на правах рукописи

УДК 533.9

ОЛЕЙНИКОВА НАТАЛЬЯ ВАСИЛЬЕВНА

ВЫРАВНИВАНИЕ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ: АНАЛИЗ СУЩЕСТВУЮЩИХ МЕТОДОВ И РАЗРАБОТКА НОВЫХ АЛГОРИТМОВ

03.00.02.-биофизика

автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Москва 2004

Работа выполнена в Институте Математических Проблем Биологии РАН

Научный руководитель:
кандидат физико-математических наук
Михаил Абрамович Ройтберг

Официальные оппоненты:
доктор биологических наук
Андрей Александрович Миронов
кандидат физико-математических наук
Всеволод Юрьевич Макеев

Ведущая организация:
Институт теоретической и экспериментальной биофизики РАН

Защита состоится 25 марта 2004 г. в 10 час. 00 мин. на заседании диссертационного совета К 212.156.03 при Московском физико-техническом институте (141700, Московская обл., г. Долгопрудный, Институтский пер. 9, МФТИ).

С диссертацией можно ознакомиться в библиотеке МФТИ.

Автореферат разослан «10» февраля 2004г.

Ученый секретарь
диссертационного совета
кандидат физико-математических наук  В.Е. Брагин

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы Современный эган разни шя молекулярной биолопж и генетики характеризуется лавинообразным ростом объема расшифрованных биологических последовательностей. Это произошло в значительной степени благодаря программе «Геном человека» и другим подобным программам. В то же время, как правило, первичные структуры белков являются лишь начальной точкой исследования, а молекулярных биологов, в конечном итоге, интересует трехмерная структура белков и их функциональная активность. Однако получение трехмерной структуры белка и определение его функции является очень трудоемкой и дорогостоящей задачей. Следовательно, нужны методы, позволяющие в идеале предсказывать структуру и функцию белка, зная только последовательность, или как минимум отнести новую последовательность к какому-нибудь из уже известных классов белков.

Одним из эффективных и удобных методов классификации и выявления сходств последовательностей является метод выравнивания. Выравнивание новой последовательности с последовательностью уже хорошо изученного белка, т.е. такого, о котором известна третичная структура и функция, дает возможность количественно определить уровень сходства этих последовательностей, а так же указать участки наиболее вероятного сходства структур или функций. Для того, чтобы предсказание было правдивым, необходимо уметь строить биологически адекватное выравнивание последовательностей. Выравнивания белков, полученные наложением их пространственных структур, рассматриваются в качестве эталонных, т.к. они в наилучшей степени отражают эволюцию белков. Наиболее часто используемыми программами для сравнения последовательностей являются BLAST, FASTA и метод Смита-Уотермана (SW) - наиболее чувствительный из перечисленных.

Сказанное выше определяет актуальность темы настоящего исследования - сравнения структурных выравниваний с выравниваниями, построенными методом SW и построения нового алгоритма выравнивания аминокислотных последовательностей.

Цель работы — Оценить качество восстановления структурных выравниваний методом SW, выявить причины неточного восстановления структурных выравниваний. На основе проведенного исследования разработать новый метод выравнивания аминокислотных последовательностей.

Основные задачи исследования.

(1) Определение степени максимально возможного приближения к эталонному выравниванию за счет индивидуального подбора параметров алгоритма SW.

(2) Детальное исследование внутренней структуры эталонных и алгоритмических выравниваний аминокислотных последовательностей для выяснения причин различий между алгоритмическими и эталонными выравниваниями.

(3) Разработка нового метода выравнивания аминокислотных последовательностей, сравнение качества восстановления эталонных выравниваний и скорости работы нового метода и стандартных методов выравнивания.

(4) Адаптация разработанного метода выравнивания двух аминокислотных последовательностей для решения близких задач (построение выравнивания последовательности и профиля, поиск гомологов по банку данных).

Научная новизна работы.

В процессе решения поставленных задач получены следующие новые научные результаты:

(1) Показано, что индивидуальный подбор штрафов за делеции существенно (более чем на 10%) увеличивает точность выравниваний Смита-Уотермана для белков с уровнем сходства 10-30% (серая зона). Однако даже в этом случае средняя точность выравниваний для этого диапазона %Ш превышает 52%, а достоверность - 70%.

(2) Показано, что 32% островов (безделеционных участков выравнивания) эталонных выравниваний имеют вес < 5, суммарная длина таких островов составляет 20% всей длины эталонных выравниваний. Острова весом < 5 оставляют 65% от всех потерянных островов. Только 5% островов такого малого веса были восстановлены алгоритмом. Это является причиной недостаточной точности алгоритмических выравниваний.

(3) Разработан новый алгоритм ANCHOR выравнивания последовательностей, который при построении выравнивания учитывает не весь вес острова, а только его положительную основу (ядро). Показано, что новый алгоритм не уступает в качестве восстановления структурного выравнивания методу Смита-Уотермана, но работает примерно в 2 раза быстрее.

(4) Новый алгоритм адаптирован для выравнивания. последовательности и профиля.' Показана эффективность нового метода для поиска гомологов по банку данных.

Научно-практическая ценность.

На основе нового метода построения выравниваний разработан программный комплекс, который позволяет строить выравнивания в 2 раза быстрее метода Смита-Уотермана, при этом, не теряя точности и достоверности. Программа и исходные коды доступны через всемирную компьютерную сеть по адресу: ftp://genetics.bwh.harvard.edu/Sunyaev/saadi/

Апробация работы. Материалы диссертации были доложены и обсуждены на

- Moscow Conference on Computational Molecular Biology, Moscow, 2003
- V International congress of mathematical modeling, Dubna, Russia, 2002
- Third International Conference on Bioinformatics of Genome Regulation and Structure (BGRS 2002), Novosibirsk, 2002
- Artificial intelligence and heuristic methods in bioinformatics, A NATO Advanced Studies Institute, San -Milanojtaly, 1-11 October 2001
- 1 фучныи совет Подпрограммы "ГЕНОМ ЧЕЛОВЕКА", 2001

- Workshop "Bioinformatics and Protein Classification", New Brunswick, 15-16 Dec 2000
- Human Genome Conference, Cernogolovka, 17-19 Dec 2000
- Научные конференции МФТИ, Долгопрудный, 1999,2000

Публикации. По материалам диссертации опубликовано 6 работ, в том числе 1 работа в реферируемом журнале.

Структура и объем диссертации. Диссертация состоит из введения, 6 глав, заключения, выводов и списка цитируемой литературы. Работа изложена на 85 страницах, содержит 19 таблиц и 26 рисунков. Список цитированной литературы включает 68 источников.

Список используемых сокращений: SW — метод Смита-Уотермана.

СОДЕРЖАНИЕ РАБОТЫ

Во введении дано краткое описание тематики диссертации, показана актуальность работы, сформулированы цель и задачи исследования, перечислены основные проблемы в области компьютерного анализа биологических последовательностей.

Первая глава диссертации - это литературный обзор, состоящий из трех.разделов.

В первом разделе литературного обзора дано формальное определение выравнивания биологических последовательностей, очерчен круг проблем, которые можно решать с помощью выравниваний.

Усовершенствование методов секвенирования белков и нуклеиновых кислот привело к лавинообразному росту числа расшифрованных структур этих биомолекул, размеры банков данных постоянно растут. Наиболее мощным и биологически обоснованным инструментом для анализа сходств первичных структур считается выравнивание цепочек символов и вычисление некоей меры по результирующему выравниванию. Выравнивание пары последовательностей является ключевым шагом в вычислительных методах обработки, классификации, предсказания трехмерной структуры и функции новых последовательностей. Для наглядного преставления выравнивания последовательности записывают друг под другом, сопоставляя похожие участки, а напротив негомологичных участков оставляя пробелы-делеции (рис.1)

```
V1   MqflAVStKKCA..
V2   Mr..AVSnKKCAlk
```

Рис.1 Выравнивание двух коротких последовательностей аминокислот V_1 = MQFLAVSTKKCA и V_2 = MRAVSNKKCALK. «.» - знак вставки/делеции в негомологичных частях последовательностей. Жирным шрифтом выделены совпадения.

Во втором разделе рассмотрены различные типы выравниваний и методы построения выравниваний последовательностей.

Выравнивание можно получить исходя из различных данных о двух сравниваемых последовательностях. Если известна трехмерная структура обоих белков (например, из рентгено-структурного анализа или метода ядерного магнитного резонанса), то можно

построить *структурное* выравнивание, в котором будут сопоставлены участки наибольшего структурного сходства (вторичные структуры, их взаимное расположение и т.д.).

