Глава 6

Сравнительный анализ информационных биополимеров

Т. В. Астахова, Н. В. Олейникова, М. А. Ройтберг

6.1. Введение. Развитие методов анализа биополимеров

Анализ биологических последовательностей имеет свою специфику — прежде всего с точки зрения постановок задач. Например, задача о распознавании «вторичной» структуры РНК очень важна для молекулярной биологии и впервые была рассмотрена еще в конце 70-х годов. Молекула рибонуклеиновой кислоты (РНК) — однонитевой полимер, состоящий из четырех видов мономеров-нуклеотидов (аденин, гуанин, урацил, цитозин). А–У и, соответственно, Г–Ц могут образовывать водородные связи, стабилизирующие молекулу. Однако образование одних связей из-за стереохимических соображений делает невозможным образование других, то есть не все комбинации межнуклеотидных связей в-молекуле РНК допустимы (правила конфликтов между связями известны). Требуется для данной нуклеотидной последовательности найти наиболее стабильную вторичную структуру, т. е. допустимый набор межнуклеотидных связей, содержащий наибольшее возможное количество элементов. Эта задача может быть переформулирована как задача построения графа специального вида с максимально возможной суммой весов ребер (вершины соответствуют нуклеотидам, ребра — установленным связям) и решена с помощью методов динамического программирования. На примере задачи распознавания наиболее стабильной «вторичной» структуры РНК, отметим следующие обстоятельства, характерные для многих важных задач анализа биополимеров:

- в описанной выше модели правильнее считать не количество связей, а их суммарную энергию, энергия каждой возможной пары считается известной. С алгоритмической точки зрения задача практически не меняется;

- модель, положенная в основу описанной выше задачи, — упрощенная и во многих случаях не согласуется с экспериментом. Полезно учитывать и вклад нуклеотидов, не участвующих в образовании водородных связей. Ограничения на множество допустимых наборов связей, принятые в задаче, слишком строгие. Различные формальные постановки задач, лучше отражающие биологическую реальность, приводят к существенному усложнению алгоритма;

- в реальности молекула РНК может принимать не ту структуру, которой мы приписали оптимальную энергию, а несколько иную, например, из-за того, что мы не знаем точных значений энергетических параметров. Поэтому полезно не искать одну «оптимальную» структуру, а проанализировать все возможные структуры и оценить вероятность образования каждой отдельной связи («статистический вес» связи). Это также можно решить методом динамического программирования.

- многие авторы пытаются выяснить вторичную структуру РНК, не сводя ее к какой-либо алгоритмической оптимизационной задаче, а путем моделирования реального процесса «сворачивания» молекулы РНК (т. е. установления и исчезновения водородных связей).

Нельзя не сказать о самой старой и, наверное, самой популярной задаче анализа биологических последовательностей — об их выравнивании. *Выравнять две последовательности — это изобразить их друг над другом, вставляя в обе пробелы так, чтобы сделать их длины равными*. Вот, например, как можно выровнять слова ПОДБЕРЕЗОВИК и ПОДОСИНОВИК:

ПОДБЕРЕЗОВИК	(1)
ПОДОСИНОВИК-	
ПОДБЕРЕЗОВИК	(2)
-ПОДОСИНОВИК	
ПОДБЕРЕЗОВИК	(3)
ПОДОСИН-ОВИК	
ПОДБЕРЕЗОВИК	(4)
ПОД-ОСИНОВИК	
ПОДБЕРЕЗ----ОВИК	(5)
ПОД-----ОСИНОВИК	

Такой способ изображения последовательностей широко распространен в молекулярной биологии. Предполагается, что выравнивание отражает эволюционную историю, то есть стоящие друг под другом символы соответствуют одному и тому же символу последовательности-предка. К сожалению, мы не знаем, как именно шла эволюция последовательностей. Поэтому в качестве «правильного» обычно выбирается выравнивание, оптимальное относительно некоторой функции качества. Но как мы можем контролировать правильность выбора этой функции? Есть ли у нас (пусть приблизительные) «эталоны»? К счастью, да. В качестве эталонных можно взять выравнивания, соответствующие наилучшему возможному совмещению их пространственных структур (такие структуры известны для нескольких сотен белков). Это связано с тем, что функционирование белка в клетке определяется прежде всего его пространственной структурой и можно ожидать, что аминокислоты, лежащие в сходных местах трехмерной структуры, соответствуют одним и тем же аминокислотам предкового белка.

В «добиологическом» анализе последовательностей (например, при сравнении файлов) использовалось понятие редактирующего расстояния. При этом фиксируется набор редактирующих операций (например, замена символа, вставка символа и удаление символа) и для каждой операции фиксируется цена. Тогда каждое выравнивание получает свою цену, определяемую как сумма цен отдельных операций.

Лучшим считается то, которое имеет наименьшую цену. Например, при цене замены 1 и цене вставки/удаления 3, лучшими в примере будут третье и четвертое выравнивания, а при цене замены 10 и той же цене вставки/удаления, лучшим будет пятое.

Довольно скоро выяснилось, что для выравнивания биологических последовательностей в эту естественную схему необходимо внести ряд важных изменений. Дело в том, что разные аминокислоты различны по-разному. Например, аланин и валин очень похожи по своим свойствам (и цена замены аланина на валин должна быть небольшой), и они оба совершенно не похожи на триптофан. Более того, даже одинаковые аминокислоты «одинаковы по-разному». Так, триптофан — редок, и сопоставление двух триптофанов более ценно, чем сопоставление весьма распространенных аланинов.

Поэтому вместо «цены замены символа» в схеме редактирующего расстояния при сравнении белков используется весовая матрица замен, где каждой паре символов соответствует вес (положительный — для похожих,

отрицательный для непохожих), а выравниванию в целом — вес $W = R - G$, где R — суммарный вес сопоставлений символов (в соответствии с выбранной весовой матрицей замен), G — суммарный штраф за удаления и вставки символов. Таким образом, оптимальное выравнивание — это выравнивание, имеющее наибольший вес (в то время как цена требовалась наименьшая). Например, пусть вес совпадения для гласных букв $+2$, вес совпадения для согласных букв $+1$, вес сопоставления двух различных гласных или двух различных согласных -1, вес сопоставления гласной и согласной -2. Далее, пусть штраф за удаление или вставку символа -5. Тогда, например, третье выравнивание имеет вес -3, а четвертое — $+1$. Таким образом, оптимальное выравнивание слов ПОДБЕРЕЗОВИК и ПОДОСИНОВИК (при выбранных матрице замен и штрафе за удаление/вставку) — четвертое. Переход от минимизации цены к максимизации качества, — это не только технический трюк. На языке максимизации качества естественно ставится задача о поиске оптимального локального сходства. Эта задача соответствует сравнению двух белков, которые в ходе эволюции стали совсем непохожи — везде, кроме относительно короткого участка.

