Computer Assisted Mechanics and Engineering Sciences, 14: 681–693, 2007.

Searching for strong structural protein similarities with EAST

Dariusz Mrozek, Bożena Małysiak
Silesian University of Technology, Institute of Informatics
ul. Akademicka 16, 44-100 Gliwice, Poland

(Received in the final form September 18, 2007)

The exploration of protein conformation can be supported by methods of similarity searching that allow seeking the 3D patterns in a database containing many molecular structures. We developed a novel search method called EAST (Energy Alignment Search Tool), which serves as a tool for finding strong structural similarities of proteins. It differs from other algorithms that concentrate on fold similarities. We use the EAST to find protein molecules representing the same structural protein family and inspect conformational modifications in their molecular structures as an effect of biochemical reactions or environmental influences. The similarity searching is performed through the comparison and alignment of protein energy profiles. Energy profiles are received in the computational process based on the molecular mechanics theory. These profiles are stored in the special database (Energy Distribution Data Bank, EDB) and can be used later by the search engine to find similar fragments of protein structures on the energy level. In order to optimize the alignment path we use modified, energy-adapted Smith-Waterman method, which is one of the main phases of the EAST. The use of fuzzy techniques improves the fault tolerance of presented method and allows to measure the quality of the alignment. In the paper, we present the main idea of the EAST algorithm and brief discussion on its basic parameters. Finally, we give an example of the system usage regarding proteins from the RAB family that play an important role in intracellular reactions in living organisms.

Keywords: bioinformatics, proteins, soft computing, data mining, protein structure comparison

1. Introduction

Proteins are biological molecules that play very important role in all biological reactions in living cells. They are involved in many processes, e.g.: reaction catalysis, energy storage, signal transmission, maintaining of cell structure, immune response, transport of small biomolecules, regulation of cell growth and division [11]. Appropriate activity of proteins depends usually on many factors influencing their spatial structures. Therefore, the knowledge of protein structure allows to understand and to predict protein function in organisms. Furthermore, it allows to model the function, e.g. in drug design processes, and plan intended modifications of the structure to influence the function, e.g. in protein engineering. For all these reasons, the analysis of protein structures became very important to study the entire chains of complex processes proteins are involved in [5].

The very interesting and very important types of proteins are enzymes that catalyze cellular reactions. The activity of enzymes in catalytic reactions depends on the exposition of some typical parts of their 3D structures called active sites [11–13]. Conformation and chemical features of active sites allow to recognize and to bind substrates during the catalysis [5, 12, 19]. For these reasons, the study of active sites and spatial arrangement of their atoms is essential while analyzing activity of proteins in particular reactions [8, 10]. Having a group of proteins indicating strong similarity of selected structural modules we can explore the atomic arrangement of the interesting fragments taking part in respective reactions. The exploration can be supported by techniques of similarity searching. The similarity searching that base just on the protein amino acid sequence (like BLAST [3] or FASTA [25]) can be supportive but it does not consider geometrical features. More advanced, structural methods allow seeking the 3D structural patterns in a database containing many protein structures. This is a very complicated task since proteins are usually composed of thousands of atoms. Searching for similar proteins through the comparison of their spatial structures requires efficient and fully automated methods and become an area of dynamic researches in recent years. The most popular methodologies developed so far base on various representations of protein structures in order to reduce the search space, e.g. secondary structure elements (SSE) in VAST [14], locations of the Cα atoms of a protein body and intermolecular distances in DALI [17], aligned fragment pairs (AFPs) in CE [31], or 3D curves in CTSS [7]. However, these methods are more appropriate for homology modeling or function identification.

During the analysis of small parts of protein structures that can be active sites in cellular reactions it is required to use more precise methods of comparison and searching. For this reason, in our research we benefit from the dependency between the protein structure and the conformational energy of the structure [6]. Scientists made a huge progress developing theories that describe relation between chemical variety of protein sequence, structure and the shape of energy. There is also a group of algorithms of similarity searching based on the atomic potentials. They use Molecular Interaction Potentials (MIPs) or Molecular Interaction Fields (MIFs), e.g. [15, 18, 21, 28, 33]. MIPs/ MIFs are results of interaction energies between the considered compounds and relevant probes [28]. This group of algorithms is usually used in the process of drug design.

In our research, we calculate so called energy profiles (EPs) that are distributions of various potential energies along polypeptide chains of proteins. Energy profiles represent protein structures in the similarity search process. In this way, we reduce the complexity of the search process and we improve the efficiency of a "pattern vs. database" comparison. Although EPs and MIPs/MIFs base on the identical background theory of molecular mechanics [6] they are not the same. In the paper, we briefly describe the idea of energy profiles (Section 2), the method of similarity searching of proteins in different biological states EAST (Section 3), its parameters and usage (Sections 4 and 5, respectively).

2. Protein structure energy profiles

In our research on active sites of enzymes, we calculate distributions of conformational energy along amino acid chains of proteins. However, amino acids (peptides) are not directly considered in the calculation process. On the contrary, all performed calculations base on Cartesian coordinates of small groups of atoms constructing each peptide. Therefore, energy distributions can be seen as forms of representation of protein structures. The distribution of energy of various types along the protein/enzyme polypeptide chain may be very descriptive for protein function, activity and may reflect some distinctive properties.

Definition 1. Let R be an ordered set of m residues in the polypeptide chain $R = \{r_1 r_2 r_3 \ldots r_m\}$, and X_i^{ni} be a set of atomic coordinates building the ith residue r_i (n_i is a number of atoms of the ith residue r_i depending on type of the residue), then simplified protein structure can be expressed as the ordered set of small groups of atoms $X = \{X_1^{n_1} X_2^{n_2} X_3^{n_3} \ldots X_m^{n_m}\}$.

The *energy characteristics* or *energy distribution* E^t, where t is a type of energy, is an ordered set of energy values (*energy points*) e^t calculated for groups of atoms X_i^{ni} constructing the consecutive residues r_i in the protein polypeptide chain R:

$$E^t = \{e_1^t e_2^t e_3^t \ldots e_m^t\} \tag{1}$$

where: e_i^t is energy point of the ith residue, t is a type of energy (described later in the section).

