proteins, secondary structure,
protein similarity, alignment

Dominika WIECZOREK[1], Bożena MAŁYSIAK-MROZEK[1], Stanisław KOZIELSKI[1], Dariusz MROZEK[1]

A METHOD FOR MATCHING SEQUENCES OF PROTEIN SECONDARY STRUCTURES

Alignment of specific regions of two biological molecules is a basic method for determination how similar these two molecules are. There are several methods of optimal alignment that were developed through many years. However, they are dedicated for nucleotide sequences of DNA/RNA or amino acid sequences of proteins. Since the construction of proteins can also be analyzed at the level of secondary structure (and higher), we need a comparative method, which would allow us to determine the similarity between biological particles at this level and express it through the appropriate similarity measure. For this reason, we have modified an existing Smith-Waterman method towards matching sequences of secondary structures elements (SSEs). In the paper, we present our modification to the method. We also describe how we find several alternative and equally optimal alignment paths on the basis of the characteristics of compared sequences. Presented alignment method is used in the PSS-SQL language, which allows searching a database in order to find proteins having secondary structures similar to the structural pattern specified by a user.

1. INTRODUCTION

Proteins are biological molecules made up of amino acids (peptides), joined consecutively to each other by peptide bonds and thus, forming linear amino acid chains. Internal structure of proteins is determined at four different representation levels – from primary structure to quaternary structure [1]. Primary structure determines amino acids, which construct the protein, and the order of amino acids in the linear chain. For this reason, the primary structure is usually just called *amino acid sequence*. Secondary, tertiary and quaternary structures define so-called spatial structure [2]. These three representation levels are related to the way of folding the linear chain of protein in the cellular environment and, hence, the location of atoms building particular amino acids [3, 4].

One of the basic tools for biochemical analysis of proteins is similarity searching [5]. The process can be implemented at the level of amino acid sequence or at the level of spatial structure. Searching for similar proteins may have different applications. Depending on the application we can analyze proteins at the level of amino acid sequence or with respect to various features of their structures. Comparative analysis of protein sequences is essential for the identification of proteins, identification of their functions and determination of their fundamental physical-chemical properties. On the other hand, comparative analysis of protein structures brings much more information and is extremely important in processes such as predicting the function of newly discovered proteins that are difficult to identify on the basis of amino acid sequence [6, 7].

Protein similarity searching is usually carried out by *comparison* of a specific protein with a group of proteins and mutual *alignment* of specific regions of compared molecules. By alignment we mean the process of juxtaposition of two or more sequences in such a way that as a result we obtain the maximum number of identical or similar elements (e.g. amino acids). Since the number of possible juxtapositions of two sequences of proteins is very large, the alignment is the process of optimization. The aim of this process is to find regions of similarity between biological molecules, usually expressed by the largest number of identical or similar positions matched to each other. As a result of optimal alignment and on the basis of relevant alignment measures it is possible to assess the degree of similarity between proteins.

In the paper, we show the method for aligning sequences of protein *secondary structure elements* (SSEs). Sequences of secondary structure elements describe how the chain of amino acids is folded, i.e. which amino acids are part of particular secondary structures (in one-to-one relationship). In Fig. 1 we

[1] Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland.

show the amino acid sequence of the *6-phosphogluconolactonase* in the *Escherichia coli* and the corresponding sequence of SSEs. Particular SSEs have the following meaning: H denotes α-helix, E denotes β-strand, C (or L) stands for loop, turn or coil.

