Информационные процессы, Том 11, № 4, 2011, стр. 510–519.

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ

Использование программы PARCA для глобального выравнивания аминокислотных последовательностей[1]

И. В. Поверенная*, М. А. Ройтберг, В. В. Яковлев****

**Московский Государственный Университет им. М. В. Ломоносова, Москва, Россия*
***Институт математических проблем биологии, Российская академия наук, Пущино, Россия*

Поступила в редколлегию 21.11.2011

Аннотация—Наиболее точным из широко используемых в настоящее время алгоритмов глобального выравнивания последовательностей является алгоритм Смита-Ватермана. Однако, в случае выравнивания слабогомологичных последовательностей, даже этот метод не дает хороших результатов. Ранее мы предложили новый подход к задаче глобального выравнивания и на модельных выравниваниях продемонстрировали, что, как правило, он строит выравнивания более близкие к эталонным, чем выравнивания Смита-Ватермана. В настоящей работе мы сравниваем работу нашего алгоритма и алгоритма Смита-Ватермана, используя базу эталонных выравниваний реальных белков PREFAB-P; выравнивания в этой базе получены наложением пространственных структур белков. Результаты сравнения подтверждают преимущества предложенного подхода по сравнению с алгоритмом Смита-Ватермана.

1. ВВЕДЕНИЕ

В настоящее время основным способом сравнения аминокислотных последовательностей в целом является глобальное выравнивание последовательностей. Для решения этой задачи известно несколько алгоритмов [1]. Для приложений важно, насколько алгоритмически построенное выравнивание близко к «биологически корректному» выравниванию, которое отражает ход эволюции. Здесь имеется в виду, что в выравнивании сопоставлены те позиции сравниваемых последовательностей, которые соответствуют одной и той же позиции общего предка этих последовательностей. Так как ход эволюции в точности неизвестен, в качестве эталонов выравниваний белковых последовательностей обычно используются выравнивания, полученные наложением пространственных структур белков [2].

Под *точностью* алгоритмического выравнивания понимается доля таких позиций эталонного выравнивания, которые успешно восстановлены алгоритмом [2]. Другой характеристикой близости алгоритмического выравнивания к эталону является достоверность – доля таких позиций алгоритмического выравнивания, которые присутствуют в эталонном выравнивании [2]. Наши исследования (см. вспомогательные материалы к статье по адресу [3]) показали, что в случае глобального выравнивания значения этих характеристик скоррелированы. Поэтому в настоящей работе мы исследуем только точность выравниваний.

Наиболее точным из широко используемых в настоящее время алгоритмов парного глобального выравнивания является алгоритм Смита-Ватермана [4]. Он основан на поиске оптимального значения функции $W = m - \alpha \cdot g - \beta \cdot d$, где m – это суммарный вес за сопоставления символов, g и d – соответственно количество удаленных фрагментов и их суммарная длина, α и β – соответствующие штрафы. Эти штрафы являются числовыми параметрами, которые определяются эмпирически, теоретической базы для их выбора не существует [5]. Однако, для

[1] Работа выполнена при финансовой поддержке государственного контракта № 07.514.11.4004

слабогомологичных последовательностей (доля совпадений не более 30%) даже выравнивания Смита-Ватермана имеют невысокую точность [2, 6, 7].

В работах [8,9] и ряде других было показано, что точность выравнивания аминокислотных последовательностей можно существенно повысить, используя данные о вторичной структуре белков. Другим источником повышения точности выравниваний является одновременное сравнение нескольких последовательностей [10]. В то же время, как с практической, так и с методической точки зрения, представляет интерес задача разработки алгоритма, который позволяет строить более точные выравнивания двух слабогомологичных последовательностей, чем алгоритм Смита-Ватермана, и не использует сведений о вторичной структуре белков. Этот интерес, в частности, связан с выравниванием последовательностей не белковой природы.

Для решения указанной задачи в работах [11, 12] была предложена идея использовать различные весовые функции на различных участках выравнивания. На нескольких примерах были показаны возможности такого подхода, однако ни в одной из этих работ не было проведено массовое тестирование предложенных алгоритмов. В работе [2] представлен анализ выравниваний более 200 пар последовательностей белков различной степени сходства. На основании этого анализа был предложен новый иерархический алгоритм выравнивания, обобщающий подходы из [11, 12]. Однако, тестирование показало, что этот алгоритм, превосходя алгоритм Смита-Ватермана по быстродействию примерно в 2 раза, в среднем строит выравнивания примерно той же точности, что и алгоритм Смита-Ватермана.

