УДК 591.95

# ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ БИОИНФОРМАЦИОННЫХ ИССЛЕДОВАНИЙ

© А.А. Арзамасцев, В.А. Аверков, Т.И. Горбачева

Arzamastsev A.A., Averkov V.A., Gorbacheva T.I. The program complex for bioinformation researches. The authors suggest carrying out visualization of the genetic information by development of the binary file representing translation result of DNA code. Algorithms of such visualization which are realized in the program «Visual analyzer» are discussed. Computing experiments have confirmed presence of a large number of regular fragments in genetic material of a person.

**Введение.** В настоящее время широкую известность получило новое направление исследований, использующее математические и алгоритмические методы для решения молекулярно-биологических задач. Это направление получило название биоинформатика.

Среди основных задач этой области важно отметить следующие: исследование эволюции живой природы с помощью средств информатики и математики; компьютерное и математическое моделирование информационных процессов в биологических системах; компьютерная генетика (расшифровка и моделирование структурной организации генов и геномов, а также кодируемых генами белков, корреляционный анализ мутаций); компьютерная нейробиология (моделирование природных нейронных систем, разработка и приложение нейросетей) [1].

Одним из приоритетных направлений биоинформатики является сравнительная геномика, занимающаяся системным анализом нуклеотидных последовательностей ДНК и РНК, а также аминокислотных последовательностей белков. Сравнительная геномика позволяет рисовать разные сценарии того, как менялись и меняются различные функциональные системы в ходе эволюции [2, 3].

Также можно отметить следующие направления современной биоинформатики: создание и поддержка баз данных регуляторных последовательностей и белков; компьютерные методы анализа и распознавания в геноме регуляторных последовательностей; компьютерные технологии для изучения генной регуляции; предсказания структуры генов и др. [4].

Огромную роль исследования в области биоинформатики играют в медицине. Благодаря этим исследованиям уже сейчас разрабатываются и внедряются молекулярно-генетические тесты для своевременного выявления опасных заболеваний, вводятся ДНК-чипы, благодаря которым диагностируются формы туберкулеза, устойчивые к определенным лекарственным препаратам. В последнее время фиксируется множество болезней человека, связанных с генетическими нарушениями. Анализ и расшифровка таких генетических сбоев возможны только с помощью компьютерных технологий. Особенно важно, что биоинформатика занимается созданием вакцин против различных вирусов, например, против вируса гепатита С. Работа в этой области может создать новый класс лекарств, который позволит бороться с вирусными инфекциями [2, 5].

С помощью разработок в области биоинформатики предполагается создание так называемой индивидуализированной медицины, т. е. стратегии лечения и профилактики с учетом индивидуальных генетических особенностей. В основе такого подхода лежит генетический паспорт человека, информация о том, что несет именно его ДНК. По ней специалисты могут судить о предрасположенности индивида к различным заболеваниям, а также диагностировать их. По мнению американского эксперта в области биоинформатики Евгения Колкера, индивидуализированная медицина может стать реальностью уже через пять лет [2].

Компьютерный анализ во всех этих исследованиях играет наиболее важную роль. Без компьютерных биоинформационных технологий развитие геномных исследований было бы невозможным. Ведь экспериментальный поиск одного гена занимает недели и месяцы работы целой лаборатории. Компьютерные методы позволяют сделать это за считанные минуты, если просеквенирована ДНК организма и если есть хорошие алгоритмы поиска. В связи с этим за последние годы создано программное обеспечение, позволяющее проводить компьютерный анализ нуклеотидных и аминокислотных последовательностей, организованы базы данных для систематизации геномной и протеомной информации [3, 6].

Цель данной работы – разработка компьютерной программы для биоинформационных исследований, в частности, для поиска регулярных последовательностей в коде ДНК. Указанная программа должна включать в себя алгоритмы трансляции генетической информации, записанной четырехбуквенным кодом в двоичный код, алгоритмы поиска изображений регулярных последовательностей при неизвестной ширине изображения и визуализации таких последовательностей.

