

An Improved Method for Protein Similarity Searching by Alignment of Fuzzy Energy Signatures*

Dariusz Mrozek[†], Bożena Małysiak-Mrozek

*Institute of Informatics, Silesian University of Technology
Akademicka 16, 44-100 Gliwice, Poland*

*E-mail: Dariusz.Mrozek@polsl.pl, Bożena.Małysiak@polsl.pl
http://www.polsl.pl*

Accepted: 12-01-2009

Received: 19-09-2010

Abstract

Describing protein structures in terms of their energy features can be a key to understand how proteins work and interact to each other in cellular reactions. This can be also a base to compare proteins and search protein similarities. In the paper, we present protein comparison by the alignment of protein energy signatures. In the alignment, components of energy signatures are represented as fuzzy numbers. This modification improves the decision making while establishing the alignment path and guarantees the approximate character of the method, at the same time. The effectiveness of the developed alignment algorithm is tested by incorporating it in the new FS-EAST method (Fuzzy Signatures – Energy Alignment Search Tool), which allows to seek structurally similar regions of proteins.

Keywords: bioinformatics, protein structure, similarity searching, force fields, molecular mechanics, fuzzy numbers.

1. Introduction

Estimating similarity between two or more protein structures requires comparative techniques such as alignment that allow for the character of the information that has to be processed. The similarity searching is a fault-tolerant process, which allows seeking molecules with identical or similar structures to the given query molecule. Furthermore, the similarity searching may concern the whole structure of a protein or just selected protein regions and it must consider evolutionary changes and possible mutations that could appear in protein structures through many years.¹⁻⁴

Alignment is a valuable tool for the comparison of two or more sequences of data. The alignment is a way of arranging sequences to identify mutual similarities of their particular elements. The purpose of the process is to find and show similarity relationships between elements of two compared sequences. Gaps and mismatches can occur between elements in the final alignment with the intention that identical or similar elements can be assigned as corresponding.^{5,6} Since proteins are built up with hundreds amino acids and thousands of atoms, for the efficiency reasons they are usually represented in much reduced form in the alignment process. Two most popular forms of the representation include: amino acid sequences, if the

*Scientific research supported by the Ministry of Science and Higher Education, Poland in years 2008-2010. Grant No. N N516 265835: *Protein Structure Similarity Searching in Distributed Multi Agent System*

[†]To whom the entire correspondence should be directed: Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland, e-mail: Dariusz.Mrozek@polsl.pl

comparison occurs at the primary structure level (sequence alignment), and sequences of alpha carbon positions, if the comparison occurs at the tertiary structure level (structural alignment).^{7,8}

In our research on protein activities in cellular reactions Refs. 9-12, we usually seek regions that are biologically important modules, like active sites of enzymes^{1,7}, or we evaluate a quality of predicted protein structures. For this purpose, we have developed the EAST method of similarity searching.^{13,14} The EAST stands for Energy Alignment Search Tool. This means the EAST represents protein structures as sequences of different energy features, called *energy profiles*, and consequently, it uses the alignment process during the similarity searching. Before the alignment we fuzzify the input sequences of energy features. The fuzzification influences the decision making during the calculation of values for the similarity matrix in the alignment phase.

In the paper we present an improved alignment of energy profiles (Section 5), which we incorporated into the new FS-EAST method (Fuzzy Signatures – Energy Alignment Search Tool). The FS-EAST is the successor of the EAST method. To ensure the approximate character of the similarity searching we treat energy profiles as sequences of fuzzy energy signatures. In consequence, we have eliminated some weaknesses of previous versions of the EAST method. Performance tests and discussion on the FS-EAST algorithm with the new alignment method are presented in Section 6. Before the detailed description of the alignment method, we give a short overview of popular methods used in the area of protein similarity searching in Section 2. In Section 3, we present a brief idea of protein construction followed by the explanation, how we represent protein structures as sequences of fuzzy energy signatures (Section 4).

2. Related Works

Although, protein structure similarity searching has been explored for the last two decades, efficient and accurate methods of protein alignment are still a challenging topic. Similarity searching methods developed so far use various representations of protein molecules, depending on the purpose the method will be used for. Nevertheless, existing algorithms for protein similarity searching are usually grounded in principles of approximate retrieval and heuristics. For molecules,

such as proteins, two trends can be distinguished: (1) similarity searching based on the alignment of protein amino acid sequences, (2) similarity searching based on the alignment of three-dimensional molecular structures.

In the first group, there are two leading competitive methods – FASTA¹⁵ and BLAST¹⁶. Both methods apply the paradigm of cutting sequences into shorter fragments called *words*. Having the list of words, they find entire proteins or their regions with the best word hits and use dynamic programming to establish the optimal alignment of input sequences. Similarity searching by protein sequence is usually one of the first steps in many studies on biological molecules, e.g., gene or protein identification.

In the second group, protein structures, originally represented by atomic coordinates and interatomic covalent bonds, are first transformed to the simpler form in order to reduce the search space. There are three main reasons of this:

1. protein structures are very complex; they are usually composed of thousands of atoms;
2. the similarity searching is usually carried through the comparison of a given structure to all structures in a database;
3. the number of protein structures in databases, like Protein Data Bank (PDB)¹⁷ rises exponentially every year and is now 61 695 (November 24, 2009).

The reduced representation of proteins depends on the method. E.g. well-known VAST¹⁸ algorithm identifies secondary structure elements (SSE) in compared proteins and maps them into set of representative vectors. Afterwards, it tries to match pairs of vectors using the bipartite graph. The SSE representation of protein structures is also used in the comparison method applied in the LOCK2.¹⁹

The idea of popular DALI²⁰ method bases on the calculation of a distance matrix for each compared protein. Single matrix includes intramolecular distances between coordinates of the C_α atoms representing each residue of the protein in the comparison process. The DALI method seeks similar regions in distance matrices of two compared proteins. The DALI belongs to the group of methods known as clustering-based methods.

The CE²¹ algorithm use the *combinatorial extension* of alignment path formed by *aligned fragment pairs* (AFPs) of both compared proteins. AFPs are fragments of both structures indicating the clear structural

similarity and are described by local geometrical features. The distance between corresponding residues in two AFPs is calculated considering the positions of their C_{α} atoms. The idea of AFPs is also used in the FATCAT.^{22,23}

The CTSS²⁴ method is grounded in the theory of differential geometry on 3D space curve matching. It first calculates splines to approximate positions of the C_{α} atoms in compared proteins and afterwards, for each residue it calculates shape signatures that incorporate curvature, torsion and secondary structure type. The pairwise comparison is performed with the use of distance matrices that store the distance between shape signatures. Different shape features are also used in the PFSC method.²⁵

Presented methods are frequently used in the protein function identification, homology modeling or protein structure prediction with the use of the threading.

There is also a group of algorithms that use Molecular Interaction Potentials (MIPs) or Molecular Interaction Fields (MIFs), like these presented in Refs. 28-30. They use atomic coordinates of biological molecules to calculate component, nonbonded interaction energies. MIP/MIFs are results of interaction energies between the considered compounds and relevant probes. MIPs are often calculated with the popular GRID³¹ program and are used for the comparison of series of compounds displaying related biological behavior. However, this group of very precise algorithms is not appropriate for our purposes. MIPs-based methods are frequently used to study ligand-receptor interactions, which is crucial for the pharmacology and development of new drugs. Moreover, since MIPs-based methods usually represent molecular structures in the form of 2D grids, they are too computationally complex and time-consuming for big molecules.

