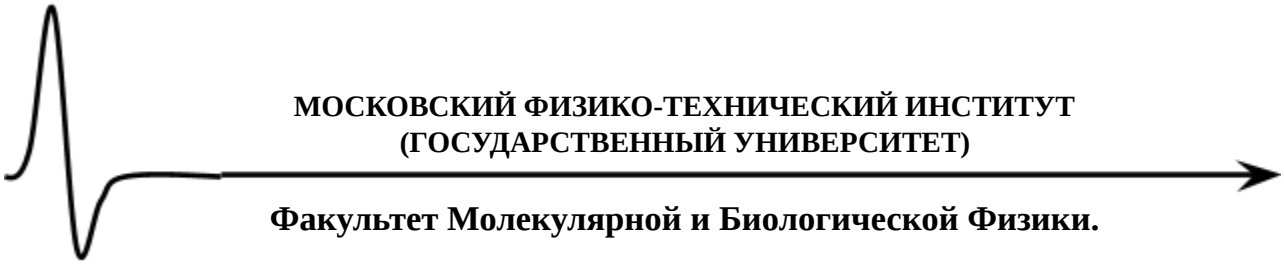


МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)

Факультет Молекулярной и Биологической Физики.

“Выравнивание аминокислотных последовательностей белков и выравнивание их пространственных структур”

Магистерская диссертация

студентки 492 гр.
Олейниковой Н.В.

научный руководитель:
к.ф.-м.н. Ройтберг М.А. (ИМПБ РАН).

Рецензент:
к.ф.-м.н. Макеев В. Ю. (ИМБ РАН)

Заведующий кафедрой
ак. Мирзабеков А.Д

- Долгопрудный -
- 2000 -

СОДЕРЖАНИЕ:

ВЕДЕНИЕ.

Компьютерная революция и революция в молекулярной биологии произошли примерно одновременно, весьма удачно для тех, кто был занят анализом последовательностей ДНК и белка. Усовершенствование методов секвенирования белков и нуклеиновых кислот привело к лавинообразному росту числа расшифрованных структур этих биомолекул. Для систематизации и упорядочения имеющейся информации, а также обеспечения возможности оперативной работы с последовательностями созданы специальные банки данных, в которых хранится информация о первичной и пространственной (для белков) структуре молекул, сведения об организмах, из которых выделены данные последовательности и т.п. Нужны алгоритмы, позволяющие выделять характерные черты семейств и групп белков и на основе выделенных признаков (определённые участки последовательности, консервативные позиции и т.п.) производить поиск по всей базе, определяя белки с аналогичными чертами.

Наиболее мощным и биологически обоснованным инструментом для анализа сходств первичных структур является выравнивание цепочек символов и вычисление некоей меры по результирующему выравниванию. При изучении биологических последовательностей требуется строить выравнивания, сопоставляющие друг другу функционально и/или эволюционно близкие участки. Нас будут интересовать выравнивания, отражающие сходство пространственных структур.

Цель работы – понять, насколько стандартный метод выравнивания последовательностей, предложенный Smith T.F. and Waterman M.S. [1], позволяет восстановить выравнивание пространственных структур, как выбирать параметры в этом методе.

1. ОБЗОР ЛИТЕРАТУРЫ.

1.1 Задача выравнивания последовательностей.

Выравнивание последовательностей - традиционный способ сравнения первичных структур биополимеров. При изучении биологических последовательностей, требуется построить выравнивание, сопоставляющее друг другу функционально или эволюционно близкие участки.

Выровнять две последовательности - значит расположить их друг над другом, при этом между некоторыми символами могут быть вставлены пробелы, так чтобы получившиеся подпоследовательности имели равную длину [2]. Символы, расположенные в полученных последовательностях на одинаковых позициях, называются сопоставленными друг другу. Сопоставленные символы могут быть одинаковыми (образуют «совпадение») или нет (образуют «несовпадение»). Возьмем, например, две небольшие последовательности аминокислот: V_1 = MQFLAVSTKKCA и V_2 = MRAVSNKKCALK. Выравнивание:

V_1 - **Mqfl**AVS**t**KKCA--

V_2 - **Mr--**AVS**n**KKCA**lk**

В этом выравнивании 8 совпадений, 2 замены и 4 удаленных символа. Необходимо отметить, что при построении выравниваний невозможно различать между собой вставки и делеции символов. В рассмотренном примере вставку ***FL*** в последовательности V_1 можно рассматривать и как удаление этого символа из последовательности V_2. На рисунке 1 дано графическое изображение выравнивания двух последовательностей.

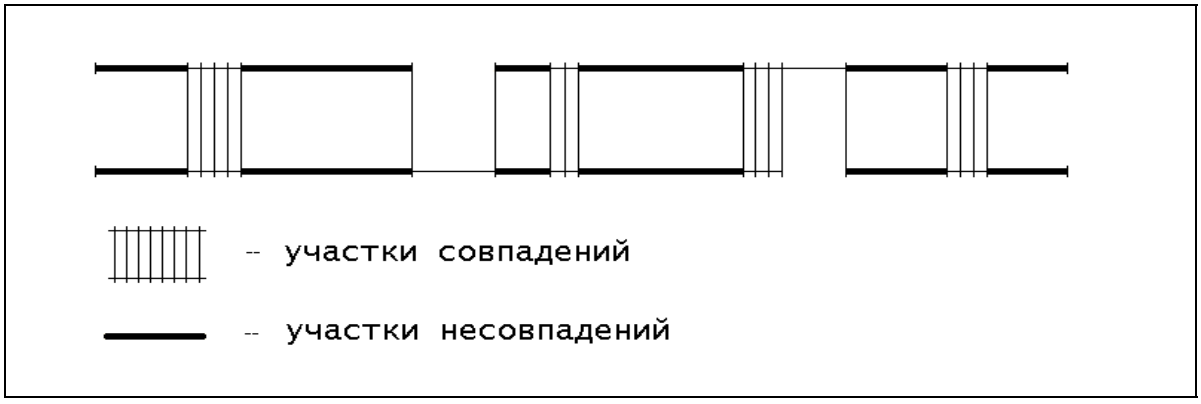


Рис. 1. Пример выравнивания двух последовательностей.

Выравнивания бывают глобальные и локальные. В глобальных

выравниваниях последовательности сопоставляются от начала до конца, а в локальных – концевые участки последовательностей можно не сопоставлять.

1.2 Метод динамического программирования в задачах выравнивания

Для оценки качества различных выравниваний вводится понятие «веса» выравнивания (alignment score). Вес W выравнивания двух последовательностей определяется как число совпадающих символов в этих последовательностях минус суммарный штраф за замены и вставки символов:

$$W(\mu,\delta,\sigma) = k_m\mu - k_s\delta - k_{id}\sigma ,$$

где k_m - число совпадений, k_s - число замен, k_{id} - количество вставок/делеций; а μ - стоимость совпадения сопоставленных символов в выравнивании, δ - штраф за замену, σ - штраф за вставку/делецию одного символа.

Первый алгоритм для выравнивания двух биологических последовательностей был предложен Нидльманом и Вуншем [3]. Их алгоритм, основанный на методе динамического программирования, заключался в построении оптимального выравнивания, т.е. выравнивания, имеющего максимальное значение весовой функции:

$$W^* = \max_{\text{all alignments}} W .$$

Весовую функцию W иногда называют функцией сходства или функцией подобия.

При использовании алгоритмов динамического программирования для сравнения биологических последовательностей одной из наиболее значительных трудностей является выбор параметров алгоритма, а именно: штрафов за замены и делеции символов. На практике вместо пары параметров μ и δ используется симметричная квадратная матрица весов сопоставлений символов S_{N*N}, где размерность матрицы N определяется размерностью алфавита (в случае ДНК $N = 4$, в случае белков $N = 20$). Пример: фрагмент матрицы замен BLOSUM62 [4]:

	A	L	V	W
A	4	-1	0	-3

L	-1	5	1	-2
V	0	1	4	-3
W	-3	-2	-3	11

Матрицы весов сопоставлений (матрицы замен), как правило, вычисляются по частотам замен символов в среднем по банку структурных выравниваний белковых семейств, то есть в основе этих матриц косвенно лежат физико-химические свойства аминокислот. Очевидно, что такая матрица зависит от анализируемого банка и от метода оценки частот замен. На сегодняшний день существует несколько семейств матриц замен аминокислот, наиболее известные из них: PAM, BLOSUM, GONNET [5]. Матрицы в этих семействах получены исходя из частот замен аминокислот в гомологичных фрагментах эволюционно близких белков; различные матрицы в серии различаются степенью близости использованных фрагментов.

Другой способ составления матриц замен следует из анализа физико-химических свойств аминокислот. Например, боковая цепь триптофана (W) содержит цикл и имеет огромные размеры по сравнению с боковыми цепями аланина (A) или глицина (G). Мутация триптофана на эти остатки приводит к стерическим затруднениям, следовательно, соответствующие веса замены должны быть сильно отрицательными. Таким образом, чем критичнее для пространственной структуры замена, тем более отрицательный вес она имеет в матрице, и наоборот. Сравнение матриц, вычисленных по банкам структурных выравниваний, и матриц, составленных на основе физико-химических свойств аминокислот, показывает, что друг от друга они отличаются незначительно.

Матрица замен как один из параметров выравнивания имеет обоснованную биологическую мотивацию. Хуже обстоит дело со штрафом за удалённый символ, в том смысле, что нет никаких разумных доводов в пользу того или иного значения этого параметра. С точки зрения эволюционных событий, наличие вставки или делеции длиной в несколько символов соответствует единичному событию, причем длинные делеции для родственных последовательностей менее вероятны, чем короткие, поэтому логично ввести штраф за множественную делецию, как функцию от ее длины:

$$\sigma(l) = A + B \cdot l,$$

где ***A*** – штраф за открытие, а ***B*** – штраф за продолжение множественной делеции.