В качестве объектов для исследований взяты хромосомы человека: митохондриальная хромосома и Y-хромосома.

**Алгоритмы. Алгоритм трансляции.** Генетический код записан четырехбуквенным алфавитом. Так как при поиске регулярных последовательностей предполагается визуализация информации, то для упрощения зрительного восприятия эта информация должна быть

транслирована в простейшую систему, т. е. в двоичную систему счисления. Таким образом, трансляция должна осуществляться путем перевода каждого четверичного символа в два двоичных знака. Всего существуют 24 варианта такой трансляции. Например, выбрав в качестве первой буквы генетического кода «*a*» – аденин, получим для него четыре возможных варианта трансляции: «00», «01», «10» и «11». Тогда для следующей буквы – «*g*» – гуанина, число вариантов ограничится тремя, для «*c*» – цитозина – двумя, для «*t*» – тимина – одним. Таким образом, общее число вариантов трансляции есть: $4 \times 3 \times 2 \times 1 = 24$. Возможные варианты трансляции показаны в табл. 1.

Алгоритм трансляции заключается в посимвольном чтении файла источника (с кодом ДНК) и записи информации в соответствии с табл. 1 в 24 служебных файла в соответствии с вариантами кодировок.

**Алгоритм поиска изображений при неизвестной ширине растра, точках входа и выхода.** Одной из задач данной работы является визуализация информации, представленной на биологических носителях. Поскольку при визуализации такие понятия, как ширина блока отображаемой информации, точка начала чтения и т.д. не являются определенными, вначале проанализируем, как влияют эти параметры на изображение.

Для этой цели использовали растровый рисунок (см. рис. 1). Было проверено, как меняется данный рисунок при изменении его ширины (увеличении или уменьшении). Для этого был создан шаблон – текстовый файл из нулей и единиц, в котором белый квадрат – это ноль, а черный – единица, соответствующий рис. 1. Шаблон читался и отображался на экране при различной ширине блока. Результаты данного вычислительного эксперимента приведены на рис. 2.

Таблица 1

24 варианта трансляции кода ДНК в двоичный алфавит

| Основания | Варианты трансляции | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| *a* | 00 | 00 | 00 | 00 | 00 | 00 | 01 | 01 | 01 | 01 | 01 | 01 | 10 | 10 | 10 | 10 | 10 | 10 | 11 | 11 | 11 | 11 | 11 | 11 |
| *g* | 01 | 01 | 10 | 10 | 11 | 11 | 00 | 00 | 10 | 10 | 11 | 11 | 01 | 01 | 00 | 00 | 11 | 11 | 01 | 01 | 10 | 10 | 00 | 00 |
| *c* | 10 | 11 | 01 | 11 | 01 | 10 | 10 | 11 | 00 | 11 | 00 | 10 | 00 | 11 | 01 | 11 | 01 | 00 | 10 | 00 | 01 | 00 | 01 | 10 |
| *t* | 11 | 10 | 11 | 01 | 10 | 01 | 11 | 10 | 11 | 00 | 10 | 00 | 11 | 00 | 11 | 01 | 00 | 01 | 00 | 10 | 00 | 01 | 10 | 01 |

