Mini-BLAST: computer systems to search for the pattern sequences in the bioinformatics databases

Mini-BLAST: sistema computacional para encontrar similitudes entre secuencias de ADN en bases de datos bioinformáticas

Gennadiy Burlak,[1]* Christian Eduardo Martínez-Guerrero,[1]
Enrique Merino-Pérez[2]

[1] Centro de Investigación en Ingeniería y Ciencias Aplicadas, Universidad Autónoma del Estado de Morelos.
Av. Universidad 1001, col. Chamilpa, CP 62210, Cuernavaca, Morelos, México.
[2] Instituto de Biotecnología, Universidad Nacional Autónoma de México.
Av. Universidad 2001, col. Chamilpa, Cuernavaca, Morelos, CP 62210, México.
* E-mail: burlak@uaem.mx

KEYWORDS:

bioinformatics, computer systems for searching DNA sequences, parallel algorithms, GUI.

ABSTRACT

Bioinformatics focuses on developing and applying computationally intensive techniques to increase the understanding of biological processes. In this report we describe the compact computer system Mini-BLAST, designed to find DNA sequences in bioinformatics databases placed in local or web configurations. Our system allows the identification of gene sequences relating to new patterns (metagenome) that are not yet identified in such databases, containing data on known nucleotides. Such a task is a quite expensive and time consuming operation; therefore, for large genomes parallel algorithms are required. We developed a user-friendly graphical interface that allows the simple input of query data and outputs representative statistical analysis. Additionally, users can select particular databases for cases when a specific alignment is required. Although the package is developed in MS .NET 3.5/4.0 Visual C# system, it works with no limitations in Linux through the Mono framework.

PALABRAS CLAVE:

bioinformática, sistemas para la búsqueda de secuencias de ADN, algoritmos paralelos, GUI.

RESUMEN

La bioinformática se enfoca en el desarrollo y aplicación de técnicas de cómputo intensivo con la finalidad de incrementar el entendimiento de procesos biológicos. En este reporte describimos un sistema computacional compacto (Mini-BLAST) que encuentra similitudes entre secuencias de ADN (ácido desoxirribonucleico) en bases de datos bioinformáticas almacenadas de forma local o en configuraciones web. Nuestro sistema permite identificar secuencias de genes nuevas, tales como metagenomas, a partir de secuencias nucleotídicas conocidas. Dado que esta tarea consume mucho tiempo y recursos computacionales, para genomas grandes se requiere procesamiento en paralelo. Desarrollamos una interfaz gráfica amigable al usuario que permite introducir de manera simple datos y devuelve análisis estadísticos representativos a la salida. Adicionalmente, el usuario puede seleccionar determinados organismos de las bases de datos, cuando requiere alineamientos más específicos. Este paquete fue diseñado en MS .NET 3.5/4.0 Visual C#, sin embargo funciona sin limitaciones en Linux a través del sistema Mono.

Recibido: 19 de septiembre de 2011 • **Aceptado**: 19 de noviembre de 2012 • **Publicado en línea**: 6 de noviembre de 2013

1 INTRODUCTION

Bioinformatics focuses on developing computationally intensive techniques (e.g., pattern recognition, data mining, machine learning algorithms, and visualization) to increase the understanding of biological processes. Major research efforts in the field include sequence alignment, gene finding, genome-wide association studies and the modeling of biological evolution. Bioinformatics now entails the creation and advancement of databases, algorithms, and computational and statistical techniques to solve practical problems arising from the management and analysis of biological data. Bioinformatics was applied in the creation and maintenance of a database (DB) designed to store biological information such as nucleotide and amino acid sequences. The creation of this type of database involved not only design issues but the development of complex interfaces whereby researchers could access existing data.

Therefore, the field of bioinformatics has evolved in such a way that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences. The actual process of analyzing and interpreting data is referred to as computational biology. Bioinformatics and computational biology include the development and implementation of tools that enable efficient access to, and use and management of, various types of information, which allows to assess relationships among components of large data sets, such as methods to locate a gene within a sequence.

Nowadays the DNA sequences of thousands of organisms have been decoded and stored in databases. This sequence information is analyzed to determine genes that encode polypeptides (proteins), RNA genes, regulatory sequences, structural motifs, and repetitive sequences. Today, computer programs such as BLAST [1] are used to search sequences from more than 260 000 organisms, containing over 190 billion nucleotides. These programs can compensate for mutations in the DNA sequence, in order to identify sequences that are related, but not identical. They can also be used to reconstruct the complete genome.

