MINI-BLAST: Computer Systems to Search for the Pattern Sequences in the Bioinformatics Databases

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, Cuernavaca, Mor. México.
gburlak@uaem.mx
[2]IBT, Universidad Nacional Autónoma de México, av. Universidad 2001,
Cuernavaca, Mor., CP 62210, México

Abstract. The bioinformatics focus on developing and applying computationally intensive techniques to increase the understanding of biological processes. In this report we create the compact computer systems mini-blast and methagraph finding the dna sequences in the bioinformatics databases (dbs) placed in local or web configurations. Our system allows identify the gene sequences relating to new pattern (metagenome) that is not identified yet in such dbs containing data on known nucleotides. Such a task is quite expensive and time consuming operation; therefore for large genomes the parallel algorithms are required. We develop a graphics user-friendly interface (gui) that allows simple input the query data and representative statistical analysis in the output. Additionally, user can select the particular dbs 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 in the mono framework.

Keywords: Bioinformatics, computer finding DNA sequences, parallel algorithms, GUI.

1 Introduction

The bioinformatics focus 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, 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 to store biological information such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues but the development of complex interfaces whereby researchers could both access existing data.

Therefore, the field of bioinformatics has evolved such 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

M.A. Cruz-Chávez (Ed): CICos 2011, ISBN. 978-607-00-5091-6. pp. 359 – 365, 2011

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 assess and relationships among members 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 (Altschul,S.F. et al. 1990) 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, to identify sequences that are related, but not identical. It also can 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 the 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 doing database searches for the protein or nucleic sequences uses a full alignment procedure like Smith-Waterman (Smith et. al 1981) that performs local sequence alignment for determining similar regions between two nucleotide or protein sequences. In this a dynamic programming algorithm the backtracking starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment.

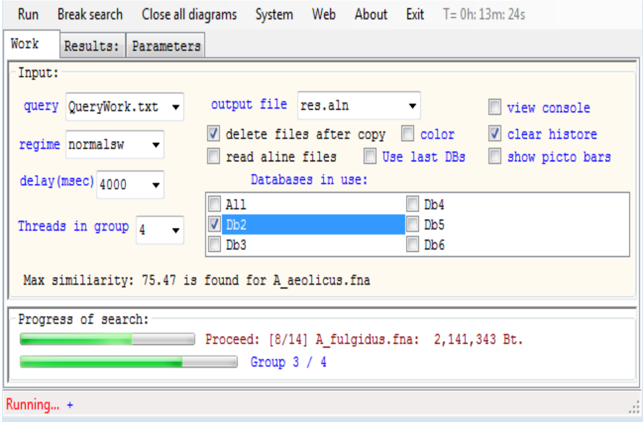
In MINI-BLAST we have used a SW (Smith, Waterman) algorithm (adopted for our purpose), which accounts for this uncertainty of homopolymer length by allowing for gaps preferentially in homopolymers (R. Kofler 2009). The basic idea of such approach is to adjust the gap-introduction penalty (gap-opening 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 create an advanced computer system with graphics user interface (GUI) that perform the parallel processing of the nucleotide databases sequence to analyze a new patterns (metagenoms). We provide the description of our results and benchmark parameters, which show that such a system can be use in both local and WEB configurations. Such a system can be easily reconfigured for situation when the WEB connection to large bioinformatics databases is slow or even cannot be established.

2 System.

Searching in the bioinformatics database normally consists of some steps: (i) input the patterns of interest to analyze a metagenome, (ii) open the connection to DNA databases, (iii) the procedure of searching, (iv) analysis and display the results. For illustrative purposes we show here the case with 13 databases, however the system easily proceeds with more than 500 databases for time about or less than 1 hour.

Fig. 1 shows the GUI of program for a typical session. Various visual elements allow choosing optimal parameters to control by calculations. Main of them are following: 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

regime normalsw

delay(msec) 4000

Threads in group 4

output file res.aln

☑ delete files after copy
□ read aline files
□ color
□ Use last DBs
□ view console
☑ clear histore
□ show picto bars

Databases in use:

□ All
☑ Db2
□ Db3
□ Db4
□ Db5
□ Db6

Max similiarity: 75.47 is found for A_aeolicus.fna

Progress of search:

