

GOCE DELCEV UNIVERSITY - STIP
FACULTY OF COMPUTER SCIENCE

ISSN 2545-479X print
ISSN 2545-4803 on line

**BALKAN JOURNAL
OF APPLIED MATHEMATICS
AND INFORMATICS
(BJAMI)**



YEAR 2018

VOLUME I, Number 1

GOCE DELCEV UNIVERSITY - STIP, REPUBLIC OF MACEDONIA
FACULTY OF COMPUTER SCIENCE

ISSN 2545-479X print
ISSN 2545-4803 on line

BALKAN JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS



BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS

(BJAMI)

AIMS AND SCOPE:

BJAMI publishes original research articles in the areas of applied mathematics and informatics.

Topics:

1. Computer science;
2. Computer and software engineering;
3. Information technology;
4. Computer security;
5. Electrical engineering;
6. Telecommunication;
7. Mathematics and its applications;
8. Articles of interdisciplinary of computer and information sciences with education, economics, environmental, health, and engineering.

Managing editor

Biljana Zlatanovska Ph.D.

Editor in chief

Zoran Zdravev Ph.D.

Technical editor

Slave Dimitrov

Address of the editorial office

Goce Delcev University – Štip
Faculty of philology
Krstev Misirkov 10-A
PO box 201, 2000 Štip,
R. of Macedonia

**BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI), Vol 1**

**ISSN 2545-479X print
ISSN 2545-4803 on line
Vol. 1, No. 1, Year 2018**

EDITORIAL BOARD

- Adelina Plamenova Aleksieva-Petrova**, Technical University – Sofia,
Faculty of Computer Systems and Control, Sofia, Bulgaria
- Lyudmila Stoyanova**, Technical University - Sofia , Faculty of computer systems and control,
Department – Programming and computer technologies, Bulgaria
- Zlatko Georgiev Varbanov**, Department of Mathematics and Informatics,
Veliko Tarnovo University, Bulgaria
- Snezana Scepanovic**, Faculty for Information Technology,
University “Mediterranean”, Podgorica, Montenegro
- Daniela Veleva Minkovska**, Faculty of Computer Systems and Technologies,
Technical University, Sofia, Bulgaria
- Stefka Hristova Bouyuklieva**, Department of Algebra and Geometry,
Faculty of Mathematics and Informatics, Veliko Tarnovo University, Bulgaria
- Vesselin Velichkov**, University of Luxembourg, Faculty of Sciences,
Technology and Communication (FSTC), Luxembourg
- Isabel Maria Baltazar Simões de Carvalho**, Instituto Superior Técnico,
Technical University of Lisbon, Portugal
- Predrag S. Stanimirović**, University of Niš, Faculty of Sciences and Mathematics,
Department of Mathematics and Informatics, Niš, Serbia
- Shcherbacov Victor**, Institute of Mathematics and Computer Science,
Academy of Sciences of Moldova, Moldova
- Pedro Ricardo Morais Inácio**, Department of Computer Science,
Universidade da Beira Interior, Portugal
- Sanja Panovska**, GFZ German Research Centre for Geosciences, Germany
- Georgi Tuparov**, Technical University of Sofia Bulgaria
- Dijana Karuovic**, Tehnical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia
- Ivanka Georgieva**, South-West University, Blagoevgrad, Bulgaria
- Georgi Stojanov**, Computer Science, Mathematics, and Environmental Science Department
The American University of Paris, France
- Iliya Guerguiev Bouyukliev**, Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences, Bulgaria
- Riste Škrekovski**, FAMNIT, University of Primorska, Koper, Slovenia
- Stela Zhelezova**, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
- Katerina Taskova**, Computational Biology and Data Mining Group,
Faculty of Biology, Johannes Gutenberg-Universität Mainz (JGU), Mainz, Germany.
- Dragana Glušac**, Tehnical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia
- Cveta Martinovska-Bande**, Faculty of Computer Science, UGD, Macedonia
- Blagoj Delipetrov**, Faculty of Computer Science, UGD, Macedonia
- Zoran Zdravev**, Faculty of Computer Science, UGD, Macedonia
- Aleksandra Mileva**, Faculty of Computer Science, UGD, Macedonia
- Igor Stojanovik**, Faculty of Computer Science, UGD, Macedonia
- Saso Koceski**, Faculty of Computer Science, UGD, Macedonia
- Natasa Koceska**, Faculty of Computer Science, UGD, Macedonia
- Aleksandar Krstev**, Faculty of Computer Science, UGD, Macedonia
- Biljana Zlatanovska**, Faculty of Computer Science, UGD, Macedonia
- Natasa Stojkovik**, Faculty of Computer Science, UGD, Macedonia
- Done Stojanov**, Faculty of Computer Science, UGD, Macedonia
- Limonka Koceva Lazarova**, Faculty of Computer Science, UGD, Macedonia
- Tatjana Atanasova Pacemska**, Faculty of Electrical Engineering, UGD, Macedonia