Такая схема называется аффинным штрафом за вставку/делецию. В некоторых специфических задачах используются другие схемы штрафования множественных делеций: $\sigma(l) = A + B\log(l);\quad \sigma(l) = A + Bl^C$. В любом случае пользователю приходится подбирать подходящие значения для параметров ***A*** и ***B***. Никаких точных биологически оправданных оценок на эти параметры нет [6]. Известно только то, что для разных матриц замен надо использовать разные штрафы и что для разных семейств параметр ***A*** варьируется

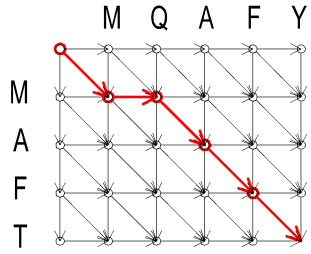
В некоторых случаях небольшие изменения весов аминокислот или штрафной функции вставок/делеций приводят к существенным изменениям в результирующих выравниваниях. В других случаях выравнивания остаются устойчивыми к изменениям алгоритмических параметров. Однако, единого множества "правильных" параметров не существует [7, 8]. Подход, впервые предложенный в [9], лишен этих недостатков, но требует дополнительного анализа результата.

Выравнивание Смита – Уотермена.

Даны две последовательности символов α_1 , α_2 длиной *m* и *n* соответственно, матрица замен $S = \{s_{ij}\}_{i,j=1}^{k}$ и штрафная функция $\sigma(l) = A + B \cdot l$. Требуется найти выравнивание ***V*** с максимальным весом $W(V, S, \sigma(l))$.

Задача максимизации ***W*** и нахождения оптимального выравнивания решается методом динамического программирования [1].

Рис. 2 Граф Ниделмана – Вунша соответствующее ему выравнивание.



MQAFY
M-AFT

Граф Ниделмана – Вунша [2] - множество вершин $V = \{v_{ij}\}; i = \overline{1,m}; j = \overline{1,n}$. Паре сопоставленных символов (α_i, α_j) ставится в соответствие вершина v_{ij} .

Множество ориентированных дуг $L = \{l_k\}; k = \overline{1,3 \cdot (m+1)(n+1)}$. Вес дуги определяется следующим образом: все вертикальные и горизонтальные дуги имеют

одинаковый вес *B*, вес диагональной дуги равен весу сопоставления двух символов, соответствующих вершине, в которую входит данная дуга.

Матрица связности: $C = \{c_{ij,pq}\}; i = \overline{0,m}; j = 0,n; p = \overline{0,m}; q = \overline{0,n}$

$$c_{ij,pq} = \begin{cases} 1, p = i+1, q = j+1 \vee p = i+1, q = j \vee p = i, q = j+1 \\ 0, \text{otherwise} \end{cases}$$

то есть множество дуг можно определить так:

$$L = \{v_{ij} \to v_{i+1\,j}\} \bigsqcup \{v_{ij} \to v_{i\,j+1}\} \bigsqcup \{v_{ij} \to v_{i+1\,j+1}\}; \; i = \overline{0,m}; j = \overline{0,n}$$

Глобальному оптимальному выравниванию соответствует путь на графе от вершины $V_{0,0}$ к вершине $V_{m,n}$ с максимальным весом. Очевидно, что таких путей может быть несколько.

Чаще бывает нужно найти локальное сходство между последовательностями. Принимая во внимание доменную организацию белков, глобальное выравнивание, например, не имеет смысла для двух многодоменных белков из одного суперсемейства, но из разных семейств. Схематически локальное выравнивание изображено на рисунке 3.

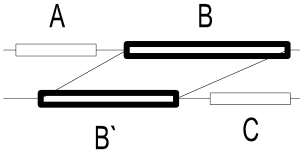


Рис. 3 Схематическое изображение локального выравнивания.

Найти локальное выравнивание - значит найти путь с максимальным весом в графе Нидлмана – Вунша такой, что: $v_{ij} \to \ldots \to v_{pq}$, где $0 \le i < p \le m; 0 \le j < q \le n$. Если существует несколько таких путей с одинаковым весом, то выбирается путь с наибольшим числом сопоставлений символов. Эта задача так же решается методом динамического программирования [8].

Парето-оптимальные выравнивания.

М.А. Ройтбергом [9, 10] был предложен многокритериальный подход 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 [11] 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

$$V(A)=(Comp(A), Gap(A)),$$

где Comp(A) – сумма весов сопоставлений в ***A***, а Gap(A) – число множественных делеций в ***A***.

Или**:**

$$V(A) = (Match(A), -Del(A)),$$

где Match(A) - количество совпадающих символов, Del(A) - количество удаленных символов в ***A***.

Определение 1. Пусть S_1 и S_2 - две последовательности длиной, соответственно, n и m; k >= 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.*

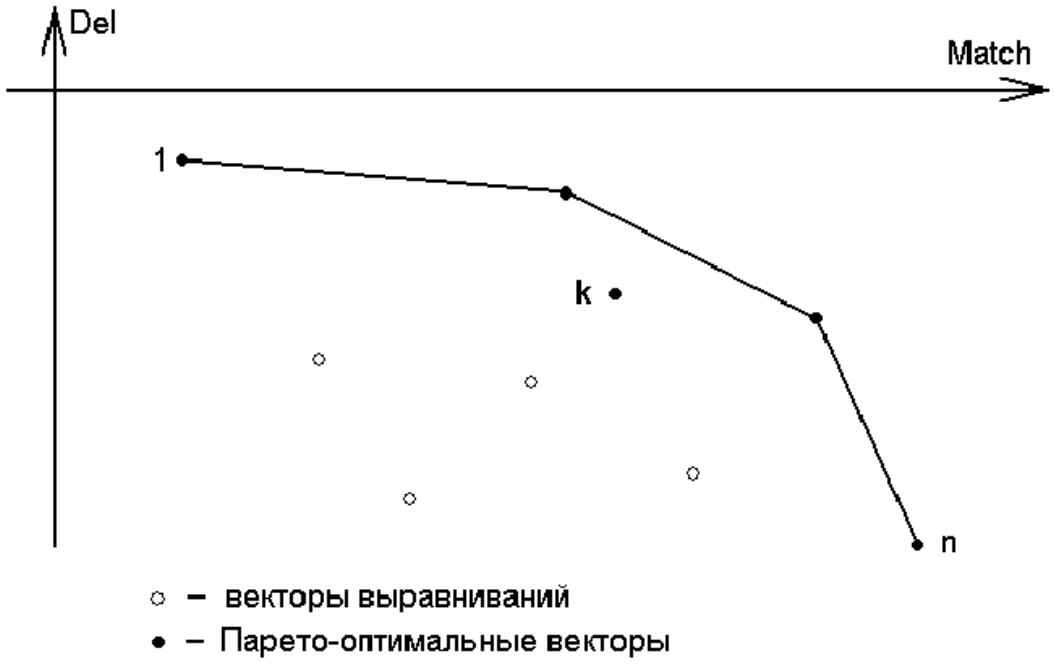


Рис. 4. Графическое изображение Парето-подмножества

Определение 3. Пусть M - множество k-мерных векторов. Вектор ***v***, лежащий во множестве M, называется *Парето-оптимальным* в M, если никакой другой вектор ***u*** из M не доминирует над ***v*** (рис.4). *Парето-оптимальное подмножество* в M - это подмножество, состоящее из всех векторов, Парето-оптимальных в M.

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 $V(A) = \{x_1(A), .., x_k(A)\}$ и скалярная весовая функция $W(A) = W(x_1(A),..,x_k(A))$, где функция $W(x_1, .., x_k)$ монотонно возрастает относительно каждого из аргументов $x_1, .., x_k$.

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) [12].

Пример: последовательность S_1 = MNTPFVCFIM

последовательность S_2 = MNAALTPSRFCFVV

Множество Парето-оптимальных выравниваний для весовой функции V(A)=(Comp(A),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	**MN...TP**F.V**CF**IM **MN**AAL**TP**SRF**CF**VV	35	2	3
4	**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 $m_* Match(A) - d_* Del(A)$ (вектор **k** на рис. 4.). 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 -Del(A) и максимальным значением компоненты Match(A) (векторы **1 - n** на рис. 4.).

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

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

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.

1.2 Молекулярно-биологические банки данных

Всего 20 лет назад знания о структуре генома на молекулярном уровне ограничивались информацией о порядке следования генов в геноме. Сегодня секвенирование нуклеиновых кислот, то есть определение генетической информации на более фундаментальном уровне, является основным инструментом биологических исследований. Секвенированные нуклеиновые кислоты охватывают диапазон всех известных функций: кодирование белков, кодирование структурной РНК и регуляторных участков ДНК, а так же матричной ДНК. Более того, секвенируется все больше и больше областей, функции которых пока не известны. С определением первичной структуры белков дело обстоит сложнее. Для подавляющего большинства белков, первичная структура получена по кодирующей ДНК. Этот способ имеет один серьезный недостаток. Дело в том, что после синтеза белок претерпевает пост трансляционную модификацию, так что реальный состав может отличаться от определенного по ДНК, однако для нескольких процентов белков имеются точные данные по первичной и даже по пространственной структуре.

Преимущества централизованного структурированного хранения данных по последовательностям очевидны. Формирование молекулярно-биологических банков данных началось в 1980 году. На сегодняшний день число таких банков уже приближается к сотне. Хранимая в них информация весьма разнообразна: от первичных структур до мутаций, связанных с заболеваниями. Условно банки последовательностей можно разделить на две группы: основные и производные. Основные банки

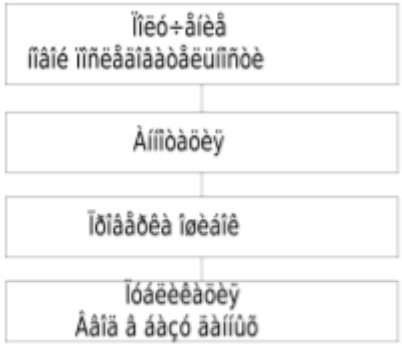
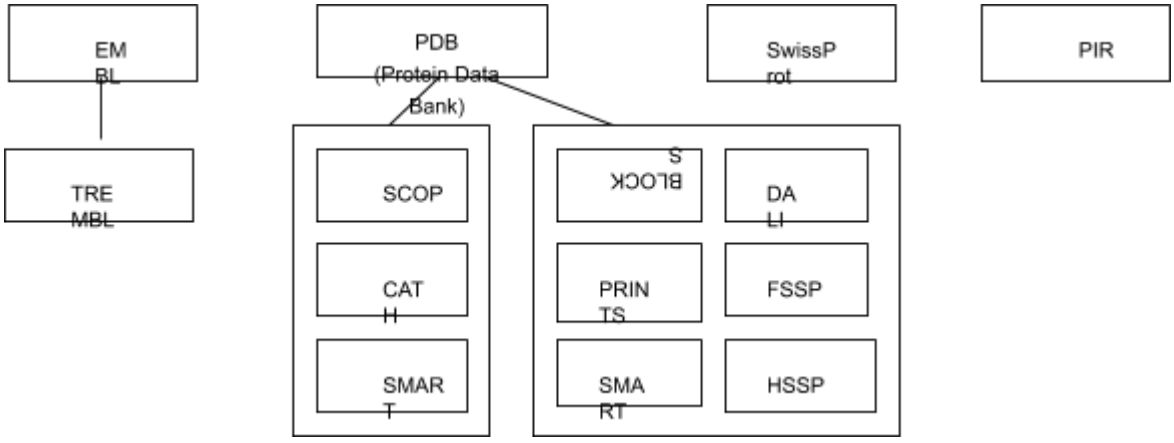


Рис.5 Схема поступления данных в первичный банк.

