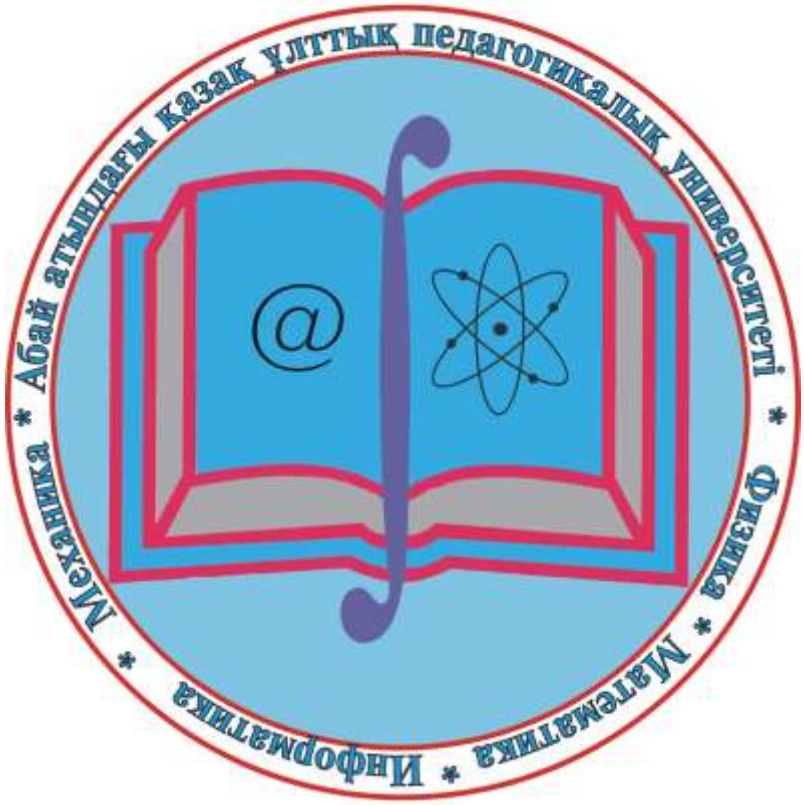
ISSN 1728-7901

Абай атындағы
Қазақ ұлттық педагогикалық университеті

Казахский национальный педагогический
университет имени Абая

ХАБАРШЫ
ВЕСТНИК



Серия «Физико-математические науки» • «Физика-математика ғылымдары» сериясы

№ 3 (51)
2015

Алматы

ИНФОРМАТИКА. ИНФОРМАТИКАНЫ ОҚЫТУ ӘДІСТЕМЕСІ. БІЛІМ БЕРУДІ АҚПАРАТТАНДЫРУ
ИНФОРМАТИКА, МЕТОДИКА ПРЕПОДАВАНИЯ ИНФОРМАТИКИ. ИНФОРМАТИЗАЦИЯ ОБРАЗОВАНИЯ