From the viewpoint of the computational procedure, the newly developed EAST uses similar techniques to methods mentioned in the first and the second group. However, oppositely to rough methods, like VAST, CE, DALI or other that focus on the fold similarity, the EAST concentrates on stronger regional similarity of protein substructures and grasps small structural deformations. In the view of the structure representation, our method is similar to techniques that use MIPs/MIFs. However, the EAST is less computationally complex.

3. Protein Construction

Analyzing their general construction, proteins are macromolecules with the molecular mass above 10 kDa ($1 \text{ Da} = 1.66 \times 10^{-24} \text{ g}$) built with amino acids (>100 amino acids, aa). Amino acids are linked in linear chains by peptide bonds.¹ In the construction of proteins we can distinguish four description (or representation) levels: primary structure, secondary structure, tertiary structure and quaternary structure. The last three levels define the protein conformation or protein spatial structure, which is determined by location of atoms in the 3D space.² The biochemical analysis is usually carried on one of the description levels.

Primary structure is defined by amino acid sequence in protein linear chain.³ Example of sequence of myoglobin molecule is presented in Fig. 1. Each letter in a sequence corresponds to one amino acid in the protein chain. There are 20 standard amino acids found in most living organisms.⁴

Secondary structure describes spatial arrangement of amino acids located closely in the sequence. This description level distinguishes in the spatial structure some characteristic, regularly folded substructures.^{1,2} The examples of the secondary structures are α -helices (visible in Fig. 2a) and β -sheets.

Tertiary structure (Fig. 2) refers to spatial relationships and mutual arrangement of amino acids located closely and distantly in the protein sequence.⁴ Tertiary structure describes the configuration of a protein structure caused by additional, internal forces, like: hydrogen bonds, disulfide bridges, attractions between positive and negative charges, and hydrophobic and hydrophilic forces. This description level characterizes the biologically active spatial conformation of proteins.³

Quaternary structure refers to proteins made up of more than one amino acid chain (Fig. 3). This level describes the arrangement of subunits and the type of their contact, which can be covalent or not covalent.⁴

```
>1MBN:A|PDBID|CHAIN|SEQUENCE
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFRDFK
HLKTEAEMKASEDLKKGVTVLTAALGAILKKGHHEAELKPLAQSHA
TKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGGAMNKALELFRKD
IAAKYKELGYQG
```

Fig. 1. Protein sequence of the myoglobin (PDB ID: 1MBN) in the FASTA format.

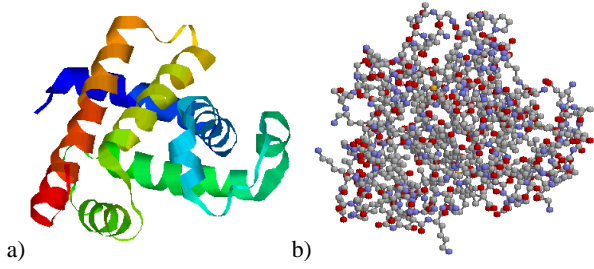


Fig. 2. Spatial structure of the myoglobin (PDB ID: 1MBN): a) secondary structure representation, b) atomic representation of the tertiary structure in the RasMol viewer.³²

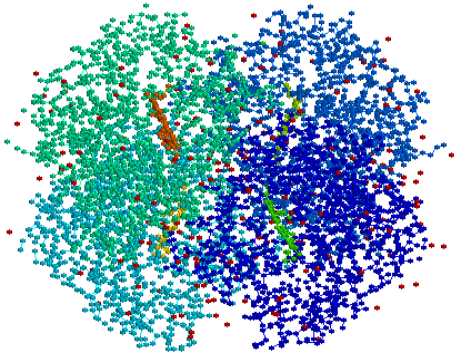


Fig. 3. Quaternary structure of the human hemoglobin (PDB ID: 4HHB, all four chains) in the atomic representation, colored by chain.

4. Energy Profiles for Protein Structures

Let's consider a simple protein P built up with m amino acids (residues). The primary structure of the protein P will have the following form: $P=(p_1, p_2, \dots, p_m)$. The tertiary structure (spatial structure) will be symbolized by a set of N atoms A^N . The structure A^N can be also represented as a sequence: $A^N=(A_1^{n_1}, A_2^{n_2}, \dots, A_m^{n_m})$, where each $A_i^{n_i}$ is a subgroup of atoms corresponding to the i^{th} residue p_i of the protein P , n_i is a number of atoms in the i^{th} residue p_i depending on the type of the residue, and:

$$A^N = \bigcup_{i=1}^m A_i^{n_i}, \text{ and } N = \sum_{i=1}^m n_i. \quad (1)$$

Locations of atoms in the structure A^N are described in the 3D space by the (x, y, z) Cartesian coordinates.

The method that we have developed benefits from the dependency between the protein structure and the conformational, potential energy of the structure.^{33,34} In our

research, we calculate *energy profiles* E^{Ξ} , which describe energy properties for all substructures $A_i^{n_i}$ in the amino acid chain of the protein structure A^N . Energy profiles are calculated according the rules of molecular mechanics^{33,34} and on the basis of Cartesian coordinates of small groups of atoms that constitute each peptide p_i . Therefore, energy profiles represent energy features distributed in protein structures.

The energy profile for a single protein structure A^N can be presented in form of matrix:

$$E^{\Xi} = [\bar{e}_1 \bar{e}_2 \bar{e}_3 \dots \bar{e}_m], \quad (2)$$

where each \bar{e}_i is an *energy signature*, which is a vector of energy features for the i^{th} peptide p_i (Fig. 4) and respective subgroup of atoms $A_i^{n_i}$ of the protein P :

$$\bar{e}_i = (e_i^{st}, e_i^{ben}, e_i^{tor}, e_i^{vdw}, e_i^{cc})^T \quad (3)$$

Vector components correspond to appropriate energy types for the i^{th} peptide p_i : e_i^{st} represents bond stretching energy feature, e_i^{ben} represents angle bending energy feature, e_i^{tor} represents torsional angle energy feature, e_i^{vdw} represents van der Waals energy feature, e_i^{cc} represents electrostatic (charge-charge) energy feature.

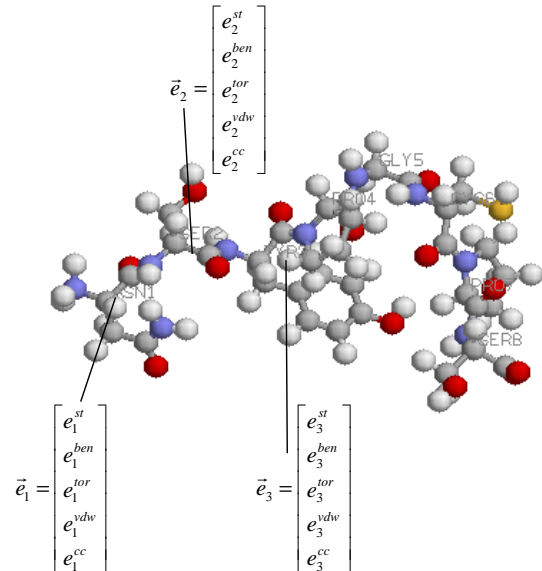


Fig. 4. Part of sample protein spatial structure (eight amino acids) with representative energy signatures for consecutive residues.