пополняются непосредственно данными молекулярно-биологических лабораторий (рис.5). Производные банки составляются из основных путем выборки данных по какому-либо критерию, но чаще производные банки – это результат обработки информации из первичного банка (например, множественные выравнивания). Наиболее известные первичные и их производные банки приведены на рис.6.

Рис.6 Некоторые первичные и их производные молекулярно биологические банки.



Для банков биополимеров характерно то, что они не сопровождаются системами управления базами данных (СУБД). Запросы пользователей к банкам последовательностей – это специфические задачи поиска, сортировки, группировки и фильтрации, и само понятие запроса в этой области отличается от традиционного понятия запроса в СУБД. Для извлечения полезной информации требуются специальные программы. Строго говоря, пользователи хотят получать выборку записей, связанных некими неявными отношениями. Эти отношения на последовательностях имеют как качественные, так и количественные характеристики и не представлены в базе данных явным образом. Для установления интересуемых отношений необходим нетривиальный анализ самих записей базы данных. Мы хотим выявить эволюционное родство между биополимерами. Эволюционное родство между предъявляемой в качестве запроса последовательностью и записями базы данных – это то, что интересует многих пользователей, хотя не единственное. В любом случае, всякий запрос, связанный с информацией, заключенной в последовательностях, подразумевает применение математических методов анализа последовательностей.

Примеры укладки полипептидной цепи в пространстве.

Имеющееся многообразие структур и функций белка - это результат длительного эволюционного процесса. Между всеми белками существует иерархическая древовидная родственная связь. Это обстоятельство учитывается при составлении банков белковых структур. Классификация белков включает четыре или пять (в зависимости от автора) уровней иерархии. Мы придерживаемся классификации SCOP [13], в которой присутствуют следующие уровни: а) класс, б) архитектура, в) суперсемейство, г) семейство, д) представитель.

Класс определяет преобладание того или иного элемента вторичной структуры. Существует четыре класса:

I. В основном, альфа (mainly alpha);

II. В основном, бета (mainly beta);

III. Поровну альфа и бета (alpha/beta);

IV. Нерегулярные структуры (irregular structures).

Архитектура определяется пространственной ориентацией элементов вторичной структуры друг относительно друга. Несколько примеров архитектур приведено на рисунках 7(а, б, в).

Рис. 7 а. ***4-Бета-пропеллер.***

В этой архитектуре 4 бета-структуры расположены так, что напоминают крылья пропеллера. Бета-структуры направлены от центра и повернуты на 90 градусов друг относительно друга. Каждый бета-слой сильно изогнут, так что угол между первым и последним листом может быть более 90 градусов.

Рис. 7 б. ***2-Солленоид***

Эта структура содержит два плоских бета листа, плотно упакованных друг против друга. Она имеет явную периодичность, что отличает ее от других структур. Цепь последовательно переходит от одного бета-листа к другому, что напоминает намотку провода в соленоиде. Концы бета-структур соединены альфа-спиральным участком. В литературе такая архитектура часто называется бета-спиралью.

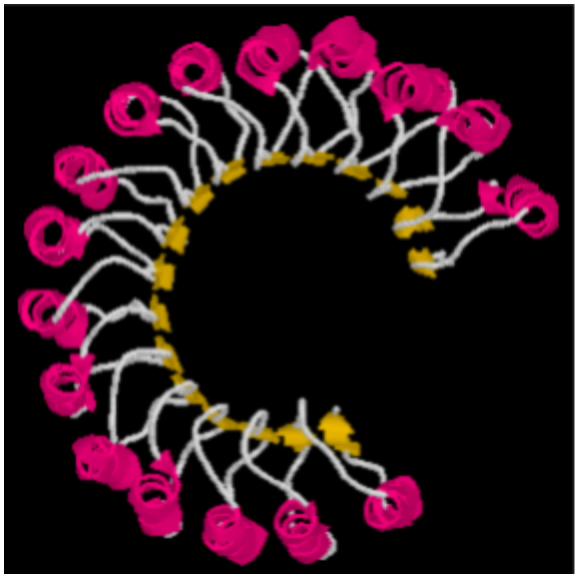


Рис. 7 в. ***Конская подкова.***

Пример высокой упорядоченности плотно упакованных бета-структур с антипараллельными альфа-спиралями снаружи.

Суперсемейство объединяет белки со схожими функциями, например, перенос

электронов осуществляется белками из суперсемейства цитохромов. Так как одна и та же функция может осуществляться несколькими способами,

суперсемейства решили разделять на семейства в зависимости от функциональной специфичности. Представители семейства – это белки с одинаковыми структурами и функциями, присутствующие в разных организмах, органах или клеточных тканях.

Использование банков биополимеров в задаче выравнивания.

Белок – это очень необычный биополимер, в том смысле, что только белку присуща уникальная укладка полипептидной цепи в пространстве, определяемая первичной структурой. В нашей работе кроме анализа первичной структуры мы пытаемся сопоставить пространственную структуру с последовательностью. Такое сопоставление можно сделать, например, через выравнивания.

Выравнивание – это некоторый биологически и математически осмысленный объект, получаемый из двух или более последовательностей с помощью определенной вычислительной процедуры или вручную. Выравнивания как такового не существует в природе (в живой клетке), скорее это плод деятельности исследователей. Выравнивание на уровне первичных структур – это сопоставление аминокислот одного белка с аминокислотами другого с возможным пропуском или вставкой некоторого числа не сопоставимых ни с чем аминокислот, как в первой, так и во второй последовательностях. Например:

LS**AS**Q**P**T___

LR**AS_P**YNHT

Совпадения на позициях выравнивания подтверждают эволюционное родство последовательностей. Несовпадения говорят о том, что на данных позициях в одном из белков произошли мутации (мутация - замена одной аминокислоты на другую). Пропущенные символы соответствуют ошибкам транскрипции.

Когда для белка известна его пространственная структура, например, по данным рентгеноструктурного анализа, то можно сделать выравнивание по пространственной структуре. В этом случае сопоставляются элементы вторичной структуры или отдельные атомы и группы атомов. С точки зрения физико-химических свойств белка, такое выравнивание более корректно, так как учитываются особенности строения конкретного белка. В связи с этим, структурные выравнивания можно принять за эталон для выравниваний по первичной структуре. Структурные выравнивания обычно проверяются и корректируются по многим

критериям. В дальнейшем будем называть структурные выравнивания структурно-верными, подчеркивая их биологическую адекватность.

1.3 Существующие пакеты программ для сравнения последовательностей.

Необходимость компьютерных программ сканирования банков и анализа последовательностей очевидна. Такие программы появились почти сразу после возникновения первых банков последовательностей, и сейчас их уже насчитывается порядка нескольких сотен. В основу каждой программы заложена та или иная математическая вычислительная идея. Часто эти программы используют схожие методы, например, выравнивание последовательностей. Успех и популярность программы в этой области, главным образом, зависит от качества поиска и распознания гомологов последовательностей, необходимых пользователю. Время работы, удобство интерфейса и т.п. важны в меньшей степени.

На сегодняшний день хорошо себя зарекомендовали следующие теоретические разработки:

- FASTA (Fast Alignment Search Tool)
- PSI-BLAST (Position Specific Iterated Basic Local Alignment Search Tool)
- SAM-T98: HMM Search (Hidden Markov Models)
- ISS (Intermediate Sequence Search)
- Bioccelerator

FASTA

Уже в конце 80-х готов банки разрослись до таких размеров, что поиск последовательностей был сопряжен с большими затратами времени и усилий специалистов. Программа поиска и сравнения последовательностей FASTA [14] – это одна из первых разработок в области поиска гомологий по банкам первичных структур биополимеров.

Эвристическая процедура сравнения последовательностей проста с точки зрения математики. Она состоит из 4-х этапов. Для сравнения пары последовательностей строится матрица сопоставления символов. По столбцам откладываются символы первой последовательности, а по строкам – символы второй. Элемент ***(i, j)*** отмечается знаком “+”, если ***i***-ый символ первой

последовательности равен ***j***-ому второй (рис. 8). В матрице отмечаются 10 диагональных сегментов, соответствующих точному совпадению длиной более чем ***ktup*** (***ktup*** – параметр, задаваемый пользователем).

	S	K	L	V
K		+		
L			+	
M				
S	+			
V				+
R				

Рис.8 Матрица сопоставления символов пары последовательностей.

Далее для каждого такого сегмента диагонали вычисляется вес по матрице замен. Программа пытается расширить каждый сегмент в обе стороны, если, конечно, расширение дает положительный прирост в весе. С помощью специального алгоритма отсеиваются сегменты малого веса так, что остаются только те сегменты, проекции которых не пересекаются (рис. 6 а). Для построения результирующего выравнивания из оставшихся сегментов применяется алгоритм динамического программирования, описанный в [8]. Результат сборки выравнивания из сегментов показан на рис. 6 б.

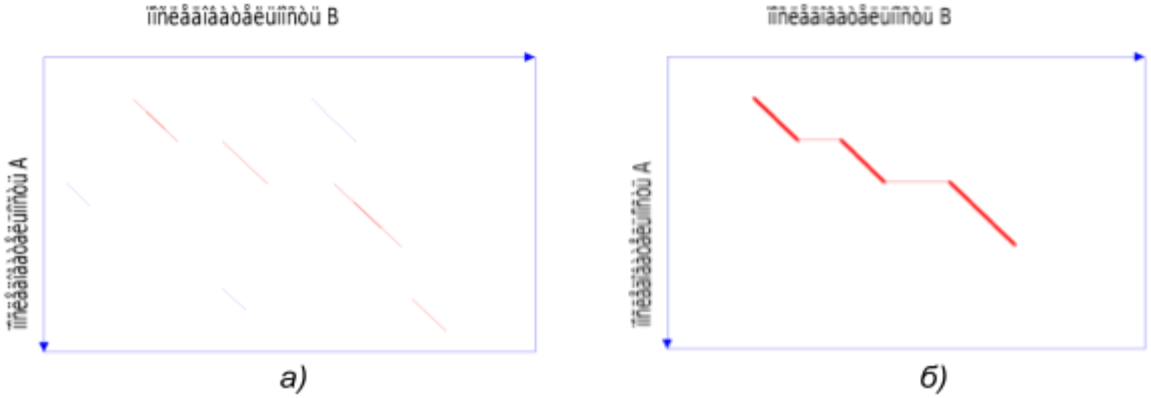


Рис.9 Последовательная сборка выравнивания с помощью расширения и склейки положительных сегментов матрицы сопоставления.