К.А. Айдаров, Г.Т. Балакаева Обеспечение требовании QoS в веб-серверных фермах 188
К.С. Апаев, М.Е. Мансурова, А.Б. Нугуманова Визуальное представление больших коллекций текстов с помощью тематического моделирования 194
А.Т. Бектемесов Ақпараттарды үлкен көлемді құжаттардан BDD әдісімен іздеу 201
Е.Ы. Бидайбеков, Н.И. Пак Академическая мобильность и международная деятельность в педагогической магистратуре по программе «Информатика в образовании / информатизация образования» 205
Е.Ы. Бидайбеков, Б.Ғ. Бостанов, Қ.Ү. Үмбетбаев Әл-Фарабидің салу есептерін заманауи математикалық білім беруде ақпараттық технологияларды пайдаланып оқытудың өзіндік ерекшеліктері 210
С.Н. Боранбаев, А.Б. Нурбеков Создание информационной системы для моделирования функционирования обрабатывающих отраслей промышленности Республики Казахстан 214
В.В. Гриншкун Об участии и роли педагогов в совместном создании образовательных электронных изданий и ресурсов 221
Ф.Р. Гусманова, М.А. Скиба, А.Т. Турганбаева Білім берудегі басқарылатын шешім: маңыздылығы, жіктелуі және қабылдау технологиясы 225
Е.А. Дадыкина, І. Қашқымбай, М.Е. Мансурова Кодирование RDF словарей с применением технологии MapReduce Hadoop 229
С.Н. Конева, Д.М. Амирканова Организация виртуальных лабораторных работ по компьютерным сетям в среде Microsoft Visio 235
Б.А. Нупбаев, А.А. Аманбаев Исследования методов оптимизации SQL-запросов 241
А.С. Омарбекова, А.Б. Закирова, А.О. Сейфуллина Реализация рефлексии в интеллектуальной обучающей системе посредством моделирования поведения студента и построения эффективной обратной связи 248
Anna Yu. Pyrkova, Anatoly T. Ivashchenko, Olga A. Berillo Development and implementation of parallelization algorithm for prediction of miRNA binding sites in mRNA. mirtarget program ... 254
А.Е. Сағымбаева, Ә.Е. Жақсылықов Білімді бақылауды компьютерлік көпдеңгейлі ақпараттық модель негізінде іске асырудың қажеттілігі 259
А.Е. Сағымбаева, А.С. Назарбекова Информатиканы жобалар әдісін қолданып оқытудың ерекшеліктері 264
С.М. Сарсимбаева, Б.Б. Камаш Анализ и использование инструментов в программировании игровых программ для операционной системы Android 269
М.А. Скиба, А.Р. Турганбаева Нечеткая модель оценки качества результатов образовательного процесса 273
Ш.Т. Шекербекова Информатиканы дамыта оқыту әдістері . 279

UDC 577.29

Anna Yu. Pyrkova, Anatoly T. Ivashchenko, Olga A. Berillo

DEVELOPMENT AND IMPLEMENTATION OF PARALLELIZATION ALGORITHM FOR PREDICTION OF MIRNA BINDING SITES IN MRNA. MIRTARGET PROGRAM

(Al-Farabi Kazakh National University)

Abstract: *Today in the human genome of the person more than 2500 miRNA are known, it is necessary for each miRNA to find target genes among 30 thousands genes of the human. Large amount of calculations demands creation of the program, allowing processing these huge data files. In presented article the solution of the problem of gene scanning for the purpose of prediction of sites of binding of miRNA with matrix RNA (mRNA) is proposed. During the conducted research by authors the following results were received:*

- the mathematical model of optimum process of scanning of genes and miRNA sequences is constructed;

- the constructed algorithm of scanning of genes with miRNA is parallelized on the cluster platform with use of MPJ tools (Java MPI);

- the developed program was used for performing researches by search of sites of binding of miRNA from matrix RNA (mRNA).

Keywords: *miRNA and mRNA, mathematical model, algorithm, Java MPI, the cluster platform, the parallelized algorithm, scanning, complementarity.*