Алгоритм построения оптимального выравнивания основан на методе динамического программирования, введенном в широкую практику Ричардом Беллманом в 1957. Идея метода состоит в следующем: чтобы решить основную задачу, нужно придумать множество промежуточных и последовательно их решить (в каком порядке — отдельный вопрос). При этом очередная промежуточная задача должна «легко» решаться, исходя из уже известных решений ранее рассмотренных задач. Множество промежуточных задач удобно представлять в виде ориентированного ациклического графа. Его вершины соответствуют промежуточным задачам, а ребра указывают на то, результаты решений каких промежуточных задач используются для основной. Таким образом, исходная задача сводится к поиску оптимального пути в графе Аналогично можно переформулировать различные варианты задач выравнивания, предсказания вторичной структуры РНК и белков, поиска белок-кодирующих областей ДНК и других важных проблем анализа биологических последовательностей. При построении оптимального выравнивания (мы рассматриваем простейший случай, когда удаление и вставка отдельных символов штрафуются независимо) промежуточные задачи — это построение оптимальных выравниваний начальных фрагментов исходных последовательностей. При этом задачи нужно решать в порядке возрастания длин фраг-

ментов. Граф зависимости между промежуточными решениями для сравнения слов «ПАПКА» и «ПАПАХА», а также последовательность промежуточных шагов, приводящих к оптимальному выравниванию, показаны на рис. 6.1a.– рис. 6.1c.

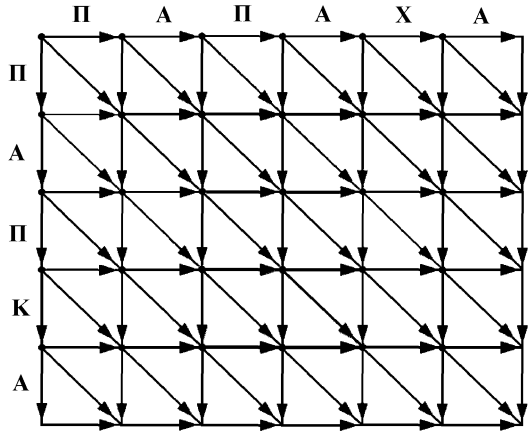

Рис. 6.1a. Граф зависимостей между промежуточными задачами для выравнивания слов ПАПКА и ПАПАХА. Каждая вершина соответствует паре начальных фрагментов указанных слов. Диагональное ребро, входящее в вершину, соответствует сопоставлению последних букв сравниваемых начальных фрагментов (случай 1) горизонтальное ребро — удалению буквы в слове ПАПАХА, вертикальное ребро — удалению буквы в слове ПАПКА (случаи 2 и 3). Правая верхняя вершина — начальная и соответствует выравниванию пустых слов, левая нижняя вершина — конечная, соответствует выравниванию полных слов ПАПКА и ПАПАХА

На двух примерах — распознавания вторичной структуры РНК (бегло) и выравнивания белковых последовательностей (более подробно) мы проследили за эволюцией постановок задач анализа биополимеров. Упомянем кратко еще несколько аспектов. Пожалуй, с практической точки зрения самым важным является поиск в базах данных последовательностей, сходных с изучаемой. Определяющую роль начинают играть проблемы вычислительной эффективности, решаемые, в частности, с применением алгоритмов хеширования. Для предсказания пространственной структуры белков важны алгоритмы выравнивания последовательности со структурой (при этом

ПАП-КА
ПАПАХА

Рис. 6.1б. Оптимальное выравнивание слов ПАПКА и ПАПАХА при следующих параметрах: вес совпадения букв: 1, штраф за замену гласной на гласную или согласной на согласную: 1, штраф за замену гласной на согласную или согласной на гласную: 2, штраф за удаление символа: 3

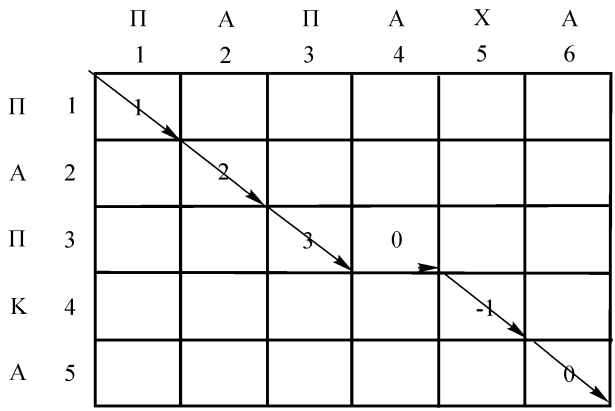

Рис. 6.1с. Траектория, соответствующая оптимальному выравниванию. В клетках указаны веса промежуточных оптимальных выравниваний. Например, вес оптимального выравнивания для «ПАП» и «ПАПА» равен 0, а для «ПАПК» и «ПАПАХ» равен -1

используется тот факт, что из-за разницы физико-химических свойств аминокислоты встречаются с разной частотой на поверхности белка и в структурном ядре). Наконец, мы полностью оставили в стороне задачи построения эволюционных деревьев по белковым последовательностям. Подчеркнем, что во всех случаях происходит интенсивная «притирка» постановок задач — как с биологической (большая адекватность), так и с алгоритмической (возможность построения более эффективных алгоритмов) точки зрения.

6.2. Другой подход к проблеме выравнивания аминокислотных последовательностей. Парето-оптимальные выравнивания

При традиционных подходах каждому выравниванию сопоставляется число — вес выравнивания и строится оптимальное т. е. имеющее наибольший возможный вес, выравнивание. Вес выравнивания является функцией «элементарных» характеристик выравнивания (количества совпадений, удаленных символов и т. п.) и параметров, которые должны отражать специфику сравниваемых последовательностей. К сожалению, выбор значений этих параметров часто весьма затруднен, Вид весовой функции (как правило, линейный) также часто не имеет достаточного основания. Для того, чтобы преодолеть указанные трудности был предложен многокритериальный подход [1,2] к проблеме выравнивания: в качестве веса выравнивания используется не число, а вектор в некотором k-мерном пространстве. Компонентами такого вектора могут быть, например, количество совпадений, количество удаленных символов, число множественных делеций («дырок») — групп идущих подряд удаленных символов. В работе [3] был предложен алгоритм выделения из множества всех векторов выравниваний специального подмножества, векторам которого соответствуют оптимальные выравнивания последовательностей. Пусть S_1, и S_2 — две последовательности длин n и m, соответственно, построенные из символов некоторого конечного алфавита. Каждому выравниванию **A** этих последовательностей ставится в соответствие k-мерный вектор $V(A)$. Например, в случае $k = 2$

$$V(A) = (\mathrm{Comp}(A), \mathrm{Gap}(A)), \tag{1}$$

где $\mathrm{Comp}(A)$ — сумма весов сопоставлений в **A**, а $\mathrm{Gap}(A)$ — число множественных делеций в **A**. Или:

$$V(A) = (\mathrm{Match}(A), -\mathrm{Del}(A)), \tag{2}$$

где $\mathrm{Match}(A)$ — количество совпадающих символов, $\mathrm{Del}(A)$ — количество удаленных символов в **A**.

