SOFTWARE FOR BIOLOGICAL DATA STORAGE AND SEQUENCE COMPARISON

Janis Eiduks, Liga Paura
Latvia University of Agriculture
janis.eiduks.llu@gmail.com, liga.paura@llu.lv

Abstract. Bioinformatics is the science, where mathematical, statistical and information technology methods are used to solve various problems related to biology. Typical research problems are development of database for the measured biological sample information and manipulation of biological data. Thereby the most common problem is DNA or protein sequence alignment. The sequences are compared, determining their similarity in each position. There are two categories of pairwise sequence alignments: global and local alignments, which are based on dynamic programming algorithms. Needleman-Wunsch and Smith-Waterman algorithm is a general method for global and local two sequence alignments. Within the project a software prototype was developed. The functionality of the software is to store data about animals and compare DNA or protein sequences according to Needleman-Wunsch and Smith-Waterman algorithms. For the implementation of database PostgreSQL (release 9.4.) and for the implementation of the dynamic programming algorithms C# (release 2010) environment were used. The software provides global-linear, local-linear, global-affine and local-affine DNA and protein sequence alignment algorithms. One of the directions of the software development is to include multi sequence alignment algorithms for phylogenetic trees construction.

Keywords: dynamic programming algorithm, sequence alignment, database.

Introduction

One of the popular alignment algorithms for sequence comparison is the dynamic programming algorithm [1]. The question which will be solved is how similar two sequences are or from biological point of view how similar the new sequenced gene's function is to genes of known function.

There are two categories of alignments: global alignments and local alignments. By the term of global comparison we understand that two full-length sequences are compared. A global comparison can be time consuming[2]. In turn, by the term of local comparison we understand that one sequence segment is compared with the sequence of other segment. Summary of the sequence comparison algorithms is described in Table 1.Needleman-Wunsch algorithm is a general global alignment method and Smith-Waterman algorithm is a local alignment method, both methods are based on dynamic programming algorithms [3; 4].

Table 1

Two sequence alignment algorithms

Alignment	Global	Local
Linear	1: Linear global (Needleman-Wunsh)	2: Linear local (Smith-Waterman)
Affine	3: Affine global (Needleman-Wunsh)	4: Affine local (Smith-Waterman)

Global and local alignments are possible to be divided in linear and affine. Linear comparison determines that the penalty points for a gap are not related to the length of the gap – in turn, comparison of affine comparison long inserting or deleting provides less penalty points [5].

All algorithms calculate sequence similarity.The statistical estimates for sequence similarity are one part of the biological sequence comparison and have fundamentally changed the practice of biochemistry and molecular biology [6].The scientists want to find biologically significant relationships between sequences.

The LUA realizes a research project, within it animal milk and meat quality data and animal genotype data will be analysed. Based on the above mentioned approaches, we present the software. The software is designed for animal phenotypic and genetic data storage and comparing of two nucleotide or protein sequences.

Materials and methods

For storage of the project data set a database was developed that stores information about productivity and genotypes of animals. For each animal species (pigs and cows) a separate database, was created because the range of the controlled traits is different.

For implementation of the database PostgreSQL (release 9.4.) was used. Often it is referred as simply Postgres, it is a free object-relational database management system (DBMS). PostgreSQL is based on the SQL language and it supports various features and capabilities of the SQL2003 standard (ISO/IEC 9075) [7; 8].

For implementation of the dynamic programming Needleman-Wunsh and Smith-Waterman bioinformatics algorithm for the two sequences alignment software, C# (release 2010) environment was used. C# is a programming language developed by the company Microsoft. Initially, it was developed specifically for the .NET execution environment, but later it was approved by Ecma and ISO standards. C# is a programming language developed for Common Language Infrastructure (CLI) [9].

For processing of genetic information by bioinformatics algorithms a system was developed to compare two nucleotide or protein sequences and to compute the alignment results. The dynamic programming Needleman-Wunsh (1) and Smith-Waterman (2)linear sequence alignment algorithms were implemented as follows:

$$M(i,j) = \max\begin{cases} M(i-1,j-1) + S(x_i, y_i) \\ M_x(i-1,j) + g \\ M_y(i,j-1) + g \end{cases} \tag{1}$$

$$M(i,j) = \max\begin{cases} M(i-1,j-1) + S(x_i, y_i) \\ M_x(i-1,j) + g \\ M_y(i,j-1) + g \\ 0 \end{cases} \tag{2}$$

where x – first sequence $x_1 \dots x_n$;
y – second sequence $y_1 \dots y_m$;
M_{ij} – (n+1)×(m+1) matrixindexed by (i, j);
$S(x_i,y_i)$ – score of the best possible alignment between x and y sequence initial segments;
g – gap penalty value.