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

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.*

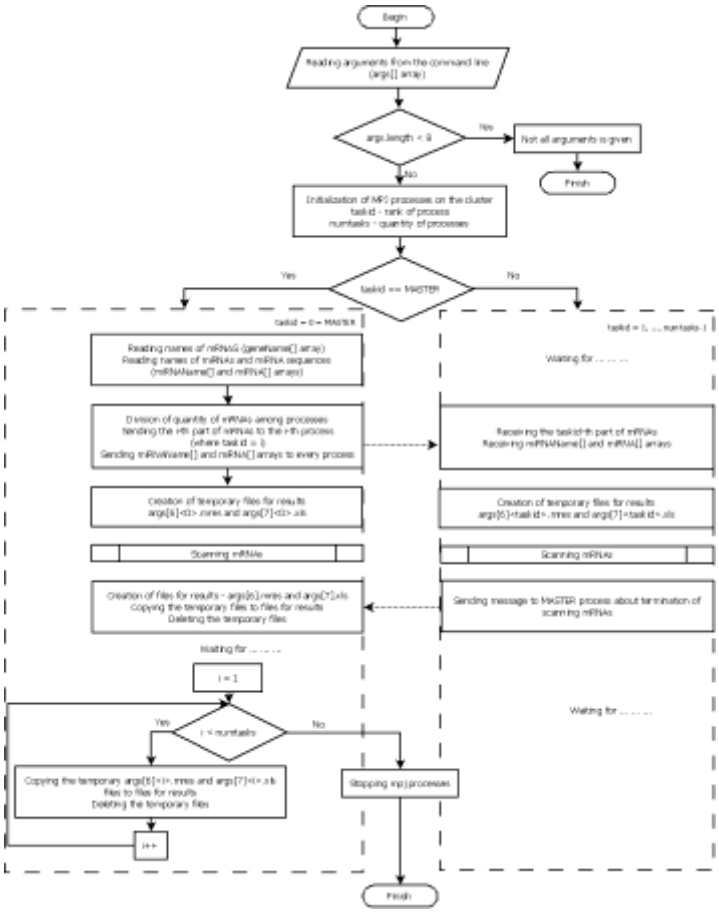


Fig. 1. Flowchart of the scanning program (parallel algorithm).

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]<task_id>.mres and args[7]<task_id>.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.*

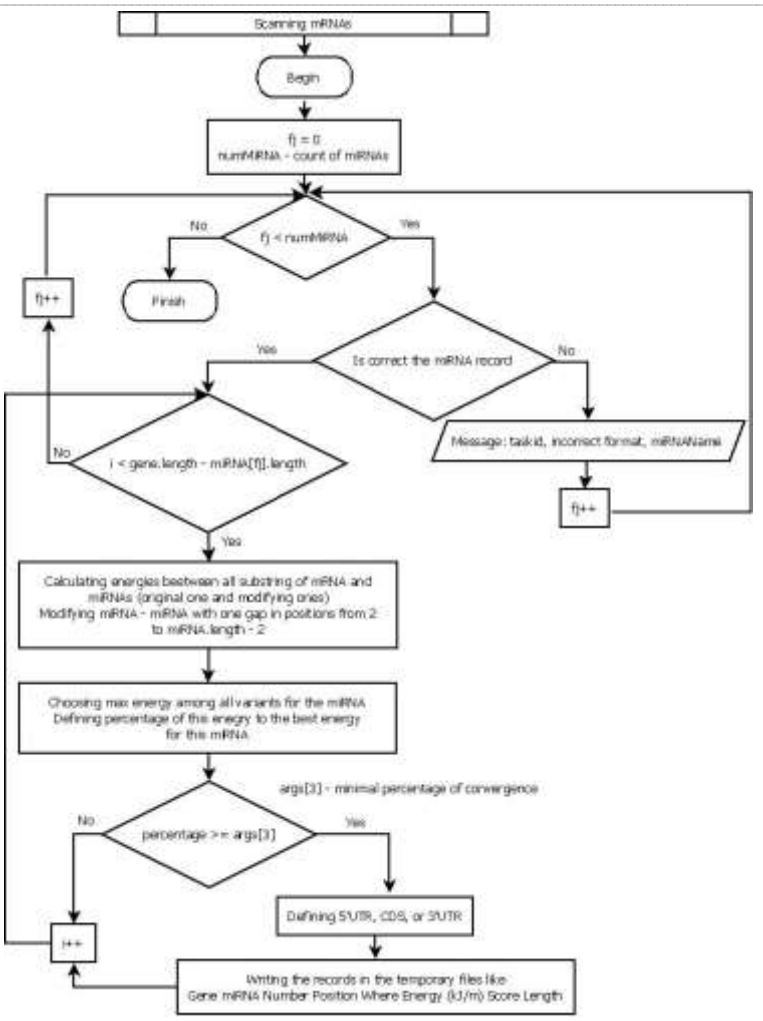


Fig. 3. Flowchart of procedure of scanning mRNAs.