Определение 1. Пусть S_1 и S_2 — две последовательности длиной, соответственно, n и m; $k \geqslant 2$ — некоторое число. Сопоставим каждому выравниванию **A** этих последовательностей k-мерный вектор $V(A)$. Функция V будет называться *весовой* (или оценочной) функцией, а вектор $V(A)$ — *весом выравнивания* **A**.

Определение 2. Вектор V_1 доминирует над вектором V_2, если каждая компонента вектора V_1 больше или равна соответствующей компоненте вектора V_2, и имеет место хотя бы одно строгое неравенство. Если вектор V_1 не доминирует над вектором V_2, и выравнивание V_2 также не доминирует над V_1, то выравнивания называются *несравнимыми*.

Определение 3. Пусть M — множество k-мерных векторов. Вектор $\boldsymbol{\nu}$, лежащий во множестве M, называется *Парето-оптимальным* в M, если никакой другой вектор $\mathbf{u}$ из M не доминирует над $\boldsymbol{\nu}$ (рис. 6.2). *Парето-оптимальное подмножество* в M — это подмножество, состоящее из всех векторов, Парето-оптимальных в M.

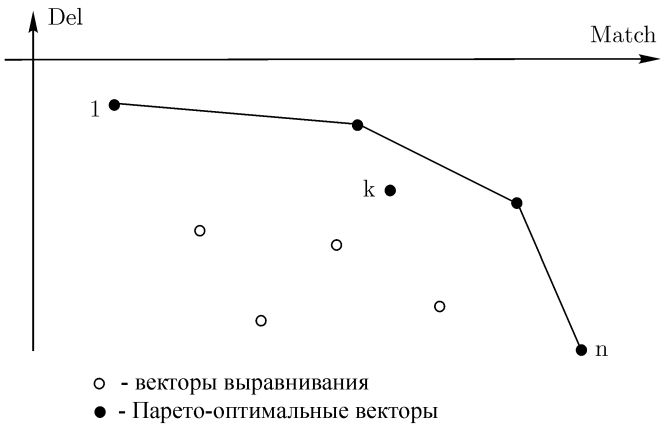

Рис. 6.2. Графическое изображение Парето-подмножества

Определение 4. Пусть S_1 и S_2 — последовательности, V — весовая функция. Выравнивание **A** последовательностей S_1 и S_2 называется *Парето-оптимальным* относительно весовой функции V, если вектор-вес $V(A)$ является Парето-оптимальным вектором в множестве весов всех выравниваний последовательностей S_1 и S_2.

Множество весов всех выравниваний S_1 и S_2, Парето-оптимальных относительно весовой функции V, будем называть множеством *Парето-оптимальных весов* для S_1, S_2 и V или просто множеством Парето-оптимальных весов.

Утверждение 1. Пусть задана векторная весовая функция выравнивания $V(A) = \{x_1(A), \ldots, x_k(A)\}$ и скалярная весовая функция $W(A) = W(x_1(A), \ldots, x_k(A))$, где функция $W(x_l, \ldots, x_k)$ монотонно возрастает относительно каждого из аргументов $x_1, \ldots, x_k$.

Пусть **A** — оптимальное выравнивание последовательностей S_1, S_2 относительной весовой функции $W(A)$. Тогда **A** является Парето-оптимальным выравниваниваем относительно весовой вектор-функции $V(A)$.

Пример: последовательность S_1 =MNTPFVCFIM
последовательность S_2 =MNAALTPSRFCFVV

Множество Парето-оптимальных выравниваний для весовой функции $V(A) = (\text{Comp}(A), \text{Gap}(A))$ показано на таблице 1.

Таблица 1. Множество Парето-оптимальных выравниваний для весовой функции $V(A) = (\text{Comp}(A), \text{Gap}(A))$

№	**Выравнивание (локальное)**	**Comp**(A)	**Gap**(A)	**число «островов»**[1]
1	**CF**IM **CF**VV	17	0	1
2	**TP**F.V**CF**IM **TP**SRF**CF**VV	27	1	2
3	**TP . . . TP**F.V**CF**IM **MN**AAL**TP**SRF**CF**VV	35	2	3
3	**MN . . . TP..F**V**CF**IM **MN**AAL**TP**SR**F.CF**VV	39	3	4

Замечание 1. Парето-оптимальное выравнивание может не быть оптимальным ни для какой линейной функции $m \cdot \text{Match}(A) - d \cdot \text{Del}(A)$ (вектор k на рис. 6.2). То есть, множество всех возможных оптимальных выравниваний для данной пары последовательностей представляет собой выпуклую оболочку, натянутую на Парето-оптимальное подмножество.

Замечание 2. Интерес для изучения представляет поведение только той части выпуклой оболочки, которая натянута на верхнюю правую часть Парето-подмножества, соединяющая векторы с максимальным значением компоненты $-\text{Del}(A)$

[1] «Остров» — это участок выравнивания между делециями.

и максимальным значением компоненты Match(A) (векторы $1-n$ на рис. 6.2). В традиционном подходе к анализу сходств последовательностей обычно получают одно оптимальное в смысле весовой функции выравнивание с фиксированным штрафом за делецию. Новая идея заключается в том, что мы рассматриваем несколько выравниваний. Выбор веса сопоставлений и числа множественных делеций в качестве компонент вектора имеет принципиальное значение. Легко показать, что выравнивание с фиксированным аффинным штрафом находится среди выравниваний Парето-оптимального множества, то есть, применяя Парето-технику с выбранным вектором весов мы, как минимум ничего не теряем.

Для Парето-выравниваний справедливо следующее ключевое утверждение:

Если структурно-верное выравнивание (идеальное решение задачи) достижимо методом выравнивания символьных последовательностей с какой-либо схемой штрафов, то оно входит во множество Парето-оптимальных выравниваний. Это означает, что, имея способ выбора правильной точки из Парето множества, мы фактически имеем метод построения биологически правильного выравнивания.

