КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ имени АЛЬ-ФАРАБИ

ИНСТИТУТ ВЫЧИСЛИТЕЛЬНЫХ ТЕХНОЛОГИЙ
СИБИРСКОГО ОТДЕЛЕНИЯ РАН

ISSN 1560-7534
ISSN 1563-0285

СОВМЕСТНЫЙ ВЫПУСК

по материалам международной научной конференции
"Вычислительные и информационные технологии в науке, технике и образовании"
(CITech-2015)
(24-27 сентября 2015 года)

ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ

Том 20

ВЕСТНИК КАЗНУ им. АЛЬ-ФАРАБИ

Серия математика, механика и информатика № 3 (86)

ЧАСТЬ I

АЛМАТЫ – НОВОСИБИРСК, 2015

ВЕСТНИК КАЗНУ 2015

Серия математика, механика и информатика № 3 (86)

Научное издание

Вестник КазНУ
Серия математика, механика, информатика
№ 3(86) 2015

ИБ № 8514

Подписано 14.04.2015 г. Формат 60x84 1/8. Бумага офсетная.
Печать цифровая. Объем 30,5 п.л. Тираж 500 экз. Заказ № 2617.
Казахского национального университета им. аль-Фараби.
050040, г. Алматы, пр. аль-Фараби, 71, КазНУ.
Отпечатано в типографии издательского дома «Қазақ университеті».
Издательский дом «Қазақ университеті»

Table of Contents

Session I. Technological Process Automation and Control

Session II. Information Management, Processing and Security

Parallelization of Algorithm of Prediction of miRNA Binding Sites in mRNA on The Cluster Computing Platform

Anna Yu. Pyrkova[1], Anatoli T. Ivashchenko[2], and Olga A. Berillo[2]

[1] Mechanic-mathematical faculty, Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2] Faculty of biology and biotechnology, Al-Farabi Kazakh National University, Almaty, Kazakhstan
{Anna.Pyrkova,Anatoli.Ivashenko}@kaznu.kz,%Devolia18@mail.ru

Abstract. In presented article the solution of the problem of gene scanning for the purpose of prediction of binding sites of miRNA with matrix RNA (mRNA) is proposed. During the research by authors the following results were received: the mathematical model of optimum process of scanning of genes and miRNA sequences is developed; the algorithm of scanning of genes with miRNA with one gap in miRNA and maximum (in a percentage ratio) free energy is developed and analyzed at coincidence of miRNA and a gene site on the basis of complementarity properties; the constructed algorithm of scanning of genes with miRNA is parallelized on the computational cluster with use of MPJ tools - the MirTarget program; the assessment of overall performance of the parallelized algorithm on the cluster computing platform with consecutive algorithm is performed when processing large volumes of data; the developed program was used for performing researches by search of binding sites of miRNA with matrix RNA (mRNA).

Keywords: parallelized algorithm, cluster computing platform, Java MPI, miRNA, mRNA.

1 Introduction

After opening of an important role of microRNA (miRNA) in regulation of an expression of genes the problem of a prediction of binding sites of miRNA with matrixRNA (mRNA) has arisen. Some programs which predicted binding sites of miRNA were created. However many of them had unreasonable restrictions for search of binding sites. Earlier it was claimed that binding sites are localized only in 3'UTR. It was established later that binding sites are localized in 5'UTR and CDS. Other programs were based on identification of binding sites with the obligatory requirement to have complementary interactions of a guanine (G) and an adenine (A) in a site of "seed"which corresponds 5'-end of miRNA. Many such programs predicted a large number of false positive sites and did not allow revealing the binding sites located in 5'UTR and CDS. On this and other reasons it is inexact the beginning of binding sites was established and incorrectly schemes of interaction of miRNA with mRNA were formed. Now, in a genome of the human more than 2500 miRNAs are known and it is necessary for each of them to find target genes among 30 thousand genes of the human. Large volume of calculations demands creation of the program, allowing processing these huge data files. We created the MirTarget program which has no shortcomings given above and with big reliability finds binding sites of miRNA with mRNA.

2 Scanning gene problem definition

Scanning genes [1], [2] is a process of consecutive comparison of sites of a gene with miRNA with possibility of adding one gap in miRNA in positions with the 3rd on n-2-th, where by n – nucleotide number (length) of miRNA. Thus there is an assessment all of possible comparisons on one site of mRNA with miRNA which is defined according to the value of free energy of compared sequences. It is considered the best that option which is closer (in a percentage ratio) on free energy for coincidence of miRNA and a gene site on the basis of a complementarity.

Scanning of a genome allows to reveal hundreds possible targets for therapy of various diseases. Such scanning is important because knowledge the interacting genes will allow to define that, for what this or that protein and, respectively, what intracellular processes answers are broken at this disease.

