УДК 577.29

А.Ю. Пыркова

Механико-математический факультет, Казахский национальный университет им. аль-Фараби, Алматы, Казахстан; e-mail: Anna.Pyrkova@kaznu.kz

Моделирование ресурсоёмких задач в области биоинформатики *

В представленной статье рассмотрены задачи множественного выравнивания нуклеотидных последовательностей и построения дендограмм. В ходе проведённого исследования автором были получены следующие результаты:

- разработана математическая модель множественного выравнивания нуклеотидных и аминокислотных последовательностей;
- разработан и проанализирован алгоритм множественного выравнивания, построенный на основе алгоритма Нидлмана-Вунша, который был модифицирован для обработки больших массивов данных с помощью распараллеливания процесса обработки средствами MPJ (Java MPI);
- разработан алгоритм построения дендограмм, представляющий собой модификацию алгоритмов UPGMA (Unweighted Pair Group Method with Arithmetic Mean) и NJ (Neighbour Joining) с возможностью распараллеливания обработки данных;
- выполнена программная реализация алгоритма множественного выравнивания и построения дендограмм на языке Java с использованием средств MPI;
- результаты работы программы были протестированы на данных о нуклеотидных последовательностях, предоставленных сотрудниками кафедры биотехнологии КазНУ имени аль-Фараби.

Ключевые слова: математическая модель, алгоритм, Java MPI, выравнивание, нуклеотидные и аминокислотные последовательности, дендограммы.

А.Ю. Пыркова

Биоинформатика саласындағы көпресурсты есептерді пішіндеу

Бұл мақалада нуклеотид тізбектерінің көптік теңестіруі есептері мен дендограмма құру қарастырылған. Зерттеу жұмыстарын жүргізу барысында автор төмендегідей нәтижелерге қол жеткізген:

- нуклеотид тізбектерінің көптік теңестіруі мен аминқышқылды тізбектердің математикалық пішіні өңделген;

*Работа выполнена при поддержке грантового финансирования научно-технических программ и проектов Комитетом науки МОН РК, грант № 1619/ГФ, 2012г.-2014г.

- MPJ (Java MPI) құралдарымен үлкен массивті деректерді параллельдеу көмегімен процессті өңдеу үшін модификацияланған, Нидлман-Вунш алгоритмі негізінде құрылған, көптік теңестіру алгоритмі өңделіп және талқыланған;
- деректерді өңдеуде параллельдеу мүмкіндігі мен UPGMA (Unweighted Pair Group Method with Arithmetic Mean) және NJ (Neighbour Joining) алгоритмдерін модификациялауды ұсынатын, дендограмма құратын алгоритм өңделген;
- MPI құралдарын қолданып Java тілінде дендограмма құру және көптік теңестіру алгоритмін программалық іске асыру орындалған;
- программа жұмыстарының нәтижесі әл-Фараби атындағы ҚазҰУ-нің биотехнология кафедрасының қызметкерлері ұсынған, нуклеотид тізбектері жөніндегі деректерде тестіленген.

A.Yu. Pyrkova

Modelling of resource-intensive problems in the field of bioinformatics

In presented article the problems of multiple alignment of nucleotide sequences and dendrogram construction are considered. During the conducted research by the author the following results were received:

- the mathematical model of multiple alignment of nucleotide and amino-acid sequences is developed;
- the algorithm of multiple alignment, constructed on the basis of algorithm of Needleman-Wunsch which was modified for processing of big data files with help of parallelization of treatment process by means of MPJ (Java MPI), is developed and analyzed;
- the algorithm of dendrogram construction, representing modification of algorithms of UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and NJ (Neighbour Joining) with possibility of parallelization of data processing, is developed;
- program realization of algorithm of multiple alignment and dendrogram construction in the Java language with use of means of MPI is executed;
- results of work of the program were tested on data on the nucleotide sequences provided by staff of the biotechnology department of Kazakh NU named al-Farabi.

Разработка математической модели множественного выравнивания нуклеотидных и аминокислотных последовательностей

Выравнивание аминокислотных или нуклеотидных последовательностей /1/ - это процесс сопоставления сравниваемых последовательностей для такого их взаиморасположения, при котором наблюдается максимальное количество совпадений аминокислотных остатков или нуклеотидов. Различают два вида выравнивания: парное (выравнивание двух последовательностей ДНК, РНК или белков) и множественное (выравнивание трёх или более последовательностей).