6.3. Распознавание белок-кодирующих областей в последовательностях ДНК — важная задача анализа биологических последовательностей

Будучи одной из традиционных задач компьютерной генетики, распознавание белок-кодирующих областей становится сейчас особенно актуальным в связи с появлением большого количества протяженных неаннотированных фрагментов геномной ДНК из различных организмов. Современные успехи в крупномасштабном секвенировании ДНК открывают возможность нового подхода к распознаванию генов, основанного на последовательном использовании данных о гомологичных белках. Доля новосеквенированных генов, имеющих уже известных родственников, составляет в настоящее время долее 60% и постоянно растет. Однако прямое использование этой информации или приводит к вычислительным проблемам и, как следствие, к неприемлемым временам счета, или не исчерпывает всех возможностей подхода. Алгоритм сплайсированного выравнивания [4, 5], позволяет среди всех потенциальных генов, закодированных в рассматриваемом участке ДНК, эффективно найти наиболее сходный с родственным геном из банка данных. Эксперименты на тестовых выборках показали, что возможно добиться 99%-ного распознавания генов человека, если известен родствен-

ный белок млекопитающих, и 91%-ного распознавания, если известен белок птиц или холоднокровных позвоночных. Качество работы алгоритма не ухудшается на длинных фрагментах ДНК (15–25 тыс. нуклеотидов) содержащих 10 экзонов и более.

В основе этого алгоритма лежит идея максимально полного использования информации о гомологах распознаваемого гена. Алгоритм сплайсированного выравнивания работает следующим образом. Сначала строятся все потенциальные белок-кодирующие экзоны. Затем осуществляется слабая фильтрация, не приводящая к потерям истинных экзонов. Это означает, что правило отсева является довольно слабым, и полученное в результате множество экзонов относительно велико. Затем с помощью подходящего варианта метода динамического программирования находится потенциальный ген, имеющий наилучшее выравнивание с родственным белком из банка данных. Так как этот шаг делается за полиномиальное (относительно числа возможных экзонов) время, то слабая фильтрация не препятствует достаточно эффективной работе алгоритма.

Компьютерный анализ участков последовательностей ДНК, связанных с экспрессией генов (белок-кодирующих областей, фукциональных сайтов) — одна из «классических» областей современной компьютерной генетики. Алгоритмы распознавания функциональных сайтов основываются на предположении, что все сайты определенного типа обладают неким общим свойством (сигналом), которое распознается соответствующим ДНК- (или РНК-) связывающимся белком. В качестве элементов сигналов рассматриваются как нуклеотиды (предпочтения нуклеотидов в различных позициях), так и более сложные объекты, например, участки «аномального» олигонуклеотидного распределения или элементы вторичной структуры Наиболее распространены подходы, использующие метод весовых матриц, и различные методы распознавания образов. В последние годы были предложены новые интересные подходы, в частности, использующие нейронные сети, и получены интересные результаты. По-видимому, можно считать решенной задачу выделения белок-кодирующих областей в отсутствие интронов (у прокариот). Для ряда сигналов (в основном, различных промоторов) показана корреляция между предсказанной силой сайта и уровнем экспрессии соответствующего гена.

Распознавание белок-кодирующих участков ДНК эукариот также быстро развивающаяся и актуальная область исследований. Известно, что кодирование белка приводит к статистическим ограничениям на последовательность ДНК. Таким образом, можно опредедить «кодирующий потен-

циал», измеряющий статистическую похожесть участка ДНК на известные кодирующие участки.

Такой подход оказался весьма плодотворен для прокариот, однако в эукариотическом случае его применение наталкивается на существенные трудности: кодирующий потенциал нельзя вычислять ни для открытых рамок считывания (интроны нарушают открытую рамку считывания), ни для скользящего окна (экзоны часто бывают короче, чем длина окна, которая не может быть уменьшена из-за статистического шума). Дополнением к «глобальному» подходу, основанному на использовании того или иного кодирующего потенциала, служит «локальный» подход, состоящий в предсказании сайтов сплайсинга и, самих, границ экзонов. Такие предсказания осуществлялись при помощи простых весовых матриц, поиска (вырожденных) паттернов, более продвинутого «синтаксического анализа паттернов», алгоритмов типа «кора» и «перцептрон», Байесовского дискриминантного анализа сложных признаков, нейронных сетей. Однако ни один из предложенных алгоритмов не обладает достаточной надежностью и избирательностью. В зависимости от строгости решающего правила (например, величины порога для методов, сводящихся к вычислению распознающей функции) либо пропускаются истинные сайты, либо в предсказание включается большое число ложных сайтов. Следует отметить, что в рассматриваемом случае вообще не совсем ясно, что такое ложный сайт, поскольку известно, что мутации, блокирующие сплайсинг в истинном сайте, часто приводят к оживлению молчавшего ранее расположенного рядом скрытого (cryptic) сайта (особенно много примеров такого рода известно для гена и-глобина человека). Как бы то ни было, полученное в результате множество вероятных сайтов не может автоматически рассматриваться как задающее истинные границы экзонов.

В основном программы распознаваия генов используют «комбинированный» подход, состоящий в следующем. Сначала отбирается множество «возможных экзонов», затем среди множества «возможных генов», т. е. цепочек возможных экзонов, отыскивается оптимальный. При этом функция качества возможного гена учитывает как «качество» возможных сайтов сплайсинга, так и кодирующий потенциал гена (обзор см. в [6]). В настоящее время пользователю доступны множество программ и электронных серверов, предсказывающих отдельные экзоны (точнее, транслируемые экзоны или транслируемые части экзонов) или целые гены. Качество распознавания для лучших из них, понимаемое как средний коэффициент корреляции между предсказанными и истинными генами, не превышает 70%

и нет оснований надеяться, что оно может быть существенно улучшено без существенного продвижения в понимании молекулярно-биологических механизмов сплайсинга.

Другой проблемой является то, что методы, основанные на применении нейронных сетей или алгоритмов распознавания образов, используют сложные статистические параметры и поэтому нуждаются в больших обучающих выборках, состоящих из хорошо описанных последовательностей. Такие выборки доступны при работе с традиционными геномами (млекопитающих и птиц, в меньшей степени нематоды Caernorabditis elegans, возможно, дрозофилы), однако они отсутствуют для многих важных геномов холоднокровных, многих беспозвоночных, растений.