These components are calculated with the use of molecular mechanics methods:

- bond stretching (e^{st})

$$e^{st}(A^n) = \sum_{i=1}^{bonds} \frac{k_i}{2} (d_i - d_i^0)^2, \quad (4)$$

where: k_i is a bond stretching force constant, d_i is a distance between two atoms, d_i^0 is an optimal bond length;

- angle bending (e^{ben})

$$e^{ben}(A^n) = \sum_{i=1}^{angles} \frac{k_i}{2} (\theta_i - \theta_i^0)^2, \quad (5)$$

where: k_i is a bending force constant, θ_i is an actual value of the valence angle, θ_i^0 is an optimal valence angle;

- torsional angle (e^{tor})

$$e^{tor}(A^n) = \sum_{i=1}^{torsions} \frac{V_m}{2} (1 + \cos(m\omega - \gamma)), \quad (6)$$

where: V_m denotes the height of the torsional barrier, m is a periodicity, ω is the torsion angle, γ is a phase factor;

- van der Waals (e^{vdw})

$$e^{vdw}(A^n) = \sum_{i=1}^N \sum_{j=i+1}^N \left[4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \right], \quad (7)$$

where: r_{ij} denotes the distance between atoms i and j , σ_{ij} is a collision diameter, ϵ_{ij} is a well depth;

- electrostatic (charge-charge, e^{cc}), described by the Coulomb's law

$$e^{cc}(A^n) = \sum_{i=1}^N \sum_{j=i+1}^N \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}, \quad (8)$$

where: q_i, q_j are atomic charges, r_{ij} denotes the distance between atoms i and j , ϵ_0 is a dielectric constant.³⁴

The number of components in the energy signature \vec{e}_i depends on the force field parameter set used in the computation of the energy profile E^E . In our computations, we used the Amber94³⁵ force field, which generates five mentioned types of potential energy. Therefore, a single energy profile is a $5 \times m$ matrix, where m is a length of the protein (in amino acids). Rows of the matrix are called *energy patterns* (or *energy characteristics*) and columns are called *energy signatures*. In Fig. 5 we can observe five energy characteristics for three molecules of HIV-1 transcriptase.

However, in further considerations we will look at energy profiles as sequences of energy signatures.

In our approach, we compute energy profiles on the basis of protein atomic coordinates (x, y, z) of protein structures retrieved from the macromolecular structure database Protein Data Bank (PDB¹⁷). During the calculations we used TINKER³⁶ application of molecular mechanics and Amber94 force field, which is a set of physical-chemical parameters.

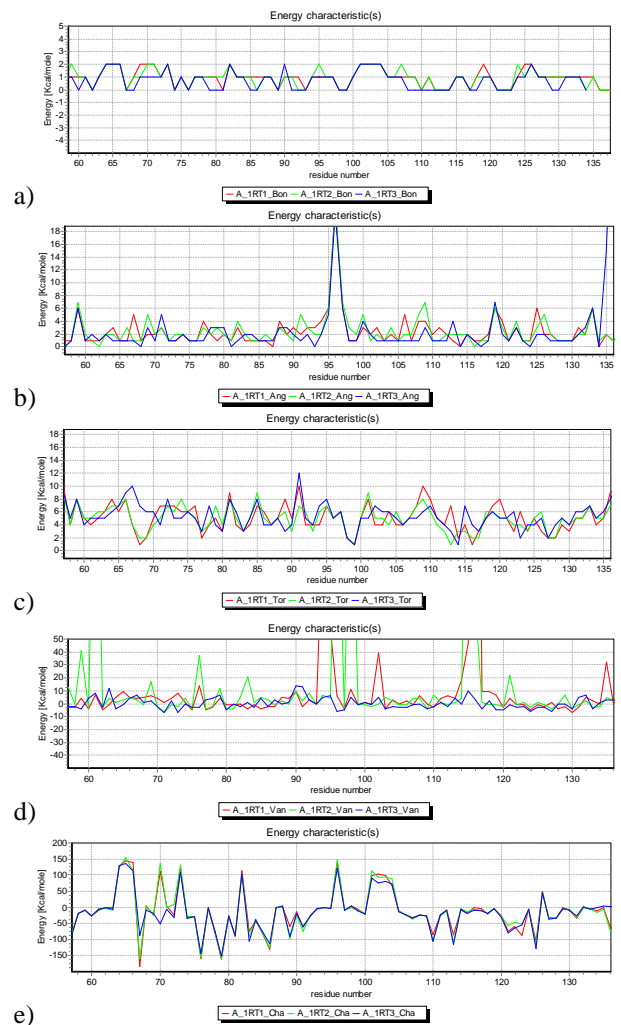


Fig. 5. Different energy characteristics: a) bond stretching, b) angle bending, c) torsional angle, d) van der Waals, e) electrostatic, for three molecules of the HIV-1 DNA transcriptase (PDB ID: 1RT1, HIV-1 Reverse Transcriptase Complexed with MKC-442, PDB ID: 1RT2, HIV-1 Reverse Transcriptase Complexed with TNK-651, and PDB ID: 1RT3, AZT Drug Resistant HIV-1 Reverse Transcriptase Complexed with 1051U91).

We had computed complete energy profiles for more than 34 000 protein structures from the PDB (November 10, 2009) and we store them in a special database. To this purpose, we have designed and developed the Energy Distribution Data Bank (EDB),³⁷ which is available to the public with no costs under the following address <http://edb.aei.polsl.pl>.

5. Optimal Alignment of Fuzzy Energy Profiles

The FS-EAST (*Fuzzy Signatures – Energy Alignment Search Tool*) that we have developed aligns energy profiles in order to find strong similarities between proteins or between parts of these proteins. In the similarity searching, a user specifies the energy profile as a sequence of energy signatures representing an input protein molecule. This profile will be compared and aligned to profiles stored in the EDB database. This is a pairwise comparison. The FS-EAST incorporates the alignment method, which treats components of energy signatures as fuzzy numbers. Therefore, the alignment process is carried on the sequences of fuzzy numbers in 5-dimensional energy space. The alignment method is described in next sections.

5.1. Preliminaries

Let $E_A^{\bar{e}} = (\bar{e}_{A,1}, \bar{e}_{A,2}, \dots, \bar{e}_{A,n})$ and $E_B^{\bar{e}} = (\bar{e}_{B,1}, \bar{e}_{B,2}, \dots, \bar{e}_{B,m})$ are two energy profiles of molecules *A* and *B*. The length of the $E_A^{\bar{e}}$ is *n* and the length of the $E_B^{\bar{e}}$ is *m*.

We transform each energy profile to the fuzzyfied energy profile:

$$E^{\bar{e}} = [\bar{e}_1 \ \bar{e}_2 \ \bar{e}_3 \ \dots \ \bar{e}_m] \rightarrow \Phi^{\bar{e}} = (\bar{\varphi}_1, \bar{\varphi}_2, \dots, \bar{\varphi}_m) \quad (9)$$

The transformation proceeds as follows. Each component e_i^t of any energy signature \bar{e}_i is represented as a triangular fuzzy number $\varphi_i^t = (e_i^t - \alpha, e_i^t, e_i^t + \alpha)$, where e_i^t becomes a modal value of its fuzzy representation, *t* is one of five types of potential energy and α is a spread (Fig. 6).

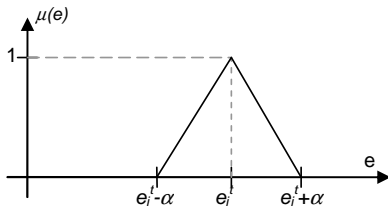


Fig. 6. Representation of a single component in the energy signature as a fuzzy number.

Values of spreads are specific for the type of the energy *t*. They are the same for all energy components φ_i^t of the same type of potential energy *t* in all energy signatures in any energy profile. Values of the spread for different energy types are discussed in Section 5.3.