Sequence databases can be searched using a variety of methods. The most common is searching for a sequence similar to a certain target protein or gene whose sequence is already known to the user. BLAST is one of the most widely used bioinformatics programs, because it addresses a fundamental problem and its algorithm emphasizes speed over sensitivity. This emphasis on speed is vital to making the algorithm practical on the huge genome databases currently available, although subsequent algorithms can be even faster.

Other algorithms used in database searches for protein or nucleotid sequences employ a full alignment procedure, like Smith-Waterman's [2], that performs local sequence alignment for determining similar regions between two nucleotide or protein sequences. In this a dynamic programming algorithm starts the backtracking at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment.

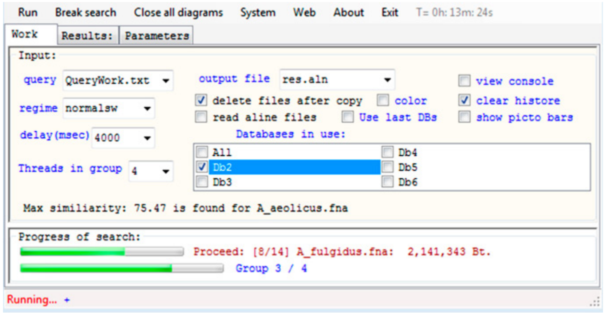
In Mini-BLAST we have used a SW [2] algorithm (adapted for our purpose), which accounts for the uncertainty of homopolymer length by allowing for gaps, preferentially in homopolymers [3]. The basic idea of such approach is to adjust the gap-introduction penalty (gap-opening penalty) dynamically to the "homopolymer-terrain" of a nucleotide sequence, i.e. to use a position specific gap-introduction penalty, which decreases linearly within homopolymers.

In this paper we describe the creation of an advanced computer system with a graphical user interface (GUI) that performs parallel processing of sequences in nucleotide databases to analyze new patterns (metagenomes). We provide the description of our results and benchmark parameters, which show that such a system can be used both in local and web configurations. Such a system can be easily reconfigured for situations when the connection to large bioinformatics databases in the web is slow or even cannot be established.

2 SYSTEM

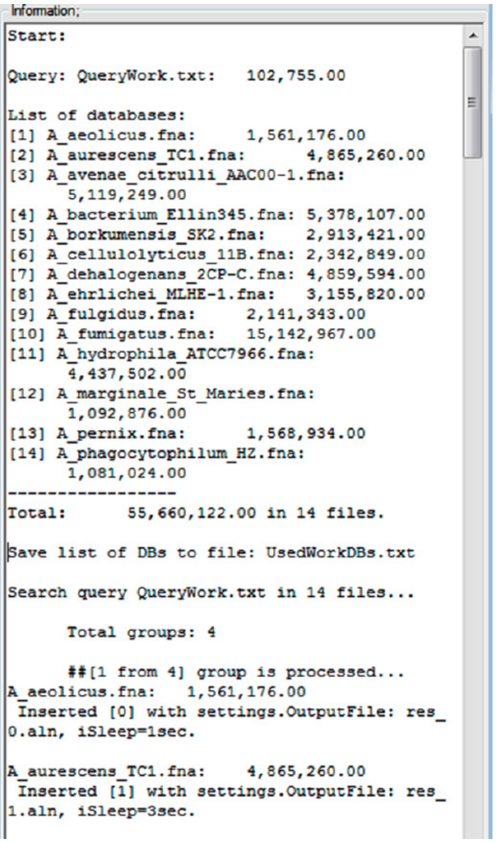
Searching in bioinformatics databases normally consists of the following steps: (i) inputting the patterns of interest to analyze a metagenome, (ii) opening the connection to DNA databases, (iii) the procedure of searching, (iv) the analysis and display of results. For illustrative purposes we show here the case with 13 databases, however the system easily processes 500 databases or more in about an hour or less.

Figure 1 shows the GUI of the program during a typical session. Various visual elements allow choosing optimal parameters to control calculations. Some of the most important are: the name of the metagenome file (pattern query), the regime of Smith-Waterman algorithm, the list of databases in use for current session, the number of threads in group to work in parallel, etc.