Для оценки значимости выравнивания применяется традиционная эмпирическая формула Z-score:

$$\hat{S}_i = \frac{O(S_i) - E(S)}{\sqrt{\hat{D}}}$$, где

$O(S_i)$ - наблюдаемый вес выравнивания последовательности запроса и ***i***-ой последовательности банка, *E(S)* - ожидаемый вес на случайных последовательностях, $\hat{D}$ - оценка дисперсии ожидаемого веса, вычисленная по выборке выравниваний случайных последовательностей. Предполагается, что все S_i распределены нормально, и формула Z-score следует из нормировки нормально распределенных случайных величин.

Вышеописанная процедура построения выравнивания и оценки значимости выполняется для каждой последовательности из банка. После того, как весь банк просмотрен, программа выдает список последовательностей, отсортированный по $\hat{S}_i$, а точнее говоря, первую часть списка, где $\hat{S}_i > S_{treshold}$. Параметр порога значимости $S_{treshold}$ задается пользователем.

PSI-BLAST

Программа PSI-BLAST [15] характеризуется высокой скоростью сканирования по банкам. Эвристический подход к сравнению последовательностей основан на быстром поиске участка локального совпадения и выполнении выравнивания Смита – Уотермена [1] только на этом участке.

Быстрый поиск участка локального совпадения.

Входная последовательность разбивается на блоки длиной *T* (*T* – параметр, задаваемый пользователем, по умолчания для белков *T* = 4, для ДНК *T* = 11). Далее для каждого блока генерируется синоним – мутант с заданным процентом сходства ***S*** по матрице замен. Например, для блока B = MYCH, будут сгенерированы следующие синонимичные блоки с $S \geq 70$:

MYCH
QYCH
M**F**CH
MY**I**H
MYC**E**

Далее программа осуществляет поиск сгенерированными блоками по последовательностям банка. Алгоритм поиска подстроки в строке реализован как поиск в суффиксном дереве. Для повышения скорости работы программы все последовательности сканируемых банков хранятся в предобработанном виде (в виде суффиксных деревьев, рис. 10).

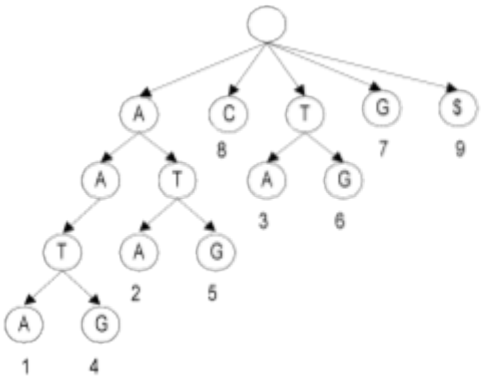


Рис. 10. Пример суффиксного дерева для последовательности AATAATGC

Вторым этапом в сравнении двух последовательностей является расширение найденного блока (называемого инициирующим локальное выравнивание) и построение локального выравнивания с делециями. Необходимое условие выполнения расширения – существование двух инициирующих блоков на определенном расстоянии друг от друга (рис. 11).

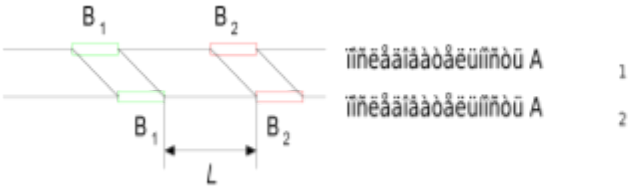


Рис. 11 Последовательности, имеющие два блока точного совпадения.

Между блоками точного совпадения (B_1 и B_2 на рисунке 11) выполняется выравнивание типа жадных расширений (Greedy extension) [15]. Далее считается, что построенное выравнивание является оптимальным среди всех локальных выравниваний. Строго говоря, это не всегда так. На практике вышеописанная процедура выравнивания дает оптимальный результат только на последовательностях с процентом сходства ~ 30% и выше.

Оценка значимости

Для оценки значимости сходства применяется стандартная формула пересчета веса выравнивания в нормализованный вес:

$$W' = \frac{\lambda W - \log K}{\log 2}$$

где log – натуральный логарифм, K, λ - параметры распределения, W – вес выравнивания, W' измеряется в битах. Нормализованный вес переводится в другой, более информативный показатель E-value. E-value(t) – это математическое ожидание числа последовательностей из банка, которые имеют сходство с запросом $\geq W'$, тем самым учитывается размер пространства поиска (банка последовательностей).

$$E(W') = N / 2^{W'}$$

где $N = m \cdot n$, m – размер банка в аминокислотных остатках или нуклеотидах, n – длина последовательности. Например, сравнивая белок длиной 250 остатков с банком размера ~ 50`000`000 остатков, уровню значимости 0,05 соответствует нормализованный вес ~ 38 бит.

HMM Search

Скрытые Марковские модели (HMM – Hidden Markov Models [16]) – это класс стохастических моделей, являющийся мощным инструментом для анализа информации, заключенной в символьных последовательностях. CMM может рассматриваться как “черный ящик”, который генерирует последовательность наблюдений. Скрытые внутренние состояния – случайные величины, которые образуют цепь Маркова с конечным числом состояний. Наблюдаемые выходные значения – символы на позициях являются стохастическими переменными с распределением, определяемым текущим состоянием цепи.

Применение цепей происходит в два этапа. Сначала модель обучается на эталонных последовательностях. Затем на вход предъявляется последовательность, не принадлежащая обучающей выборке, и вычисляется вероятность следующего события: данная последовательность была порождена той же статистической моделью, что и эталонные последовательности. С точки зрения молекулярной эволюции Марковские цепи едва ли могут быть корректными моделями белковых последовательностей. Дело в том, что отношения родства между белками имеют иерархическую древовидную структуру. Не смотря на это, CMM дают корректные

множественные выравнивания, не противоречащие пространственной структуре выравниваемых белков.

Bioccelerator и ISS (Intermediate Sequence Search)

Программы Bioccelerator и ISS [16] используют стандартный метод выравнивания Смита-Уотермена с фиксированными штрафами за вставки/делеции. Bioccelerator выполняет парные сравнения заданной последовательности с каждым

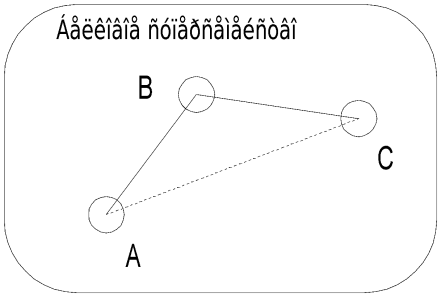


*Рис. 12 Построение суперсемейства последовательности **A**, включающее явные (**B**) и неявные (**C**) её гомологи.*

представителем из банка. ISS осуществляет поиск в две итерации. Сначала по последовательности ***A*** находятся все ее промежуточные гомологи H_i. Далее процедура поиска по банку повторяется для каждой H_i, что позволяет найти неявных гомологов H'_j последовательности ***A***. Схематически, правило $A \approx H_i, H_i \approx H'_j \Rightarrow A \approx H'_j$ показано на рис. 12. Конечным результатом поиска считается объединение $\{H_i\} \cup \{H'_j\}$.

В программе Bioccelerator статистическая значимость сравнений оценивается по эмпирической формуле, полученной В.Пирсоном [17, 18]. Вес выравнивания растет логарифмически с ростом длин сравниваемых последовательностей. Известно также, что вес оптимального выравнивания зависит от матрицы замен и от частот символов, но при одном прогоне программы поиска эти параметры фиксированы. Принимая во внимание все эти факторы, можно написать зависимость веса оптимального выравнивания от длины сравниваемых последовательностей:

$$\mu(W_{\max}) = \rho \log(mn) + \varphi$$

Параметры ρ и φ зависят от банка последовательностей и матрицы замен. Сравнивая заданную последовательность с банком, мы фактически имеем репрезентативную выборку весов оптимальных выравниваний $W^i{}_{\max}(m,n)$.

Считая каждое сравнение независимым, а распределение величины $W_{\max}(m,n)$ одинаковым для всех сравнений, можно оценить параметры ρ и φ с помощью линейной регрессии μ на $\log(mn)$ по реальной выборке весов, полученных выравниванием заданной последовательности с банком. Конечная формула нормализованного веса такова:

$$W' = \frac{W - \rho \log(mn) - \varphi}{\sigma}$$

Считается, что дисперсия веса не зависит от длин и вычисляется непосредственно по выборке. На реальных данных Пирсоном было показано, что оценки линейной регрессии дают биологически адекватную значимость.

2. ПОСТАНОВКА ЗАДАЧИ.

Нас будут интересовать выравнивания, отражающие сходство пространственных структур. Цель работы – понять, насколько стандартный метод выравнивания последовательностей, предложенный Смитом и Уотерменом [1], позволяет восстановить выравнивание пространственных структур, как выбирать параметры в этом методе.

Алгоритм выравнивания Смита – Уотермена строит оптимальное (т.е. имеющее наибольший возможный вес) выравнивание двух данных последовательностей. Чтобы определить вес выравнивания, априорно выбираются параметры – матрица весов замен, штраф за открытие делеции, штраф за удаление символа (штраф за продолжение делеции).

В литературе изучался вопрос, какие значения параметров являются лучшими “в среднем”. С помощью компьютерных экспериментов было показано, что наилучшие результаты дает матрица замен, предложенная Gonnet [5], и значения штрафов GapOpenPenalty=12.0; DeletionPenalty = 1.0 [4]. Мы перепроверили эти результаты для базы структурных выравниваний BaliBase [19], которая использовалась нами в качестве источника эталонных выравниваний, и получили подтверждение этих результатов (подробнее в результатах и обсуждении).

