced by a doctor elaborating a case are different guidelines and procedures concerning the clinical cases to be developed. The eventually administered therapies may have a slightly different course. For example, in cases similar to the one discussed earlier, there are different procedures in different EU countries. There are countries where patients with gastrointestinal bleeding are assigned to surgical wards by default, whereas in other countries such patients are initially treated by internists (Fig. 1). What course is taken is often influenced by guidelines and hints published by the respective specialists associations. Another problem faced when elaborating a case is connected with medical documentation. On some occasions, the principles of medical practice do not require taking images that might be didactically relevant to be recorded during diagnostics. At other times, for reasons beyond the researchers' control, the visual documentation is not sufficient (for example, in the case discussed earlier endoscopic images could not be taken during the first gastroscopic examination). It seems that a case is the most consistent when using documentation for only one patient. However, there are times when didactic goals make it necessary to use the available images taken from similar cases, if the original documentation may not be accessed.

Not only does solving a case teach the student the skills necessary to take therapeutic action, but, by studying basic subjects, they should assimilate knowledge on various pathomechanisms, which in turn would allow them to act in a judicious and well-thought-out way, and to anticipate certain phenomena.

It appears that the application of e-Learning platforms fills the gap between textbook theory and practice. It may also serve to prepare the student to make diagnostic and therapeutic recommendations. What should be emphasized, however, is the role of practice, which must complement the theoretical background.

References

1. Nørgaard K., Ringsted C., Dolmans D.: Validation of a checklist to assess ward round performance in internal medicine, Med. Educ. 2004; 38(7): 700-707.

2. Wray N. P., Friedland J. A., Ashton C. M., Scheurich J., Zollo A. J.: Characteristics of house staff work rounds on two academic general medicine services, Med. Educ. 1986; 61: 893-900.
3. Montague M., Hussain S. S. M.: Patient perceptions of the otolaryngology ward round in a teaching hospital, J. Laryngol. Otol. 2006 Apr; 120(4): 314.
4. Nikendei C., Kraus B., Schrauth M., Briem S., Jünger J.: Ward rounds: how prepared are future doctors?, Med. Teach. 2008; 30(1): 88-91.
5. Davis M. H., Harden R. M.: AMEE Medical Education Guide No. 15: Problem-based learning: a practical guide, Med. Teach. 1999; 21(2): 130-140.
6. Ellaway R., Masters K.: AMEE Guide 32: e-Learning in medical education Part 1: Learning, teaching and assessment, Med. Teach. 2008; 30(5): 455-473.
7. Ellaway R., Poulton T., Fors U., McGee J. B., Albright S.: Building a virtual patient commons, Med. Teach. 2008 30(2): 170-174.
8. Harden R. M.: E-learning – caged bird or soaring eagle?, Med. Teach. 2008; 30(1): 1-4.
9. Hege I., Kononowicz A., Pfahler M., Adler M., Fischer M. R.: Implementation of the MedBiquitous standard into the learning system casus, Bio-Algorithms and Med-Systems 2009; 5(9): 51-55.
10. Conn H. F. and others: Current therapy: latest approved methods of treatment for the practicing physician, Saunders Elsevier, 2007.
11. Herold G., Innere Medizin, 2009.
12. Oberle S., Huber S., Tonshoff B., Nawrotzki R., Huwendiek S.: Repurposing virtual patients for the preclinical years – A pilot study, Bio-Algorithms and Med-Systems 2009; 5(9): 79-82.
13. Integration of basic and clinical sciences – AMEE Conference archive 2008, http://www.amee.org/index.asp?llm=27

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 9–13

VOGE – JAVA APPLICATION FOR PRESENTATION OF NEVER BORN PROTEINS GENE STRUCTURE ELEMENTS

Monika Piwowar[1], Ewa Matczyńska[1], Filip Pomański[2], Adrian Kośmider[2], Damian Kość[2], Michał Swatowski[2], Piotr Więcek[2], Tomasz Szepieniec[3]

[1] *Department of Bioinformatics and Telemedicine, Jagiellonian University, Collegium Medicum, Sw. Anny 12, 31-008 Krakow, Poland*
[2] *Faculty of Physics, Astronomy and Applied Informatics, Jagiellonian University, Reymonta 4, Krakow, Poland*
[3] *Academic Computer Center CYFRONET, Nawojki 11, 30-950 Krakow, Poland*

Abstract: VoGE application for visualizing gene structure elements in nucleotide sequences is presented in the paper. Genomic analyzing of genome sequences was continuation of proteomic efforts in the EuChinaGrid project that was oriented on the structure prediction of never born proteins, but probably with pharmacological application (there were 10**4 protein sequences generated which were used to create structures of that sequences). Finding of gene traces of never born proteins in all accessible sequenced genomes was one of the aim. As a results of searching genome materials there were found regions including protein-coding gene fragments that VoGE (Visualiser of Gene Elements) application presents. Graphical presentation of particular sequences enable user to see localization of coding NBP sequence and gene element composition in wider sequence context. Application interface and menu is intuitive so it seems to be easy to use.
Keywords: never born proteins, gene structure elements, gene finding

Introduction

Genomic part of the efforts in Euchina grid was the continuation of proteomic achievements in folding of never born protein with pharmacological potential [1, 2].

Motivation for finding never born proteins traces was suspicion that many proteins not occurring in nature as a real proteins can have sequences representation in genetic material. Such hypothesis suggest that DNA can accumulate information about proteins or fragments of proteins that may have existed during ancient time, but now are withdrawn from nature.

Innovation of genomic analysis rely on finding information about proteins in genomic regions where it theoretically should not be (in case of humane genome it is big amount genetic materials (about 97%) that does not encode any known proteins).

Searching all accessible completely sequenced genetic information to identify stretches of genomic sequence about potential biological function and results presentation was the aim. To obtain it all genetic information was translated to aminoacids sequences (DNA AA) in three reading frames. Then DNA AA were searched by random generated sequences (represent never born proteins [3, 4]), the same sequences which were used to creation theirs three dimensional structures (fig. 1).

The most interesting in searching DNA AA was identifying never born coding stretches that represent structural gene elements especially exon. GENSCAN software was used For identification gene elements [8] (fig. 2). Results were put into MySQL database. They are used by VoGE application as VoGE gets all data from the database.

Because there was assumed that exons can be part of the query sequences so for viewing localization of coding sequences there were taken lager length of DNA than query sequence (fig. 3).

Materials and methods

Genome data:

DNA from National Center of Biotechnology Information were taken for analysis (ftp.ncbi.nih.gov). Entire genetic information was searched not only humane genome but also other Eukaryotes genomes e.g. genomes of animals, plants, fungi and Protists genomes and organelles.

Never Proteins data:

Never born proteins were obtained from Roma Tre University proteomic group [13].

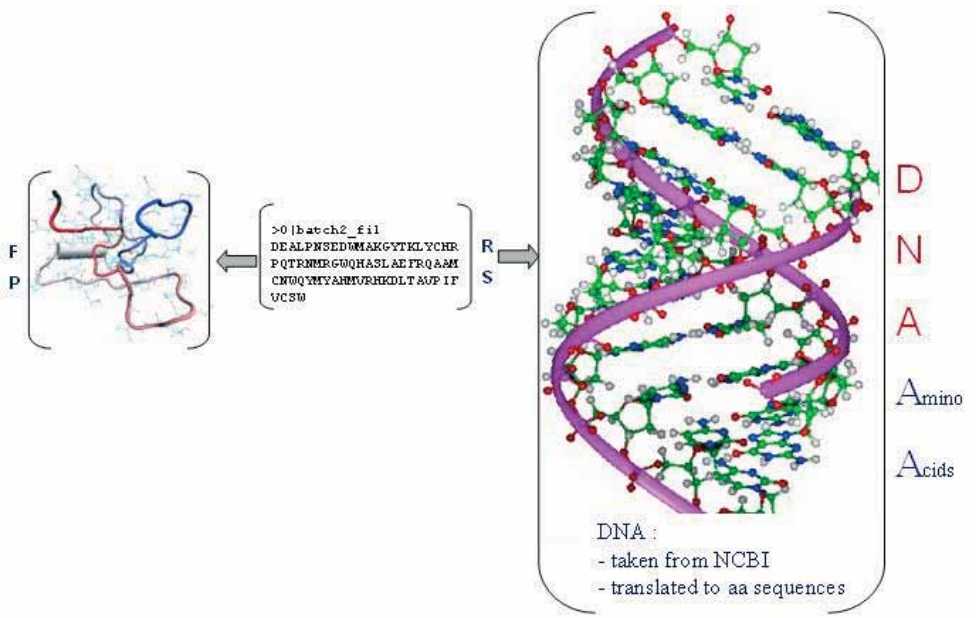


Fig. 1. DNA Amino Acids – DNA AA. DNA AA created on the basis of translation of genome sequences to aminoacid sequences. FP: Folding proteins, RS: Random sequences

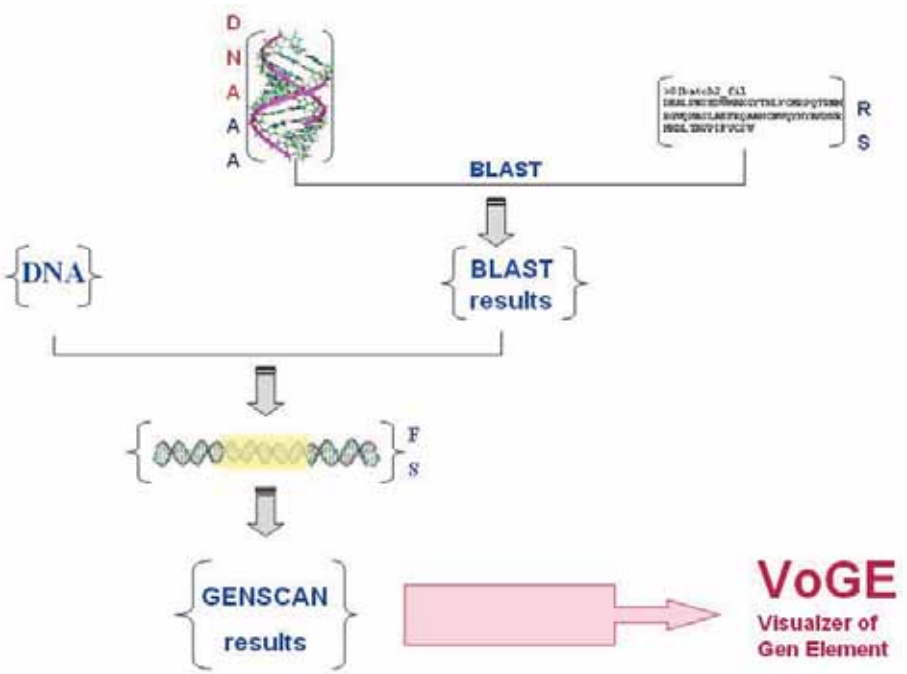


Fig. 2. VoGE input. Data to VoGE application comes form GENSCAN results. DNA AA: DNA Amino Acids (DNA translated to aminoacid sequences), FP: Folded proteins, RS: Random sequences; FS: Flanked Sequences

Fig. 3. Gene structural elements in analyzing nucleotide sequences. Block without description shows coding information of never born protein which traces was found in genome sequences.

Software that ware used:

BAST: Program for searching DNA AA database by RS database was BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) [6].
GENSCAN: Program for identification gene elements – [5]

Technology used for creating application:

C: Scripts for translating nucleotide sequences to aminoacide sequences are created in C programming language.

Java – Java 1.6 was applied for creation graphical interface of VoGE by which user can analyzed data.

BioPython – a package of freely available Python tools for computational molecular biology. It helps parsing the BLAST output files, making it simple and quick.

MySQL – a relational database management system. It is very useful in storing and processing large amounts of data. Outputs of BLAST and GENESCAN are parsed and stored in a relational database called GENOMIC. VoGE connects with database through Java Servlet. All needed data is quickly gathered and calculated thanks to SQL queries. Results from database are sent back to VoGE, which presents them in a simple and accessible way.

Results

Workflow

The VoGE graphical user interface enables user to specify his request by filling special form. While user is waiting for results, the request with values entered by user is sent to the Java Servlet, which is running under Tomcat Server (fig. 4).

Servlet process the user request and builds specific SQL query, which is executed against the database. Results are sent back to the Servlet, where they are checked and formatted, so they can be visualized easily. Eventually result data is sent to VoGE GUI and presented to the user.

VoGE options

On the beginning VoGE give window with name of the experiments that should be choose. Kind of experiment defined which genomes and with parameters were processing. During next step user can explicitly enter the sequence FASTA name (for specifying the sequence, which was used for BLAST searching) and results for this sequence will be presented.

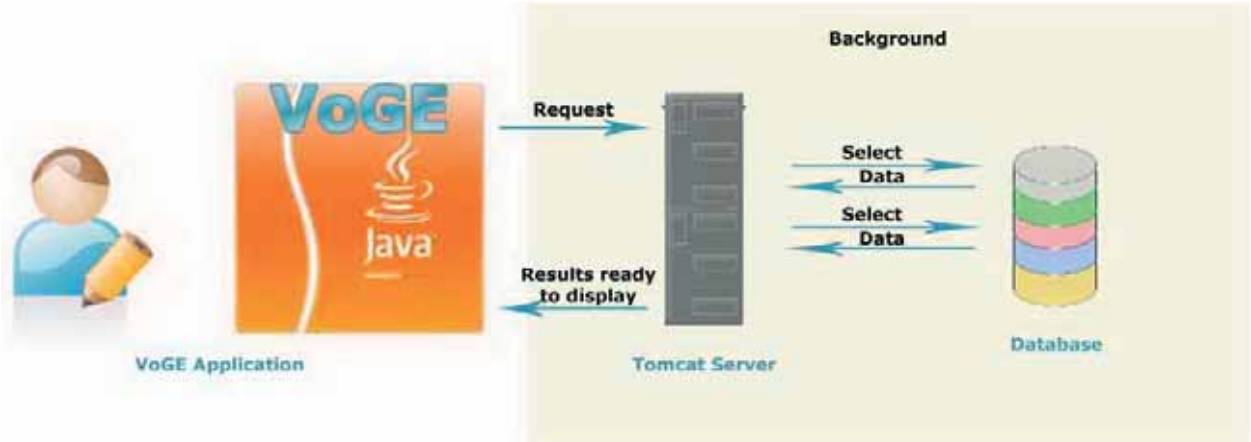


Fig. 4. Application created in Java technology connects with Tomcat server due to request. Java servlet connects to database, selects results and sends it back to user.

Fig. 5. Entry query form for choosing experiment results.

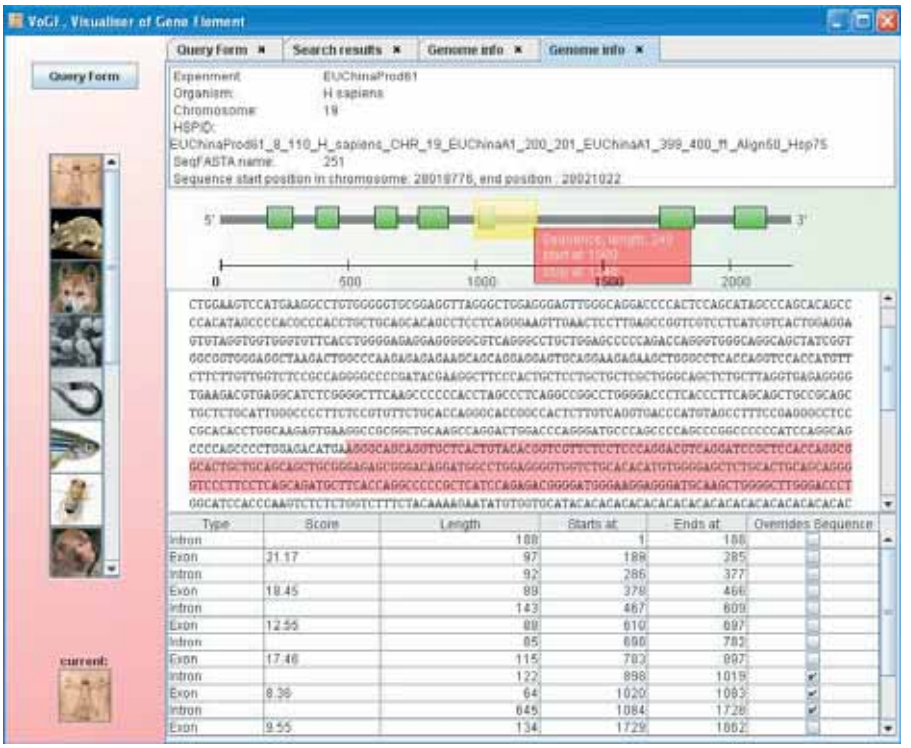


Fig. 6. One of the alignments for particular sequence view.

Fig. 7. Summarized results of experiment view.

The other possibility is to define the organism and result sequences properties (sequence length, exon number, exon score, polyA, etc.) (fig. 5). The list of matching results are returned.

VoGE enable user to see the result alignment sequence properties (fig. 6).

All coordinates of the sequence are in the upper frame. The visualization of gene structure elements are presented below the coordinates frame. In addition the raw sequence fragment and detailed information about gene elements are provided.

Summarized data of experiment for particular genomes and its chromosomes is possible to present by use VoGE (fig. 7). The participation of all gene structure elements in genome and each chromosome can be visualized.

Sequences, pictures and more detailed information can be downloaded in cvs file.

Conclusion

VoGE is application for visualization localization of gene structure elements. Data obtained on the basis of searching nucleotide sequences (especially whole genome material) translated to the protein language with protein sequences are possessed by common known detection method of gene structure elements (incorporated to the Genscan tool).

VoGE helps analyse and visualise stretches of genomic regions encoding NBP that can be important in planning expression of NBP in vivo.

In the future construction of hierarchical tree of similar sequences is planned as well as statistical analysis of obtained results. It is expected revealing groups of sequences which are related and give information about amount sequences in particular cluster, in the same way like in analysis that were done during genome-wide expression patterns analysis.

References

1. Brylinski M., Jurkowski W., Konieczny L., Roterman I.: Limited conformational space for early-stage protein folding simulation, Bioinformatics, 20, 199-205, 2004.
2. Brylinski M., Konieczny L., Roterman I.: Fuzzy-oil-drop hydrophobic force field – a model to represent late-stage folding (in silico) of lysozyme, J Biomol Struct Dyn, 23(5), 519-528, 2006.
3. Chiarabelli C., Vrijbloed J. W., Thomas R. M., Luisi P. L.: Investigation of de novo totally random biosequences. Part I: A general method for in vitro selection of folded domains from a random polypeptide library displayed on phage, Chem Biodivers, 3, 827-839, 2006.
4. Chiarabelli C., Vrijbloed J. W., De Lucrezia D., Thomas R. M., Stano P., Polticelli F., Ottone T., Papa E., Luisi P. L.: Investigation of de novo totally random biosequences. Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display, Chem Biodivers, 3, 840-859, 2006.
5. Burge C. and Karlin S.: Prediction of complete gene structures in human genomic DNA, J Mol Biol, 268, 78-94, 1997.
6. Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J.: Basic local alignment search tool, J Mol Biol, Oct 5; 215(3): 403-10, 1990.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 15–23

UNFOLDING SIMULATION TO VERIFY THE CONCEPT OF LIMITED CONFORMATIONAL SUB-SPACE FOR EARLY-STAGE INTERMEDIATE

[1,2] PIOTR KIEŁKOWICZ, [1] IRENA ROTERMAN

[1] Department of Bioinformatics and Telemedicine, Collegium Medium – Jagiellonian University, Lazarza 16, 31-530 Krakow, POLAND
[2] Faculty of Physics, Astronomy, Applied Computer Science – Jagiellonian University, Reymonta 4, 30-059 Krakow, POLAND

Abstract: Model introducing the limited conformational sub-space for early-stage intermediate definition for protein folding process presented formerly is verified in respect to the unfolding process treated as reverse process to folding. It was expected to receive the step-wise unfolded structure keeping the structural alphabet. It is shown that as long as the secondary structure is present in the gradually unfolded structures, the codes for structural alphabet are changed for relatively low number of residues. The high temperature molecular dynamics simulations revealed the structures with significantly increased distance versus the limited conformational sub-space and large change of alphabet codes. The test was performed for ubiquitine in 300K, 350K, 400K, 500K, 700K and 1000K. It suggests that the structural codes found for crystal structures can not be treated rigorously to be kept during the folding process simulation. Although some tendencies for structural codes changes are observed suggesting the corrections for the definition of early stage structural forms.

Introduction

The multi-step process of protein folding has been recognized experimentally [1-17]. The introduction of the early-stage intermediate seems to be necessary [18]. In consequence the definition of the limited conformational sub-space is expected [18]. In consequence, the multi-step model to simulate the protein folding process was introduced [19]. The early-stage [ES] intermediate was defined according to the backbone conformations with side chain-side chain interaction excluded. The backbone conformation described by two geometric parameters: V-angle and R-radius of curvature which appeared to be dependent on V-angle in form of second degree polynomial. The structures satisfying this equation – assumed to represent the relaxed backbone conformations – revealed the limited conformational sub-space as the fragment of Ramachandran map. This space appeared to be of ellipse path linking all secondary structural forms (α_R-helix, β-structural area and α_L-helix). Additionally it was proved that the amount of information carried by amino acid sequence appeared to be balanced with the amount of information necessary to define the structure of ES intermediate [19].

The calculations represented in this paper were aimed to check whether the unfolding process (molecular dynamics simulation of ubiquitine in different temperatures: 300, 350, 400, 500, 700 and 1 000 K) indicates the approach toward the limited conformational sub-space assumed to represent the early-stage intermediate conformations.

Materials and Methods

Data: the protein ubiquitine (PDB – 1UBQ) was taken as the example. The protein without the SS-bonds and structurally differentiated (α-helix – 23%, β-structure – 34% and RC – 43%) and of medium size is the very convenient object for simulation procedure. The secondary structure characteristics was performed according to the procedure available on the PDB webpage [20].

The structural parameters

The solvent accessible area for post-dynamics structures was calculated according to program Surface Racer 5.0 [21].

The number of non-bonding contacts was calculated according to the program prepared specially for the project. Program was written according to standards (parameters) given in [22].

The structural codes

The Phi, Psi angles distributed on the Ramachandran map as they appear in protein moved (according to shortest distance criterion) toward the ellipse path reveal the probability distribution characterized by seven maxima. Some of them represent the secondary structure elements: Code C – α_R-helix, E,F – β-structural forms, G – α_L-helix. Others (A, B, D) represent the structures belonging to Random Coil. The detailed presentation of the structural codes according the ellipse shaped limited conformational sub-space is given in [23].

The changes of structural codes were estimated according to the structural codes as observed in all post-dynamics structures.

Molecular dynamics simulation:

The AMBER program was used to simulate the molecular dynamics. The following temperatures were applied: 300 K, 350 K, 400 K, 500 K, 700 K, 1000 K.

The implicit solvent model for the ff03 force field applying the standard parameterization for 0.1 M solution with the rigid positions of hydrogen atoms was used for simulations. The cut-off distance was taken to be 16 A. Simulations was performed in three runs: equilibration until the stabilization of the energy was achieved; the heating process – 50 000 iterations with the time step 0.002 ps (the total time range 100 ps). The output data (structure and parameters) was collected every 1 ps.

The starting (equilibration) simulation was continued until the stability of temperature and energy was achieved.

The heating dynamics was performer for the appropriate temperature until the stabilization characteristics for the temperature was achieved (about 50 000 steps).

The 500 000 steps of effective simulation with the stable temperature was performed with the time step 0.002 ps, which was common for all simulations.

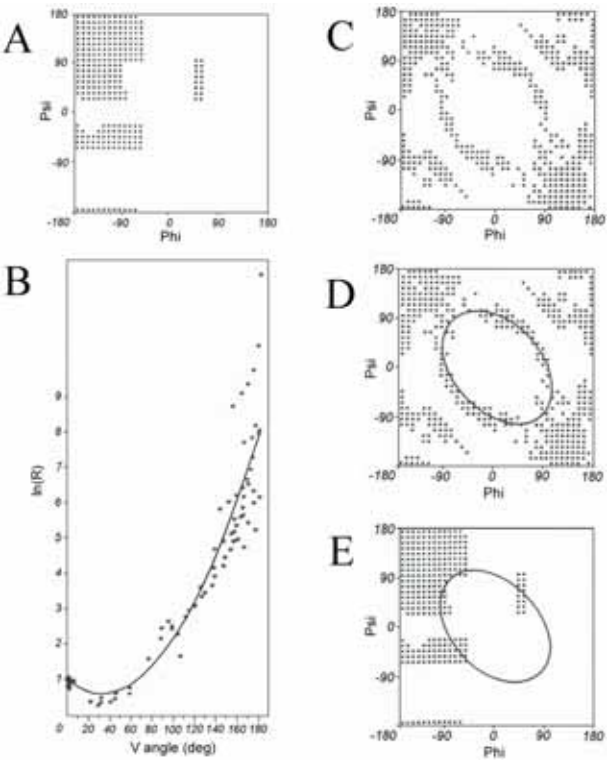


Fig. 1. The ellipse path definition: A – the low energy areas on Ramachandran map, B – the relation between V-angle (dihedral angle between two sequential peptide bond planes) and R – resultant radius of curvature (in log scale). The structures satisfying the approximation function are distributed as shown in C. The approximation to the ellipse path (D). The relation of ellipse path (early-stage conformational sub-space) in relation to low energy areas on Ramachandran map (E).

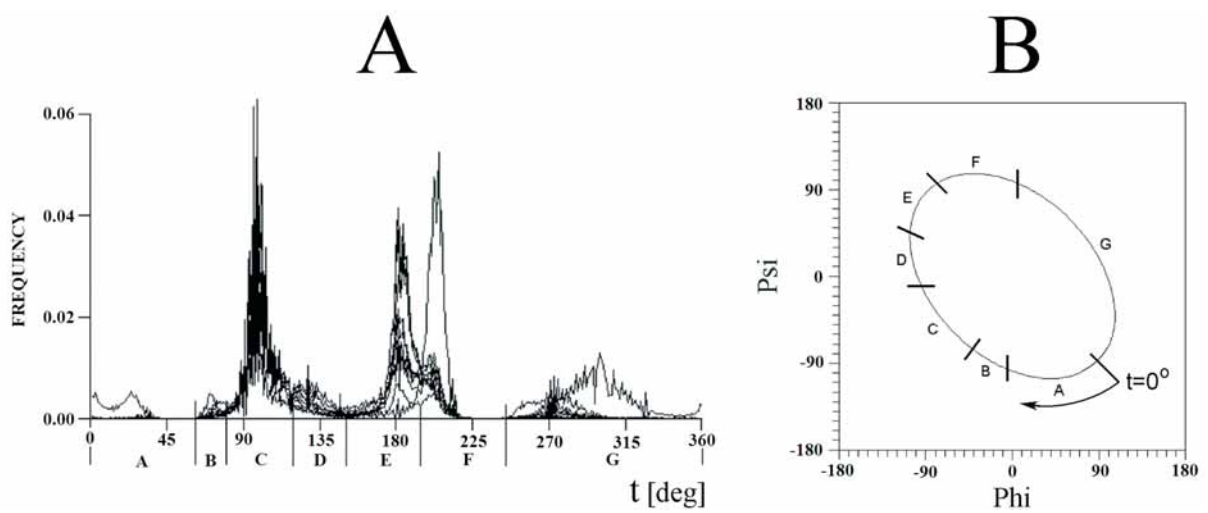


Fig. 2. The probability distribution A - along the ellipse path after moving all Phi, Psi angles (found in proteins) toward the ellipse path. The letter code is attributed to each local maximum. B – the starting point and the direction of the walk along the ellipse is shown.

Fig. 2A

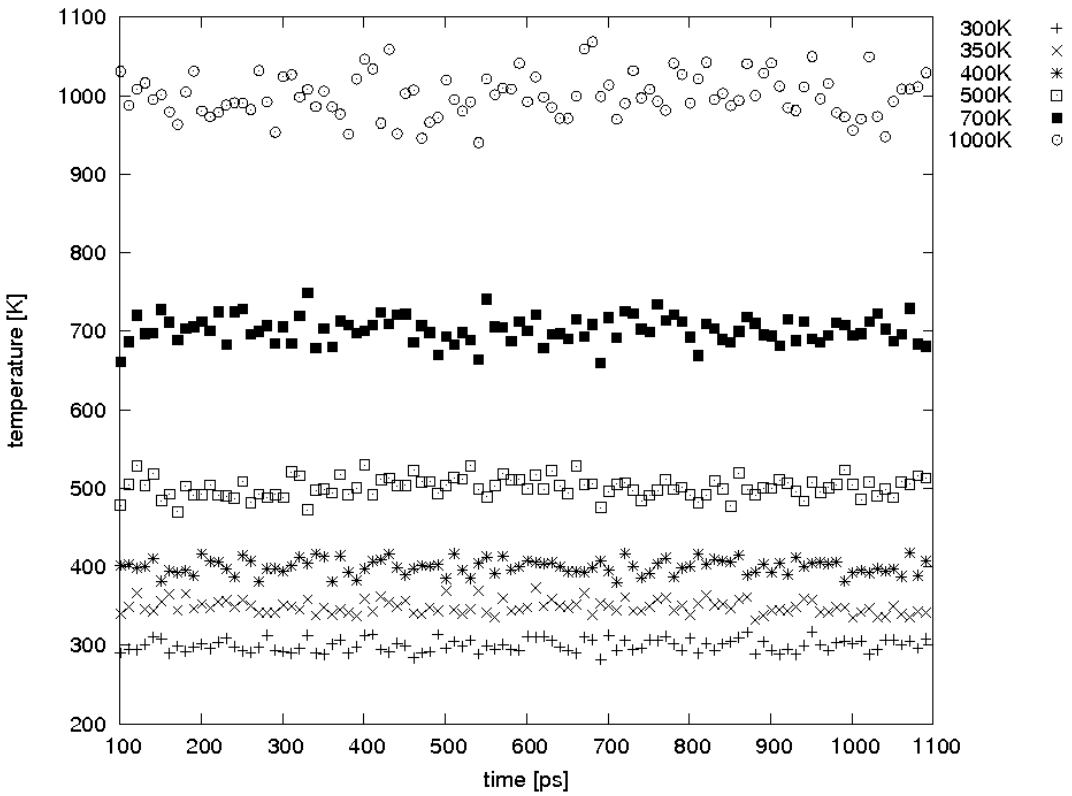


Fig. 2B

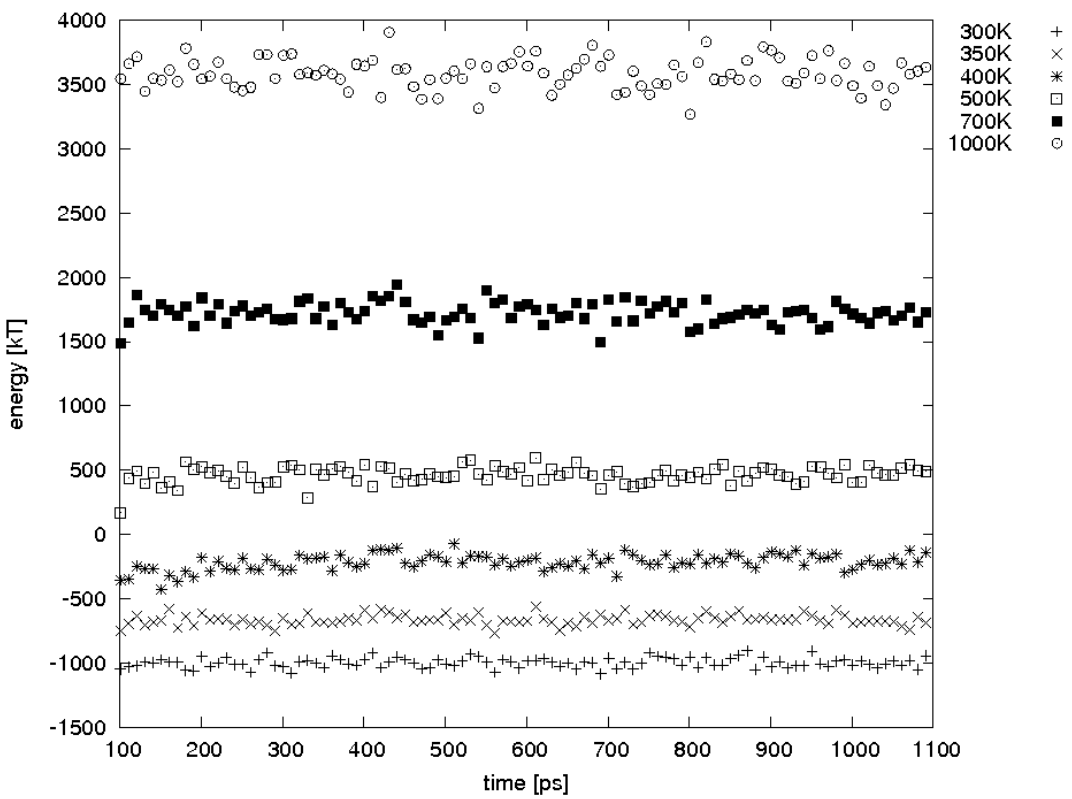


Fig. 3. The stability of the dynamics process. A –stability of temperatures for particular processed and B – stability of total energy of the molecule – ubiquitin

The early-stage conformational sub-space

The ellipse path assumed to represent the early-stage related conformational sub-space is shown in Fig. 1. The ellipse-path was defined according to the analysis of geometric parameters describing the backbone conformation. The main assumption is that all structural forms can be treated as helices including β-structure. These two parameters are as follows: V- angle between two sequential peptide bond planes and the resultant radius of curvature which for V angles close to zero takes the low values and for values close to 180 degs the radius becomes very large (this is why the logarithmic scale was introduced) (Fig. 1. B). The β-structure is treated as helix with large value of radius of curvature. The structures generated according to the approximated function are shown in Fig. 1 C. The ellipse path can be distinguished as the well ordered part of the distribution (Fig. 1.D). The relation of ellipse path to the low energy areas on the Ramachandran map (Fig. 1.A.) are shown in Fig. 1.E.