Парное выравнивание является базовым методом сравнения биологических последовательностей /2/. Выровнять две последовательности - это поместить их друг над другом, возможно, вставляя в обе последовательности пробелы, так, чтобы сделать их длины равными. При этом позиции, оказавшиеся друг над другом, считаются сопоставленными друг другу, а остальные символы (расположенные напротив пробелов) - удаленными. Две данные символьные последовательности можно выровнять многими способами. Алгоритмический выбор нужного выравнивания двух данных последовательностей основан на понятии веса выравнивания - строится оптимальное выравнивание, т.е. выравнивание, имеющее максимально возможный вес. Вес выравнивания определяется как сумма весов сопоставленных символов минус сумма штрафов за удаленные фрагменты.

Математическую модель задачи множественного выравнивания можно сформулировать в следующем виде:

Пусть $\{u_l\}, l = \overline{1,N}$ - набор нуклеотидных или аминокислотных последовательностей, которые нужно выровнять, N - количество нуклеотидных или аминокислотных последовательностей, тогда $\langle u_l, u_k, S\rangle_{l,k=\overline{1,N},k\neq l}$ - множественное выравнивание последовательностей, где $S = \{\langle i_1, j_1\rangle, ..., \langle i_n, j_n\rangle\}$ - набор пар позиций в последовательностях u_l и u_k соответственно, таких, что $1 \leq i_1 < ... < i_n \leq |u_l|$ и $1 \leq j_1 < ... < j_n \leq |u_k|$, т.е. i_m позиция последовательности u_l сопоставлена с j_m позицией последовательности u_k, $m = \overline{1,r}$, r - длина полученной выровненной последовательности (после выравнивания все последовательности имеют одинаковые длины), а фрагменты вида $u_l[i_m+1, i_{m+1}-1]$ и $u_k[j_m + 1, j_{m+1} - 1]$ заполнены пробелами, $k = \overline{0,r}$; $i_0 = j_0 = 0$; $i_{r+1} = |u_l| + 1$; $j_{r+1} = |u_k| + 1$. Вес выравнивания определяется как функция весов сопоставлений букв и штрафов за добавление пробелов.

В качестве весовых функций была выбрана кусочно-линейная система весовых функций /2/. Кусочно-линейная система весовых функций - это система, которая обладает следующими свойствами:

- штрафы за добавления пробелов на концах последовательностей могут быть произвольными;
- штраф за добавление пробела может зависеть от граничных позиций добавляемого пробела;
- зависимость штрафа от длины фрагмента может задаваться произвольной кусочно-линейной функцией;
- штрафы за добавление пробелов в каждой из сравниваемых последовательностей могут задаваться по-своему;
- вес сопоставления символов $u_l[i]$ и $u_k[j]$ задаётся произвольной функцией $\eta(i, j, u_l, u_k)$.

Вес выравнивания последовательностей определяется как разность $V - D$, где V - сумма весов сопоставлений букв, D - сумма штрафов за удаление фрагментов.

Одним из недостатков алгоритмов динамического программирования /2/, предназначенных для выравнивания нуклеотидных последовательностей, является их относительно невысокое быстродействие. Даже на современных компьютерах невозможно

за приемлемое время выровнять последовательности длиной более миллиона символов (как при сравнении геномов) или провести сотни тысяч сравнений последовательностей длиной несколько тысяч.

Разработка алгоритма множественного выравнивания

На сегодняшний день известно немало биоинформационных программ, занимающихся поиском родственных последовательностей в базе данных нуклеотидных и аминокислотных последовательностей; множественным выравниванием нуклеотидных и аминокислотных последовательностей; редактированием филогенетических деревьев; филогенетическим анализом; таких как BLAST, ClustalW, ClustalX, UGENE и многие другие /3/. Главная проблема, возникающая при обработке больших массивов данных - это, прежде всего, нехватка вычислительных средств. В данном исследовании предпринята попытка разработать программное приложение для выравнивания нуклеотидных последовательностей и построения филогенетических деревьев путём применения алгоритма выравнивания Нидлмана-Вунша и алгоритмов NJ и UPGMA для построения филогенетических деревьев с использованием возможностей MPJ при построении множественного выравнивания и дендограмм.

В результате проведённого исследования /4, 5/ был разработан распараллеленный алгоритм множественного выравнивания нуклеотидных и аминокислотных последовательностей, где каждая пара выравнивается с использованием алгоритма Нидлмана-Вунша.

Алгоритм Нидлмана-Вунша /3, 6/ - это алгоритм динамического программирования для выполнения выравнивания двух последовательностей A и B. В этом алгоритме соответствие выровненных символов задается матрицей их похожести друг на друга и используется линейный штраф за разрыв d. Для нахождения выравнивания с наивысшей оценкой назначается двумерная матрица F, содержащая столько же строк, сколько символов в первой последовательности, и столько же столбцов, сколько символов во второй последовательности.