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.*

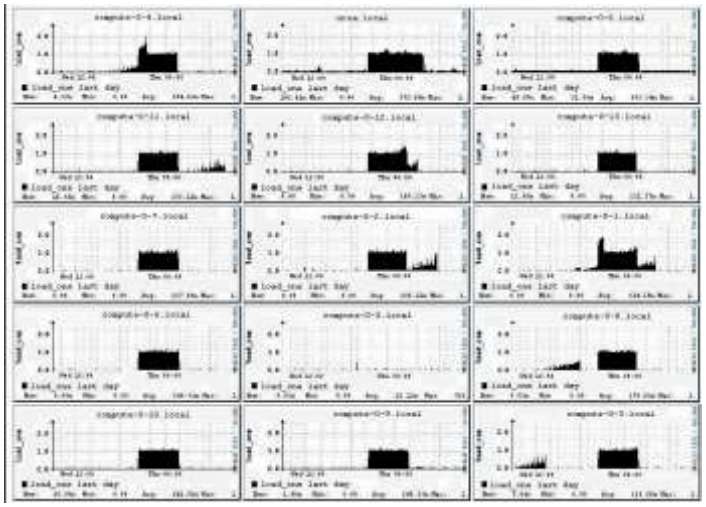


Fig. 4. Load of the cluster (For the database containing 13016 genes and 100 miRNAs operating time

is 2 hours 15 minutes 19 seconds).

8 step. *Processes form the args[6]<task_id>.mres and args[7]<task_id>.xls files and after completion of processing of their part of genes send messages about completion to the main process.*

9 step. *The main process (MASTER) receives messages from parallel processes, copies information from the corresponding args[6]<task_id>.mres and args[7]<task_id>.xls files in the args[6].mres and args[7].xls files, and deletes the args[6]<task_id>.mres and args[7]<task_id>.xls files.*

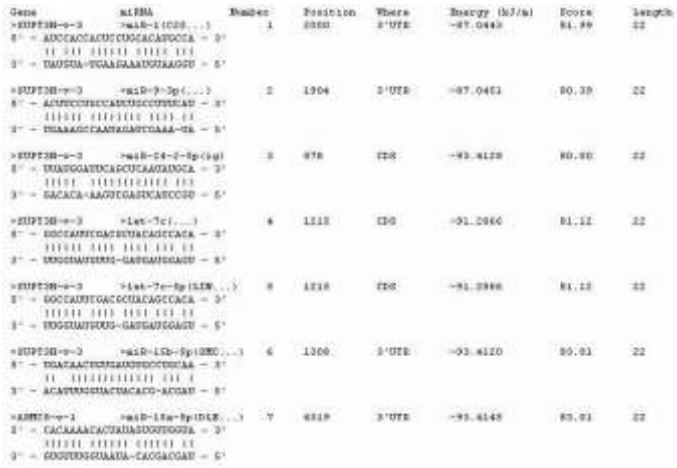


Fig. 5. Result of scanning genes on the cluster (the amount of the processed sequences makes 13016 mRNA and 100 miRNA). Speed of work of the algorithm parallelized on 18 nodes is 66.5 times higher than the speed of work of serial algorithm.

Comparative analysis of linear and parallel algorithms of scanning genes

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

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.

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 (ΔG, 100 kJ/mole) of miRNA with mRNA and the scheme of their interactions, calculation of the relations ΔG/ΔGm, 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. ΔGm 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 ΔG 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.

1. Lesk, A.M.: Introduction to Bioinformatics. Oxford University Press, Oxford (2002)
2. Jones, N.C., Pevzner, P.A.: An Introduction to Bioinformatics Algorithms. Massachusetts Institute of Technology Press, Massachusetts (2004)
3. Guckian, K.M., Schweitzer, B.A., Ren, R.F., Sheils, C.J., Paris, P.L.: Experimental measurement of aromatic stacking in the contex to duplex DNA, Vol. 118. J.Am.Chem.Soc. (1996) 8182–83
4. Turner, D.H., Sugimoto, N., Kierzek, R., Dreiker, S.D.: Free energy increments for hydrogen-bonds in nucleic acid base pairs, Vol. 109. J. Am. Chem. Soc. (1987) 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, Vol. 14. Biochemistry (1987) 4554–58
6. Boland, T., Ratner, B.D.: Direct measurement of hydrogen bonding in DNA nucleotide bases by atomic force microscopy, Vol.92. Proc. Natl. Acad. Sci., USA (1995) 5297-5301
7. Kool, E.T.: Hydrogen bonding, base stacking, and steric effects in DNA replication, Vol. 30. Annu. Rev. Biophys. Biomol. Struct. (2001) 1–22