The post-dynamics structures were characterized by Phi, Psi angles distribution and the approach toward the ellipse path was calculated in particular.

The equation for ellipse path can be expressed as follows:

$$\Phi = -A\cos(t) - B\sin(t)$$
$$\Psi = A\cos(t) - B\sin(t) \qquad \text{eq (1)}$$

Where t expresses the angular movement along the ellipse starting from the lower right quarter of the Ramachandran map.

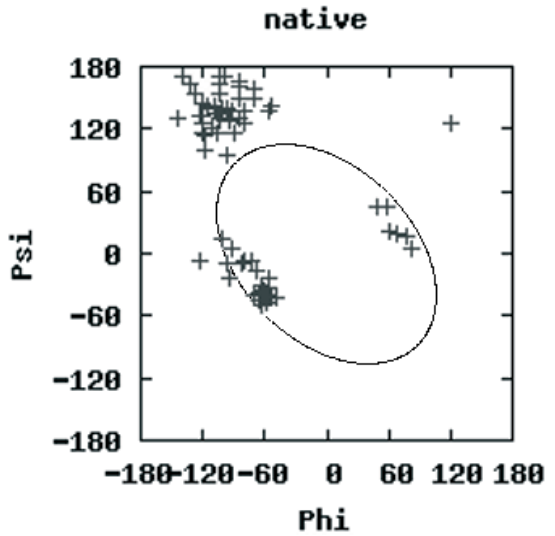


Fig. 4. The distribution of Phi, Psi angles for native form of ubiquitine in relation to the ellipse path.

Structural alphabet

The probability distribution along the ellipse path after moving the Phi, Psi angles toward ellipse path is shown in Fig. 2. The presence of seven maxima with attributed letter codes may be distinguished producing the “structural alphabet”.

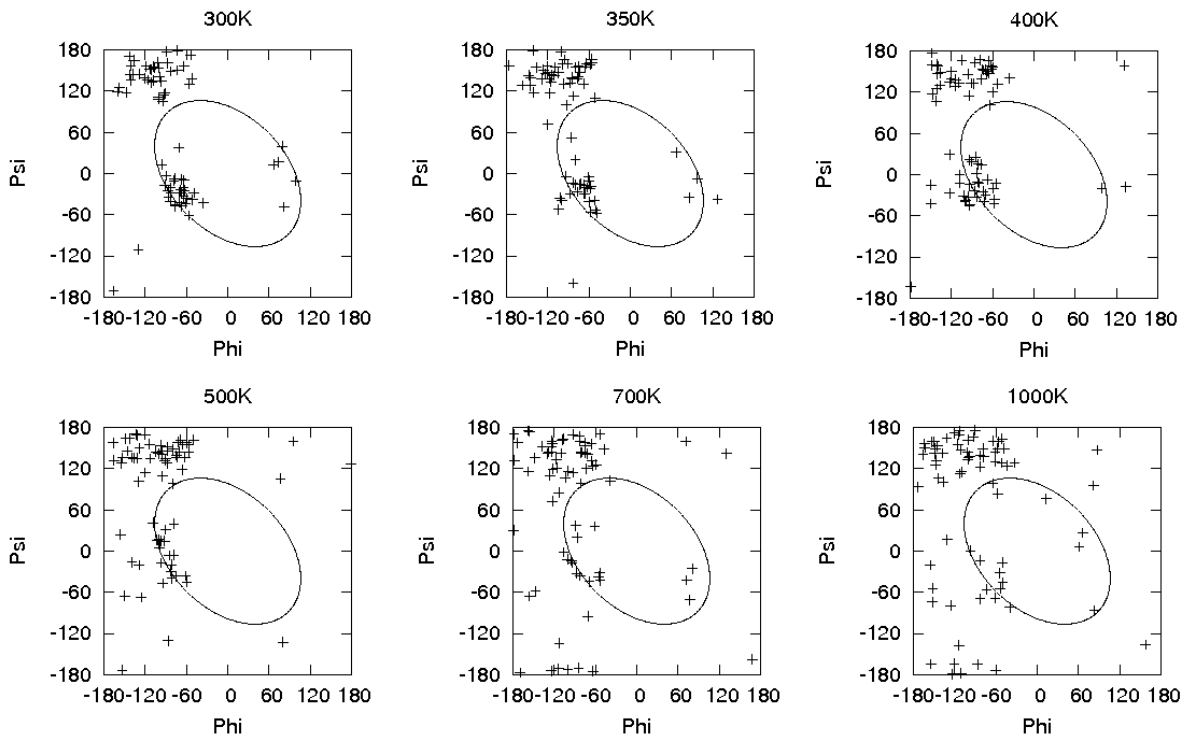


Fig. 5. The Phi, Psi angles distribution for step-wise unfolding of ubiquitine in respect to ellipse path.

Tab. 1. The distance (in degs units) for step-wise unfolding and for crystal structure. The percentage of changed structural codes during unfolding with the number of stable alphabet codes is given in two columns on right. The last column on right presents the decreasing number of side chain-side chain interactions

TEMPERATURE	DISTANCE (angular scale)	Standard deviation of distance	Percentage of stable structural alphabet codes	Number of structural alphabet codes changed	Number of side chain – side chain interactions
300 K	39.44	31.59	63.51%	27	211
350 K	41.18	26.76	60.81%	29	176
400 K	40.70	30.22	41.89%	43	79
500 K	46.81	32.24	36.49%	47	52
700 K	54.91	37.05	33.78%	49	45
1000 K	56.12	33.76	33.78%	49	37
NATIVE – crystal	35.45	23.48	-	-	247

Results:

The molecular dynamics simulation

The stabilization of energy and temperature is shown in Fig. 3.A and Fig. 3.B. respectively. The given results prove the stabilization of the protein molecule in dynamics process for the period of time 100-1100 ps. These structures averaged are taken for the analysis of the unfolding process of ubiquitine.

The Phi, Psi re-localization for all structural forms (including crystal structure) was characterized by the parameter measuring the distance versus the nearest point belonging to the ellipse path (shortest distance criterion).

The distribution of Phi, Psi angles for crystal structure is given in Fig. 4 and for post-dynamics structures in Fig. 5.

The significantly larger dispersion of the points representing the Phi, Psi angles can be seen in respect to the crystal structure Phi, Psi angle distribution. The gradual disappearance of points in the regions of secondary structure related positions can be observed.

The relative disappearance of secondary structure can be seen also in Fig. 6 A and Fig. 6.B. The post-dynamics structures above 450 K do not represent any form of secondary structure although the low representation of secondary structure is seen also in 400 K post-dynamics structure. The qualitative measurements of these changes are given in Tab. 1. The mean distance between Phi, Psi points and nearest point belonging to ellipse seems to be relatively stable for structures below 500 K

The analysis of Phi, Psi angles distribution suggests that no approach toward the ellipse was observed during unfolding procedure. The mean distance between the Phi, Psi angle observed and the nearest point belonging to ellipse path seems to be almost constant for temperatures below 450 K. Te stability of the distribution versus the ellipse path for this range of temperatures seems to be related to the range of temperature close to physiological temperatures. The stability of mean distance (and standard deviation values) suggests the rearrangement of the structure without any significant changes using the Phi, Psi angles as criterion for structural changes.

The post-dynamic structure at 1000 K obviously representing the limit case in extremely high temperature is used just for comparison with the highest distance versus the ellipse-path,

Tab. 2. The changes of structural codes according to the crystal structure. On the left: number of residues disappearing from particular region; on the right: number of total residues moving from the region under consideration.

STRUCTURAL CODES–TEMP	A	B	C	D	E	F	G
A 300 K	0/0	0/0	0/0	0/0	0/0	0/0	0/0
350 K	0/0	0/0	0/0	0/0	0/0	0/0	0/0
400 K	0/0	0/0	0/0	0/0	0/0	0/0	0/0
500 K	0/0	0/0	0/0	0/0	0/0	0/0	0/0
700 K	0/0	0/0	0/0	0/0	0/0	0/0	0/0
1 000 K	0/0	0/0	0/0	0/0	0/0	0/0	0/0
B 300 K	0/2	0/2	2/2	0/2	0/2	0/2	0/2
350 K	0/2	0/2	1/2	0/2	0/2	1/2	0/2
400 K	0/2	0/2	2/2	0/2	0/2	0/2	0/2
500 K	0/2	0/2	0/2	1/2	1/2	0/2	0/2
700 K	0/2	0/2	1/2	1/2	0/2	0/2	0/2
1 000 K	0/2	0/2	2/2	0/2	1/2	1/2	0/2
C 300 K	0/20	1/20	17/20	2/20	0/20	0/20	0/20
350 K	0/20	3/20	14/20	2/20	0/20	0/20	1/20
400 K	0/20	1/20	7/20	2/20	5/20	4/20	1/20
500 K	0/20	2/20	6/20	2/20	7/20	2/20	1/20
700 K	0/20	1/20	5/20	2/20	5/20	6/20	1/20
1 000 K	0/20	3/20	6/20	0/20	4/20	4/20	3/20
D 300 K	0/2	0/2	2/2	0/2	0/2	0/2	0/2
350 K	0/2	0/2	1/2	0/2	1/2	0/2	0/2
400 K	0/2	0/2	1/2	1/2	0/2	0/2	0/2
500 K	0/2	0/2	0/2	0/2	1/2	1/2	0/2
700 K	0/2	1/2	0/2	0/2	1/2	0/2	0/2
1 000 K	0/2	0/2	0/2	0/2	1/2	1/2	0/2
E 300 K	0/32	1/32	2/32	1/32	22/32	6/32	0/32
350 K	0/32	0/32	1/32	0/32	22/32	9/32	0/32
400 K	0/32	0/32	7/32	3/32	15/32	7/32	0/32
500 K	0/32	1/32	6/32	3/32	16/32	6/32	0/32
700 K	0/32	6/32	3/32	1/32	15/32	6/32	1/32
1 000 K	0/32	5/32	4/32	1/32	14/32	5/32	3/32
F 300 K	0/10	2/10	0/10	0/10	5/10	3/10	0/10
350 K	0/10	0/10	1/10	0.10	3/10	6/10	0/10
400 K	0/10	0/10	2/10	1/10	1/10	6/10	0/10
500 K	0/10	0/10	0/10	3/10	2/10	4/10	1/10
700 K	0/10	0/10	2/10	0/10	5/10	2/10	1/10
1 000 K	0/10	2/10	0/10	0/10	5/10	3/10	0/10
G 300 K	0/8	0/8	2/8	0/8	1/8	0/8	5/8
350 K	0/8	1/8	1/8	0/8	2/8	1/8	3/8
400 K	0/8	0/8	3/8	0/8	2/8	1/8	2/8
500 K	1/8	1/8	1/8	1/8	2/8	1/8	1/8
700 K	0/8	1/8	0/8	0/8	1/8	3/8	3/8
1 000 K	0/8	1/8	1/8	0/8	3/8	1/8	2/8

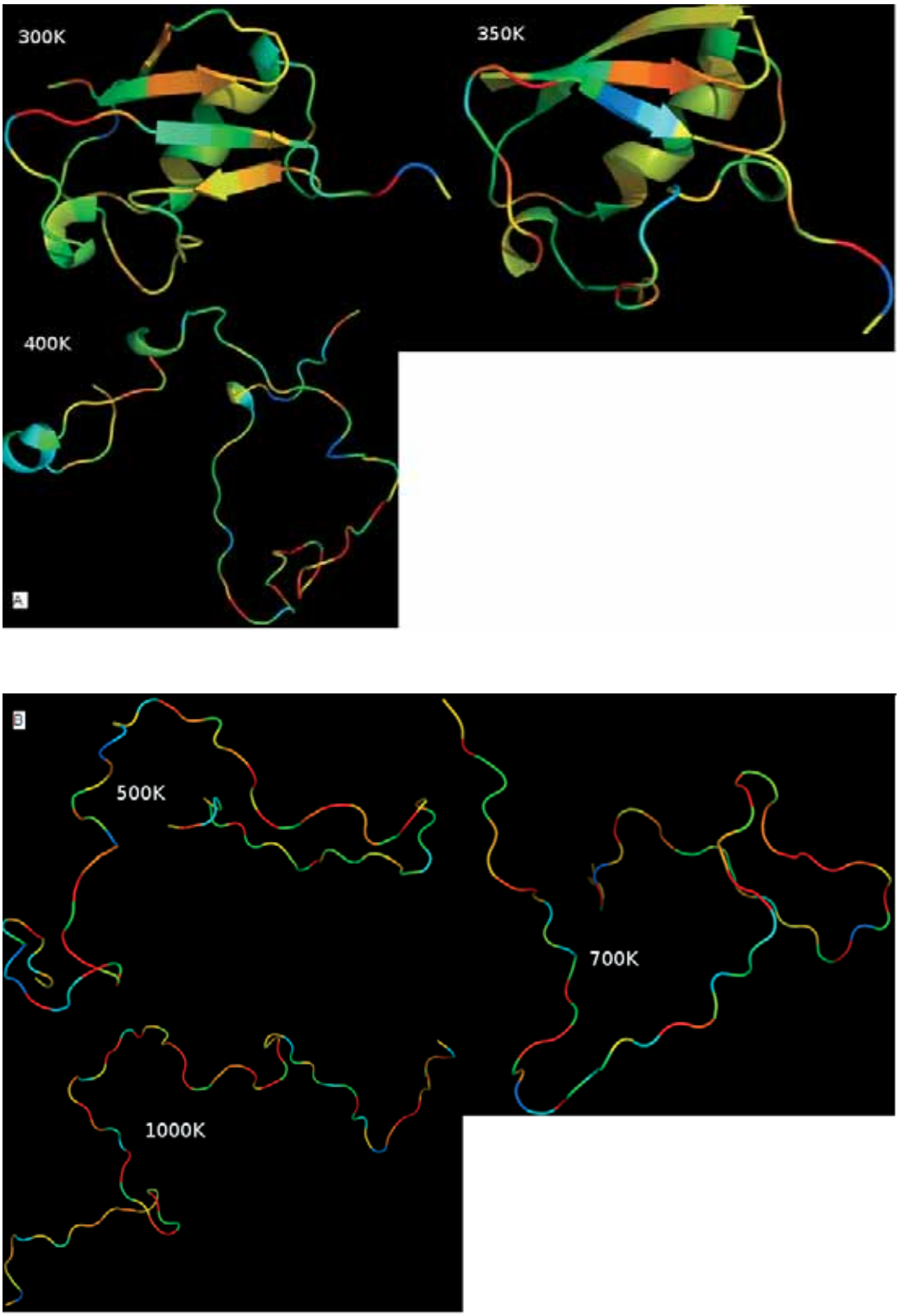


Fig. 6. The 3-D presentation of the post-dynamic structures of ubiquitine. The color scale applied shows the scale of Phi, Psi distances. The red color indicates the increase of distance (higher than 60 degs), the blue color – decrease of distance (higher than 60 deg), green/yellow – more or less stable (in between).
A – structures with secondary structure preserved
B – structures secondary structure lost

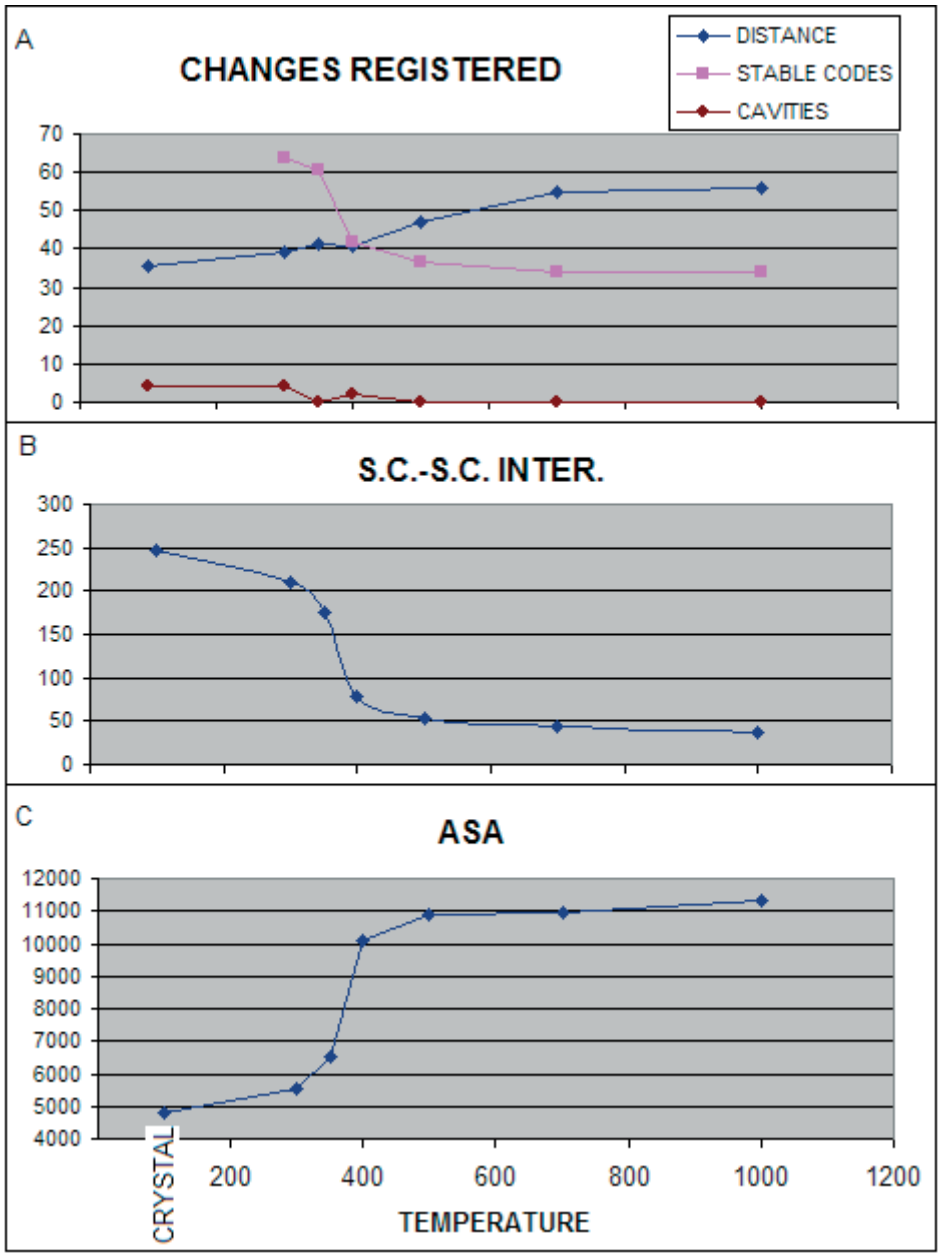


Fig. 7. The changes in the post-dynamics structures (increasing temperatures) displayed versus the crystal structure:
A – distance versus ellipse, # of stable structural codes (%), the number of cavities recognized according to own program prepared on the basis of [21].
B – # of side chain-side chain contacts
C – solvent accessible area (according to Surface Racer [21]

showing to what extent this distance may increase and what range of dispersion of distance values is possible for the protein under consideration.

The post-dynamics structures (300, 350, 400 K) represent the changes occurring slowly. The 500 K post-dynamics structure reveals significant changes in all parameters (Fig. 7).

The most important parameter, which is the distance between Phi, Psi angles versus the ellipse path is accordant with the changes of all other parameters. The interesting is the decrease of number of cavities for the post-dynamics structures above 350 K.

The structural alphabet

The distribution of Phi, Psi angles along the ellipse path received by the moving the observed Phi, Psi angles toward the ellipse path (the shortest distance criterion) revealed the presence of seven maxima [23]. Each of them attributed by the letter [from A to G] allowed the construction of the structural alphabet for ES intermediate. The comparison of post-dynamics (step-wise unfolding] structures of ubiquitine classified according the structural alphabet revealed the fragments of polypeptide with different structural stabilization. The fragments of high

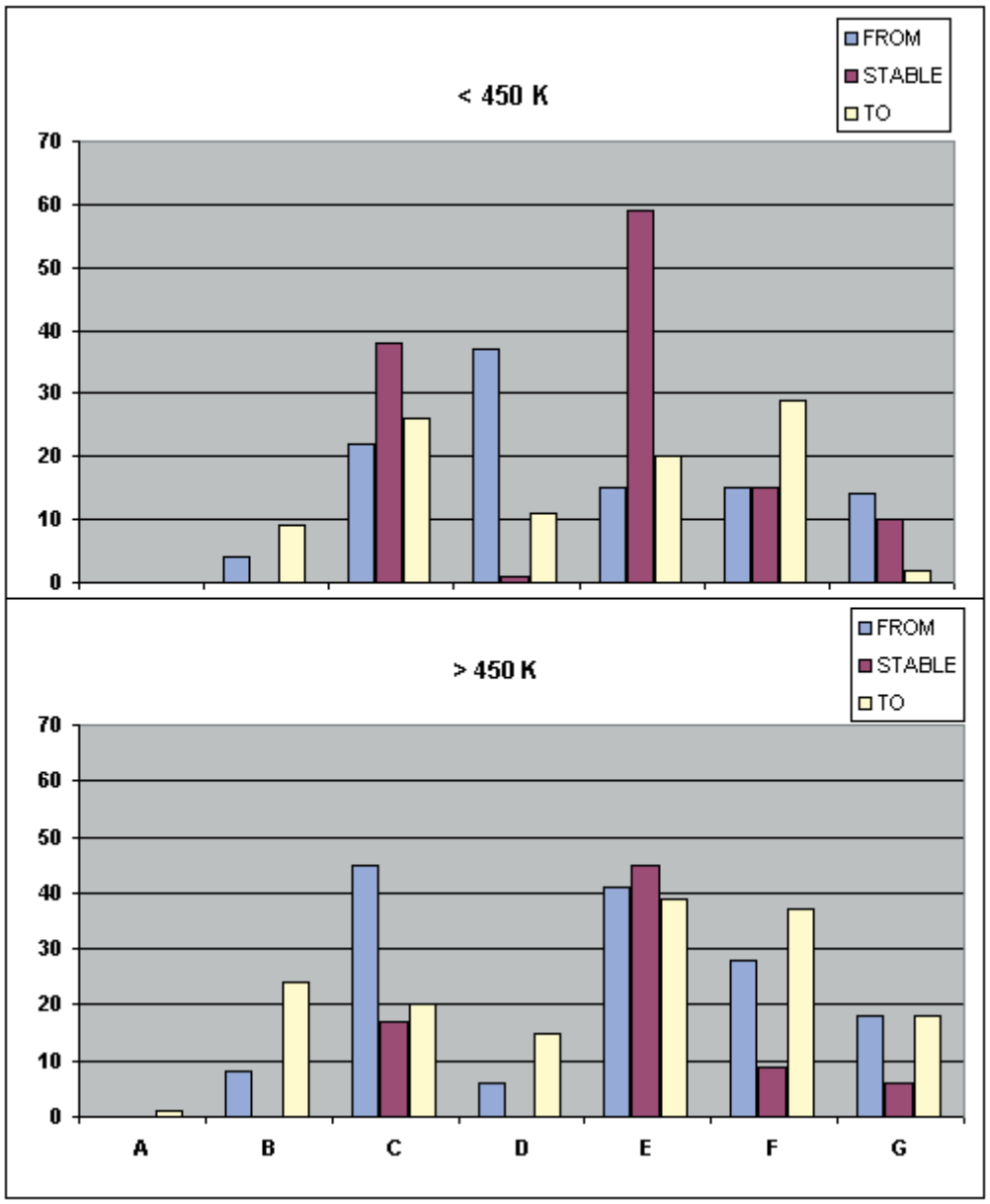


Fig. 8. The stability versus mobility of the codes. „Stable" denotes number of letter codes not changed during dynamics, ("TO" denotes the target letter code, "FROM" denotes the original letter code. The bars are shown for lower temperatures (<450 K) and upper (> 450 K).

and low stability (the minimum and maximum of changed letters per pentatpetide) are listed in Tab. 2.

The observed structural codes changes in post-dynamics structures are given in Tab. 2. and in Fig.8.

The analysis of Tab. 2. and Fig. 8. suggests that the melting of helix takes place in upper temperature. The stability of C (helix) and E (β-structure) seems to be the highest in relation to other codes. No stable B code and very low stability of D code were observed.

Generally there is no significant tendency to regular pattern of structural changes categorized according to letter codes.

Conclusions:

The simulation of molecular dynamics was performed to mimic the process of unfolding, which was treated as the reverse process to folding. The structures: crystal one and all post-dynamics ones were classified according to letter codes distinguishing the particular Ramachandran areas (final structures) and particular ellipse fragments (early stage structures). It was expected that some of structural codes appear stable. According to the results shown in Tab. 2. and Fig. 8. there is no recognizable patter for structural codes changes. The codes C and E seem to be the most stable. The Phi, Psi angles of molten helix move toward the E / F zone and toward the B zone (which is the neighbor structure in the sense of ellipse localization).

The post-dynamic structures are planned to be used as the starting forms for folding process (energy minimization procedure) to reveal whether any of these structures are able to find the proper way to recreate its native structural form.

The similar simulation was performed for immunoglobulin molecule. The approach toward the ellipse was observed there [24]. Immunoglobulin is highly β-structural protein. The

movement of Phi, Psi maximum of concentration on the Ramachandran map was observed there to identify the approach of β-structural maxima toward the α-helix Phi, Psi area on the Ramachandran map. The ubiquitine was selected this time to analyze the behavior of mixed secondary structure protein. The gradual increase of the mean distance between ellipse and ellipse path is observed accordant to the increased temperature although the structures for temperatures below 450 K seem to be change slowly and above 450 K this difference was found to be larger.

The ellipse path and structural alphabet was used for protein folding simulation using the ellipse path belonging Phi, Psi angles to define the starting structure for folding simulation [25-27]. The results of unfolding simulation of BPTI was presented and discussed in [28,29]. The movement of Phi, Psi angles in that paper was following the ellipse path moving toward the left-handed helix in the highest temperature simulation although visual analysis of these results suggests also the increase of Phi, Psi distances versus the ellipse path for all analyzed temperatures.

The general tendency for structural alphabet changes during unfolding process will be taken under consideration for larger number of proteins.

References:

1. Creighton T. E. (1977) Conformational restrictions on the pathway of folding and unfolding of the pancreatic trypsin inhibitor. J. Molec. Biol. 113, 275-293.
2. Creighton T. E. & Goldenberg D. P. (1984) Kinetic role of a meta-stable native-like two-disulphide species in the folding transition of bovine pancreatic trypsin inhibitor. J. Molec. Biol. 179, 497-526.
3. Creighton T. E. (1978) Experimental studies of protein folding and unfolding. Prog. Biophys. Molec. Biol. 33, 231-297.
4. Creighton T. E. (1974) The single-disulphide intermediates in the refolding of reduced pancreatic trypsin inhibitor. J. Molec. Biol. 87, 603-624.
5. States D. J., Creighton T. E., Dobson C. M. & Karplus M. (1987) Conformations of intermediates in the folding of the pancreatic trypsin inhibitor. J. Molec. Biol. 195, 731-739.
6. States D. J., Dobson C. M., Karplus M. & Creighton T. E. (1984) A new two-disulphide intermediate in the refolding of reduced bovine pancreatic trypsin inhibitor. J. Molec. Biol. 174, 411-418.
7. Creighton T. E. (1980) Experimental elucidation of pathways of protein unfolding and refolding. In: Protein Folding (ed. Jaenicke R.) 427-446. Elsevier/North-Holland Biomedical, Amsterdam.
8. Creighton T. E. (1977) Effects of urea and guanidine-HCl on the folding and unfolding of pancreatic trypsin inhibitor. J. Molec. Biol. 113, 313-328.
9. Creighton T. E. (1980) Role of the environment in the refolding of reduced pancreatic trypsin inhibitor. J. Molec. Biol. 144, 521-550.
10. Creighton T. E. (1974) The single-disulphide intermediates in the refolding of reduced pancreatic trypsin inhibitor. J. Molec. Biol. 87, 603-624.
11. Creighton T. E., Kalef E. & Arnon R. (1978) Immunochemical analysis of the conformational properties of intermediates trapped in the folding and unfolding of bovine pancreatic trypsin inhibitor. J. Molec. Biol. 123, 129-147.
12. Creighton T. E. (1875) Reactivities of the cysteine residues of the reduced pancreatic trypsin inhibitor. J. Molec. Biol. 96, 777-78.
13. Creighton T. E. (1985) The problem of how and why proteins adopt folded conformations. J. Phys. Chem. 89, 2452-2459.
14. Kosen P. A., Creighton T. E. & Blout E. R. (1983) Circular dichroism spectroscopy of the intermediates that precede the rate-limiting step of the refolding pathway of bovine pancreatic trypsin inhibitor. Relationship of conformation and the refolding pathway. Biochemistry 22, 2433-2440.
15. Goldenberg D. P. & Creighton T. E. (1985) Energetics of protein structure and folding. Biopolymers 24, 167-182.
16. Creighton T. E. (1978) Refolding of bovine pancreatic trypsin inhibitor modified at methionine-52. J. Molec. Biol. 119, 507-518.
17. Hollecker M. & Creighton T. E. (1983) Evolutionary conservation and variation of protein folding pathways. Two protease inhibitor homologues from black mamba venom. J. Molec. Biol. 168, 409-437.
18. Alonso D. O., Daggett V. (1998) Molecular dynamics simulations of hydrophobic collapse of ubiquitin. Protein Sci. 7, 860-874.
19. Jurkowski W., Brylinski M., Konieczny L., Wiiniowski Z., Roterman I. (2004) Conformational subspace in simulation of early-stage protein folding. Proteins 55, 115-127.
20. http://www.rcsb.org/pdb/explore/remediatedSequence.do?structureId=1UBQ].
21. http://apps.phar.umich.edu/tsodikovlab/index_files/Page756.htm
22. http://www.cs.rutgers.edu/pub/seredin/DomainRewievEng.doc
23. Brylinski M., Konieczny L., Czerwonko P., Jurkowski W., Roterman I. (2005) Early-Stage Folding in Proteins (In Silico) Sequence-to-Structure Relation. J. Biomed. Biotechnol. 2, 65-79.
24. Roterman I., Konieczny L. (1995) Geometrical analysis of structural changes in immunoglobulin domains' transition from native to molten state. Comput. Chem. 19, 247-252.
25. Jurkowski W., Brylinski M., Konieczny L., Roterman I. (2004) Lysozyme folded in silico according to the limited conformational sub-space. J. Biomol. Struct. Dyn. 22, 149-158.
26. Roterman I. (1995) Modelling the optimal simulation path in the peptide chain folding-studies based on geometry of alanine heptapeptide. J. Theor. Biol. 177, 283-288.
27. Roterman I. (1995) The geometrical analysis of peptide backbone structure and its local deformations. Biochimie 77, 204-216.
28. Daggett V., Levitt M. (1993) Protein unfolding pathways explored through molecular dynamics simulations. J. Molec. Biol. 232, 600-619.
29. Daggett V., Levitt M. (1992) Molecular dynamics simulations of helix denaturation. J. Molec. Biol. 223, 1121-1138.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 25–30

COMPUTATIONAL ANALYSIS OF PROSTATE PERFUSION IMAGES – A PRELIMINARY REPORT

[1] JACEK ŚMIETAŃSKI, [2] RYSZARD TADEUSIEWICZ

[1] Institute of Computer Science
Jagiellonian University, ul. Łojasiewicza 6, 30-438 Kraków, e-mail: jacek.smietanski@ii.uj.edu.pl
[2] Department of Automatics
AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, e-mail: rtad@agh.edu.pl

Abstract: Currently used diagnostic procedures for identification of the prostate cancer (PCa) are insufficient. It occurs quite often that the existing PCa cannot be detected. Therefore scientists search for other methods enabling a better efficacy of diagnosis. The perfusion computed tomography technique (p-CT), which measures some parameters of blood flow within diagnosed organs, is supposed to avoid such problems, even in particularly hard cases.