Однако остались неизученными более тонкие вопросы:

- можно ли повысить точность восстановления пространственного выравнивания, подбирая специальные значения параметров отдельно для каждой пары сравниваемых белков;
- какие именно части структурных выравниваний восстанавливаются, а какие – теряются;
- каким участкам выравнивания последовательностей можно доверять в большей, а каким – в меньшей степени.

Эти вопросы и были предметом нашего исследования.

3. МЕТОДИКА.

3.1 Источник структурно адекватных выравниваний.

Тестовую выборку составили множественные структурно адекватные выравнивания белковых доменов, представленные в базе BAliBASE [19].

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 ещё 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 (рис.13).

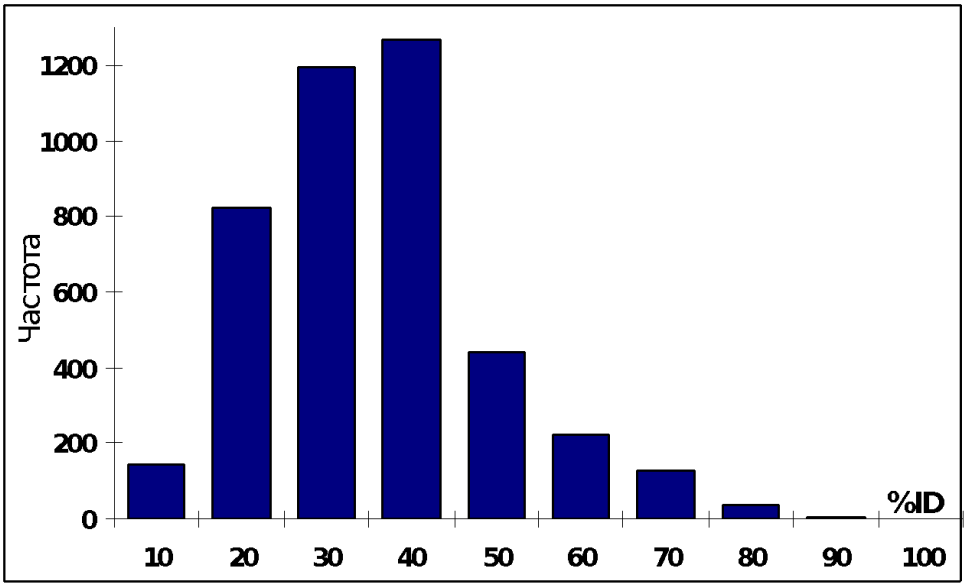


Рис. 13 Распределение выравниваний, содержащихся в базе BaliBase, по %ID.

3.2 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).

Мы решили выяснить взаимосвязь этих величин. Рис.14 сделан на случайных последовательностях (Модификация консенсуса профиля Igb4 с заданным уровнем гомологии, матрица замены Blosum62). Подобные диаграммы получились и на реальных белках. Получилось, что зависимость линейная с небольшим разбросом точек, следовательно, можно считать %ID и S-фактор идентичными величинами гомологии. В дальнейшем будем рассматривать %ID как более распространенную меру сходства последовательностей.

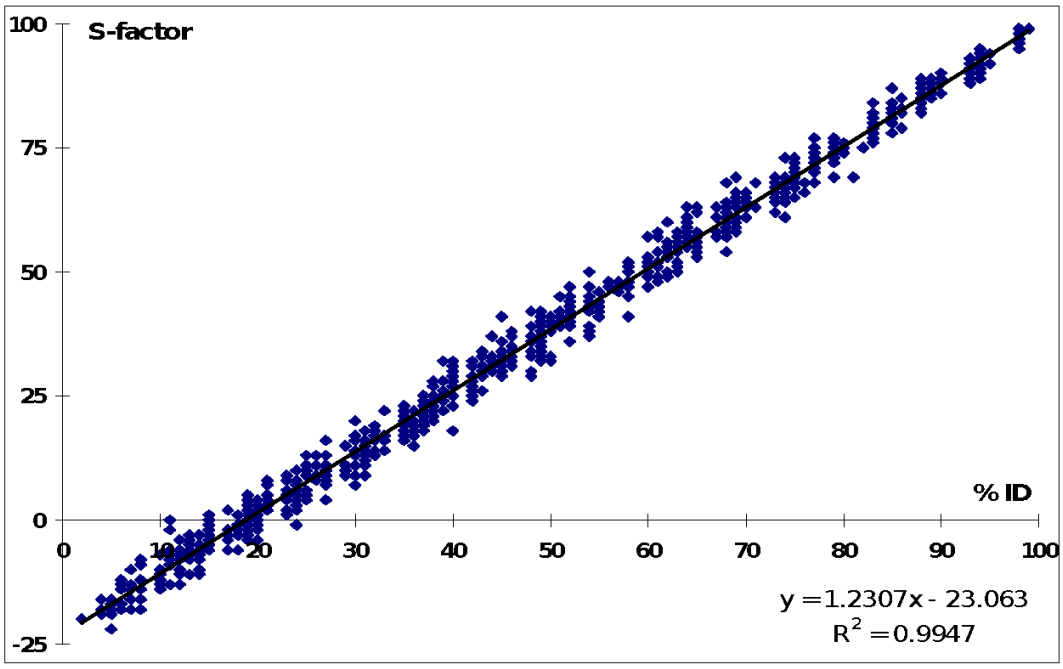


Рис.14 Взаимосвязь между S-factor и % ID.

3.3 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. 15 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:

Aln_Sim% = 42 / 58 = 0.72 = 72%

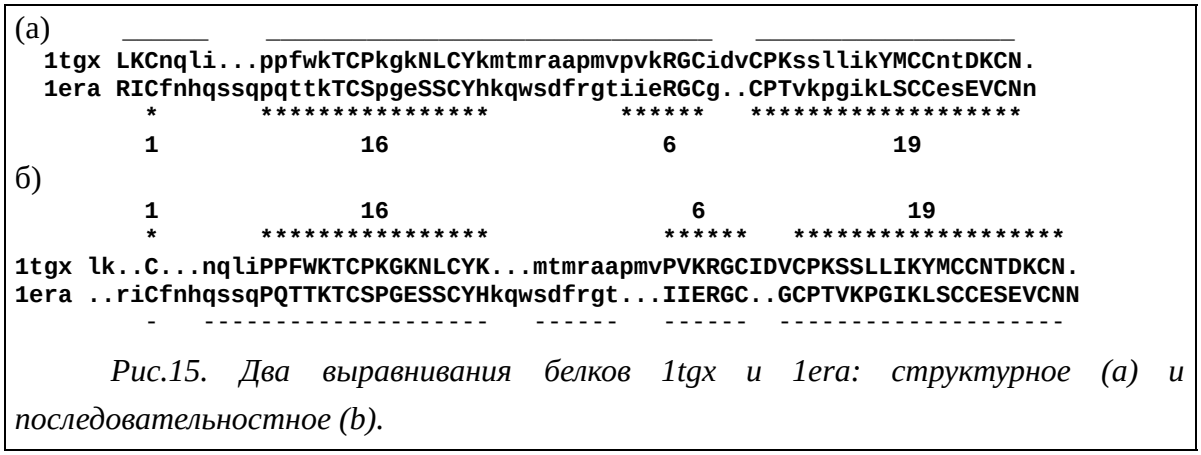


Рис.15. Два выравнивания белков 1tgx и 1era: структурное (а) и последовательностное (b).

Для более детального рассмотрения выравнивания введем понятие *«острова»*. «Остров» - это участок выравнивания между двумя делециями. Структурное выравнивание (а) содержит 3, а выравнивание (б) – 5 «островов».

3.4 Парето-оптимальные выравнивания. Субэталонное выравнивание.

Строится множество выравниваний, каждое из которых соответствует определённому количеству «дырок» (Gaps) - несколько подряд идущих удаленных символов в какой-либо из выровненных последовательностей.

Отличие реализованного метода от других широко применяемых заключается в том, что строится множество выравниваний, каждое из которых соответствует определённому количеству «дырок». Выравнивание аминокислотных последовательностей осуществляется по методу динамического программирования. Из полученного множества впоследствии можно выбирать одно выравнивание.

Будем называть лучшее (наиболее близкое к эталонному) выравнивание из получаемого множества ***субэталонным*** (или, в случае если мы сравниваем с биологически адекватным выравниванием, ***субадекватным***). Таким образом, близость субэталонного выравнивания к эталонному отражает степень «восстанавливаемости» верного выравнивания (отставим пока в сторону вопрос о наличии методов, позволяющих верно выбрать такое выравнивание из множества

построенных).

4. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ.

4.1. Можно ли с помощью выравнивания последовательностей правильно сопоставить структуру 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?

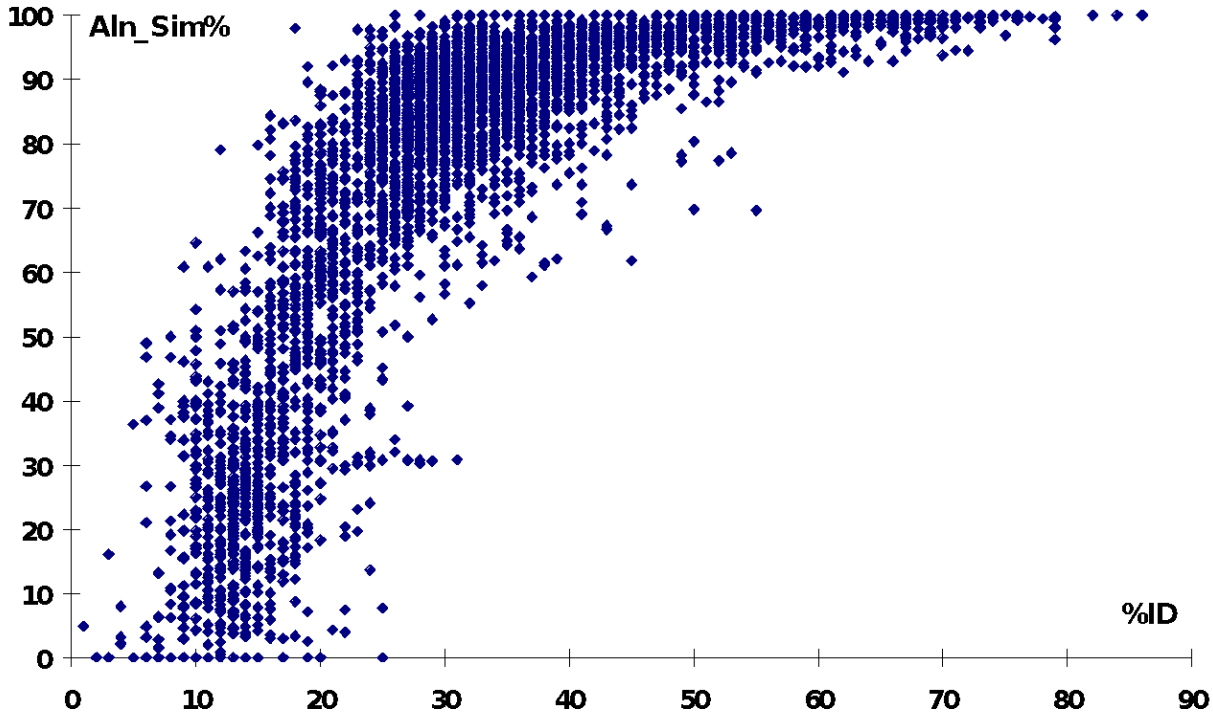


Рис.16 Зависимость угаданности структурно верного выравнивания от % ID.