The sequence alignment result in the software is represented by sequence similarity and sequence complexity.

Sequence similarity is (S) calculated by formula:

$$S = \sum_{i=1}^{m} S_{ij} \tag{3}$$

where S_{ij} values are equal to +1 in case of a match or -1 in case of a mismatch for nucleotide alignment and for protein alignment S_{ij}elements are equal to BLOSUM matrix value.

Evaluation of the sequence complexity in a text region was implemented by Wootton and Federhen [10] formula:

$$C = \frac{1}{N}\log_K\left(\frac{N!}{\Pi_{i=1}^{K} n_i}\right) \tag{4}$$

where N – length of the sequence;
K – number of possible letters in the sequence (for DNA $K = 4$, for protein $K = 20$);
n_i – number of occurrences of each letter in the sequence.

Results and discussion

General description of the software

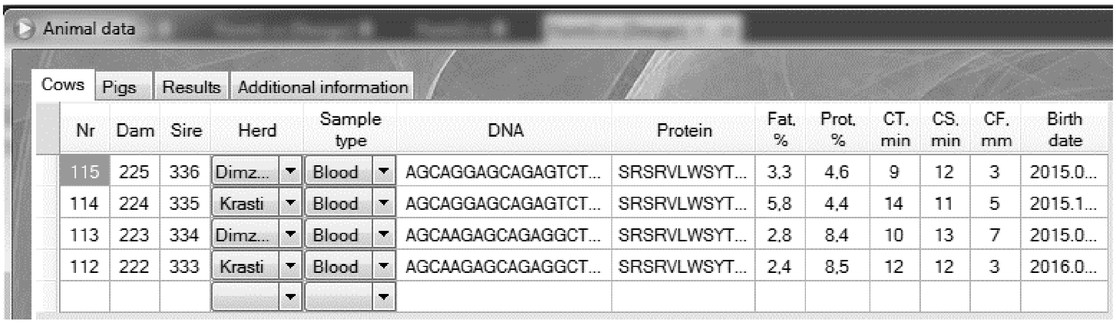
The user has the access to the following functionality of the software: adding, correction and deleting the data, to alignment of two sequences and data filtration according to one or more parameters (Table 2).

Table 2

A summary of the software functionality

Data processing	Purpose of the form	Functionality
Basic form	To introduce user with functional possibilities of the software	The user chooses between possibilities to start work and to read a brief description of the program activities and functional possibilities
Data processing form for the cows	To provide to the user possibilities for data processing using information and data of the cows	Database synchronizing with C# ensures data of the cows adding, deleting, modifying and filtering
Data processing form for the pigs	To provide to the user possibilities for data processing using information and data of the pigs	Database synchronizing with C# ensures data of the pigs adding, deleting, modifying and filtering
Data processing form for the compared individuals	To provide to the user possibilities for data processing using information and data of the sequence alignment results	Database synchronizing with C# ensures data of the results deleting and filtering by statistical and comparison method parameters
Algorithm execution form for the linear algorithm	To compareDNA/protein sequences according to the linear algorithm	Software calculates the values of matrix, trace back and provides statistical evaluation of the results
Algorithm execution form for the affine algorithm	To compareDNA/protein sequences according to the affine algorithm	Software calculates the values of 4 matrices, trace back of the main matrix and provides statistical evaluation of the results

The database was created to store the data of cows and pigs (Fig. 1). The database form consists of the fields that contain additional text, selection boxes, and date-time picker cells.

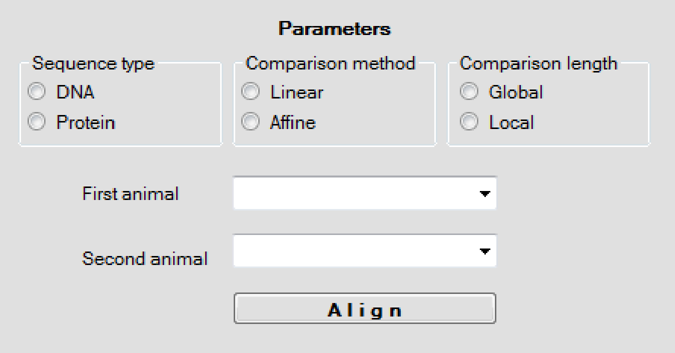


Nr	Dam	Sire	Herd	Sample type	DNA	Protein	Fat. %	Prot. %	CT. min	CS. min	CF. mm	Birth date
115	225	336	Dimz...	Blood	AGCAGGAGCAGAGTCT...	SRSRVLWSYT...	3,3	4,6	9	12	3	2015.0...
114	224	335	Krasti	Blood	AGCAGGAGCAGAGTCT...	SRSRVLWSYT...	5,8	4,4	14	11	5	2015.1...
113	223	334	Dimz...	Blood	AGCAAGAGCAGAGGCT...	SRSRVLWSYT...	2,8	8,4	10	13	7	2015.0...
112	222	333	Krasti	Blood	AGCAAGAGCAGAGGCT...	SRSRVLWSYT...	2,4	8,5	12	12	3	2016.0...