При построении выравнивания только по первичным структурам белков используется метод динамического программирования, который заключается в нахождении выравнивания максимально возможного веса из всех возможных выравниваний.

Вес W выравнивания двух последовательностей определяется как с\ммарный вес сопоставления (вычисляется по матрице замен S) минус суммарный штраф за делеции/вставки символов:

$$W(S,\sigma) = \sum_{ij} S_{ij} - k_d\sigma ,$$

где S_{ij} – веса сопоставлений i-го символа первой последовательности и j-го символа второй последовательности, сумма производится по всем сопоставлениям выравнивания, k_d – количество вставок/делеций, σ - штраф за вставку/делецию одного символа.

Матрицы весов сопоставлений S_{ij} (матрицы замен) вычисляются по частотам замен символов в среднем по банку структурных выравниваний белковых семейств. На сегодняшний день существ) ет несколько семейств матриц замен аминокислот. Наиболее известные из них: PAM, BLOSUM. GONNET. Штрафы за вставки/делеции подбираются так, чтобы оптимальные выравнивания были максимально похожими на структурные.

На настоящий момент наибольшее распространение получила три программы выравнивания: метод Смита-Уотермана, FASTA и BLAST. Метод Смита-Уотермана находит самые достоверные выравнивания, однако работает очень медленно, программные комплексы FASTA и BLAST в своей основе несут эвристические процедуры, что значительно (в 10 раз для FASTA и в 15-20 раз для BLAST) ускоряют построение выравниваний, однако полученные выравнивания менее надежны и не всегда оптимальны.

Для построения более точных выравниваний последовательностей и возможности обнаружить более дальних гомологов применяются множественные выравнивания. Если построить два парных выравнивания одной последовательности с двумя другими, то эти две последовательности так же можно сопоставить между собой опосредованно, через общую последовательность. В итоге получается выравнивание более чем двух последовательностей, где друг под другом записываются похожие участки во всех последовательностях, а несопоставимые заменяются делециями/вставками. Такие выравнивания более надежны, т к. они проверяются сразу на нескольких последовательностях.

В свою очередь по такому множественному выравниванию можно построить *профиль* - матрицу замен длиной равной длине множественного выравнивания и шириной равной ко шчсстиу аминокислот (20). Веса замен в такой матрице вычисляются в зависимости от частоты появления того или иного остатка на данной позиции.

Третий раздел литературного обзора посвящен краткому обзору существующих баз данных биологических последовательностей: простые базы данных (PDB, Swiss-Prot), классификационные (SCOP) и базы данных множественных выравниваний (BaliBase, FSSP, SMART, Clustal).

Методы и определения

1. Эталонное выравнивание. В идеале, при сравнении последовательностей желательно строить эволюционное выравнивание, т.е. выравнивание, в котороми сопоставлены аминокислотные остатки, соответствующие одному и тому же остатку общего предка сравниваемых белков. Общий предок нам, к сожалению, не известен, и, даже зная все **о** двух белках, нет никакой возможности **в** точности воссоздать ход эволюции. Однако множественные структурные выравнивания можно считать хорошим приближением к такого рода эталону.

В данной работе **в** качестве эталонных выравниваний использовалась база данных BaliBase (http://wvyw-igbmc.u-strasbg.fr/BioInfo/BAliBASE/). Для устранения неоднозначностей мы удалили из рассмотрения последовательности, для которых еще неизвестна трехмерная структура, **а** также те белки, которые были одновременно в двух семействах. В результате, из этой базы данных тестировалось 583 парных выравнивания. В. качестве-контрольной базы данных использовалась база FSSP (http://www.ebi.ac.uk/dali/fssp/fssp.html). Из общей базы данных были выбраны семейства с одноименными названиями в BaliBase, из каждого семейства к рассмотрению допускались только те последовательности, которые имели Z-score >20. Итого использовалось ~ 13000 парных выравниваний из этой базы данных.

2. Алгоритм SW был взят из исходных кодов программного пакета FASTA (модуль dropnswJc). По нашим данным наиболее точные выравнивания SW получаются для матрицы замен Gonnet250, однако отличия для BLOSUM62 и других матриц достаточно малы.

Использовались два набора штрафов за делеции. Наибольшая точность восстановления эталонного выравнивания достигается при значениях 10 за открытие делеции (GOP) и 0.5 за её удлинение (GEP). Однако стоит заметить, что эти параметры относятся к *линейному* домену, т.е. они соответствуют линейному росту длины оптимального выравнивания относительно длины случайной последовательности, следовательно, не могут быть применимы для работы с многодоменными белками и для поиска по банкам данных. Чтобы исключить зависимость результатов от параметров, мы также использовали параметры из *логарифмического* домена (длина оптимального выравнивания растет как логарифм от длины случайной последовательности): G0P=15, GEP=1, - обычно используемые для поиска по банкам данных.

3. Метод ANCOR. Программная реализация нового алгоритма доступна по адресу ftp://genetics.bwh.harvard.edu/Sunvaev/saadi/.OcHOBHbie идеи алгоритма изложены ниже в

результатах и обсуждении.

Мм использовали 2 набора параметров для парного выравнивания метолом AN'COR: *целочисленный* - матрица замен Gonnet (целочисленная), минимальный вес якоря =20, разрешенная «яма» = -10, вес парных затравок - 8, штраф за сход с ядра = 15, за увеличение дслеции =1; *дробный* - матрица замен Gon250 (дробная), минимальный вес якоря =20, разрешенная «яма» = -10, вес парных затравок = 8, штраф за сход с ядра = 15, за увеличение делеции - 0.5, параметры метода SW между якорями: открытие делеции = 15, удлинение = 1, используемый вес >-25. Оба эти набора принадлежат к так называемой «логарифмической облает» шачений параметров, которая используется при поиске в базах данных.

4. Мера сходства двух белков ***(%ID)*** принята равной отношению числа совпадений к числу сопоставлений в конечном выравнивании. В данной работе для подсчета *%ID* всегда использовалось эталонное выравнивание, взятое из базы данных структурных выравниваний.

5. Меры качества алгоритмических выравниваний. Для того чтобы количественно определить качество алгоритмического выравнивания, были использованы две меры:

Точность выравнивания. ***Ali Acc*** (аналогично [1,2]), равна отношению количества одинаковых сопоставлений в обоих выравниваниях к общему количеству сопоставлений в эталонном выравнивании.

Достоверность выравнивания. Ali conf. равна- отношению количества одинаковых сопоставлений в обоих выравниваниях к общему количеству сопоставлений в.алгоритмически построенном выравнивании.

Неформально говоря, ***Ali_Acc*** показывает, какую долю эталонного выравнивания удалось восстановить, а ***Ali_Conf*** – насколько можно доверять построенному выравниванию.

I

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ состоят из 4 глав.

Вторая глава диссертации содержит данные по исследованию выравниваний^Смита-Уотермана и их отличиям от эталонных выравниваний. В ней проводится подробное исследование качества алгоритмического выравнивания.

Мы исследовали, насколько выравнивания, построенные методом Смита-Уотермана, соответствуют эталонным- выравниваниям. Нами была получена зависимость качества выравниваний от уровня гомологии сравниваемых белков (рис. 2).

Нами была исследована зависимость точности ***(Ali_Acc)*** и достоверности ***(Ali_Conf)*** выравниваний (см. определения в разделе Методы), построенных методом Смита-Уотгрмана, or cieiicmi сходства между сравниваемыми последовательностями.

Как видно на рис.2 (зависимость, аналогичная представленной на рис.2а, были получена ранее [1]), алгоритм SW может строить выравнивания, дос га точно точно совпадавшие с

-малонными, только при уроннс юмолопш срлшипасмыч бслкон более 30-40% Ого хорошо согласуется с давно известным пороюм *%ID*, выше ко юрою можно доенжерно носи аноним, выравнивание, зная только последовательное!и При уровне юмолопш меньше 10% метод SW совсем не может восстановить правильное выравнивание, а диапазон 10-30% является так называемой «серой зоной», в которой наблюдается очень широкий разброс *Ali_Acc* и *Ali_Conf*.