In this paper some methods of automatic analysis of prostate perfusion tomographic images are presented and discussed. Although the work concentrates only on one image derived from one patient, we can see the complexity and importance of the task. The proposed algorithms and methods based on the Haralick's co-occurrence matrices seems to be the appropriate technique to point out the cancerous lesions.

In the further work described algorithms will be tested on a large set of patients. This goal needs close cooperation between radiologists, pathologist, computer scientists and engineers. The final goal is to develop a professional diagnostic system used in computer aided prostate diagnosis.

Introduction

Prostate Cancer (PCa) is the second most popular men's cancer in Poland and the most popular in West Europe and in the USA. In 2006 in Poland there was 7154 new registered cases and 3861 deaths from the PCa [1, 2]. Such high mortality percentage is because that malignancy is often diagnosed too late. Meanwhile the PCa detected in early stage can be successfully treated and increase the lifetime of patients or even lead to cure. This fact causes that the regular comprehensive diagnosis is very important.. There are many methods which could help to detect the tumor in early stage, for example the DRA examination, PSA measure, transrectal ultrasound, and biopsy (fig. 1) [3, 4, 5]. However the sensitivity and specificity of those methods are dissatisfying. Conventional computed tomography (CT) can help only with detection of metastasis in advanced PCa. In view of this the need of other, more effective method, is obvious.

It is supposed that the effectiveness of detecting early PCa can be improved using the perfusion computed tomography (p-CT) method. In this technique some parameters of blood flow within diagnosed organs are measured. The patient has injected the bolus and repeated scans of the minor pelvis using the multislice CT scanner. One of the measured parameters is blood flow (BF) (fig. 2).

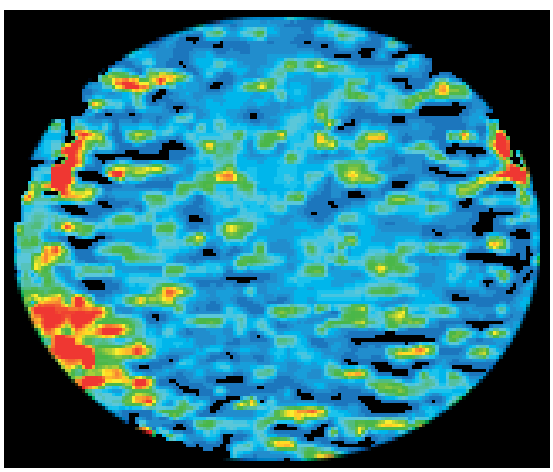


Fig. 2. Analyzed perfusion prostate image – blood flow (BF).

The presumption that this method can be helpful in detecting early PCa is based on the documented fact that the growth of cancer needs many nutrients. To secure their supply to cancerous lesion, new blood vessels are created in this area [6]. It is supposed that this effect (named angiogenesis) is visible on the p-CT prostate images.

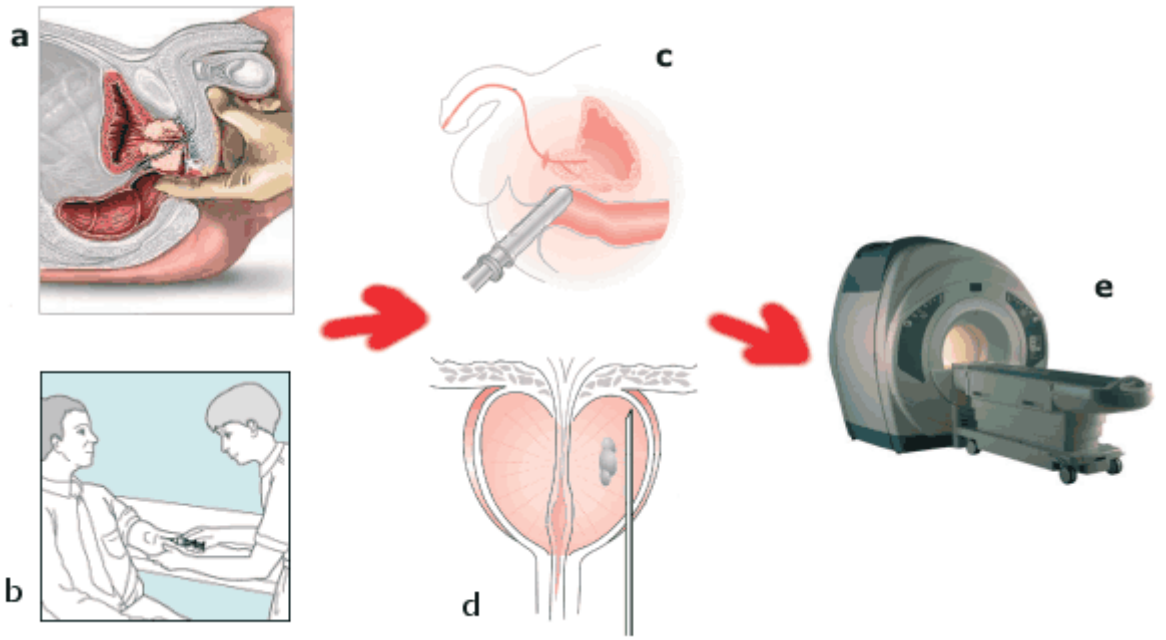


Fig. 1. PCa diagnosis: a) per rectum examination (DRE); b) blood examination (PSA measure); c) transrectal ultrasound (TRUS); d) biopsy; e) additional radiology diagnostic.

Materials and methods

The p-CT examination was held in the Cracow branch of the Oncology Center for a 61 years old patient with suspected PCa (the PSA level 13,60). The p-CT scans were started about 10 s after administration of 50 ml non-ionic contrast medium (370 mgI/ml) at the rate 5 ml/s, and lasted 50 s. The parametric map was drawn using the *CT Perfusion 3 application on the Advantage Workstation. The received image is shown on the fig. 2. The task is to automatically point out the suspicious regions, with the existence of PCa.*

The image analysis was held using different methods and algorithms from the field of image processing and pattern recognition. The most of our algorithms, tested in this work, were based on the Haralick's co-occurrence matrices (GLCM) [7] (fig. 3) and 21 coefficients derived from them (tab. 1) [8].

Let $I : Z^2 \supset D \rightarrow G = \{1, ..., N_g\}$ (where Z denotes set of integers) be a two-dimensional discrete image with N_g gray levels. For the given image I we define the GLCM:

$$P(i, j \mid d, \theta) = \frac{\#\{k, l \in D : I(k) = i, I(l) = j, \mid k - l \mid = d, \angle(k - l) = \theta\}}{\#\{m, n \in D : \mid m - n \mid = d, \angle(m - n) = \theta\}} \quad (1)$$

where: i,j ∈ G – gray levels of points k and l, respectively; $\angle(k-l)$ - the angle between vector kl and axe 0X; d – distance between k and l; θ – direction of co-occurrence, #X – power (number of elements) of set X.

Notation used in the table:

$$\mu_x = \sum_i i \sum_j P(i, j), \quad \mu_y = \sum_j j \sum_i P(i, j),$$

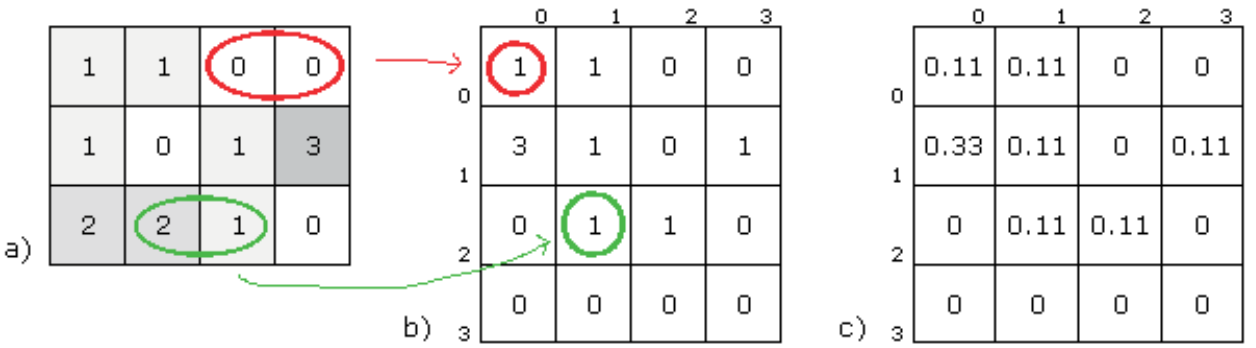


Fig. 3. Example of GLCM: a) source image with 4 gray levels; b) illustration of counting co-occurrences for d=1, θ=0°; c) GLCM, d=1, θ=0° (counted co-occurrences are divided by number of all considered pairs of points (here 9); in this example the values were rounded to two places after comma).

$\sigma_x = \sum_i (i-\mu_x)^2 \sum_j P(i,j)$, $\sigma_y = \sum_j (j-\mu_y)^2 \sum_i P(i,j)$,

$P_x(i) = \sum_j P(i,j)$, $P_y(j) = \sum_i P(i,j)$, $P_{x+y}(k) = \sum_{i,j:i+j=k} P(i,j)$,

$P_{x-y}(k) = \sum_{i,j:|i-j|=k} P(i,j)$,

HX – entropy $P_x(i)$, HY – entropy $P_y(j)$,

$HXY_1 = -\sum_{i,j} P(i,j)\log(P_x(i)P_y(j))$.

Those coefficients were calculated for each GLCM characterized by displacement d in range from 1 to the mask size (described below), and angle θ with values 0°, 45°, 90° and 135°.

Second features analysis

The mentioned above coefficients are the second statistical features of the image. Of course it is unlikely that all of them will be useful to differentiate between healthy and cancerous area. In addition, the more features we calculate the more computationally expensive the analysis is. Hence we need to select only several features with the best ability to distinguish healthy and suspicious region.

In order to compute those features there were selected regions of interest (ROI) within the analyzed image. Each ROI represents a part of it and all ROIs together covers whole prostate area. The values of GLCM and coefficients strongly depends on the shape and size of the analyzed ROIs, so we tested regions with different shape and size. In each experiment the image was covered by the fixed masks (ROIs). For each of them GLCM and finally second features were evaluated.

On the example below we used the square mask sized 40x40 pixels. The mask was moved within the image by ¼ of its size (10 pixels). So, to cover the whole image, we used 130 positions of the mask. For each position we evaluated GLCM for each displacement from 1 to 39 and for each of the 4 mentioned above angles. So we evaluated 425880 coefficients (21 for each of 20280 GLCM matrices). Within image two example areas (ROI) were selected. The first in the upper-right part of the prostate represents the healthy area, and the second – near the lower-left corner covers the cancerous lesion (fig. 4).

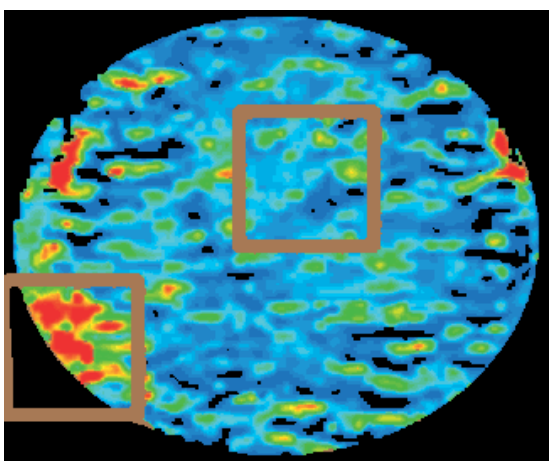


Fig. 4. The localization of selected ROIs – cancerous area near the lower-left corner, and the healthy area in the upper-right region.

Table 1. Coefficients of GLCM

no.	name	abbr.	value
f_1	energy	ENE	$f_1 = \sum_{i,j} P(i,j)^2$
f_2	entropy	ENT	$f_2 = -\sum_{i,j} P(i,j)\log P(i,j)$
f_3	homogeneity	IDM	$f_3 = \sum_{i,j} \frac{1}{1+(i-j)^2} P(i,j)$
f_4	inertia	CON	$f_4 = \sum_{i,j} (i-j)^2 P(i,j)$
f_5	correlation	COR	$f_5 = -\sum_{i,j} \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x\sigma_y} P(i,j)$
f_6	variance	VAR	$f_6 = \sum_{i,j} (i+j-\mu_x-\mu_y)^2 P(i,j)$
f_7	shade	SHA	$f_7 = \sum_{i,j} (i+j-\mu_x-\mu_y)^3 P(i,j)$
f_8	prominence	PRO	$f_8 = \sum_{i,j} (i+j-\mu_x-\mu_y)^4 P(i,j)$
f_9	sum average	SA	$f_9 = \sum_{i=2}^{2N_g} P_{x+y}(i)$
f_{10}	sum entropy	SE	$f_{10} = -\sum_{i=2}^{2N_g} P_{x+y}(i)\log P_{x+y}(i)$
f_{11}	sum variance	SV	$f_{11} = -\sum_{i=2}^{2N_g} (i-f_9)^2 P_{x+y}(i)$
f_{12}	difference average	DA	$f_{12} = \sum_{i=0}^{N_g-1} P_{x-y}(i)$
f_{13}	difference entropy	DE	$f_{13} = -\sum_{i=0}^{N_g-1} P_{x-y}(i)\log P_{x-y}(i)$
f_{14}	difference variance	DV	$f_{14} = -\sum_{i=0}^{N_g-1} (i-f_{12})^2 P_{x-y}(i)$
f_{15}	information measure	IMC1	$f_{15} = \frac{f_2 - HXY_1}{\max(HX, HY)}$
f_{16}	coefficient of variation	COV	$f_{16} = \frac{\sigma(P(i,j))}{\mu(P(i,j))}$
f_{17}	peak transition probability	MAX	$f_{17} = max(P(i,j))$
f_{18}	diagonal variance	DIAV	$f_{18} = \sigma^2(P(i,j))$
f_{19}	diagonal moment	DIAM	$f_{19} = \sum_{i,j} \left(\frac{1}{2}\lvert i-j\rvert P(i,j)\right)^{\frac{1}{2}}$
f_{20}	second diagonal moment	DSM	$f_{20} = \sum_{i,j} \frac{1}{2}\lvert i-j\rvert P(i,j)$
f_{21}	triangular symmetry	TRS	$f_{21} = \lvert P(i,j) - P(j,i)\rvert$

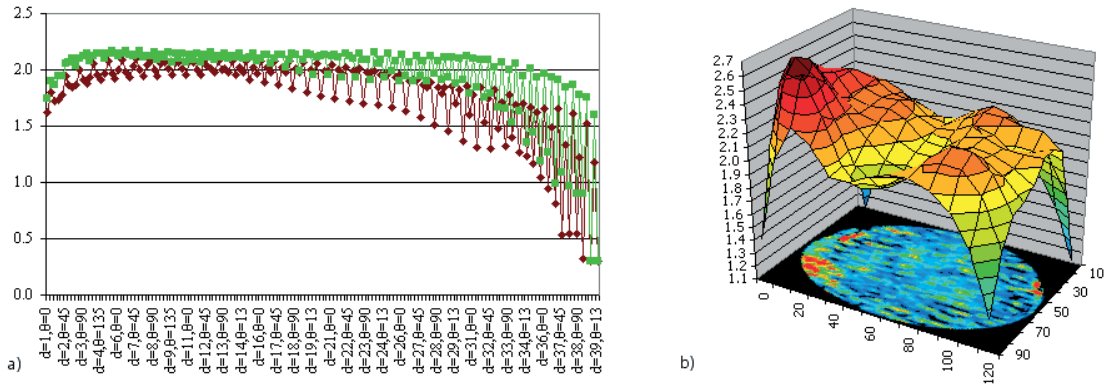


Fig. 5. The entropy (f_2) of analyzed image: a) The value of entropy in dependence on parameters d and θ for two selected ROIs; green – healthy area; red – cancerous area; b) Entropy map, d=10, θ=0º.

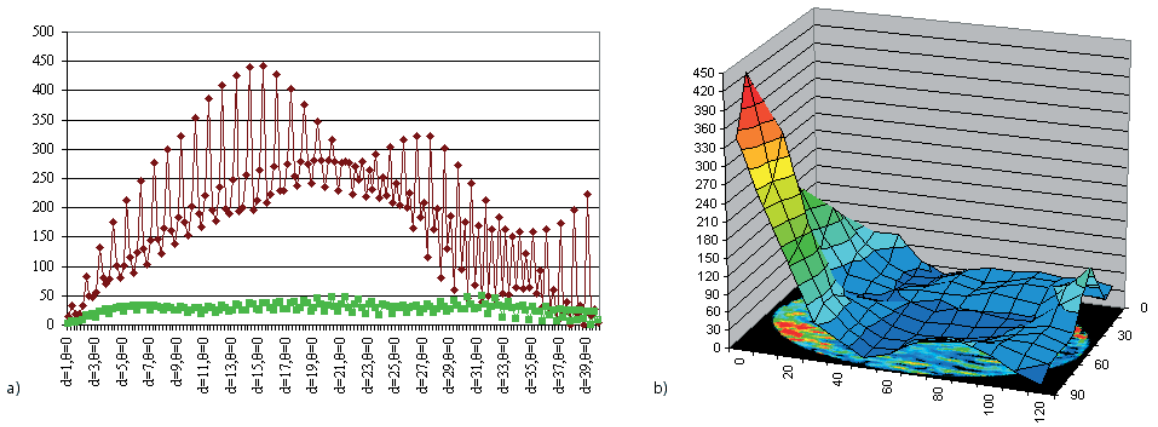


Fig. 6. Inertia (f_4): a) Dependence on d and θ for selected ROIs; green – healthy area; red – cancerous area b) Inertia map, d=15, θ=45º.

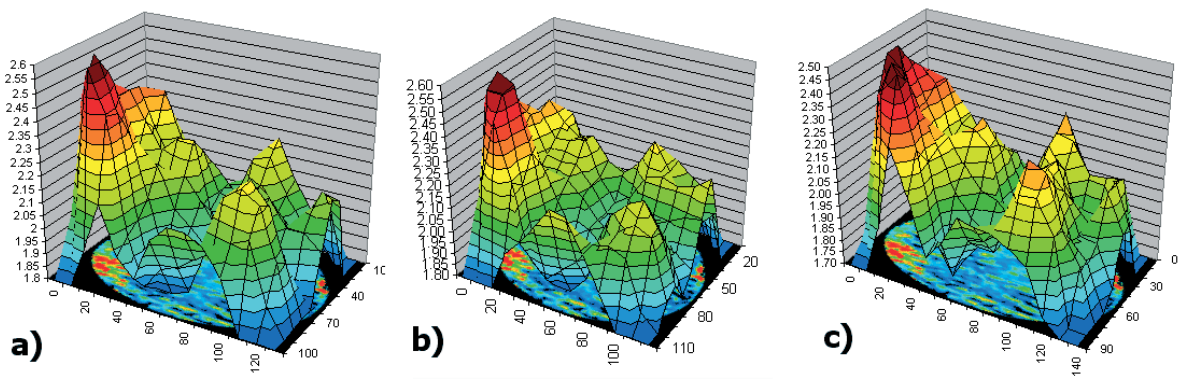


Fig. 7. The influence of mask size. Entropy for d=10, θ=0°. On fig.5b mask 40x40. Here: a) 30x30; b) 20x20; c) 10x10 (in this case, in term of small mask, d=5).

For each ROI the graph illustrating dependence the value of the analyzed coefficient on displacement d and angle θ were drawn. For further analysis such d and θ were chosen, where the differences between the healthy and cancerous area were remarkable and they were enough stable (with reference to the neighboring parameters). On the graph below (fig. 5) you can see that entropy could be a good determinant of the cancerous region (the highest value).

The result for some coefficients and parameters are more spectacular (fig. 6) but here the effect of border area occurs. The problem will be discussed in next part.

Another aspect to be considered is the size of the ROI (mask). Presented in the above examples size 40x40 pixels is about the size of the cancerous area. The example below (fig. 7) shows that small mask can point out too many local differences in texture instead of the key, pathological change.

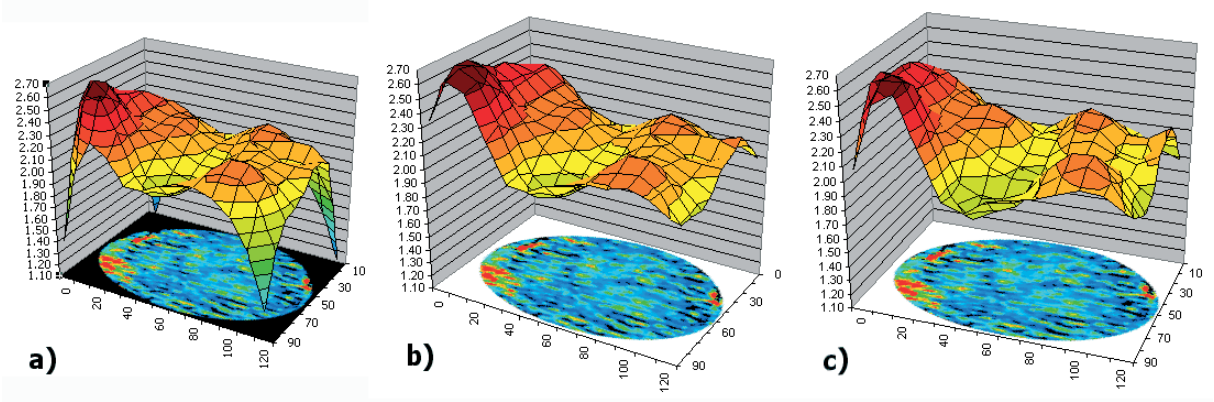


Fig. 8. The influence of mask shape. Entropy for d=10, θ=0°: a) square mask 30x30; b) horizontal rectangle mask 40x20; c) vertical rectangle mask 20x40.

Also the direction and shape of the ROI is important. We observed that texture in healthy area is rather horizontal than vertical. So, it is recommended to choose the shape in such a way to enable emphasis of those anisotropy, for example rectangle higher than wider (fig. 8).

The boundary problem

As mentioned, one of the key problems was to select the proper ROI for the analysis. In the above quoted examples we concentrated on the rectangular (squared or not) mask. But the shape of the prostate is almost oval. That's why the "boundary problem" in the corners occurs – the resulted values depends on the size and location of outside-prostate area within ROI.

According to previous examples it is worth to see that entropy – indeed higher on the left side of the image, where the pathological change is visible – is not the highest exactly in that ROI which covers all prostate points belonging to the cancerous area (fig. 9). Also comparing presented earlier graphs of entropy with the values for two example ROIs (fig. 5a) give us food of thought. On the figure 5a higher values are for the healthy mask (green line), while in the second (cancerous, red line) values are almost always smaller.

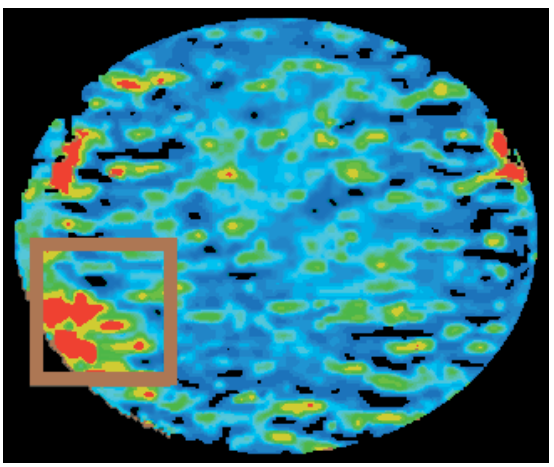


Fig. 9. The mask for which the entropy pointed out on the fig. 5b is the highest.

To explain this disagreement we must notice that almost half of the area indicated on fig. 4 as cancerous region are points outside the prostate. So it is the effect of "boundary problem". During the analysis, the black points, visible on the

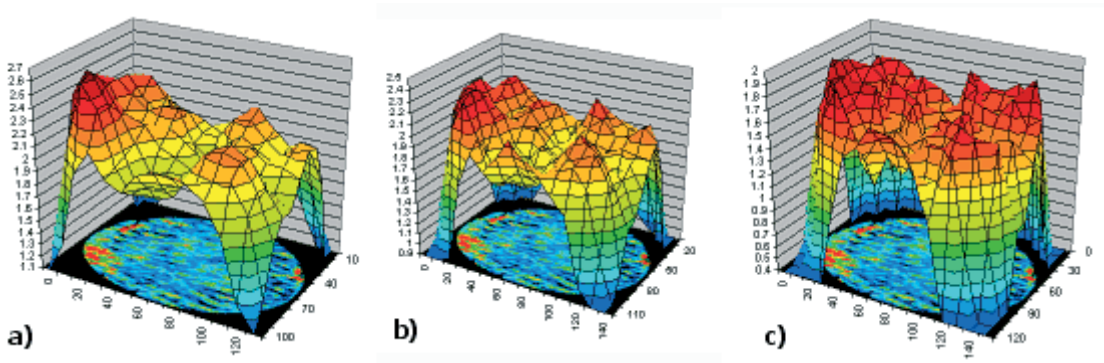


Fig. 10. Entropy calculated for different shapes of the mask and different approaches to the outside-prostate area (d=10, θ=0°): a) outside-prostate area counted (treated as region with no perfusion), rectangular mask; b) outside-prostate area omitted, rectangular mask; c) outside-prostate area omitted, circle mask.

corners of the image were treated as points without observed perfusion, exactly as black pixels within prostate. Now it is obvious that this way is unacceptable. To solve this problem, two alternative methods were proposed:

Leave out points outside the prostate. In this approach matrices calculated for border ROIs are remarkably sparse. It could influence on the values of coefficients.

Analyze ROIs with non-rectangular shape, eg. oval or circle. In this solution some border effects are eliminated but not for all cases – if the middle of the circle is almost on the border, still many analyzed pixels are outside the prostate. Additionally, the computational cost of selecting a circle mask is higher than a rectangular one.

The next graph (fig. 10) presents entropy for three different mask sized 40x40 pixels. We see that the best solution is rectangular mask where the pixels outside the prostate are not analyzed.

Conclusion

In this paper it was shown that it is possible to select such parameters of perfusion prostate image, which are deterministic and independent from personal assessment. Some algorithms which enable a fast, automatic and correct selection of cancerous regions were presented and discussed. The results assure us that the p-CT technique may be very useful in early prostate cancer diagnosis and it is worth to continue the exploration of this field.

In the Oncology Center in Cracow the p-CT method is being used to examine other patients. Thanks to that it will be possible to verify the usefulness of the proposed algorithm. In the further work the researches will be also expanded to other perfusion parameters in order to determine the effectiveness of each one.

Although it seems that the p-CT method has a big potential to recognize PCa and to point out the cancerous regions, the full verification of the proposed algorithms and evaluation of sensitivity and specificity of this diagnostic method needs a lot of work and a close cooperation between radiologists, pathologist, computer scientists and engineers.

References

1. Estimated New Cancer Cases and Deaths by Sex, US, 2008, http://www.cancer.org/docroot/MED/content/downloads/MED_1_1x_CFF2008_Estimate-d_Cancer_Cases_Deaths_All.asp
2. Krajowy Rejestr Nowotworów, Raporty na podstawie danych Centrum Onkologii, http://85.128.14.124/krn
3. Hricak H., Choyke P., Eberhardt S. et al., Imaging Prostate Cancer. A Multidisciplinary Perspective, Radiology 2007; 243(1): 28-53.
4. Roscigno M., Scattoni V., Bertini R. et al., Diagnosis of prostate cancer. State of the art, Minerva Urol Nefrol 2004; 56(2): 123-145.
5. Simon H. (ed.), Prostate Cancer, [in:] Lifespan's A – Z Health Information Library 2006, http://www.lifespan.org/adam/indepthreports/10/000033.html
6. Miles K. A., Functional computed tomography in oncology, European Journal of Cancer 2002; 38: 2079-2084.
7. Haralick R. M., Shanmugam K., Dinstein I., Textural features for image classification, IEEE Transactions on Systems, Man and Cybernetics 1973; 3: 610-621.
8. Walker R. F., Adaptive multi-scale texture analysis with application to automated cytology, University of Queensland 1997.

COMPUTER SCIENCE	⟷	DIAGNOSTICS
↕		↕
ENGINERING	⟷	RADIOLOGY

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 31–34

USAGE OF NAIVE BAYES CLASSIFIER IN DECISION MODULE OF E-LEARNING DECISION SUPPORT SYSTEM

MARCIN CHABIOR, ANNA NOGA, MAGDALENA TKACZ

University of Silesia
Faculty of Computer and Materials Science
Institute of Computer Science

Abstract. This article describes adopted, within the scope of one of e-learning system creation, a method of diagnosis generation based on defined symptoms with the use of the Bayes naive algorithm. It is intended for Medical Faculties students, inference module shall facilitate determination of probability of the influence of individual symptoms on diagnosis. Practical application of such approach presented in the article gives the possibility of distant verification of student's knowledge taking into account not only diagnosis, but also will facilitate gaining skills of appearance materiality determination of symptoms in relation to diagnosed affection. For the testing purposes of adopted assumptions the hundred element database has been applied depending on spine affection in discopathy form and factors, which can affect the origin of the affection – obesity and hard physical work. The operation of the Bayes algorithm has been presented within the scope of influence evaluation of enumerated factors on affection origin.
Keywords: Bayesian network, naïve Bayes estimation, elearning, decision support systems, medical e-diagnosis.

Introduction

Initial assumptions of e-learning system planned to be developed in Medical University of Silesia in Katowice resulted in in-depth study in forming a concept of the system. Successively with clearly defined computer scientists a certain assumptions has been formed for all modules at the same time. But the second thought about module intended for online diagnosis basing on symptoms (concerning the method of determining the influence of a certain symptoms in final diagnosis of a patient, which a decision-based module in the decision support system) appeared.