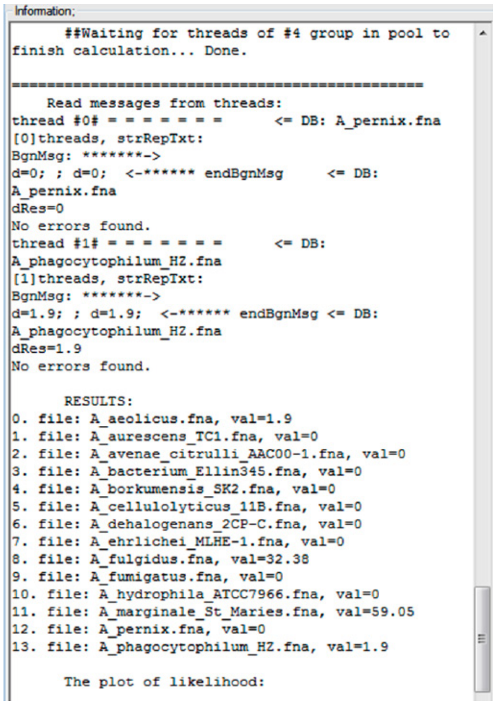
```
T= 0h: 13m: 24s
Input:
query QueryWork.txt      output file res.aln          □ view console
regime normalsw          ☑ delete files after copy  □ color      ☑ clear histore
                         □ read aline files  □ Use last DBs    □ show picto bars
delay(msec) 4000                 Databases in use:
                         □ All                □ Db4
Threads in group 4       ☑ Db2                □ Db5
                         □ Db3                □ Db6

Max similiarity: 75.47 is found for A_aeolicus.fna

Progress of search:
   Proceed: [8/14] A_fulgidus.fna:  2,141,343 Bt.
   Group 3 / 4
Running... +
```

Figure 1. Example search of the pattern file QueryWork.txt in DNA databases in a parallel regime. The case when the number of threads in the searching group is 4 is shown. The top progress bar depicts the progress of searching through the sequences of all databases, while the bottom progress bar shows the progress of parallel processing for current thread group



```
Information;
Start:

Query: QueryWork.txt:    102,755.00

List of databases:
[1] A_aeolicus.fna:      1,561,176.00
[2] A_aurescens_TC1.fna:       4,865,260.00
[3] A_avenae_citrulli_AAC00-1.fna:
      5,119,249.00
[4] A_bacterium_Ellin345.fna: 5,378,107.00
[5] A_borkumensis_SK2.fna:    2,913,421.00
[6] A_cellulolyticus_11B.fna: 2,342,849.00
[7] A_dehalogenans_2CP-C.fna: 4,859,594.00
[8] A_ehrlichei_MLHE-1.fna:    3,155,820.00
[9] A_fulgidus.fna:       2,141,343.00
[10] A_fumigatus.fna:    15,142,967.00
[11] A_hydrophila_ATCC7966.fna:
      4,437,502.00
[12] A_marginale_St_Maries.fna:
      1,092,876.00
[13] A_pernix.fna:       1,568,934.00
[14] A_phagocytophilum_HZ.fna:
      1,081,024.00
------------------
Total:       55,660,122.00 in 14 files.

Save list of DBs to file: UsedWorkDBs.txt

Search query QueryWork.txt in 14 files...

      Total groups: 4

      ##[1 from 4] group is processed...
A_aeolicus.fna:    1,561,176.00
 Inserted [0] with settings.OutputFile: res_
0.aln, iSleep=1sec.

A_aurescens_TC1.fna:     4,865,260.00
 Inserted [1] with settings.OutputFile: res_
1.aln, iSleep=3sec.
```

Figure 2. Log information from the initial input configuration (query) when the pattern search starts. In the top of the window form the name of the pattern query file (metagenome) is shown. In the bottom part the list of databases (with sizes) included in the current session is demonstrated



```
Information;
      ##Waiting for threads of #4 group in pool to
finish calculation... Done.

================================================
    Read messages from threads:
thread #0# = = = = = = =      <= DB: A_pernix.fna
[0]threads, strRepTxt:
BgnMsg: *******->
d=0; ; d=0;  <-****** endBgnMsg     <= DB:
A_pernix.fna
dRes=0
No errors found.
thread #1# = = = = = = =      <= DB:
A_phagocytophilum_HZ.fna
[1]threads, strRepTxt:
BgnMsg: *******->
d=1.9; ; d=1.9;  <-****** endBgnMsg <= DB:
A_phagocytophilum_HZ.fna
dRes=1.9
No errors found.

      RESULTS:
0. file: A_aeolicus.fna, val=1.9
1. file: A_aurescens_TC1.fna, val=0
2. file: A_avenae_citrulli_AAC00-1.fna, val=0
3. file: A_bacterium_Ellin345.fna, val=0
4. file: A_borkumensis_SK2.fna, val=0
5. file: A_cellulolyticus_11B.fna, val=0
6. file: A_dehalogenans_2CP-C.fna, val=0
7. file: A_ehrlichei_MLHE-1.fna, val=0
8. file: A_fulgidus.fna, val=32.38
9. file: A_fumigatus.fna, val=0
10. file: A_hydrophila_ATCC7966.fna, val=0
11. file: A_marginale_St_Maries.fna, val=59.05
12. file: A_pernix.fna, val=0
13. file: A_phagocytophilum_HZ.fna, val=1.9

      The plot of likelihood:
```