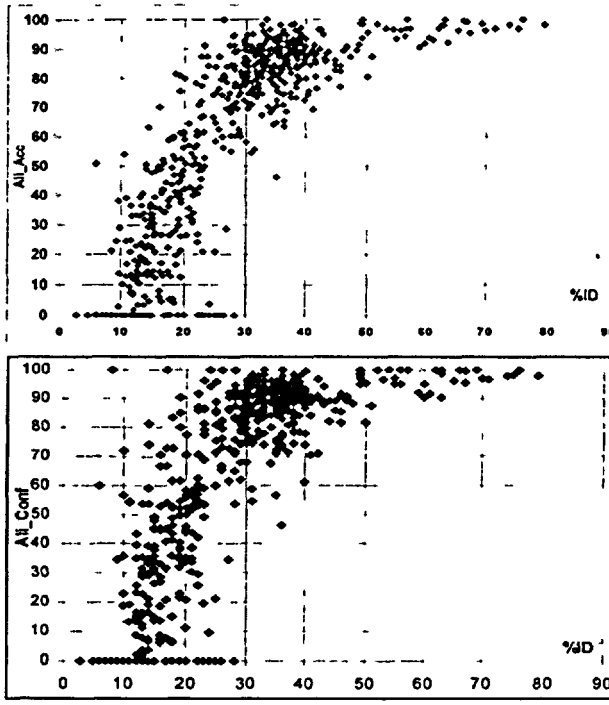


Рисунок 2.

а) Зависимость точности восстановления эталонных выравниваний (*Ali_Acc*) методом SW По оси X отложен *%ID* равный отношению точных совпадений к общему количеству сопоставлений в эталонном выравнивании.

б) Зависимость достоверности (*Ali_Conf*) выравниваний, посі роенных методом SW.

При построении диаграмм использовались линейные параметры

Для логарифмических параметров (см. методы) общая картина такая же.

Нами были получены аналогичные зависимости для выравниваний, построенных программой BLAST (данные здесь не представлены) Общий вид диаграмм сохраняется, однако они проходят несколько ниже Достаточно точными и достоверными можно считать выравнивания для последовательностей с *ID* > 40%.

Выравнивания, полученные с помощью метода SW, зависят от параметров алгоритма, в частности, от штрафов за делении Для разных пар сравниваемых белков оптимальные параметры могут отличаться. Однако до сих пор не существует техники, позволяющей подбирать параметры алгоритма индивидуально для каждой пары белков. Поэтому обычно используются стандартные значения. В зависимости от целей исследователя используются два набора параметров: *линейный и логарифмический* (см. Методы) *Логарифмические параметры* используются для поиска гомологов по банкам данных, они обладают высокой избирательностью, но более низкой чувствительностью, чем *линейные,* которые используются в

	0	0..5	5..10	10..15	15..20	20..25	25..50	50..100
□ 30< ID	124	82	29	10	4	2	2	0
□ 10< ID ≤ 30	57	64	49	43	32	16	31	6
■ ID ≤ 10%	7	5	5	6	3	2	4	0

Рисунок 3 Распределение возможного увеличения ***Ali_Acc*** за счет подбора индивидуальных параметров для каждой пары сравниваемых белков. Левый столбец соответствует случаям, когда никакого улучшения не было, следующий - случаям, когда улучшение было не более 5% и т.д. На гистограмме отдельно отмечен вклад в общую сумму каждого из 3-х диапазонов ***%ID***. Субэталонное выравнивание сравнивалось с выравниванием SW при линейных параметрах.

тех случаях, когда нужна высокая чувствительность, а избирательность не важна.

Мы исследовали возможность улучшения точности выравнивания за счет индивидуального подбора штрафов за делецию. При этом использовалась матрица замен Gonnet250 (дробная). Были просканированы все варианты штрафов за открытие делеции в диапазоне 0 - 25 с шагом 1 и штрафы за удлинение делеции в диапазоне 0 - 215 с шагом 0.1.

Для каждой пары белков были построены все оптимальные выравнивания, и выбраны те из них, в которых достигался максимально возможный ***Ali_Acc*** для этой пары (субэталонные выравнивания). ***Ali_Conf*** считались для выравниваний с максимальным ***Ali_Acc.***

Таблица 1. Средние ***Ali_Acc*** и ***Ali_Conf*** для субэталонных выравниваний (Субэт.) и выравниваний Смита-Уотермана (SW) и увеличение ***Ali_Acc*** и ***Ali_Conf*** за счет оптимального подбора параметров по сравнению со стандартными параметрами SW: линейными и логарифмическими (см. Методы). Данные представлены как для всей базы данных, так и для каждого диапазона уровня гомологии.

	Все %ID			ID ≤10%			10%< ID ≤ 30%			30% < ID		
Ali_Acc, %	Субэт	SW	Разн	Субэт	SW	Разн	Субэт	SW	Разн	Субэт	SW	Разн
Линейные	67	60	7	20	8	12	52	41	11	89	86	3
Логарифмические	67	56	11	20	6	14	52	35	17	89	85	4
Ali_Conf, %												
Линейные	79	64	15	29	13	17	70	45	24	93	89	4
Логарифмические	79	65	14	29	10	19	70	49	21	93	88	5

Распределение разницы ***Ali_acc*** субэталонного выравнивания и выравнивания SW со стандартными параметрами представлено на рис. 3. Данные для разностей ***Ali_Conf*** выглядят в общих чертах также. Видно, что для белков с высоким уровнем гомологии возможное увеличение качества выравнивания очень невелико, и большинство из них достигают своего субэталонного выравнивания на стандартных параметрах. Среднее увеличение ***Ali_Acc*** для этого диапазона ~3%. Для пар белков из серой зоны потенциальное увеличение ***Ali_Acc*** значительно больше и достигает 11%. Однако этого увеличения недостаточно для точного

восстановления эталонного выравнивания. Средняя точность субэталонных выравниваний *m* этого диапазона не превышает 52%, а достоверность - 70%.

В Таблице I представлены средние неличины ***Ali_Acc*** и ***Ali_Conf*** для субэталонного выравнивания и выравнивания SW для обоих стандартных параметров, а так же увеличения ***Ali_Acc*** и ***Ali_Conf*** при индивидуальном подборе параметров. По сравнению с линейными параметрами, логарифмические дают несколько меньшую точность выравниваний, но даже в этом случае возможное увеличение качества выравниваний не очень велико. Из таблиц видно, что зависимость точности и достоверности субэталоинных выравниваний от уровня сходства сравниваемых последовательностей качественно не отличается от аналогичных зависимостей для выравниваний SW. Таким образом, никаким подбором параметров невозможно добиться полного совпадения выравнивания SW с эталонным.

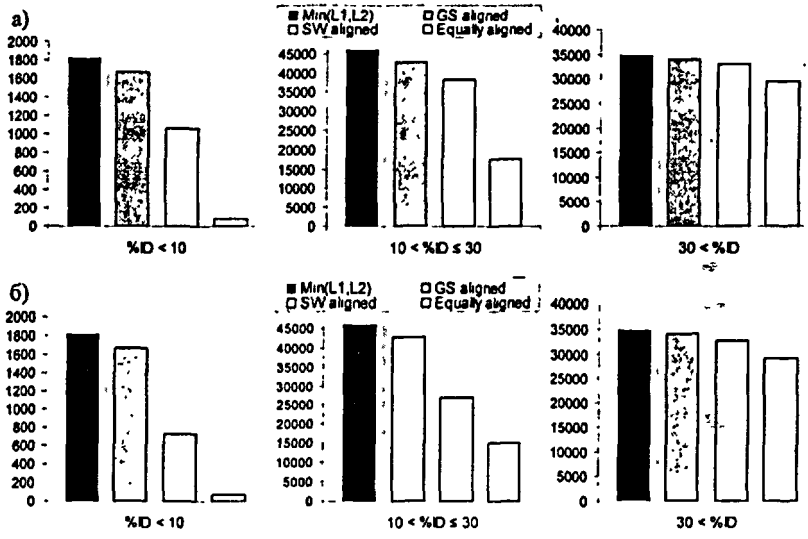


Рисунок 4. Соответствие между эталонными выравниваниями и выравниваниями SW для 3-х диапазонов *%ID*: а) линейные параметры: GOP=10; GEP=0.5 ; Ь) логарифмические параметры: GOP=15; GEP=1. Столбцы отображают 4 характеристики выравниваний. Черный столбец - максимально возможное количество позиций, которые могут быть выровнены, т.е. сумма по всем парам величины $L = \min(L_1, L_2)$, где L_1, L_2 – длины последовательностей в паре. Темно-серый - суммарное количество выровненных позиций в эталонных выравниваниях. Светло-серый - суммарное количество выровненных позиций в выравниваниях SW. Белый -суммарное количество позиций, выровненных одинаково в эталонном выравнивании и выравнивании SW.