Searching the solution for that problem, we decided to use the Bayes naive estimation. As opposed to quantitative approach which is basic form of probability estimation where population parameters are constants and randomness depends on data, the Bayes approach is based on other assumptions. In the Bayes statistics, parameters are considered to be random variable, however data are treated as fixed constant [1].

The Bayes network is a consistent representation of total distribution of a random variable which applies (in the process of compression), a part of marginal and conditional independence between variables [2]. The Bayes network is defined as acyclic direction graph, where knots represents variables and edges – direct reason dependence. The syntax of the Bayes network is as follows – for each random variable graph knots are formed, graph edges constitute connections and reflect the dependence between variables and the distribution of conditional probability. For each knot at known values of a distribution of parents probability is recorded. Conditional distribution is presented in the form of Conditional Probability Tables (*CPT*) which includes a distribution of conditional probability for Z_i and for each combination of parents value (Figure 1).

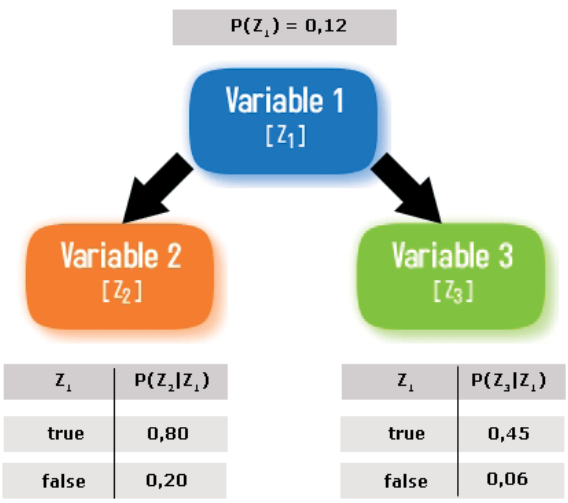


Figure 1 Example of Bayesian Network with *CPT* tables

The Bayes analysis requires so called "a priori" distribution which is the first distribution of θ, where θ represents

parameters of unknown distribution. With this distribution the expert knowledge on Θ distribution is modeled. The qualification of individual probabilities of influence about definite values the parameters usually in case the possible medicines is. It required therefore experts knowledge which was provided for in project of application. Possible it is as the application on entry of value of initial probabilities parameters and in case of lack of suitable experts knowledge will make possible automatic their the adjustment across use the of mean value (attributing to every of parameters the same value of probability). Taking under attention fact, that application will serve students teaching, mentioned solution gives the possibility of opinion of pertinence put by student of value of probabilities. In other words, possibility of checking how student be able to accurately diseases from sickness individual definite symptoms. Lack of expert knowledge does not impose the need to change this method because it is possible to assume non-information a priori distribution which allows to assign the same probabilities for all parameter values. The amount of data in analyzed data set usually has significant influence on assumed solutions of a priori distribution which causes its modification. In this way we obtain a posteriori distribution $p(\theta|X)$ where *X* represents the whole data set. Distribution a posteriori is expressed with the formula:

$$p(\theta|X) = \frac{p(X|\theta)p\theta}{p(X)} \tag{1}$$

where $p(X|\theta)$ is a likelihood function
$p(\theta)$ is a priori distribution
$p(X)$ is marginal distribution

The Bayes theorem for simple events presents as follows – let *A* and *B* will be events in testing space. Than conditional probability *P(A|B)* can be described in the following way:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2}$$

Analogously

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{3}$$

And after this simple mapping we obtain:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4}$$

Which finally is the Bayes theorem for simple events.

The problem

In the first phase of the our approach with the Bayes methods efficiency testing (in application of concerning the problem of determining levels of variables interaction on formulated diagnosis) two quality explanatory variables has been applied – *overweight* and *physical work*. The diagnosis problem was to make appropriate classification of new patients with regard to the possibility of appearing disc prominence of interverterbal disc of a spine (discopathy) basing on existing learning set of connections between variable discopathy and two descriptive variables. Discopathy variable can has two possible values – truth and false, assuming the following notation:
B *means overweight* = true
$\bar{B}$ *means overweight* = false
F *means physical work* = true
$\bar{F}$ *means physical work* = false
R *means the possibility of discopathy appearance* = true
$\bar{R}$ *means the possibility of discopathy appearance* = false
we can describe dependences with the formula:

$$\theta_{MAP} = R_{MAP} = \arg\max p(B \cap F|R)p(R) \tag{5}$$

Here the method of maximal a posteriori has been used. It is a method of a posteriori distribution estimation (estimator *MAP* – M*aximum a Posteriori method*). The search of Θ estimator with the method of maximal a posteriori consist of seeking such value Θ, which will maximize the function $p(\theta|X)$. Using formula on distribution a posteriori we obtain the following:

$$\theta_{MAP} = \arg\max \frac{p(X|\theta)p(\theta)}{p(X)} = \arg\max p(X|\theta)p(\theta) \tag{6}$$

where *X* represent the whole data set.

Use only two variables is explanatory large simplifying but it allow to analysis step by step of issue in present article. In reality the number of variables having the influence on occurrence the diseases, can be considerably larger but the pattern of analytic conduct stays the same.

Results

The next step has been to determine which variables value *discopathy* (*true* or *false*) has greater probability value to distinguish the estimator *CMAP* for the variable. For this purpose the first step was to find marginal and conditional series of probabilities. Results of estimated probabilities a posteriori are presented in the Table 1.

The next step was to found total conditional probability in a form $P(B, F|R)$. Estimation results are presented in the Table 2.

Having the knowledge of total conditional probability, the evaluation of estimator MAP of discopathia variable was possible for all four variables combination: *overweight* and *physical work* with formula (5). Because the result for each possible combination has been false, it did not correlated with expectations it was necessary to estimate the quotient of aposteriori chances – which reflect the level in support of the benefit of every possible classification. The quotient of a chance in case of a posteriori has been estimated in the following way:

$$\frac{p(\theta_R|X)}{p\bar{\theta}_R} = \frac{p(\theta_R|X)p(\theta_R)}{p(\bar{\theta}_R X)p(\bar{\theta}_R)} \tag{7}$$

Where θ_R is a specific classification of unknown target variable.

The value of quotient greater than 1 informs, that a posteriori distribution gives us a positive classification (*true*) and

Table 1 Results of estimated probabilities a posteriori

	Set size (false)	Set size (true)	Probability value
Overweight	60	40	0,4
Physical work	48	52	0,52
Discopathy	80	20	0,2
Overweight = *true* assuming that discopathy = *true*	-	13	0,65
Overweight = *false* assuming that discopathy = *true*	7	-	
Overweight = *true* assuming that discopathy = *false*	-	27	0,3375
Overweight = *false* assuming that discopathy = *false*	53	-	
Physical work = *true* assuming that discopathy = *true*	-	11	0,55
Physical work = *false* assuming that discopathy = *true*	9	-	
Physical work = *true* assuming that discopathy = *false*	-	41	0,5125
Physical work = *falsehood* assuming that discopathy = *false*	39		

Table 2 Estimation results

Patients having the obesity who perform hard physical work	
$P(B \cap F\|R)P(R) = 0{,}1$	$P(B \cap F\|\bar{R})P(\bar{R}) = 0{,}1$
Patients having the obesity who do not perform physical work	
$P(B \cap \bar{F}\|R)P(R) = 0{,}03$	$P(B \cap \bar{F}\|\bar{R})P(\bar{R}) = 0{,}17$
Patients without the obesity who perform hard physical work	
$P(\bar{B} \cap F\|R)P(R) = 0{,}01$	$P(\bar{B} \cap F\|\bar{R})P(\bar{R}) = 0{,}29$
Patients without the obesity who do not perform hard work	
$P(\bar{B} \cap \bar{F}\|R)P(R) = 0{,}06$	$P(\bar{B} \cap \bar{F}\|\bar{R})P(\bar{R}) = 0{,}22$

Table 3 The quatients of a posteriori chance

$\frac{P(B \cap F\|R)P(R)}{P(B \cap F\|\bar{R})P(\bar{R})} = 1$	Patient suffering from overweight and performing hard physical work Both classifications are supported at the same level
$\frac{P(B \cap \bar{F}\|R)P(R)}{P(B \cap \bar{F}\|\bar{R})P(\bar{R})} = 0{,}17$	Patient suffering from overweight who do not perform physical work Justification Discopathy = *true* with regard to discopathy = *false* is at the level of 17 %
$\frac{P(\bar{B} \cap F\|R)P(R)}{P(\bar{B} \cap F\|\bar{R})P(\bar{R})} = 0{,}034$	Patient who does not suffer from overweight and performs hard physical work Justification Discopathy = *true* with regard to discopathy = *false* is at the level of 3,4 %
$\frac{P(\bar{B} \cap \bar{F}\|R)P(R)}{P(\bar{B} \cap \bar{F}\|\bar{R})P(\bar{R})} = 0{,}227$	Patient who does not suffer from overweight and does not perform hard physical work Justification Discopathy = *true* with regard to discopathy = *false* is at the level of 22,7 %

the quotient value lesser than 1 on the contrary – a posteriori distribution gives us a negative classification (*false*). The quotient value equal 1 means that the information from posteriori distribution gives the same level at both classification . In the case we considered the following results has been obtained (Table 3).

The considered case (with two explanatory variables with binary values assumed, as well as variable binary of the target) does not cause large programming difficulties, because of uncomplicated mathematic theory usage. Basing on practice, if there is a possibility of arising of *p* variables with *k* values then we will have a lot of $O(k^p)$. As the practice shows probabilities which have to be estimated it is the numb. It means that when we have 20 binary variables (*k*=2) it is necessary to estimate from 2 up to 20 probabilities (1 048 576). The assumption about independences of explanatory variables has to be done here.

The two events *A* and *B* are independent under condition, if for some event *C*: $p(A \cap B|C) = p(A|C)p(B|C)$. The assumption of conditional independence can be described with the following formula,

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m|\theta) = \prod_{i=1}^{m} [p(\square)X_i = x_i|\theta) \quad (8)$$

where x_1, x_2, …, x_m are every possibile combination of *X* described factors.

However the notation

$$\theta_{NB} = argmax \prod_{i=1}^{m} p(X_i = x_i|\theta)p(\theta) \quad (9)$$

is named as the Bayes naive assumption or the Bayes first series assumption.

This approximation facilitates to show complete conditional distribution from the $O(k^p)$ quantity to the one-dimension distribution product requiring $O(kp)$ probabilities. We obtain the model of conditional independence which is linear and not exponential dependent from variables number *p*. Despite the fear of reduced model's efficiency, decreasing the number of parameters is a positive solution and is related to the problem of diagnostic evaluation based on determined symptoms. For instance, where x_i will be Medical symptom and C_K correspond to various diseases, then the assumption that if a certain person that suffer from C_K disease, then the probability of one symptom dependency (only from C_K disease and not from other symptom) will be faultless. In other words, we are modeling how symptoms appear, given each disease, as having no interactions (note that this does not mean that we are assuming marginal (unconditional) independence).

A practical implementation of such a work are still "in progress". We would like to take into account the possibility of more than binary variables in the future and in independent way-together with the other variable type. Assuming that maximal number of variables is 10 and the number of possible values is 4 than the estimation of $k \cdot m = 4 \cdot 10$ values will be needed. As the measure of justification, each variable affects the classification decision, the value of chances logarithm a posteriori were applied. The quotient is described in the following way::

$$\frac{\log\left(p(X_i = x_i|\theta)\right)}{p(X_i = x_i|\bar{\theta})} \quad (10)$$

Received results for the tested patent group is presented as follows:

$$\frac{\log\left(p(B|R)\right)}{p(B|\bar{R})} = 0{,}6378 \quad (11)$$

$$\frac{\log\left(p(F|R)\right)}{p(F|\bar{R})} = -0{,}2514 \quad (12)$$

As we can notice, the results shows that patients suffering from overweight give positive influence on the probability of affection origin in a discopathy form. Physical work does not presents mentioned tendencies.

Conclusion

The method presented in the article presents clearly the efficiency of statistical methods application, based on the Bayes theorem. It should be usable for utilization as a part of a decision support system, which is able to diagnose with the determination of the fixed probability level of each symptoms Further use of the application will be as one of the module of distant-learning complex system, and we hope that it will be quite good method for classical education in diagnosis. The system flexibility will facilitate the use of the same application regardless of the Medicine field – where the diagnosis is required together with the determination of the level of its correctness and it can be done with classical probability measure.

Bibliography

1. Kłopotek M. A.: Inteligentne Wyszukiwarki Internetowe, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001.
2. Ross K. A., Wright C. R. B., Matematyka Dyskretna, Wydawnictwo Naukowe PWN, Warszawa 2006.
3. Bøttcher S. G. Dethlefsen C.: Learning Bayesian Networks with R, Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March 20-22, Vienna, Austria, 2003.
7. Larose D. T.: Data Mining Methods and Models, John Wiley & Sons, Hoboken 2006.
8. Moczko J. A.: Wybrane metody analizy danych jakościowych na przykładzie wyników badań kardiologicznych, StatSoft Polska, Kraków 2008.
9. Hand D., Mannila H., Smyth P.: Principles of Data Mining, Massachusets Institute of Technology, Cambridge 2001.
10. Friedman N., Linial M., Nachman I., Peer D.: Using Bayesian Networks to Analyze Expression Data, Hebrew University, J. Computational Biology, Vol. 7, No. 3-4, pp. 601-620, 2000.Kwiatkowska A. M.: Systemy wspomagania decyzji, Wydawnictwo Naukowe PWN, Warszawa 2007.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 35–37

E-LEARNING WITH USE OF VIRTUAL PATIENT IN PHARMACY EDUCATION

KRZYSZTOF NESTEROWICZ, SEBASTIAN POLAK

Unit of Pharmacoepidemiology and Pharmacoeconomics, Faculty of Pharmacy Jagiellonian University Medical College, Medyczna 9 Street, Kraków 30-688, Poland
contact author: krzysztof.nesterowicz@gmail.com

Abstract: e-learning is an approach to facilitate and enhance learning through, and based on both computer and communications technology. At first it is necessary to acknowledge an important growth of blended and collaborative learning applications. Many institutions try to make universal learning modules which promote cooperative methods of work. Other initiatives focus on the idea of building a common e-learning system.
One of available distance learning techniques in studying pharmacy is a virtual patient application. The practical example of such educational attitude is presented based on the in-house developed patient case.
Keywords: e-learning, e-education, blended learning, virtual patient

Introduction

E-learning or e-education is becoming increasingly popular and the Internet is laden with the ever growing number of applications of the subject matter. The European Community is fully aware of the importance of these developments and supports them in many ways. Today one can pinpoint the direction of the progress which has so much influence on the European investigations. At first it is necessary to acknowledge an important growth of blended and collaborative learning applications. Many institutions try to make universal learning modules which promote cooperative methods of work. Other initiatives focus on the idea of building a common e-learning system [1].

E-learning is a term which is commonly used, but does not have a common definition [2]. Most frequently it seems to be used for web-based distance education, with no face-to-face interaction. However, also much broader definitions are common. For example, it may include all types of technology enhanced learning, where technology is used to support the learning process. Although pedagogy is usually not part of the definition, some authors do include it [3]. For example in this definition, where e-learning is said to be: *pedagogy empowered by digital technology* [4].

It cannot be forgotten that e-learning means student-student, student-teacher or teacher-teacher interaction. Participants of an e-learning course should be always aware about results of their education process and their knowledge should be evaluated during the course. Therefore uploading some materials on a website, like lectures or exercises is still not real e-learning because it lacks the component of interactivity [5].

Today e-education is often connected with traditional learning. In this way, the first direction of e-learning development in Europe comes into existence – blended learning [1]. The method combines advantages of e-learning with advantages of traditional learning. Main advantages of the e-learning are listed below:

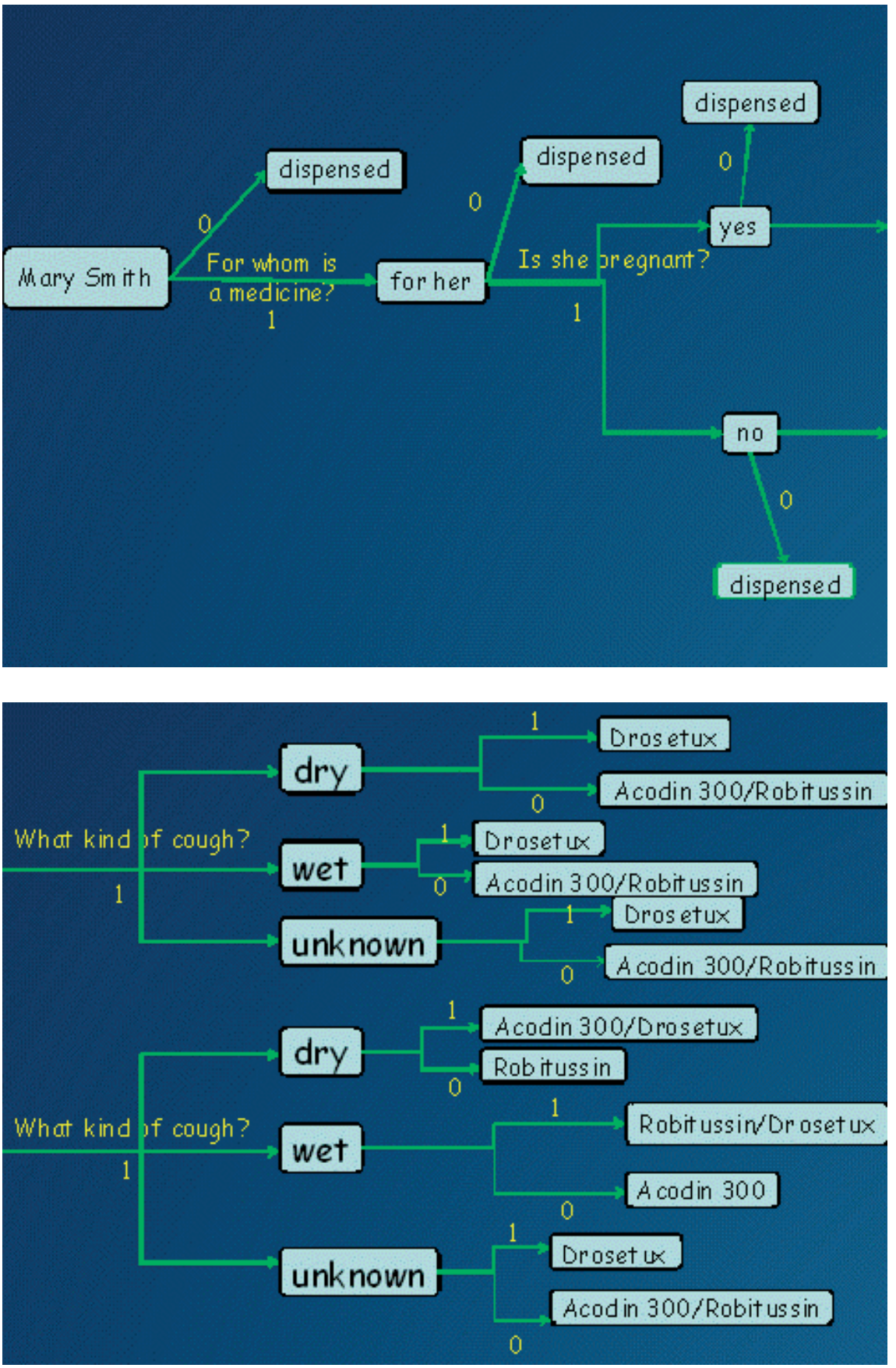
- differentiation of learning,
- cost reduction,
- time flexibility,
- integrated assessment tools,
- multimedia forms,
- high interactivity [5].

On the other hand, there are undisputable advantages of traditional learning, including:

- direct interpersonal relations,
- live contact with the tutor,
- exact definite time and place of training,
- evaluation of knowledge,
- contact with real experiment,
- training of interpersonal abilities [5].

One of available distance learning techniques which can be used during the pharmacy studies is a virtual patient application. It has become a useful way of teaching pharmaceutical care where students need to challenge with many situations related to pharmacotherapy of patients with chronic diseases. By studying with virtual patients students are involved in decision making procedures.

Simulated patients increase the availability of training opportunities for pharmacy students, making them less depen-



Picture 1. An example of a decision making tree [8].

dent on actual cases to learn how to handle different situations in pharmaceutical care. Unlike real patients, simulated patients can be accessed on demand and they can be endlessly replayable to allow the user to explore different options and strategies. They can be structured with narratives that represent real situations while challenging the user with a wide range of tasks [6]. Students learn how to handle with some cases which later they can meet in reality while providing pharmaceutical care to their own patients.

E-education in Poland and abroad

Unfortunately there are still not suitable legal regulations in the current education system. This is the basic obstacle for the development of distance learning in Poland. Nevertheless the situation has improved recently. Constructing a didactic process with the use of the Internet potentially requires the competent application of both educational regulations and baseline programmes, in schools [1].

The level of technical culture and ability to use computer is another barrier. The majority of users confess that their computer activity is limited to simple applications, frequently created especially for beginners.

In the Faculty of Pharmacy, Medical College, Jagiellonian University in Krakow, Poland there is a course "Practical Pharmacy in a Community Pharmacy" where students learn by using blended learning method. Participants listen to the interview with a patient which is stored in the computer and provide a patient an adequate pharmaceutical care. During the course they are provided with some interactive presentations and they are allowed to check all medical information in online databases. At the end of the course attendees need to write a final test. They do it online and receive their results immediately.

This innovative approach challenges students and promotes interactive learning. Student evaluations indicates achievement the objective of creating a course that more closely simulates the actual provision of pharmaceutical care.

Pharmaceutical care laboratory courses offer students the opportunity to learn and practice pharmaceutical care skills in a controlled environment. These courses usually include instruction in dispensing as well as clinical activities. Typically, a new patient case is presented in each laboratory section and no "follow up" care of patients from previous laboratory sections is discussed [7].

The authors of the above course have created several virtual patients whose responses to care vary based on students' input and recommendations. For example, if a wrong drug or dose is recommended, an adverse event may occur; if patient counseling is not provided, an error in administration may be encountered; if patient compliance is not evaluated; drug levels may be unexpectedly altered. The application of virtual patients who change in ways appropriate to the recommendations students make creates scenarios that more closely parallel the reality of pharmaceutical care. Since students must search online databases and study other resources to gather the information needed to make informed decisions about their patients, the course also builds the data collection skills.

Decision making tree

Nowadays we are to implement in the subject for Pharmacy students "Practical Pharmacy in a Community Pharmacy" exercises with a decision making tree tool to teach pharmaceutical care. Below there is an example of such decision making tree (Picture 1).

The Virtual Patient Mary Smith asks for a medicine against cough and says she is not taking other medicines now.

A participant is granted with one point for each right question given before dispensing a medicine (Picture 1).

Exercises like that help students to improve their decision making skills which they need to possess during their everyday work as pharmacists. Besides students learn how to get proper information from a patient by asking right questions.

In the literature teachers sometimes complain that preparing such materials is very time consuming and require special computer skills. But from the other hand this form of teaching is also convenient for educators because it enables to check students' skills in shorter time comparing to essays or formularies they would need to fill.

References

1. Meger Z., E-EUROPE, International Conference on e-learning in education, E-Learning – European Trends, 9, Warsaw 2006.
2. Dublin L., If You Only Look Under the Street Lamps... Or Nine e-Learning Myths, The eLearning Developers' Journal, 1-7, 2003.
3. http://en.wikipedia.org/wiki/E-Learning
4. EC, Communication from the Commission: E-Learning – Designing "Tejas at Niit" tomorrow's education, Brussels: European Commission, 2000.
5. Nesterowicz K., E-learning in Pharmacy Education, European Pharmaceutical Students' Association Newsletter, Vol. 16, Ed. 3, July 2009.
6. http://en.wikipedia.org/wiki/Virtual_patient
7. Hussein G., Kawahara N., Adaptive and Longitudinal Pharmaceutical Care Instruction Using an Interactive Voice Response/Text-to-Speech System, Am. J. Pharm. Educ., 70(2): 37, April 15, 2006.
8. Nesterowicz K., Short communication, International Conference on Virtual Patient, Kraków, Poland, June 5-6, 2009.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 39–43

INTERACTIVE KNOWLEDGE BASE FOR EXPERT SYSTEM

ANNA NOGA, MARCIN CHABIOR, GRZEGORZ SAPOTA

The Silesian University
IT and Science Department

Summary: In this article the project of interactive knowledge enabling the record of disease symptoms is presented by doctors to prepare learning set for expert system based on self-organizing networks. Knowledge base will enable the formulation of a diagnosis of disease changes caused by a discopathy and degenerative changes of the backbone.
The SpineMedical system can be also used for education purposes for doctor and Medicine faculty students. It enables to complete database by doctors who formulate a diagnosis while examining patients. The SpineMedical system was equipped with the possibility of faulty records' filtering. In the article results of expert system operation which diagnoses patient morbidity was presented.
Keywords: Knowledge modelling, expert system, discopathy

Introduction

The SpineMedical program is formed with the cooperation of the Silesian University in Katowice and the Silesian Medicine University.

The purpose of the application is to facilitate the communication, team work as well as processing, archiving and files' creation. The SpineMedical gathers data on discopathy and degenerative changes of the backbone for quality and quantity diagnosis of patients' affection.

The backbone is a central anatomical structure of a human being and diseases are currently one of major health problems because they hinder normal living, family living, work, rest and many other daily activities.

Familiar backbone pain last increased significantly. The fact that more and more children and youth has similar problems is considered.

The discopathy can be divided into two morbidities. In the first phase the discopathy were treated as needed and in the later phase of the disease the diagnostic imaging were used that is examination of magnetic resonance type and computer-assisted tomography as well. Unfortunately, a diagnosis in the disease advanced stage is connected with the operation this is nucleus removal [5, 6].

Intervertebral disc disease can appear in cervical, thoracic and lumbar section. The Discopathy symptoms in initial stage manifest in thoracic-lumbar section pain. It can appear pain radiation to one or both lower limbs (hips, genua, lower legs or feet), paraesthesia. Despite a sick person can have the impression of increased tension of backbone muscles, limitation in this section movement caused by pain. The following symptoms of a disease are neurology deficits. In case of more advanced degenerative changes the neurology symptoms' defects can occur this is a dysesthesia in various places of lower limbs, weakening of foot, lower leg muscles and pareses of lower leg nerves where the most often is paresis of peoneral nerve and in the final stage a paralysis of peoneral nerve. However, in advanced discopathy disorders of urinary bladder constrictor muscles, anus or virility and libido disorders can occur.

Description of interactive program operation

The SpineMedical is an application which is computer scientist education system which is intended for doctors and students and covers the diagnostics in discopathy and backbone's degenerative changes. This application allows to patient's „conversation" to formulate a diagnosis with administer method and a choice from the list of proposed questions. The program is based on database which can be updated, viewed and corrected. A doctor has a possibility to complete data with records connected with new cases and has a possibility of free modification of base's resources (Fig. 1).

Application operation is based on dialogue windows. When the application is activated the user chooses basic identification data like the following: age, sex, growth, weight (Fig 2.) or load data in case of first visit in a database. When data are loaded the program classifies a person in proper weight group (underweight, appropriate weight, overweight).

In the following step a patient chooses a type of pain (sporadic, transient, continuous, hard, low, moderate) and indicates the place of pain appearance (cervical, thoracic, lumbar sec-

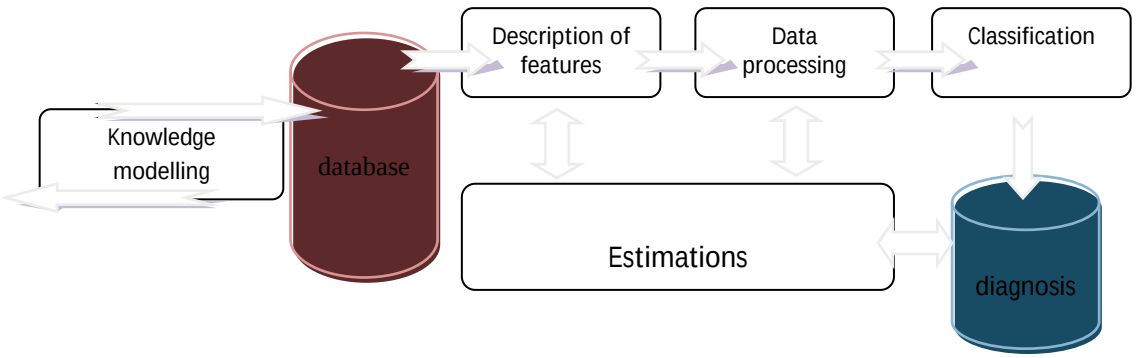


Fig. 1. Pictorial operation of the SpineMedical application

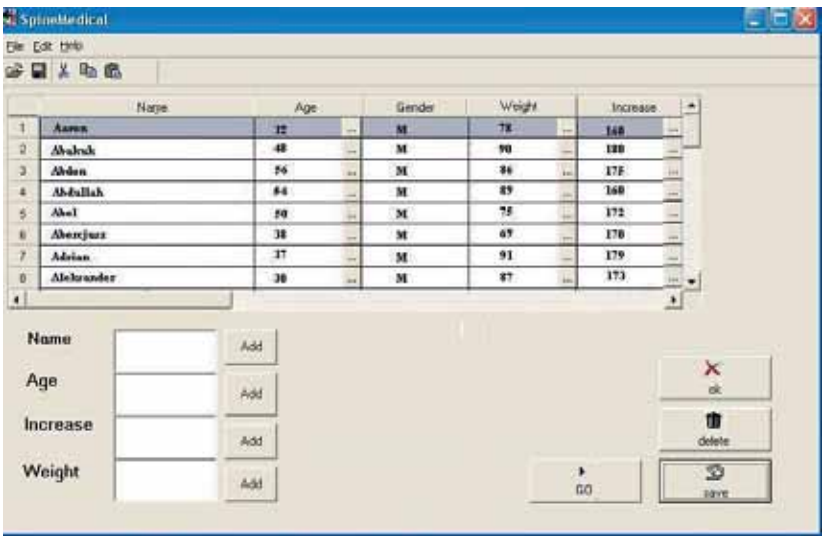


Fig. 2. The window presenting basic identification data

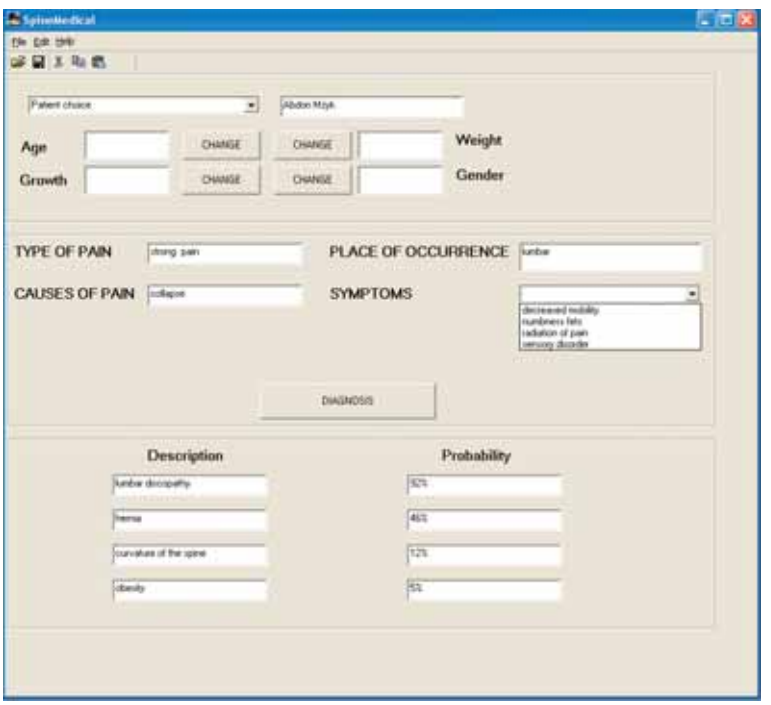


Fig. 3. The decision module of Spine Medical program

tion) what defines the number of a disc. Afterwards, he gives probable reason of pain (injury, lifting heavy weights, sedentary work, vibrations, fall, osteoporosis) [9].