CONTENT

Aleksandar, Velinov, Vlado, Gicev PRACTICAL APPLICATION OF SIMPLEX METHOD FOR SOLVING LINEAR PROGRAMMING PROBLEMS	7
Biserka Petrovska , Igor Stojanovic , Tatjana Atanasova Pachemska CLASSIFICATION OF SMALL DATA SETS OF IMAGES WITH TRANSFER LEARNING IN CONVOLUTIONAL NEURAL NETWORKS	17
Done Stojanov WEB SERVICE BASED GENOMIC DATA RETRIEVAL	25
Aleksandra Mileva, Vesna Dimitrova SOME GENERALIZATIONS OF RECURSIVE DERIVATES OF k-ary OPERATIONS	31
Diana Kirilova Nedelcheva SOME FIXED POINT RESULTS FOR CONTRACTION SET - VALUED MAPPINGS IN CONE METRIC SPACES	39
Aleksandar Krstev, Dejan Krstev, Boris Krstev, Sladzana Velinovska DATA ANALYSIS AND STRUCTURAL EQUATION MODELLING FOR DIRECT FOREIGN INVESTMENT FROM LOCAL POPULATION	49
Maja Srebrenova Miteva, Limonka Koceva Lazarova NOTION FOR CONNECTEDNESS AND PATH CONNECTEDNESS IN SOME TYPE OF TOPOLOGICAL SPACES	55
The Appendix	
Aleksandra Stojanova , Mirjana Kocaleva , Natasha Stojkovic , Dusan Bikov , Marija Ljubenovska , Savetka Zdravevska , Biljana Zlatanovska , Marija Miteva , Limonka Koceva Lazarova OPTIMIZATION MODELS FOR SCHEDULING IN KINDERGARTEN AND HEALTHCARE CENTRES	65
Maja Kukuseva Paneva, Biljana Citkuseva Dimitrovska, Jasmina Veta Buralieva, Elena Karamazova, Tatjana Atanasova Pacemska PROPOSED QUEUING MODEL M/M/3 WITH INFINITE WAITING LINE IN A SUPERMARKET	73
Maja Mijajlovikj1, Sara Srebrenkoska, Marija Chekerovska, Svetlana Risteska, Vineta Srebrenkoska APPLICATION OF TAGUCHI METHOD IN PRODUCTION OF SAMPLES PREDICTING PROPERTIES OF POLYMER COMPOSITES	79
Sara Srebrenkoska, Silvana Zhezhova, Sanja Risteski, Marija Chekerovska Vineta Srebrenkoska Svetlana Risteska APPLICATION OF FACTORIAL EXPERIMENTAL DESIGN IN PREDICTING PROPERTIES OF POLYMER COMPOSITES	85
Igor Dimovski, Ice Gjumandeloski, Filip Kochoski, Mahendra Paipuri, Milena Veneva , Aleksandra Risteska COMPUTER AIDED (FILAMENT WINDING) TAPE PLACEMENT FOR ELBOWS. PRACTICALLY ORIENTATED ALGORITHM	89