The *i*th fuzzyfied energy signature $\bar{\varphi}_i$ will have the following form:

$$\bar{\varphi}_i = (\varphi_i^{st}, \varphi_i^{ben}, \varphi_i^{tor}, \varphi_i^{vdw}, \varphi_i^{cc})^T \quad (10)$$

Therefore, the fuzzyfied energy profile of the protein *P* will be a sequence of fuzzyfied energy signatures:

$$\Phi^{\bar{e}} = (\bar{\varphi}_1, \bar{\varphi}_2, \dots, \bar{\varphi}_m) \quad (11)$$

5.2. Alignment method

Let $\Phi_A^{\bar{e}} = (\bar{\varphi}_{A,1}, \bar{\varphi}_{A,2}, \dots, \bar{\varphi}_{A,n})$ and $\Phi_B^{\bar{e}} = (\bar{\varphi}_{B,1}, \bar{\varphi}_{B,2}, \dots, \bar{\varphi}_{B,m})$ are two fuzzyfied energy profiles for molecules *A* and *B*. We are looking for the best adjustment of these two energy profiles, which indicates the best structural similarity of proteins. The adjustment allows some mismatches and gaps to occur, if this leads to the best solution. To accomplish this task we can use dynamic programming methods. We considered different methods, like: Dynamic Time Warping,³⁸⁻⁴⁰ Needleman-Wunsch,⁴¹ and Smith-Waterman.⁴² Finally, we have chosen Smith-Waterman algorithm, since it concentrates on local alignments and reduces the influence of evolutionary noise and produces more meaningful comparisons.

We have modified the Smith-Waterman method to align sequences of energy signatures, which are vectors of fuzzy numbers. The modified method generates the similarity matrix *S* according to the following rules:

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

$$S_{i0} = S_{0j} = 0, \quad (12.1)$$

$$S_{ij} = \max \left\{ \begin{array}{l} S_{i-1,j-1} + \delta(\bar{\varphi}_{A,i}, \bar{\varphi}_{B,j}) \\ \max_{k \geq 1} \{ S_{i-k,j} - \omega_k \} \\ \max_{l \geq 1} \{ S_{i,j-l} - \omega_l \} \\ 0 \end{array} \right\}, \quad (12.2)$$

where: ω_k, ω_l are gap penalties for horizontal and vertical gaps of length k and l , respectively, and $\delta(\bar{\varphi}_{A,i}, \bar{\varphi}_{B,j})$ is a progression function (or delta function):

$$\delta(\bar{\varphi}_{A,i}, \bar{\varphi}_{B,j}) = \begin{cases} 1 + \bar{\mu}_{ij}, & \text{when } \bar{\mu}_{ij} > 0 \\ -1/3, & \text{when } \bar{\mu}_{ij} = 0 \end{cases} \quad (13)$$

The progression can be positive or negative. This depends on the similarity of energy signatures $\bar{\varphi}_{A,i}$ and $\bar{\varphi}_{B,j}$ from compared energy profiles. If two energy signatures $\bar{\varphi}_{A,i}$ and $\bar{\varphi}_{B,j}$ match to each other, the progression is positive and equal to the $1 + \bar{\mu}_{ij}$. Actually, the $\bar{\mu}_{ij}$ parameter is the weighted mean compatibility degree of two fuzzy energy signatures $\bar{\varphi}_{A,i}$ and $\bar{\varphi}_{B,j}$. The $\bar{\mu}_{ij}$ parameter quantifies similarity between these two energy signatures and it is calculated according to the following expression:

$$\bar{\mu}_{ij} = \frac{\sum_{t \in T} \lambda^t \mu_{ij}^t}{\sum_{t \in T} \lambda^t}, \quad (14)$$

where μ_{ij}^t is the compatibility degree of t^{th} components of compared energy signatures, t is one of the energy type from the set $T = \{st, ben, tor, vdw, cc\}$, λ^t is the participation weight specific for the energy type.

The calculation of the compatibility degree for t^{th} components of energy signatures $\bar{\varphi}_{A,i}$ and $\bar{\varphi}_{B,j}$ is presented in Fig. 7a. For mismatching components of energy signatures (Fig. 7b) the progression is always negative and has the constant value $(-1/3)$.

One of the key problems during the calculation of the similarity matrix S is to make an appropriate decision, how to derive values for the current matrix cell based on cells calculated previously. Certainly, this is done according to the eq. (12.2), which shows that for the cell $S_{i,j}$ we can derive the value from (Fig. 8):

- (a) the cell $S_{i-1,j-1}$ in case of matching energy signatures $\bar{\varphi}_{A,i}$ and $\bar{\varphi}_{B,j}$, which gives positive progression – eq. (13), and also mismatching energy signatures $\bar{\varphi}_{A,i}$ and $\bar{\varphi}_{B,j}$, which gives negative progression,
- (b) left side cells $S_{i-k,j}$ providing k gaps ($k \geq l$)

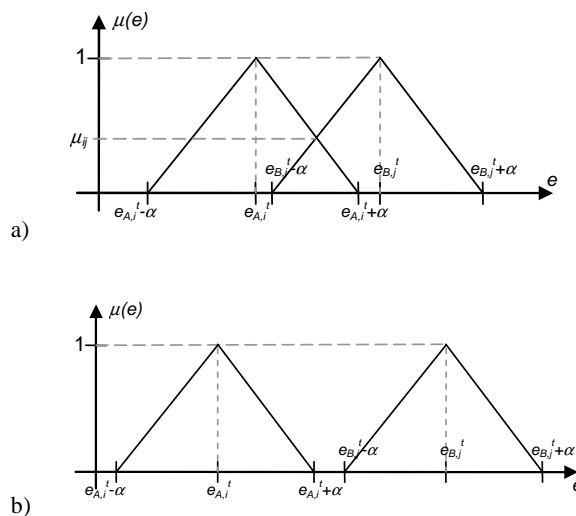


Fig. 7. Calculation of the compatibility degree for matching component values of energy signatures of molecules A and B (a). Mismatching energy components (b).

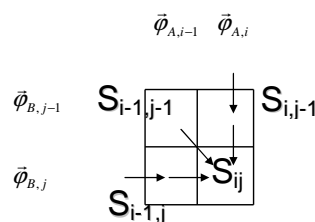


Fig. 8. Derivation directions during the calculation of cell values in the similarity matrix S . Axes x and y of the matrix consists of sequences of energy signatures for two compared proteins A and B.

- (c) upper cells $S_{i,j-l}$ providing l gaps ($l \geq l$)
- (d) we can also set 0, if values derived from above rules decreases below zero

In each of the cases (a-d), we have to decide what is more profitable, i.e. what gives the higher score. Is it more advantageous to give mismatch penalty (negative progression) or insert a gap (with the gap penalty)? How to measure the accordance of two compared energy signatures $\bar{\varphi}_{A,i}$ and $\bar{\varphi}_{B,j}$ and decide whether they match or mismatch to each other? Should we use hard or soft computing approach to determine the similarity between energy signatures? These were the questions that we have considered in our research. Finally, we have decided to fuzzify energy signatures, since we observed small component energy discrepancies in families of

similar protein structures. Therefore, the similarity of component energies in energy signatures is not resolved in the context of crisp values 0 or 1. This would exclude many molecules from the group of similar proteins. Instead of this, we allow a bit of tolerance and imprecision for components of energy signatures and we measure their compatibility with the use of eq. (14). If the compatibility degree is profitable, the diagonal derivation direction is promoted in the decision process instead of inserting a gap or penalizing for a mismatch.

Filled similarity matrix S consists of many possible paths how two energy profiles can be aligned. In the set of possible paths the modified Smith-Waterman method finds and joins these paths that give the best alignment and the highest number of aligned energy signatures. Backtracking from the highest scoring matrix cell and going along until a cell with score zero is encountered gives the highest scoring alignment path. The backtracking is possible, since for each cell we remember how the value of the cell was derived – from left, up or diagonal.

In Fig. 9 we present two sample cases how the alignment path can look like and how we interpret the similarity of proteins based on it.