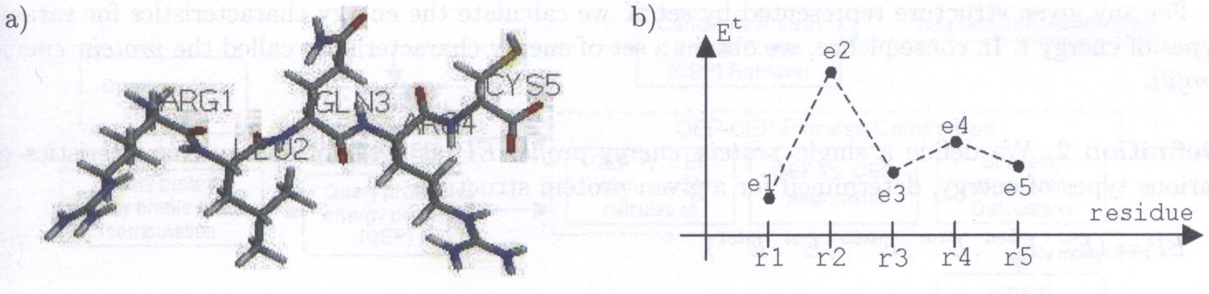

Fig. 1. The structure of example polypeptide containing five amino acids a), symbolic distribution of energy E^t for the molecule b)

The molecule presented in Fig. 1a is an example of simple polypeptide R with five peptides:

R	r_1	r_2	r_3	r_4	r_5
Residue type	ARG	LEU	GLN	ARG	CYS

The structure of the polypeptide is represented by set X. For the set X we calculate energy characteristics E^t (Fig. 1b):

X	X_1	X_2	X_3	X_4	X_5
E^t	e_1	e_2	e_3	e_4	e_5

In our approach, we compute energy characteristics base on the protein atomic coordinates retrieved from the macromolecular structure database Protein Data Bank (PDB) [4]. During the calculation we use TINKER [26] application of molecular mechanics and one of several standard force fields that are sets of physical-chemical parameters, usually Amber [9], and also Charmm [20] and Amoeba [27].

Energy characteristics represent protein structures in much reduced form of ordered sets of energy points (Fig. 1b), just like other algorithms as sets of positions of C_α atoms. The simplification of protein structure does not consider some elements of tertiary structure, like disulfide bonds. However, during the computation process of energy characteristics we include also cross-group interactions of atoms, so any energy point is calculated in the context of entire molecular structure. This reduces the search space during the comparison of two protein structures and during the search process that affects entire database of protein structures. The electrostatic energy characteristics for the real molecule of the Crystal Structure of Human RAB5A taken from the PDB is presented in Fig. 2. The RAB5A molecule is one of the proteins analyzed in our research on similarity searching and signal transduction phenomena [24].

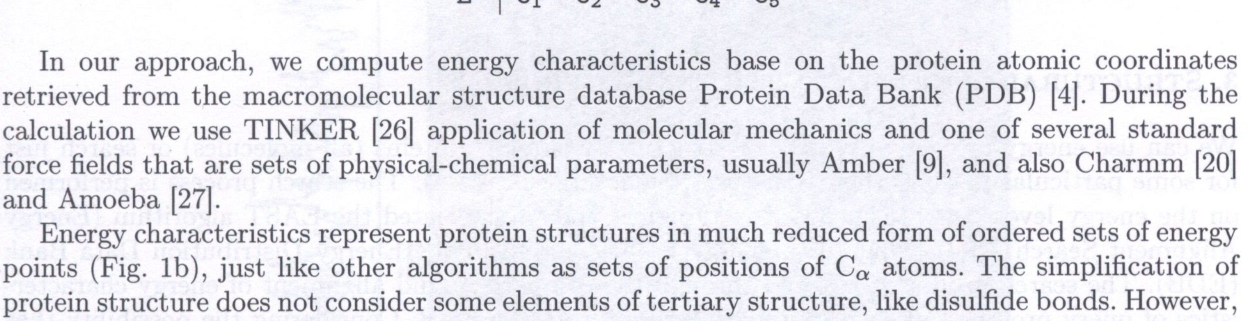

Fig. 2. Electrostatic energy characteristics for sample molecule 1N6H (Crystal Structure of Human RAB5A)

For any given structure represented by set X we calculate the energy characteristics for various types of energy t. In consequence, we obtain a set of energy characteristics called the *protein energy profile*.

Definition 2. We define a single *protein energy profile* EP as a set of energy characteristics of various types of energy, determined for a given protein structure,

$$EP = \langle E^{st}, E^{ben}, E^{tor}, E^{vdw}, E^{el}, E^{tot} \rangle \tag{2}$$

where: E^{st} denotes a distribution of the bond stretching energy which is a result of a deformation of optimal bond lengths, E^{ben} is a distribution of the angle bending component energy which is a consequence of changes of the optimal angles between each pair of adjacent covalent bonds, E^{tor} is a distribution of the torsional angle energy, E^{vdw} is a distribution of the van der Waals energy that is caused by correlated motion of the electron clouds of interacting atoms, E^{el} denotes a distribution of the electrostatic energy which is represented by the Coulomb's law. E^{tot} is a distribution of the total energy, which is a summary of all component energies in each residue.

We had computed complete energy profiles for about 5 000 protein structures from the PDB and we store them in a special database. To this purpose, we designed and developed the Energy Distribution Data Bank (EDB). The EDB is a foundation for the similarity search task. With the use of the EDB protein structures can be compared to each other based on their energy profiles, in order to find strong structural similarities, places of discrepancies, or possible mutations.