Figure 3 Typical output with results of searching and analysis of the pattern structure. The names of databases and the percent of the pattern likelihood for current session are also shown

Typical initial log information is shown in Figure 2, where the structure of the current DB is depicted. Figure 3 shows the information that every thread brings up from the class threadPool at parallel calculations.

In Figure 4 the results of analysis and the diagram of evaluated likelihoods of the pattern metagenome for current session are shown. From Figure 4 we observe that the pattern of interest has a compound structure where the particular similarity exceeds 75% for the organism *A. aeolicus*. Nevertheless other organisms are presented as well in this metagenome in different fractions. Organisms *A. aeolicus*, *B. aphidicola* and *Buch-*

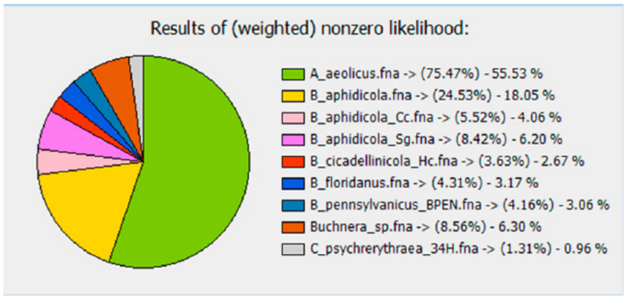


Figure 4. Structure of evaluated likelihood fractions of various organisms in this pattern. We observe that for this metagenome the pattern similarity exceeds 75% for the organism *A. aeolicus*

nera are of most likelihood, and other organisms have similarity of about 5.6% or less. In Figure 4 the number in parenthesis indicates the likelihood evaluated at calculations, the number without parenthesis shows the corresponding fraction in the diagram.

3 DECRYPTION OF THE PROGRAM STRUCTURE AND FLOWCHART

To map ESTs (expressed sequence tags) to genes or whole genomes we have used the library Bioinformatics [3]. Similarly to BLAST [4], such an approach uses an heuristic algorithm to find approximate hits between the database and the query sequence and then extends these hits with dynamic programming.

The program code was developed in MS .Net 3.5/4.0 in Visual C#. While we used this code also in Linux in the Mono system, we did not employ the advanced parallel programming possibilities implemented in MS .Net version 4.0. The code of the program has rather complicated structures, therefore only its main parts, having self-representative shapes, are shown in flowchart in Figure 5.

After parameters initialization and choosing of desired databases of organisms (from menu or in the server side) the system starts the main cycle with threads from the class threadPool. In such a cycle all threads assume and perform the instances of the class SearchInBlast, realizing the parallel mode of calculations with the use of the available PC cores. The class SearchInBlast wraps the connection to the library Bioinformatics [3] in order to implement (normal or homopolymer) Smith-Waterman algorithm to evaluate the likelihood. Besides, our system can recognize the number of cores in PC (that was 4 in our case) and then prompts the value of parameter “Threads in group” of GUI. At such a cycle every thread in group analyzes in parallel the structure of the metagenome and compares the latter with corresponding DB file (organism) to evaluate the likelihood. After last thread finishes the system starts to analyze and then displays the statistics, histograms and diagrams that allow to have an insight into the results. After that the system begins reinitialization and then becomes ready to accept new input data from GUI to start a new session.