WEB SERVICE BASED GENOMIC DATA RETRIEVAL

Done Stojanov

Faculty of computer science, Goce Delcev University, Stip, Macedonia
done.stojanov@ugd.edu.mk

Abstract: An application of web service for genomic data retrieval is considered. Records for nucleotide sequences are retrieved from the European Nucleotide Archive and preprocessed locally in order to be able to apply local host based analysis. This analysis is required because many of the built-in EMBL-EBI services pose restrictions upon the metrics with which will operate the user, what in some cases may also affect the structure of the solution, such as when performing pairwise or multiple DNA or mRNA comparison.

Keywords: web, service, data, retrieval, ENA

1. Introduction

In recent years, a huge volume of genomic data has been accessed. This was possible due to the advances in the DNA sequencing techniques, such as shotgun sequencing and bridge PCR technique. The value of the collected DNA information increases by sharing in the scientific community. This idea drove the construction of the central repository of nucleotide data: *ENA* or *The European Nucleotide Archive* [1].

ENA or The European Nucleotide Archive (web access: <https://www.ebi.ac.uk/ena>) [1] is a public database hosted under the European Bioinformatics Institute (EMBL-EBI) [2] that contains records for nucleotide sequences. ENA allows upload/access/download of nucleotide data derived from different sources (organisms) applying different sequencing techniques.

ENA evolved from the EMBL Data Library which was officially released in 1982 [3] and it contained 568 records with total size of 500,000 base pairs [4]. Since then the volume of the data in this repository exponentially grows [5].

Data stored in this repository is functionally annotated. This means that in spite of the exact order of nucleotides in the sequence, data linkage details are also available. For instance, when accessing a whole genome, each CDS (coding sequence) is differentiated from other coding frames, as well as from non-coding frames. By knowing the start and the end of each CDS, the structure of the mRNA being transcribed and the protein being translated can be exactly determined.

Regardless record's functional association, there is metadata common for all. This means that regardless the record is associated to a whole genome, chromosome or partial CDS (partial coding region or partial mRNA) there are descriptors which are common for all records.

For each record there is information that describes: the source (organism) of nucleotide sequence, type and topology of the molecule, taxonomic division or taxonomic rank (ex. *HUM* (*Human*), *PRO* (*Prokaryote*)...etc.), length of the sequence or the number of base pairs, sequence version, date when the sequence was made public and the date of last update of the sequence. Keywords that describe the sequence and secondary accession information are also available.

However, the most important feature of each record in ENA is that there is no restriction in the accession of exact structure of nucleotide sequence. This information can be accessed in three data formats: *TEXT*, *FASTA* and *XML*.

The European Bioinformatics Institute (EMBL-EBI) also hosts web implementations of the most popular algorithms for genomic data analysis, such as: EMBOSS Needle – an implementation of the Needleman-Wunsch algorithm [6], EMBOSS Water – an implementation of the Smith-Waterman algorithm [7], EMBOSS Matcher – an

implementation of the LALIGN algorithm [8], EMBOSS backtranambig – reverses protein into nucleotide sequence...etc. Most of these services work with data from the European Nucleotide Archive under some programming constraints. For instance when comparing two nucleotide sequences using EMBOSS Water, only 8 values for gap opening are available: {1,5,10,15,20,25,50,100} and there is no option to choose value out of the list. But since the structure of the solution depends on the metrics being chosen, especially in terms of accuracy and comprehensiveness, there is a real need to develop a program which would be able to retrieve and filter ENA data locally in order to apply algorithms that overcome the limitations posed by EMBL-EBI web service implementations.