Наиболее полное (к настоящему моменту) воплощение идея различной обработки высокогомологичых и низкогомологичных участков сравниваемых последовательностей получила в работах Huang и соавторов [13,14]. Однако, как было показано в работе [7], программа GAP3, реализующая подход Huang и соавторов, строит выравнивания, превосходящие по точности выравнивания Смита-Ватермана только тогда, когда сходство между сравниваемыми последовательностями носит локальный характер. В работе [7] это было показано при сравнении модельных последовательностей. Мы перепроверили это, сравнивая реальные белки из базы эталонных выравниваний PREFAB-P [15], и получили аналогичный результат (см. сопроводительные материалы [3]). Отметим, что во всех цитированных работах, как и в алгоритме Смита-Ватермана, выбор параметров, определяющих штрафы за удаление фрагментов, основан на эмпирических соображениях.

Другой подход к задаче построения точных глобальных выравниваний основан на построении не одного выравнивания, а семейства выравниваний. При таком подходе предполагается, что биолог-исследователь сможет проанализировать все предложенные выравнивания и, если среди них есть «хорошее» выравнивание, распознать его, используя дополнительные, возможно, экспериментальные методы. Поэтому под точностью такого набора естественно понимать наибольшую точность среди выравниваний набора. Говоря неформально, решается следующая задача

Дано:

Символьные последовательности S_1, S_2.

Построить:

Набор выравниваний (желательно, упорядоченный) последовательностей S_1, S_2 такой, что в нем содержится как можно более точное выравнивание. При этом, желательно, чтобы размер набора был небольшим, а «хорошее» выравнивание находилось как можно ближе к началу списка (если набор упорядочен).

Впервые такой подход был предложен в работах [16, 17]. В этих работах было указано на то, что мы не можем быть уверены в правильности выбора параметров выравнивания

и, следовательно (единственное!) построенное оптимальное выравнивание может быть нерелевантно решаемой биологической задаче. В качестве альтернативы была сформулирована задача построения всех выравниваний, вес которых мало отличается от веса оптимального выравнивания и предложен алгоритм решения этой задачи. К сожалению, в практике этот алгоритм почти не используется, так как для последовательностей длины 300 (характерная длина белков) могут существовать сотни и даже тысячи выравниваний, имеющих максимально возможный вес.

Нашей целью являлась разработка метода, который позволяет построить небольшое (~ 10 выравниваний) упорядоченное множество выравниваний, которое содержит выравнивание, имеющее большую точность, чем выравнивание Смита-Ватермана. Другой целью было исследование возможности отказа от использования штрафов за удаление фрагментов. Такой алгоритм был описан нами в работе [18]. На модельных данных было показано, что при рассмотрении множеств, содержащих не более 6 выравниваний, наш метод имеет в среднем лучшие значения точности, чем метод Смита-Ватермана. Целью настоящей работы является исследование поведения предложенного метода при выравнивании аминокислотных последовательностей реальных белков.

2. ПОСТАНОВКА ЗАДАЧИ И АЛГОРИТМ

2.1. Построение множества выравниваний-кандидатов

Построение набора выравниваний-кандидатов основано на многокритериальном подходе [19]. При этом подходе выравнивание характеризуется вектором, компонентами которого являются две величины – суммарный вес сопоставлений и общее количество удаленных фрагментов. Отметим, что, в отличие от метода Смита-Ватермана, в качестве веса выравнивания рассматривается вектор, а не линейная комбинация его компонентов. Поэтому наш подход не требует использования штрафа за удаление фрагмента (*Gap Opening Penalty*, GOP). Второй из используемых в алгоритме Смита-Ватермана параметров-штрафов, штраф за удаление символа (*Gap Elongation Penalty*, GEP) может быть учтен за счет модификации матрицы сопоставления символов. Таким образом, единственным параметром нашей постановки задачи является выбранная матрица сопоставления символов, подробнее см. [18]

Предлагаемая программа PARCA (от *PAReto CAndidate*) строит искомый упорядоченный набор выравниваний в два этапа. На первом этапе строится множество Парето-оптимальных выравниваний [20] относительно указанного векторного веса. На втором этапе в множестве Парето-оптимальных выравниваний отбирается упорядоченное множество «основных» выравниваний. В качестве множества выравниваний-кандидатов предъявляется набор из нескольких первых по списку основных выравниваний.