Since existing alignment methods are dedicated for DNA/RNA and amino acid sequences, we have modified one of the methods in order to align SSE sequences. The method is used as one of the last steps in the execution of queries in the PSS-SQL language (*Protein Secondary Structure – Structured Query Language*). We have developed the PSS-SQL in order to search databases against proteins having secondary structure similar to the structure specified in the user's query [8].

```
A7ZY23
6PGL_ECOHS
6-phosphogluconolactonase OS=Escherichia coli O9:H4 (strain HS) GN=pgl PE=3 SV=1

MKQTVYIASPESQQIHVWNLNHEGALTLTQVVDVPGQVQPMVVSPDKRYLYVGVRPEFRVLAYRIAPDDGALTFAAESAL
PGSPTHISTDHQGQFVFVGSYNAGNVSVTRLEDGLPVGVVDVVEGLDGCHSANISPDNRTLWVPALKQDRICLFTVSDDG
HLVAQDPAEVTTVEGAGPRHMVFHPNEQYAYCVNELNSSVDVWELKDPHGNIECVQTLDMMPENFSDTRWAADIHITPDG
RHLYACDRTASLITVF

CCCEEEEECCCCEEEEEEECCCCEEEEEEEEEECCCCCEEEECCCEEEEECCCCCEEEEEEEECCCCCHHHHHHHCC
CCCCCCEEECCCCEEEEECCCCCCEEEEEEECCCCCCEEEEEEECCCCCCCCCCCCCCEEEECCHHHHHEEEEECCCCC
CEEECCCCCEEEECCCCCCCEEECCCCEEEEEECCCCCCEEEEEECCCCCCEEEEEECCCCCCCCCCCCCEEEECCCCC
CEEEECCCCCCCEEEE
```

Fig. 1. Sample amino acid sequence of the protein *6-phosphogluconolactonase* in the *Escherichia coli* with the corresponding sequence of secondary structure elements.

2. POPULAR ALIGNMENT METHODS

The best-known methods of optimal alignment of biological sequences are: the Needleman-Wunsch method [9], implementing global matching strategy, and the Smith-Waterman method [10], implementing local matching strategy. Both methods belong to the class of dynamic programming methods [11].

Alignment, which covers the whole range of examined sequences, is called global. Global alignment methods, like Needleman-Wunsch, can be used for this type of similarity that occurs along the whole chain. Therefore, methods of global alignment are primarily used for testing the similarity of protein fragments with single functional regions, so-called functional domains, or proteins slightly differing in the process of evolution. A global alignment takes no account of the important features of proteins that is their modular construction with the possibility of internal rearrangement of some parts of the amino acid chain (translocations), and duplication of some functional regions inside the chain. Despite some drawbacks, these methods are very much needed, and their subsequent versions, such as the Needleman-Wunsch-Sellers [12], are successfully used as one of the phases in heuristic search algorithms, such as FASTA [13].

In addition to these methods, in many cases we can use methods of local alignment, which match only certain parts of sequences and therefore, allow to find similarities between sequences, which might seem to be not related. One of the most popular methods to establish optimal local alignments is the Smith-Waterman method [10]. Alignment is considered to be locally optimal, if the value of similarity measure, calculated for the matched fragments of both sequences, cannot be improved by shortening or extension of matched fragments. The Smith-Waterman method was originally developed to align nucleotide sequences of DNA/RNA or amino acid sequences of proteins. For many years this method went through several upgrades [12, 14-16] and its assumptions formed the foundation for the development of very popular BLAST algorithm [17].

3. ALIGNMENT METHOD FOR SEQUENCES OF SECONDARY STRUCTURE ELEMENTS

In PSS-SQL queries [8], when we search proteins having secondary structure descriptor similar to the specified structural pattern, we make use of the alignment performed by the Smith-Waterman method. We modified the Smith-Waterman method, originally destined to align nucleotide sequences of DNA/RNA and amino acid sequences of proteins, to align sequences of SSEs. The modified version of the Smith-Waterman method returns more than one optimal solution, by reason of the approximate character of the specified pattern.

In PSS-SQL queries, the pattern is represented by blocks of segments, where each segment can be defined precisely or by an interval. For example, in the pattern *h(4),e(2;5), c(2;4)* we can distinguish an α-helix containing exactly 4 elements, followed by β-strand of the length 2 to 5 elements, and loop of the length between 2 and 4 elements. During the alignment phase the pattern is expanded to the full possible length, e.g. for the given pattern it takes the following form *HHHHEEEEECCCC*. In this form it may take part in comparison with candidate SSEs sequences from the database. In the section, we describe how the alignment method works.