В большинстве существующих алгоритмов задача поиска наилучшего из возможных генов интерпретируется как задача поиска оптимального пути в ориентированном графе и использовании метода динамического программирования. Ребра в модельном графе соответствуют возможным экзонам, а объединение ребер в путь — образованию экзонной структуры. Однако, применение метода динамического программирования в его классической форме оказалось невозможным, в частности, потому, что критерием оптимальности в задаче являются не суммарные характеристики, а средние веса доноров, акцепторов и используемых кодонов.

Рассмотрим алгоритм в котором «весами» путей являются вектора вида

$$(A,\, D,\, N,\, C,\, L) \tag{3}$$

где A - суммарный вес акцепторных сайтов (напомним, что путь в модельном графе соответствует цепочке экзонов), D — суммарный вес донорных сайтов, N — количество экзонов, C — суммарный кодонный потенциал, L — длина закодированного белка. Несколько упрощая ситуацию, можно сказать, что структура $S1$ с весом $(A1,\, D1,\, N1,\, C1,\, L1)$ «абсолютно лучше» структуры $S2$ с весом $(A2,\, D2,\, N2,\, C2,\, L2)$, если она лучше по всем 5 компонентам, т. е.

$$A1 > A2; \quad D1 > D2; \quad N1 < N2; \quad C1 > C2; \quad L1 < L2. \tag{4}$$

Алгоритм отсеивает такие структуры, которые абсолютно хуже некоторых других структур. Оставшееся множество и будет множеством структур — кандидатов в правильные структуры: если структура $S1$ абсолютно лучше структуры $S2$, то по любому «разумному» критерию, ориентированному на средние значения силы сайтов сплайсинга и кодирующего потенциала

структура $S1$ будет лучше структуры $S2$. Подчеркнем, что на этом этапе нам не важно — с какими коэффициентами учитываются указанные средние. Далее, отобранные структуры — кандидаты подвергаются дополнительному отсеву и оставшиеся структуры ранжируются — уже в соответствии с конкретной целевой функцией $f(A, D, N, C, L)$. Этот подход реализован в программе GREAT [7]. Алгоритм уверенно распознает кодирующие сегменты в ДНК человека. С вероятностью близкой к 100%, среди небольшого числа предсказанных сегментов находится заданное количество истинно кодирующих. Простота статистической базы алгоритма позволяет использовать его при анализе слабо изученных геномов, для которых нет возможности составить большие обучающие выборки.

Современные программы распознавания белок-кодирующих областей в геномах эукариот обеспечивают корреляцию между истинным и предсказанным геном на уровне около 70%. В то же время, оказывается, что в гене почти всегда есть участки, которые можно предсказать достоверно. Подход к поиску этих участков основан на простом наблюдении над результатами программы GREAT: сегменты, присутствующие во многих субоптимальных потенциальных генах, как правило являются кодирующими.

Это дает возможность вместо расплывчатого распознавания полных генов, надежно предсказывать белок-кодирующие области, которые будут использованы для конструирования олигонуклеотидных зондов, праймеров ПЦР и т. д. Показано, что 30-нуклеотидный кодирующий сегмент в последовательности ДНК человека можно распознать с 100%-ной надежностью, а для того, чтобы иметь два таких сегмента с промежутком между ними не менее заданного, достаточно рассмотреть пять кандидатов (при этом в 95% случаев достаточно ровно двух кандидатов).

Это наблюдение формализовано в следующем алгоритме. Сначала с использованием алгоритма векторного динамического программирования строится множество субоптимальных потенциальных генов. В отличие от других алгоритмов, этот позволяет применять нелинейные распознающие функции и за счет этого добиваться хорошего распознавания с использованием простых статистических параметров. Затем множество потенциальных генов используется для вычисления вероятностей кодирования для отдельных сегментов; при этом используется известная двойственность между поиском оптимального пути на графе и вычислением статистической суммы весов путей. Данный подход лег в основу создания программы CASSANDRA [8–10].

Нельзя не упомянуть о новом, успешно развиваемом в настоящее время подходе к проблеме определения кодирующих областей. Этот подход представлен в программе GeneScan [11], где используется метод Фурье-преобразования. Метод распознавания основан на том, что кодирующие районы ДНК во всех организмах имеют периодичность одинаковой длины. Этот метод хорошо себя показал на предсказании генов прокариот (точность предсказания около 90%), однако, при предсказании экзон-интронных границ эукариот применение его встретило определенные трудности.

Вероятностная модель программы GeneScan используется и программе GenomeScan [12] предназначенной для распознавания экзон-интронной структуры генов эукариот. В программе также учитывается множество особенностей структуры гена: типичное число экзонов в гене, распределение длин экзонов и интронов, информация о GC содержании, характерном для кодирующиющих и некодирующих областей различных организмов (в основном, человека и позвоночных) и т. д. При предсказании генов также привлекается информация о гомологичных белках.

6.4. Современные задачи сравнительного анализа биологических последовательностей, предпосылки для применения параллельных вычислений

Значение разработки новых высокоэффективных методов анализа биологических последовательностей особенно возрастает в настоящее время в связи с осуществлением проектов тотального секвенирования геномов различных организмов и совершенствованием методов определения первичных структур белков. Так, в настоящее время (начало 2002 г.) полностью расшифрована последовательность нуклеотидов в геномной ДНК нескольких десятков организмов. Практически завершена работа по расшифровке генома человека (около 3 миллиардов нуклеотидов), известна значительная часть генома мыши.

Ниже представлены новые методы сравнительного анализа первичных структур биополимеров и примеры применения этих методов для решения актуальных задач молекулярной биологии. Для белков — это задача определения типа пространственной структуры белка по его первичной структуре, для нуклеиновых кислот — задача иерархического анализа протяженных участков ДНК (в том числе — полных геномов), поиск функционально значимых (в частности, белок-кодирующих) областей.

В случае сравнения белков рассматривается (совместно с Институтом молекулярной биологии и EMBL Germany) [13, 14] задача распознавания пространственной структуры белка путем сравнения его аминокислотной последовательности с аминокислотными последовательностями белков, для которых пространственная структура была определена экспериментально.

Эта задача допускает естественное распараллеливание, т. к. сравнивать тестовый белок с различными белками с известной пространственной структурой можно независимо и, следовательно, одновременно. То есть, в принципе, имеет смысл одновременное выполнение сотен заданий.

В следующем разделе рассказывается о результатах исследования достоверности выравниваний первичных последовательностей белков (по отношению к «эволюционно-правильному» выравниванию, полученному сравнением пространственных структур). Такое исследование является необходимым предварительным этапом при разработке методов распознавания пространственной структуры белка по его аминокислотной последовательности.