Рис. 4 дает общее представление о соответствии между эталонными и алгоритмическими выравниваниями. Увеличение штрафов за удаление/вставку несколько уменьшает суммарную длину правильно сопоставленных участков, существенно уменьшает длину неправильных сопоставлений, и, следовательно, сильно увеличивает достоверность выравнивания в целом.

Третья глава посвящена исследованию внутренней структуры алгоритмических и эталонных выравниваний.

Невозможность построить правильное выравнивание обусловлена внутренними различиями эталонных и алгоритмических выравниваний. Внутреннее строение выравниваний удобно описывать в терминах «островов». «Островом» назовем непрерывный участок сопоставлений, в котором отсутствуют делеции/вставки в любой из последовательностей. Выравнивание можно представить как цепочку островов, соединенных делециями (выравнивание, представленное на Рис. 1, содержит 2 острова длиной 2 и 8 сопоставления).

Мы исследовали характеристики, аналогичные точности и достоверности выравниваний в целом, для отдельных островов. Для острова в эталонном выравнивании определим точность восстановления (*Isl_Acc*) равной отношению количества позиций, одинаково выровненных в SW и эталонном выравниваниях, к полной длине эталонного острова. В алгоритмическом выравнивании определим достоверность острова (*Isl_Conf*) как отношение количества правильных сопоставлений к полной длине построенного острова.

Как видно из Рис. 5, остров эталонного выравнивания может быть полностью восстановлен или полностью потерян. Частичное восстановление острова случается достаточно редко. Дальнейший анализ показал, что это в целом верно для всех диапазонов *%ID*. Следовательно, острова можно поделить на «потерянные» (не имеющие ничего общего с эталонными выравниванием) и «найденные» (содержащие хотя бы одно правильное сопоставление). Тем не менее, ситуация «все или ничего» не верна для *Ali_Acc* по всему выравниванию. Например, при линейных параметрах (см. Рис 3) для 69 выравниваний SW *Ali_Acc* = 0; *Ali_Acc* больше либо равное 80% наблюдается у 249 выравниваний (229 из них соответствует ID > 30%), а 264 выравнивания имеют *Ali_Acc* в диапазоне 0 - 80%.

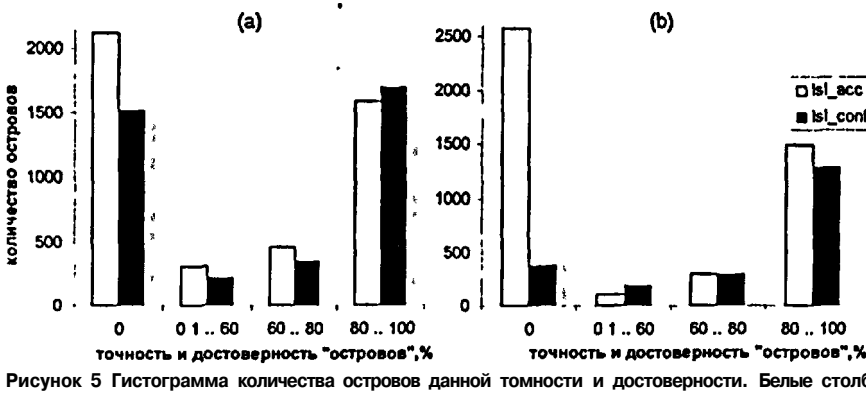


Рисунок 5 Гистограмма количества островов данной томности и достоверности. Белые столбики относятся к точности восстановления эталонных островов, черные - достоверности островов построенных методом SW. Точность и достоверность острова мы определили аналогично точности и достоверности BCCI O выравнивания, а, Ь - линейные и логарифмические параметры соответственно.

Свойства эталонных островов очень важны для возможности их восстановления каким-либо алгоритмом выравнивания последовательностей. Например, если суммарный вес острова (определяется как сумма весов всех сопоставлений в этом острове) очень низкий, то вероятность того, что остров может быть обнаружен методом SW очень низка.

Алгоритм SW не может восстановить остров с малым весом по двум причинам: 1) если штраф за открытие делеции достаточно большой, алгоритму не выгодно вставлять новую делецию, чтобы построить этот остров; 2) если штраф низкий - найдется альтернативный путь более высокого веса. Со стандартными параметрами первый случай типичен для пар белков с высокой степенью гомологии, второй - для дальних гомологов.

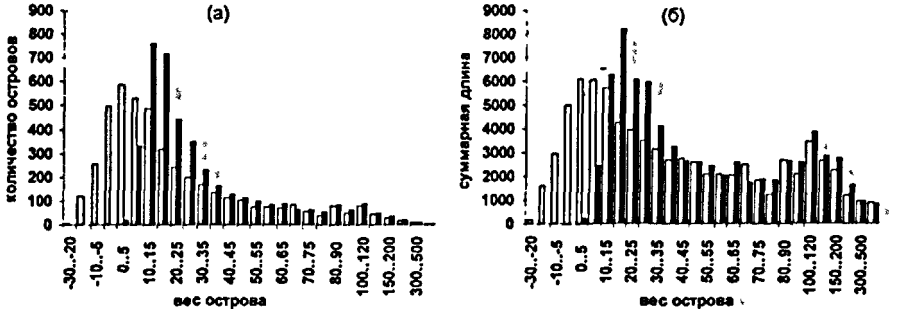


Рисунок б. а) Гистограмма распределения количества островов в эталонных выравниваниях (белый) и выравниваниях SW (черный) по весу острова, б) Суммарная длина островов, имеющих вес в пределах диапазонов. Эталонные острова - белый, SW - черный. Выравнивания SW получены с линейными параметрами. Данные для логарифмического домена параметров такие же.

На Рис 6а представлены гистограммы распределения весов островов эталонных выравниваний и выравниваний SW. Оказалось, что довольно много эталонных островов имеют очень низкий или даже отрицательный вес, в то время как алгоритмические выравнивания совсем не содержат островов малого веса. Стоит отметить, что суммарная длина таких «слабых» островов довольно велика (Рис. бб). Эталонные острова веса меньше 5 составляют

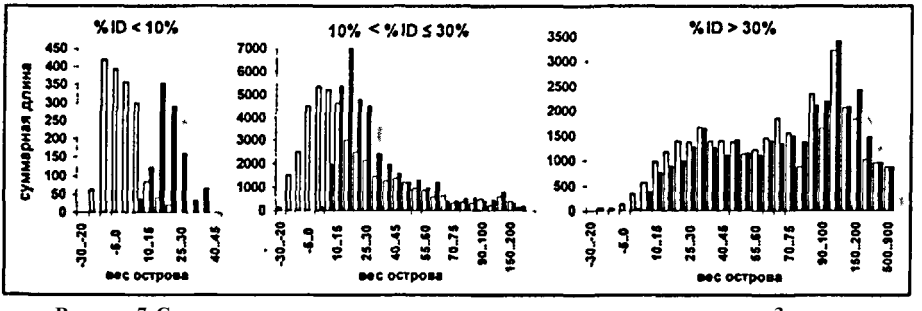


Рисунок 7 Суммарная длина островов, имеющих вес в пределах диапазонов, для 3-х диапазонов *%ID* (более детальное рассмотрение рисунка ба). Данные по эталонным островам показаны белым, SW - черным. Выравнивания SW получены с «линейными» параметрами».

32Ko от всех островов и покрывают 20% всей длины эталонных выравниваний, острова веса < 5 осшвляют 65% от всех потерянных островов и покрывают 63% суммарной длины потерянных оарэовов. Только *5%* островов такого малого веса были восстановлены алгоритмом. Для вцвавниваний из серой зоны **(10 < %ID ≤ 30)** картина еще более критическая - восстановлено всвго 2.5% островов веса меньше 5. В частности, наличие эталонных островов малого веса гоктрит о неприменимости матрицы замен для всей длины выравнивания.

На Рис. 7 видно, что с увеличением гомологи сравниваемых белков различие в весе этжионных и построенных островов уменьшается.