In the following phase of loading data to the application a patient determines symptom connections with the disease. Final step of data loading is the quality and quantity diagnosis of patient's affection (Fig. 3).

Expert system construction

The SpineMedical application is equipped with a module enabling data filtration and expert system intended for a definition of affection type.

The purpose of data filtration module consist in verification of identification data in the first stage of application operation. Filtration module divides records into false and probably true ones.

The classification of incorrect and uncertain records has been realised by means of decision trees.

Decision trees are one of the most popular techniques used for data analysis. The advantages of the described technique are: the possibilities of creating the trees using the algorithmic techniques "divide and do"; the perfect protection against disturbance of data; the possibilities if using this technique to select and extract the features. The following coefficients have been used to create a decision tree"

- Age
- Weight
- Height
- Sex
- The type of pain
- The place of occurrence
- Accompanying symptoms

On the basis of the introduced data the system generates decisions, which give the possibility to classify properly the introduced records.

At the moment of loading faulty data the application displays information window with data incompatibility.

In the application the expert system based on neural networks was applied (Fig 4.). Expert system is a part of knowledge where loaded or inserted data, in the following steps, are rules and facts. In this work, to determine the affection type the neural networks were used because they facilitate to apply sophisticated modeling technique which can proceed complex functions' mapping. Neural networks enable forming of nonlinear model in easy way as well as allow to a control of complex multidimensionality problem and with the use of other methods it hinders modeling trials of nonlinear functions having large number of independent variables.

Neural networks are used to associate many factors and as identifying tool and predicating on this basis other factors and characteristics. These factors can be for example weakening of foot muscles and foot paraesthesia as well. Neural expert system consists of few neural networks where each intends for formulation of other diagnosis. Neural networks were performed on the basis of multidimensional perceptron where input data included the following parameters: pain type, identification data, section of pain appearance, sections' disc, pain reasons. The purpose of the network is determining quality diagnosis of patient's affection [4, 8].

Utilizing data set obtained with doctor's diagnosis the process of learning and neural networks testing was applied. For the purposes of evaluation and anticipation of disease type based on experimental results it was necessary to prepare explanatory data set and the algorithm which will allow to detect rules describing mutual connections [1].

Many factors like the quality of collecting data, a method of testing, initial data processing. The most important thing is the highest quality. These characteristics can have linear or nonlinear type, this is normal expansion, but from time to time exist features of exponential or logarithmic data dispersion to some dimension. Such features provide lots of information after performing transformation which are reverse to those observed in some dimension [2]. That is why data transformation is carried out to level large disproportions between these values in the dimension.

Data standardization aims to put values within the limit [0;1] after transformation process. For this purpose, at assumptions

$$x_{min} = \min_i x_i \quad (1.1)$$

$$x_{max} = \max_i x_i \quad (1.2)$$

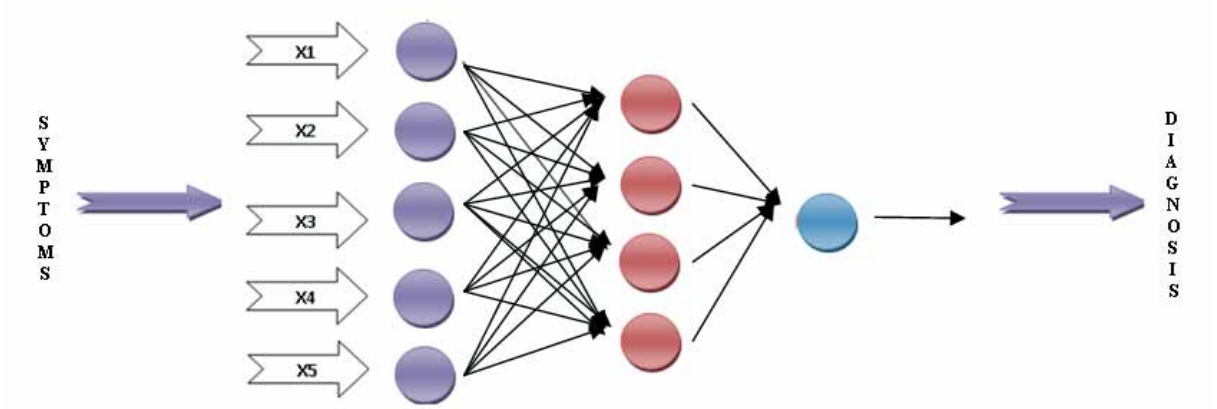


Fig. 4. Neural network construction

mapping for each characteristic is performed where i varies from 1 to n:

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1.3)$$

It is necessary to remember that at the moment we deal with faulty values considering some characteristics the result can show that they are outside of some characteristic and this can be the reason of pressing standardized value of the characteristic which can include lots of significant information. In such a case it is necessary to apply a standardization and a cut. With the use of cut standardization of values x_{min} and x_{max} selects from the set $S = \{x_1, x_2, \ldots x_n\}$ after rejecting S k% the lowest and highest values where k takes values from 5 to 10.

Selection of existent characteristics among all dimensions of input data has significant influence on learning and final level of generalization. There is a lot of methods for characteristics' selection and they vary with ways of quantity selection and the characteristics sequence in the analysis. Data comprising list of exemplary attributes declared as proper solutions are named as training set [3].

While network learning under the supervision we deal with two phases. In the first phase this is learning and the second one is the knowledge recovery. Prepared attributes' sets determine all inputs and also outputs required in the input data presentation. Current output signal of the network was compared with input signal and after processing the whole subset of learning samples we correct weight which connects network neurons was corrected. In other words, correlation aims at reducing of error measure of network operation. Unfortunately, applied methods of weighting disintegration or their elimination give positive results. It should be mentioned that even with small weights small error can be obtained. It is, thus important to determine so called significance parameter which allows to weighting elimination testing where significance coefficients are the lowest and do not cause much changes in error functions. Error function was estimated by the Taylor series [7].

$$\delta E = \left(\frac{\delta E}{\delta w}\right)^T * \delta w + \frac{1}{2}\delta w^T * \frac{\delta^2 E}{\delta w^2} * \delta w + O(\| \delta w^2 \|^3) \quad (1.4)$$

Where $\frac{\delta^2 E}{\delta w^2}$ forms Hessian H.

In case weights are in the minimum of function error we omit first term of the sum on the right side of equation and the third summand which results in equation reduction to the following form:

$$\delta E = \frac{1}{2}\sum_i^m H_{ii}\delta w_i^2 \quad (1.5)$$

When the i weight is eliminated we obtain $w_i = \delta w_i$ and this can indicate significance coefficients of the Si for each weight

$$S_i = H_{ii} w_i^2 \quad (1.6)$$

However, while learning the situation can happen where some weight or neurons stop to play important roles. Such circumstances cause that the model must proceed in adaptation of such neurons despite the fact that they are suitable and can be the reason of the model over-learning. The increase and reduction of model's structure in learning process causes the situation that a model can be increased and reduced as needed with part of the structure elimination or replacement with smaller one. The algorithm of proper verification of neurons should be defined to increase and reduce network architecture by proper selection of learning model's complexity to the model's data loading. With the assumption that density function and described neurons of hidden layer are similar, however significance coefficients of neurons of hidden value are defined as the ration of weight amount to the weight variation [7]. The relation favor only strong weights this is such which have significant influence. Defined coefficients are written with the formula:

$$S_i = \frac{w_i^2}{\delta w_i} \quad (1.7)$$

We can eliminate only such neurons where significance coefficients are the lowest and are written with the formula:

$$L_i = \min_i S_i = \min_i \frac{w_i^2}{\delta w_i} \quad (1.8)$$

Equation 1.8 can be reduced to the other form:

$$L_i = \min_i S_i = \min_i \frac{w_i^2}{[Pw]_{ii}} \quad (1.9)$$

For the purpose of 1.9 dependences one must use expanded Kalman filter as the estimator of model's parameters to determine δw_i variance with covariance matrix P_n, which is as follows:

$$P = \begin{bmatrix} P_w & P_{wv} \\ P_{wv} & P_v \end{bmatrix} \quad (2.0)$$

where P_W is a covariance matrix between weight and P_{WV} this is covariance matrix between weight and other parameters but P_V is a covariance matrix between other parameters. Decision on neurons elimination brings to the criterion:

$$\frac{L_i}{R_y} < X^2_{i,\upsilon} \quad (2.1)$$

where $x^2_{n,u}$ is the chi-aquare distribution with trust level equal to **u**%

$$R_y = R_n + d_n^T P_{n-1} d_n \quad (2.2)$$

If the criterion is met the neuron which obtained the lowest significance coefficient should be eliminated.

Evaluation of the efficiency of expert system operation

To increase the efficiency of the expert system in estimating the type of illness it needs an enormous database. The base, which is used to study the neural network consists of hundreds of records introduced by the doctors who co-operated with the authors. Additionally, on the basis of experts' (doctors') knowledge and experience the record set has been generated. This set has been used to test the system.

In the table 1 presented the comparison of the efficiency of expert system operation including diagnoses formulated by the experienced doctor.

Table 1. Efficiency of network operation

	Sensitivity	Specificity	Precision
Doctor	88%	92%	89%
Network	80%	82%	82%
Networks and a doctor	92%	95%	92%

Comparing the efficiency of neuron network operation and doctor's diagnosis we can notice that networks deal with quality evaluation of a patient. It can be noticed that difference between doctor's diagnosis and expert system diagnosis amounts to few percents. Better result can be obtained if the doctor in the initial phase of the diagnosis uses expert system to formulate own diagnosis.

Summary

Today's advanced progress can be observed not only in techniques but also in science and medicine. Current database have gigabytes of data and include hidden information of significant value. This fast increase in size of database caused limitations in analysis and interpretation of collected data. It is harder now to analyze efficiently big amount of data and results of various tests determining the patient health.

That is why, expert systems based on computer science technique are formed and become helpful for doctors as can use appreciable sets of current and archive databases as well as present deviations of proper values. Computer science applications are designed and are more often applied tool for doctors in formulation a diagnosis based on provided data. The SpineMedical application is intended for doctors and medicine students and patients as well. In the future the application can have larger scope of application for example for the purpose of analyzing and medical imaging processing like X-rays and computer tomography images enabling to more accurate determination of diseases changes in the backbone.

Bibliography

1. Osowski S.: Sieci neuronowe do przetwarzania informacji, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2000.
2. Cottrell M., Girard B., Girard Y., Mangeas M., Muller C.: Neural modeling for time series: a statistical stepwise method for weight elimination, IEEE Transaction on Neural Networks IEEE Transaction on Neural Networks, 1995.
3. Cichosz P.: Systemy uczące się, Wydawnictwa Naukowo-Techniczne, Warszawa 2000.
4. Tadeusiewicz R., Lula P.: Wprowadzenie do sieci neuronowych, StatSoft & C.H.Beck, Kraków 2001.
5. Rąpała K.: Zespoły bólowe kręgosłupa. Zagadnienia wybrane, Warszawa 2004.
6. Stodolny J.: Choroba przeciążeniowa kręgosłupa, Kielce 2004.
7. Le Cun Y., Denker J., Solla S., Kauffman M.: Optimal brain damage: Advances in Neural Information Processing Systems 2, San Mateo 1990.
8. Duch W., Korbicz J., Rutkowski L., Tadeusiewicz R.: Sieci neuronowe, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2000.
9. Bubnicki Z.: Wstęp do systemów ekspertowych, Wydawnictwo Naukowe PWN, Warszawa 1990.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 45–49

SPEECH PERCEPTION – TOWARD UNDERSTANDING OF CONSCIOUSNESS

JAN TRĄBKA, PIOTR WALECKI, WOJCIECH LASOŃ, WIESŁAW PYRCZAK, KRZYSZTOF SARAPATA

Department of Bioinformatics and Telemedicine, Jagiellonian University, Medical College, Kopernika 7, 31-034 Krakow

Abstract. This paper presents a project executing within the framework of COST Action BM0605: Consciousness: A Transdisciplinary, Integrated Approach. The research is centred on neurophysiological experiments to image brain function in diseases in which the phenomenon of consciousness plays an important role. The fundamental problem studied in this project will be an investigation of the relation between consciousness and speech perception. The study will be conducted using electroencephalographic methods on subjects with Auditory Processing Disorders.
Keywords: speech perception, consciousness, Auditory Processing Disorders, event-related potential

Introduction

The problem of consciousness is regarded as one of the fundamental problems in contemporary science. Understanding the mechanisms which contribute to the creation of states of consciousness such as perception, sensation, cognition and action requires a highly-specialized and interdisciplinary approach which combines research and discoveries from various branches of science (ranging from neuroscience and artificial intelligence to philosophy and psychology), from various experimental methods (such as behavioural observation, brain activity imaging or simulations and numerical methods) and from research of various populations (both animal and human).

The project description presented in this article pertains to research carried out by Working Group 3 (WG3) within the framework of COST Action BM0605 – *Consciousness: A Transdisciplinary, Integrated Approach.* The research is centred on neurophysiological experiments to image brain function in diseases in which the phenomenon of consciousness plays an important role. The fundamental problem studied in this project will be an investigation of the relation between consciousness and speech perception. The study will be conducted using electroencephalographic methods on subjects with Auditory Processing Disorders (APD).

APD is characterised by an inability to learn from auditory stimuli and a difficulty in understanding speech in the following conditions: a) around normal auditory thresholds, b) in poor acoustic conditions, c) distorted or unclear speech.

The studies are based upon an analysis of elicited exogenous and endogenous potentials. Measurements are made both of phonological capabilities such as auditory memory volume and sequence and temporal discrimination and of other behavioural phenomenon that are responsible for central auditory processing such as sound localization and lateralization, sound differentiation, auditory pattern recognition, temporal aspects of hearing (discrimination, masking, integration, organization), the capacity to distinguish competing acoustic signals and the capacity to recognize degraded acoustic signals.

The results of the study are relevant for the development of effective therapeutic strategies for patients diagnosed with APD. In addition to medical intervention (pharmacotherapy, surgical procedures) these strategies will include auditory reception training (e.g. computer games which strengthen or modify temporal concentration disorders in children), compensatory techniques (strengthening receptive reactions and improvement of capacities such as auditory discrimination and analysis, phoneme synthesis, auditory memory, hearing in noise, temporal processing) and cognitive training (teaching how to actively monitor and autoregulate one's ability to understand the speech, linguistic and metalinguistic training).

The results of the study also contribute to the formulation of a widely accepted definition of APD, to the creation of a battery of tests for the diagnosis of this disorder and to the development of therapy guidelines. A measurable effect of this project which is related directly to COST Action BM0605 is the increase of knowledge about the relation between speech perception and consciousness. The application in this research of objective tests to measure specific physiological parameters will allow us to study processes related to conscious data processing and of their influence on particular cognitive process disorders.

Neuropsychological characterization of APD patients – audiological tests and evoked potentials analysis

During the first stage of the study the characteristics and degree of severity of APD in the studied patients will be determined. Research methods employed during this stage are: interview, observation of behavioural responses to auditory stimuli, audiological tests: audiometry (cortical audiometry, evoked response audiometry, pure tone liminal audiometry, supraliminal tonal audiometry, speech audiometry, vocal audiometry), temporal process assessment tests, localization and lateralization assessment tests, monaural low-redundancy speech comprehension, comprehension of binaurally separated stimuli, binaural audition interaction, and speech-language pathology tests. The research tools used in these studies will use otoacoustic emission test (OAE): Spontaneous Otoacoustic Emissions (SOAE), Transiently Evoked Otoacoustic Emission (TEOAE) and Distortion-Product Otoacoustic Emission (DPOAE), and exogenous and endogenous components of auditory evoked potentials (AEP): Auditory Brainstem Response (ABR) and Medial Latency Response (MLR). The possibility of peripheral auditory system damage, for example conductive or neurosensory damage, is eliminated during this stage (using tympanometry and tonal audiometry for frequencies in the range of 250 to 8000 Hz measured in octave intervals).

The ABR measurement is an important element of the study which consists in a series of neurological responses that successively image the activity of the auditory nerve and of the neural fibres and nuclei lying on each level of the auditory pathway. Wave peaks labelled with the Roman numerals from I to VII are analyzed. They reflect bridge-interbrain conduction in the nerve fibres and provide a good measurement of central auditory processing on the level of the brain stem. Numerous studies have confirmed the diagnostic usefulness of the ABR test in different groups of patients. The correlations between results of the Staggered Spondaic Word Test (SSW) test and MLR disorders have also been established. These studies are carried out using sensitized speech verbal tests in which certain verbal stimuli have been deformed in such a way as to reduce intelligibility. These include tests of filtered, interrupted and time-compressed speech perception, a binaural filtered-speech signal composition test using both low and high-pass filters turned on alternatingly or increasingly/decreasingly and dichotic CV and spondaic tests. The basic assumption behind these tests is the thesis that a person with healthy audition and without central auditory pathway disorders will be able to understand distorted speech, but that if there are disorders then the comprehension will be worsened. Because both central auditory system factors and peripheral auditory system factors can influence distorted speech comprehension, an assessment of auditory thresholds (sensitivity to different amplitude and frequency stimuli) is made before interpreting the results of the central auditory system tests.

Cognitive event-related potential (CERP) assessment in APD patients

The second stage of the study takes advantage of cognitive event-related potential (CERP) advanced mathematical methods (see Fig. 1). Relation to APD is analysed and we determine the roles of cognitive components, such as the P300 wave, which are related to concentration, of components related to semantic analysis such as the N400 wave and of selective attention indicators such as the late Contingent Negative Variation (CNV) wave. The Mismatch Negativity (MMN) wave is also analysed. This expresses automatic brain activity connected to the perception of a difference between a distinguished auditory stimulus and a series of preceding identical standard auditory stimuli.

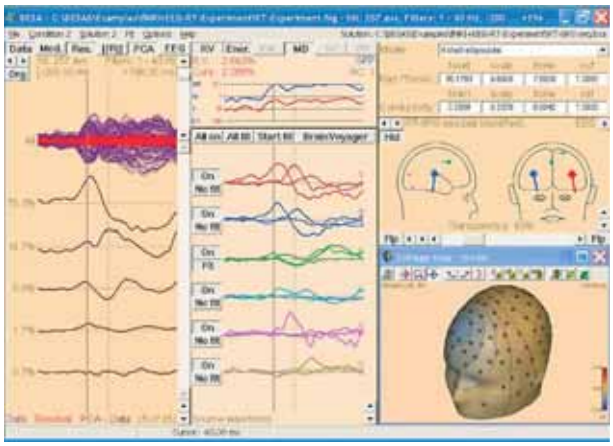
This part of the project primarily analyses endogenous potentials, which are an expression of the cognitive or emotional reaction to a stimulus, to a change of its parameters or to an unexpected lack of the stimulus. P3b is the name given to the wave which is most clearly registered in the central-parietal leads when two stimuli are discriminated though it is more often referred to as the P300 potential because when auditory stimuli are used, this wave occurs at a latency of about 300-350 ms.

The analysis and study of the connection of this potential with APD is important because it arises in a situation when the stimulus is unexpected or when it carries new and important information. Then, the latency is a measure of the time dedicated to the processing of the stimulus (decoding, recognition, classification) and the amplitude reflects the size of engaged cognitive structures (the wave itself arises at the moment when the cognition problem is solved). Factors which impact the amplitude and latency of the P300 wave are: the patient's state of consciousness, the type of task presented to the patient during registration of the signal, concentration of attention (motivation to complete the task) and meaning of the stimuli for the patient.

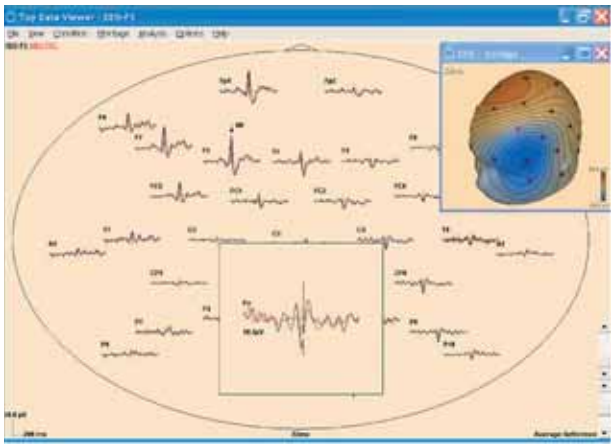
The semantic potential N400 will also be analysed during the study. This potential only arises when sentences are presented in which the last word does not fit into the preceding context. The signal will be registered while words or sentences are presented aloud. Sensitized speech tests will also be applied in this case. The main experimental part has been planned assuming a tonotopic organization of the first segment of the auditory pathway (precisely determined map of the preferred stimulus frequencies).

The activity of the primary auditor cortex (A1) which is located at the centre of the superior temporal gyrus (Brodmann areas 41 and 42) will be mapped. The registration of the EEG signal and of elicited potentials, especially cognitive potentials, from specific brain generators (small areas or centres of the cerebral cortex such as the primary auditory cortex A1) will require very precise placing of the electrodes and an exact determination of their location.

A specialized digitizer (Polhemus) has been employed in order to overcome the problem of placing the electrodes on the patient's head as exactly as possible. Thanks to this device, it is possible to determine points in three-dimensional Cartesian space with an accuracy of 0.025 angular degrees. The measurements of the coordinates of the electrodes on the patient's head will be imported into the EGG signal analysis software.

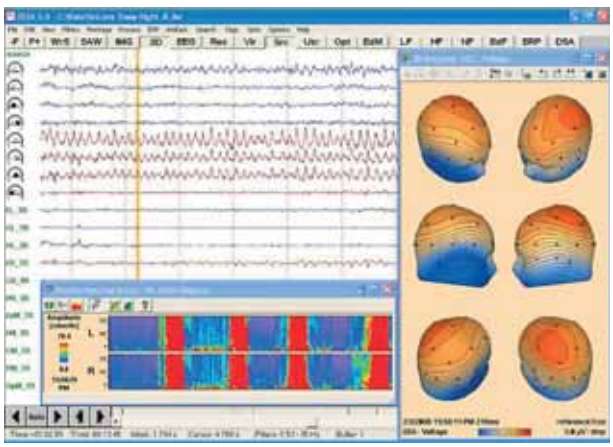


A



B

C



D

Fig. 1. Megis BESA (Brain Electrical Source Analysis). A. Source waveforms separate the activities in auditory, visual and motor cortex in a reaction time experiment. B. Averaged frontal spike in top view. Peaks can be automatically analyzed. C. Automated artifact scan – 2D artifact scanning tool for fast decision on bad channels and sweeps prior to averaging. D. EEG review – 3D whole-head maps and hemispheric comparison of density spectral arrays (DSA). Source: www.besa.de

Speech perception model assessment based on invariant signal characteristics

During the project speech perception models based on invariant signal characteristics will be verified. Initial speech signal processing which includes filtration, suppression, adaptation and phase synchronization will be studied as well as an analysis of the functioning detectors of acoustic properties (turning on, spectral changes, formant frequency and periodicity) and of phonetic characteristics (sonority or nasality).

The last stage of this process is segment analysis and lingual search. The research methods employed during this stage are the short-time Fourier transform (STFT) and Daubechies discrete wavelet transform (DWT). The FT calculated for the time period of the ERP elicited potentials will show us the spectral complexity of the signal. An ERP consists of suppressed oscillations so knowledge of the spectral components gives only a partial understanding and does not permit one to reconstruct the signal.

The application of short-time Fourier transform allows us to observe the changes in the spectrum over time and thus to obtain the frequency-time structure of the course, but the resolution of this method is low and limited by the length of the transformation window. Therefore the signal will be analysed using wavelet transform (the analysis program will be implemented in the MATLAB environment). In this way we will obtain a time resolved spectrum of the signal with a resolution higher than in the case of short-time Fourier transform. DWT also permits time-localization of a sought pattern.

The underlying assumption behind this analysis is that it is possible to find a relatively invariable relation between the course of acoustic signals and the perception of speech sounds which in turn relates to the possibility of obtaining relatively stable patterns of neuronal activity corresponding to specific acoustic signal patterns.

Conclusions

The presented project is carried out in the Department of Bioinformatics and Telemedicine of the Jagiellonian University Medical College. The project is a continuation and development of European programs COST Action B27: *Electric Neuronal Oscillations and Cognition* (ENOC) and COST Action BM0601: *Advanced Methods for the Estimate of Human Brain Activity and Conectivity* (NeuroMath) which were previously carried out by the same team.

The results of the study will have a practical significance (contribution to the development of improved diagnostic and therapeutic techniques for auditory processing disorders) and a theoretical significance (increasing our knowledge about the role of consciousness in speech perception). The study of phenomenon related to speech perception contributes to our understanding of the nature of consciousness and its function. Both consciousness disorders and auditory processing are related to each other although the essence of this relation has not been sufficiently understood yet. The studies in the presented research project based on measurements of neurophysiological parameters will serve to verify speech perception models. This is the approach which is suggested by the organizers of the COST BM0605 project because with the enormous amount of speculative approaches to consciousness, there are few objective neurophysiological studies at documenting its role. The process of speech perception is a particularly crucial mechanism because it is connected to consciousness, its origin and function.

The application of precise methods of neuronal activity measurement and advanced mathematical methods of signal analysis as well as the use of high-power computing devices makes it possible to gain new knowledge which can then be used in effective medical intervention. APD's represent a pathological sector which is not sufficiently understood nor treated. Therefore, the precise determination of a widely accepted definition of APD and the development of effective diagnostic and therapeutic methods is so important. The results of the study will contribute to the development of effective therapeutic strategies for patients diagnosed with APD. In addition to medical intervention (pharmacotherapy, surgical procedures) these strategies will include auditory reception training (e.g. computer games which strengthen or modify temporal concentration disorders in children), compensatory techniques (strengthening receptive reactions and improvement of capacities such as auditory discrimination and analysis, phoneme synthesis, auditory memory, hearing in noise, temporal processing) and cognitive training (teaching how to actively monitor and autoregulate one's ability to understand speech, linguistic and metalinguistic training).

This area of research also has particular significance in the context of Poland. Because there so few centres specializing in speech perception disorders, this project can contribute to increasing general awareness of this problem and to activate environments dedicated to helping people suffering with APD.

References

1. Barlow J. S., Trąbka J., The relationship between photic driving in the EEG and responses to single flashes, Fifth International Congress of Electroencephalography and Clinical Neurophysiology, Rome, Italy, 7-13 Sept. 1961, Exc. Med. Int. Congr. Ser., No. 37, 182.
2. Chermak G. D., Musiek F. E., Managing central auditory processing disorders in children and youth, American Journal of Audiology, 1, 1992: 61-65.
3. Cole R., Jakimik L., A model of speech perception, in: Cole R. (Ed.), Perception and production of fluent speech, Erlbaum, Hillsdale, NJ, 1979, pp. 133-160.
4. Colson K., Robin D., Luschei E., Auditory processing and sequential pitch and timing changes following frontal opercular damage, Clinical Aphasiology, 20, 1991: 317-325.
5. Craig C. H., Kim B. W., Rhyner P. M., Chirillo T. K. B., Effects of word predictability, child development, and aging on time-gated speech recognition performance, Journal of Speech and Hearing Research, 36, 1993: 832-841.
6. Jaśkowski P., Verleger R., Amplitudes and latencies of single-trial ERP estimated by maximum likelihood method, IEEE Transactions on Medical Engineering, 46, 1999: 987-993.
7. Jaśkowski P., Verleger R., An evaluation of methods for single-trial estimation of P3 latency, Psychophysiology, 37, 2000: 153-162.
8. Jerger J., Johnson K., Jerger S., Coker N., Pirozzolo R., Gray L., Central auditory processing disorder: A case study, Journal of the American Academy of Audiology, 2, 1991: 36-54.
9. Jirsa R. E., Clontz K. B., Long latency auditory event-related potentials from children with auditory processing disorders, Ear and Hearing, 11, 1990: 222-232.
10. Jirsa R. E., The clinical utility of the P3 AERP in children with auditory processing disorders, Journal of Speech and Hearing Research, 35, 1992: 903-912.
11. Keith R. W. (Ed.), Central auditory and language disorders in children, San Diego: College-Hill, 1981.
12. Sęk A., Auditory filtering at low frequencies, Archives of Acoustics, 25, 2000: 291-316.
13. Sęk A., Moore B. C. J., Detection of quasitrapezoidal frequency and amplitude modulation, J. Acoust. Soc. Am., 107, 2000: 1598-1604.
14. Sęk A., Moore B. C. J., Testing the concept of a modulation filter bank: The audibility of component modulation and detection of phase change in three-component modulators, J. Acoust. Soc. Am., 113, 2003: 2801-2811.
15. Skrodzka E. B., Sęk A. P., Application of BEM to modeling loudspeaker's directivity patterns based on its dynamic behavior, Archives of Acoustics, 26, 2001: 75-91.
16. Szczuka M., Wojdyłło P., Neuro-wavelet classifiersfor EEG signals based on rough set methods, Neurocomputing, 36, 2001: 103-122.
17. Trąbka J., et al., EEG Signals Described by the Automatic Linguistic Analysis, w: Rother M., Zwiener U., Quantitative EEG Analysis, Univ. Jena, 1993, 114-117.
18. Trąbka J., Przewłocki R., Siuta J., The influence of topical administration of the carboline derivatives on direct cortical response (DCR), Diss. Pharm. Pharmacol., 1969, 6, 515-522.
19. Trąbka J., Sekuła J., Fenczyn J., Warchołek J., Efekt Bezold-Bruckego w obrazie uśrednionych słuchowych odpowiedzi wywołanych, The Bezold-Brucke effect in the pattern of averaged auditory evoked responses, Otolaryng. Pol., 1975, 29, 1.