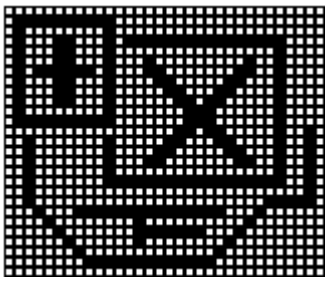


**Рис. 1.** Исходный растровый рисунок размером 32×27 пикселей

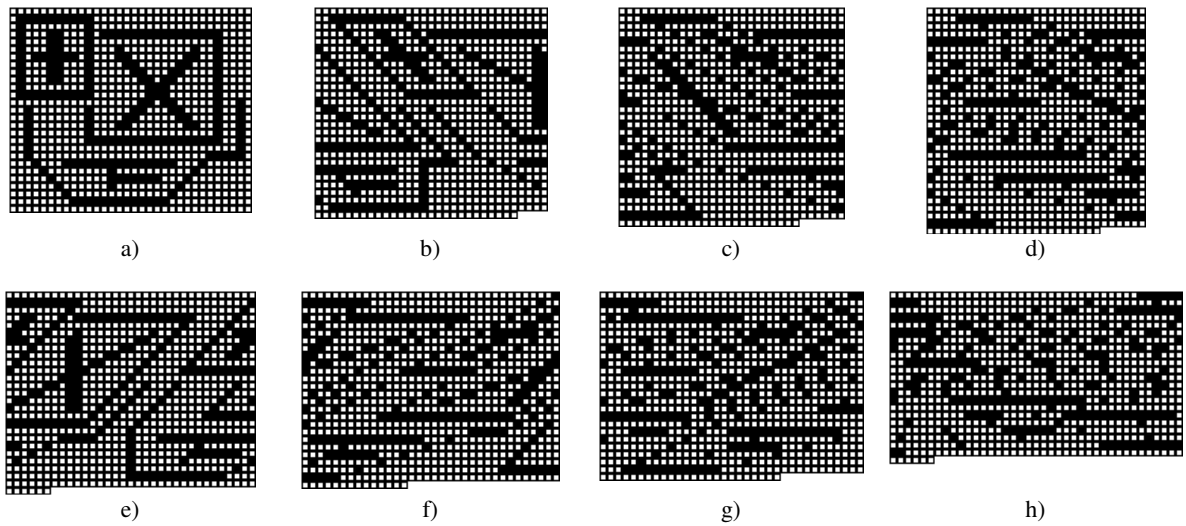



**Рис. 2.** Исходное изображение (ширина 32 пикселя) – a) и его трансформации при изменениях ширины растра: b) – 31 пиксель; c) – 30 пикселей; d) – 29 пикселей; e) – 33 пикселя; f) – 34 пикселя; g) – 35 пикселей; h) – 39 пикселей

Видно, что при изменении ширины рисунка вертикальные полосы становятся диагональными, горизонтальные, как правило, остаются, меняя своё местоположение. Причем небольшое изменение ширины (на 1–2 пикселя) позволяет увидеть искаженный рисунок. Дальнейшее изменение ширины приводит к потере рисунка. Можно сделать вывод, что при наблюдении диагональных последовательностей следует попробовать изменить ширину на несколько пикселей, чтобы проверить рисунок на наличие вертикальных последовательностей.

Далее проверяли, как изменится данный рисунок, если удалить несколько символов (нулей и единиц) в начале или в конце или внутри исходного текстового файла. Результаты показаны на рис. 3.

Видно, что отсутствие начала или конца записи приводит к потере его части, но при правильном выборе ширины та часть, что осталась, не искажается, однако происходит фрагментация рисунка.

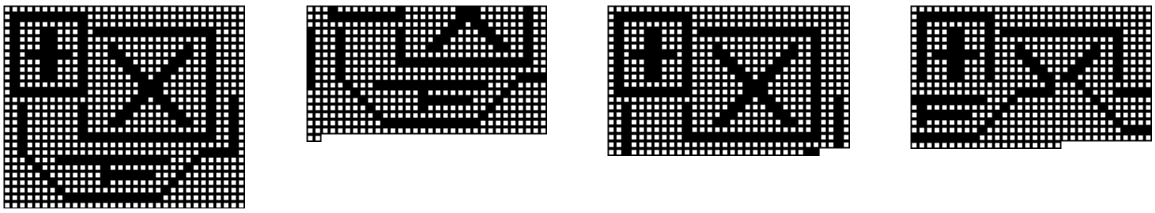
На практике для более точной «подгонки» последовательности необходимо учитывать начальный символ, с которого происходит рисование. Должна быть также введена возможность смещения начального пикселя в программе визуализации.