3. Structural alignment with energy profiles

We can use energy profiles to search for structurally similar proteins (all molecules) or search just for some particular parts of their structures (structural patterns). The search process is performed on the energy level. To this purpose, we designed and implemented the EAST algorithm (Energy Alignment Search Tool), which uses energy profiles stored in the Energy Distribution Data Bank (EDB). The search process is realized through the comparison and alignment of energy characteristics of query protein and each candidate protein from the EDB. Considering the possibility that any two compared molecules could diverge from common ancestor molecule through evolutionary changes over a time we use the alignments with gaps [2] in order to obtain better results. In the comparison we consider only one selected type of energy from each EP. The alignment can be thought as the juxtaposition of two energy characteristics that gives the highest number of identical or similar energy points (residues). A *similar* word means that energy points do not have to be identical but they should indicate similarity with the given range of tolerance. This tolerance is different for different types of energy considered in the search process. As a consequence of a suitable juxtaposition and based on appropriate similarity measures it is possible to evaluate a similarity degree of compared molecules and optionally find regions of dissimilarity. Therefore, the EAST algorithm examines identities, similarities (both are qualified as matches) and disagreements (mismatches) of compared energy points of both energy characteristics. Analogically to the nucleotide or amino acid sequences alignment, some mismatching positions and gaps can appear in the best alignment of energy characteristics.

General idea of the EAST algorithm is presented in Fig. 3 (the algorithm was described in our previous work [22] in more details). During the search process a user specifies a query protein (or a part of a molecule) with a known structure. This query-structure is then transformed into the energy profile (QEP). Query-molecule EP (QEP) can be compared with other energy profiles stored in the EDB. This is a pairwise comparison – each pair is formed by the query energy profile QEP and energy profile of the successive molecule from the EDB (candidate energy profile, CEP).

In the pairwise comparison, both QEP and CEP are represented by only one chosen energy characteristics (e.g. torsion angle energy, bond stretching energy, electrostatic, or other). During this

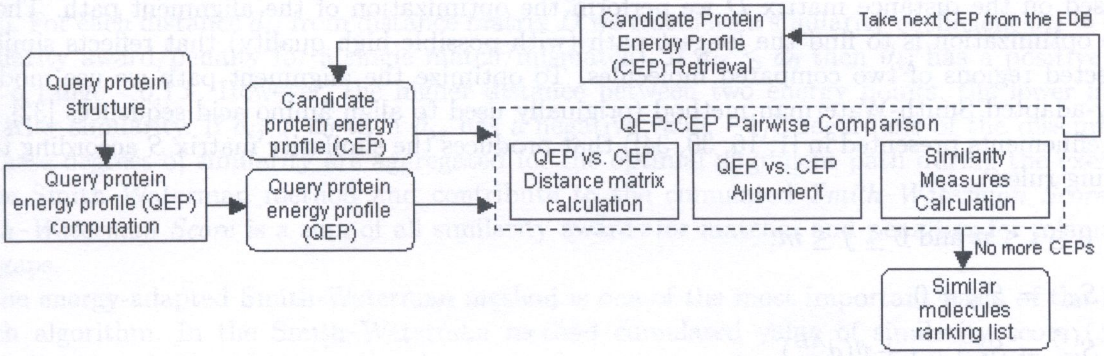

Fig. 3. Overview of the similarity search process with the EAST algorithm

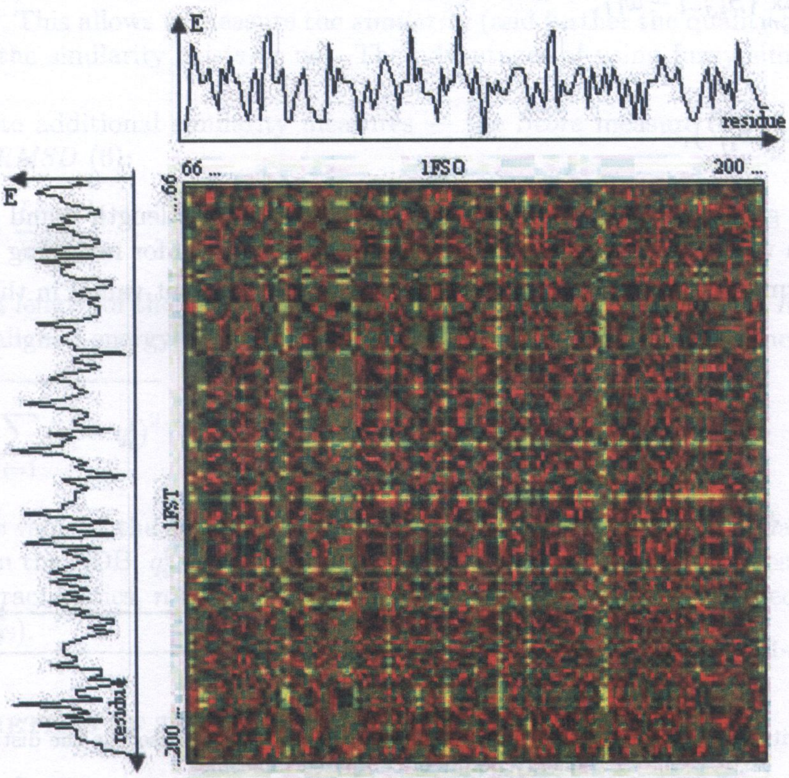

Fig. 4. Distance matrix for comparison of torsional angle energy characteristics for two similar molecules – query molecule 1FSO and candidate molecule 1FST from EDB. The matrix is visualized with the use of EAST toolkit – darker colors indicate stronger similarity of compared regions

phase we build a distance matrix D (Fig. 4) to check the distance between energy characteristics of query molecule and all candidate molecules.

The distance matrix allows to compare all energy points of the query molecule energy characteristics $E_A^t = (e_{A,1}^t e_{A,2}^t \ldots e_{A,n}^t)$ (part of the QEP) to all energy points of the candidate molecule energy characteristics $E_B^t = (e_{B,1}^t e_{B,2}^t \ldots e_{B,m}^t)$ (part of the CEP) and is filled according to the expression (3).