The mathematical model of a problem of scanning genes can be formulated in the following view:

Let $\{u_l\}, l = \overline{1,N}$ is a set of nucleotide or amino-acid sequences $miRNA$, N is amount of sequences $miRNA$, and $\{v_g\}, g = \overline{1,M}$ is a set of sequences of genes $mRNA$, M is amount of sequences $mRNA$, then $\langle u_l, v_g, Number, Position, Where, Energy, Score, Length\rangle_{l\overline{1,N},g=\overline{1,M}}$ is scanning genes, where $Number$ is order number, $Position$ is a position of $\{u_l\}$ in $\{v_g\}$, $Where$ is an element from a set $\{5'UTR, CDS, 3'UTR\}$, defining site arrangement area $\{u_l\}$ in $\{v_g\}$, $Energy$ is value of free energy on the basis of a complementarity, $Score$ is value of $\triangle G/\triangle Gm$, $Length$ is length of $\{u_l\}$.

3 Algorithm of the program of scanning genes on the cluster

1 step. Verification of attributes in a command line.

```
mpjrun.bat -np
             number\_of\_parallel\_processes
             /path/to/Base
             minimal\_percentage\_of\_convergence
             directory\_with\_genes
             directory\_with\_miRNAs
             file\_name\_of\_result\_mres
             file\_name\_of\_result\_xls
```

For example,

```
mpjrun.bat -np 10 Base 80 gene mir ResultsFull ResultsBrief
```

2 step. The main process (MASTER) invokes the function which forms the geneName array of the files containing genes.

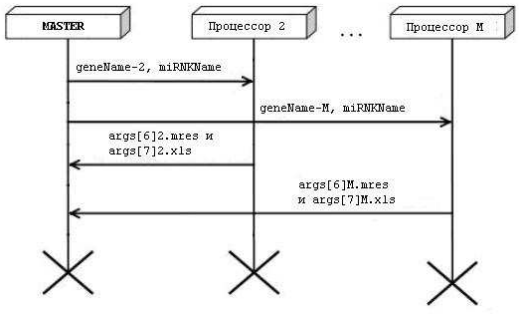

Fig. 1. The protocol of exchanging messages between MPJ processes.

3 step. The main process (MASTER) invokes the function which forms two arrays for miRNAs:

1) miRNAName array of the miRNA names;

2) miRNA array of miRNA sequences. Thus there is a check of the format of the file with miRNA.

```
>let-7a-2-3p              MIRLET7A2-ex-cod  11
CUGUACAGCCUCCUAGCUUUCC
```

Fig. 2. File structure with miRNA (.mir).

4 step. The main and parallel processes create files with the names $args[6]\ \langle task\_id\rangle.mres$ and $args[7]\langle task\_id\rangle.xls$ for saving result.

5 step. The main process (MASTER) divides genes between parallel processes and dispatches a certain part of the geneName and the miRNAName and miRNA arrays. The main process also receives part of genes for scanning.

6 step. Processes read out consistently genes from files of the geneName array and check the file format with gene.

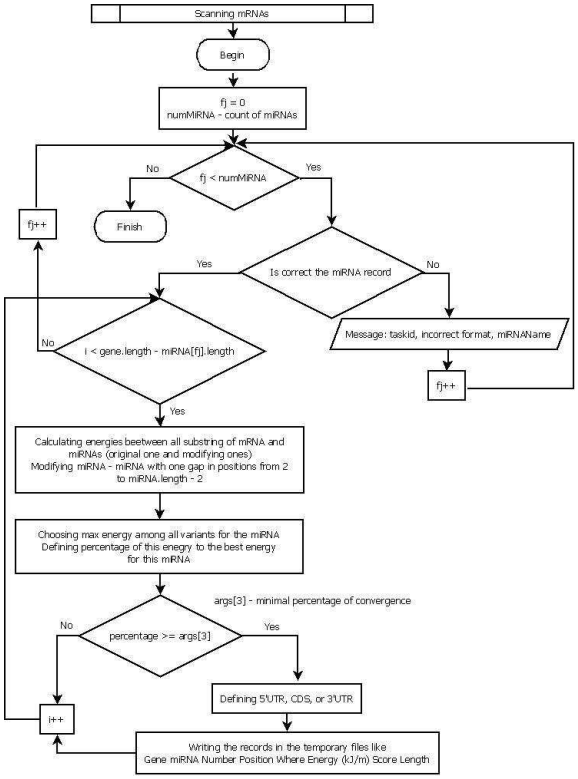

Fig. 3. Flowchart of procedure of scanning mRNAs.

7 step. Processes scan their genes consistently with miRNA, allowing one gap in miRNA in positions from the 3rd to the n-2-th where by n – nucleotide number (length) of miRNA, and choosing that option which is closer (in a percentage ratio) on free energy for coincidence of miRNA and a gene site on the basis of a complementarity.