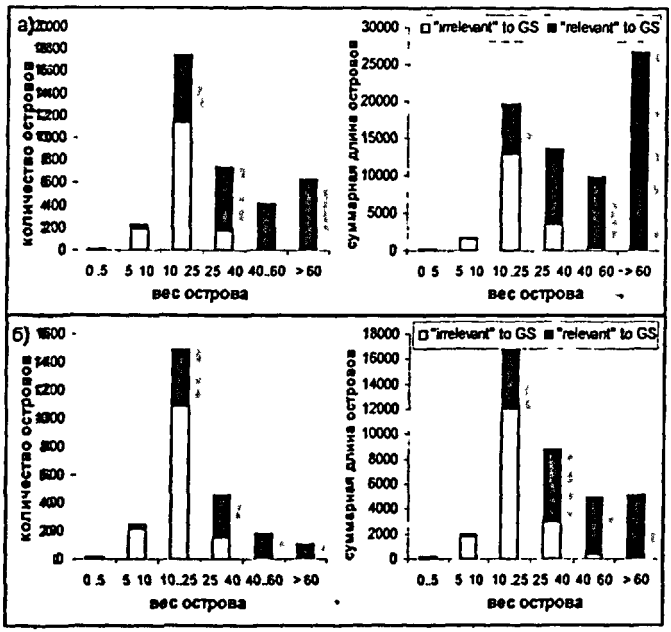


Рисунок 8.
Количество и полная длина похожих (черный) и непохожих (белый) островов выравниваний SW на эталонные острова, которые имеют вес в указанных диапазонах. Похожими островами SW мы считаем такие, у которых хотя бы одно сопоставление идентично сопоставлению в эталонном выравнивании, непохожими - все остальные.
Данные посчитаны для линейного домена параметров, логарифмические параметры дают похожие результаты.

Из Таблицы 2 следует, что «найденные» и «потерянные» острова сильнее различаются по весу, чем по длине. Наши данные показывают, что острова большого веса в основном достоверны. Даже если выравнивание относится к серой зоне, существуют острова, которым можно доверять (см. Рис 8). Для выравниваний уровня гомологии 10-30% в случае линейного набора параметров только 4% островов SW с весом больше 40 не имеют ничего общего с эталонным выравниванием, и только 15% из них содержит меньше чем 2/3 правильно выровненных позиций.

Рисунки 5,6,7 демонстрируют, что качество выравниваний определяется суммарной длиной потерянных островов эталонного выравнивания. В свою очередь, вероятность восстановления эталонного острова обычно зависит только от его веса (см. Таблицу 2).

Таблица 2 Зависимость количества «найденных» и «потерянных» островов от их веса и длины.

	1 - 5		6 - 10		11 - 15		16 - 20		21 – 25		26 - 30		31 - 35		Длина
30 .. 35			23	0	48	0	40	0	19	0	8	1	13	0	
25 .. 30	2	0	29	1	55	2	46	5	18	2	11	1	13	2	
20 .. 25	17	2	46	3	54	10	36	4	26	5	14	3	7	1	
15 .. 20	20	7	89	22	60	22	39	6	18	4	9	6	3	3	
10 .. 15	20	64	89	69	57	50	50	24	11	13	12	2	3	3	
5 .. 10	15	79	46	138	34	99	13	43	11	15	7	8	4	1	
0 .. 5	6	92	18	217	13	129	14	45	0	16	1	9	0	2	
-5 .. 0	6	92	3	189	5	110	3	51	3	13	1	1	0	0	
-10 .. -5	0	23	0	100	3	68	0	29	0	12	0	4	0	4	
-15..-10	0	6	1	21	0	27	0	22	0	9	0	2			
-20..-15			0	7	0	5	0	8	0	4	1	0			
-25..-20					0	1	0	2	0	1					
-30..-25									0	1					
вес	Н	П	Н	П	Н	П	Н	П	Н	П	Н	П	Н	П	

Комментарии к Таблице 2: Данные для «потерянных» островов находятся в колонках, помеченных буквой «П», «найденные» - помечены буквой «Н». В таблице представлены данные для островов с весом и длиной меньше 35. Острова весом больше 35 все найдены. Количество островов длины больше 35 и весом меньше 35 пренебрежимо мало. Пустые ячейки означают отсутствие островов с таким весом и длиной. В затемненных ячейках количество потерянных островов превосходит количество найденных.

Достоверность островов (*Isl_Conf*) напрямую связана с достоверностью (*Ali_Conf*) выравнивания SW (Таблица 3). 411 выравниваний SW содержат, по меньшей мере, один остров с весом больше 40, для 338 из них *Ali_Conf*> 60%, и только у 73 таких выравниваний *Ali_Conf*< 60%. Среди выравниваний с *Ali_Conf* > 60% нет островов, не имеющих правильных сопоставлений, и только 3% из них имеют достоверность меньше 2/3. В то время как выравнивания с *Ali_Conf* < 60% содержат 14% островов с достоверностью 0, и 31% островов содержит меньше чем 2/3 правильно выровненных позиций. Дм выравниваний, относящихся к серой зоне, данные примерно такие же.

Таблица 3. Связь между достоверностью островов и достоверностью выравниваний SW. Данные представлены для линейных параметров.

а) Ali_Conf <60%	*Кол-во выравниваний*	*Кол-во островов веса > 40*	*Кол-во островов с Isl_Conf=0*	*%%*	*Кол-во островов с Isl_Conf < 2/3*	*%%*
All	73	101	14	13.9	31	30.7
10% - 30%	69	97	13	13.4	29	29.9
б) Ali_Conf >60%	*Кол-во выравниваний*	*Кол-во островов веса > 40*	*Кол-во островов с Isl_Conf=0*	*%%*	*Кол-во островов с Isl_Conf < 2/3*	*%%*
All	338	870	0	0	26	3.0
10% - 30%	81	168	0	0	8	4.8

Назовем весом ядра острова (в работе [3] есть похожие объекты - HSP) максимальный вес фрагмента острова. Острова малого веса можно разделить на два класса:

1) вес острова распределен равномерно, так что невозможно выделить в нем участок с существенно положительным весом. Следовательно, никакой алгоритм, использующий данную матрицу замен, не способен воспроизвести этот остров даже частично;

2) острова, в которых есть непрерывные участки существенно положительного веса (якоря).

Алгоритм SW не находит большинство островов малого веса вне зависимости от их типа, однако можно предложить алгоритм выравнивания последовательностей, который сможет восстанавливать хотя бы эталонные ядра.

а) Веса сопоставлений: -3+1-2-4+1+7+1-1+3-4-4-1-2+1-3-4 *Полный вес острова=-14*

v s f k d G d a i i n v q a i d
q a a w q G q i v g w y c t n l

вес ядра = +11

б)

	< 0	0 .. 5	5 .. 10	10 .. 15	15.. 20	20.. 25	25..30	30..35	35+	Вес ядра
> 35									655 1	
30 .. 35								85 0	75 1	
25 .. 30							74 9	88 4	14 1	
20 .. 25						83 12	86 14	32 1	6 1	
15 .. 20					103 43	101 21	30 6	8 1	0 0	
10 .. 15				71 158	106 57	53 10	10 0	2 1	0 2	
5 .. 10			27 153	*60 199*	*29 31*	*11 5*	*3 0*			
0 .. 5		5 124	21 266	*22 113*	*4 7*	*0 1*				
-5 .. 0	0 16	8 230	9 174	*2 37*	*2 2*					
-10 .. -5	0 14	2 148	1 65	*0 13*	*1 1*					
-15 .. -10	1 6	0 53	0 23	*0 5*	*0 0*					
-20 .. -15	0 1	0 18	0 3	*0 2*	*1 0*					
-25 .. -20		0 2	0 1	*0 1*						
-30 .. -25		0 1		*0 1*						
Вес острова	Н П	НП	Н П	Н П	Н П	Н П	Н П	Н П	Н П	

Рисунок 9.
а) Остров эталонного выравнивания белков lark и Ivie (коды PDB). Веса сопоставлений указаны согласно матрице замен Gonnet250 с округлением до целых чисел.
б) Таблица зависимости количества «найденных» (Н) и «потерянных» (П) островов от полного веса острова и веса его ядра. Жирным шрифтом выделены острова веса меньше 10, но имеющие ядра веса > 10.

На Рис. 9а представлен остров эталонного выравнивания белков lark и Ivie (коды PDB). Этот остров имеет отрицательный вес равный -14, и таким образом не восстановлен методом SW. Однако этот остров содержит ядро длиной 5 сопоставлений, вес которого равняется +11 (выделено подчеркиванием).