In the energy distance matrix D, the entry d_{ij}^{AB} denotes the distance between energy points of the ith residue of protein A and the jth residue of protein B, and can be expressed as

$$d_{t,ij}^{AB} = \left|e_{A,i}^t - e_{B,j}^t\right| \tag{3}$$

where: t is a type of energy that is considered in the searching.

Based on the distance matrix D we perform the optimization of the alignment path. The aim of the optimization is to find the longest path (with possible high quality) that reflects similarity of selected regions of two compared molecules. To optimize the alignment path we use modified, energy-adapted Smith-Waterman method (originally used to align amino acid sequences [32], with later refinements presented in [1, 16, 30, 34]) that produces the similarity matrix S according to the following rules:

for $0 \le i \le n$ and $0 \le j \le m$:

$$S_{i0} = S_{0j} = 0, \tag{4a}$$

$$S_{ij}^{(1)} = S_{i-1,j-1} + \vartheta(d_{ij}^{AB}), \tag{4b}$$

$$S_{ij}^{(2)} = \max_{1 \le k \le n} \{S_{i-k,j} - \omega_k\}, \tag{4c}$$

$$S_{ij}^{(3)} = \max_{1 \le l \le m} \{S_{i,j-l} - \omega_l\}, \tag{4d}$$

$$S_{ij}^{(4)} = 0, \tag{4e}$$

$$S_{ij} = \max_{v=1..4} \{S_{ij}^{(v)}\}, \tag{4f}$$

where: ω_k, ω_l are gap penalties for horizontal and vertical gaps of length k and l, respectively, and $\vartheta(d)$ is a function which takes a form of similarity award $\vartheta^{+}(d_{ij}^{AB})$ for matching energy points ($e_{A,i}^{t}$ and $e_{B,j}^{t}$) or a form of mismatch penalty $\vartheta^{-}(d_{ij}^{AB})$ (usually constant value) in the case of mismatch (Fig. 5).

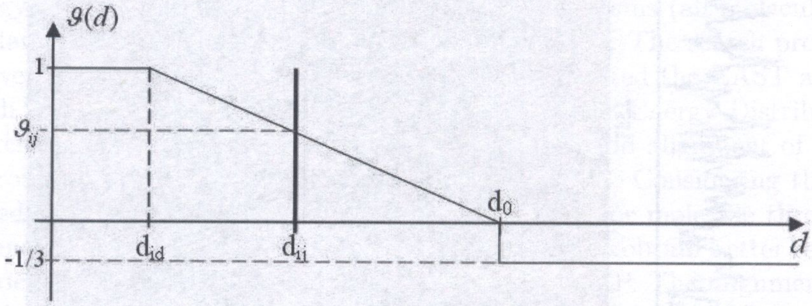

Fig. 5. Similarity award/penalty $\vartheta(d)$ allows to measure the similarity based on the distance between two energy points of compared molecules

The match/mismatch is resolved based on distance d_{ij}^{AB} between considered points (stored in the distance matrix D) and additional parameter d_0 called cutoff value. The **cutoff value** is the highest possible difference between energy points $e_{A,i}^{t}$ and $e_{B,j}^{t}$ when we treat these two points as similar. Cutoff value determines the range of tolerance for energy discrepancies.

The cutoff value can differ for various types of energy, e.g. for torsional angle energy default value was established based on *a priori* statistics to $d_0 = 1.4$ kcal/mole. Default settings for the energy-adapted Smith–Waterman method are (taken directly from the original method [32]): mismatch penalty $\vartheta^{-}(d_{ij}^{AB}) = -1/3$, affine gap penalty $\omega_k = 1.2 + 1/3 * k$ [2], where k is a number of gaps and d_{ij} is single cell value of the distance matrix D. The similarity award $\vartheta^{+}(d_{ij}^{AB})$ for a match depends on distance d_{ij}.

In the scoring function $\vartheta(d)$ we have to define two values: a cutoff value d_0, and identity threshold d_{id}. The **identity threshold** d_{id} is the highest possible difference between energy points when we treat two points as the same. The value of identity threshold was chosen arbitrary to 0.3–0.5 kcal/mole during observations of energy characteristics performed for many proteins.

The calculation of similarity between two energy points based on their distance is presented in Fig. 5. For each distance d_{ij} from distance matrix D we calculate a similarity coefficient $\vartheta_{ij} = \vartheta(d_{ij})$ (similarity award/penalty for a single match/mismatch). If $d_{ij} \leq d_0$ then ϑ_{ij} has a positive value from a range <0,1>. However, the higher distance between two energy points, the lower is their similarity similarity. If $d_{ij} > d_0$ then ϑ_{ij} has a negative value $-1/3$ regardless of the dissimilarity. All these degrees of similarity are aggregated for the optimal alignment path during the execution of the Smith–Waterman method and contribute to the cumulated *Smith–Waterman Score*. The *Smith–Waterman Score* is a sum of all similarity awards for matches and penalties for mismatches and gaps.

The energy-adapted Smith-Waterman method is one of the most important parts of the EAST search algorithm. In the Smith–Waterman method cumulated value of similarity score (*Smith-Waterman Score*) rises when considered energy points match to themselves (are equal or similar with the given range of tolerance), and decreases in regions of dissimilarity (mismatch). Moreover, the energy-adapted Smith–Waterman algorithm with the similarity award/penalty given by a function (not constant values) considers both, number of matching energy points and a quality of the match in the final alignment. This allows to measure the *similarity* (and further the quality of the alignment), not only to claim the similarity exists or not. The advantages of using fuzzy similarity award are presented in [23].

We also calculate additional similarity measures — the *Score* measure (5) and the *Root Mean Square Deviation RMSD* (6):

$$SCORE_{EAST} = \frac{Length_t}{RMSD_t} \tag{5}$$

where: $Length_t$ is a length of the alignment frame (in residues) and $RMSD_t$ is a *Root Mean Square Deviation* for the aligned energy characteristics calculated in the alignment frame;

$$RMSD_t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i^t - q_i^t)^2} \tag{6}$$

where: e_i^t is a single value of the energy (of the type t) for ith residue of the candidate protein energy characteristics from the EDB, q_i^t is a value of the energy (of the type t) for ith residue of the query protein energy characteristics, n is a number of aligned residues (without gaps, equal to alignment frame in some cases).