8 step. Processes form the $args[6]\langle task\_id\rangle.mres$ and $args[7]\langle task\_id\rangle.xls$ files and after completion of processing of their part of genes send messages about completion to the main process.

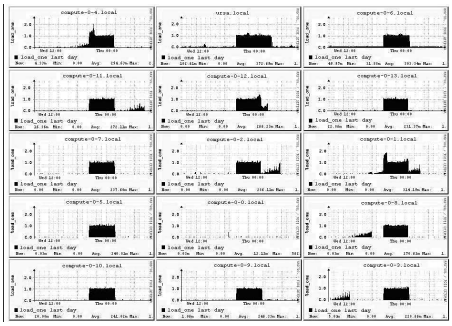

Fig. 4. Load of the cluster (For the database containing 13016 genes and 100 miRNAs operating time is 2 hours 15 minutes 19 seconds).

9 step. The main process (MASTER) receives messages from parallel processes, copies information from the corresponding $args[6]\langle task\_id\rangle.mres$ and $args[7]$ $\langle task\_id\rangle.xls$ files in the $args[6].mres$ and $args[7].xls$ files, and deletes the $args[6]\langle task\_id\rangle.mres$ and $args[7]$ $\langle task\_id\rangle.xls$ files.

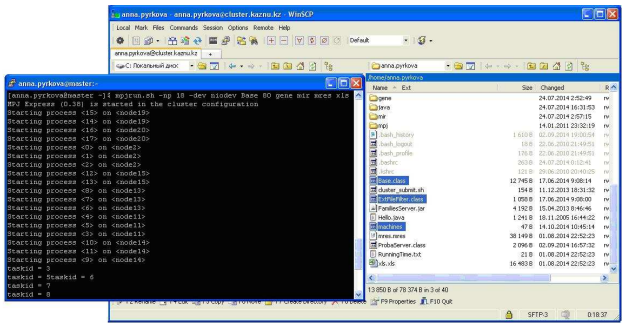

Fig. 5. Running the application on the cluster computing platform cluster.kaznu.kz.

4 Conclusion

The MirTarget program has advantages which are not present in known programs of predicting of binding sites of miRNA with mRNA. In literature there are many data about the value of free energy of hydrogen bond between nucleotides in water solution [3]. However there is a wide spacing of value of free energy of this bond and it is difficult to give preference to certain data [4], [5]. It is important to know the relative relations of free energy of hydrogen bond between nucleotides as they are necessary at formation of RNA of secondary and tertiary structures. The analysis of free energy of the hydrogen bond arising between nucleotides at intramolecular interaction of mRNA at formation of its secondary structure showed that between nucleotides of G-C is formed three, between A-U – two and between G-U and A-S – on one hydrogen bond. The relation of free energy of hydrogen bond in G-C and A-U pairs approximately corresponds to the relation of forces of their interaction 3:2 (0.188 nNewton and 0.125 nNewton) [6]. The value of free energy of one hydrogen bond between nucleotides changes in the range from -0.7 to -1.6 kcal/mol [7]. In the MirTarget program free energy of interaction of nucleotides due to hydrogen communications was considered as equal to 6.368 kJ/mol and 4.246 kJ/mol for G-C and A-U pairs, and 2.123 kJ/mol for G-U and A-S pairs, respectively.

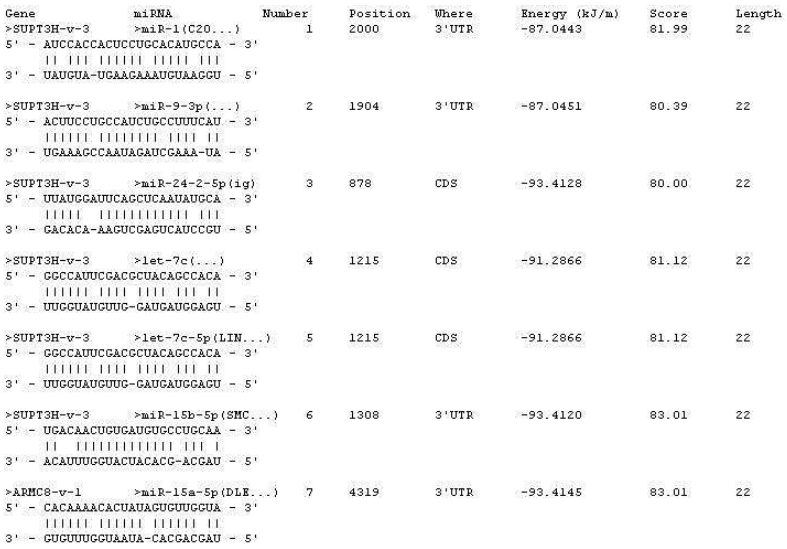

Fig. 6. Result of scanning genes on the cluster (the amount of the processed sequences makes 13016 mRNA and 100 miRNA). Speed of work of the algorithm parallelized on 18 nodes is 66.5 times higher than the speed of work of serial algorithm.