Множество Парето-оптимальных выравниваний строится методом динамического программирования [19]. Отбор основных выравниваний и их ранжирование описаны в следующем подразделе.

2.2. Отбор основных выравниваний и их ранжирование

Как было указано выше, после построения набора Парето-оптимальных выравниваний, в нем отбирается упорядоченное подмножество *основных* выравниваний. В качестве множества выравниваний-кандидатов предъявляется набор из N первых по списку основных выравниваний (N – параметр алгоритма).

Рассмотрим поведение зависимости суммарного веса за сопоставление символов m от числа удаленных фрагментов g. В большинстве случаев график зависимости $m(g)$ имеет выраженный излом (см. рис. 1). Наличие этого излома объясняется тем, что после того, как вос-

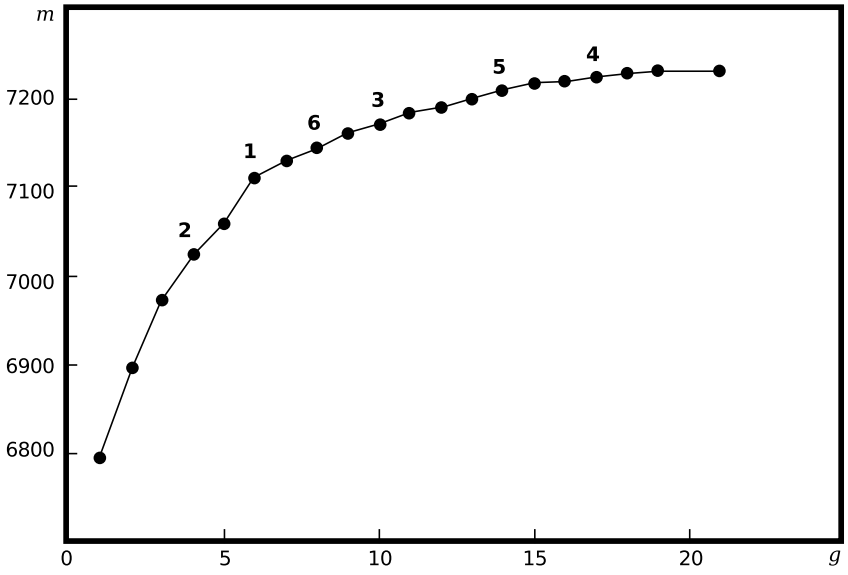

Рис. 1. Вид кривой зависимости суммарного веса за сопоставления m от числа удаленных фрагментов g. Числами от 1 до 6 отмечены выравнивания, являющиеся основными. Число указывает номер выравнивания в списке основных выравниваний.

становлено соответствие гомологичных фрагментов сравниваемых последовательностей, увеличение числа удаленных фрагментов не влияет существенным образом на суммарный вес сопоставлений. Следовательно, выравнивание, соответствующее точке излома можно считать искомым [19].

К сожалению, на практике точку излома не всегда легко выделить. Мы используем следующую эвристику.

Пусть $\mathrm{T(g)} = \langle \mathrm{m, g} \rangle$ – точка на графике $m(g)$. Положим (для $g > 1$)

$$\mathrm{dm(g)} = \mathrm{m(g)} - \mathrm{m(g-1)}.$$

Тангенсы углов наклона отрезков, примыкающих к точке T(g), будут равны

$$\mathrm{tg}_{left} = \frac{\mathrm{dm(g)}}{\sigma}, \qquad \mathrm{tg}_{right} = \frac{\mathrm{dm(g+1)}}{\sigma},$$

где σ – это масштабный коэффициент, эмпирически принятый равным 20. Значение этого параметра можно менять в достаточно широком диапазоне значений без изменения результатов работы программы.