В результате данных исследований выведена зависимость надежности восстановления выравнивания пространственных структур по последовательностям от степени сходства самих последовательностей. Исследована зависимость качества восстановления структурного выравнивания от параметров алгоритма выравнивания методом Смита-Уотермана. Показано, что наилучшие результаты получаются для матрицы Gonnet, при этом в 49% случаев выравнивание со штрафом за открытие делеции 12 является наилучшим среди выравниваний, которые могут быть получены с какими-либо значениями этого параметра, а в 29% случаев оно отличается от наилучшего из возможных не более чем на 5%. В то же время показано, что метод выравнивания Смита-Уотермана обладает недостатками, которые не могут быть устранены путем подбора значений параметров. Это связано с тем, что естественный отбор по-разному действует на разные участки белка, а метод Смита – Уотермана предполагает использование всюду одних и тех же значений параметров.

В случае выравнивания геномной ДНК следует упомянуть разработанный совместно с группой А. С. Кондрашова, (NCBI USA) [15] новый подход к задаче выравнивания геномов высших эукариот, то есть организмов, чьи клетки имеют ядра. Для таких организмов (в отличие от бактерий) характерно наличие протяженных (сотни миллионов нуклеотидов) участков, в которых сходные гены следуют в одном и том же порядке в разных орга-

низмах. Эти участки, т. н. области синтении, и являются объектом нашего изучения.

Предложенный подход ориентирован на достаточно распространенную ситуацию, когда в областях синтении участки достаточно сильного сходства между сравниваемыми геномами, чередуются с участками, в которых сходство отсутствует (именно так обстоят дела, например, для пары «человек» — «мышь»). То есть, глобальное выравнивание синтенных участков геномов представляется как цепочка локальных сходств.

Как правило, конфликт между двумя локальными сходствами (т. е. ситуация, когда 2 сходства не могут быть одновременно включены в итоговое глобальное выравнивание) может быть разрешен исключительно на основе рассмотрения этих сходств, не прибегая к оптимизации какой-либо глобальной весовой функции. Это наблюдение (кардинально отличающее задачу выравнивания геномов от рассмотренной выше и более традиционной задачи выравнивания белков) позволило предложить иерархический подход к построению геномных выравниваний. Эксперименты с пробной реализацией этого подхода показали, что он работает существенно быстрее, чем существующие программы, основанные на глобальной оптимизации, а его результаты биологически оправданны.

В рамках предложенного подхода сравниваемые области синтении динамически делятся на фрагменты, которые обрабатываются иерархически независимо друг от друга.

Использование параллельных вычислений позволяет одновременно обрабатывать участки, находящиеся на одном уровне иерархии. Таким образом, общее время вычислений оказывается пропорциональным количеству уровней иерархии, а не количеству образовавшихся фрагментов, что означает многократное ускорение вычислений по сравнению с «непараллельным» алгоритмом.

6.5. Исследование достоверности выравнивания аминокислотных последовательностей

6.5.1. Источник структурно адекватных выравниваний

Тестовую выборку составили множественные структурно адекватные выравнивания белковых доменов, представленные в базе BAliBASE.Данная база доступна через Internet по адресу:

http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/.

В базе представлены как белки, содержащиеся в банке PDB, так и еще не вошедшие в него (что предполагает отсутствие общедоступной информации о пространственной структуре). Для тестов взята группа выравниваний, в которых от исходных последовательностей были «отрезаны» концевые фрагменты. Всего в базе содержится 23 семейства (множественных выравниваний) с длинами последовательностей от нескольких десятков до нескольких сотен аминокислотных остатков, каждое семейство составлено примерно из полутора десятков доменных последовательностей. Уровень их сходства (%ID) варьируется от нескольких процентов до ~80 процентов. На рис. 6.3 показано распределение выравниваний в базе BAliBase по %ID.

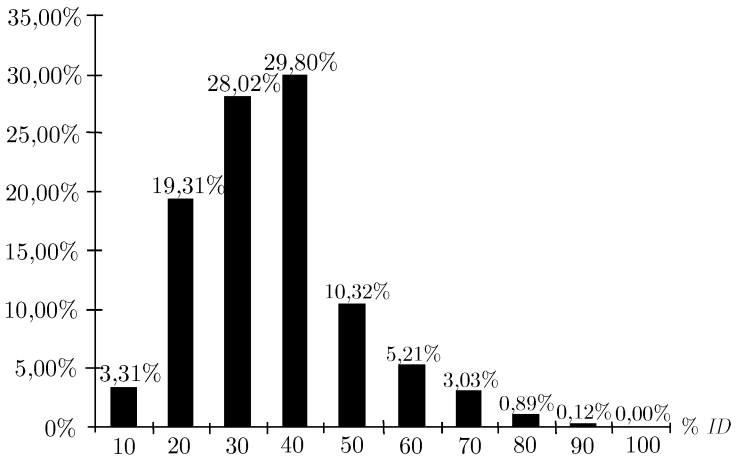

Рис. 6.3. Распределение выравниваний в базе BAliBase по %ID

6.5.2. Мера сходства последовательностей

Степень похожести двух последовательностей называют уровнем гомологии. Его можно охарактеризовать двумя величинами:

1. % ID (identity — идентичность)=число совпадений / число сопоставлений, т. е. процент совпадающих букв среди всех сопоставляемых в структурно верном выравнивании аминокислотных последовательностей белков (делеция сопоставлением, естественно, не считается).

2. S-фактор = вес выравнивания / max (вес выравнивания последовательности самой на себя), т. е. отношение веса выравнивания (по матрице аминокислотных замен) к максимально возможному весу (максимум из двух весов выравниваний последовательностей самих с собой).

Была выяснена взаимосвязь этих величин. Рис. 6.4 отображает связь между S-factor и % ID для случайных последовательностей. (Модификация консенсуса профиля Igb4 с заданным уровнем гомологии, матрица замены Blosum62).

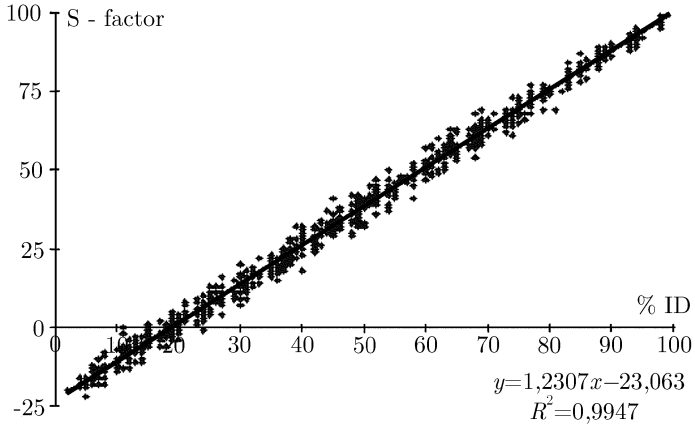

Рис. 6.4. Взаимосвязь между S-factor и % ID.