Во всех 23 семействах, содержащихся в базе данных Bali Base, были построены Парето-множества выравниваний для всех пар белков семейства (использовалась матрица замен Gonnet). Из полученных Парето-множеств были выбраны выравнивания, имеющие максимальный Aln_Sim% (субадекватные выравнивания). На рис.16 представлена зависимость Aln_Sim% субадекватного выравнивания от % ID.

Результат: чем больше гомология последовательностей, тем больше вероятность для них обладать похожей структурой, и, следовательно, построить структурно верное выравнивание. Близкое (AlnSim% > 50%) к структурно верному

выравниванию можно построить при %ID > 20%.

4.2. Детальное изучение выравниваний. Угаданые «острова».

Для более детального понимания, что можно угадать с помощью выравнивания последовательностей, мы решили проверить, как хорошо угадываются участки в выравнивании без делеций («острова»). «Остров» эталонного выравнивания считается угаданным, если в построенном последовательностном выравнивании присутствует хотя бы одно такое же сопоставление, как в эталонном «острове». Мера, отражающая степень угаданности «острова» (Isl_Sim%), равна отношению числа угаданных сопоставлений к общему количеству сопоставлений, присутствующих в эталонном «острове».

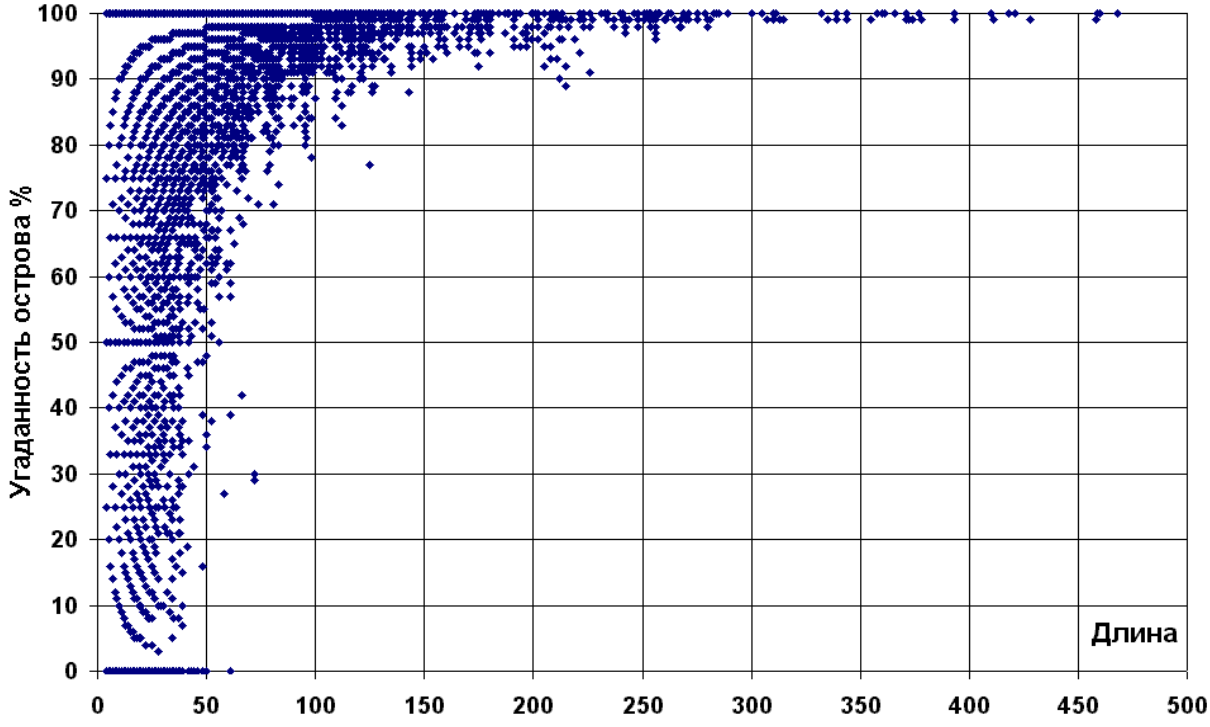


Рис. 17 Зависимость процента угаданности «острова» от его длины.

Была построена диаграмма процента угаданности острова в зависимости от его длины (рис.17). На диаграмме видно, что при длине острова больше 50 аминокислот он угадывается более чем на 70%. Более короткие острова могут быть как хорошо угаданными, так и потерянными вовсе (Isl_Sim% = 0%).

На гистограмме рис.18 показано распределение «островов» субадекватных выравниваний по проценту похожести на эталонные. Всего потерянных «островов» (в которых не угадано ни одного сопоставления) 37%, что достаточно много. Однако из 62% угаданных «островов» 42% приходится на острова, которые угаданы на 90%

и более. Получается, что если остров угадан, то он угадан почти целиком.

Рис.18 Распределение «островов» субадекватных выравниваний по проценту угаданности.

Неугаданные «острова» имеют длину менее 50 символов и небольшой или даже отрицательный вес (рис.19). В то время как угаданные «острова» имеют положительный вес, который растет с длиной «острова».

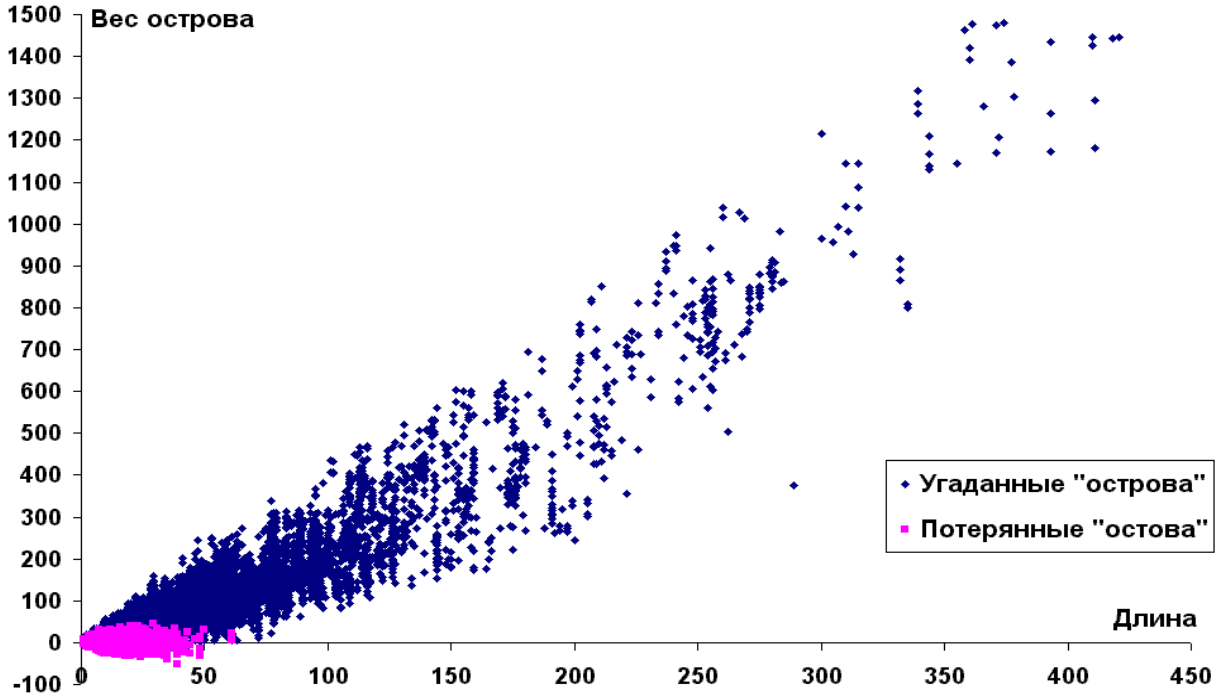


Рис. 19. Зависимость веса «острова» от его длины.

Рассмотрим подробно выравнивание аминокислотных последовательностей белков 1ark и 1vie (рис.20). Сходство этих белков (%ID) равно 17%.

Структурное выравнивание (рис.20 а) содержит три «острова», обозначенных соответственно *A*, *B* и *C*. «Остров» *C* входит во все Парето-оптимальные выравнивания данных последовательностей (рис. 20 б), «остров» *B* частично угадан в выравнивании с 1-й множественной делецией, а «остров» *A* отсутствует во всех последовательностных выравниваниях.

```
а)      ---A---         --------B-------  ------------C------------
1ark tagKiFRAmydymaadadevSfkDGDaIinvqAid..EGWmygTvQRTGRTgMLPANyVEaI.
1vie ...DrVRKks.........gAawQGQiVgwycTnltpEGYaveSeAHPGSVqIYPVAaLErIn

б)
0    0.52    36.00   SW               ------------------------
(tagkiframydymaadadevsfkdgdaiinvqaid) EGWmygTvQRTGRTgMLPANyVEaI.
(drvrkksgaawqgqivgwyctnltp)           EGYaveSeAHPGSVqIYPVAaLErIn

1    0.63    40.00   Best -----         ------------------------
(tagkiframydymaadadevsfk) DGDaI..inVQaidEGWmygTvQRTGRTgMLPANyVEaI.
(drvrkksgaaw)             QGQiVgwycTNltpEGYaveSeAHPGSVqIYPVAaLErIn

2    0.52    51.00                            -------------------------
(TAGKIFRAMYDYMAADA)DEVSfKDGDA....IIN...VQaidEGWmygTvQRTGRTgMLPANyVEaI.
                   DRVRkKSGAAwqgqIVGwycTNltpEGYaveSeAHPGSVqIYPVAaLErIn
```

Рис.20 Различные выравнивания белков 1ark и 1vie: а) структурно адекватное выравнивание; б) Различные Парето-оптимальные выравнивания.