20. Trąbka J., Badania EEG u dzieci z zaburzeniami ortostatycznymi, EEG examination in the children with orthostatic disturbances, Pediatr. Pol., 1965, 40, 1333-1337.
21. Trąbka J., Behavioral and EEG changes caused by the substituted derivatives of the gamma-butyrolacton, Abstracts First Meeting of the German Neuropharmacolog. Society, Magdeburg – GFR, 1968.
22. Trąbka J., Easy making own EMG glossary and knowledgebase – is it possible? IX International Congress of Electromyography and Cl. Neuroph., Jerusalem, Israel, 1992, 132-132.
23. Trąbka J., EEG observations of the alterations of consciousness, Electroenceph. Clin. Neurophysiol., 1959, 11, 175.
24. Trąbka J., Electrophysiological approach to the problem of brain hemisphere asymetry, EEG Abstracts 6th International Congress of EEG, Vienna, Austria, 1965, Elsevier, 307-310.
25. Trąbka J., High frequency components in brain wave activity, Electroenceph. Clin. Neurophysiol., 1962, 14, 453-464.
26. Trąbka J., Steering function of the consciousness in the language decoding process, Proceedings of the First International Aphasia Rehabilitation Congress, 1990, 29-35.
27. Trąbka W., Hamuda G., Trąbka J., The desing kind and order of stochastic model for EEG signals, XII International Congress of Electroenceph. and Clinical Neuroph., Rio de Janeiro, Brazil, 1990.
28. Trąbka W., Stanuch H., Trąbka J., Automatic analysis of the evoked potentials using harmonical functions, XIIIth Annual Joint Meeting of Electroenceph. and Clinical Neuroph., Prague, Czechoslovakia, 1990.
29. Trąbka W., Trąbka J., Fractal Consciousness, Third IBRO World Congress of Neuroscience, 1991.
30. Trąbka J., Walecki P., Sarapata K., Pyrczak W., Roterman-Konieczna I., Percepcja mowy – analiza potencjałów wywołanych w zaburzeniach procesów przetwarzania słuchowego (Speech perception – analysis of event-related potential in Central Auditory Processing Disorders), EPISTEME, 7/2008, s. 83-94.
31. Wróbel A., Beta activity: a carrier for visual attention, Acta Neurobiol. Exp., 60, 2000: 247-260.
32. Wróbel A., Kublik E., Modification of evoked potentials in the rat's barrel cortex induced by conditioning stimuli, in: Kossut M. (Ed.), Barrel Cortex, Graham Publ. Corp., New York, 2000, pp. 229-239.
33. Wróbel A., Kublik E., Musiał P., Gating of sensory activity within barrel cortex of the awake rat, Exp. Brain Res., 123, 1998: 117-123.
34. Wypych M., Kublik E., Wojdyłło P., Wróbel A., Sorting functional classes of evoked potentials by wavelets, Neuroinformatics, 1, 2003: 193-202.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 51–58

NEUROINFORMATIC MODELLING OF OCULOMOTOR SYSTEM

PIOTR WALECKI

Department of Bioinformatics and Telemedicine, Jagiellonian University, Medical College, Kopernika 7, 31-034 Krakow, e-mail: pwalecki@cm-uj.krakow.pl

Abstract. Neuroinformatics is a new branch of science that uses informatics tools to modelling some parts of neural system. This paper presents several models related to oculomotor system, both educational and research models. Most of the models presented in this paper are our own work. The educational models created for medical purposes demonstrate interactively the effects of damage to different neuronal structures and also simulate disorders of eyeball movement dynamics and synchronization in various diseases. The research models may find application both in clinical tests and in experimental study.
Keywords: oculomotor system, modelling, neuroinformatics

Introduction

Using measurements and modelling of eyeball movement we can very precisely analyse one of the most important neurobiological mechanisms of living beings: sensory-motor integration [16]. The oculomotor system possesses a small and well defined set of degrees of freedom in comparison to other motor systems [1], [2]. Thus it seems to be particularly well suited for modelling neuronal function.

Although human behaviour is characterized by an immense complexity far beyond that found in other organisms, this behaviour is nevertheless often realized through simple sensory-motor systems which have a relatively simple construction [17]. An analysis of the construction of sensory and motor systems thus contributes to a better understanding of the function of more complex structures of the central nervous system [15], [18].

The knowledge provided to us by an understanding of the cooperation of sensory-motor systems in processing sensory stimuli and behavioural expression finds an application in neuropsychology and medical diagnostics [1]. Eyeball movement irregularities have been observed in neurological disorders connected to central nervous system or cranial nerve damage [2] as well as in the case of patients suffering from schizophrenia [3], autism [4], Parkinson's disease [5], dyslexia [6] or ADHD [7].

Material and methods

We presently possess a large set of neurobiological data from experimental studies. Thanks to this we can model the oculomotor system with varying degrees of accuracy, distinguishing a certain function depending on the goal of the study. Even a fragment of the nervous system, like the oculomotor system, is so complex, that it would be difficult to model it in its entirety [18]. Thus, it is usually more advantageous to pick for modelling only certain subsystems or neuroanatomic structures which are related to a specific function. In this way, based on state-of-art advances in neurophysiology, we aim to model the functioning of the oculomotor system on the level of cellular structure [14], [15]. We also have the possibility of modelling only those characteristics which pertain to a specific function.

In choosing the right model, it is important determine for whom and based upon which branch of science the given model is created. This is due to the fact that, for example, a physician requires different information from the model than a neurobiologist. A separate group are educational models which place less emphasis on numerical calculations and are more focused on illustratively transmitting specific content.

Most of the models presented in this work are our own models which reflect specific neurophysiological functions or neuroanatomic structures under various aspects and with varying amount of detail. A specific type of model are the educational models created for medical purposes (see Fig. 1-3). These models interactively and dynamically demonstrate the effects of damage to different neuronal structures and also simulate disorders of eyeball movement dynamics and synchronization in various diseases.

Fig. 1 shows a model which simulates palsy of different extraocular muscles and cranial nerves. The adaptation of neurological data [8] from this programme contributed to the creation of two versions of an eyeball movement disorder simulator. Both versions use a widely available open-source programming environment (JavaScript, HTML, PHP), thanks to which they are suitable for further development and for running on different operating systems.

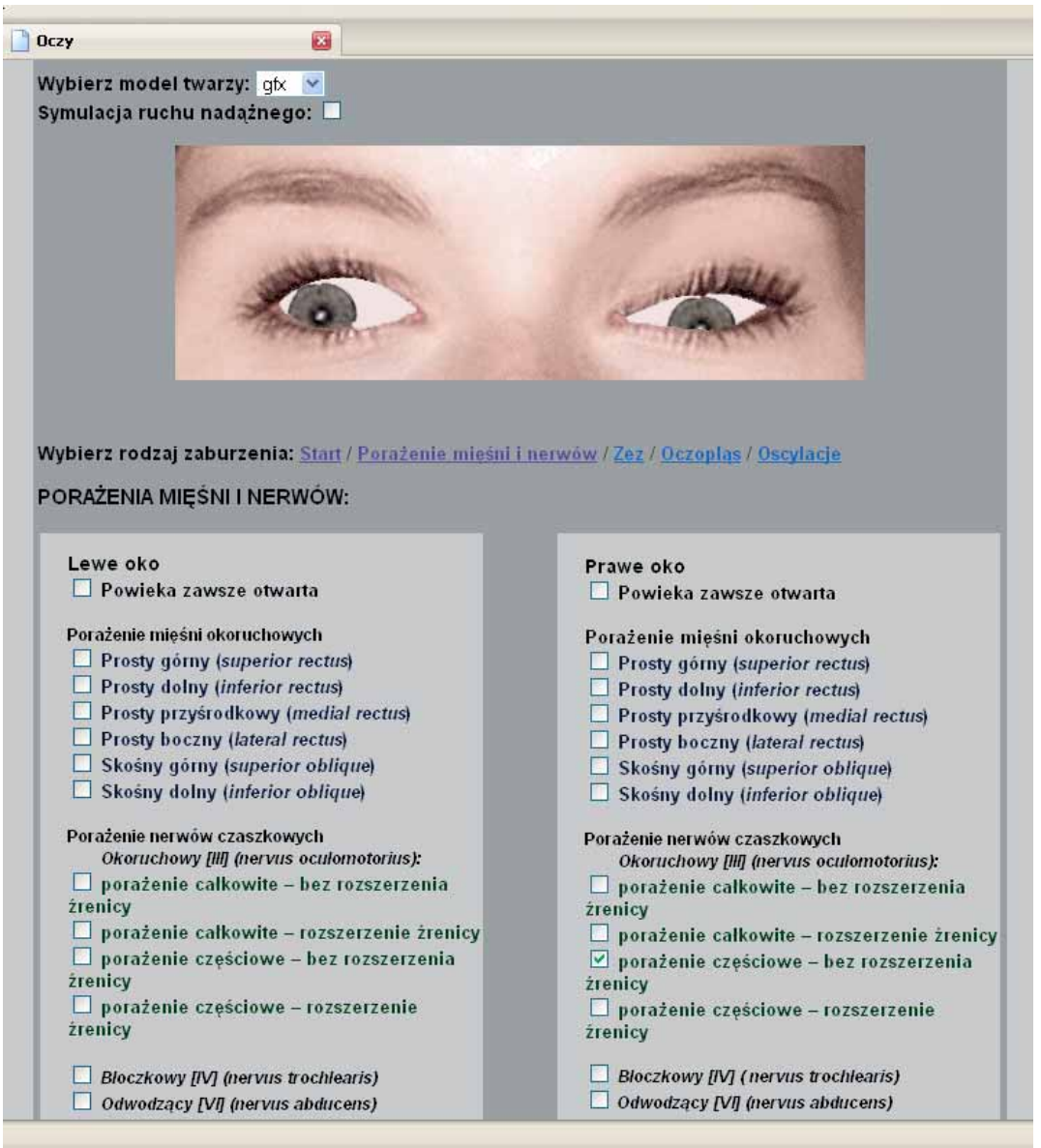


Fig. 1. The model used in simulation of palsy of different extraocular muscles and cranial nerves. Executors of the project: Jakub Baka, Piotr Olchawski. Source: www.neurobiologia.pl.

The model is meant for doctors who, while conducting medical examinations, encounter oculomotor disorders and abnormal functioning of the mechanisms responsible for motor control of the eyelid and pupil. The computer system simulates natural eyeball movement (movement is triggered by fixing eyesight on the mouse cursor) by means of an interactive animated patient. The patient's behaviour has been imaged using dynamically changing and interactive models of the face based on series of photographs which can be freely changed and new versions of which can be added. Below the photograph of the face is a panel in which one can choose (separately for each eye) from among various types of oculomotor damage.

The initial version of the model simulated movement disorders of the eyes, eyelids and pupils which occur in the case of palsy of the following oculomotor muscles. Later, the model was enriched with other types of eyeball movement disorders such as pathologic nystagmus, oscillations (see Fig. 2) and strabismus (see Fig. 3).

Modelling of eye movement

Modelling eye movement entails making an analysis of eyeball position and of changes in movement velocity in response to exterior stimuli or to an interior state. The precision of a model which reflects the real functioning of the oculomotor system has particular significance because changes in the values of eye location parameters occur on a millisecond timescale [2], [14].

Nevertheless, the processing of visual information in the cerebral cortex is a very complicated process. Moreover, eyeball movement, especially saccades, are elicited as a result of the engagement of many neuronal centres which are tied to cognitive, motivational and emotional processes and to attention and memory mechanisms [15], [16]. Therefore, despite the large amount of collected neuroanatomical and neurophysiological knowledge about the oculomotor system and despite the gigantic calculation capacities of computer systems, no attempt is made at the moment to create a holistic model of the

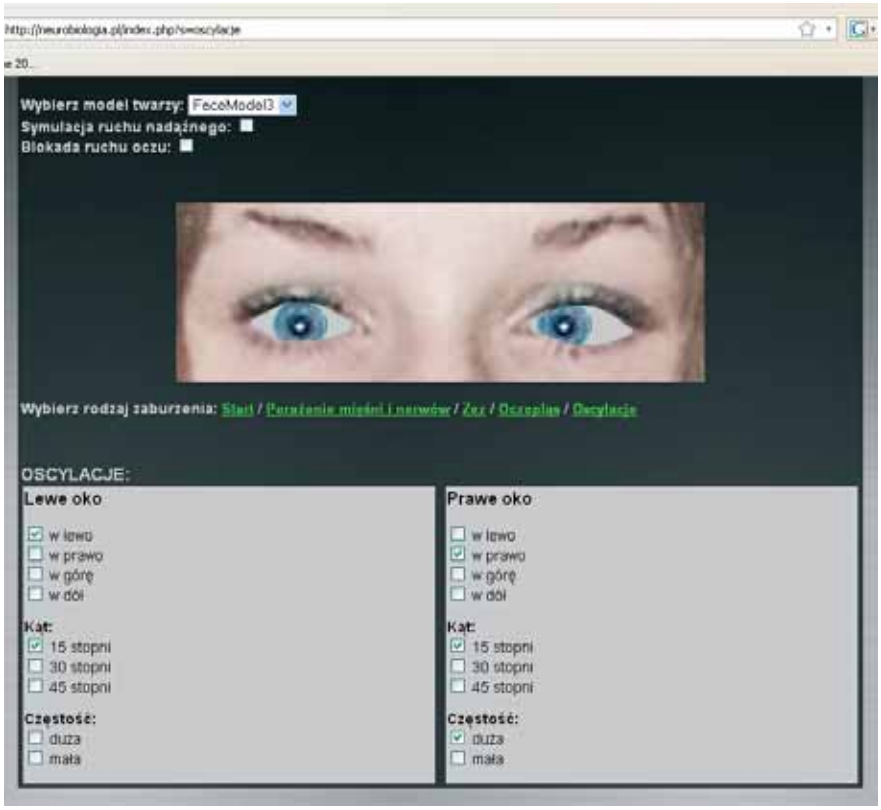


Fig. 2. The model used in simulation of pathologic nystagmus and oscillations.
Executors of the project: Jakub Baka, Piotr Olchawski. Source: www.neurobiologia.pl.

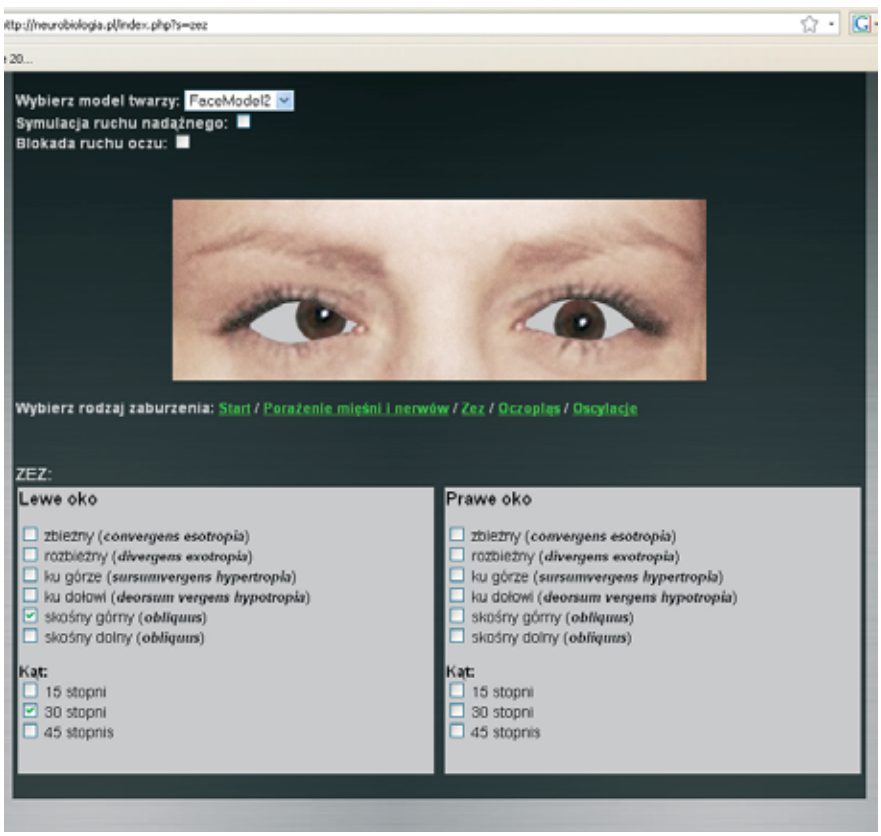


Fig. 3. The model used in simulation of strabismus.
Executors of the project: Jakub Baka, Piotr Olchawski. Source: www.neurobiologia.pl.

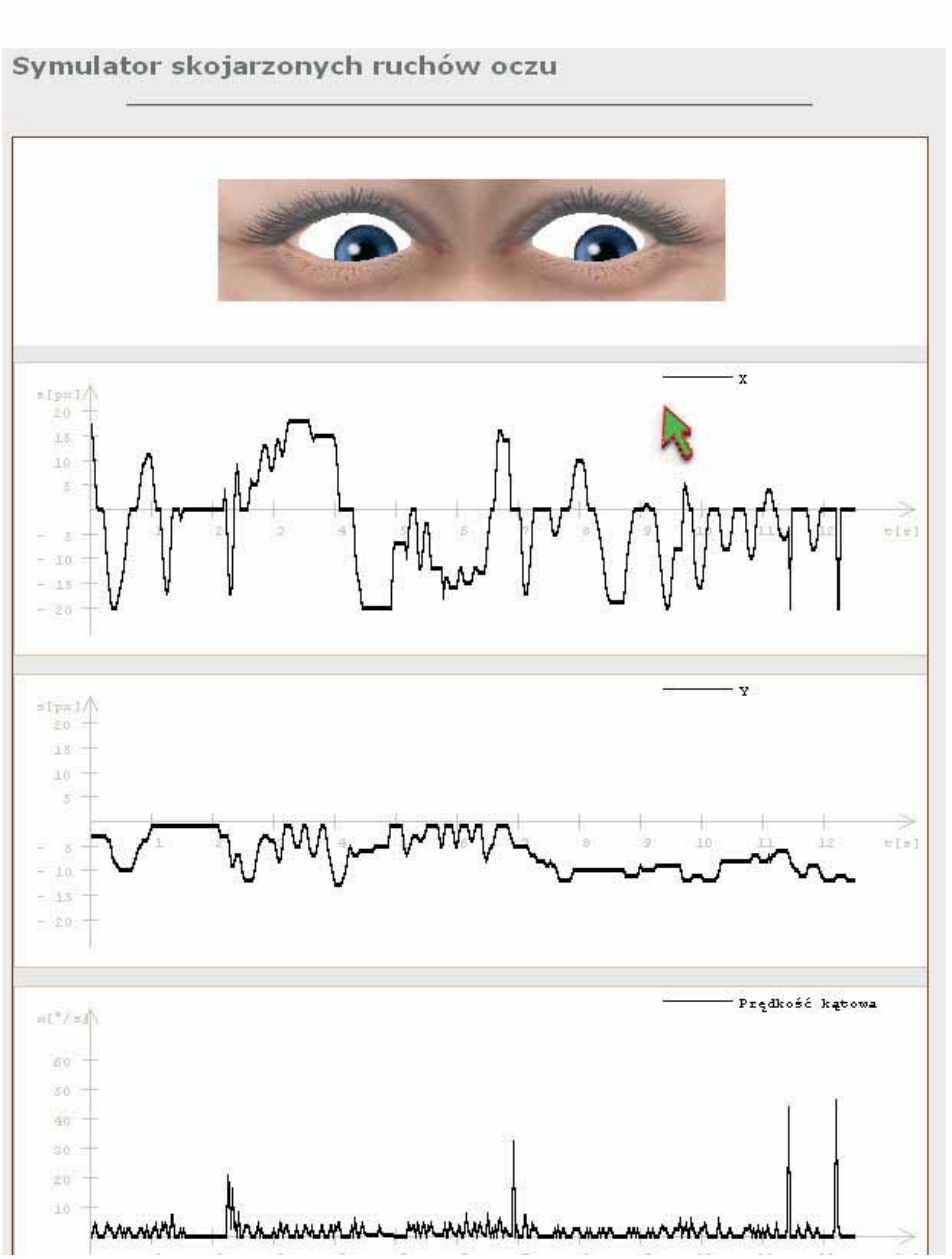


Fig. 4. The model used in processing of the virtual movement of eyes following the mouse cursor into data about their position on the x- and y-axis and which also calculates their angular velocity at a given moment. Executor of the project: Jan Paluch. Source: www.informatyka.cm-uj.krakow.pl.

oculomotor system which would faithfully represent the course of biological processes [17], [18]. It seems, however, that such an undertaking is to a large extent unnecessary, because every model is used only in order to explain certain aspects of a given system in accordance with accepted assumptions and applied methodology.

Fig. 4 shows a model which processes the virtual movement of eyes following the mouse cursor into data about their position on the x- and y-axis and which also calculates their angular velocity at a given moment. In the model, the calculation of the movement dynamics values is explicitly shown. The model was programmed in Java language.

Fig. 5 shows an actual recording of free eyeball movements in a studied person along with the eye velocity and x- and y-axis position calculations. The data was collected with the Jazz eye tracker (Ober Consulting LTD.).

Fig. 6 presents a model of saccade generator created in Simulink MATLAB that was implemented by Ansgar Koene [10]. The model was based on the *Model With Distributed Vectorial Premotor Bursters Accounts for the Component Stretching of Oblique Saccades* [9]. This model uses information form the most recent discoveries in neurobiology regarding how the saccade generator functions [19]. The initial stage of the model's construction was the creation of a block scheme to describe the dynamics of eyeball movement based on selected parameters (change in position and time). Using mathematical equations, this model simulates eye movement along a chosen trajectory and precisely describes changes in movement

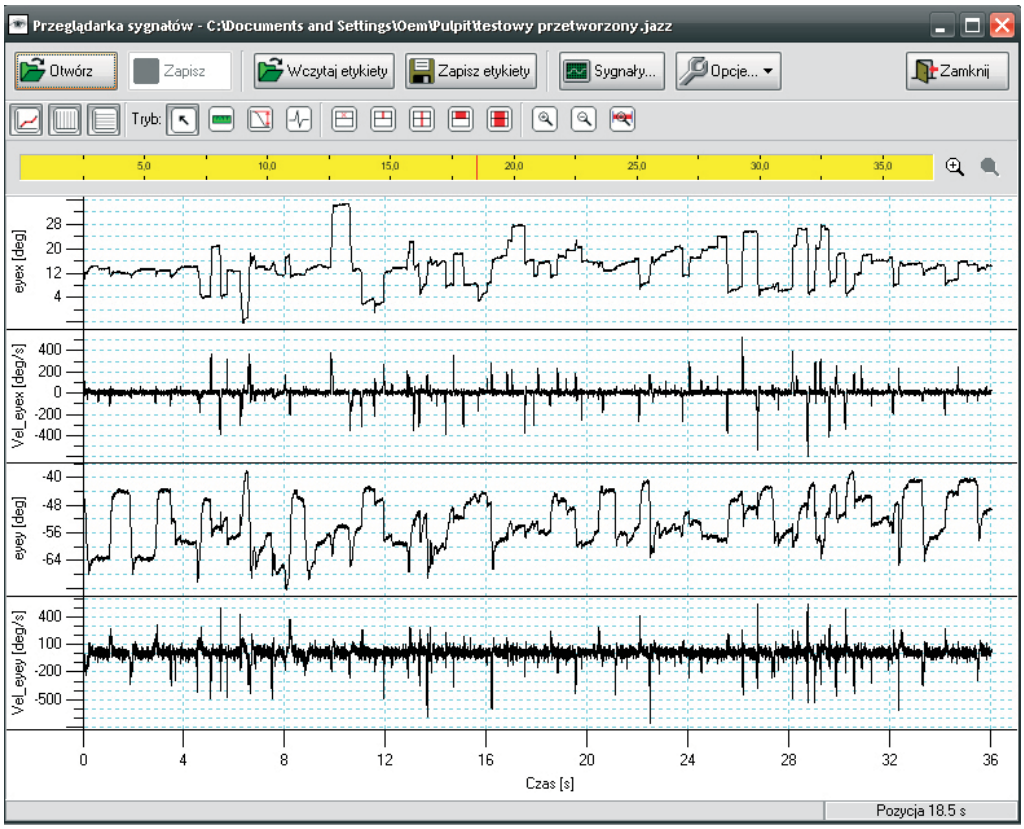


Fig. 5. The window of JazzManager program shows an actual recording of free eyeball movements in an examined person along with the eye velocity and x- and y-axis position calculations. The data was collected with the Jazz eye tracker (Ober Consulting LTD.).

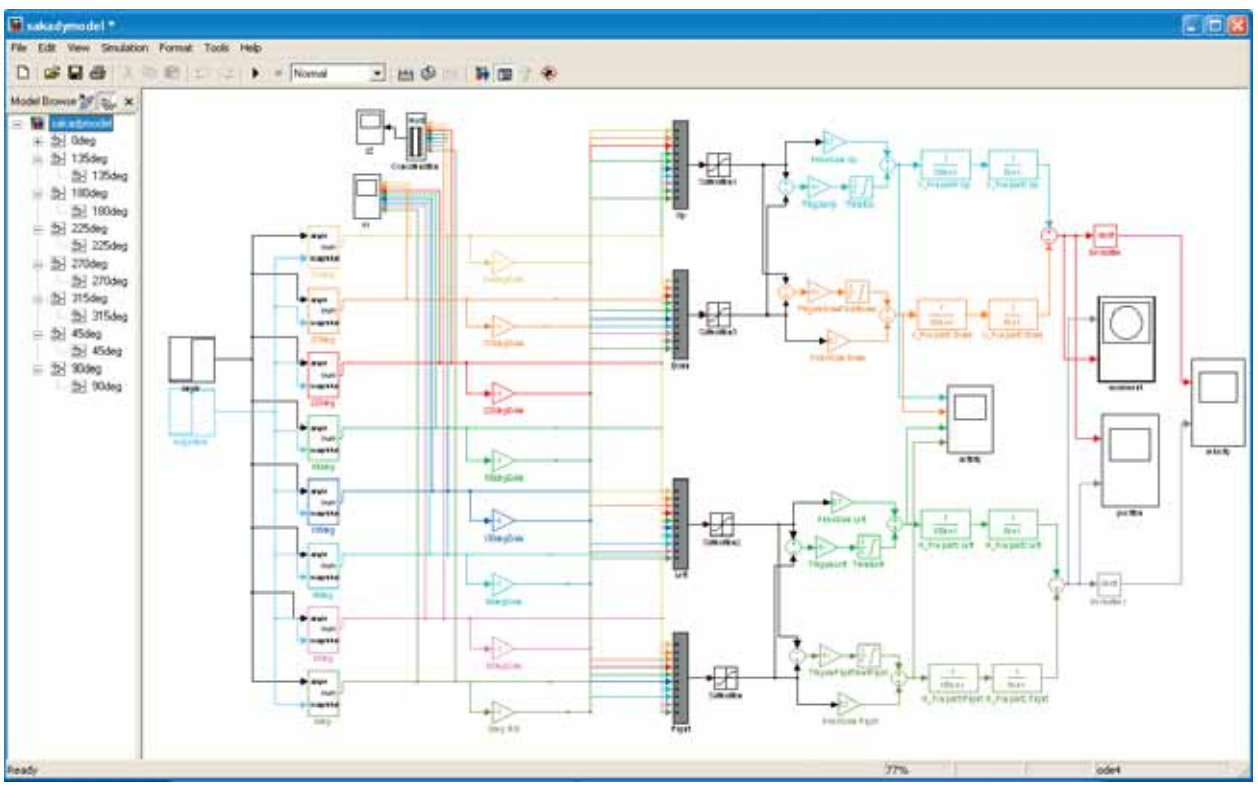
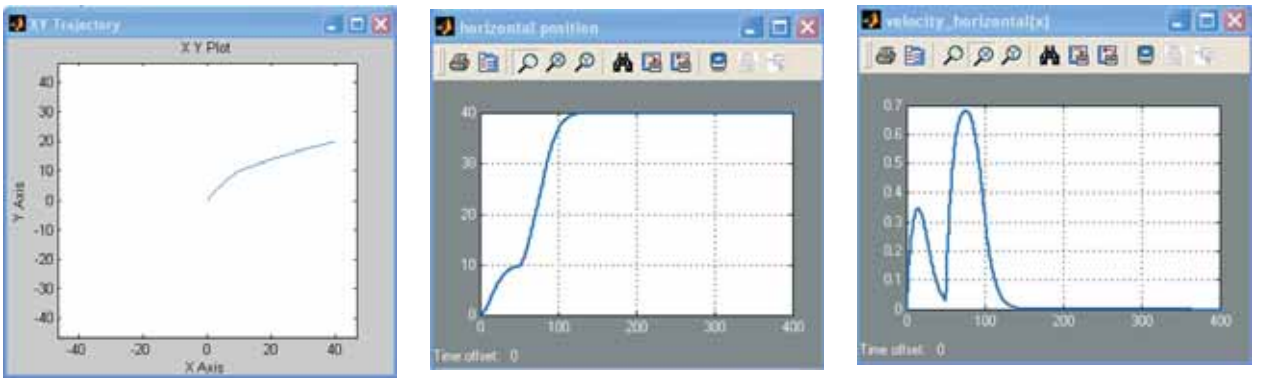


Fig. 6. The model of saccade generator created in Simulink MATLAB that was implemented by Ansgar Koene. Source: [10].



A B C

Fig. 7. The graphs present results of simulation. The model simulated an eye movement along a chosen trajectory (A) and precisely described changes in movement velocity (B) and position (C). Source: own research.

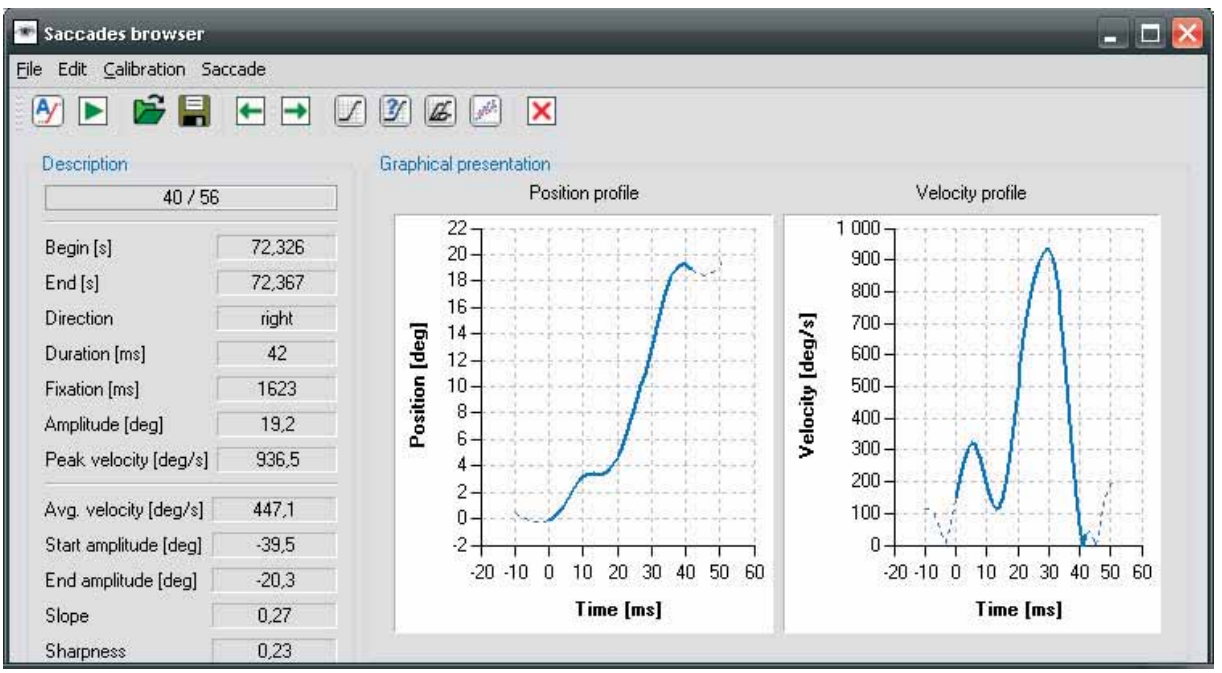


Fig. 8. The graphs present an actual recording of saccadic movement. The data was collected with the Jazz eye tracker (Ober Consulting LTD.). Source: own research.

velocity (see Fig. 7). The model was then tested by comparing it to actual measurements of saccade dynamics. The measurements were made using an eye tracker (see Fig. 8). After calibration it was confirmed that the values calculated in the simulation correlate to the experimental values. This model will find use both in clinical tests and in experimental research.

Modelling of oculomotor neuronal connections

By modelling neuronal connections we can display actual biological structure of the connections, or we can focus only on the function of a chosen system. However, regardless of which approach we choose, knowledge about the construction and function of the nervous system will be indispensable. Usually the choice of a given modelling method and of how detailed it will be is based on the amount of available neurobiological data. The less we know about a certain neuronal system, the more we will focus on its function. Sometimes we will completely abstract from the real biological construction and treat it as a so-called "black box" [17]. Such an approach is often used when modelling psyche or consciousness function [18].

A holistic diagram of neuronal connections has been created as part of the effort to model the oculomotor system (see Fig. 9) [2], [20]. The next step will be the construction of an interactive version of this diagram which will be used for educational purposes and also in the creation of functional models of selected neuronal functions and in the simulation of certain subsystems which participate in eyeball movement.