In Fig. 9a the best alignment is represented by the longest path on the main diagonal. This means two proteins are similar on the corresponding positions, e.g., signature $\vec{\varphi}_{A,i}$ is similar to $\vec{\varphi}_{B,j}$, where $i=j$. For two proteins A and B , it means part F_1 of the molecule A is similar to part F_1 of molecule B .

We can also observe shorter paths besides the main diagonal. They represent local similarities of some regions of one molecule to appropriate regions of the second molecule. However, any reconstruction of the alignment path that would lead through these shorter paths will not give better alignment than the main diagonal.

In Fig. 9b we can observe many local similarities represented by short paths. Therefore, the best alignment is determined by joining appropriate paths in order to obtain the highest scoring alignment path – according to the eq. (12). This is more complex situation. However, it expresses the nature of biological molecules, such as proteins, which differentiated during the evolution process. Spaces between joined paths, which in Fig. 9b are marked with dashed lines, reflect gaps in the final alignment. We can notice the similarity of regions F_1 ,

F_2 , F_3 , and F_4 of molecules A and B , but the location of these regions in the construction of both molecules is different. E.g., modules F_1 , F_2 in the molecule A are connected and in the molecule B are disjointed.

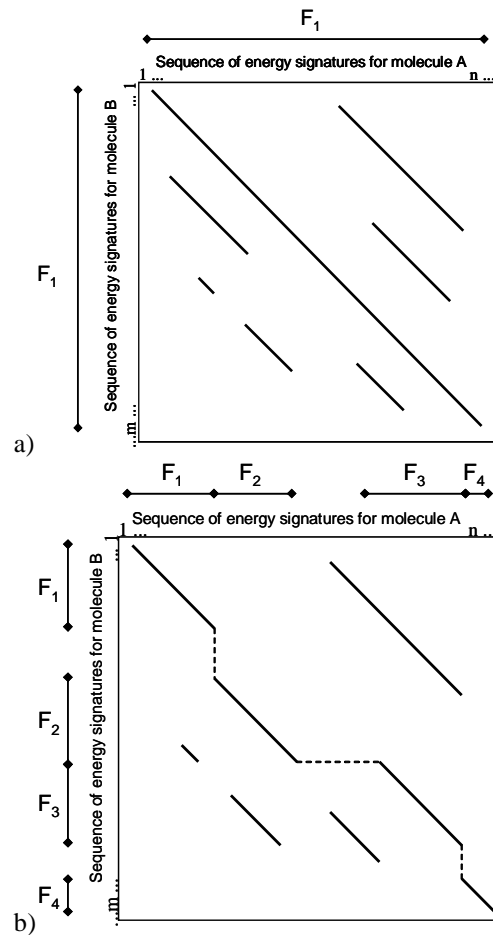


Fig. 9. Similarity matrix with the best alignment path traced through it: a) alignment of consecutive residues, b) alignment with gaps.

For real proteins there can be many local similarities as it is shown in Fig. 10.

Presented matrix shows similarities between sequences of energy profiles for two molecular structures representing different conformations of the Human Kinase CDK2. Dark color indicates high similarity – high compatibility degree between signatures – and bright color indicates weak similarity. The matrix of compatibility degrees CD in Fig. 10 is not exactly the similarity matrix S used in the modified Smith-Waterman method. Each cell of the matrix CD holds values of the compatibility degree between a pair of

energy signatures, while the matrix S contains cumulated values of the similarity for each cell.

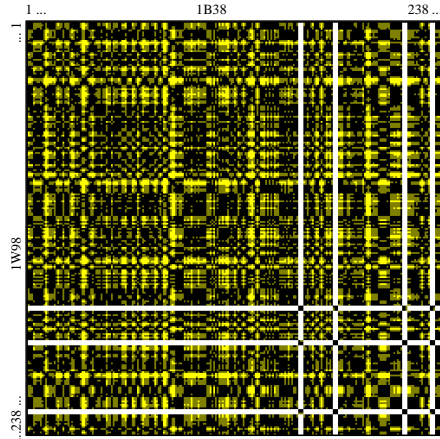


Fig. 10. Matrix of compatibility degrees between sequences of energy profiles for two molecules PDB ID: 1B38 (Crystal Structure of Human CDK2 with ATP) and PDB ID: 1W98 (Structural Basis of CDK2 Activation by Cyclin E).

The cumulated value includes all match awards (positive progressions), mismatch and gap penalties for the path that stops in particular cell of the matrix S . For the best alignment path this cumulated value determines the S - W Score similarity measure:

$$S-W \text{ Score} = \sum_{k=0}^{matches} \delta_k^+(\bar{\varphi}_{A,i}, \bar{\varphi}_{B,j}) + \sum_{k=0}^{mismatches} \delta_k^-(\bar{\varphi}_{A,i}, \bar{\varphi}_{B,j}) + \sum_{k=0}^{gaps} \omega_k, \quad (15)$$

where: $\delta_k^+(\bar{\varphi}_{A,i}, \bar{\varphi}_{B,j})$ is a similarity award (positive progression) for matching energy signatures $\bar{\varphi}_{A,i}$ and $\bar{\varphi}_{B,j}$, $\delta_k^-(\bar{\varphi}_{A,i}, \bar{\varphi}_{B,j})$ is a mismatch penalty (negative progression) for mismatching energy signatures $\bar{\varphi}_{A,i}$ and $\bar{\varphi}_{B,j}$, and ω_k is a gap penalty.

The modified Smith-Waterman method allows horizontal and vertical gaps to appear in the final alignment. These gaps are related to evolutionary changes in protein molecules. However, there are penalties for entering a gap and extending it. In our solution, we use affine gap penalty:

$$\omega_k = -(\omega_o + k\omega_e), \quad (16)$$

where: ω_o is a gap open penalty, ω_e is a gap extension penalty, k is a number of gaps.

Parameters for the modified Smith-Waterman method that we use in our FS-EAST algorithm are provided in the next section.

5.3. Parameters of proposed alignment method

Fault-tolerance and approximate character of the modified Smith-Waterman method is regulated by spreads α , which affect the calculated values of compatibility degrees μ'_{ij} , and final average compatibility degree $\bar{\mu}_{ij}$.

Spreads decide how distant two energy values can be to treat them as similar. The higher value of the spread, the more tolerant the method is. Certainly, increasing the value of the spread causes a danger of accidental alignments. In the paper Ref. 43 we presented results of the research carried for families of protein molecules. The research constitutes the statistical foundation for parameters of the previous versions of the EAST. On the basis of the research, we derived the current values of spreads and participation weights for different energy types. These values are presented in table 1.

For the gap penalty, we applied parameters from the original implementation of the Smith-Waterman method⁴²: gap open penalty $\omega_o=1$, gap extension penalty $\omega_e=1/3$. We made a group of tests that confirmed they work fine in our modified version. Therefore, the gap penalty is $\omega_k = -(1 + k / 3)$, where k is a gap length.

Table 1. Values of spreads and participation weights for different energy types.

Energy type	Range		Default α	Weight λ'
	From	To		
Bond stretching	0.30	0.80	0.50	0.5
Angle bending	0.40	2.00	0.80	0.5
Torsional angle	0.40	0.95	0.70	1.0
Van der Waals	1.55	3.55	2.55	0.2
Electrostatic	1.20	4.20	3.05	1.0

5.4. Graphical User Interface for FS-EAST algorithm

Using, calibrating and tuning parameters for the FS-EAST require friendly access to its particular steps. For this reason, we have developed an advanced GUI in order to preview selected phases of the similarity searching process with the use of energy profiles.