Тангенс угла между отрезками, примыкающими к точке T(g), равен

$$\mathrm{tg}\left(\mathrm{T\,(g)}\right) = \frac{\mathrm{tg_{left}} - \mathrm{tg_{right}}}{1 + \mathrm{tg_{left}} \cdot \mathrm{tg_{right}}} = \frac{\mathrm{dm(g)} - \mathrm{dm(g+1)}}{\sigma^2 + \mathrm{dm(g)} \cdot \mathrm{dm(g+1)}}.$$

Парето-оптимальное выравнивание, определяемое вектором $\langle m, g \rangle$, будем называть *основным*, если тангенс угла между отрезками, примыкающими к соответствующей этому выравниванию точке T(g) является локальным максимумом последовательности $\left\{\mathrm{tg}\left(\mathrm{T\,(g)}\right)\right\}$. Значение $\mathrm{tg}\left(\mathrm{T\,(g)}\right)$, определяет порядок данного выравнивания в наборе основных выравниваний: первое выравнивание в списке имеет самое большое значение $\mathrm{tg}\left(\mathrm{T\,(g)}\right)$ и так далее.

Ограничивая размер N полученного упорядоченного набора основных выравниваний, мы сокращаем число потенциальных кандидатов на оптимальное выравнивание. Вычислительные эксперименты на модельных данных, описанные в [18], показали, что при $N = 6$ точность множества кандидатов, которое строит программа PARCA, как правило, превосходит точность выравнивания Смита-Ватермана. Целью настоящей работы является апробация метода на эталонных выравниваниях, полученных с помощью наложения пространственных структур реальных белков.

Сервер, реализующий описанный подход, находится по адресу [21].

3. КОМПЬЮТЕРНЫЕ ЭКСПЕРИМЕНТЫ. МЕТОДИКА И РЕЗУЛЬТАТЫ

3.1. Эталонные выравнивания

Для апробации предложенного метода мы использовали базу данных эталонных выравниваний PREFAB-P [22]; эта база получена в результате анализа и обработки базы эталонных выравниваний PREFAB [10]. В частности, были удалены выравнивания доменов, которые принадлежат к разным семействам с точки зрения классификации SCOP [23], выравнивания, в которых исключены внутренние фрагменты исходных последовательностей, дважды встречающиеся выравнивания. Подробнее см. [15].

3.2. Методика

Нами были рассмотрены все пары последовательностей из базы PREFAB-P.

1. Для каждой такой последовательности предварительно были построены выравнивания методом PARCA. При этом для каждой последовательности были апробированы весовые матрицы из семейства PAM [24]: PAM120, PAM240, PAM360, PAM480 и штрафы за удаление символа из диапазона от 0 до 3 с шагом 0.5. В результате этих экспериментов были определены оптимальные матрицы для каждого уровня сходства сравниваемых последовательностей (см. ниже).

   В качестве штрафа за удаление символа было взято значение 1.0, стандартно применяемое при использовании метода Смита-Ватермана с весовыми матрицами семейства PAM. Наши эксперименты показали, что при таком выборе штрафа результаты применения метода PARCA близки к оптимальным по достигаемой точности. Полные результаты проведенных экспериментов находятся по адресу [3].
2. На следующем этапе для каждой пары последовательностей рассматривался результат работы программы PARCA только для матрицы PAM, соответствующей уровню сходства этой последовательности. В построенном множестве Парето-оптимальных выравниваний было выделено упорядоченное подмножество основных выравниваний. Рассмотрены 10 наборов выравниваний-кандидатов – начальные фрагменты набора основных выравниваний размером от 1 до 10 элементов. Подсчитаны точности этих наборов, наборов всех основных и всех Парето-оптимальных выравниваний, а также точность выравнивания Смита-Ватермана. Полные результаты находятся в сопроводительных материалах [3].
3. Чтобы убедиться, что наши результаты в целом не зависят от выбора весовых матриц, к каждой паре последовательностей из базы PREFAB-P были применены алгоритм Смита-Ватермана и алгоритм PARCA с весовой матрицей BLOSUM62. Штраф за удаление символа в обоих случаях был взят равным 1, это значение было определено на основе предварительных экспериментов с модельными последовательностями так, чтобы оптимизировать среднюю точность выравниваний. Штраф за удаление фрагмента для алгоритма Смита-Ватермана был взят равным 14; это значение максимизирует среднюю точность выравниваний на множестве всех пар последовательностей из базы PREFAB-P при выбранном

штрафе за удаление символа. Отметим, что, например, что значения 10 и 0.5, рекомендованные для реализующей алгоритм Смита-Ватермана программы `needle` пакета EMBOSS [25] приводят к худшей (в среднем) точности выравниваний Смита-Ватермана.