Таблица, представленная на Рис. 96, отражает зависимость количества «найденных» и «потерянных» островов от их веса и веса содержащихся в них ядер. Жирным шрифтом выделены данные для островов, содержащих ядра значительного веса, но имеющие низкий суммарный вес. Алгоритм выравнивания последовательностей, который основан на работе с ядрами, в принципе, может распознавать такие ядра. Следовательно, такая техника не должна приводить к уменьшению качества восстановления эталонного выравнивания и даже может немного улучшить его. В то же время, безделяционные участки, вес которых больше обычно используемого штрафа за открытие делеции, составляют малую часть матрицы Нидельмана-Вунша [4]. Таким образом, алгоритм, который будет сканировать разреженную матрицу, может работать значительно быстрее.

В четвертой главе представлен новый алгоритм выравнивания последовательностей, основанный на использовании понятия ядер.

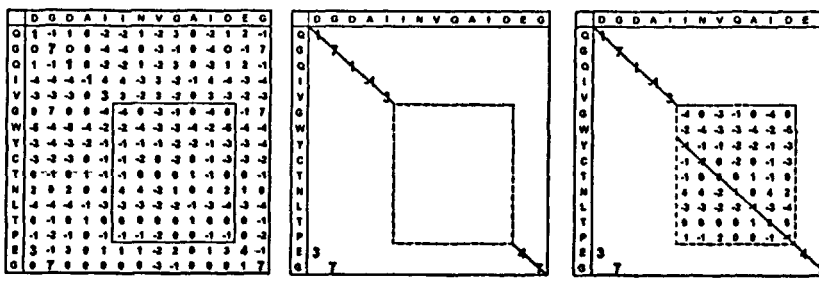
Алгоритм, основанный на ядрах, схематически представлен на рис. 10:

1. Строим все безделяционные участки с весом больше специального порога T > 0 (подобно алгоритму, описанному в [5]). Далее будем называть такие участки якорями. В конечном итоге некоторые из этих якорей сформируют ядра островов построенного выравнивания. Построение якорей начинается с нахождения начальных пар (затравок) с весом больше заданного, которые потом расширяются до полных якорей. Этот шаг похож на

процедуры, используемые в BLAST и FASTA. Посторонние якорей занимает большую часть времени построения выравнивания.

2. Находим оптимальный путь через якоря (все элементы матрицы Нидельмана-Вунша кроме якорей обнуляются). Есть две особенности, отличающие весовую функцию, которую мы оптимизируем, от стандартной. Первая: матрица замен используется только для покрытых якорями участков. Вторая: вместо штрафа за количество делеций штрафуется количество используемых якорей на пути выравнивания. В частности, мы штрафуем соединение между якорями, даже если они находятся на одной диагонали матрицы Нидельмана-Вунша.

3. Строится путь выравнивания в участках матрицы Нидельмана-В} нша между якорями, закрепленными на предыдущем шаге. На этом этапе используется изначальная матрица замен и стандартный метод SW.



I. II. III.

Рисунок 10. Основные шаги нового предложенного алгоритма: I. Построение якорей; II. Вычисление оптимального пути через якоря; III. Установление пути выравнивания между якорями.

Построение якорей может производиться двумя способами. При первом, более точном, но существенно более медленном, просматриваются все диагонали с тем. чтобы обнаружить участки, обладающие следующими свойствами: суммарный вес не меньше порога (*Ancor_Weight*), никакое расширение не приносит увеличение веса, внутри якоря не допускаются участки суммарного веса меньше *Anc_Hole_Weight*. Время работы этой процедуры пропорционально произведению длин выравниваемых последовательностей.

Более быстрое построение якорей (второй способ) в свою очередь идет в 3 этапа:

а) Для каждой пары аминокислот X_1, X_2 находим все такие пары Y_1, Y_2, чтобы сумма $M(X_1, Y_1) + M(X_2,Y_2)$ весов их сопоставлений по заданной матрице замен M превосходила порог *Cseed*. Результаты заносятся в таблицу *Pair_Table*. Этот шаг не зависит от сравниваемых последовательностей, следовательно таблицу *Pair_Table* можно вычислить один раз для используемой матрицы замен.

б) Для каждого дипептида X_1, X_2 при помощи таблицы *Pair_Table* находим все такие позиции в первой последовательности, сопоставление которых с данным дипсптидом дает вес не менее порога *Cseed*. Чтобы получить таблицу вхождения дипептидов из *Pair_Table* в первую

последовательность (*Seq1_Table*), достаточно одного прохода вдоль последовательности, при ко юром » *Seq1_Table* заносятся соответствующие данные согласно таблице *Pair_Table*. Этот iiiai дополыо продолжительный по времени, но при поиске по банку данных выполняется всего один раг

в) При помощи таблицы *Seq1_Table* находим позиции всех парных затравок первой последовательности для каждой позиции Seq2[k] во второй последовательности. Затравкой мы называем такую пару позиций <m,k>, для которой M(Seq1[m], Seq2[k])+ M(Seq1[m+1], Seq2[k+1]) > Cseed '(M - используемая матрица замен). Каждую такую пару пытаемся расширить до полного якоря. В данном случае также требуется всего один проход вдоль второй последовательности.

Точность и достоверность нашего метода были протестированы на ~13 000 парных структурных выравниваний, извлеченных из баз данных Balibase и FSSP. База Balibase использовалась для оптимизации алгоритма и его параметров, FSSP - как независимый тест для того, чтобы исключить зависимость от одной базы данных.

Результаты тестов показали, что новый метод примерно равен методу SW и в точности, и в достоверности. Время работы нашей программы сравнивалось со стандартной реализацией алгоритма SW, извлеченной из программного пакета FASTA. В таблице 4 • представлены сравнительные характеристики нашего метода (ANCHOR) и метода Смита-Уотермана. Видно, что предложенный метод требует примерно в два раз меньше времени, чем SW.

Таблица 4. Сравнение среднего времени работы, точности и достоверности парных выравниваний, построенных методом Смита-Уотермана (SW) и новым предложенным алгоритмом ANCHOR (Anch).

%ID	N пар	Ali_Acc, %			Ali_Conf, %			Время работы		
		Anch	SW	Anch/SW	Anch	SW	Anch/SW	Anch	SW	Anch.SW
а)										
10-30%	298	36±20	35±23	1.02	52±26	47±27	1.11	324	734	0.44
>30%	253	84±7	84±7	1.00	88±6	87±7	1.03	126	406	0.31
ALL	583	41±24	40±25	1.01	54±26	50±28	1.10	226	560	0.40
б)										
10-30%	2605	52±16	54±16	0.96	68±16	65±16	1.06	72	179	0.40
>30%	10049	99±1	99±1	1.00	99±1	99±1	1.00	64	174	0.37
ALL	12671	77±19	78±19	0.98	85±15	83±15	1.01	66	175	0.38

Данные представлены для двух баз данных: BaliBase (а) и FSSP (б).

Время работы алгоритмов технически измерялось двумя способами: используя стандартную функцию *clock()* макроязыка Си (результаты представлены в Таблице 4), и при помощи встроенного измерителя времени работы процедур *profiler* системы разработки программ MS-VisualStudio (VisualC++ 6.O). Известно, что процедура *clock()* дает сильную погрешность в измерении времени. Поэтому процедуры построения выравниваний были заключены в цикл с повторением до 100 раз для увеличения точности измерения. Из Таблицы 5 видно, что расхождение данных по измерению времени *profiler* и *clock()* невелико.

Таблица 5. Сравнение времени работы алгоритмов ANCHOR и SW для парных выравниваний при помощи стандартной функции *clock()* макроязыка Си (столбец clock() Time) и встроенного измерителя времени работы процедур *profiler* системы разработки программ VisualC++ 6 0 (столбец Func+Child Time). Использовалась база данных BaliBase и логарифмические параметры (см. Методы).

Функция	Кол-во вызовов	Func+Child Time	clock() Time
_compute_sw_alignment (SW)	5830	12812	559 8
_hexograph_build (Anch)	5830	4835	225 6
Anch/SW		0 38	0 40

Мы так же сравнили точность **и** достоверность нашего алгоритма **с** программой BLAST [ftp://ftp.ncbi.nih.gov/blast/executables/blastz.exe] (Таблица 6). Видно, что для белков из серой зоны наша программа значительно превосходит BLAST как по достоверности, так и по точности выравниваний. Однако известно, что BLAST требует примерно в 15 раз меньше времени **на** построение выравнивания, чем метод SW.