The EAST execution panel is presented in Fig. 11. In the upper part of the panel users specify the query protein (by PDB ID identifier), for which the EAST algorithm searches similar molecules. The structure of the query protein can be also provided in the form of the PDB molecular structure file or using already prepared energy profile in the EDML⁴⁴ format.

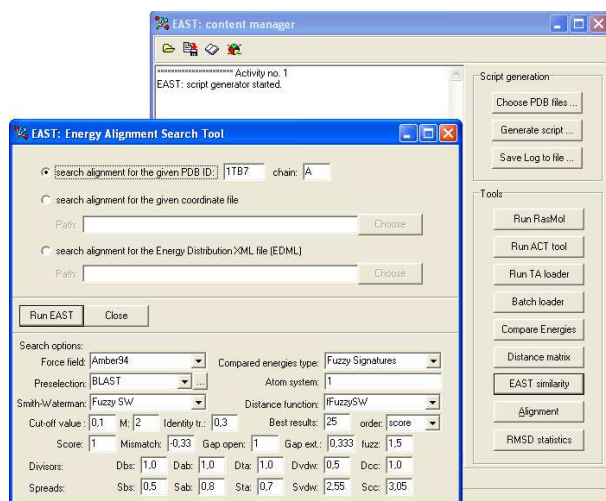


Fig. 11. EAST similarity searching execution panel.

Results of the similarity searching are presented in the *EAST statistics* window (Fig. 12). Here, we can observe, which proteins are similar to the given query protein (PDB ID) and what is the level of the similarity.

PDB ID	Chain	Length	Matches	Mismatches	RMSD	Score	S-W Score	
1TB7	A	319	319	100	0	0.718	444.28	542.77
2PW3	A	320	319	99	1	0.661	482.45	521.48
1Q9M	A	320	319	99	1	0.624	510.89	509.43
1Q7N	A	320	319	99	1	0.624	511.46	509.20
1PTW	A	320	320	100	0	0.606	528.24	506.07
2QYL	A	320	317	99	3	0.578	548.18	489.89
1R06	A	320	316	98	4	0.581	543.66	488.94
1R08	A	319	316	99	3	0.584	541.49	488.91
1R09	A	319	317	99	2	0.567	559.00	485.67
1XNU	A	318	313	98	5	0.572	547.42	477.39
1TB5	A	318	309	97	9	0.546	565.84	465.48

Fig. 12. Results of the similarity searching using the EAST.

Alignment of protein structures (query vs. resultant) represented by energy profiles can be visualized at the level of chosen component energy characteristics and also at the level of protein amino acid sequences (Fig. 13). Although, we concentrate on strong protein similarities in our research, it is worth noting that amino acid sequences of compared proteins do not have to be the same. In Fig. 13 we can see the alignment of electro-

static energy characteristics for query protein 1TB7 and sample resultant molecule 2QYL.

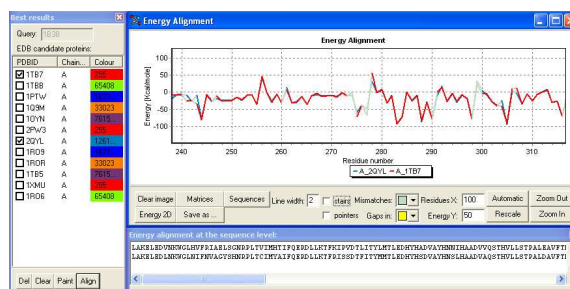


Fig. 13. Alignment of protein structures visible at the level of chosen component energy characteristics and the level of protein amino acid sequences.

Compatibility degree matrices, similarity matrices and path matrices can be visualized with the use of the *EAST: matrix analyzer* window presented in Fig. 14. In the presented example, we show the matrix of compatibility degrees for molecules 1TB7 and 2QYL. Dark regions indicate higher similarity of energy signatures according to assumed thresholds.

The *EAST: matrix analyzer* is very useful in the observation of similarities of particular regions of proteins, verification of similarity matrix during the alignment phase and verification of value derivation directions for the backtracking. It can also support studies on similarity changes as an effect of changes of the EAST parameters.

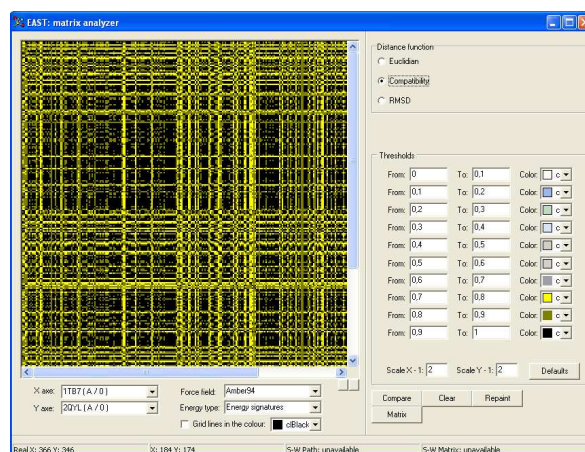


Fig. 14. *EAST: matrix analyzer* window showing sample compatibility degree matrix for specified thresholds.

6. Effectiveness Analysis

Results of many searching processes that we have performed with the use of presented alignment method show the new implementation of the EAST is more perceptive than previous versions. This can be observed in Fig. 15, which shows partial results of an example similarity searching process with the use of the new FS-EAST and its direct predecessor EAST.⁴³ Searching was executed for molecule 1TB7, which represents catalytic domain of Human Phosphodiesterase 4D in complex with the AMP. For the new FS-EAST (Fig. 15a) we can observe wider alignment frames (*Length*), higher values of matching positions (*Matches*) and higher percentage of matching positions (*Match%*), e.g. for molecules 1TBB, 2PW3, 1Q9M, 1OYN (*PDB ID*).

Best results for job: 2009-11-23 12:02:09
S-W type: Fuzzy SW; Energy type: Fuzzy signatures
Mismatch: -0.3334; gap open: 1; gap ext.: 0.3334

PDB ID	Chain	Length	Matches	Match%	S-W Score
1TBB	A	319	317	99	471.32
2PW3	A	320	317	99	458.27
1Q9M	A	318	318	100	449.70
1OYN	A	319	317	99	449.41
1PTW	A	320	318	99	447.18
2QYL	A	319	314	98	429.31
1ROR	A	319	313	98	428.29
1RO9	A	319	313	98	426.34
1RO6	A	319	313	98	426.23
1XMU	A	318	308	96	415.34
1TB5	A	318	301	94	402.47
1ZKL	A	314	289	92	358.70
1T9S	A	311	296	95	347.07
a) 1TAZ	A	312	285	91	343.52

Best results for job: 2009-11-23 17:56:18
Cut-off: 1.0; Energy type: Energy signatures
Mismatch: -0.3334; gap open: 1; gap ext.: 0.3334

PDB ID	Chain	Length	Matches	Match%	S-W Score
1TBB	A	314	292	92	269.54
2PW3	A	317	266	83	229.56
1Q9M	A	314	262	83	221.61
1OYN	A	314	262	83	221.37
1PTW	A	314	262	83	218.42
2QYL	A	313	224	71	170.53
1ROR	A	313	214	68	161.13
1RO9	A	312	212	67	155.37
1XMU	A	307	212	69	153.38
1TB5	A	313	210	67	149.59
1RO6	A	316	202	63	145.81
1TAZ	A	262	115	43	31.70
1ZKL	A	236	101	42	28.54
b) 1T9S	A	112	55	49	24.87

Fig. 15. Results of the searching process with the use of the new FS-EAST algorithm (a) and its predecessor EAST (b) executed for the molecule 1TB7 (Human Phosphodiesterase 4D with the AMP).

Moreover, the modified value of positive progression (similarity award) in the FS-EAST results in higher values of the similarity measure *S-W Score* and higher stability of the FS-EAST.