Fig. 1. **Database user interface**

Identification numbers for animals and their ancestors are defined as cells, where only numbers are allowed to be added. The herd and sample type are defined as selection boxes, where multiple choices are accessible. DNA and protein sequences are described as cells, where only text without numbers is allowed to be added. The birth date of the animals is defined as date-time picker cells, where only specific dates are allowed to be chosen. In cells for productivity traits (fat, prot.,etc.)only numbers are allowed to be added and cells had limits to minimum and maximum values according to the range of the productivity traits.

The sequence alignment form consists of three parts: sequence type, sequence comparison method and length (Fig. 2). The user has possibility to choose the animals from the database, to identify the sequence type (DNA or protein) and to select the pairwise sequence comparison method (local or global; linear or affine).



Parameters

Sequence type: ○ DNA ○ Protein

Comparison method: ○ Linear ○ Affine

Comparison length: ○ Global ○ Local

First animal:

Second animal:

Fig. 2. **Sequence alignment user interface**

Pairwise alignment of DNA/protein sequences using the dynamic programming algorithms

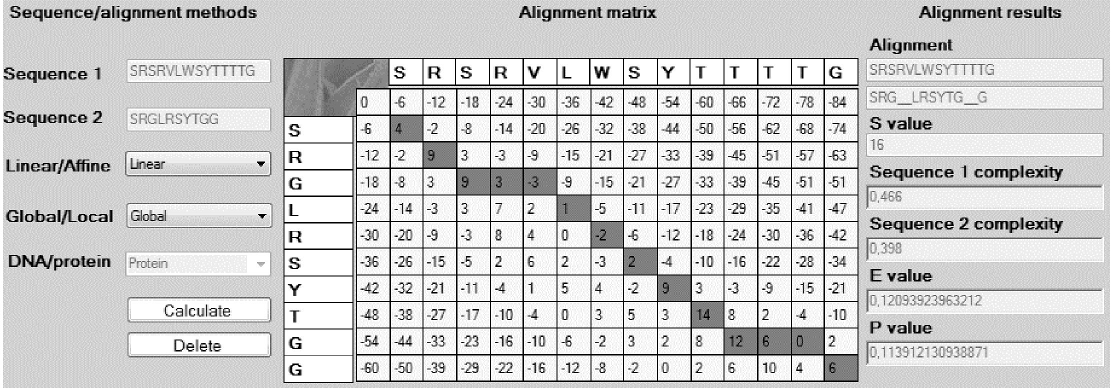
Dynamic programming algorithm for the optimal two DNA or protein sequence alignment includes three steps [1]:

- Initialization of the dynamic programming matrix for the linear sequence alignment or matrices for the affine sequence alignment;
- Definition of optimal score and filling the matrix by solving the smallest subproblems;
- To find the traceback – optimal solution of the sequences alignment that gives the optimal score.

In the linear global comparison (DNA/protein sequence) both sequences are compared in all their length. Using the method of the linear global comparison, the matrix M is developed (1), where the number of the rows and columns is the number of the nucleotides of the sequence +1. The matrix is completed from the left upper corner to the right bottom corner. M(i,j) is the score of the best possible alignment between two sequences. The matrix traceback is startedfrom the right bottom corner to the upper corner by diagonal and identifies the optimal path by the maximum M(i,j)values [3].

In the software presented, the algorithm evaluating sequence similarity (*S* value) (3) and complexity (4) [11] and statistical estimates for sequence similarity was included. Evaluation of statistical estimates for sequence similarity is given by *E* and *P* values, where the E value is the expected number of high-scoring segment pairs (HSPs) with the score at least S score and P value the probability of finding exactly HSPs with the score ≥ *S* [12].

As an example, two animals were compared, which have the following protein sequences – SRSRVLWSYTTTTG and SRGLRSYTGG. The calculation was done by linear global Needleman-Wunsch algorithm (Fig. 3).