Knowledge about the neurophysiological basis of eyeball movement is indispensable for a doctor. It also provides important information for a neuropsychologist by demonstrating the

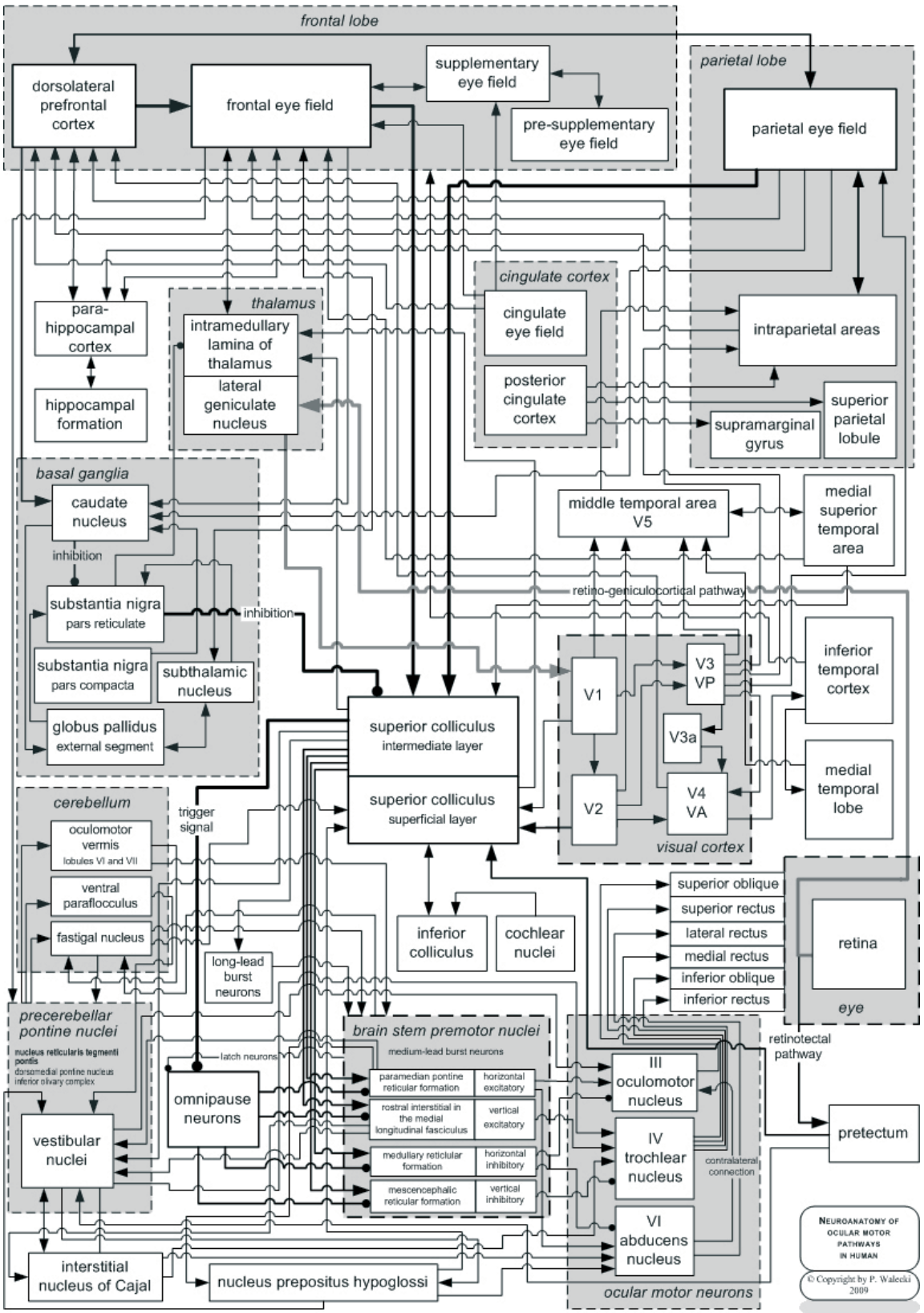


Fig. 9. The diagram of oculomotor neuronal pathways. Author: Piotr Walecki. Source of neurobiological data: [20].

basis of the eye attention mechanism and also for a neurobiologist by showing the evolutionary organizational structure of the various sensory-motor systems [11]. The visual and oculomotor systems are an excellent example of sensory-motor integration by means of which sensory signals are transformed into motor output [12]. The difficulty in studying eyeball movement lies in the dynamics of these movements because these events occur on a very short, millisecond timescale [2]. Nevertheless, the use of modern diagnostic equipment such as devices for measuring eyeball movement and of computers for data processing and analysis is contributing an increased dedication in this field of research [13]. Eyeball movement diagnostics are applied not only in medicine but also ever more frequently in industry and in the military where some movement characteristics contain key information for the success and effectiveness of certain actions [14].

Conclusion

The modelling of the oculomotor system is significantly contributing to a better understanding of the processes which take place in the nervous system. The oculomotor system models presented here develop these ideas on different planes. The utility of all of the mentioned models are currently being developed further and improved. The implementation of the model which simulates eye movement disorders is being applied for educational purposes in the formation of medical professionals. Further development of the model with the addition of eyeball movement disorders which occur in the case of patients suffering, for example, from schizophrenia or Parkinson's disease, is planned.

References

1. Karatekin C., 2007, Eye tracking studies of normative and atypical development, Developmental Review 27, 283-348.
2. Leigh J., Zee D., 2006, The neurology of eye movements, Oxford University Press.
3. Jacobsen L. K., Hong W. L., Hommer D. W., Hamburger S. D., Castellanos F. X., Frazier J. A., et al., 1996, Smooth pursuit eye movements in childhood-onset schizophrenia: Comparison with attention-deficit hyperactivity disorder and normal controls, Biological Psychiatry, 40, 1144-1154.
4. Holzman P. S., 2000, Eye movements and the search for the essence of schizophrenia, Brain Research Reviews, 31, 350-356.
5. Broerse A., Crawford T. J., den Boer J. A., 2001, Parsing cognition in schizophrenia using saccadic eye movements: A selective overview, Neuropsychologia, 39, 742-756.
6. Avila M. T., Hong E., Moates A., Turano K. A., Thaker G. K., 2006, Role of anticipation in schizophrenia-related pursuit initiation deficits, Journal of Neurophysiology, 95, 593-601. Levy D., Holzman R., Matthyse S., Mendeil N., 1993, Eye tracking dysfunction and schizophrenia: a critical perspective, Schizophr. Buli., 19, 3, 461- 536.
7. Dalton K. M., Nacewicz B. M., Johnstone T., Schaefer H. S., Gernsbacher M. A., Goldsmith H. H., et al., 2005, Gaze fixation and the neural circuitry of face processing in autism, Nature Neuroscience, 8, 519-526.
8. Hodgson T. L., Tiesman B., Owen A. M., Kennard C., 2002, Abnormal gaze strategies during problem solving in Parkinson's disease, Neuropsychologia, 40, 411-422.
9. Fischer B., Hartnegg K., 2000, Effects of visual training on saccade control in dyslexia, Perception, 29, 531-542.
10. Karatekin C., 2006, Improving antisaccade performance in adolescents with Attention-Deficit/Hyperactivity Disorder (ADHD), Experimental Brain Research, 174, 324-341.
11. Lasslo R., Henderson G., and Keltner J., Eye Simulator version 2.0, UC Davis School of Medicine, http://cim.ucdavis.edu/EyeRelease/
12. Quaia C., Optican L. M., 1997, Model With Distributed Vectorial Premotor Bursters Accounts for the Component Stretching of Oblique Saccades, J. Neurophysiol., 78: 1120-1134.
13. Ansgar Koene, http://arkoene.googlepages.com/matlabscripts
14. Krauzlis R. J., 2005, The Control of Voluntary Eye Movements: New Perspectives, Neuroscientist, 11(2): 124-137.
15. Becker W., 1989, The neurobiology of saccadic eye movements (eds Wurtz and Goldberg), Elsevier, Amsterdam.
16. Hoffman J. E. & Subramaniam B., 1995, The role of visual attention in saccadic eye movements, Perception and Psychophysics, 57, 787-795.
17. Walecki P., 2007, Neurofizjologia ruchu gałek ocznych (Neurophysiology of eye movement), Episteme, 4/2007, 159-176.
18. Kien J., McCrohan C. R., Winlow W., 1992, Neurobiology of Motor Programme Selection, Elsevier Science Pub Co.
19. Shadmehr R., Wise S. P., 2005, The Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning, The MIT Press.
20. Arbib M. A. (Ed.), Grethe J. S. (Ed.), 2001, Computing the Brain: A Guide to Neuroinformatics, Academic Press.
21. Moss F. (Ed.), Gielen S. (Ed.), 2001, Neuro-informatics and Neural Modelling, North Holland.
22. Scudder C. A., Kaneko C. S., Fuchs A. F., 2002, The brainstem burst generator a modern synthesis, Exp. Brain Res., 142, 439-62.
23. Becker W., 1989, The neurobiology of saccadic eye movements (eds Wurtz and Goldberg), Elsevier, Amsterdam.
24. Karatekin C., 2007, Eye tracking studies of normative and atypical development, Developmental Review 27, 283-348.
25. Krauzlis R. J., 2005, The Control of Voluntary Eye Movements: New Pers 38. pectives, Neuroscientists, 11(2): 124-137.
26. Missal M., Heinen S. J., 2004, Supplementary eye fields stimulation facilitates anticipatory pursuit, J. Neurophysiol, 92: 1257-62.
27. Schiller P. H., Chou I. H., 1998, The effects of frontal eye field and dorsomedial frontal cortex lesions on visually guided eye movements, Nat. Neurosci., 1: 248-53.
28. McPeek R. M., Keller E. L., 2004, Deficits in saccade target selection after inactivation of superior colliculus, Nat. Neurosci. 7: 757-63.
29. Sommer M. A., Tehovnik E.J., 1997, Reversible inactivation of macaque frontal eye field, Exp. Brain Res., 116: 229-49.
30. Sparks D., Rohrer W. H., Zhang Y., 2000, The role of the superior colliculus in saccade initiation: a study of express saccades and the gap effect, Vis. Res., 40: 2763-77.

BIO-ALGORITHMS AND MED-SYSTEMS
JOURNAL EDITED BY MEDICAL COLLEGE – JAGIELLONIAN UNIVERSITY
Vol. 5, No. 10, 2009, pp. 59–69

HEARTFAID'S ECRF: LESSONS LEARNT FROM USING A TWO-LEVEL DATA ACQUISITION AND STORAGE SYSTEM FOR KNOWLEDGE DISCOVERY TASKS WITHIN AN ELECTRONIC PLATFORM FOR MANAGING HEART FAILURE PATIENTS

[1] Andrzej A. Kononowicz, [1] Katarzyna Styczkiewicz, [1] Bogumiła Bacior, [2] Matko Bošnjak, [2] Rajko Horvat, [2] Marin Prcela, [2] Dragan Gamberger, [3] Angela Sciacqua, [5] Maria Consuelo Valentini, [1] Kalina Kawecka-Jaszcz, [4,5] Gianfranco Parati, [6] Domenico Conforti

[1] Jagiellonian University Medical College, Kraków, Poland
[2] Rudjer Boskovic Institute, Zagreb, Croatia
[3] University "Magna Graecia" of Catanzaro, Department of Experimental and Clinical Medicine, Italy
[4] University of Milan – Bicocca, Department of Clinical Medicine and Prevention, Milan, Italy
[5] Department of Cardiology, S.Luca Hospital, Istituto Auxologico Italiano, Milano, Italy
[6] University of Calabria, Department of Electronics, Informatics, Systems (DEIS), Italy

Abstract: Case report forms are important sources of medical knowledge in all clinical studies. Electronic versions of these forms have several advantages compared to traditional paper-based questionnaires, and they have been adopted in many contemporary research projects in medicine. This paper presents a framework for creating case report forms designed with a two-level approach. Data at the generic information model level is stored in EAV (entity-attribute-value) tables and extended by tables facilitating specification of the questionnaire layout. The second layer (knowledge model) specifies the domain specific concepts describing the field of application of the questionnaire. This framework has been applied and tested in the frame of an EU FP6 research project – HEARTFAID – the objective of which was to build a knowledge-based platform supporting the management of elderly patients suffering from heart failure. Data collected by the electronic case report form (eCRF) was used in the project's knowledge discovery and decision support tasks. The work presents a new way for effective extraction of the data necessary for the integration with the knowledge discovery process in a distributed, service oriented framework of the HEARTFAID platform. It is demonstrated that it is feasible to implement these tasks using the two-level EAV table design.
Keywords: Electronic Data Capture, Remote Data Entry, EAV, Two Layers Modelling, eCRF, Spring Framework

1. Introduction

Electronic Data Capture (EDC) techniques have been used in clinical trials for a long time [8]. The first EDC systems (also known under other names e.g. Remote Data Entry (RDE) Systems) date back to the early 1970s [7]. Since that time a huge amount of applications (either academic or commercial) for creating, managing and publishing medical on-line forms has been developed. The electronic versions of questionnaires seem to have lot of advantages in comparison to their paper-based counterparts [18]. Among the assets of EDC are cost savings, faster dissemination of forms and collection of data, built-in validation mechanisms, easy maintenance and export to statistical packages.

The conventional method of designing database schemes for questionnaires is to map a form to a single table or a set of tables in a relational database in which each attribute (question from the form) is stored into an individual column [10, 14]. Even though this technique works fine for many applications, it has become apparent that this method is not always effective [14], especially in bio-medical research or electronic health records. This problem pertains to databases with a large, heterogeneous list of fields from which many are optional and can be omitted. In such databases new fields are often added, altered or removed after the database has been deployed, and this introduces additional complication in its structure. Designing such databases with the conventional approach is possible but often troublesome and ineffective due to the limitations of tra-

ditional RDBM systems (e.g. a maximum limit of 255 columns in a single database table in some RDBM systems) or to the need to frequently update the database structure.

An alternative approach is to store records as association lists containing *(attribute name, attribute value)* pairs of variables [13]. A database that stores information in that form is called an entity-attribute-value (EAV) database. This storage method by itself is not new since it dates back to at least the time when the LISP programming language was created. However, its application in relational databases has not yet been very widespread. Classical EAV-databases contain one large table with just three columns: identifier of the described object, identifier of the attribute, value of the attribute. Additionally, dictionaries are required which contain metadata describing the attributes applied.

This simple design technique enables a very flexible method of space-efficient storage of heterogeneous data. However, it should be acknowledged that also this approach is not free from flaws. It is well suited for one object-at-a-time queries in which all information about a single object (e.g. patient) is returned, but it is less efficient in complex attribute-centric queries [14]. For such situations special frameworks facilitating more advanced searches in this model are implemented (as e.g. QAV: *querying entity-attribute* framework [13]), thus empowering the user to browse the data more easily. The overhead that is needed to organise the data in an EAV manner is often not worth the effort for the simple and static databases used in many business applications. The queries are also not time-effective, rendering them less suitable for commercial usage. These drawbacks are, however, not as obvious in research projects and clinical trials where more emphasis is put on the flexibility of the tool than its efficiency.

The EAV model can be considered the first generic layer of a database. This tier may be used in virtually any field of application, and can be extended by additional tables supporting more complex data design. An example of a model with such additions is represented by the EAV/CR by Nadkarni et al. [14], which enhances the EAV by structures for the representation of classes and relationships. Other approaches customize the EAV to store clinical forms [5]. The EAV model with its extensions represents an information model of the database which is domain independent. In a two-level approach to database design, a second layer (i.e. the knowledge model), is added on top of the first [11]. This model specifies the domain specific concepts describing the field of application of the questionnaire. It may consist of terminologies and ontologies related to a given specialization field. The values that can be entered into the information model can be constrained by knowledge model archetypes [2] – i.e. special templates that specify at runtime the way data can be entered. Archetypes may be specified autonomously by subject matter experts (e.g. clinicians) without the need to consult database specialists. A clear separation between the first and second level of the database makes the architecture flexible and reusable.

The aim of this paper is to report on the information obtained while implementing a vast two-level electronic case report form (eCRF) which was designed for the cardiology domain. The eCRF is part of a large knowledge-based platform called HEARTFAID supporting the management of elderly heart failure patients, developed in the frame of the EU FP6 research program. It was required by the HEARTFAID project that the eCRF system implements insertion, modification and querying of large forms (containing over 700 attribute values for each patient). The system needed to be well integrated with the remaining services of the platform.

2. The HEARTFAID Platform

Heart failure (HF) occurs when the heart fails to pump enough blood to meet the metabolic needs of the body's tissues and/or organs. The prevalence of this pathological condition is very high – approximately 10 million patients suffering from HF in Europe. Chronic (C)HF is a disease of older people; the Framingham study noted a doubling of prevalence with each advancing decade, reaching a rate ranging from 7% to 10% in those aged 80 and older. The mortality of patients with severe HF is also high, approaching 50% in the course of one year in NYHA IV[1] class patients. However, it is believed that through regular monitoring and personalised management of patients affected by this condition, their survival rate and quality of life can be significantly improved.

The role of the HEARTFAID platform is to support physicians and healthcare personnel (e.g. nurses) in managing heart failure patients, while at the same time empowering patients to self-monitor their health condition [3, 4]. HEARTFAID is a web-based platform of services integrating several diverse modules (Fig. 1). Its basic function is to collect patient-related biomedical data from different sources (e.g. mobile and wearable measurement devices or medical imaging systems) and enable access to previously collated data from electronic health records. Part of the system includes declarative and procedural knowledge taken from evidence-based sources such as medical guidelines and carefully selected research papers [6, 16]. The users of the platform may securely access the data it contains in a standardised manner. The platform gives access to data taking into account the different roles and rights of the users. New knowledge can be discovered based on the data collected on the platform by employing newly developed artificial intelligence tools. The system has the potential to support physicians in making clinical decisions in the workplace also by alerting them if a dangerous situation is detected. All HEARTFAID services are integrated by an enterprise service bus (ESB). The system utilises a single sign-on mechanism. Users interact with the system through a customisable web portal. The anticipated results of integrating the platform into clinical practice include a reduction in the re-admission of HF patients to hospital, improvement in the quality of treatment and a decrease in management costs [3, 4].

A knowledge-based platform like HEARTFAID requires various forms of medical data acquired from different sources. Data collected from mobile and wearable devices are covered by the AmI service. The role of the HEARTFAID's electronic case report form (eCRF) is to handle all data required by clinicians that need to be inserted manually by the medical personnel. Beyond the scope of the eCRF is the storage of medical knowledge in the form of rules or ontologies which are used for inference in knowledge discovery and decision support systems. However, these modules exploit the data collected by the

1 NYHA – New York Heart Association Functional Classification – A four scale classification of heart failure extent

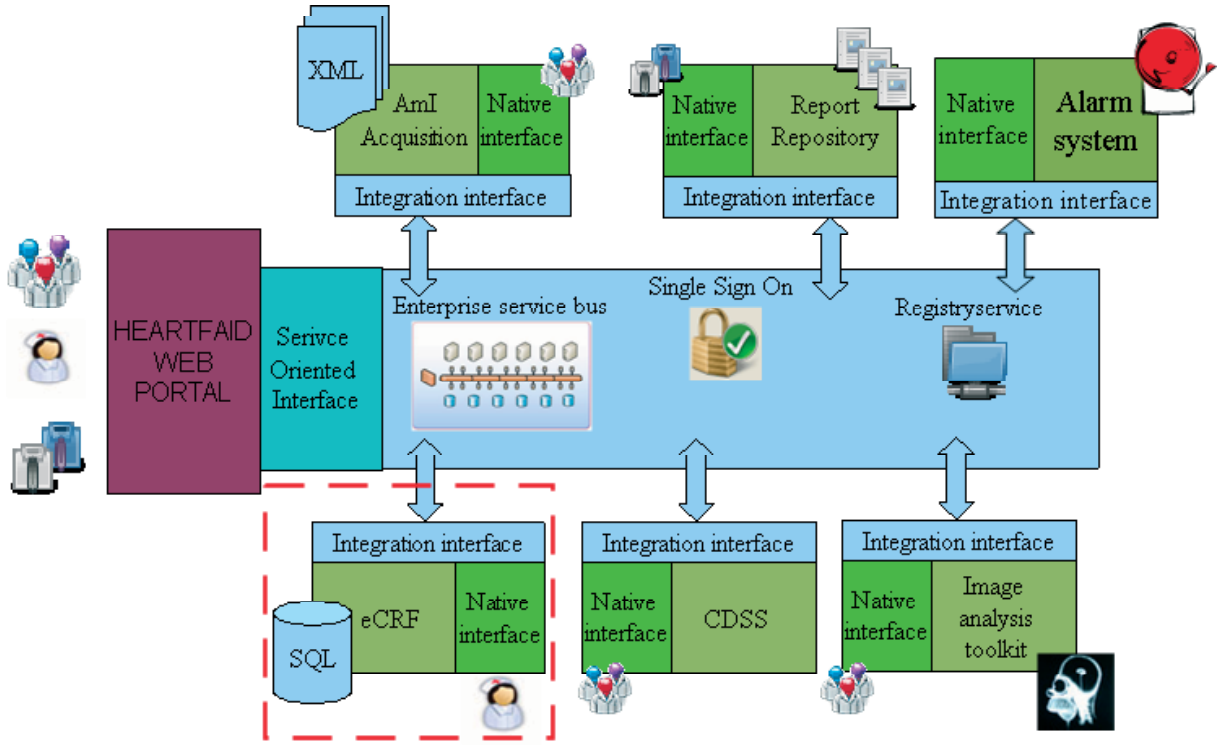


Fig. 1. General overview of the HEARTFAID services

eCRF. The eCRF is intended to be used by medical personnel in the hospital and is not accessible by patients. It plays the role of a specialised electronic health record, collecting heart failure data from a multitude of sources. From a medical (i.e. cardiographic) perspective the eCRF is useful because it gives easy access to the results of lab tests, to treatment schedules and to the prognostic assessment of HF patients.

The HEARTFAID eCRF comprises three parts: the baseline, additional visits and final evaluation forms. Each of these forms is uniquely assigned to a patient and can be filled out only once, with the exception of the additional visit form which may be repeatedly compiled without limitations. Questions in the eCRF questionnaire may be combined into groups. The activation of a group may be triggered in real-time by the input of a specific value by the user. Question groups may be nested to an unlimited depth. Most of the questions are of simple types: Boolean values, text strings, numerical values (integer, real numbers) and dates. However, there are also more complex types of questions which involve, for instance, the selection of a value from a controlled vocabulary, or the use of a special tool to specify a medication and its dosage from a hierarchical list of products (drug class, international name and generic name). It is also possible to add new drugs to the list. Some questions are grouped into matrices (tables) of values of simple types. The forms also contain rules for validating inserted values.

3. Method

While planning the implementation of the eCRF we looked for off-the-shelf products that were web-based, available free, open source, flexible enough to add new question types, able to support large questionnaires with nested groups of questions, based on XML and J2EE technologies, and easy to integrate into the HEARTFAID platform. None of the existing tools we found for designing web-based questionnaires (e.g. ArchiMed [5], Form Handler [21], Instant Survey [24], Survey Monkey [26], WebEAV [15], Zoomerang [27]), fully met our demands. For that reason it was decided to implement the eCRF from scratch. Since the number of questions was large (more then 700), quite diverse, potentially changeable and the efficiency of the tool was not a critical factor, it was decided to employ a two level architecture. The idea of a two-level approach emerged in electronic health records development [11]. Following this approach database structures are divided into two separated models: information model and knowledge model. The information model represents stabile and generic concepts, whereas the knowledge model depicts the dynamics of the problem field [11]. In the HEARTFAID's eCRF the information model expresses a generic database for storing clinical forms following the EAV paradigm. The classical EAV data model has been extended to facilitate the usage of complex web-based forms. In Fig. 2 the ERD (entity-relationship-diagram) of the information model underlying the HEARTFAID eCRF is presented. The model is generic – i.e. it does not contain any information specific to the heart failure domain and can be used in diverse multi-centred clinical trials. The EAV model was implemented in a RDBM system. Basic EAV tables were extended by additional tables for storing hierarchical question groups and for user management. Similar approach was taken in other EAV projects (as e.g. in EAV/CR representation by Nadkarni at al [14]). The *User Center* and *User* tables enable the separation of patients coming from different research institutions and enable access to the data only by entitled users. The *Patient*

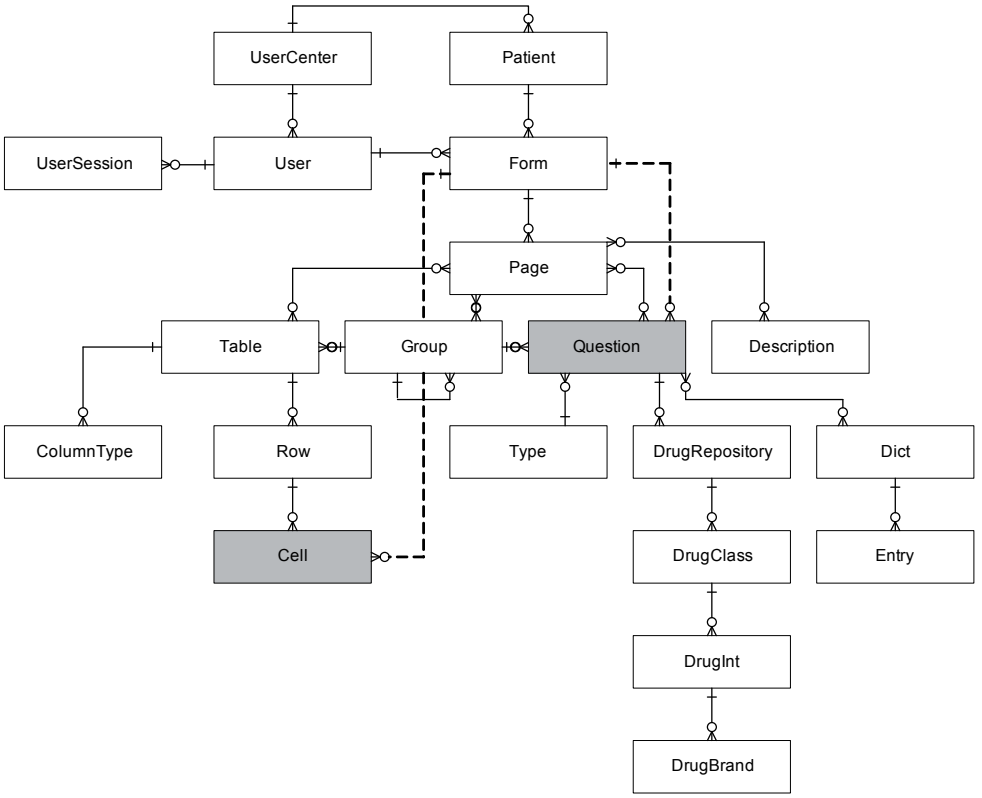


Fig. 2. ERD of the information model under laying the HEARTFAID eCRF

```
<bean id="physical_exam_systolic_blood_pressure" class="org.javs.ecrf.mvc.model.Type" singleton="true">
     <property name="type" value="integer"/>
     <property name="html">

     <value><![CDATA[Systolic blood pressure:]]></value>
     </property>
     <property name="cui" value="C1306620"/>
</bean>

<bean id="question_hfcrfsa_1_g1_1" class="org.javs.ecrf.mvc.model.Question" singleton="false">
     <property name="order" value="1"/>
     <property name="type" ref="physical_exam_systolic_blood_pressure"/>
     <property name="value" value=""/>
</bean>
```

Fig. 3. XML archetypes specifying the values that can be inserted into the form

table contains basic patient data. Since the questions are assigned to pages and these pages may contain many levels of nested question groups or tables this structure is reflected by the *Page*, *Table* and *Group* entities. The grey-shaded *Question* and *Cell* tables are classic EAV tables containing a reference to the type of the question, the owning entity (i.e. *Page*, *Group*, or *Row*) and the value. Additionally, these tables contain information about the time of creation and last modification of the values, version number, as well as the identity of the user that modified the value. The dashed line between the *Question* or *Cell* tables and the *Form* table is a redundant connection added for efficiency reasons to accelerate queries with fields nested deep in many subgroups. The *Description* table contains textual information needed as additional description in the forms. The *Drug[X]* ($X \in \{Repository, Class, Int, Brand\}$) and *Dict* tables represent respectively the pharmacological treatment and values from controlled vocabularies.

The archetypes (second layer of the model) constraining the values that can be inserted into the database are specified in XML syntax, compatible with the bean definition syntax of the Spring Application Framework [9]. An example of the specification of a question type and its instance is presented in Fig. 3. The first archetype bean example defines a type representing a patient's systolic blood pressure taken during a physical examination. This question has its description in HTML format (attribute *html*), stating that it accepts only integer values (attribute *type*) and a mapping to a concept in the UMLS ontology explaining its semantics (attribute *cui*). The second bean

is an instantiation of the previously mentioned question type (attribute *type*). The position at which the question is displayed in the question group is specified by the attribute order, and its default value may be specified by the attribute value. The archetypes often also contain lists of questions or subgroups aggregated by group type, or information about question groups being activated or deactivated based on specific values of the questionnaire fields inserted by the user.

Both archetype beans (i.e. type declaration bean and question instantiation bean) are mapped to POJO (plain old java objects) elements and are stored on demand in the eCRF database using the Hibernate Framework [12]. The way the archetypes are specified enables easy extension of the list of constraining rules (e.g. by information about the soft or hard limits for ranges of accepted values).

4. Results

The eCRF has been implemented in the course of the second year of the HEARTFAID project in the Java 5 programming language. The development has been accelerated by the usage of the Spring Application Framework [9] version 1.2 and Hibernate 3 [12]. The final knowledge model of the eCRF specified by XML archetypes included 735 question instances of 364 semantic types. Archetypes were created using a general purpose XML editor (Altova XMLSpy 2008 [20]). Data were stored in a MySQL 5.1 RDBM system. A simplified structure of the eCRF is presented in the figures included in the Appendices 1 and 2. In order to make the schemes legible, the number of fields for each object was limited to a maximum of 10 fields. The letters *b,a* and/or *f* denote in which eCRF type of form this question is located (i.e. *baseline*, *additional form* or *final visit*). The forms are presented online as HTML views created with XSLT transformation of XML archetypes and data retrieved from the database (Fig. 4). The *top bar* contains the questionnaire's name, patient id and page number. The pages can be changed either through the list of pages in the *table of content* panel in the right part of the form or through the *backward* and *forward* buttons in the navigation bar. The form is automatically saved after changing a page or after clicking on the *submit* button. Activation of the *cancel* button rejects the last changes and exits the form. Question groups are marked by red boxes and activated by trigger questions (e.g. in Fig. 4 the group containing the *max. ST depression* question is activated by setting the value "yes" in the *ST depression* field). In Fig. 5 a 3x3 question table (matrix) of integers is presented. In addition, above the main form a list of detected validation errors is presented.

Communication with the eCRF with the HEARTFAID platform is established through an XML protocol implemented by one of the partners in the project (SYNAPSIS) including all the necessary information of an HL7 message [22] and following the transactions suggested by IHE [23]. The HEARTFAID middleware implements Patient Demographic Query HL7 V3 (PDQ) and Patient Identifier Cross-Reference HL7 V3 (PIX) profiles. In order to integrate the patient-related data into the platform a MPI (Master Patient Index) service is used which manages patient's demographic information and guarantees their unique identification in the environment. For instance, while registering a new patient on the platform, a message is sent from the HEARTFAID portal to the ESB which was implemented using the Mule open-source framework [25]. Mule descriptors for routing the information to a MIDA Graph (a workflow engine implemented by SYNAPSIS [19]) are read and transformed into information that is stored in the MPI and transmitted as HTTP XML messages to the eCRF service. The eCRF receives the messages, enrols the patient and sends back a confirmation message [19].