We compared results of the EAST similarity searching to results of the VAST algorithm available at the National Center for Biotechnology Information (NCBI) web site. The verification confirmed successfulness of our method. In some specific cases the VAST can verify results of the EAST, i.e. results are not contradictory. However, the VAST focuses on the fold similarity, while EAST concentrates on stronger similarity. Therefore, the EAST is more sensitive for conformational changes caused e.g., by the activation of a protein in cellular reaction, while the VAST is perfect for the homology searching and modeling.

In Fig. 16 we can observe structural alignment of the query molecule 1TB7 and resultant molecule 2QYL (crystal structure of PDE4B2B in complex with inhibitor NPV) performed by the VAST and visualized by the Cn3D⁴⁵ confirming results of the FS-EAST. We can observe the reduced representation of both structures in the form of backbones (red color for 1TB7 and yellow for 2QYL). Mismatching positions were marked using blue lines.

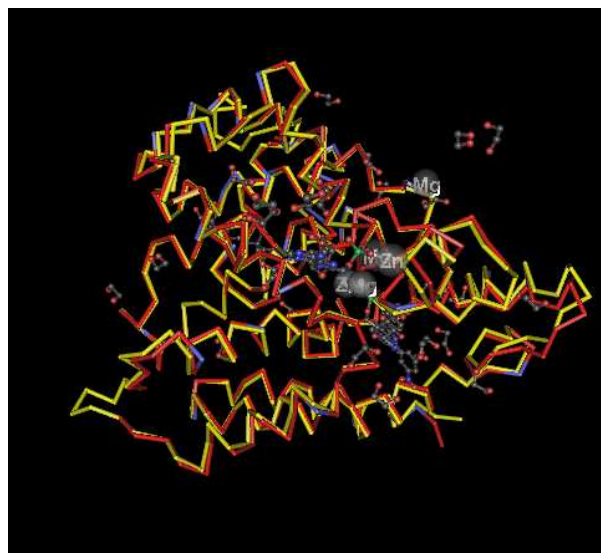


Fig. 16. Structural alignment of the query molecule 1TB7 and resultant molecule 2QYL.

We also tested performance of presented alignment algorithm built-in in our FS-EAST method. Tests were

prepared for the Energy Distribution Data Bank storing 34 372 protein energy profiles. For our tests we chose query molecules with different lengths and representing different structural classes in the SCOP:⁴⁶

- 1QUZ (solution structure of the potassium channel Scorpion toxin HSTX1) – 32 aa,
- 1QPM (NMR structure of the Mu Bacteriophage repressor DNA-binding domain) – 80 aa, class: all α ,
- 1QZ8 (crystal structure of SARS coronavirus NSP9) – 110 aa, class: all β ,
- 1R3U (crystal structure of Hypoxanthine-Guanine Phosphoribosyltransferase from *Thermoanaerobacter tengcongensis*) – 178 aa, class: $\alpha&\beta$,
- 1QPS (crystal structure of a post-reactive cognate DNA-Eco RI complex at 2.50 Å in the presence of Mn^{2+} ion) – 256 aa, class: $\alpha&\beta$,
- 3ERK (complex structure of the MAP kinase ERK2/SB220025) – 349 aa, class: $\alpha+\beta$,
- 1R9O (crystal structure of P4502C9 with Flurbiprofen bound) – 454 aa, class: all α ,
- 1QQA (Purine repressor mutant-Hypoxanthine-Palindromic operator complex) – 674 aa, class: all α ,
- 1QQW (crystal structure of human Erythrocyte Catalase) – 996 aa, class: $\alpha&\beta$.

The FS-EAST search executed without any additional acceleration takes about 6-20 min depending on the size of the user's query molecule. All tests were performed using the PC CPU Intel 3.2 GHz, 2GB RAM. We compared these results to VAST and DALI for the same set of query structures. The VAST search can take up to several hours, when carried out against database containing all structures from the PDB, e.g. for the 1QQW molecule it took 90 min. Using DaliLite⁴⁷ we obtained results after 4-30 min. However, in order to speed up its execution, DaliLite uses feature filters, like BLAST¹⁶ or GTG,⁴⁸ and narrowed database of protein structures PDB90. The FS-EAST also incorporates the BLAST as a preselection filter. The BLAST preselection speeds up the entire process of similarity searching. In our method, the BLAST eliminates molecules, which amino acid sequences completely differ from the user's molecule. The acceleration is noticeable – the FS-EAST with the preselection phase runs about 1-3 minutes, which is about 10 times faster than DALI.

Comparison of the average execution times for three similarity searching algorithms: DALI with the GTG preselection, VAST and FS-EAST without preselection is presented in Fig. 17.

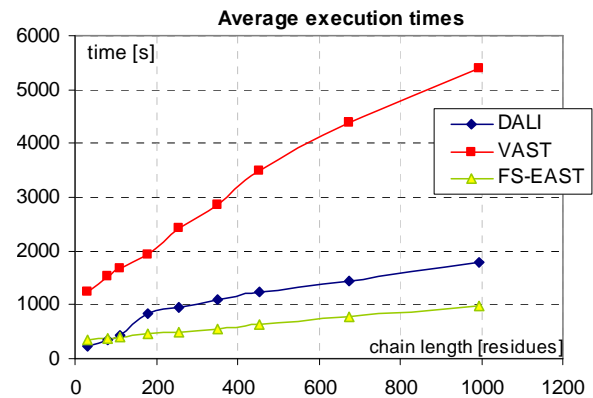


Fig. 17. Comparison of the execution times for three similarity searching algorithms: DALI with preselection, VAST and FS-EAST without any preselection.

7. Concluding Remarks

In the paper, we studied a fuzzy representation of energy signatures in the alignment of protein energy profiles. Representing energy components of energy signatures as fuzzy numbers brings several advantages. It improves the decision making process in determination of the alignment path. Performed tests showed the alignment method is more perceptive, which is reflected in wider alignment frames. We have introduced a new progression function (known as similarity award in previous implementations). This increased a stability of the alignment path and eliminated a tendency to jump between diagonals in the similarity matrix S built by the energy adapted Smith-Waterman algorithm. It was one of the weaknesses of previous versions of the EAST. Moreover, the new FS-EAST method with the presented alignment method continues good traditions in measuring the quality of the alignment and self-compensating small dissimilarities of some components of energy signature vectors by higher similarities of other components. Finally, performance tests show the FS-EAST is as fast as its direct predecessor and faster than rough methods used in the fold similarity search, like VAST and DALI.