В процессе работы алгоритма величина F_{ij} будет принимать значения оптимальной оценки для выравнивания первых $i = 0, ..., n$ символов в первой последовательности и первых $j = 0, ..., m$ символов во второй последовательности. Используемый в этом алгоритме принцип оптимальности Беллмана формулируется следующим образом:

$$\begin{cases} F_{0j} = d * j, \\ F_{i0} = d * i, \\ F_{ij} = max(F_{i-1,j-1} + S(A_j, B_j, F_{i,j-1} + d, F_{i-1,j} + d). \end{cases}$$

где S - матрица похожести, а $S(A_k, B_l)$ - величина, определяющая похожесть k элемента первой последовательности A и l элемента второй последовательности B.

Когда матрица F рассчитана, её элемент $F_{i,j}$ дает максимальную оценку среди всех возможных выравниваний. Для вычисления самого выравнивания, которое получило такую оценку, нужно начать с правой нижней клетки матрицы F и сравнивать значения в ней с тремя возможными источниками (соответствие, вставка или делеция), чтобы увидеть, откуда оно появилось. В случае соответствия A_i и B_j выровнены, в случае делеции A_i выровнено с разрывом, а в случае вставки с разрывом выровнено уже B_j. Алгоритм для линейных штрафов имеет время работы $O(m * n)$; алгоритмы построения оптимального выравнивания для выпуклых весов делеций и для произвольных весов

делеций имеют временную сложность соответственно $O(m*n*(m+n))$ и $O(m^2*n^2)$ (m и n - длины последовательностей).

Среди современных инструментов построения множественного выравнивания наибольшей популярностью пользуются программы ClustalW, Muscle, T-Coffee, самый точный из которых T-Coffee, но, по сравнению с другими, существенно медленнее. Подход, который используется практическими всеми программами множественного выравнивания, состоит в том, что они пытаются найти лучшее выравнивание методом последовательных попарных выравниваний. Недостаток метода последовательных попарных выравниваний в том, что производимое выравнивание не гарантирует достижения оптимального выравнивания. Поэтому дополнительно предлагаются программы для редактирования результатов множественного выравнивания. Эти программы применяются для подготовки отчетов по выравниванию к публикации, а также ручного редактирования полученных автоматически программами результатов.

В данном исследовании /4, 5/ для достижения более точного результата множественного выравнивания и оптимизации временных затрат, требуемых для обработки данных при множественном выравнивании, набор последовательностей разбивается на M самостоятельных групп, обрабатываемых M параллельными процессами, каждый из которых будет независимо от других выполнять выравнивание своей группы, а само выравнивание происходит не попарно, а по суммарным последовательностям группы.

Пусть в группе $\{u_k\}$ n_k последовательностей, $k = \overline{1, M}$, где M - число процессов. В начале с помощью алгоритма Нидлмана-Вунша выравниваются первые две последовательности, затем по полученному выравниванию строится некоторая суммарная последовательность, полученная путём слияния двух ранее выровненных последовательностей. Например, если выравнивались две последовательности:

miR-1203 CCCGGAGCCAGGAUGCAGCUC

и

miR-1224-3p CCCCACCUCCUCUCUCCUCAG,

после выравнивания которых были получены следующие последовательности:

CCCGG-AGCCAGGAUGCAG- - - - - - -CUC- -

CCC- -CA-CC- - - -U-C- -CUCUCUCCUCAG,

то суммарной последовательностью будет последовательность:

CCCGGCAGCCAGGAUGCAGCUCUCUCCUCAG,

которая и будет выравниваться с третьей последовательностью в группе, в случае же внесения разрыва при выравнивании в суммарную последовательность, этот разрыв будет добавляться в каждую из ранее выровненных последовательностей, слиянием которых была получена суммарная, участвующая в текущем выравнивании. Процесс продолжается до тех пор, пока не будет выровнены n_k последовательностей в группе $\{u_k\}$, $k = \overline{1, M}$, где M - число процессов.

Процесс выравнивания происходит одновременно во всех группах, параллельность обработки которых достигается за счёт в первую очередь независимости обрабатываемых данных и, конечно же, с технической точки зрения с помощью средств MPJ. Когда каждый из процессов заканчивает выравнивание своей группы, он формирует свою суммарную последовательность, полученную слиянием всех выровненных последовательностей в группе, и, наконец, главный процесс программного приложения завершает выравнивание выравниваем этих суммарных последовательностей.