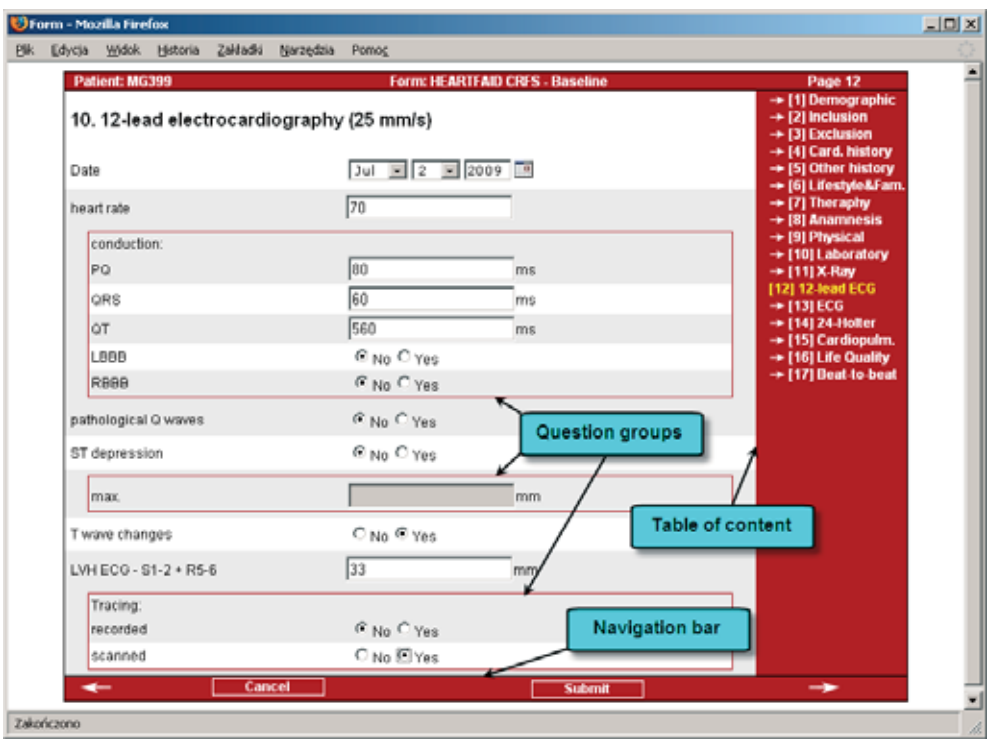


Fig. 4. User interface of the HEARTFAID eCRF

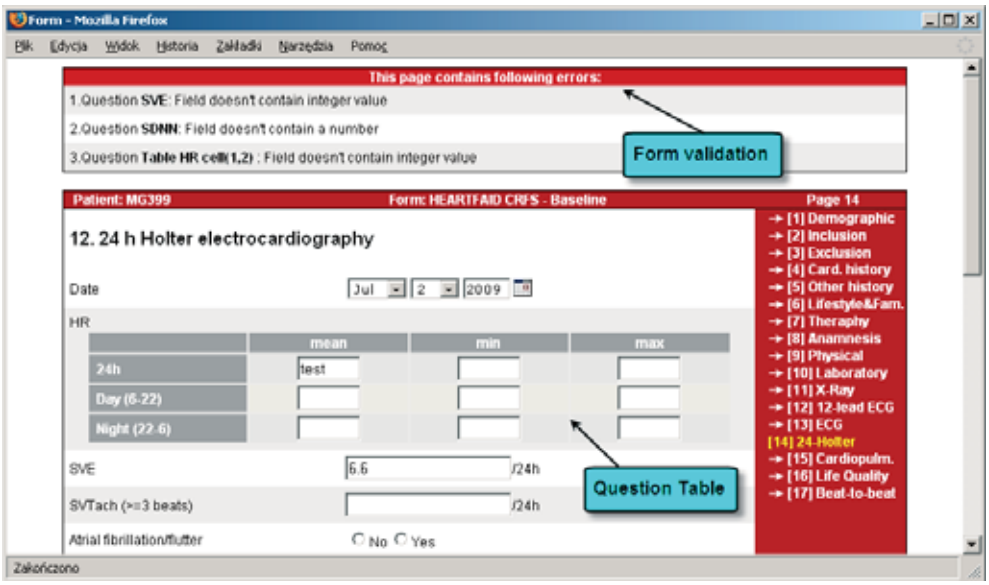


Fig. 5. Question tables and form validation in eCRF HEARTFAID

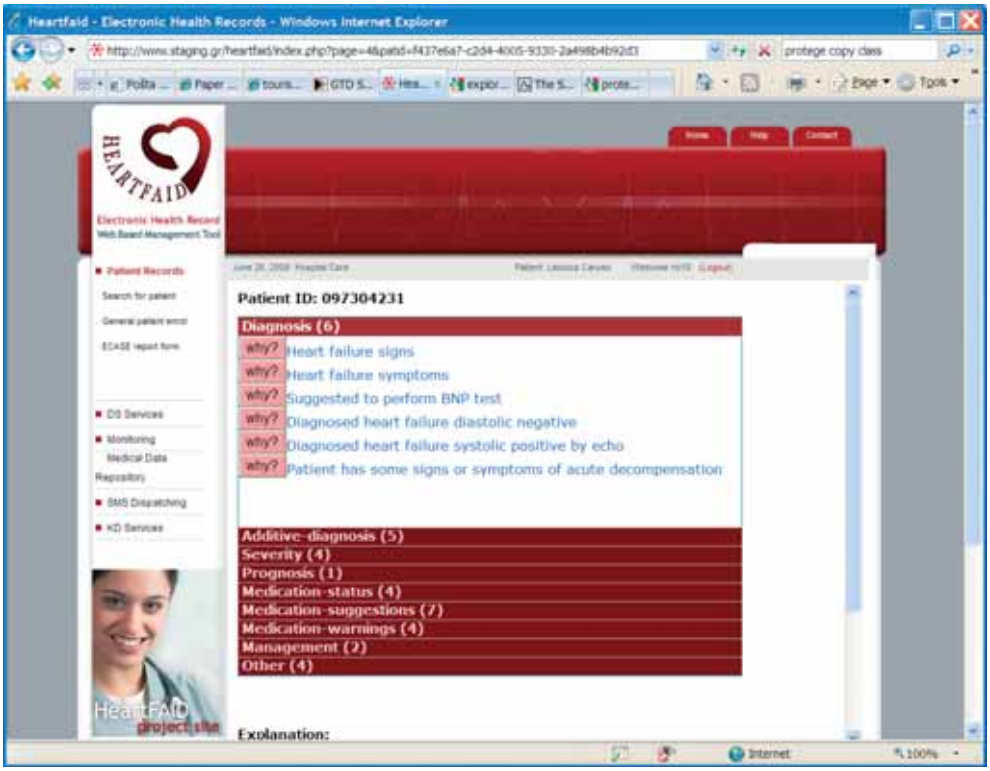


Fig. 6. The result of reasoning based on the data collected by the eCRF

The eCRF was deployed on the HEARTFAID platform in 2007 and since then it has been in constant use. Data from approximately 100 patients from four clinical centres [Università degli studi Magna Graecia, Catanzaro (Italy), Università degli studi di Milano Bicocca, Milan (Italy), Jagiellonian University Medical College, Kraków (Poland) and S. Luca Hospital, Istituto Auxologico Italiano, Milan (Italy)] have been collected.

The eCRF has been integrated with the HEARTFAID's Knowledge Discovery Service (KDS) and Decision Support System (DSS) developed by Rudjer Boskovic Institute in Zagreb (Croatia). Both services require tight integration with the large amount of patient data collected by eCRF, however these services require substantially different data access types. DSS is always focussed on one patient while KDS requires information about all or most of available patient data that has been collected by the eCRF. Additionally, it must be noted that DSS requires effective access to the most recent information for all potentially relevant measurements regardless of when they

were collected and with a clear indication about when the data was acquired. In contrast to this, actual data collection time is not relevant for KDS, but it requires access to the data grouped according to the time of its collection, that data should be ordered by its historical order, and that it is identified by the time interval from previous measurements. Fig. 6 demonstrates a typical result from the decision support service while Fig. 7 and 8 illustrate the knowledge discovery service.

A unique property of the currently implemented KDS is that it integrates knowledge discovery algorithms with direct database access into one web-based service. This is not a simple task due to the complexity of the KD process [16]. The HEARTFAID service implements the modern random forest based machine learning algorithm [1] that has been reimplemented by Rudjer Boskovic Institute. The service has been built as a series of projects so that every project consists of different datasets with many tasks that can be performed for every dataset. Access to projects, datasets, and tasks is enabled though a web interface (Fig. 7).

Computationally, the most complex part of the service is the construction of the classifier and the preparation of a report (Fig. 8) based on the results of this process. The value of this newly implemented service is the realisation of direct access to the data in the eCRF and its automatic transformation into a form that can enter the KD process. Direct access to the relational database containing EAV tables by a traditional SQL interface is very laborious. This problem can be solved by implementing a special query module for external analytical services. An interface consisting of four generic functions has been implemented for the purpose of the knowledge discovery task. Table 1 contains a description of these functions:

These functions are available through a HTTP GET interface. In the following line a command is shown that starts a query involving the *getLastValue* function http://localhost:8080/heartfaid/query.html?function=getLastValue&uuid=312&sid=physical_exam.weight

After execution the interface searches in the eCRF database for all values of the *physical_exam.weight* attribute regarding the patient with the id 312. The result of the function is returned in simple XML syntax. This allows a clear separation of the data collection and query tool located at one centre (currently at Jagiellonian University Medical College, Kraków, Poland), from the knowledge discovery system located at a remote centre (currently located at Rudjer Boskovic Institute, Zagreb, Croatia).

Discussion

After the generic framework for designing questionnaires had been developed, the process of implementation of the HEARTFAID's eCRF knowledge model by specifying the XML archetypes took little time and did not cause any difficulties. The structure of the eCRF turned out to be more stable than initially anticipated, so the benefit of the flexibility of the architecture has not been fully used (with the exception of a few minor changes). On the other hand, the drawbacks of decreased database efficiency in this type of application are hardly noticeable. In the production database containing just a few users and approximately 100 forms installed on a Intel Core 2 Duo T5450 1,66Ghz,1GB RAM computer, loading a whole form from a database took in average 1360 ms, saving a modified

Fig. 7. The main page of the HEARTFAID knowledge discovery service with three current projects: "platform-test-worsening", "Iris-test-project", and "platform-demographic"

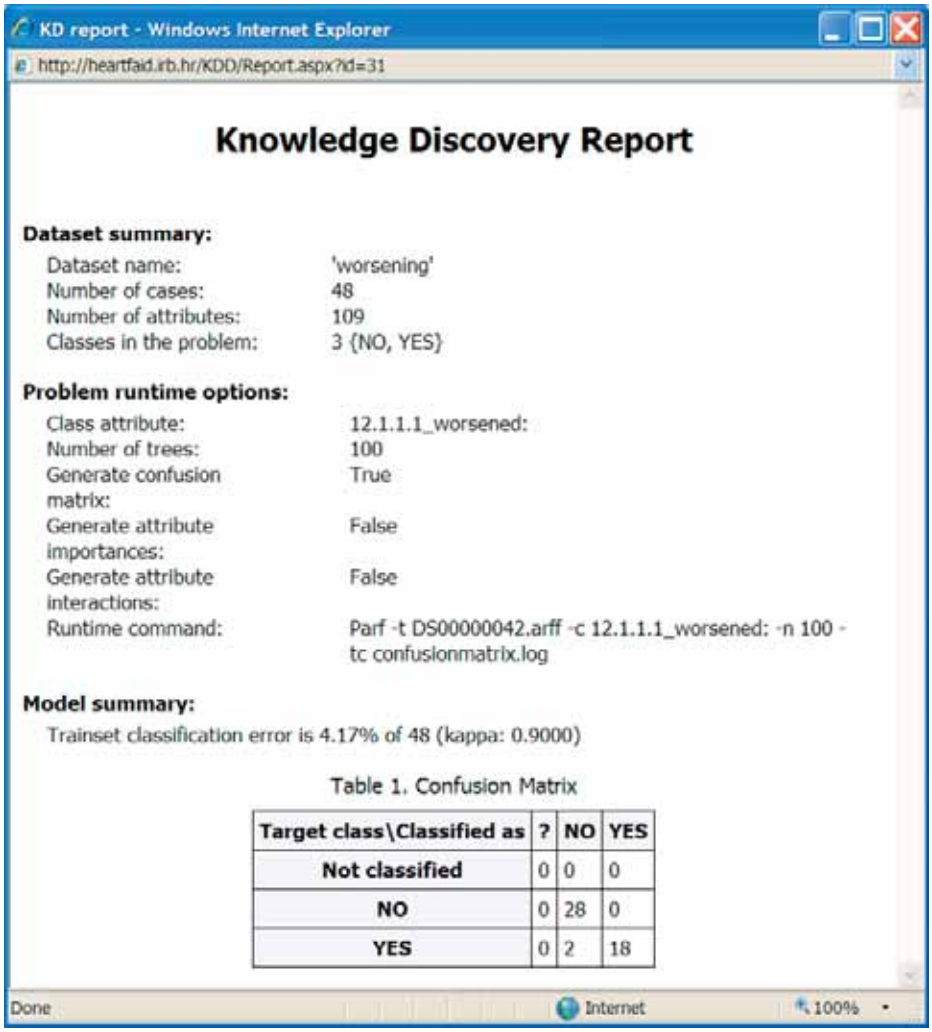


Fig. 8. The result of any KD task is a report. The figure presents a report for a two-class domain obtained after constructing a random forest with 100 trees. The main part of the report is the confusion matrix demonstrating the predictive accuracy measured by cross-validation on the training set.

form 600 ms, querying the last value of a selected parameter 84 ms. . Thanks to the application of XML technology the integration of the eCRF to the platform's enterprise service bus was easy and fulfilled the requirements of current medical informatics standards.

The future plan for the proposed architecture includes implementation of a graphical editor for the XML archetypes and extension of the list of constraints that can be used for the knowledge model's specification. Tighter integration of the eCRF with knowledge engineering and data mining tools through the proposed interface also seems to be important.

It is not easy to give definite advice about when to use EAV tables instead of traditional relational database design. If our highest priority is flexibility, and the number of collected attributes is very large and potentially often changeable, this suggests that a two-level EAV design should be used. In all other cases, a more traditional design would probably be more advantageous. When designing frameworks with EAV databases for knowledge discovery tasks it is imperative to also offer a special query module with an interface similar to that presented in this paper, or to export the data to an external system with a different data model.

Conclusions

This paper presents a practical implementation of a two-level database system for a medical research project. The generic layer of this database uses EAV tables which are useful for designing large heterogeneous and frequently changeable database schemas, as are often found in research studies. In this system a method for implementing the concept of two-level architecture in a modern application framework (Spring Framework) has been demonstrated. The fact that the system has been in use for almost two years in the HEARTFAID project and has delivered useful data for other modules like a knowledge discovery module and decision support services proves the feasibility and the effectiveness of this solution. The significance of our work consists in the proposal of a new type of direct interface for accessing complex data structures with the

Tab. 1. eCRF interface for knowledge discovery tasks

Function Name	Description
getLastValue	Returns the last known descriptor value available for the patient. If all values are unknown the returned value is also unknown.
getAnyValue	Returns information concerning all previous visits. For numerical measurements it returns two values: minimum and maximum while for categorical attributes it returns most frequent (mode) value. If all values are unknown the returned value is also unknown.
getDifference	Returns the difference between the last available piece of data and the penultimate piece. If there are not two available entries the value is unknown. For numerical attributes (e.g. laboratory values) it is the difference (+/- value). For categorical attributes it is 0 (no change) and 1 (value changes) [or -1 improved, 0 no change, 1 worsening]
getFlattenedTable	For categorical values it returns the number of known values and the most frequent value. For numerical it returns mean, minimal and maximum value, range, standard deviation and slope.

output already prepared for artificial intelligence applications. Additionally, it is also clearly stated that this model is not appropriate for every database, especially not for large commercial databases, and therefore its adoption needs to be carefully considered.

References

1. Breiman L., Random Forests, Machine Learning 45(1), pp. 5-32, 2001.
2. Bird L., Goodchild A., Tun Z., Experiences with a Two-Level Modelling Approach to Electronic Health Records, Journal of Research and Practice in Information Technology, 35(2), pp. 121-138, 2003.
3. Chiarugi F. et al., Support for the Medical-Clinical Management of Heart Failure within Elderly Population: the HEARTFAID Platform, Proc. of ITAB, Ioannina, Greece, 26-28 October 2006.
4. Conforti D. et al., HEARTFAID: A Knowledge Based Platform for Supporting the Clinical Management of Elderly Patients with Heart Failure, The Journal on Information Technology in Healthcare, 4(5), pp. 283-300, 2006.
5. Duftschmid G., Gall W., Eigenbauer E., Dorda W., Management of data from clinical trials using the ArchiMed system, Med. Inform. Internet, 27(2), pp. 85-98, 2002.
6. Gamberger D., Prcela M., Jović A., Šmuc T., Parati G., Valentini M., Kawecka-Jaszcz K., Kononowicz A. A., Candelieri A., Conforti D., Guido R., Medical Knowledge Representation Within Heartfaid Platform, Healthinf, Funchal, Madeira – Portugal, 2008.
7. Helms R. W., Entering Data from Remote Terminals in Clinical Centers using IBM's OS/TSO in the Kidney Transplant Histocompability Study, Technical Report 007, Chapel Hill, NC University of North Carolina, KTHS Statistics and Data Management Center, Department of Biostatistics, 1973.
8. Helms R. W., Data Quality Issues in Electronic Data Capture, Drug Information Journal, 35, pp. 827-837, 2001.
9. Johnson R., Hoeller J., Arendsen A., Risberg T., Sampaleanu C., Professional Java Development with the Spring Framework, John Wiley & Sons, 2005.
10. Merzweiler A., Weber R., Garde S., Haux R., Knaup-Gregori P., TERMTrial – terminology-based documentation systems for cooperative clinical trials, Comput. Meth. Programs Biomed., 78, pp. 11-24, 2005.
11. Michelsen L., Pedersen S. S., Tilma H. B., Andersen S. K., Comparing different approaches to two-level modelling of electronic health records., Stud. Health Technol. Inform., 116, pp. 113-118, 2005.
12. Minter D., Linwood J., Hibernate From Novice to Professional, Apress, 3 edition, 2006.
13. Nadkarni P., QAV: querying entity – attribute – value metadata in a biomedical database, Comput. Meth. Programs Biomed., 53, pp. 93-103, 1997.
14. Nadkarni P. et al., Organization of Heterogeneous Scientific Data Using the EAV/CR Representation, J. Am. Med. Inform. Assoc., 6(6), pp. 478-493, 1999.
15. Nadkarni P., Brandt C., Marenco L., WebEAV: Automatic Metada-driven Generation of Web Interfaces to Entity-Attribute-Value Databases, J. Am. Med. Inform. Assoc., 7(4), pp. 343-356, 2000.
16. Prcela M., Gamberger D., Bogunovic N., Developing Factual Knowledge from Medical Data by Composing Ontology Structures, MIPRO 2007, Opatija, Croatia.
17. Sonicki Z., Gamberger D., Smuc T., Sonicki D., Kern J., Data mining server: On-line knowledge induction tool, in: Proc. of Medical Informatics Europe, IOS press, pp. 330-334, 2002.
18. Wyatt J. C., When to Use Web-based Surveys, J. Am. Med. Inform. Assoc., 7(4), pp. 426-430, 2000.
19. HEARTFAID Consortium, D28 – Integration and Interoperability middleware prototype, 2008.
20. Altova XMLSpy, http://www.altova.com/xml-editor/
21. Form Handler, http://www.formhandler.net
22. HL7, Health Level 7, http://www.hl7.org
23. IHE, Integrating the Healthcare Enterprise, http://www.ihe.net
24. Instant Survey, http://www.instantsurvey.com
25. Mule, ESB http://www.mulesoft.org/display/COMMUNITY/Home
26. Survey Monkey, http://www.surveymonkey.com
27. Zoomerang, http://www.zoomerang.com

Appendix 1 – Knowledge model of the HEARTFAID eCRF, Simplified – Part 1 of 2

Anamnesis		
blood_pressure_change	(a,f)	enum[change]
bradycardia	(b)	boolean
bradycardia_change	(a,f)	enum[change]
chest_pain	(b)	boolean
chest_pain_change	(a,f)	enum[change]
chest_pain_remote	(b)	boolean
dyspnoea	(b)	boolean
dyspnoea_change	(a,f)	enum[change]
dyspnoea_remote	(b)	boolean
fatigue	(b)	boolean
and 19 more fields		

Echocardiography		
aorta_ascending_aorta_diameter	(b,a,f)	double
aorta_root_diameter	(b,a,f)	double
contractility_akinesis	(b,a,f)	boolean
left_atrium_anteroposterior_diameter	(b,a,f)	double
left_ventricle_end-diastolic_diameter	(b,a,f)	double
left_ventricle_end-diastolic_volume	(b,a,f)	integer
mitral_valve_deceleration_time	(b,a,f)	integer
mitral_valve_emax-amax	(b,a,f)	double
mitral_valve_mitral_regurgitation	(b,a,f)	integer
pulmonary_artery_pressure	(b,a,f)	integer
and 16 more fields		

24 h Holter Electrocardiography		
atria_fibrillation_flutter	(b,a,f)	boolean
conduction_abnormalities	(b,a,f)	boolean
conduction_abnormalities_details	(b,a,f)	textfield
date	(b,a,f)	date
heart_rate_HF	(b,a,f)	double
heart_rate_LF	(b,a,f)	double
heart_rate_pNN50	(b,a,f)	double
heart_rate_rMSSD	(b,a,f)	double
heart_rate_SDANN	(b,a,f)	double
heart_rate_total_power	(b,a,f)	double
and 9 more fields		

Final Visit		
date	(f)	double
required_hospitalization_date	(f)	double

Rehabilitation		
model	(f)	textfield
required	(f)	boolean
time	(f)	integer

Chest X-ray		
cardio-thoracic_ratio	(b,a,f)	integer
comment	(b,a,f)	textarea
date	(b,a,f)	date
pulmonary_congestion_or_oedema	(b,a,f)	boolean

Quality of Life Questionnaire		
date	(b,f)	date
minnesota_total_score	(b,f)	integer
sf36_bodily_pain	(b,f)	integer
sf36_general_health	(b,f)	integer
sf36_mental_component_summary	(b,f)	integer
sf36_mental_health	(b,f)	integer
sf36_physical_component_summary	(b,f)	integer
sf36_role_emotional	(b,f)	integer
sf36_role_physical	(b,f)	integer
sf36_social_functioning	(b,f)	integer
and 2 more fields		

Family History		
primary_cardiomyopathy	(b)	boolean

Beat-to-beat Blood Pressure Monitoring		
baseline_finger_BP_SBP	(b,f)	integer
baseline_finger_HR	(b,f)	integer
comments	(b,f)	textarea
cuff_size	(b,f)	enum[cuff_size]
date	(b,f)	date
device	(b,f)	enum[device]
end_standing_CB_finger_BP_SBP	(b,f)	integer
end_standing_CB_finger_HR	(b,f)	integer
finger	(b,f)	enum[finger]
hand	(b,f)	enum[hand]
and 13 more fields		

Drug Theraphy		
drug_theraphy	(b,a,f)	drug
drug_theraphy_change	(a)	drug

Laboratory Assessment		
ALT	(b,a,f)	double
AST	(b,a,f)	double
blood_samples_for_DNA-RNA	(b,a,f)	boolean
BNP	(b,a,f)	pmol_mg
creatinine	(b,a,f)	umol_mg
creatinine_clearance	(b,a,f)	double
date	(b,a,f)	date
glucose	(b,a,f)	mmol_mg
glycated_hb	(b,a,f)	double
hb	(b,a,f)	double
and 12 more fields		

Physical Examination		
body_temperature	(b,a,f)	double
diastolic_blood_pressure	(a,f)	integer
heart_murmurs	(b,a,f)	boolean
heart_murmurs_apex	(b,a,f)	boolean
heart_murmurs_base	(b,a,f)	boolean
heart_murmurs_diastolic	(b,a,f)	boolean
heart_murmurs_systolic	(b,a,f)	boolean
heart_sounds	(a,f)	boolean
heart_sounds_bilateral	(b,a,f)	boolean
heart_sounds_fourth	(b,a,f)	boolean
and 24 more fields		

Cardiopulmonary Exercise Testing		
AT	(b,f)	double
BP_baseline_DBP	(b,f)	integer
BP_baseline_SBP	(b,f)	integer
BP_end_DBP	(b,f)	integer
BP_end_SBP	(b,f)	integer
BP_peak_ex_DBP	(b,f)	integer
BP_peak_ex_SBP	(b,f)	integer
data_recorded	(b,f)	boolean
O2_pulse	(b,f)	double
RQ	(b,f)	double
and 10 more fields		

Additional Visit		
date	(a)	date
next_scheduled_visit_date	(a)	date
other_than_chf_reasons_of_visit	(a)	boolean
required_advice	(a)	boolean
required_advice_details	(a)	boolean

Appendix 2 – Knowledge model of the HEARTFAID eCRF – Simplified – Part 2 of 2

Six-minute walking test		
BP_baseline_DBP	(a)	integer
BP_baseline_SBP	(a)	integer
BP_end_DBP	(a)	integer
BP_end_SBP	(a)	integer
date	(a)	date
HR_baseline	(a)	integer
HR_end	(a)	integer
SpO2_baseline	(a)	integer
walking_distance	(a)	integer

Lifestyle Information		
alcohol_use	(b)	boolean
physical_activity	(b)	enum[ph_activity]
smoking	(b)	boolean
smoking_cessation	(a,f)	boolean
smoking_cessation_date	(a,f)	date
smoking_duration	(b)	integer
smoking_no_cigarettes	(b)	integer

Non Cardiovascular Medical History		
anemia	(b)	boolean
anemia_worsening	(a,f)	boolean
bronchial_asthma	(b)	boolean
connective_tissue_diseases	(b)	boolean
diabetes	(b)	boolean
diabetes_type	(b)	enum[type12]
diseases_not related_to_hf	(b)	boolean
diseases_potentially related_to_hf	(b)	boolean
endocrine_disorders	(b)	boolean
exposure_to_endemic_diseasesy	(b)	boolean
and 29 more fields		

Demographic Data		
birthday	(b)	date
death	(a,f)	boolean
death_cause	(a,f)	textfield
death_date	(a,f)	date
sex	(b)	enum[sex]
status	(b)	enum[pat_status]

Cardiovascular Status		
aortic_regurgitation	(b)	boolean
aortic_stenosis	(b)	boolean
CABG	(b,a,f)	boolean
cardiovascular_reason_of_death	(a,f)	boolean
cerebrovascular_events	(b,a,f)	boolean
changes_in_therapy	(a)	boolean
chf_status_improved	(a)	boolean
chf_status_requires_hospitalization	(a)	boolean
congenital_heart_disease	(b)	boolean
congestive_heart_failure	(b)	boolean
and 41 more fields		

12-Lead Electrocardiography		
conduction_LBBB	(b,a,f)	boolean
conduction_PQ	(b,a,f)	integer
conduction_QRS	(b,a,f)	integer
conduction_QT	(b,a,f)	integer
conduction_RBBB	(b,a,f)	boolean
date	(b,a,f)	date
heart_rate	(b,a,f)	integer
heart_rate_24h_max	(b,a,f)	integer
heart_rate_24h_mean	(b,a,f)	integer
heart_rate_24h_min	(b,a,f)	integer
and 14 more fields		

Substudy 1 - Inclusion Criteria		
age_gt_65	(b)	boolean
chf	(b)	enum[diag_chf]
diastolic_dysfunction	(b)	boolean
ef_lt_40p	(b)	boolean
functional_capacity	(b)	boolean
hypertension	(b)	boolean
idcm	(b)	boolean
ihd	(b)	boolean
informed_consent	(b)	boolean
sinus_rhythm_presence	(b)	boolean
and 2 more fields		

Substudy 1 - Exclusion Criteria		
AIDS	(b)	boolean
autoimmune_disorders	(b)	boolean
cardiac_resynchronization_therapy	(b)	boolean
drug_or_alcohol_abuse	(b)	boolean
gfr_lt_30	(b)	boolean
hepatic_disease	(b)	boolean
immunosuppressive_therapy	(b)	boolean
malignancy	(b)	boolean
no_informed_consent	(b)	boolean
pacemaker	(b)	boolean
and 3 more fields		

Patient	
id	long
uuid	string
initials	string
usercenter	integer
createTime	date
updateTime	date
createUser	string
updateUser	string

INSTRUCTION FOR AUTHORS

1. The original and one Photostat copy of the manuscript should be mailed to: Managing Editor Zdzisław Wisniowski. Authors are strongly urged to submit a CD containing the manuscript in Word for PCs format along with the hard copies. Articles already published or those under consideration for publication in other journals or periodicals should not be submitted.
2. Manuscripts should be in English. Usage of correct language is the responsibility of the author.
3. Submission of manuscripts to us implies that the copyright of the entire material except those already copyrighted by other publications like tables, figures etc and those permission the author(s) have already obtained, belongs to the publisher.

PRESENTATION

4. The text may be prepared from a good quality laser printer or electronic typewriter. It is to be typed (one side only) on a regular sheet of paper A4 size preferred with single line spacing. New Times Roman font of 12 point size is recommended. Authors are requested to paste figures and tables (if possible) at appropriate places in the text. If this is not possible, they may be given separately, in which case their exact locations in the text may be marked by pencil.
5. The article should begin with a short ABSTRACT followed by a short INTRODUCTION. The rest of the article ma be titled and arranged as per the wishes of the author.
6. REFERENCES should continue the last section of the article. In the text, references to other papers or books should be cited using consecutive numbers in parenthesis (e.g. [1,2]) and they should be listed numerically in the last section.
   Examples:
   I. Pyrczak W., Sarapata K.: Instructions for authors: references. Bio-Alghoritms and Med-Systems, 1, 386, 2005
   II. Tadeusiewicz R., Ogiela M.R.: Medical Image Understanding Technology. Springer Verlag, Heidelberg 2004
   Authors are to make sure that each reference appearing in the text appears in the list of references at the end and vice versa.
7. TITLE PAGE AND FIRST PAGE – The title page of the manuscript, the name(s) of author(s) and institution(s) where the work has been carried out should be typed out on a separate sheet of paper.
8. Section headings (ABSTRACT, INTRODUCTION etc.) are to be typed in upper case letters and placed on a separate line
1. Scientific names should be typed in italics or underlined
2. Tables should be numbered consecutively using Arabic numerals.
3. Illustrations should be limited to materials essential for the text. All figures should be numbered consecutively and should be submitted as sharp, high-quality prints. Figures that are to appear as a group should be photographed and mounted together. All figures and groups of figures should be trimmed at right angles and pasted directly on the typescript in the appropriate places. They should be of a size permitting photographic reduction by 25% together with text (not more than 15 cm across for page width and not higher than 20 cm). Legends to figures should appear directly below the respective figures.
4. In line drawings, all lines should be of uniform thickness; letters and numbers should be of professional quality and proper dimensions, approximately 1mm high when reproduced.
5. RUNNING TITLE – Please make sure to mention the running title (short title) of the article separately in the first page where the title, names of authors, their affiliations, etc. are given.
6. PREFFERED LENGTH – The length of the manuscript may be limited to 20-25 A4 size pages (20 cm width – 26 cms height) through this is not binding.
7. PERMISSION – Authors must obtain permission from other publishers if they use already published and copywrighted materials.
8. COLOUR FIGURES – Cost of printing color figures will be charged to the author at the rate of U.S. $ 125 per plate of four color pages.
9. REPRINTS – The order for reprints should be placed at the time of submission of the article itself, so that it may be out for printing without loss of time, if found suitable for order for not less than 100 reprints is to be made.

MANUSCRIPT SUBMISSION:

1. Electronic – the www.bams.cm-uj.krakow.pl address is available for all authors to submit manuscript. The instruction is available at each step of submission process.
2. The pdf or doc file with figures and tables incorporated into the manuscript body can be sent by email – address: bams@cm-uj.krakow.pl

www.akademicka.pl

ISSN 1895-9091