4. Main parameters of the EAST algorithm

The contribution of particular component characteristics in the search process can differ. Thus, their support for the process is not the same. Preliminary studies show some characteristics are more predestined to distinguish one group of proteins from other molecules. In the section we discuss the usage of particular component characteristics in the search process run with the EAST algorithm. In addition, we present results of our research on basic parameters of our method.

The cutoff value d_0 and identity threshold d_{id} are different for particular component characteristics and can differ for particular group of proteins. In the search process carried out with our EAST method these parameters can change according to the users' will – users have an option to specify any value. However, based on research and a priori statistics (Table 1) we established a set of default values of the cutoff and identity threshold for all component characteristics contributing to energy profile (Table 2).

In the first phase, we carried our studies with the use of groups of molecules representing the same proteins in different biological states. These groups were separated based on the protein annotations, sequence similarity or literature data. For all possible pairs of molecules in the particular group of similar molecules we compared each type of energy characteristics and we calculated values of the

Table 1. Average values of the RMSDs for each type of energy [kcal/mole]

Component energy of EP	MRMSD_0	MRMSD_5	MRMSD_{10}	MRMSD_{15}
bond stretching	1.617	0.950	0.622	0.596
angle bending	4.104	1.599	1.212	1.037
torsional angle	1.891	1.421	1.170	1.014
van der Waals	7.106	5.101	3.771	3.114
electrostatic	8.413	6.093	5.145	4.003

Table 2. Recommended ranges of cutoff values and default values of the cutoff and threshold identity for searching with each type of the energy characteristics of the EP

Component energy of EP	Cutoff value [kcal/mole]		Default value of cutoff [kcal/mole]	Default value of identity threshold [kcal/mole]
	from	to		
bond stretching	0.6	1.6	1.0	0.3
angle bending	0.8	4.0	1.6	0.3
torsional angle	0.8	1.9	1.4	0.3
van der Waals	3.1	7.1	5.1	0.5
electrostatic	2.4	8.4	6.1	0.5

Root Mean Square Deviation (RMSD). Calculations were done for all energy points of the compared pair of characteristics. Therefore, for one pair of molecules and selected type of energy we obtained the value of the RMSD_0 . In the next step, we cut off 5% pairs of energy points indicating the highest differences and we computed the RMSD once again (we got the RMSD_5). Twice more we cut off 5% pairs of energy points indicating the highest differences and each time we computed the RMSD. As a result we got RMSD_{10} and RMSD_{15} . After calculations for all possible pairs of molecules within groups, we calculated average of the RMSD_0 , RMSD_5 , RMSD_{10} , RMSD_{15} for each type of energy regardless of the group of molecules (across the groups). In consequence, we got the values of the Mean RMSD (MRMSD_0 , MRMSD_5 , MRMSD_{10} , MRMSD_{15}). Based on the values we assumed boundary values of the cutoff d_0 (from MRMSD_{15} to MRMSD_0) and the default values of the cutoff d_0 (MRMSD_5). The values of identity threshold d_{id} were chosen arbitrary.

In the second phase, we performed a series of search processes and observations of the distance matrices changing the cutoff value with the 0.1 kcal/mole within the assumed ranges. We also carried out additional search processes beyond the boundary values of the ranges. In this way, we decreased the left boundary value for angle bending energy and torsional angle energy from 1.0 to 0.8 kcal/mole and charge-charge (electrostatic) energy from 4.0 to 2.4 kcal/mole. After the second phase of the research we obtained a final set of boundary and default values of the cutoff d_0 (Table 2).

The wide range of the cutoff value for the van der Waals component energy is flexible (Table 2). The usage of the component energy in the search process is limited due to high fluctuations of energy values as a consequence of some structural changes observed in the groups of the same molecules. We noticed strong energy changes for small conformational discrepancies. On the other hands, there are regions of the energy characteristics where the distribution of the van der Waals component energy changes in the small range. For all these reasons, it is difficult to set appropriate, common cutoff values for this type of energy.

Generally, all default values of the EAST algorithm can be modified before the execution of the search process. Changes of the default cutoff value d_0 influence the sensitivity of our method. If we set higher value of the d_0 the algorithm will qualify more energy points as similar. This will also increase the similarity measures for compared molecules and spread the similarity frames. However, too high value of the d_0 (beyond the recommended range) causes a danger to get more molecules in the result set that are not structurally similar but indicate this with high values of similarity measures. On the other hand, when we set the d_0 beneath the left endpoint of the cutoff range we will observe narrow frames of alignment, a decrease of the percentage of matching energy points and low values of the *Score* and *SW-Score* measures.

5. Experiments on protein molecules

We heavily exploit presented algorithm in our research on active sites of proteins (parts of protein structures). Moreover, the dependency between a protein structure and its conformational energy allows to observe: conformational changes of proteins in particular reactions, conformational discrepancies in protein structural mutants, and compare predicted proteins to existing molecules in order to verify results. In our experiments we performed search tests for 100 molecules (or their parts) against the EDB database containing about 5 000 EPs of protein structures from the Protein Data Bank (containing more than 40 000 proteins). The subset was chosen arbitrary. The size of the EDB containing profiles generated with 3 different force fields is approximately 600 MB.