In this paper a web service based application that retrieves and filters data from the European Nucleotide Archive is considered.

2. Materials and methods

The user interface of the European Nucleotide Archive provides search services for nucleotide sequences based on: *sequence ID*, *descriptor* and it also provides search against a *specific pattern*, Figure 1.



Figure 1. ENA user interface

Each sequence in ENA has a unique ID which is a combination of letters and numbers. For instance, *CP025048* is ENA ID of *Escherichia coli strain SC516 chromosome, complete genome (bacteria)*. Unlike the use of ENA ID which provides search for specific record, descriptors provide general search. Descriptors are keywords associated to the structure, function or taxonomic division of the nucleotide sequence. In this way, we can search ENA for

complete genome or chromosome or for human, virus, bacteria...samples. The search against specific pattern identifies all ENA sequences that contain the referred pattern.

Once the record has been accessed, data views in: *TXT*, *FASTA* and *XML* are available, Figure 2. FASTA is the most suitable data format for easy separation of metadata and genomic data. Given FASTA view for a specific record, only the first line contains metadata regarding the name of the sequence and ID, while the rest is the structure of the nucleotide sequence, Figure 3.

Sequence: CP025048.1

Escherichia coli strain SC516 chromosome, complete genome.

View: [TEXT](#) [FASTA](#) [XML](#)

Figure 2. Options to view record

```
>ENA|CP025048|CP025048.1 Escherichia coli strain SC516 chromosome, complete genome.
AACAGAGGGCTATTACTTGCACAGATTAAAGATTGTGAATAGTTACCAGCAGTCATT
TACCCGCTTATAACAAGCGAGGCAGTTGTAATGATAGCTCAGAAGGATTATGCAAGGCTT
CGTAAGGGAGAACGCATATAACCACTTCTGTGCATACTGTTGAGCTGAAAAAAGTACGAAT
TATGATAAACTCCAGCCAACCTTATTTCATATCATTGAGGGCCTGTGGCTGATGGCACAG
CTATATTTCTACTATTCCGCAATGAATGCGGGTAAGTCTACAGCATTGTTGCAATCTTCA
TACAATTACCAGGAAACGCGGCATGCGCACTGTCTATATACGGCAGAAATTGATGATCGC
```

Figure 3. Record view in FASTA preview

In order to retrieve genomic data locally, a web client application in C# was developed (we work with object derived from the *WebClient()* class), Figure 4. This application exploits FASTA data format.

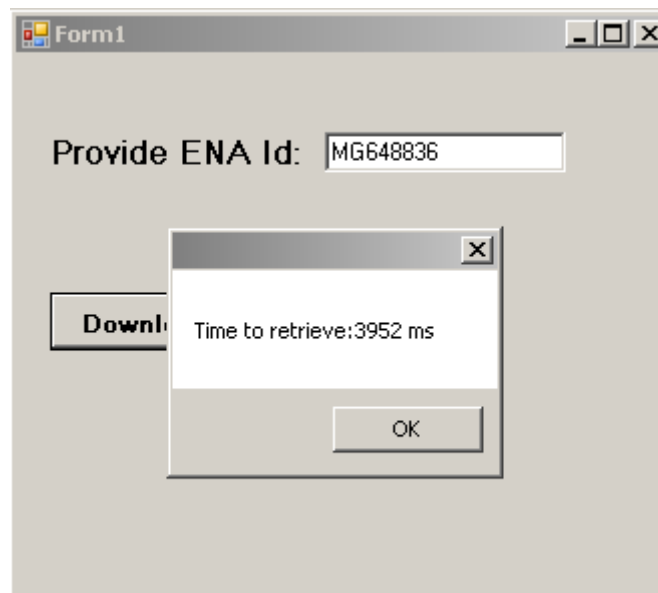


Figure 4. Web client application

First, the user must provide correct *ENA ID* of sequence. This parameter is provided in *TextBox* control, Figure 4. Since the ENA URL of each record in FASTA data format follows this pattern:

"<http://www.ebi.ac.uk/ena/data/view/>" + *ENA_ID_of_Record* + "&display=fasta", this value has to be appended between "<http://www.ebi.ac.uk/ena/data/view/>" and "&display=fasta".