Таблица 6. Сравнение средней *Ali_Acc* **и** *Ali_Conf* выравниваний, построенных новым алгоритмом (Anch), методом Смита-Уотермана (SW) и программой BLAST для белков из базы BaliBase.

%ID	N nap	Ali_Acc, %			Ali_Conf, %		
		Anch	SW	BLAST	Anch	SW	BLAST
10% - 30%	298	36±20	35±23	27±20	52±26	47±27	45±29
> 30%	253	84±7	84±7	82±8	88 ±6	87±7	87±6
ALL	583	41±24	40±25	32±24	54±26	50±28	48±29

В пятой главе дается расширение нашего алгоритма для **выравнивания последовательности и профиля.**

Данная методика также применима для построения выравниваний профиля и последовательности. Основное отличие построения парного выравнивания и выравнивания последовательности **с** профилем **в** этом случае заключается в построении якорей. В случае выравнивания профиля и последовательности применяется следующая процедура:

а) Для каждого дипептида X_1, X_2 находим все такие позиции профиля, сопоставление которых с данным дипептидом дает вес не менее порога *Cseed.* Результаты заносятся в таблицу *Prof_Table,* аналогичную *Seq1_Table* для парного выравнивания. Для более быстрой работы алгоритма в каждой колонке профиля элементы упорядочиваются по убыванию. Так как в профиле как бы совмещены матрица замен и первая последовательность, то этому пункту соответствуют одновременно пункты *«а»* и *«б»* из алгоритма для последовательностей.

б) Этот пункт полностью аналогичен пункту *«в»* для парных выравниваний. При проходе вдоль второй последовательности находим позиции всех парных затравок из таблицы *Prof_Table.* Каждую такую пару пытаемся расширить до полного якоря. В данном случае также требуется один проход вдоль последовательности.

Параметры для выравниваний профиля и последовательности были получены из логарифмических параметров парных выравниваний. Они были уменьшены пропорционально отношению дисперсии матрицы замен Gon250 к средней дисперсии «стандартного» профиля.

Мм использовали профили, полеченные in множественных выравниваний Gai данных HaliBase и FSSI* метолом, предложенным *Сюняевыи Ш.Р. и др.* [6]

Таблица 7. Сравнение среднего времени работы, точности и достоверности выравниваний профиля с последовательностью, построенных методом Смита-Уотермана {SW) и алгоритмом ANCHOR (Anch).

%ID	N	Ali_Acc, %			Ali_Conf, %			Время работы		
		Anch	SW	Anch/SW	Anch	SW	Anch/SW	Anch	SW	Anch/SW
а)										
10-30%	85	31±12	28±14	1.09	60±18	60±23	1.00	498	1974	0 25
>30%	45	79±11	84±11	0.94	90±6	88±8	1.02	98	617	0.16
ALL	135	41±16	39±19	1.03	68±19	69±24	0.98	357	1486	0.24
б)										
10-30%	42	40±11	43±13	0 93	80±13	82±9	0 98	26	74	0 35
>30%	261	99±2	99±1	1 00	99±1	99±1	1.00	23	82	0 28
ALL	304	76±17	75±17	1 01	93±9	94±8	0.99	24	81	0 29

Данные представлены для двух баз данных: BaliDase (а) и FSSP (б).

Таблица 7 для выравниваний профиля и последовательности аналогична Таблице 4 для парных выравниваний. Она демонстрирует, что. качество выравниваний профиля и последовательности, построенных методом Смита-Уотермана и новым методом, практически совпадает, а скорость работы метода ANCHOR существенно выше.

Для надежности, замеры времени работы алгоритмов также проводились двумя методами: функцией ***clock()*** и методом ***profiler*** оболочки VisualC++ **(Таблица 8).** Оба метода измерения времени дают погрешность -10% для операционных систем семейства Windows. В итоге время работы программы может варьироваться в существенных пределах, однако, как видно из Таблицы 8, даже в самом худшем случае время работы нашего алгоритма как минимум в 2 раза быстрее стандартного SW.

Таблица 8. Сравнение времени работы алгоритмов ANCHOR и SW для выравниваний профиля и последовательности при помощи стандартной функции ***clock()*** макроязыка Си (столбец clockO Time) и встроенного измерителя времени работы процедур ***profiler*** системы VisualC++ 6.0 (столбец Func+Child Time). Использовалась база данных BaliBase и логарифмические параметры (см. Методы).

Функция	Кол-во вызовов	Func+Child Time	clock() Time
compute_sw_prof_alignment (SW)	1350	5790	1485.9
hexograph_prof_build (Anch)	1350	2215	356.9
Anch/SW:		0 38	0 24

В тестон главе исследуется применимость нового алгоритма выравниваний для поиска гомологов по банку данных.

Основным применением программ выравнивания в современной биоинформатике является поиск гомологов данного белка (запроса) по банку данных. Мы проверили применимость нашего метода для этих целей. При этом были внесены следующие изменения: (1) использовалась целочисленная матрица замен (как и во всех основных программах поиска); (2) учитывался только вес выравнивания якорей, последняя стадия работы алгоритма, построение выравнивания между якорями, не выполнялось.

В качестве логарифмических параметров использовались: матрица, замен Gonnet (целочисленная), минимальный вес якоря =20; вес парных затравок - 8; штраф за сход с ядра = 15; за увеличение делеции = 1.

Для тестирования был использован банк данных SCOP40. В качестве запросов использовались последовательности, которые имеют не менее 10 гомологов в этом банке. Гомологами считаются все члены суперсемейства по классификации SCOP.

Процедура тестирования программы состояла в следующем. Каждый белок-запрос выравнивался со всеми остальными последовательностями из банка SCOP40. По полученным весам выравниваний высинивались ечнчвистующие *E-value* (аналогично процедуре, используемой в FASTA). Все последовательности банка упорядочивались в список по возрастанию *E-value.* В идеале, все гомологи запроса должны оказаться в верхней части списка.

Для оценки - качества списка использовались 3 меры: ***Ml, M2, M3.*** Во всех случаях возможные значения мер лежат между 0 и 1, идеальному списку соответствует значение 1. Ниже даны определения каждой из этих мер.

Положим - $True(n) = 1$, если n-ый элемент списка - гомолог данного белка-запроса, и $True(n) = 0$ в противном случае. Пусть N_{hom} - количество гомологов запроса в базе данных. *Верхней частью списка* будем называть его первые Nhom элементов.

Величина *Ml* - это доля гомологов запроса в верхней части списка. Т.е. $M1 = \frac{\sum_{n=1}^{N_{hom}} True(n)}{N_{hom}}$.

Величина *M2* - это отношение количества гомологов, стоящих подряд в начале списка, к величине N_{hom}. Более формально: пусть $n_{first\ wrong}$ - номер в списке первого из белков, негомологичных запросу. Тогда $M2 = (n_{first\ wrong} - 1)/N_{hom}$.

Величина *M3* учитывает насколько далеко от начала списка расположены гомологи запроса. Пусть N_{Main} - длина выходного списка гомологов, т.е. количество «значимых» первых позиций в списке. Белок, оказавшийся на n-ом месте в списке $(n \le N_{Main})$, получает оценку $S(n) = N_{Main} - n + 1$, т.е. N_{Main} за первое место, $N_{Main}-1$ - за второе и т.д. Мера *M3* равняется отношению суммы оценок, полученных гомологами запроса, к максимально возможной сумме оценок (т.е. когда все гомологи находятся и начале списка): $M3 = \frac{\sum_{n=1}^{N_{Main}} (N_{Main} - n) * True(n)}{\sum_{n=1}^{N_{Main}} (N_{Main} - n)}$.

Мы полагали $N_{Main} = 3*N_{hom}$. В этом случае, как правило, все гомологи оказывались в значимом списке. Для всех программ, использовавшихся в тестах, N_{Main} передавался как параметр. Исключение составляет BLAST, и котором ого число нычисляскя анюмашчески внутри программы и. как правило, меньше $3*N_{hom}$.

Усредненные по всем запросам данные приведены в Таблице 9. Мы сравнивали наш метод поиска с полной реализацией метода Смита-Уотермана в пакете FASTA (модуль dropnsw.c), с программой SSEARCH - упрощенным вариантом метода SW, строящим, как правило, такое же выравнивание (предложен *Phil Green,* реализован в программном комплексе FASTA, модуль dropgsw.c), а так же с программами FASTA и BLAST.

Таблица, 9. Сравнение времени обработки запроса и качества полученных списков гомологов программами для поиска по банку данных FASTA, BLAST, SSEARCH (упрощенный метод SW, используемый в FASTA пакете), SW_FASTA (строгая реализация метода Смита-Уотермана в пакете FASTA) и нового метода Search-Anchor.