Из рис. 3 d) видно, что при утрате фрагмента внутри файла исходное изображение изменилось. Если не знать вид оригинала, то невозможно было бы понять, что оно неправильное. Утраченную часть невозможно восстановить. Поэтому для более тщательной подборки в алгоритме визуализации необходимо предусмотреть сдвиг, начинаемый с любого пикселя по любой введенной строчке.

**Алгоритм визуализации**. Для визуализации транслированной последовательности использовали модуль MathImage, присоединенный к общим модулям Delphi. Визуализация представляет собой построение графических объектов по текстовому файлу и учитывает некоторые особенности.

**Программный продукт «Визуализатор»** версии 1.0 предназначен для трансляции генома ДНК в двоичный код с последующей визуализацией.

Поле программы (рис. 4) можно разделить на две части: блок управления и область просмотра результата. Первая часть включает блоки трансляции, визуализации и область сохранения полученной информации.



a) b) c) d)

**Рис. 3.** Исходное изображение (ширина 32 пикселя) – a) и его трансформации при отсутствии нескольких символов вначале – b), в конце – c) и внутри – d) записи

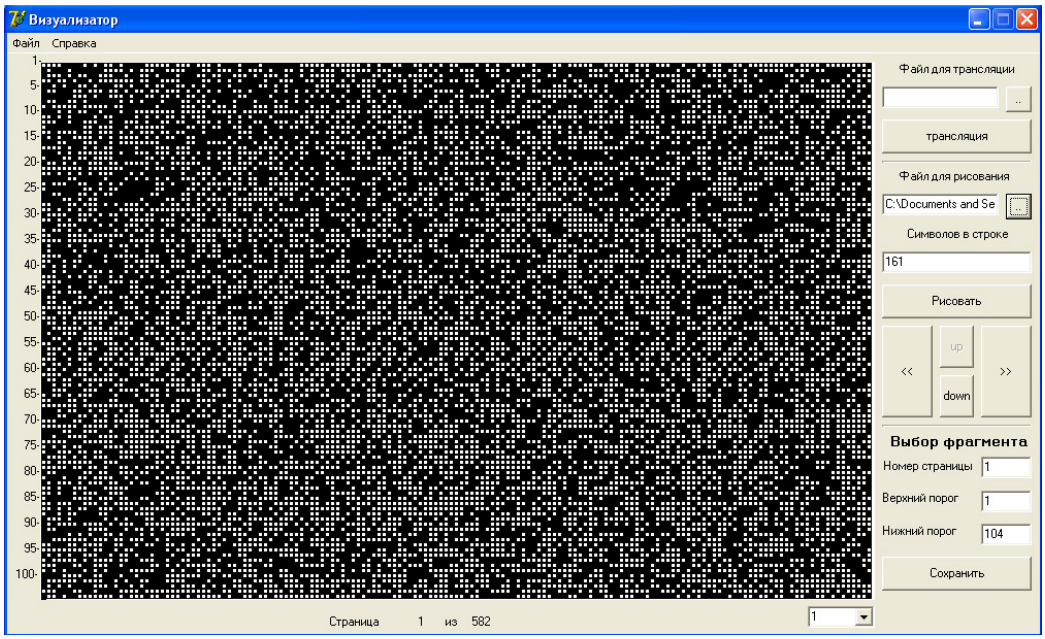



**Рис. 4.** Общий вид программы

В правом верхнем углу находится область, предназначенная для транслирования кода ДНК в двоичную систему.

Выбрав файл для трансляции и нажав кнопку «трансляция», программа запускает механизм перевода кода ДНК в двоичную систему, создавая в каталоге файла 24 транслированных варианта. Таблица трансляции указана в трех первых строках каждого файла в виде служебной информации.