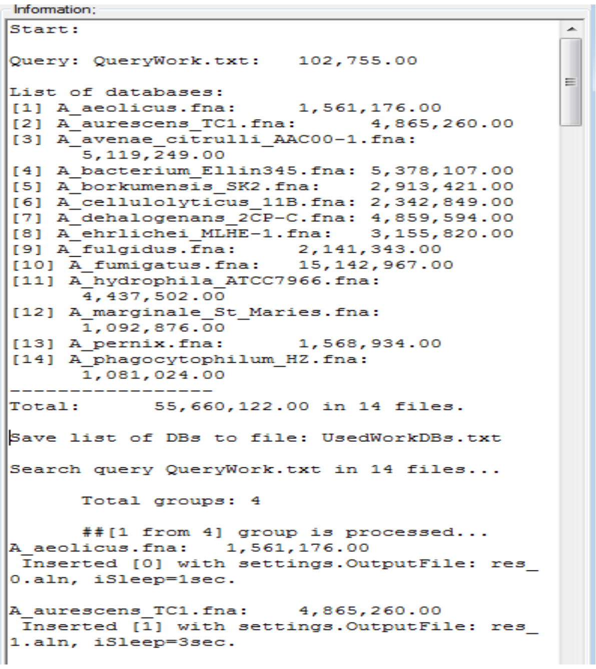
Proceed: [8/14] A_fulgidus.fna: 2,141,343 Bt.

Group 3 / 4

Running... +

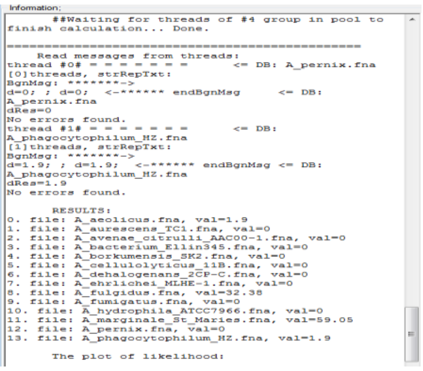
Fig. 1. Searching of the pattern file QueryWork.txt in the DNA databases in a parallel regime. The case when the number of threads in searching group is 4 shown. The top ProgressBar depicts the progress of searching throw the sequence of all databases, while the bottom ProgressBar shows the progress of parallel processing for current thread group.

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



```
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.
```

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



```
      ##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:
```

Fig.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.

In Fig.4 the results of analysis and the diagram of evaluated likelihoods of the pattern metagenome for current session are shown. From Fig.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 as well presented in this metagenome in different fractions. Organisms A. aeolicus, B. aphidicola and Buchnera are of most likelihood, and other organisms have similarity about and less than 5.6%. In Fig. 4 the number in brackets indicates the likelihood evaluated at calculations, the number without brackets show corresponding fraction in the diagram.

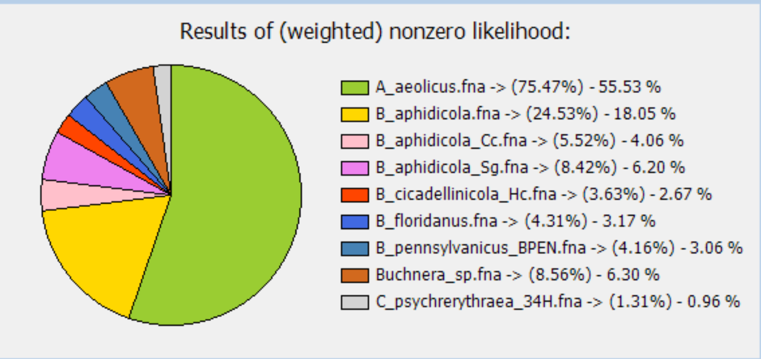


Fig. 4. The structure of evaluated likelihood fractions of various organisms in this pattern. We observe that for this metagenome the pattern similarity exceeds 75% for the organism A. aeolicus.

3 Decryption of the program structure and flow-chart.

To map ESTs (Expressed sequence tag) to genes or whole genomes we have used the library Bioinformatics (R.Kofler et. al. 2009). Similarly to Blast (Altschul,S.F. et al. 1997), such an approach uses a heuristic algorithm to find approximate hits between the database and the query sequence and then extends these hits with dynamic programming.

The program code is developed in MS.Net 3.5/4.0 in Visual C#. While we use 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 main parts having self-representative shapes are shown in flow-chart in Fig.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 (R.Kofler et. al. 2009) 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 allowing the insight of the results. After that the system begins re-initialization and then becomes ready to accept new input data from GUI to start a new session.