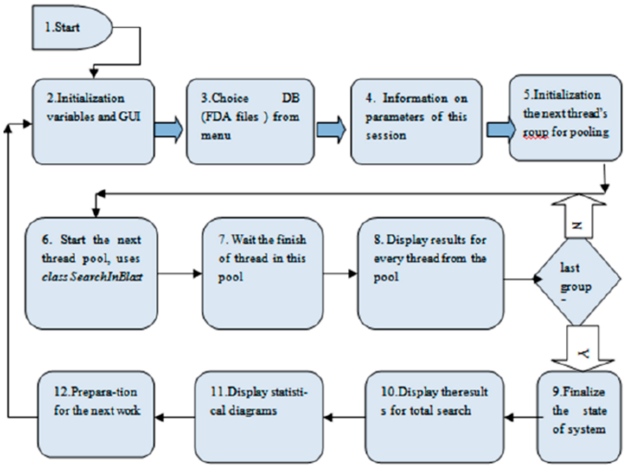

1.Start
2.Initialization variables and GUI
3.Choice DB (FDA files) from menu
4. Information on parameters of this session
5.Initialization the next thread's goup for pooling
6. Start the next thread pool, uses class SearchInBlast
7. Wait the finish of thread in this pool
8. Display results for every thread from the pool
last group
N
Y
9.Finalize the state of system
10.Display theresults for total search
11.Display statistical diagrams
12.Prepara-tion for the next work

Figure 5. The main parts of code structure and flowchart. In case of client-server (web geometry) parts 1-4, and 10-12 belong to the client, while parts 5-9 are placed in the server side

4 EXAMPLE AND BENCHMARKS

In our calculations we have used PC Intel® Xenon ®, 2.64GHz, RAM 4GB, 4 cores. We have calculated an example test of Metagenomics to explore the genomic content in a compound sample. The primary goals of this approach are (i) to characterize the organisms present in a sample and (ii) to identify what roles each organism has within a specific environment. As a sample, a query file was preparede with size about 2Mb that was compared to group organisms from the bioinformatics database. The results are shown in Table 1.

5 CONCLUSION

We have developed an advanced bioinformatics computer package Mini-BLAST with a graphical user interface (GUI) to analyze the pattern structure of metage-

Table 1. Typical time evaluation of the metagenome likelihood in various configurations of the organism’s database.

The pattern query file (metagenome) size is **2.096 Mb**

Files in database	Time	Max. likelihood
Files: 14; total size: 55 Mb; 4 groups	15 min	75.4%
Files: 22; total size: 126 Mb; 6 groups	15 min	75.4%
Files: 112; total size: 455 Mb; 28 groups	46 min	75.4%
Files: 464; total size: 1.6 Gb; 116 groups	1 h 16 min	75.4%

nomes. The advantage is that such a system performs the processing of FDA databases to analyze a pattern in a parallel regime that allows to sharply increment the speed of processing. We provided the description of Mini-BLAST, with test results and benchmarks, which show that such a system is promising to use in both local and web configurations. The system described cannot replace BLAST in general, but it can be useful in situations when connection to large, web-based, bioinformatics databases is slow or even when it cannot be established.

REFERENCES

1. Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403-410.
2. Smith T. F., Waterman M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147: 195-197.
3. Kofler R., Teixeira Torres T., Lelley T, Schlötterer C. (2009) PanGEA: Identification of allele specific gene expression using the 454 technology. BMC Bioinformatics. 10, 143. Available at: http://www.biomedcentral.com/1471-2105/10/143
4. Altschul S. F., Madden T. L., Schäffer A. A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389-3402.

About the authors

Gennadiy Burlak is professor-Investigator C, scientific adviser of magister and PhD students, and SNI level III. In 1975 he received its BCs degree from Taras Shevchenko National University of Kyiv, Faculty of physics, department of Theoretical Physics. In 1975, his Magister degree. In 1979, his Ph. D. in Physics and Mathematics. And in 1988, his PhD of Science, all of them from the same university.

Christian Eduardo Martínez Guerrero is bioinformartic research assistance. In 2003 he received his BS in Computer Engineering at Instituto Tecnológico de Zacatepec. In 2011 his MS in Engineering and Science at Centro de Investigaciones en Ingeniería y Ciencias Aplicadas UAEM. And in 2012 he became Research Assistance at Laboratorio Nacional de Genómica (LANGEBIO-CINVESTAV).

Enrique Merino Pérez is researcher and group leader, SNI level III. He works at Molecular Microbiology Department. In 1982 he received his BS Civil Engineering from UNAM. In 1988 his MS in Biomedical Research at Colegio de Ciencias y Humanidades-CEINGEBI-UNAM. And in 1993, his PhD in Biotechnology from Colegio de Ciencias y Humanidades-CEINGEBI-IBT-UNAM.