Обозначения: 1 - количество «дырок»; 0.63 - сходство со структурным выравниванием; 40.00 - "полу-вес" выравнивания (вес без учета штрафа за открытие «дырок»).

Рассмотрим детально каждый из «островов» структурного выравнивания белков 1ark и 1vie (рис.21 а):

«Остров» *A* имеет нулевой вес (рис.21 б), и поэтому не входит ни в одно из Парето-оптимальных выравниваний (см. рис.20).

«Остров» *B* имеет отрицательный вес, равный -14, но в нем есть «основа»:

```
+1+7+1-1+3
 D G D a I
 Q G Q i V
 +++++++++
```

Вес этой основы равен 11. Основа входит в последовательностное Парето-оптимальное выравнивание с 1-й дыркой. В этом выравнивании правая часть

«острова» *B* «переброшена» к «острову» *C*, т.к. тогда выравнивание имеет больший вес (рис.21 г).

Вес «острова» *C* положителен и равен 36 (рис.21 д), следовательно, данный «остров» входит во все Парето-оптимальные выравнивания белков 1ark и 1vie.

а) *структурно адекватное выравнивание*

```
        ---A---         --------B-------  ------------C------------
1ark tagKiFRAmydymaadadevSfkDGDaIinvqAid..EGWmygTvQRTGRTgMLPANyVEaI.
1vie ...DrVRKks.........gAawQGQiVgwycTnltpEGYaveSeAHPGSVqIYPVAaLErIn
```

б) Остров A:

```
0-2 0 5 0-1-2
K I F R A m y
D r V R K k s
```

Вес = 0

в) Остров B:

```
-3+1-2-4+1+7+1-1+3-4-4-1-2 1-3-4
 v S f k D G D a I i n v q A i d . .
 g A a w Q G Q i V g w y c T n l t p
 ------- +++++++++ -------------
    -8       11          -17
```

Вес = -14

г) Альтернативное выравнивание правой части острова B:

```
-1-2 0 1-1-1-1
. . i n V Q a i d
g w y c T N l t p
```

Вес = -5

д) Остров C:

```
4 7 4-1-1-1 2-2 0 1 0 7 0 0-1 2 0 8 0 0-2 2 4-1 4
E G W m y g T v Q R T G R T g M L P A N y V E a I
E G Y a v e S e A H P G S V q I Y P V A a L E r I
```

Вес = 36

Рис.21 Структурно адекватное выравнивание белков 1ark и 1vie веса его «островов».

Итак, главная причина потери «островов» – это их малый вес. Такое возможно потому, что:

а) неадекватная матрица замен – т.е. эта матрица дает малый или отрицательный вес функционально значимым «островам», может быть надо использовать разные матрицы для каждого отдельного «острова»;

б) вероятно существует различие между структурным и эволюционным выравниванием.

4.3. Выбор наилучших параметров для выравнивания последовательностей методом Смита-Уотермана.

Задача Смита-Уотермана [1] ставится так: требуется найти выравнивание V с максимальным весом $W(V,S,\sigma(l))$, где матрица замен $S=\{s_{ij}\}_{i,j=1}^{k}$ и штрафная функция $\sigma(l)=A+B\cdot l$ (A – штраф за открытие, B – штраф за продолжение делеции) являются параметрами, влияющими на качество построенного выравнивания.

Критерием качества построенного выравнивания в нашем случае является степень его похожести на адекватное выравнивание (Aln_Sim %) из базы данных BAliBase.

Парето-техника выравниваний позволяет нам избавиться от параметра A (штрафа за открытие «дырки»), от него правильность выравнивания зависит в наибольшей степени. Поэтому сначала мы подобрали матрицу замен и штраф за продолжение делеции так, чтобы субадекватное выравнивание было максимально похожим на адекватное. Нами было рассмотрено 4 матрицы: Blosum62, Blosum30, Pam250, Gonnet (рис.22 а). Далее для лучшей матрицы (Gonnet) было проверено, что лучший штраф за продолжение делеции = 1.0 (рис.22 б).

Для этих исследований использовались двухфакторные ранговые методы (критерий Фридмана и критерий Пейджа[20]).

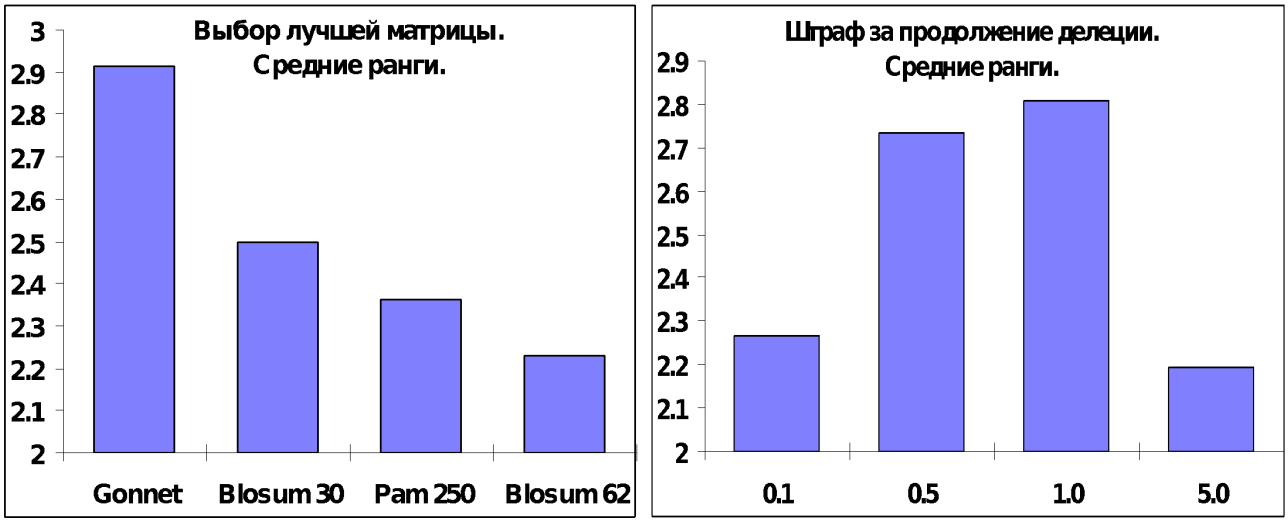


Рис.22 Средние ранги. а)Выбор матрицы замен; б) Для матрицы Gonnet выбор штрафа за продолжение делеции.

Затем были построены выравнивания по матрице Gonnet со штрафом за продолжение делеции = 1.0, а штраф за открытие делеции изменялся от 1 до 50. На рис.23 представлена гистограмма средних рангов для этого параметра в диапазоне от 6 до 20. Видно, что наилучшими являются значения 11 и 12.

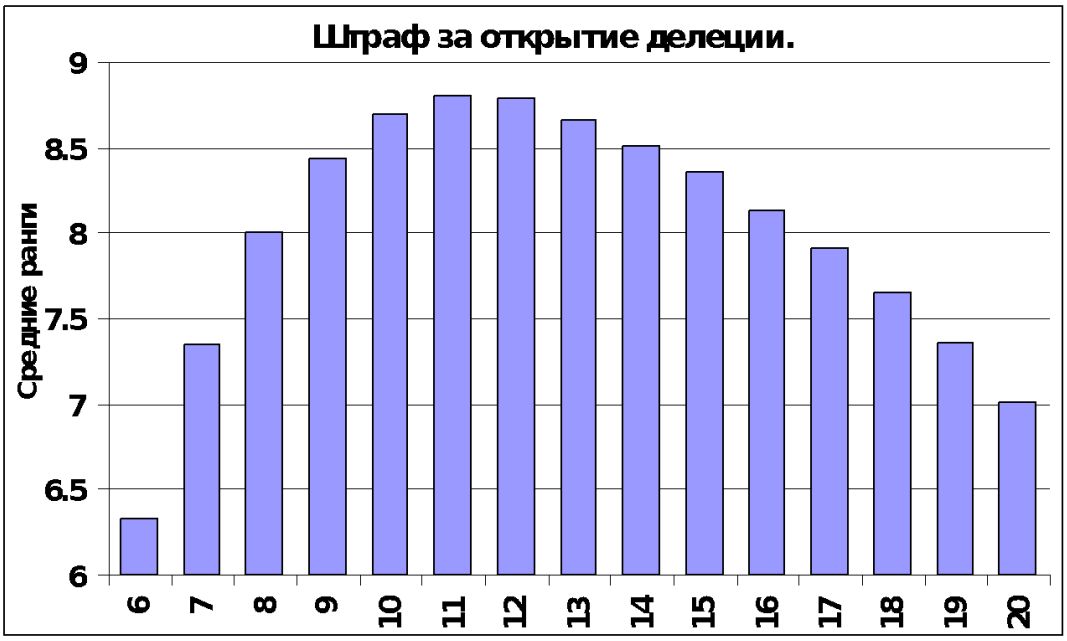


Рис. 23 Подбор наилучшего штрафа за открытие делеции.

Итак, для базы данных BAliBase лучшей матрицей замен является Gonnet; ***B*** (штраф за продолжение делеции) = 1.0; ***A*** (штраф за открытие «дырки») = 11 или 12 – что соотносится с ранее известными данными. Для дальнейших исследований мы решили использовать значение 12 для параметра штраф за открытие множественной делеции как более распространенный.

4.4 Различия между субадекватным и выравниванием Смита-Уотермана при оптимальных параметрах.

Далее нами был изучен вопрос – насколько можно повысить “процент угаданности” структурного выравнивания (Aln_Sim%), если вместо универсального штрафа за продолжение делеции = 12 брать субадекватное выравнивание. Для наглядности была введена величина Delta, равная разности процента угаданности (Aln_Sim%) наилучшего (субадекватного) выравнивания и Aln_Sim% выравнивания Смита-Уотермана при заданных параметрах. Эта величина отражает максимально возможное улучшение качества распознания пространственной структуры с помощью данного алгоритма выравнивания последовательностей. На гистограмме

рис. 24 показано распределение величины Delta.