Необходимо учитывать, что при транслировании файлов больших объемов процесс может занять достаточно длительное время (например, Y-хромосома объемом около 9 Мбайт на Turion64 3500+ транслировалась порядка 5–6 мин.).

Далее возможны визуализации одного из полученных вариантов. Для этого используется второй блок программы, изображенный на рис. 5.

После выбора файла для визуализации пользователю доступны следующие действия: задать ширину страницы (в символах) в строке 1, изменить ее посимвольно кнопками 3 и 5, «листать» страницы кнопками 4, перерисовать страницу кнопкой 2 (например, при ручном вводе ширины). Изменение ширины зоны рисования объясняется необходимостью более точной передачи регулярности на рисунке, т. е. привод ее к более удовлетворительному виду путем смещения пикселей. Для быстрого перехода по страницам также предусмотрено окно выбора страниц, находящееся внизу рабочего пространства.

Третий блок программы предназначен для сохранения результата в текстовом и/или графическом файле (рис. 6). Для этого указывается номер страницы, верхний и нижний пороги найденной регулярности. Это дает возможность впоследствии работать с выбранным фрагментом отдельно, в этой же программе, для более детального его изучения.

Область просмотра результата вмещает на странице 104 строки по вертикали и 161 элемент по горизонтали. Как показали наши вычислительные эксперименты, большего обычно не требуется, т. к. зависимости находятся в основном на кратных четырем ширинах, т. е. по 40, 80, 120 и 160 символов в строке, и имеют высоту порядка 10–40 строк.

Для удобства выбора области сохранения и отслеживания высоты последовательностей слева от панели рисования осуществлена градация строк. Имеется также и счетчик страниц – текущей и их общего количества.

**Примеры работы программы.** В качестве примера были найдены несколько регулярных последовательностей в Y-хромосоме человека. Их изображения приведены на рис. 7.

Из полученных результатов видно, что регулярные последовательности особо часто наблюдаются при ширинах в 40, 80, 120 и 160 пикселей. Все эти числа кратны 4. Регулярности в митохондриальной ДНК выражены не так явно и требуют более детального изучения.

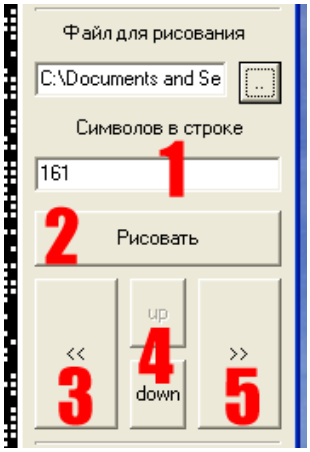



**Рис. 5**. Блок визуализации

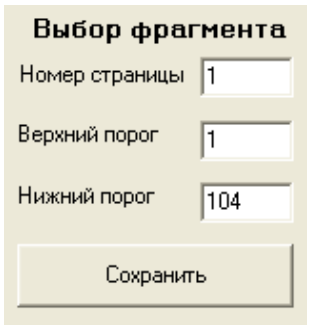



**Рис. 6.** Блок сохранения результата

Таким образом, в результате выполнения данной работы:

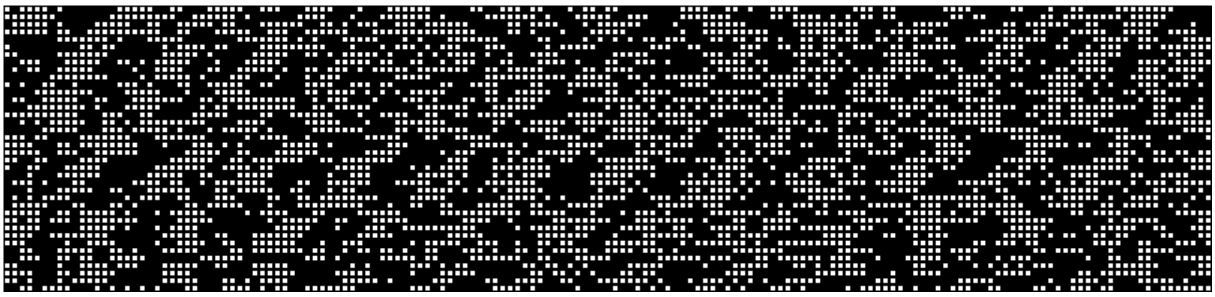
– апробированы алгоритмы, предназначенные для визуализации биологической информации;