Sequence/alignment methods

Sequence 1: SRSRVLWSYTTTTG

Sequence 2: SRGLRSYTGG

Linear/Affine: Linear

Global/Local: Global

DNA/protein: Protein

Alignment matrix

		S	R	S	R	V	L	W	S	Y	T	T	T	T	G
	0	-6	-12	-18	-24	-30	-36	-42	-48	-54	-60	-66	-72	-78	-84
S	-6	4	-2	-8	-14	-20	-26	-32	-38	-44	-50	-56	-62	-68	-74
R	-12	-2	9	3	-3	-9	-15	-21	-27	-33	-39	-45	-51	-57	-63
G	-18	-8	3	9	3	-3	-9	-15	-21	-27	-33	-39	-45	-51	-51
L	-24	-14	-3	3	7	2	1	-5	-11	-17	-23	-29	-35	-41	-47
R	-30	-20	-9	-3	8	4	0	-2	-6	-12	-18	-24	-30	-36	-42
S	-36	-26	-15	-5	2	6	2	-3	2	-4	-10	-16	-22	-28	-34
Y	-42	-32	-21	-11	-4	1	5	4	-2	9	3	-3	-9	-15	-21
T	-48	-38	-27	-17	-10	-4	0	3	5	3	14	8	2	-4	-10
G	-54	-44	-33	-23	-16	-10	-6	-2	3	2	8	12	6	0	2
G	-60	-50	-39	-29	-22	-16	-12	-8	-2	0	2	6	10	4	6

Alignment results

Alignment: SRSRVLWSYTTTTG
SRG__LRSYTG__G

S value: 16

Sequence 1 complexity: 0,466

Sequence 2 complexity: 0,398

E value: 0,12093923963212

P value: 0,113912130938871

Fig. 3. **Linear global protein sequence alignment**

The sequence alignment form is divided in three parts. The left side includes information about the sequence, alignment methods and type of the sequence. The centre of the Figure depicts the dynamic-programming matrix with traiceback. The right side shows the two sequences alignment results with maximum score of alignment (*S* value) and statistical parameters of the sequence alignment.

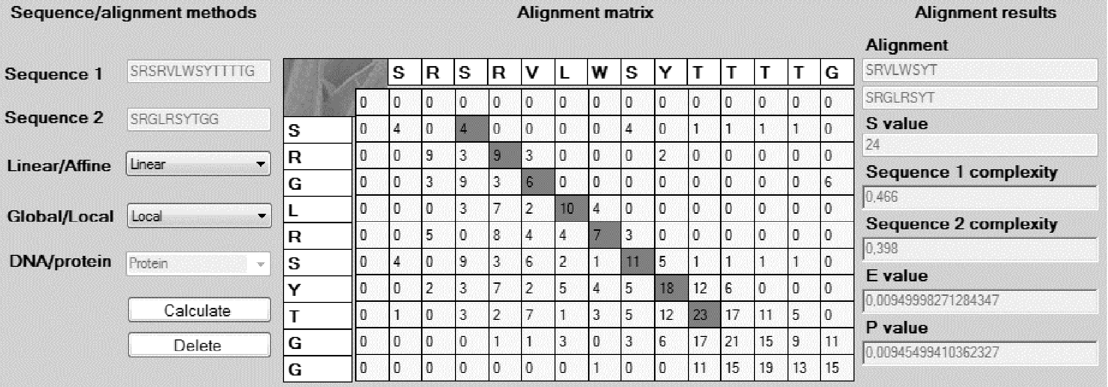
According to the results the optimal global alignment of the two sequences is:

S R S R V L W S Y T T T T G
S R G – – L R S Y T G – – G

There are mismatches in the third, seventh and eleventh columns and insertions in the fourth, fifth, twelfth and thirteenth columns. The global score of the sequences alignment *S* is equal to 16 and it is describing the overall quality of an alignment and higher number of the *S* corresponds to higher similarity of the sequences. The P value shows probability to obtain by chance the value at least equal to $S \geq 16$ for sequences of this size and it is equal to 0.1139 or 11.39 %. The closer *P* value to 0 shows better sequence alignment.

If the global alignment is completed, for the same sequences it is possible to change alignment methods from global to local (Fig. 4).

In local linear comparison case the matrix M is developed (2) and the value of the first row and column of the matrix is 0. The matrix is completed from the left upper corner to the right bottom corner. Trace back begins at the highest value and goes by the maximum values Mi from the right bottom corner to the upper corner by diagonal. The path is completed when the value 0 is achieved [9].