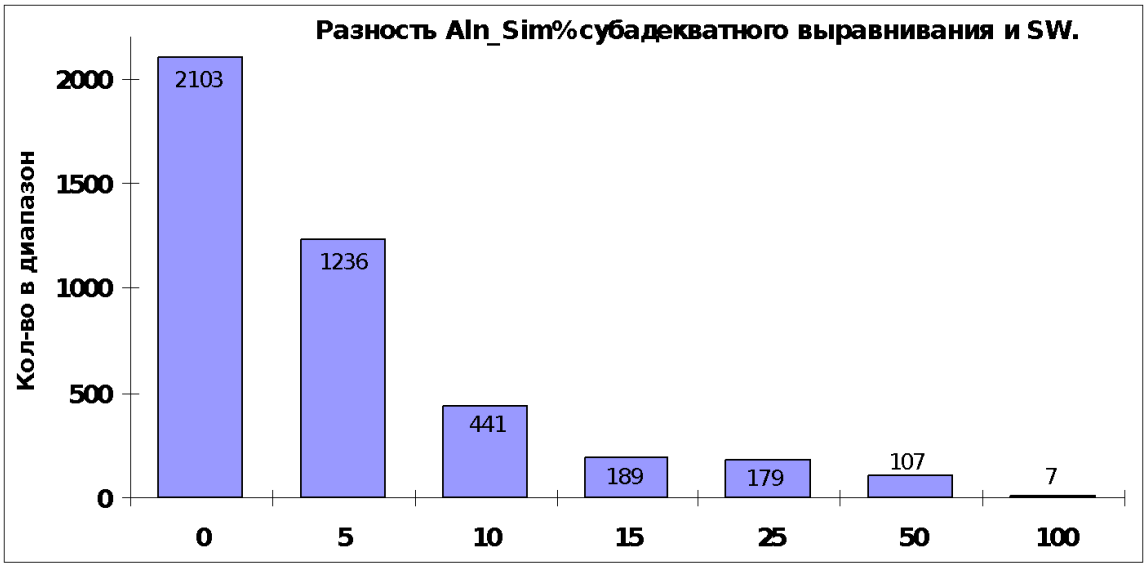


Рис.24 Распределение разности процента угаданности (Aln_Sim%) субадекватного выравнивания и выравнивания Смита-Уотермана при заданных параметрах(SW): S = Gonnet, A = 12, B = 1.0 .

Как видим, в основном, возможный выигрыш невелик. Однако, как правило, “добавка” касается структурно значимых фрагментов.

Так, на рис. 25 изображены различные выравнивания белков 1tgx и 1era из семейства кардиотоксинов: структурно адекватное выравнивание (рис. 25а), содержащее три «острова», и последовательностные выравнивания (рис. 25б) с количеством множественных делеций от 0 до 5. Концы последовательностей в скобках не выравниваются (строятся локальные выравнивания). В выравнивании Смита-Уотермана, оптимальном при указанных выше параметрах, количество «дырок» равно 1 и Aln_Sim% = 86%. В то время, как для субадекватного выравнивания Aln_Sim% = 88%. Однако эта разница на 2% означает потерю в выравнивании Смита-Уотермана «острова» *A*, содержащего сопоставление функционально значимой аминокислоты *цистеин*. Другие «острова» *B* и *C* угаданы в обоих выравниваниях.

Далее, мы детально рассмотрели случаи, когда выравнивание, оптимальное при штрафе 12, не является суб-адекватным. Предварительные результаты состоят в следующем:

Разность между полу-весом Парето-оптимального выравнивания с N+1 островами выравниванию и N островами не позволяет определить, какое из этих

выравниваний ближе к структурному выравниванию. Значимой характеристикой, возможно, является вес «нового» острова или вес его «основы» (см. рис.21)

(а) *структурно адекватное выравнивание:*

```
1tgx LKCnqli...ppfwkTCPkgkNLCYkmtmraapmvpvkRGCidvCPKssllikYMCCntDKCN.
1era RICfnhqssqpqttkTCSpgeSSCYhkqwsdfrgtiieRGCg..CPTvkpgikLSCCesEVCNn
     ---A---   ----------------B--------------  ---------C---------
```

(б) *последовательносные Парето-оптимальные выравнивания:*

```
0       0.33     63.00
 (lkcnqlippfwktcpkgknlcykmtmraapmvpvkrgcidv)CPKssllikYMCCntDKCN.
(ricfnhqssqpqttktcspgesscyhkqwsdfrgtiiergcg)CPTvkpgikLSCCesEVCNn

1       0.86    120.00     +57.0  SW(12)
   (lkcnqli)ppfwkTCPkgkNLCYkmtmraapmvpvkRGCidvCPKssllikYMCCntDKCN.
(ricfnhqssq)pqttkTCSpgeSSCYhkqwsdfrgtiieRGC..gCPTvkpgikLSCCesEVCNn

2       0.88    126.00      +6.0
(lk)C...nqlippfwkTCPkgkNLCYkmtmraapmvpvkRGCidvCPKssllikYMCCntDKCN.
(ri)CfnhqssqpqttkTCSpgeSSCYhkqwsdfrgtiieRGC..gCPTvkpgikLSCCesEVCNn

3       0.88    130.00      +4.0   Best
(lk)C.nq..lippfwkTCPkgkNLCYkmtmraapmvpvkRGCidvCPKssllikYMCCntDKCN.
(ri)CfnhqssqpqttkTCSpgeSSCYhkqwsdfrgtiieRGC..gCPTvkpgikLSCCesEVCNn

4       0.74    136.00      +6.0
(lk)C...nqlippfwkTCPkgkNLCYk...mtmraapmvpvkRGCidvCPKssllikYMCCntDKCN.
(ri)CfnhqssqpqttkTCSpgeSSCYhkqwsdfrgt...iieRGC..gCPTvkpgikLSCCesEVCNn

5       0.72    142.00      +2.0
(lk)C...nqlippfwkTCPkgkNLCYk...mtmraapmvpvkRGCidvCPKssllikYMCCntDKCN.
(ri)CfnhqssqpqttkTCSpgeSSCYhkqwsdfrgt..i.ieRGC..gCPTvkpgikLSCCesEVCNn
```

Рис.25 Различные локалные выравнивания белков 1tgx и 1era (семейство кардиотоксинов):

а) Структурное выравнивани;

б) Различные Парето-оптимальные выравнивания.

Обозначения: 3 - количество дырок; 0.88 - сходство со структурным выравниванием; 130.00 - "полу-вес" выравнивания (вес без учета штрафа за открытие дырок); +4.0 - прирост полувеса.

5. ВЫВОДЫ.

1. Выведена зависимость надежности восстановления выравнивания пространственных структур по аминокислотным последовательностям белков.

2. Исследована зависимость качества восстановления структурного выравнивания от параметров алгоритма выравнивания методом Смита-Уотермана. Показано, что наилучшие результаты получаются для матрицы Gonnet и штрафа за открытие множественной делеции = 11 или 12, штрафа за продолжение делеции = 1.

3. Показано, что в 49 % случаев выравнивание с параметрами, указанными выше, является наилучшим среди выравниваний, которые могут быть получены с другими значениями штрафа за открытие множественной делеции. А в 29% случаев оно отличается от наилучшего из возможных не более чем на 5%.

4. Изучены различия последовательносных и структурных выравниваний с помощью понятия острова. В последовательностных выравниваниях угадано 62% островов, из которых 42% приходится на острова, которые угаданы на 90% и более. Потерянных островов 37%, они имеют малый вес и длину.

5. Выявлены источники, приводящие к отличиям выравниваний Смита-Уотермана от структурно-верных (учет веса всего острова, а не его "основы"), и указаны пути их устранения.

СПИСОК ЛИТЕРАТУРЫ.

1. Smith T.F. and Waterman M.S., Comparison of biosequences, *Adv. Appl.Math.,* 2, 482, 1981.

2. Waterman M.S. (ed.) "Mathematical methods for DNA sequences.", CRC Press, Boca-Raton, FL, 1989.

3. Needleman S.B., Wunsch C.D. "A general method applicable to search for similarities in the amino acids sequences of tow proteins.", J. Mol. Biol., 1970, V.48, pp.444-453.

4. Henikoff S., Henikoff J. (1992) Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci USA 89

5. Benner SA, Cohen MA, Gonnet GH. Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng. 1994 Nov;7(11):1323-32.

6. Altschul S. "A protein alignment scoring system sensitive at all evolutionary distances.", Mol. EvoL, 1993, v.36, p.p. 290-300

7. Karlin S., Altschul S. "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.", Proc Natl. Acad. Sci. USA, 1990, v. 87, p.p. 2264-2268

8. Waterman M.S. "Introduction to computational biology.", 1995, Chapman&Hall Press, UK

9. Finkelstein A.V., Roytberg M.A. (1993) Computation of biopolymers: a general approach to different problems. Vol. 30

10. Ройтберг М.А. "Новый подход к проблеме выравнивания последовательностей: больше совпадений, меньше удалений и никаких весовых коэффициентов". - В: Труды конференции "Геном человека - 93" (Черноголовка, Март 10-12, 1993), стр. 135. М., 1993.

11. Ройтберг М.А. "Парето-оптимальные выравнивания символьных последовательностей.", ОНТИ НЦБИ, Пущино, 1994. Препринт. 10 стр.

12. Ройтберг М.А., Семионенков М.Н., Таболина О.Ю. "Парето-оптимальные выравнивания биологических последовательностей.", Ж. "Биофизика", 1999, №3, стр. 20-37

13. Murzin, A.G., Brenner, S.E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

14. Pearson W.R. (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, No. 85

15. Altschul S. F., Gish W. (1997) New generation of iterated biocomputing search tools. Nucleic Acid Research, Vol. 25, No. 17

16. Cyrus Chothia et al. (1998) Sequence comparisons using multiple sequences. Journal of Molecular Biology, 273

17. Pearson W.R. (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, No. 85

18. Pearson W.R. (1998) Empirical statistical estimates for sequence similarity searches. Journal of Molecular Biology, 274

19. Julie Thompson, Frédéric Plewniak and Olivier Poch (1999) BAliBASE: A benchmark alignments database for the evaluation of multiple sequence alignment programs, *Bioinformatics*, 15, 87-88

20. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере, Учеб. пособие для студ. вузов / Под ред. В.Э.Фигурнова, - М.: ИНФРА-М; Финансы и статистика, 1995., 384 с.