Подобные результаты (зависимость линейная с небольшим разбросом точек) были получены и на реальных белках. Следовательно, можно считать % ID и S-фактор эквивалентными мерами гомологии. В дальнейшем в качестве основной меры используется % ID.

6.5.3. Мера сходства выравниваний. Понятие «острова»

Для сравнения парных выравниваний использовалась следующая мера (Aln_Sim% — alignment similarity) — процент совпадающих сопоставлений в этих выравниваниях по отношению к общему числу сопоставлений в том из них, которое мы считаем «эталонным» (структурно адекватным).

На рис. 6.5. представлены два выравнивания белков 1tgx и 1era: структурное *а*) и последовательностное *б*). В структурном выравнивании *а*) 58

сопоставлений, из них в выравнивании *б*) присутствуют 42 (указаны звездочками).

Таким образом сходство выравниваний:

$$\text{Aln\_Sim}\% = 42/58 = 0,72 = 72\%.$$

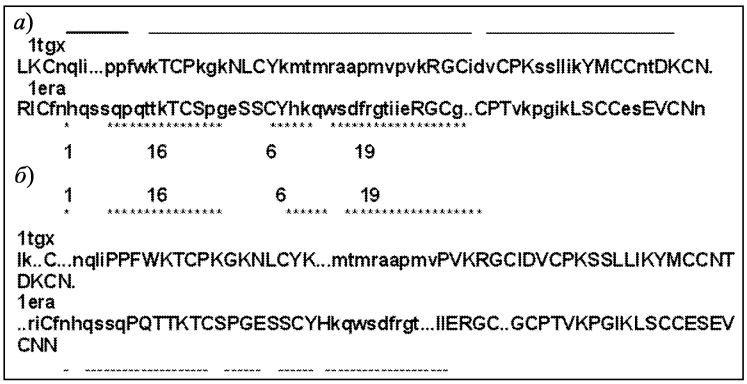

Рис. 6.5. Два выравнивания белков 1tgx и 1era: структурное *а*) и последовательностное *b*). Для более детального рассмотрения выравнивания введем понятие «острова». «Остров» — это участок выравнивания без делеций. Структурное выравнивание *а*) содержит 3, а выравнивание *б*) — 5 «островов»

6.5.4. Зависимость степени сходства структурного и последовательностного выравнивания от степени сходства исследуемых белков

Возникает вопрос, при каком уровне гомологии двух последовательностей реально предполагать, что они образуют хорошо сопоставимые структуры в пространстве. И может ли программа выравнивания аминокислотных последовательностей дать выравнивание, сколь либо хорошо совпадающее со структурно верным выравниванием, и насколько хорошим может быть это совпадение?

Для всех 23 семейств, содержащихся в базе данных Bali Base, были построены Парето множества выравниваний для всех пар белков семейства.

Из полученных Парето-множеств были выбраны выравнивания, имеющие максимальный Aln_Sim% (субадекватные выравнивания). На рис. 6.6 представлена зависимость Aln_Sim% субадекватного выравнивания от % ID.

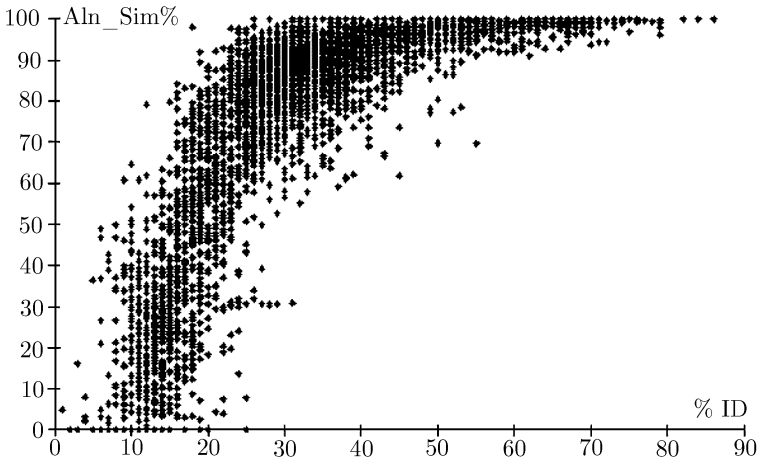

Рис. 6.6. Зависимость угаданности структурно верного выравнивания от % ID

Результат: чем больше гомология последовательностей, тем больше вероятность для них обладать похожей структурой, и, следовательно, построить структурно верное выравнивание. Выравнивание близкое (AlnSim%>50%) к структурно верному выравниванию можно построить при %ID>20%.

6.5.5. Детальное изучение выравниваний. Угаданые «острова»

Для детального исследования, того что можно угадать с помощью выравнивания, проверялось как хорошо угадываются «острова». «Остров» — это безделеционный участок в выравнивании. «Остров» эталонного выравнивания считается угаданным, если в построенном последовательностном выравнивании присутствует хотя бы одно такое же сопоставление.

Мера, отражающая степень угаданности «острова» (Isl_Sim%), равна отношению числа угаданных сопоставлений к общему количеству сопоставлений, присутствующих в эталонном «острове».

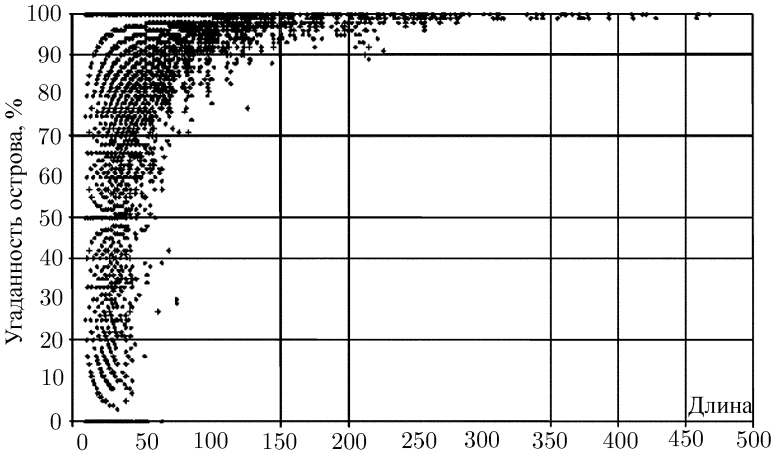

Рис. 6.7. Зависимость процента угаданности «острова» от его длины

Диаграмма угаданности острова в зависимости от его длины представлена на рис. 6.7. На диаграмме видно, что при длине острова больше 50 он угадывается, что согласуется с общепринятым мнением, что такой остров «правильный».