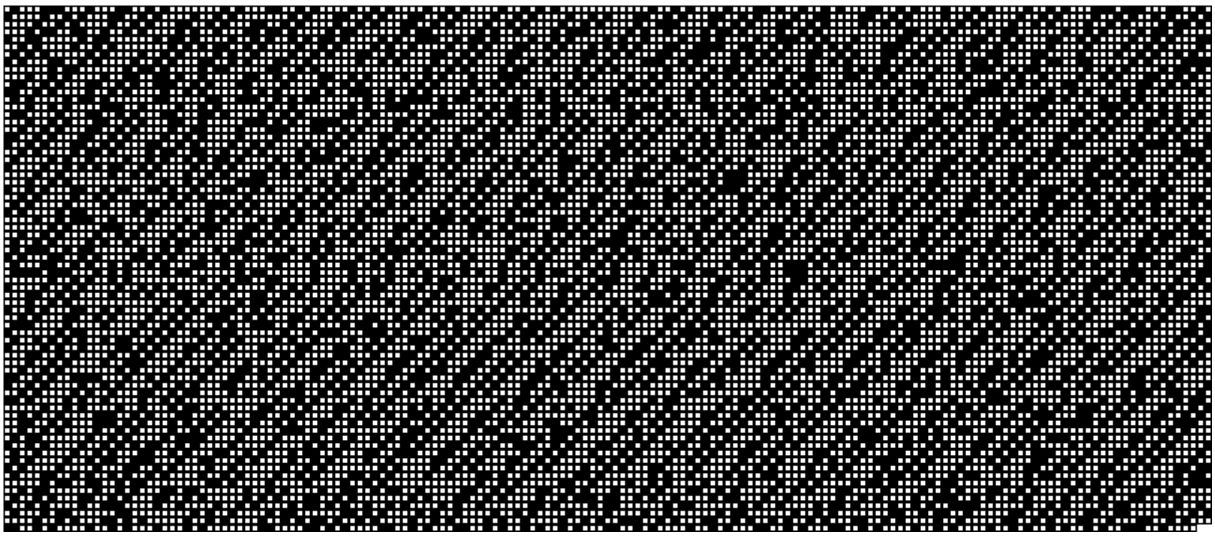
– алгоритмы реализованы в виде программы «Визуализатор»;

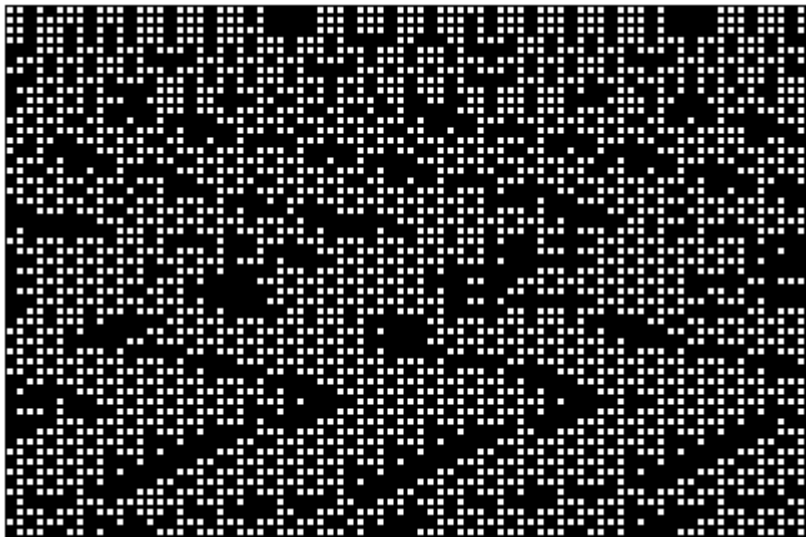
– проведенные вычислительные эксперименты подтвердили наличие большого числа регулярных фрагментов в генетическом материале человека.