To retrieve the genomic data in local *.txt* file, we call the *DownloadFile()* function. This function is called for the web client object and it has two parameters. The first one is the *URL* of the data being retrieved, while the second is the location in the host where the data is written (@*"location"*).

The code lines below summarize this discussion.

```
WebClient Object = new WebClient();  
  
string URL_String = "http://www.ebi.ac.uk/ena/data/view/" + textBox.Text +  
&"display=fasta";  
  
Object.DownloadFile(new Uri(URL_String), @"location");
```

After sequence retrieval, we face another task. If we want to process genomic data we must separate genomic data from metadata. If we analyze the structure of ENA record, Figure 3, we can notice that ' ' (single space) is used as a metadata inner separator, Figure 3. The end of metadata information (details regarding the name of the sequence and its ID) is followed by only one ' ' (single space) which is followed by genomic data without any interruption, Figure 3.

So if we want to extract genomic data, first we must read the content from the local *.txt* file. The *Split()* procedure with delimiter ' ' (single space) is called. This procedure splits record and stores into an array. The last element in this array (*array[array.Count() - 1]*) contains only genomic data.

The code lines below summarize the previous discussion.

```
string string_name = File.ReadAllText(@"location");  
  
string[] array = string_name.Split(' ');  
  
string genomic_data = array[array.Count() - 1];
```

3. Results and discussion

Applying the web client application from the previous section, the average retrieval time was analyzed on *Escherichia coli plasmid pCOV_clone COV_ samples*, Table 1, with lengths in range between 16.5 kbp and 80 kbp (kilo base pairs). Each sample was retrieved three times and the corresponding retrieval times were measured accordingly in: Test 1, Test 2 and Test 3, Table 1. Retrieval average time (Test 1+Test 2+Test 3)/3 was commuted for each sample. Tests were performed on Acer Aspire 5570Z with Genuine Intel T2080 @ 1.73 GHz and 2 GB RAM with stable net connection.

In general the retrieval time is impacted by: *the network speed, host performance* and *the size* of the sequence being retrieved. Since the first two parameters can be considered to be static (tests are performed on the same machine with constant download speed), the size of the sequence would have greatest impact on retrieval time.

Obtained results showed that longer sequence has been retrieved, more time to retrieve the sequence was spent, Table 1, Figure 5. In order to retrieve the shortest *Escherichia coli plasmid pCOV1 clone COV1_c1* sequence of 16.5 kbp 3.9 seconds were required in average, while 5.1 seconds were required to retrieve the longest *Escherichia coli plasmid pCOV9 clone COV9_c1* sample of 80 kbp, Table 1, Figure 5. Linear trend line between the length of the sequence and the retrieval time can be also established, Figure 5.

Table 1: Average Retrieval Time for different size samples

Sequeunce	Length (bp)	ENA ID	Test 1 (ms)	Test 2 (ms)	Test 3 (ms)	RAT (Retrieval Average Time) (ms)
Escherichia coli plasmid pCOV1 clone COV1_c1	16537	MG648836	3536	4301	3952	3929,666667
Escherichia coli plasmid pCOV6 clone COV6_c2	26897	MG648862	4686	3550	3962	4066
Escherichia coli plasmid pCOV26 clone COV26_c1	32494	MG649013	3594	3820	5004	4139,333333
Escherichia coli plasmid pCOV13 clone COV13_c1	53034	MG648915	3916	3864	4778	4186
Escherichia coli plasmid pCOV9 clone COV9_c1	79483	MG648907	5471	4906	5043	5140