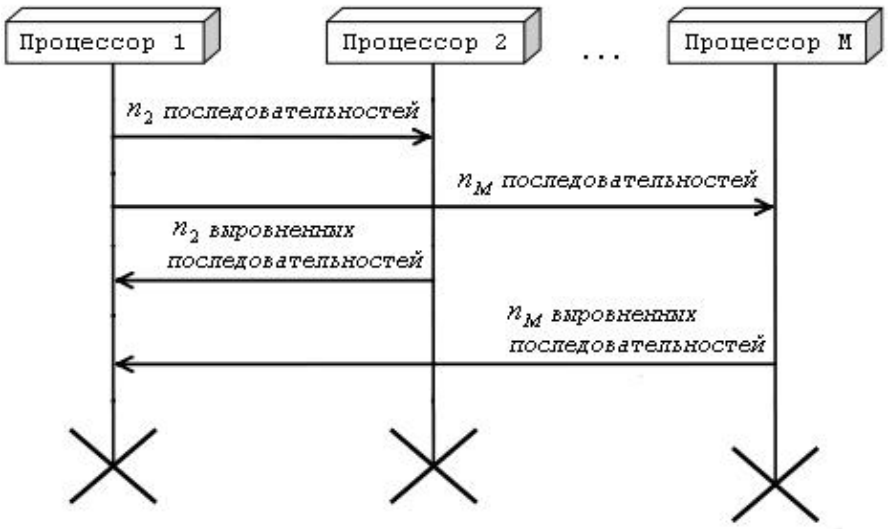

Рисунок 1. Протокол обмена сообщениями между MPJ-процессами

Таким образом, алгоритм /4, 5/ можно представить следующим образом:

Шаг 1. Разбиение набора нуклеотидных или аминокислотных последовательностей $u_l, l = \overline{1, N}$, которые нужно выровнять, на M равномерных групп: $u_1, ..., u_{l_2-1}, u_{l_2}, ..., u_{l_3-1}, \ldots\ u_{l_M}, ..., u_N$, N - количество последовательностей.

Шаг 2. Обработка каждой из M групп последовательностей соответствующим процессом с помощью модифицированного алгоритма Нидлмана-Вунша (как было предложено выше). Построение суммарной последовательности для каждой группы $u_k, ..., u_{l_{k+1}-1}$ - $\overline{u_k}$, $k = \overline{1, M}$, где M - число процессов.

Шаг 3. Отправка каждым из M процессов соответствующей группы выровненных последовательностей $u_k, ..., u_{l_{k+1}-1}$ и суммарной последовательности $\overline{u_k}$ первому процессу.

Шаг 4. Выравнивание первым процессом группы суммарных последовательностей $\overline{u_k}$, $k = \overline{1, M}$ модифицированным алгоритмом Нидлмана-Вунша с одновременным внесением необходимых разрывов в уже выровненные M процессами группы $u_1, ..., u_{l_2-1}$, $u_{l_2}, ..., u_{l_3-1}, \ldots\ u_{l_M}, ..., u_N$, N - количество последовательностей.

Шаг 5. Построение дендограммы на основе полученного выравнивания алгоритмом NJ или UPGMA. Для последовательностей длины m_1 и m_2 время работы такого алгоритма оценивается как $O(c(m_1, m_2) * m_1 * m_2)$, где коэффициент $c(m_1, m_2)$ зависит от выбранной весовой функции и определяется временем выполнения операций делеции и вставки для этой функции.

Разработка алгоритма построения филогенетических деревьев

Алгоритмов матричного построения деревьев, таких как UPGMA, WPGMA, NNM, FNM, UPGMC, WPGMC и другие /7/, известно не мало. Работа этих матричных методов построена по одному принципу, основанному на итеративной обработке матрицы расстояний между таксонами. На каждом шаге в матрице расстояний D ищется минимальный элемент D_{ij}. Найденные таксоны i и j объединяются, образуя новый таксон k. Строки и столбцы, соответствующие таксонам i и j, выбрасываются из матрицы D и добавляется новая строка и новый столбец, соответствующие таксону k. В результате матрица сокращается на одну строку и один столбец. Эта процедура повторяется до

тех пор, пока не будут объединены все таксоны. Разные матричные методы отличаются лишь способом вычисления расстояний от вновь образуемого на каждом шаге таксона k до всех оставшихся таксонов. В общей форме расстояние между таксоном k и любым другим таксоном l для основного большинства методов можно записать с помощью формулы:

$D\{lkl\}l = a\{li\}lD\{lil\}l + a\{lj\}lD\{ljl\}l + b\{lj\}lD\{lij\}l + g * D\{lil\}l - D\{ljl\}l$

где коэффициенты a_i, a_j, b_j, g - различны для разных методов.