Замечание. Как мы увидим ниже, точность выравниваний, полученных с помощью матрицы BLOSUM62 выше, чем точность выравниваний, полученных с помощью матриц семейства PAM. Однако, при сравнении модельных последовательностей, полученных с помощью имитации эволюционного процесса с заданным эволюционным расстоянием, выраженным в единицах PAM, естественно использовать матрицы семейства PAM. Мы провели выравнивание последовательностей с помощью весовых матриц семейства PAM, чтобы иметь возможность сравнивать результаты компьютерных экспериментов с реальными и модельными последовательностями.

3.3. Результаты

Мы отдельно рассматривали эталонные выравнивания с различным уровнем сходства. В качестве меры сходства использовалась величина $\%id$ – доля совпадений среди всех пар сопоставленных символов. Были выделены 4 уровня сходства, эти уровни соответствуют следующим диапазонам значений величины $\%id$ (границы диапазонов выбраны так, чтобы для выравнивания разных пар из этого диапазона оптимальной была бы одна и та же матрица весов сопоставлений семейства PAM):

- до 12% (PAM480);
- от 12% до 17% (PAM360);
- от 17% до 28% (PAM240);
- от 28% (PAM120).

При использовании матрицы BLOSUM62, мы для удобства использовали те же четыре диапазона.

Сводные результаты исследования приведены в таблицах 1–3. Полные результаты компьютерных экспериментов доступны по адресу [3]. Отметим, что столбцы таблицы 2 – это разности столбцов таблицы 1.

Таблица 3 дает представление о распределении разностей точностей наборов выравниваний программы PARCA и выравниваний Смита-Ватермана в зависимости от степени сходства сравниваемых последовательностей и используемой весовой матрицы.

Из таблиц 1 и 2 видно, что выравнивания, построенные с помощью весовой матрицы BLOSUM62, имеют более высокую точность, чем выравнивания, построенные при помощи весовых матриц семейства PAM. Это верно, как для алгоритма Смита-Ватермана, так и для программы PARCA. При этом преимущество в точности программы PARCA также выше для матрицы BLOSUM62. Так, для слабогомологичных последовательностей ($\%id < 17\%$) средняя точность набора из 6 первых основных выравниваний программы PARCA в полтора раза выше, чем точность выравнивания Смита-Ватермана. Существенно, что преимущество программы PARCA достигается именно на слабогомологичных последовательностях.

Интерес представляют также данные в столбце «Все Парето-оптимальные выравнивания». Так как (при данном значении штрафа за удаление символа и данной матрице весов сопоставлений) выравнивание, являющееся оптимальным по Смиту-Ватерману, является также Парето-оптимальным, эти данные определяют верхний предел точности, которого можно достичь, подбирая значение штрафа за удаление фрагмента в алгоритме Смита-Ватермана, адаптируясь к свойствам сравниваемых последовательностей.

%id	Кол-во	Смит-Ватерман		PARCA							
				3 первых основных		6 первых основных		Все основные		Все Парето-опт.	
		PAM	BS-62	PAM	BS-62	PAM	BS-62	PAM	BS-62	PAM	BS-62
7 . . . 12	36	16.9%	16.9%	19.8%	23.4%	22.0%	25.9%	22.4%	27.4%	25.0%	31.5%
12 . . . 17	88	27.7%	27.7%	33.0%	40.5%	35.2%	42.9%	35.5%	43.4%	38.2%	46.0%
17 . . . 28	142	58.2%	58.2%	62.7%	63.9%	63.7%	64.7%	64.5%	64.5%	66.4%	67.4%
28 . . . 100	315	90.2%	90.2%	90.1%	90.5%	90.3%	90.7%	90.4%	90.7%	91.4%	91.7%

Таблица 1. Средняя точность выравниваний Смита-Ватермана и различных наборов выравниваий-кандидатов, которые строит программа PARCA. Рассмотрены четыре вида наборов выравниваний-кандидатов: 1) 3 первых основных выравнивания в соответствии с упорядочиванием программы PARCA; 2) 6 первых основных выравниваний в соответствии с упорядочиванием программы PARCA; 3) все основные выравнивания (не более 20 выравниваний, в среднем - менее 10 выравниваний); 4) все Парето-оптимальные-выравнивания (не более 40 выравниваний). Данные приводятся отдельно по различным диапазонам степени сходства последовательностей (%id) и компьютерных экспериментов с использованием весовой матрицы BLOSUM62 (столбец BS-62) и весовых матриц семейства PAM. Точностью набора выравниваний-кандидатов считается максимальная точность, достигаемая для этого набора, см. пояснения в тексте.