Suppose we have two proteins *A* and *B*, one of which represents the given pattern and the other a candidate protein from the database. We represent primary structures of proteins *A* and *B* in the following form: $P^A = p_1^A p_2^A ... p_n^A$ and $P^B = p_1^B p_2^B ... p_m^B$, where: *n* is a length of the protein *A* (in amino acids), m is a length of the protein *B*, $p_i \in P$, and *P* is a set of 20 common types of amino acids.

We represent secondary structures of proteins *A* and *B* in the following form: $S^A = s_1^A s_2^A ... s_n^A$ and $S^B = s_1^B s_2^B ... s_m^B$, where: $s_i \in S$ is a single secondary structure element (SSE), which corresponds to the *i*-th amino acid p_i, $S = \{H, E, C, ?\}$ is a set of 3 types of the secondary structures: H denotes α-helix, E denotes β-strand, C stands for loop, turn or coil, the ? symbol corresponds to any of the mentioned SSEs.

In the alignment process we build the similarity matrix *D* according to the following rules – for $0 \le i \le n$ and $0 \le j \le m$:

$$D_{i,0} = D_{0,j} = 0, \tag{1}$$

$$D_{i,j}^{(1)} = D_{i-1,j-1} + \delta(s_i^A, s_j^B), \tag{2}$$

$$D_{i,j}^{(2)} = \max_{1 \le k \le n} \{D_{i-k,j} - \omega_k\}, \tag{3}$$

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

$$D_{i,j}^{(4)} = 0, \tag{5}$$

$$D_{i,j} = \max_{v=1..4} \{D_{i,j}^{(v)}\}, \tag{6}$$

where: $\delta(s_i^A, s_j^B)$ is an award δ^+, if two SSEs from proteins *A* and *B* match to each other, or a penalty for a mismatch δ^-, if they do not match:

$$\delta(s_i^A, s_j^B) = \begin{cases} 1 & if \quad s_i^A = s_j^B \\ -1 & if \quad s_i^A \ne s_j^B \end{cases}, \tag{7}$$

ω_k is a penalty for a gap of the length *k*:

$$\omega_k = \omega_O + k \times \omega_E, \tag{8}$$

where: $\omega_O = 3$ is a penalty for opening a gap, $\omega_E = 0.5$ is a penalty for a gap extension. In Fig. 2 we show the scoring matrix for particular pairs of SSEs. This scoring system, with such values of gap penalties, promotes longer alignments, without gaps. We assume users can determine places of possible gaps by specifying optional segments in a query pattern.

	H	E	C	?
H	1	-1	-1	1
E	-1	1	-1	1
C	-1	-1	1	1
?	1	1	1	1

Fig. 2. Scoring system for particular pairs of secondary structure elements.

Filled similarity matrix *D* consists of many possible paths how two sequences of SSEs can be aligned. In the set of possible paths the modified Smith-Waterman method finds and joins these paths that give the best alignment. Backtracking from the highest scoring matrix cell and going along until a cell with score zero is encountered gives the highest scoring alignment path. However, in the modified version of the alignment method that we have developed, we find many possible alignments by searching consecutive maxima in the similarity matrix *D*. This is necessary, since the pattern is usually not defined precisely, contains ranges of SSEs or undefined elements. Therefore, there can be many regions in a protein structure that fit the pattern. In the process of finding alternative alignment paths, the modified Smith-Waterman method follows the value of the internal parameter *MPE* (*Minimum Path End*), which defines the stop criterion. We find alignment paths until the next maximum in the similarity matrix *D* is lower than the value of the *MPE* parameter. The value of the *MPE* depends on the specified pattern, according to the following formula.

$$MPE = (MPL \times \delta^{+}) + (NoIS \times \delta^{-}), \tag{9}$$

where: *MPL* is a minimum pattern length, *NoIS* is a number of imprecise segments, i.e. segments, for which minimum length is different than maximum length. E.g. for the structural pattern *h(10;20),e(1;10),c(5),e(5;20)* containing α-helix of the length 10 to 20 elements, β-strand of the length 1 to 10 elements, loop of the length 5 elements, and β-strand of the length 5 to 20 elements, the *MPL=21* (10 elements of the type *h*, 1 element of the type *e*, 5 elements of the type *c*, and 5 elements of the type *e*), the *NoIS=3* (first, second, and fourth segment), and therefore, *MPE=18*.