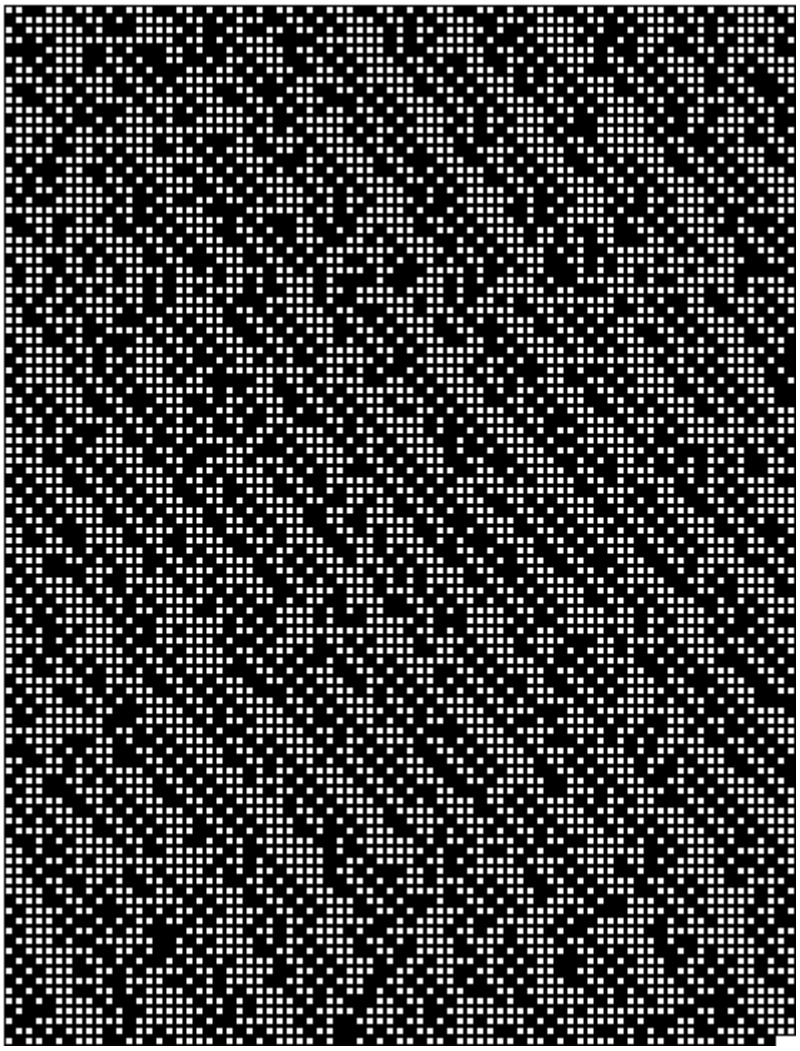
## ССЫЛКИ

1. www.lib-online.ru
2. www.newtimes.ru
3. *Фролова Л.Л.* Компьютерный анализ последовательностей. Казань: Изд-во Казан. ун-та, 2001.
4. www.biobel.bas-net.by
5. www-sbras.nsc.ru
6. www.bio.1september.ru