%id	Кол-во	3 первых основных		6 первых основных		Все основные		Все Парето-опт.	
		PAM	BS-62	PAM	BS-62	PAM	BS-62	PAM	BS-62
7 . . . 12	36	2.9%	5.1%	5.1%	7.6%	5.6%	9.1%	8.1%	13.2%
12 . . . 17	88	5.3%	5.3%	7.5%	7.7%	7.8%	8.1%	10.6%	10.8%
17 . . . 28	142	4.4%	3.8%	5.5%	4.6%	6.3%	5.3%	8.1%	7.3%
28 . . . 100	315	-0.1%	-0.1%	0.1%	0.0%	0.2%	0.1%	1.3%	1.1%

Таблица 2. Разности между средними точностями различных наборов выравниваий-кандидатов, которые строит программа PARCA и точностью соответствующих выравниваний Смита-Ватермана. Рассмотрены четыре вида наборов выравниваний-кандидатов: 1) 3 первых основных выравнивания в соответствии с упорядочиванием программы PARCA; 2) 6 первых основных выравниваний в соответствии с упорядочиванием программы PARCA; 3) все основные выравнивания (не более 20 выравниваний, в среднем - менее 10 выравниваний); 4) все Парето-оптимальные-выравнивания (не более 40 выравниваний). Данные приводятся отдельно по различным диапазонам степени сходства последовательностей (%id) и компьютерных экспериментов с использованием весовой матрицы BLOSUM62 (столбец BS-62) и весовых матриц семейства PAM. Точностью набора выравниваний-кандидатов считается максимальная точность, достигаемая для этого набора, см. пояснения в тексте.

4. ЗАКЛЮЧЕНИЕ

Несмотря на то, что задача построения глобальных парных выравниваний слобогомологичных последовательностей, имеющих точность лучше, чем выравнивания Смита-Ватермана, была сформулирована около 20 лет назад, решение этой задачи так и не было предложено. Говоря аккуратнее, были предложены лишь решения, использующие дополнительные входные данные – сведения о вторичной структуре белков, о белках-гомологах и другие данные.

Нами представлена новая постановка задачи построения глобального выравнивания двух последовательностей, которая является обобщением, с одной стороны, классической задачи глобального парного выравнивания [1], а, с другой стороны, – задачи построения всех «почти оптимальных» выравниваний [16, 17]. А именно, мы строим упорядоченный и относительно небольшой (в самом худшем случае – до 40 выравниваний, как правило – 3, 6 или 10–20) набор различных выравниваний данных последовательностей. При этом мы стремимся к тому, чтобы (а) размер набора был поменьше и (б) набор содержал как можно более точное выравнивание. Таким образом, задача выделения наиболее точного выравнивания из представленного набора выравниваний-кандидатов рассматривается, как отдельная задача, которая может решаться, в том числе, с помощью специально поставленных экспериментов.

а)

%id	3 первых основных			6 первых основных			Все основные			Все Парето-опт		
	Хуже	Равно	Лучше	Хуже	Равно	Лучше	Хуже	Равно	Лучше	Хуже	Равно	Лучше
7 . . . 12	11.1%	52.8%	36.1%	0.0%	47.2%	52.8%	0.0%	44.4%	55.6%	0.0%	36.1%	63.9%
12 . . . 17	4.5%	46.6%	48.9%	4.5%	37.5%	58.0%	3.4%	37.5%	59.1%	0.0%	25.0%	75.0%
17 . . . 28	9.9%	42.3%	47.9%	7.0%	41.5%	51.4%	6.3%	39.4%	54.2%	0.0%	33.8%	66.2%
28 . . . 100	11.1%	75.6%	13.3%	10.2%	74.3%	15.6%	9.5%	74.3%	16.2%	0.0%	77.5%	22.5%

б)