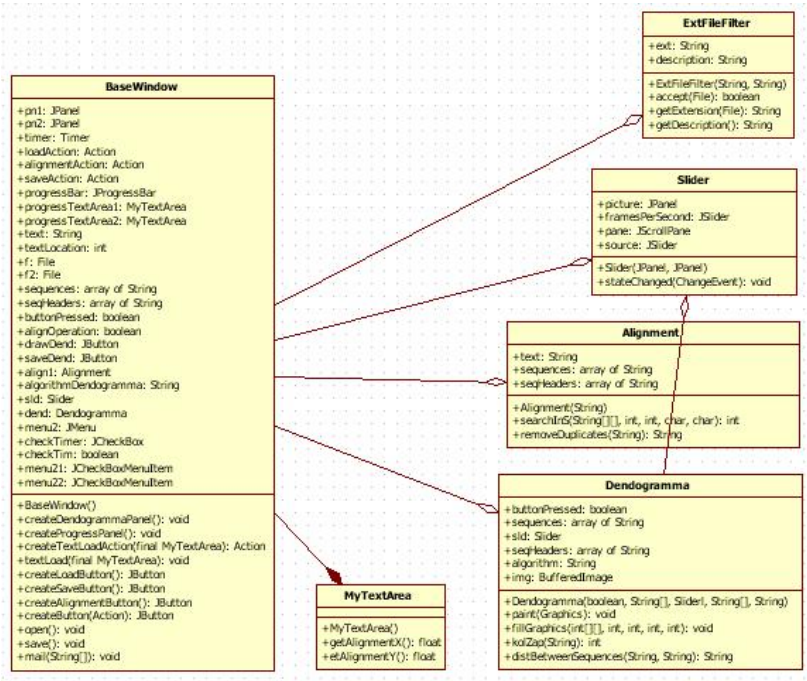

Рисунок 2. UML диаграмма классов программного приложения множественного выравнивания нуклеотидных и аминокислотных последовательностей и построения дендограмм

В ходе проведённого исследования был предложен метод построения филогенетического дерева, основанный на двух алгоритмах UPGMA и NJ, которые были модифицированы с целью оптимизации временных затрат на обработку матричных данных следующим образом: матрицы выровненных последовательностей были разбиты на кластеры родственных нуклеотидных последовательностей и обработаны M параллельными процессами с помощью средств MPJ. Полученные поддеревья были объединены в одно филогенетическое дерево главным процессом. Его вычислительная сложность составляет $O(nlk)$, где k является числом кластеров, n - размер набора данных и l - количество циклов алгоритма.

Программная реализация алгоритмов множественного выравнивания и построения дендограмм на языке Java с использованием средств MPI

Полученное программное приложение имеет удобный интерфейс (рис. 3) для выбора файла с нуклеотидными последовательностями.