The *Score* similarity measure is calculated for each of possible alignment paths and it totals all similarity awards δ^{+}, mismatch penalties δ^{-} and gap penalties ω_k according to the following formula:

$$Score = \sum \delta^{+} + \sum \delta^{-} - \sum \omega_k . \tag{10}$$

4. DISCUSSION AND CONCLUSIONS

Presented method of matching sequences of SSEs is not as fast as the heuristic BLAST method, which focuses on speed rather than on accuracy of matching. The computational complexity of presented method is O(nm(n+m)), which is certainly a drawback. However, this method returns optimal alignments, which is important from the viewpoint of executed queries.

The effectiveness of alignments was successfully confirmed by the analysis of results of hundreds of PSS-SQL queries submitted against a database containing 6 230 proteins. However, due to its complexity, we do not recommend to use this method in comparison of a query pattern to the entire content of the database. In the PSS-SQL language that we have developed, the method is used in the last phase and only for a group of proteins isolated in the preselection processes based on additional features or filtering predicates and using special indexing. Therefore, we limit the number of computationally expensive matches.

BIBLIOGRAPHY

[1] ALLEN J.P., Biophysical chemistry, Wiley-Blackwell, 2008.
[2] BRANDEN C., TOOZE J., Introduction to protein structure, Garland, 1991.
[3] DICKERSON, R.E., GEIS, I., The structure and action of proteins, 2nd ed. Benjamin/Cummings, Redwood City, Calif. Concise, 1981.
[4] CREIGHTON T.E.: Proteins: Structures and molecular properties, 2nd ed. Freeman, San Francisco, 1993.
[5] EIDHAMMER I., INGE J., TAYLOR W.R., Protein Bioinformatics: An algorithmic approach to sequence and structure analysis, John Wiley & Sons, 2004.
[6] GIBRAT J.F., MADEJ T., BRYANT S.H.: Surprising similarities in structure comparison, Curr Opin Struct Biol, Vol. 6(3), 1996, pp. 377–385.
[7] YANG J., Comprehensive description of protein structures using protein folding shape code, Proteins, Vol. 71(3), 2008, pp. 1497–518.
[8] MROZEK D., WIECZOREK D., MAŁYSIAK-MROZEK B., KOZIELSKI S., PSS-SQL: Protein Secondary Structure – Structured Query Language, Proc. 32nd Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, September 2010 (to be published).
[9] NEEDLEMAN S., WUNSCH C., A general method applicable to the search for similarities in the amino acid sequences of two proteins, Journal of Molecular Biology, Vol. 48, 1970, pp. 443–453.
[10] SMITH T.F., WATERMAN M.S., Identification of common molecular subsequences, J Mol Biol, Vol. 147, 1981, pp. 195–197.
[11] BELLMAN R., Dynamic programming, Princeton University Press, Princeton, N. J. 1957.
[12] SELLERS P.H., Pattern recognition in genetic sequences by mismatch density, Bull. Math. Biol., Vol. 46, 1984, pp. 501–514.
[13] PEARSON W.R., Flexible sequence similarity searching with the FASTA3 program package, Methods Mol Biol. Vol. 132, 2000, pp. 185–219.
[14] GOTOH O., An Improved Algorithm for Matching Biological Sequences, J. Mol. Biol. Vol. 162, 1982, pp. 705–708.
[15] ALTSCHUL S.F., ERICKSON B.W., Locally optimal subalignments using nonlinear similarity functions, Bull. Math. Biol. Vol. 48, 1986, pp. 633–660.
[16] WATERMAN M.S., EGGERT M., A new algorithm for best subsequence alignments with applications to tRNA-rRNA comparisons, J. Mol. Biol. Vol. 197, 1987, pp. 723–728.
[17] ALTSCHUL S.F., GISH W., MILLER W., MYERS E.W., LIPMAN D.J., Basic local alignment search tool, J Mol Biol, Vol. 215, 1990, pp. 403–10.