Beneath we present results and analysis of one of the search process performed in our tests. The process was run for query molecule representing the whole protein. However, it can be executed just for the smaller parts of protein structures representing active sites, biological functional domains, 3D motifs or any other structural pattern. The presented example concerns proteins from the RAB family that are members of the bigger group of GTPases – particles that have the ability to bind and hydrolyze GTP molecules. Therefore, proteins of the RAB family play an important role in intracellular reactions of living organisms. They are elements of the signal pathways where they serve as molecular controllers in the switch cycle between active form of the GTP molecule and inactive form of the GDP. In Fig. 6 we can see results of the similarity search process performed for the 1N6H molecule (Crystal Structure Of Human RAB5A). Results are sorted according to the *Score* similarity measure.

```
Best results for job: 2006-08-04 13:13:55
Cut-off: 6.1; id threshold: 0.5; energy type: Charge-charge
S-W type: Fuzzy; mismatch: -0.3334; gap open: 1.2; gap ext.: 0.3334

PDBID Chain Length Matches Match%   RMSD    Score S-W Score
----- ----- ------ ------- ------ ------- -------- ---------
 1N6R     A    159     154     96   2.091    73.64    132.84
 1N6L     A    162     155     95   2.247    68.98    134.78
 1N6O     A    128     126     98   1.914    65.85     98.01
 1N6K     A    161     155     96   2.365    65.53    130.45
 1N6I     A    154     146     94   2.391    61.06    121.69
 1HUQ     A    146     140     95   2.295    61.01    114.89
 1N6P     A    136     129     94   2.269    56.85    104.47
 1N6N     A    146     137     93   2.488    55.06    110.64
 1TU3     A    142     138     97   2.763    49.95     96.85
 1R2Q     A     43      38     88   2.914    13.04     32.13
---------------------------
 1GRN     A     28      27     96   3.259     8.29      8.96
 1GUA     A     38      32     84   3.960     8.08      4.87
 1R8Q     A     26      25     96   3.357     7.45      5.58
```

Fig. 6. Results of the similarity search process with the use of the EAST algorithm for human RAB5A molecule 1N6H (Crystal Structure Of Human RAB5A). Parameters – cutoff value: 6.1 kcal/mole, identity threshold: 0.5 kcal/mole, energy type: electrostatic

Results can be interpreted based on similarity measures: *Score* and *Smith–Waterman Score* (*SW-Score*) – the higher value the higher similarity, *RMSD* – the lower value the better quality of the alignment, and based on output parameters: *Length* – alignment frame, *Matches* – number of aligned energy points, *Match%* – percentage of matching positions in the alignment frame (the higher the better).

Molecules above the horizontal line (Fig. 6) were qualified as similar. The line was inserted manually after the search process and verification based on similarity measures, annotations of molecules and literature. Analyzing the result set we made the following interesting observations:

The molecule 1R2Q (in the result set in Fig. 6) has weak values of similarity measures. Nevertheless, it is in the group of similar molecules. In fact, it is just a part of the RAB5A – so called GTPase domain, responsible for the signal transfer in cells through GTP/GDP binding.

The molecule 1HUQ in the group of similar molecules has the same function in mouse organism (*mus musculus*) as query molecule 1N6H in human. The energy characteristics alignment of query molecule 1N6H and candidate molecule 1HUQ is presented in Fig. 7. Energy characteristics for both molecules have many matching energy points indicating structural similarity. Some parts of these characteristics cover each other what verifies a good quality of the alignment. There are also some parts indicating small conformational differences in the compared structures (e.g. residues 104–110).

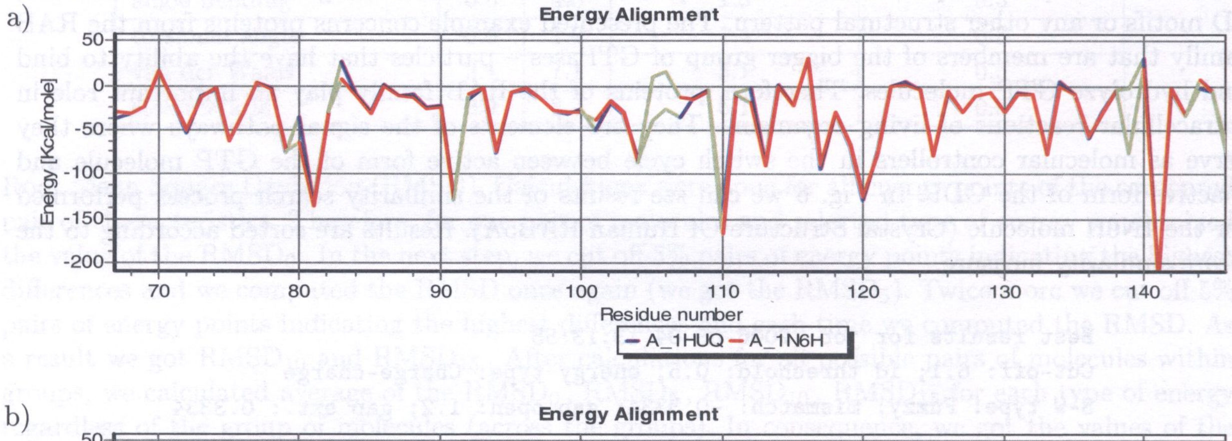

Fig. 7. Alignment of energy characteristics with the EAST toolkit for query molecule 1N6H (Crystal Structure of Human RAB5A) and resultant molecule 1HUQ (Crystal Structure of the Mouse Monomeric GTPase RAB5C): a) line representation, b) stairs representation. Visible parts: residues 68–145. Grey color (bright line) indicates mismatching parts, no gaps visible

Comparison of the 3D structures of query molecule 1N6H and resultant molecule 1HUQ is presented in Figs. 8 and 9.