%id	3 первых основных			6 первых основных			Все основные			Все Парето-опт		
	Хуже	Равно	Лучше	Хуже	Равно	Лучше	Хуже	Равно	Лучше	Хуже	Равно	Лучше
7 . . . 12	5.6%	38.9%	55.6%	2.8%	25.0%	72.2%	0.0%	25.0%	75.0%	0.0%	16.7%	83.3%
12 . . . 17	12.1%	38.7%	49.2%	8.1%	31.5%	60.5%	7.3%	30.6%	62.1%	0.0%	28.2%	71.8%
17 . . . 28	9.8%	42.1%	48.1%	6.8%	38.0%	55.3%	6.0%	36.8%	57.1%	0.0%	33.1%	66.9%
28 . . . 100	10.2%	77.5%	12.4%	9.5%	75.6%	14.9%	9.2%	75.2%	15.6%	0.0%	76.5%	23.5%

Таблица 3. Распределение разностей между точностями различных наборов выравниваий-кандидатов, которые строит программа PARCA и точностью соответствующих выравниваний Смита-Ватермана при использовании весовых матриц семкйства PAM (а) и весовой матрицы BLOSUM62 (б). Рассмотрены четыре вида наборов выравниваний-кандидатов: 1) 3 первых основных выравнивания в соответствии с упорядочиванием программы PARCA; 2) 6 первых основных выравниваний в соответствии с упорядочиванием программы PARCA; 3) все основные выравнивания (не более 20 выравниваний, в среднем – менее 10 выравниваний); 4) все Парето-оптимальные-выравнивания (не более 40 выравниваний). Данные приводятся отдельно по различным диапазонам степени сходства последовательностей (%id) и компьютерных экспериментов с использованием весовой матрицы BLOSUM62 и весовых матриц семейства PAM. Точностью набора выравниваний-кандидатов считается максимальная точность, достигаемая для этого набора, см. пояснения в тексте. В столбцах «Меньше» указано количество пар, в которых точность набора выравниваний-кандидатов, построенного программой PARCA, меньше точности выравнивания Смита-Ватермана; в столбцах «Равно» указано количество пар, в которых точность набора выравниваний-кандидатов, построенного программой PARCA, и точность выравнивания Смита-Ватермана совпадают, в столбцах «Больше» – количества пар, в которых точность набора выравниваний-кандидатов, построенного программой PARCA, больше точности выравнивания Смита-Ватермана.

Задачу построения набора выравниваний-кандидатов, таким образом, можно рассматривать, как фильтрацию пространства возможных глобальных выравниваний. Такой подход давно используется при поиске локальных сходств (см., например, [26]), однако при построении глобальных выравниваний он не применялся.

Для указанной задачи предложен алгоритм ее решения; этот алгоритм реаизован в виде программы PARCA. С использованием базы эталонных выравниваний PREFAB-P проведены компьютерные эксперименты по сравнению точности выравниваний, которые строятся методом PARCA и методом Смита-Ватермана. Показано, что для всех уровней сходства алгоритм PARCA строит множества выравниваний-кандидатов, содержащие не более 6 выравниваний, которые в большинстве случаев превосходят по точности выравнивания Смита-Ватермана. При этом для слабогомологичных последовательностей, как правило, даже среди 6-элементных наборов выравниваний-кандидатов программы PARCA есть выравнивание, существенно превосходящее по качеству выравнивание Смита-Ватермана. Таким образом, можно сказать, что программа PARCA удовлетворительно решает поставленную задачу.

С методической точки зрения достоинство программы PARCA в том, что она является первой программой глобального выравнивания, которая не использует штраф за удаление фрагмента. Используя векторную трехкомпонентную весовую функцию выравниваний $\langle m, g, d\rangle$, где m – суммарный вес сопоставлений, g – количество удаленных фрагментов, d – суммарная длина удаленных фрагментов, можно аналогичным способом избавиться и от использования штрафа за удаление символа. Однако, наши компьютерные эксперименты показывают (см. со-

проводительные материалы [3]), что зависимость результатов от выбора штрафа за удаление символа достаточно слабая. С другой стороны, результаты компьютерных экспериментов подсказывают возможные направления для развития программы: улучшение эвристики выбора основных выравниваний, улучшение ранжирования основных выравниваний. В качестве возможного применения результатов программы PARCA укажем выделение сопоставлений, общих для всех выравниваний-кандидатов из набора, скажем, шести первых основных выравниваний. Эти сопоставления входят в самое точное выравнивание набора и, с большей «вероятностью», чем «среднее сопоставление» входят и в биологически корректное выравнивание.