Future efforts will cover the improvement of the FS-EAST performance by implementation of intelligent heuristics, distribution of work and specific indexing of data. Furthermore, we think about increasing the granularity of the FS-EAST method to the level of particular atoms. This would raise the overall precision of the EAST family and extend its functionality e.g. towards drug design or clinical decision support systems.⁴⁹

References

1. J.P. Allen, *Biophysical chemistry* (Wiley-Blackwell, 2008).
2. C. Branden, J. Tooze, *Introduction to protein structure* (Garland, 1991).
3. C.R. Cantor, P.R. Schimmel, *Biophysical chemistry*, W.H. Freeman, 1980.
4. H. Lodish, A. Berk, S.L. Zipursky, et al., *Molecular cell biology*, Fourth edition. (W. H. Freeman and Company, NY, 2001).
5. C. Gibas, P. Jambeck, *Developing bioinformatics computer skills*, First edition. (O'Reilly, April 2001).
6. T.K. Attwood, D.J. Parry-Smith, *Introduction to bioinformatics*, (Prentice Hall, 1999).
7. R.E. Dickerson, I. Geis, *The structure and action of proteins*, 2nd ed. (Benjamin/Cummings, Redwood City, Calif. Concise, 1981).
8. T.E. Creighton, *Proteins: structures and molecular properties*, 2nd ed. (Freeman, San Francisco, 1993).
9. D. Mrozek, B. Małysiak, S. Kozielski, Energy profiles in detection of protein structure modifications, in *Proc. IEEE Int. Conf. on Computing and Informatics*, (Kuala Lumpur, 2006), pp. 1–6.
10. B. Małysiak, D. Mrozek, S. Kozielski, L. Znamirowski, Signal transduction simulation in nanoprocesses using distributed database environment, in *Proc. of 5th IASTED Int. Conf. on Modelling, Simulation, and Optimization*, (Oranjestad, Aruba. ACTA Press, 2005), pp. 17–22.
11. A.W. Znamirowski, L. Znamirowski, Two-phase simulation of nascent protein folding, in *Proc. of the 4th IASTED Int. Conf. on Modelling, Simulation, and Optimization*, (Kauai, Hawaii, ACTA Press, 2004), pp. 293–298.
12. D. Mrozek, B. Małysiak-Mrozek, S. Kozielski, Energy Properties of Protein Structures in the Analysis of the Human RAB5A Cellular Activity. *Advances in Intelligent and Soft Computing* 59 (2009), pp. 121–131.
13. D. Mrozek, B. Małysiak, S. Kozielski, EAST: Energy Alignment Search Tool, *Lecture Notes in Artificial Intelligence* 4223 (2006), pp. 696–705.
14. D. Mrozek, B. Małysiak, Searching for strong structural protein similarities with EAST, *Journal of Computer Assisted Mechanics and Engineering Sciences* 14 (2007), pp. 681–693.
15. W.R. Pearson and D.J. Lipman, Improved Tools for Biological Sequence Analysis, *PNAS* 85 (1988), pp. 2444–2448.
16. S.F. Altschul, et al., Basic local alignment search tool, *J Mol Biol* 215 (1990), pp. 403–10.
17. H.M. Berman, et al., The Protein Data Bank, *Nucleic Acids Res.* 28 (2000), pp. 235–242.
18. J.F. Gibrat, T. Madej, S.H. Bryant, Surprising similarities in structure comparison, *Curr Opin Struct Biol* 6(3) (1996), pp. 377–385.
19. J. Shapiro, D. Brutlag, FoldMiner and LOCK 2: protein structure comparison and motif discovery on the web. *Nucleic Acids Res.* 32 (2004): W536–41.
20. L. Holm, C. Sander, Protein structure comparison by alignment of distance matrices, *J Mol Biol.* 233(1) (1993), pp. 123–38.
21. I.N. Shindyalov, P.E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng.* 11(9) (1998), pp. 739–747.
22. Y. Ye and A. Godzik, Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19(2) (2003), pp. 246–255.
23. I. Friedberg, et al., Using an alignment of fragment strings for comparing protein structures, *Bioinformatics* 23(2) (2007), pp. 219–224.
24. T. Can, Y.F. Wang, CTSS: a robust and efficient method for protein structure alignment based on local geometrical and biological features, in *Proc. of the 2003 IEEE Bioinformatics Conference*, (2003), pp. 169–179.
25. Yang J., Comprehensive description of protein structures using protein folding shape code. *Proteins* 71(3) (2008), pp. 1497–518.
26. J.H. Zhu, Z.P. Weng, FAST: A novel protein structure algorithm, *Proteins* 58 (2005), pp. 618–627.
27. N. Krasnogor, D.A. Pelta, Measuring the similarity of protein structures by means of the universal similarity metric, *Bioinformatics* 20(7) (2004), pp. 1015–1021.
28. D.A. Thorner, et al., Similarity Searching in Files of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials. *J. Chem. Inf. Comput. Sci.* 36 (1996), pp. 900–908.
29. J. Rodrigo, M. Barbany, Comparison of biomolecules on the basis of Molecular Interaction Potentials, *J. Braz. Chem. Soc.* 13(6) (2002), pp. 795–799.
30. H. Ji, H. Li, M. Flinspach, et al., Computer modeling of selective regions in the active site of Nitric Oxide synthases: implication for the design of isoform-selective inhibitors, *J. Med. Chem.* (2003), pp. 5700–5711.
31. P.J. Goodford, Computational procedure for determining energetically favourable binding sites on biologically important macromolecules, *J. Med. Chem.* 28 (1985), pp. 849–857.
32. R. Sayle, E.J. Milner-White, RasMol: Biomolecular graphics for all, *Trends in Biochemical Sciences* 20(9) (1995), pp. 374.
33. U. Burkert, N.L. Allinger, *Molecular mechanics*, (American Chemical Society, Washington D.C., 1980).
34. A. Leach, *Molecular modelling: principles and applications*, 2nd edition. (Pearson Education EMA, UK, 2001).
35. 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 (1995), pp. 5179–5197.
36. J. Ponder, TINKER – Software tools for molecular design, User's Guide. (Dept. of Biochemistry & Molecular Biophysics, Washington University, School of Medicine, St. Louis, 2001).
37. D. Mrozek, B. Małysiak-Mrozek, S. Kozielski, A. Świerniak, The Energy Distribution Data Bank: Collecting Energy Features of Protein Molecular Structures, in *Proc.*

- 9th IEEE Int. Conf. on Bioinformatics and BioEngineering, (Taichung, Taiwan, IEEE, 2009), pp. 301–306.
38. S. Salvador and P. Chan. FastDTW: toward accurate dynamic time warping in linear time and space, *Intell. Data Anal.* 11(5) (2007), pp. 561–580.
 39. G. Jian-Kui, et al., Estimating similarity over data streams based on Dynamic Time Warping, in *Proc. of 4th Conf. Fuzzy Systems and Knowledge Discovery*, (IEEE Computer Society, 2007), pp. 53–57.
 40. Tak-chung Fu, et al., Time series subsequence searching in specialized binary tree, *LNC3 4223*, (Springer-Verlag, Heidelberg, 2006), pp. 568–577.
 41. S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J Mol Biol* 48 (3) (1970), pp. 443–53.
 42. T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, *J Mol Biol* 147 (1981), pp. 195–197.
 43. B. Małysiak, A. Momot, S. Kozielski, D. Mrozek, On using energy signatures in protein structure similarity searching, in Rutkowski, L., et al. (eds.) *Artificial Intelligence and Soft Computing, LNAI 5097*, (Springer, Heidelberg, 2008), pp. 939–950.
 44. D. Mrozek, B. Małysiak-Mrozek, S. Kozielski, S. Górczynska-Kosiorz, The EDML Format to Exchange Energy Profiles of Protein Molecular Structures. *Lecture Notes in Computer Science 5754*, (Springer, 2009), pp. 146–157.
 45. C.W. Hogue, Cn3D: a new generation of three-dimensional molecular structure viewer, *Trends Biochem Sci.* 22(8) (1997), pp. 314–6.
 46. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* 247 (1995), pp. 536–540.
 47. L. Holm, S. Kaariainen, P. Rosenstrom, A. Schenkel, Searching protein structure databases with DaliLite v.3, *Bioinformatics* 24 (2008), pp. 2780–2781.
 48. A. Heger, et al., The global trace graph, a novel paradigm for searching protein sequence databases, *Bioinformatics* 23 (2007), pp. 2361–2367.
 49. G.L. Kong, D.-L. Xu and J.-B. Yang, Clinical Decision Support Systems: a Review of Knowledge Representation and Inference under Uncertainties, *International Journal of Computational Intelligence Systems* Vol.1, No. 2 (Atlantis Press, 2008), pp. 159–167.