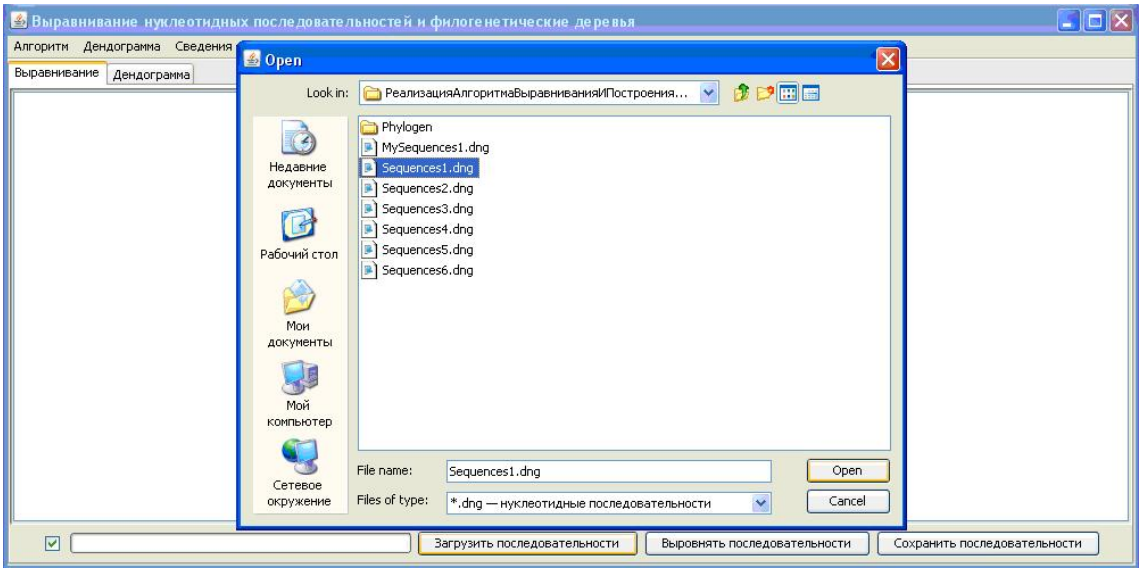

Рисунок 3. Выбор файла с расширением *.dng с нуклеотидными последовательностями

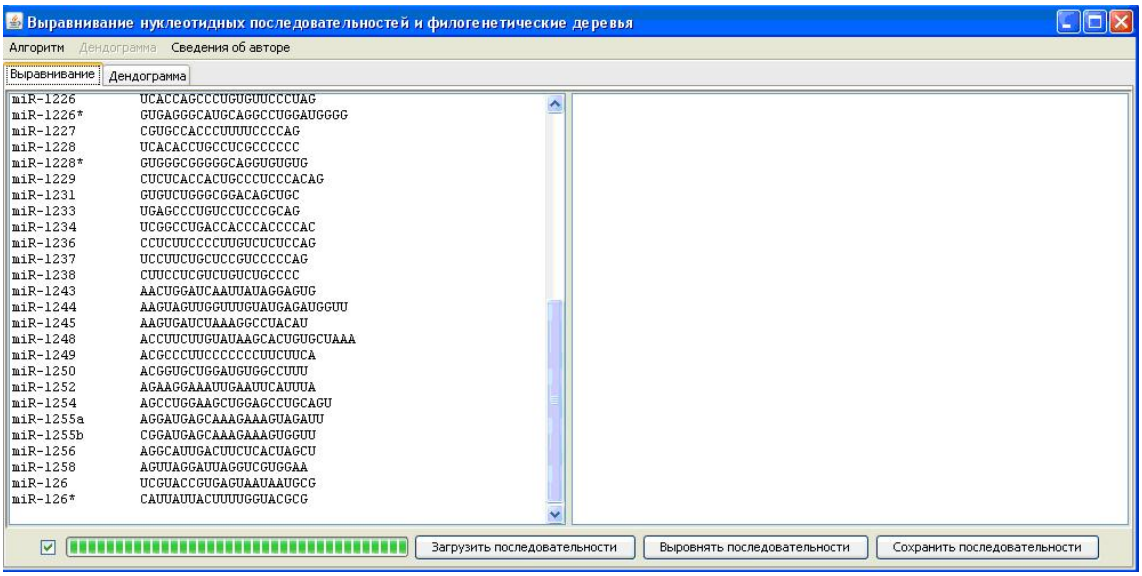

Рисунок 4. Загрузка и просмотр нуклеотидных последовательностей

При загрузке нуклеотидных последовательностей можно выполнить просмотр и редактирование (рис. 4). Помимо этого можно изменить значения матрицы схожести, т.е. матрицы, определяющей степень похожести одного нуклеотида на другой, с тем чтобы изменять параметры выравнивания. К тому же при отражении загружаемых нуклеотидных последовательностей, а также при отображении выровненных последовательностей можно установить таймер, для того чтобы отследить процесс выравнивания последовательностей.

Реализация алгоритма множественного выравнивания нуклеотидных последовательностей была реализована на языке Java с использованием MPJ.

После выполнения выравнивания нуклеотидных последовательностей также предлагается просмотр полученных выровненных последовательностей с возможностью их сохранения в файл (рис. 5).