программа	*Время работы*	N_{Main}	*M1*	*M2*	*M3*
Search-Anchor	2218.1	124.4	0.237	0.158	0.286
SSEARCH (быстрый SW)	3449.6	124.4	0.245	0.162	0.296
SW_FASTA	6627.0	124.4	0.245	0.162	0.296
FASTA	729 5	124.4	0.204	0.139	0.243
BLAST	162.9	12.9	0.159	0.131	0.236

Время работы - среднее время сканирования банка данных для одного запроса;
N_{Main} – среднее количество выдаваемых ответов (для всех, кроме BLAST, задается равным 3* Nhom);
$N_{hom} = 41$ – среднее количество гомологов для запроса в базе данных по всем запускам;
M1, M2, M3 – средние значения указанных мер качества списков (см. выше).

Из таблицы видно, что новый метод значительно превосходит программы FASTA и BLAST по качеству списка гомологов, однако работает значительно медленнее. Метод Смита-Уотермана (как в полной, так и в упрощенной реализации) дает самые точные результаты, но работает в 2 раза для полной версии и 1.5раза для упрощенной версии медленнее, чем новый метод Search-Anchor. Время работы алгоритмов (Search-Anchor, Смита-Уотермана, FASTA, BLAST) замерялось как полное время обработки одного запроса (query) вдоль всего банка последовательностей с последующим усреднением по всем запросам.

Таблица 10. Распределение времени работы между разными частями программ.

программа	процедура	% от общего времени работы, ± 5%
Search-Anchor	Процедура выравнивания (hier.c)	76
	чтение входных файлов	15
	обработка и вывод результатов	9
SW_FASTA	Процедура выравнивания (dropnsw.c)	83
	чтение входных файлов	7
	обработка и вывод результатов	10
SSEARCH (быстрый SW)	Процедура выравнивания (dropgsw.c)	70
	чтение входных файлов	18
	обработка и вывод результатов	12
FASTA	Процедура выравнивания (dropffa.c)	63
	чтение входных файлов	20
	обработка и вывод результатов	17

Такой способ измерения времени работы вполне оправдан, т.к. основное время занимает именно построение выравниваний, а остальные обслуживающие процедуры примерно равны по времени для всех программ и непродолжительны. В Таблице 10 представлено процентное

соотношение времени построения выравнивай и остальных частей программ. Время, которое алгоритм Search-Anchor тратит собственно на выравнивание последовательностей, составляет -30 % от времени выравнивания методом SW-FASTA и ~70% от времени построения-выравнивания программой SSEARCH.

ВЫВОДЫ.

1. Получена зависимость надежности восстановления выравнивания пространственных структур (эталонных выравнивании) по аминокислотным последовательностям белков методом Смита-Уотермана. Показано, что индивидуальный подбор штрафов за делеции существенно (более чем на 10%) увеличивает точность выравниваний для белков с уровнем сходства 10-30% (серая зона). Однако даже в этом случае средняя точность выравниваний для этого диапазона *%Ю* не превышает 52%, а достоверность - 70%.

2. Обнаружены различия алгоритмических и структурных выравниваний на уровне внутренней структуры «островов» (безделеционных участков выравнивания). В выравниваниях SW восстановлено 53% островов, из которых 42% приходится на острова, которые угаданы на 90% и более. Потерянных островов 47%, они имеют малый вес и длину. Показано, что 32% островов эталонных выравниваний имеет вес < 5, суммарная длина таких островов составляет 20% всей длины эталонных выравниваний, потерянные острова веса < 5 оставляют 65% от всех потерянных островов и покрывают 63% суммарной длины потерянных островов. Только 5%< островов такого малого веса были восстановлены алгоритмом. Для выравниваний из серой-зоны эти цифры аналогичны, однако восстановлено только 2.5% островов с весом меньше 5, и потерянные острова оставляют 65% от общего количества эталонных островов. Проблемы с восстановлением островов малого веса являются причиной недостаточной точности алгоритмических выравниваний.

3. Разработан новый алгоритм ANCHOR выравнивания последовательностей, который при построении выравнивания учитывает не весь вес острова, а только его положительную основу (ядро). Показано, что новый алгоритм не уступает в качестве восстановления структурного выравнивания методу Смита-Уотермана, но работает примерно в 2 раза быстрее.

4. Новый алгоритм ANCHOR адаптирован для построения выравнивания последовательности и профиля. Показано, что он так же хорошо восстанавливает эталонные выравнивания, как и SW, но работает быстрее как минимум в 2 раза.

5. На основе нового алгоритма разработана программа поиска гомологов по банку данных (Search-Anchor). Новая программа выдает более точный список гомологов, чем FASTA и BLAST, но работает медленнее более чем в 3 раза. Новый алгоритм работает быстрее алгоритма Смита-Уотермана, незначительно уступая ему в качестве.

Список публикаций по теме диссертации:

1. S.R.Sunyacv, G.A.Bogopolsky, N.V.Oleynikova, P.K.VIasov, A.V.Finkelstein, M.A.Roytberg. From Analysis of Protein Structural Alignments Toward a Novel Approach to Align Protein Sequences. PROTEINS: Structure, Function, and Bioinformatics, 2004, 54(3), 569-582.
2. S.R.Sunyaev, G.A.Bogopolsky, N.V.Oleynikova, P.K.Vlasov, A.V.Finkelstein, M.A.Roytberg. Anchor-based alignment of sequences and profiles: accurancy and effectiveness, Proceedings of the international Moscow conference on computational molecular bioiogy, MCCMB'03 Moscow Russia 22-25, July, 2003, p.222-224.
3. GA.Bogopolsky, A.V.Finkelstein, N.V.Oleinikova, M.A.Roytberg, S.R.Sunyaev, P.K.Vlasov. How similar are aminoacid sequences of the proteins with the common fold?, V International congress of mathematical modeling, Dubna, Russia, Sept. 30-Octob. 6,2002, v.2, p.191-192.
4. N.V.Oleynikova, CA.Bogopolsky, PXVlasov, Sh.R.Sunyaev, M.A.Roytberg. Accuracy of the pairwise protein sequence alignment: From the observations to a new approach. Artificial Intelligence and Heuristic Methods for Bioinformatics, NATO Advanced Studies Institute, San Milano, Italy, 1-11 October 2001, p. 19.
5. Г.В.Богопольский, П.К.Власов, Н.В.Олейникова, М.А.Ройтберг Ш.Р.Сюняев. Определение сходства пространственных структур белков на основе сопоставления их аминокислотных последовательностей. Сборник отчетов по ГНТП "ГЕНОМ ЧЕЛОВЕКА-2000", Москва, 2001.С.145.
6. Г.В.Богопольский, П.К.Власов, Н.В.Олейникова, М.А.Ройтберг, Ш.Р.Сюняев. Распознавание типа укладки белка с помощью многокритериального выравнивания первичных структур белков с профилем семейства. Сборник отчетов по ГНТП "ГЕНОМ ЧЕЛОВЕКА-1999", Москва, 2000, с.38-39.

ЛИТЕРАТУРА

1. Vogt G, Etzold T, Argos P. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.* 1995; 249: 816-831.

2. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.* 2000; 297:1003-1013.

3. Altschul SF, Gish W. Local alignment statistics. *Methods Enzymol.* 1996; 266: 460-480.

4. Needlman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. *Mol. Biol.* 1970; 48:443-453.

5. Altschul SF, Gish W, Miller W, Myers E, Lipman Ш. Basic local alignment search tool. *J. Mol. Biol.* 1990; 215:403-410.

6. Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., and Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12, 387-94.

Олейникова Наталья Васильевна

ВЫРАВНИВАНИЕ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ: АНАЛИЗ СУЩЕСТВУЮЩИХ МЕТОДОВ И РАЗРАБОТКА НОВЫХ АЛГОРИТМОВ

Подписано в печать 06.02.04. Формат 60 x 84 1/16. Печать офсетная.
Усл. печ. л. 1,0. Уч.-изд. л. 1,0. Тираж 70 экз. Заказ № ф-131

Государственное образовательное учереждение
Высшего профессионального образования
Московский физико-технический институт (государственный университет)
Отдел автоматизированных издательских систем «ФИЗТЕХ-ПОЛИГРАФ»
141700, Моск. обл., г. Долгопрудный, Институтский пер.. 9