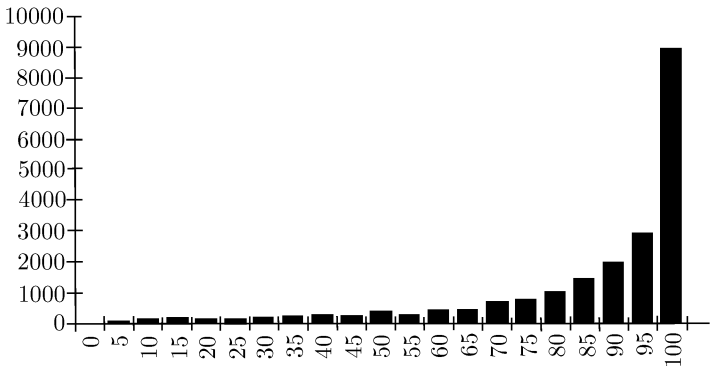

Рис. 6.8. Распределение «островов» по проценту угаданности в субадекватных выравниваниях

На гистограмме рис. 6.8 показано распределение «островов» по проценту угаданности. Видно, что «островов», угаданных более чем на 90%, гораздо больше половины. Получается, что если остров угадан, то он угадан почти целиком. Всего потерянных островов (в которых не угадано ни одного сопоставления) около трети. Они имеют длину менее 50 символов и небольшой или даже отрицательный вес (рис. 6.9).

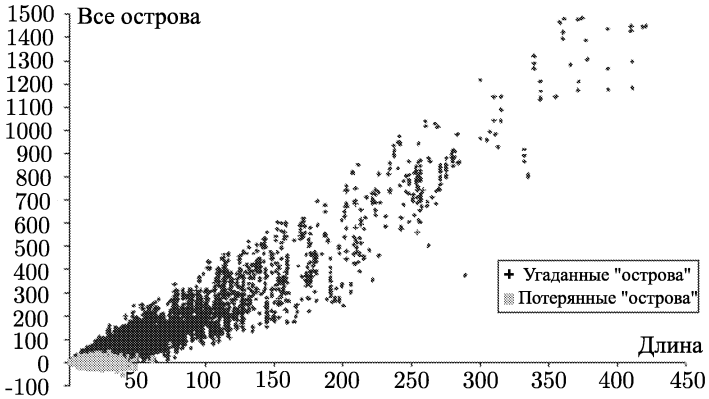

Рис. 6.9. Зависимость веса «острова» от его длины

Предполагаемая причина потери «островов» — малый вес. Это может быть связано со следующими обстоятельствами:

а) плохая матрица — т. е. эта матрица дает малый или отрицательный вес функционально значимым островам, возможно надо использовать разные матрицы для каждого отдельного острова;

б) различие между структурным и эволюционным выравниванием.

Работа выполнена при поддержке грантов РФФИ (№№ 94–04–12330, 97-04-12330, 00-04-48246) и подпрограммы «Геном человека».

Литература

[1] Ройтберг М. А. Новый подход к проблеме выравнивания последовательностей: больше совпадений, меньше удалений и никаких весовых коэффициентов // Труды конференции «Геном человека — 93», Черноголовка, март 10–12, 1993. — М., 1993, с. 135

[2] Ройтберг М. А. Парето-оптимальные выравнивания символьных последовательностей. — Препринт. — Пущино: ОНТИ НЦБИ, 1994, 10 с.

[3] Ройтберг М. А., Семионенков М. Н., Таболина О. Ю. Парето-оптимальные выравнивания биологических последовательностей // Биофизика, 1999, № 3, с. 20–37

[4] Gelfand M. S., Mironov A. A., Pevzner P. A. Gene recognition via spliced sequence alignment. 1996 // Proc. Natl. Acad. Sci. USA, 1997, v. 93, p. 334–339

[5] Mironov A. A., Roytberg M. A., Pevzner P. A., Gelfand M. S. Performance guarantee gene predictions via spliced alignment // Genomics, 1998, v. 51, p. 332–339

[6] Гельфанд М. С. Компьютерный анализ последовательностей ДНК // Молекулярная биология, 1998, т. 32, с. 103–120

[7] Gelfand M. S., Podolsky L. I., Astakhova T. V., Roytberg M. A. Recognition of genes in human DNA sequences // Journal of Computational Biology, 1996, v. 3, № 2, p. 223–234

[8] Ройтберг М. А., Астахова Т. В., Гельфанд М. С. Алгоритм высокоспецифичного распознавания белок-кодирующих областей в последовательностях высших эукариот // Молекулярная биология, 1997, т. 31, с. 25–31

[9] Roytberg A., Astahova T. V., Gelfand M. S. Combinatorial approaches to gene recognition // Computers and Chemistry, 1997, v. 1, № 21, p. 229–236

[10] Цзе Синг-Хой, Ройтберг М. А., Гельфанд М. С., Миронов А. А., Астахова Т. В., Певзнер П. А. Algorithms and software for support of gene identification experiments // Bioinformatics, 1998, v. 1, № 14, p. 14–19

[11] Tiwari S., Ramachadran S., Bhattacharya S., Bhattacharya A., Ramaswamy R. // CABIOS, 1997, v. 13, p. 263–270

[12] Yeh R. – F., Lim L. P., Burge C. B. // Genome Res., 2001, v. 11, p. 803–816

[13] Олейникова Н. В., Богопольский Г. А., Власов П. К., Сюняев Ш. Р., Ройтберг М. А. Accuracy of the pairwise protein sequence aligment: From the observations to a new approach, Artificial Intelligence and Heuristic

Methods for Bioinformatics // NATO Advanced Studies Institute, 2001, p. 49

[14] Богопольский Г. А., Власов П. К., Олейникова Н. В., Ройтберг М. А., Сюняев Ш. Р. Определение сходства пространственных структур белков на основе сопоставления их аминокислотных последовательностей. Научный совет Подпрограммы «ГЕНОМ ЧЕЛОВЕКА», Сборник отчетов за 2000 г. с. 145.

[15] Шабалина С. А., Огурцов А. Ю., Кондрашов А. С. Иерархический подход к выравниванию коллинеарных участков геномов. Научный совет Подпрограммы «ГЕНОМ ЧЕЛОВЕКА», Сборник отчетов за 2001 г. с. 57.