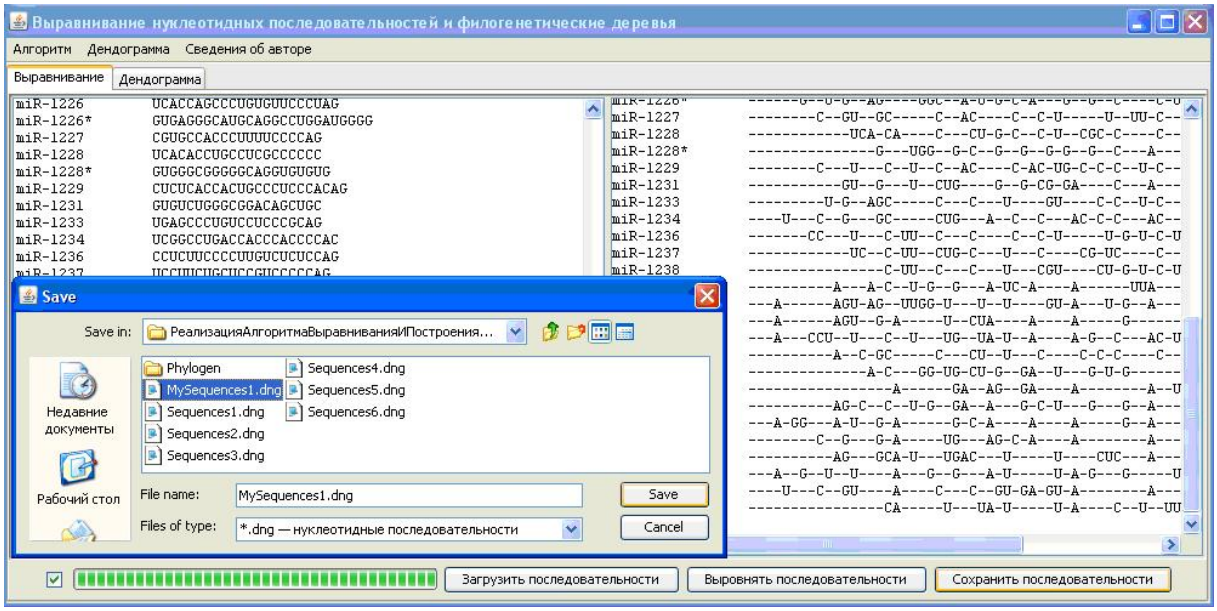

Рисунок 5. Выравнивание и сохранение выровненных нуклеотидных последовательностей в файл

По выровненным последовательностям можно построить дендограмму или филогенетическое дерево, выбрав один из кластерных методов: NJ (neighbor joining) или UPGMA (Unweighted Pair Group Method with Arithmetic Mean). В частности на рис. 6 показано филогенетическое дерево, построенное с использование кластерного метода UPGMA. Интерфейс предлагает возможность увеличивать или уменьшать построенную дендограмму, если это необходимо.

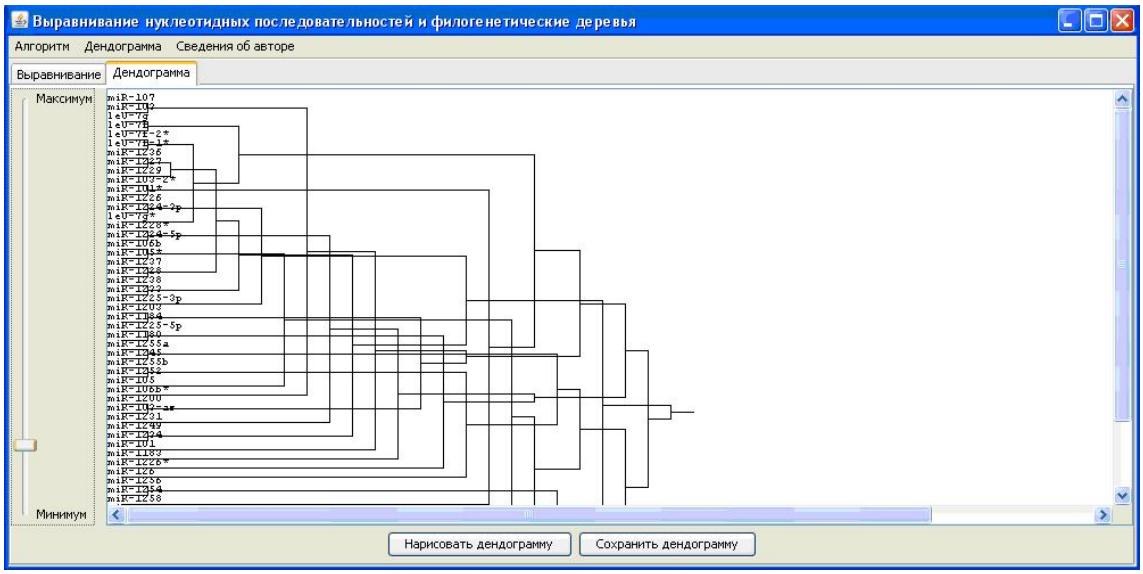

Рисунок 6. Построение филогенетического дерева с использованием кластерного метода UPGMA

Полученное программное приложение было протестировано на данных о миРНК, предоставленных сотрудниками кафедры биотехнологии КазНУ имени аль-Фараби.

В результате выравнивания было построено филогенетическое дерево кластерным алгоритмом UPGMA (рис. 7).