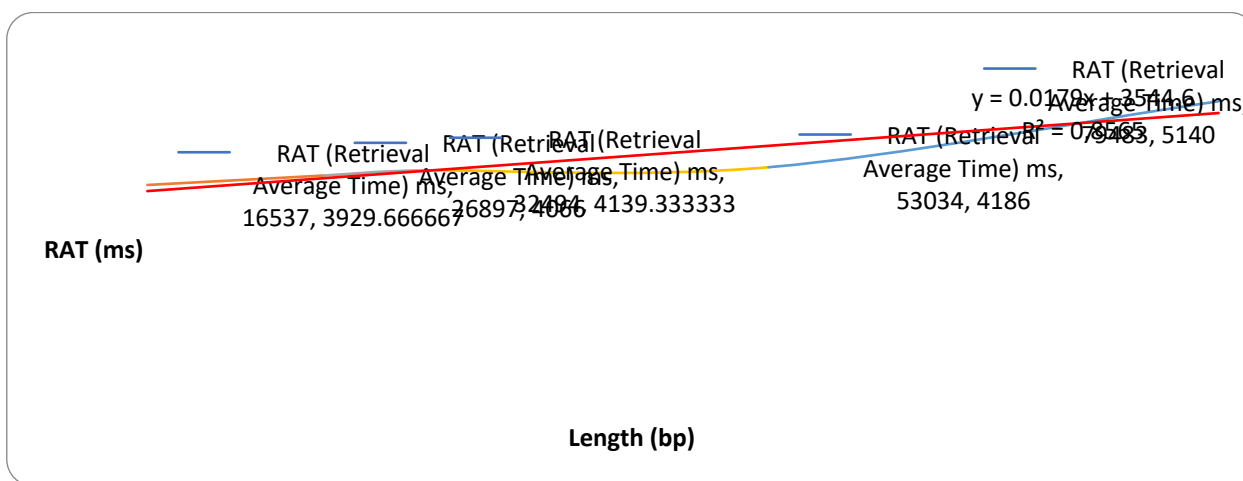


Figure 5. Best fitting line and linear trend line for RAT samples in Table 1

Concluding remarks

A web client application that retrieves and extracts genomic data directly from the world-famous repository of nucleotide sequences – the European Nucleotide Archive was considered. This solution provides genomic data retrieval for local host-based analysis and that is necessary in order to overcome the limitations posed by EMBL-EBI web services, such as limitations upon the value of the gap penalty when aligning two samples, what is some cases may affect the structure of the solution.

References

- [1] The European Nucleotide Archive. <https://www.ebi.ac.uk/ena>
- [2] The European Bioinformatics Institute. <https://www.ebi.ac.uk/>
- [3] Hamm, G.H., Cameron, G.N. (1986). "The EMBL data library": Oxford University Press, Nucleic Acids Research. 14(1): 5–9.
- [4] Kneale, G., Kennard, O. (1984). "The EMBL nucleotide sequence data library": Portland Press, Biochemical Society Transactions. 12(6): 1011–1014.
- [5] Cochrane, G., Alako, B., Amid, C., Bower, L., Cerdeno-Tarraga, A., Cleland, I., Gibson, R., Goodgame, N., Jang, M. (2012). "Facing growth in the European Nucleotide Archive": Oxford University Press, Nucleic Acids Research. 41(D1): D30–D35.

- [6] Needleman, S.B., Wunsch, C.D. (1970). "*A general method applicable to the search for similarities in the amino acid sequence of two proteins*": Elsevier, Journal of Molecular Biology. 48(3): 443–453.
- [7] Smith, T.F., Waterman, M.S. (1981). "*Identification of Common Molecular Subsequences*": Elsevier, Journal of Molecular Biology. 147: 195–197.
- [8] Pearson, W.R. (1994). "*Using the FASTA program to search protein and DNA sequence databases*": Humana Press, Computer Analysis of Sequence Data. 307-331.