There are structure mutants of the RAB5A particle in the result set (Fig. 6): 1N6I (Crystal Structure of Human RAB5A A30P Mutant Complex with GDP), 1N6K (Crystal Structure of Human RAB5A A30P Mutant Complex with GDP and Aluminum Fluoride), 1N6L (Crystal Structure of Human RAB5A A30P Mutant Complex with GTP), 1N6N (Crystal Structure of Human RAB5A A30R Mutant Complex with GPPNHP), 1N6O (Crystal Structure of Human RAB5A A30K Mutant Complex with GPPNHP), 1N6P (Crystal Structure of Human RAB5A A30E Mutant Complex with

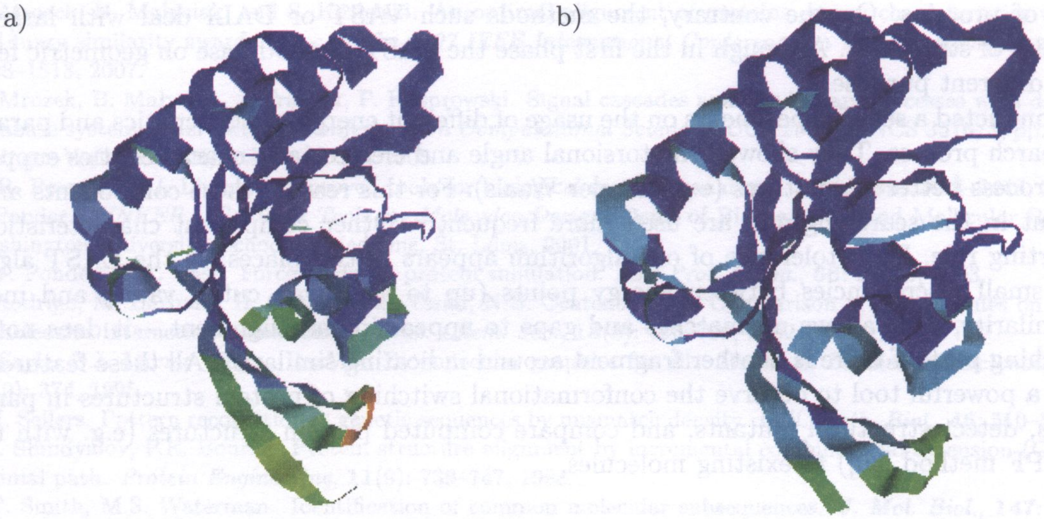

Fig. 8. Comparison of the 3D structures (ribbon representation) of molecules: a) query 1N6H (Crystal Structure of Human RAB5A), b) resultant 1HUQ (Crystal Structure of the Mouse Monomeric GTPase RAB5C). Visualization made with the use of the RasMol program [29]

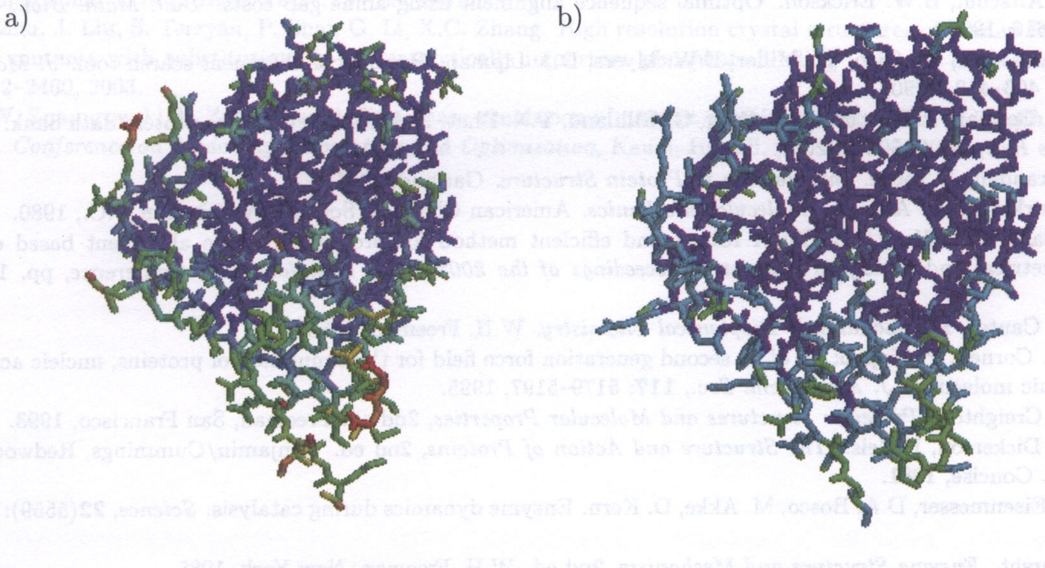

Fig. 9. Comparison of the 3D structures (sticks representation) of molecules: a) query 1N6H (Crystal Structure of Human RAB5A), b) resultant 1HUQ (Crystal Structure of the Mouse Monomeric GTPase RAB5C). Visualization made with the use of the RasMol program [29]

GPPNHP), 1N6R (Crystal Structure of Human RAB5A A30L Mutant Complex with GPPNHP). Observations of energy characteristics confirm a single mutation appears in residue Ala^{30} in the part of the molecule called *P-loop* (residues ^{27}GESAVGKS^{34}) causing structural changes in the *switch region I* (residues 47–65) and *switch region II* (residues 77–93) and the decrease in the intrinsic GTPase activity. These effects were described in the literature [35] in more details.

6. Discussion and conclusions

Similarity searching is one of the most frequent tasks performed in bioinformatics database systems. We developed a novel algorithm of searching for structurally similar proteins (or their parts) with the use of energy profiles that represent structures of proteins. The algorithm differs from other methods of structure similarity searching. The EAST focuses on strong similarities in the same families of proteins. On the contrary, the methods such VAST or DALI deal with large (fold) similarities of structures. Although in the first phase the EAST method base on geometric features, it has a different purpose.

We conducted a set of experiments on the usage of different energy characteristics and parameters in the search process. They showed the torsional angle and electrostatic characteristics support the search process better than others (e.g. van der Waals). For this reason, these components are more significant in the searching and are used more frequently. Other component characteristics have a supporting role. Fault-tolerance of our algorithm appears in two places: 1) the EAST algorithm accepts small discrepancies between energy points (up to the given cutoff value) and measures their similarity, 2) it allows mismatches and gaps to appear in the alignment – it does not reject mismatching parts if there is another fragment around indicating similarity. All these features cause we have a powerful tool to observe the conformational switching of protein structures in particular reactions, detect structural mutants, and compare computed protein structures (e.g. with the use of the NPF method [36]) to existing molecules.