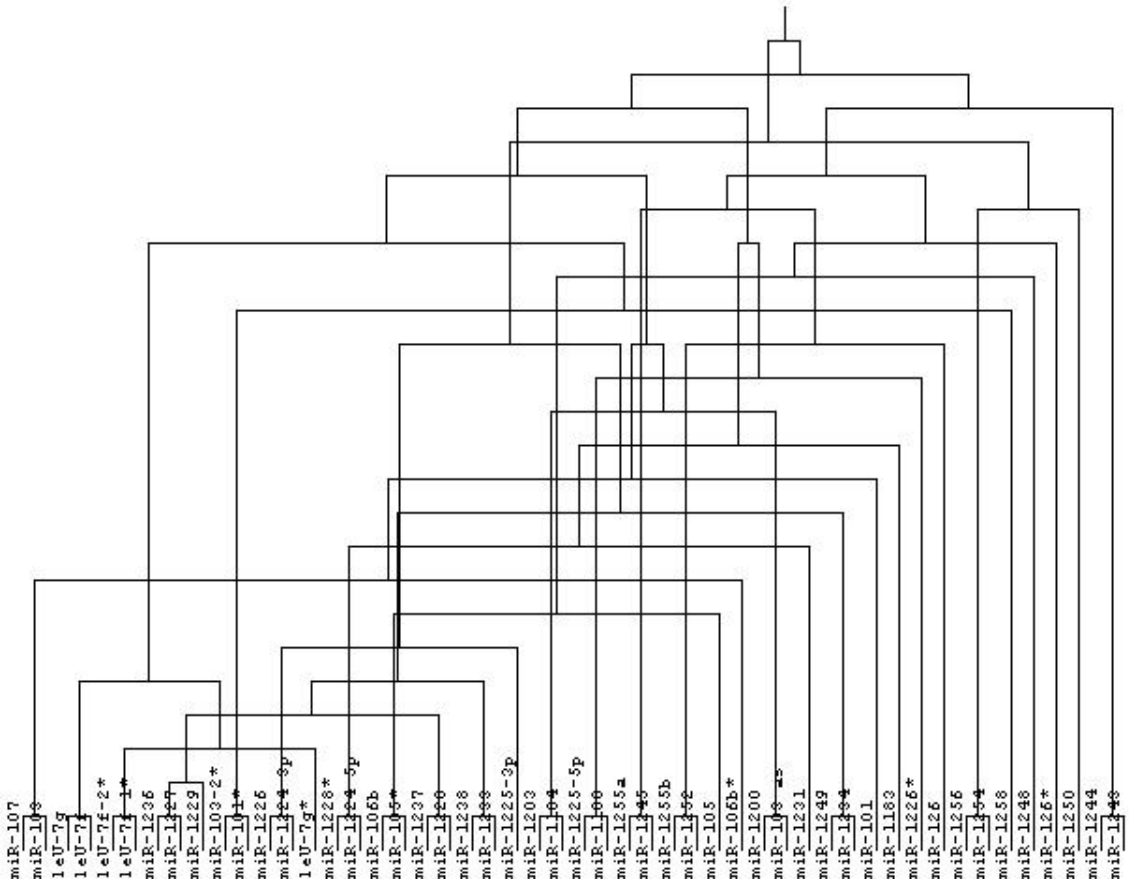

Рисунок 7. Филогенетическое дерево для выровненной выборки миРНК, полученное кластерным методом UPGMA

Список литературы

[1] *Lesk Arthur M.* Introduction to Bioinformatics. - Oxford: Oxford University Press, 2002. - 255 p.

[2] *Ройтберг М.А.* Алгоритмы сравнительного анализа первичных структур биополимеров: автореферат диссертации на соискание ученой степени доктора физико-математических наук: 03.00.28. - М.: Издательство РАН, 2009. - 43 с.

[3] *Jones Neil C., Pevzner Pavel A.* An Introduction to Bioinformatics Algorithms. - Massachusetts: Massachusetts Institute of Technology Press, 2004. - 435 p.

[4] *Пыркова А.Ю.* Множественное выравнивание нуклеотидных последовательностей и построение дендограмм с использованием средств Java MPI // Материалы IX международной научно-практической конференции "Перспективы развития информационных технологий". - Новосибирск: Издательство НГТУ, 2012. - С. 20-25.

[5] *Пыркова А.Ю.* Кластерный анализ больших массивов молекулярно-генетических данных с использование программного интерфейса MPJ // Материалы международной научно-практической конференции "Актуальные проблемы информатики и процессов управления". - Алматы: Институт проблем информатики и управления, 2012. С. 221-225.

[6] *Jonathan M. Keith* Methods in Molecular Biology. Bioinformatics: in 2 vols. - New York: Humana Press, 2008. - V. 2. - 502 p.

[7] Bioinformatics and Biological Computing [Electronic resource]. - 2012. - URL: *http : //bip.weizmann.ac.il/toolbox/overview/software_avail.html* (дата обращения: 07.09.2012)

Поступила в редакцию 18 декабря 2012 года