Sequence/alignment methods

Sequence 1: SRSRVLWSYTTTTG

Sequence 2: SRGLRSYTGG

Linear/Affine: Linear

Global/Local: Local

DNA/protein: Protein

Alignment matrix

		S	R	S	R	V	L	W	S	Y	T	T	T	T	G
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	4	0	4	0	0	0	0	4	0	1	1	1	1	0
R	0	0	9	3	9	3	0	0	0	2	0	0	0	0	0
G	0	0	3	9	3	6	0	0	0	0	0	0	0	0	6
L	0	0	0	3	7	2	10	4	0	0	0	0	0	0	0
R	0	0	5	0	8	4	4	7	3	0	0	0	0	0	0
S	0	4	0	9	3	6	2	1	11	5	1	1	1	1	0
Y	0	0	2	3	7	2	5	4	5	18	12	6	0	0	0
T	0	1	0	3	2	7	1	3	5	12	23	17	11	5	0
G	0	0	0	0	1	1	3	0	3	6	17	21	15	9	11
G	0	0	0	0	0	0	0	1	0	0	11	15	19	13	15

Alignment results

Alignment: SRVLWSYT / SRGLRSYT

S value: 24

Sequence 1 complexity: 0,466

Sequence 2 complexity: 0,398

E value: 0,00949998271284347

P value: 0,00945499410362327

Fig. 4. **Linear local protein sequence alignment**

The optimal local alignment of the two sequences was done. According to the results the optimal local alignment of the two sequences is:

S R V L W S Y T
S R G L R S Y T

There are mismatches in the third and fifth columns. The local score of the sequences alignment *S* is equal to 24 and it means the local alignment of sequences has higher similarity compared to the sequences global alignment. The *P* value in local sequence alignment is closer to 0 compared to the global alignment, and probability to obtain by chance the value at least equal to $S \geq 24$ for sequences of this size is equal to 0.0095 or 0.95 %. The P value is used for sequence alignment hypothesis testing, where the null hypothesis is that the aligned sequences are unrelated and the alternative hypothesis is the sequences are biologically related [12]. According to the local alignment results the sequences are related, because the *P* value = 0.0095 and is smaller than $\alpha = 0.05$ (*P* value <0.05).

Conclusions

As a result of the project the software for data storage and pair wise sequence alignment was developed. In the software database information about animal (cows and pigs) DNS/protein sequences, and cow milk and pig meat quality traits which will be analysed in laboratory can be stored. The main task of the software is to execute the Needleman-Wunsch and Smith-Waterman global and local pair wise sequence alignment algorithms. One of the possible scenarios of further development is to provide multi-sequence alignment. According to the results from multi sequence alignment it is possible to develop phylogenetic trees by UPGMA, Neighbour joining algorithms [13].

References

1. Eddy S. R. What is dynamic programming? Nature Biotechnology, 20014, vol. 22, pp. 909-910
2. Huang X., Chao K. A generalized global alignment algorithm. Bioinformatics, 2003, vol. 19, pp. 228-233.
3. Needleman S. B., Wunsch C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. The Journal of Molecular Biology, 1970, vol. 48, pp. 443-453.
4. Smith T.F., Waterman M.S. Identification of common molecular subsequences. The Journal of Molecular Biology, 1981, vol. 147, pp. 195-197.
5. Chakraborty A., Bandyopadhyay S. FOGSAA: Fast Optimal global sequence alignment algorithm. The Journal of Scientific Reports, 2013, vol. 3 (1746),pp. 1-9.
6. Pearson W. R., Wood T. C. Statistical significance in biological sequence comparison. [online] [10.04.2016]. Available at: ttp://faculty.virginia.edu/wrpearson/papers/stat_gen_00.pdf
7. PostgreSQL 9.4 Released. [online] [25.01.2016]. Available at: http://www.bizjournals.com/prnewswire/press_releases/2014/12/18/MN92627
8. PostgreSQL. Featureful and standards compilation. [online] [25.01.2016]. Available at: http://www.postgresql.org/about/
9. ECMA C# and Common Language Infrastructure Standards. [online] [27.01.2016]. Available at: https://www.visualstudio.com/en-us/mt639507.aspx
10. Wootton J.C., Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol., 1996, vol. 266, pp. 554-571.
11. Orlov Y. L., Potapov V. N. Complexity: an Internet resource for analysis of DNA sequence complexity. Nucleic Acids Research, 2004, vol. 32, pp. 628-633.
12. Mitrophanov A., Borodovsky M. Statistical significance in biological sequence analysis. Briefings in Bioinformatics, 2006, vol. 7 (1), pp. 2-24.
13. Feng D., Doolittle R. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. Meth Enzymol, 1996, vol. 266, pp. 368-382.