Table 1. Scanning genes (time is specified in seconds), total of genes - 13016

	Duration of processing 100 mRNA and 100 miRNA (average length - 21)	Duration of processing 1000 mRNA and 100 miRNA (average length - 21)	Duration of processing 10000 mRNA and 100 miRNA (average length - 21)
Linear algorithm	3	54	532
Parallelized algorithm (dual-core processor)	1	25	69
Parallelized algorithm (a cluster platform of 15 nodes)	1	3	8

The distance between nucleotides G-C and A-U pairs makes 1.03 nm, between nucleotides G-U pair it is equal 1.02 nm and between nucleotides A-S pair equally 1.04 nm [8]. Therefore, formation of hydrogen bonds between these couples of nucleotides allows two-chained structure of mRNA to have a spiral form similar to DNA regular structure. Such structure of mRNA except hydrogen bonds is stabilized by stacking-interactions between the nitrogenous bases [9]. Between nucleotides of couples purine-purine and pyrimidine-pyrimidine distances significantly differ from 1.03 nanometers: distances of A-A, G-A and G-G are equal to 1.23 nm, 1.25 nm and 1.25 nm, respectively. In couples of pyrimidine-pyrimidine distances between nucleotides too significantly differ from 1,03 nanometers: distances of C-C, U-U and U-C are equal to 0.85 nm, 0.81 nm and 1.18 nm, respectively. Therefore, in such couples hydrogen bonds are not formed, and these couples will break regular structure of two-chained miRNA with mRNA reducing stability of

the RISC complex (RNA-induced silencing complex). Therefore, in the program such couples of hydrogen bonds were not considered.

At alignment of the nucleotide sequences of miRNA with mRNA we assume existence of only one admission on miRNA (lack of complementary couple of hydrogen bond) that allows considering binding sites of mRNA longer miRNA on one nucleotide. In this case the regular structure of a spiral is broken and there is its bulg. Free energy of binding miRNA with mRNA of such structure is less, than in alternative case. The program determines free energy of hybridization ($\triangle G$, 100 kJ/mole) of miRNA with mRNA and the scheme of their interactions, calculation of the relations $\triangle G/\triangle Gm$, levels of reliability (p) and the mRNA areas, where the site (5'UTR, CDS or 3'UTR), since the first nucleotide 5'UTR, is located. $\triangle Gm$ is equal to free energy of binding miRNA with completely complementary to it a site of nucleotide sequence of miRNA. Level of reliability (p) was defined on the basis of value $\triangle G$ and its standard deviation. The program outputs the scheme of interaction of miRNA with mRNA, a site position in 5'UTR, CDS or 3'UTR, free energy of interaction of miRNA with mRNA, and its relative value from the maximum energy of binding miRNA. In the program the threshold value of this relation is set, this value allows not considering sites with weak free energy of binding.

References

1. Lesk Arthur M. Introduction to Bioinformatics. - Oxford: Oxford University Press, 2002. - 255 p.
2. Jones Neil C., Pevzner Pavel A. An Introduction to Bioinformatics Algorithms. - Massachusetts: Massachusetts Institute of Technology Press, 2004. - 435 p.
3. Guckian K.M, Schweitzer B.A, Ren RXF, Sheils C.J, Paris P.L, et al Experimental measurement of aromatic stacking in the contex to duplex DNA. // J.Am.Chem.Soc. - 1996. - 118: 8182–83.
4. Turner D.H, Sugimoto N, Kierzek R, Dreiker S.D. Free energy increments for hydrogen-bonds in nucleic acid base pairs. // J. Am. Chem. Soc. - 1987. - 109: 3783–85.
5. Sugimoto N, Kierzek R, Turner D.H. Sequence dependence for the energetics of dangling ends and terminal base pairs in ribonucleic acid. // Biochemistry. - 1987. - 14: 4554–58.
6. Boland T., Ratner B.D. Direct measurement of hydrogen bonding in DNA nucleotide bases by atomic force microscopy. // Proc. Natl. Acad. Sci. USA. - 1995. - V. 92. - Pp. 5297-5301.
7. Kool E.T. Hydrogen bonding, base stacking, and steric effects in DNA replication. // Annu. Rev. Biophys. Biomol. Struct. - 2001. - 30: 1–22.
8. Leontis N.B., Stombaugh J., Westhof E. The non-Watson–Crick base pairs and their associated isostericity matrices. // Nucleic Acids Res. - 2002. - 30(16): 3497–3531.
9. Richard A. Friedman, Honig B.A. Free Energy Analysis of Nucleic Acid Base Stacking in Aqueous Solution. // Biophysical Journal. - 1995. - V. 69. - Pp. 1528-1535.