References

[1] S.F. Altschul, B.W. Erickson. Locally optimal subalignments using nonlinear similarity functions. *Bull. Math. Biol.*, **48**: 633–660, 1986.

[2] S.F. Altschul, B.W. Erickson. Optimal sequence alignment using affine gap costs. *Bull. Math. Biol.*, **48**(5-6): 603–616, 1986.

[3] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, **215**: 403–410, 1990.

[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, *et al.* The protein data bank. *Nucleic Acids Res.*, **28**: 235–242, 2000.

[5] C. Branden, J. Tooze. *Introduction to Protein Structure.* Garland, 1991.

[6] U. Burkert, N.L. Allinger. *Molecular Mechanics.* American Chemical Society, Washington D.C., 1980.

[7] T. Can, Y.F. Wang. CTSS: A robust and efficient method for protein structure alignment based on local geometrical and biological features. *Proceedings of the 2003 IEEE Bioinformatics Conference*, pp. 169–179, 2003.

[8] C.R. Cantor, P.R. Schimmel. *Biophysical Chemistry.* W.H. Freeman, 1980.

[9] W.D. Cornell, P. Cieplak, *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **117**: 5179–5197, 1995.

[10] T.E. Creighton. *Proteins: Structures and Molecular Properties*, 2nd ed. Freeman, San Francisco, 1993.

[11] R.E. Dickerson, I. Geis. *The Structure and Action of Proteins*, 2nd ed. Benjamin/Cummings, Redwood City, Calif. Concise, 1981.

[12] E.Z. Eisenmesser, D.A. Bosco, M. Akke, D. Kern. Enzyme dynamics during catalysis. *Science*, **22**(5559): 1520–3, 2002.

[13] A. Fersht. *Enzyme Structure and Mechanism*, 2nd ed. W.H. Freeman, New York, 1985.

[14] J.F. Gibrat, T. Madej, S.H. Bryant. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**(3): 377–385, 1996.

[15] P.J. Goodford. Computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**: 849–857, 1985.

[16] O. Gotoh. An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**: 705–708, 1982.

[17] L. Holm, C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**(1): 123–38, 1993.

[18] H. Ji, H. Li, M. Flinspach, T.L. Poulos, R.B. Silverman. Computer modeling of selective regions in the active site of nitric oxide syntheses: Implication for the design of isoform-selective inhibitors. *J. Med. Chem.*, **46**: 5700–5711, 2003.

[19] H. Lodish, A. Berk, S.L. Zipursky, *et al. Molecular Cell Biology*, 4th ed. W.H. Freeman, NY, 2001.

[20] A.D. MacKerrell Jr., *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**: 3586–3616, 1998.

[21] K. Moffat, V. Gillet, M. Whittle, G. Bravi, A. Leach. Similarity searching using molecular interaction fields. *Proc. of the 7th International Conference on Chemical Structures*, 2005.

[22] D. Mrozek, B. Małysiak, S. Kozielski. EAST: Energy Alignment Search Tool. In: L. Wang *et al.*, eds., *Proc. of the 3rd IEEE International Conference on Fuzzy Systems and Knowledge Discovery*, LNCS 4223, pp. 696–705. Springer-Verlag, Xi'an, China, 2006.

[23] D. Mrozek, B. Małysiak, and S. Kozielski. An optimal alignment of proteins energy characteristics with crisp and fuzzy similarity awards. *Proc. of the 2007 IEEE International Conference on Fuzzy Systems*, London, UK, 1508–1513, 2007.

[24] D. Mrozek, B. Małysiak, J. Frączek, P. Kasprowski. Signal cascades analysis in nanoprocesses with distributed database system. *International Conference on Computational Science (ICCS 2005)*, LNCS 3516/3, pp. 334–341. Springer-Verlag GmbH, Atlanta, USA, 2005.

[25] W.R. Pearson, D.J. Lipman. Improved tools for biological sequence analysis. *PNAS*, **85**: 2444–2448, 1998.

[26] J. Ponder. *TINKER – Software Tools for Molecular Design.* Dept. of Biochemistry and Molecular Biophysics, Washington University, School of Medicine, St. Louis, 2001.

[27] J.W. Ponder, D.A. Case. Force fields for protein simulation. *Adv. Prot. Chem.*, **66**: 27–85, 2003.

[28] J. Rodrigo, M. Barbany, H. Gutiérrez-de-Terán, N.B. Centeno, *et al.* Comparison of biomolecules on the basis of molecular interaction potentials. *J. Braz. Chem. Soc.*, **13**(6): 795–799, 2002.

[29] R. Sayle, E.J. Milner-White. RasMol: Biomolecular graphics for all. *Trends in Biochemical Sciences (TIBS)*, **20**(9): 374, 1995.

[30] P.H. Sellers. Pattern recognition in genetic sequences by mismatch density. *Bull. Math. Biol.*, **46**: 510–514, 1984.

[31] I.N. Shindyalov, P.E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, **11**(9): 739–747, 1998.

[32] T.F. Smith, M.S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, **147**: 195–197, 1981.

[33] D.A. Thorner, D.J. Wild, P. Willett, P.M. Wright. Similarity searching in files of three-dimensional chemical structures: flexible field-based searching of molecular electrostatic potentials. *J. Chem. Inf. Comput. Sci.*, **36**: 900–908, 1996.

[34] M.S. Waterman, M. Eggert. A new algorithm for best subsequence alignments with applications to tRNA-rRNA comparisons. *J. Mol. Biol.*, **197**: 723–728, 1987.

[35] G. Zhu, J. Liu, S. Terzyan, P. Zhai, G. Li, X.C. Zhang. High resolution crystal structures of human Rab5a and five mutants with substitutions in the catalytically important phosphate-binding loop. *J. Biol. Chem.*, **278**: 2452–2460, 2003.

[36] A.W. Znamirowski, L. Znamirowski. Two-phase simulation of nascent protein folding. *Proc. of the 4th IASTED Int. Conference on Modelling, Simulation and Optimization*, Kauai, Hawaii, pp. 293–298, 2004.