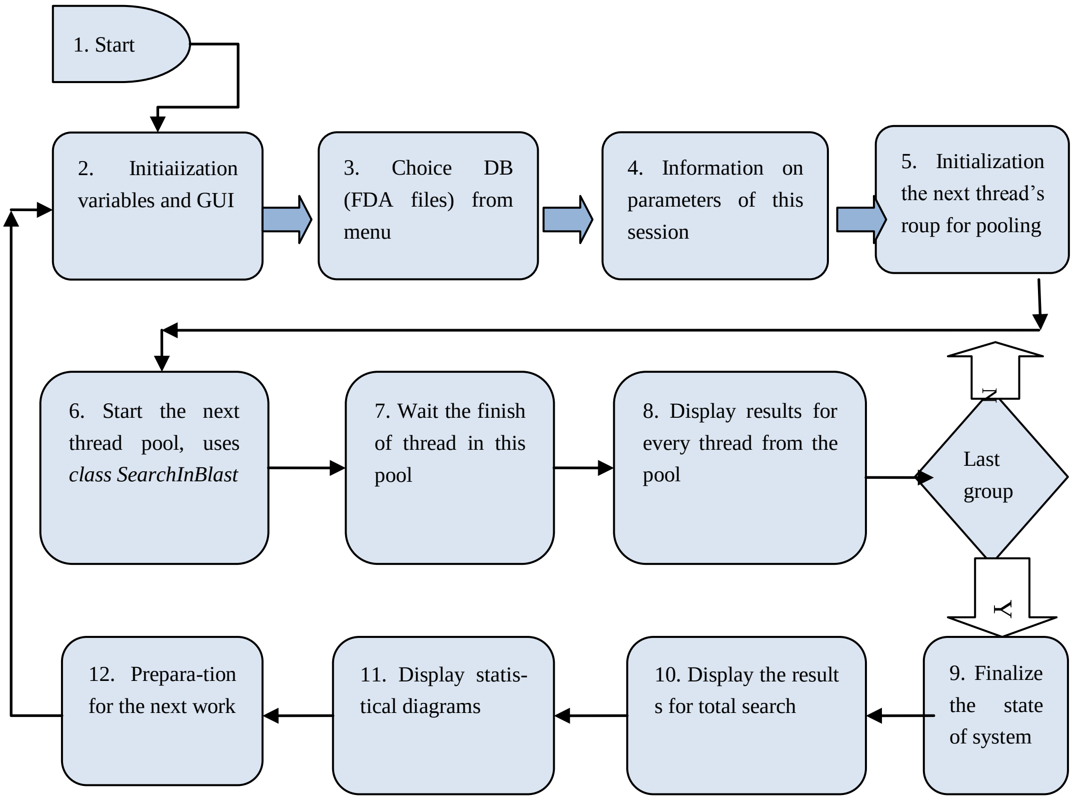

1. Start
2. Initiaiization variables and GUI
3. Choice DB (FDA files) from menu
4. Information on parameters of this session
5. Initialization the next thread's roup for pooling
6. Start the next thread pool, uses *class SearchInBlast*
7. Wait the finish of thread in this pool
8. Display results for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

Fig.5. The main parts of code structure and flow-chart. In case of client-server (WEB geometry) parts 1-4, and 10-12 belong to a 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, Cores 4. We have calculated a test example 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) identify what roles each organism has within a specific environment. As a sample was prepared the query file with size about 2MB that was compared to group organisms from bioinformatics database. The typical size of DB file was in the range 37-0.5 MB. The results are shown in Table 1.

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	14 Min	75,4 %
Files 22, total size 126MB, 6 groups	15 Min	75,4 %
Files 112, total size 455MB, 28 groups	46 Min	75,4 %
Files 464, total size 1.6 GB, 116 groups	1h. 16 min	75,4 %

CONCLUSION

We have developed an advanced bioinformatics computer package MINI-BLAST with graphics user interface (GUI) to analyze the metagenome patters structure. The advantage is that such a system performs the processing of the FDA databases to analyze a pattern in parallel regime that allows to sharply increment the speed of processing. We provided the description of MINI-BLAST, test results and benchmarks, which show that such a system is promising to use in both local and WEB configurations. Such a system cannot replace the BLAST in general, but it can be useful in situation when the WEB connection to large bioinformatics databases is slow or even cannot be established.

REFERENCES

[1]. Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

[2]. Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST:Anew generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

[3]. Smith, Temple F.; and Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences". Journal of Molecular Biology 147: 195–197.

[4]. Robert Kofler, Tatiana Teixeira Torres, Tamas Lelley, and Christian Schlötterer, PanGEA: Identification of allele specific gene expression using the 454 technology. BMC Bioinformatics 2009, 10:143. The electronic version of this article can be found online at: http://www.biomedcentral.com/1471-2105/10/143

[5]. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007, 8(7):R143.

[6]. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB: Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res* 2008, 22:22.