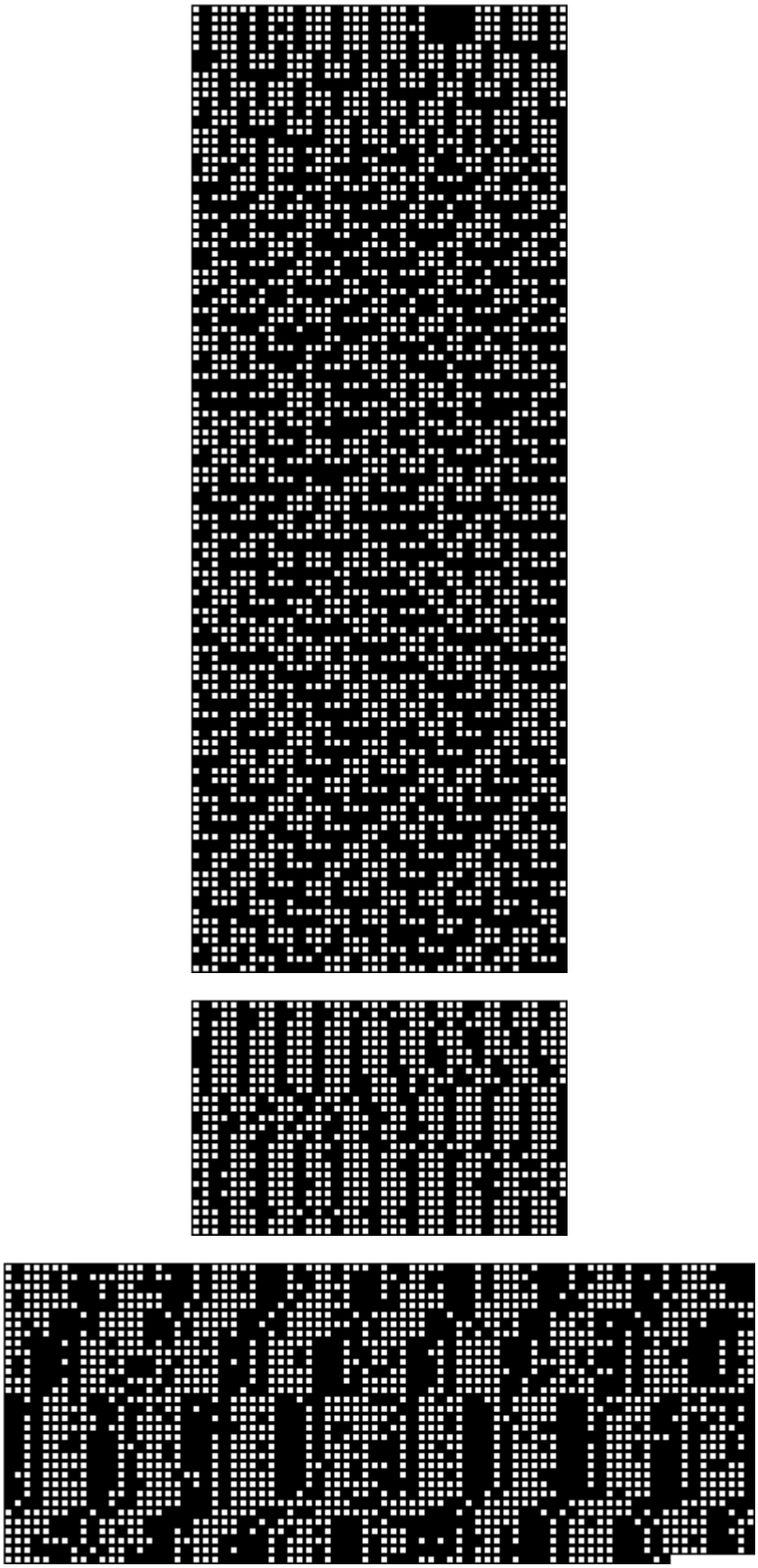


**Рис. 7**. Регулярные последовательности, найденные в Y-хромосоме человека при различных ширинах блока

Поступила в редакцию 28 ноября 2006 г.