8. Leontis, N.B., Stombaugh, J., Westhof, E.: The non-Watson–Crick base pairs and their associated isostericity matrices, Vol. 30(16). Nucleic Acids Res. (2002) 3497–3531
9. Richard, A. Friedman, Honig, B.A.: Free Energy Analysis of Nucleic Acid Base Stacking in Aqueous Solution, Vol. 69. Biophysical Journal (1995) 1528-1535

Аңдатпа. *Осы уақытта адам геномында 2500 артық miRNA белгілі, 30 мың адам геномының арасынан әрбір miRNA үшін нысана генін табу қажет. Есептеулердің үлкен көлемі, ауқымды деректер массивін өңдеуге рұқсат ететін программа құруды талап етеді. Ұсынылған мақалада РНК (mRNA) матрицасымен miRNA байланысын гендерді сканерлеу есебінің шешімі сайттарды болжау мақсатымен ұсынылған. Жұмыс барысында жүргізілген зерттеулер нәтижесінде авторлар төмендегідей нәтижелерге қол жекізді:*

- miRNA тізбегі мен гендерін сканерлеу процессі кезінде тиімді математикалық пішін құрылды;

- MPJ (Java MPI) құралдарын қолдану арқылы кластерлі платформада параллелденген miRNA гендерін сканерлеу алгоритмі құрылған;

- құрылған программа РНК (mRNA) матрицасы miRNA байланыс сайттарын зерттеу жүргізу үшін іздеу кезінде қолданылған.

Түйін сөздер: *miRNA және mRNA, математикалық пішін, алгоритм, Java MPI, кластерлі платформа, параллелденген алгоритм, сканерлеу, комплементарлық.*

Аннотация: *В настоящее время в геноме человека известно более 2500 miRNA, необходимо для каждой miRNA найти гены мишени среди 30 тысяч генов человека. Большой объём вычислений требует создания программы, позволяющей обрабатывать эти огромные массивы данных. В представленной статье предложено решение задачи сканирования генов с целью предсказания сайтов связывания miRNA с матричной РНК (mRNA). В ходе проведённого исследования авторами были получены следующие результаты:*

• построена математическая модель оптимального процесса сканирования генов и последовательностей miRNA;

• построенный алгоритм сканирования генов с miRNA распараллелен на кластерной платформе с использованием средств MPJ (Java MPI);

• разработанная программа была использована для проведения исследований при поиске сайтов связывания miRNA с матричной РНК (mRNA).

Ключевые слова: *miRNA и mRNA, математическая модель, алгоритм, Java MPI, кластерная платформа, распараллеленный алгоритм, сканирование, комплементарность.*

ӘОЖ 37.014.6:004.414.23(574)

А.Е. Сағымбаева, Ә.Е. Жақсылықов

БІЛІМДІ БАҚЫЛАУДЫ КОМПЬЮТЕРЛІК КӨПДЕҢГЕЙЛІ АҚПАРАТТЫҚ МОДЕЛЬ НЕГІЗІНДЕ ІСКЕ АСЫРУДЫҢ ҚАЖЕТТІЛІГІ

(Алматы қ, Абай атындағы Қазақ ұлттық педагогикалық университеті)

Аңдатпа. *Мақалада білімді бақылауда компьютерлік тестілеу жүйесін қолданудың мүмкіндіктері қарастырылған. Бейімделген тестілеудің көптеген нұсқалары сипатталған. Ақпараттық модель ұғымына анықтама беріліп, оның білімді бақылаудағы қолданылуы негізделген. Білімді бақылаудың көпдеңгейлі ақпараттық моделінің компьютерлік моделін жасау қажеттілігі анықталған.*

Түйін сөздер: *бақылау, бейімделген бақылау, деңгей, модель, көпдеңгейлі ақпараттық модель, компьютерлік модель*