5. БЛАГОДАРНОСТИ

Авторы благодарят анонимного рецензента за быстрое, внимательное и содержательное рецензирование.

СПИСОК ЛИТЕРАТУРЫ

1. Waterman M. S. *Mathematical methods for DNA sequences.* Boca Raton, FL: CRC Press, 1989
2. Sunyaev S., Bogopolsky G., Oleynikova N., Vlasov P., Finkelstein A., Roytberg M. From analysis of protein structural alignments towards a novel approach to align protein sequences. *Proteins: Structure, Function, and Bioinformatics*. 2004, vol. 54, issue 3, pp. 569–582
3. `http://server2.lpm.org.ru/static/iitp-paper2011/`
4. Smith T., Waterman M. Identification of common molecular subsequences. *Journal of Molecular Biology*. 1981, vol. 147, pp. 195-197
5. Mount D. W. *Bioinformatics. Sequence and genome analysis*. Cold Spring Harbor Laboratory Press, 2001
6. Vogt G., Etzold T., Argos P. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *Journal of Molecular Biology*. 1995, vol. 249, pp. 816–831
7. Polyanovsky V., Roytberg M., Tumanyan V. Reconstruction of genuine pair-wise sequence alignment. *Computational Biology*. 2008, vol. 15, N. 4. pp. 379–391
8. Литвинов И. И., Лобанов М. Ю., Миронов А. А., Финкельштейн А. В., Ройтберг М. А. Информация о вторичной структуре белка улучшает качество выравнивания. *Молекулярная биология*. 2006, Т. 40, № 3. стр. 533–540
9. Wallqvist A., Fukunishi Y., Murphy L., Fadel A., Levy R. Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics*. 2000, vol. 16. pp. 988–1002
10. Edgar R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 2004, vol. 32, issue 5, pp. 1792–1797
11. Alexandrov N, Luethy R. Alignment algorithm for homology modeling and threading. *Protein Science*. 1998, vol. 7, issue 2, pp. 254–258
12. Altschul S. Generalized affine gap costs for protein sequence alignment. *Proteins: Structure, Function, and Bioinformatics*. 1998, vol. 32, issue 1., pp.88–96
13. Huang X., Chao K.-M. A generalized global alignment algorithm. *Bioinformatics*. 2003, vol. 19, no. 2, pp. 228–233
14. Huang X., Brutlag D. Dynamic use of multiple parameter sets in sequence alignment. *Nucleic Acids Research*. 2007, vol. 35, no. 2, pp. 678–686
15. Поверенная И. В., Яковлев В. В., Астахова Т. В., Лобанов М. Н., Ройтберг М. А. Верификация базы эталонных выравниваний PREFAB. *Биофизика*. В печати

16. Byers T., Waterman M. Determining all optimal and nearoptimal solutions when solving shortest path problems by dynamic programming. *Oper Res.* 1984, vol. 32, pp. 1381–1384

17. Waterman M. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *PNAS.* 1983, vol. 80, pp. 3123–3124

18. Яковлев В. В., Ройтберг М. А. Увеличение точности глобального выравнивания аминокислотных последовательностей с помощью построения набора выравниваний-кандидатов. *Биофизика*, 2010, Т. 55, № 6, стр. 965–975

19. Ройтберг М. А., Семионенков М. Н., Таболина О. Ю. Парето-оптимальные выравнивания биологических последовательностей. *Биофизика*, 1998, Т. 44, № 4, стр. 581-594

20. Pareto V. *Manual of political economy.* New York: A. M. Kelley, 1972

21. `http://server2.lpm.org.ru/bio/online/pareto/`

22. `http://lpm.org.ru/~mroytberg/PREFAB/PREFAB-P.ZIP`

23. `http://scop.mrc-lmb.cam.ac.uk/scop/`

24. Altschul S. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology.* 1991, vol. 219, N. 3. pp. 555–565

25. `http://helixweb.nih.gov/emboss/html/needle.html`

26. Ilie L., Ilie S. Multiple spaced seeds for homology search. *Bioinformatics.* 2007, vol. 23, issue 22, pp. 2969–2977