На правах рукописи

Яковлев Виктор Вадимович

Методы выравнивания биологических последовательностей, не использующие штрафы за делеции

03.01.09 Математическая биология, биоинформатика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2012

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте математических проблем биологии РАН

Научный руководитель: доктор физико-математических наук
Михаил Абрамович Ройтберг

Официальные оппоненты: Андрей Александрович Миронов
доктор биологических наук,
кандидат физико-математических наук,
профессор,
Факультет биоинженерии и биоинформатики
федерального государственного бюджетного
образовательного учреждения высшего
профессионального образования
«Московский государственный университет
им. М. В. Ломоносова»,
профессор

Александр Владимирович Фаворов
кандидат физико-математических наук,
Государственный научный центр
Российской Федерации
ФГУП Государственный
научно-исследовательский институт
генетики и селекции промышленных
микроорганизмов,
научный сотрудник

Ведущая организация: Федеральное государственное бюджетное
учреждение науки
Институт теоретической и
экспериментальной биофизики РАН

Защита состоится «_____» _____2012 г. в _____ч. на заседании диссертационного совета Д 002.077.04 на базе Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича РАН по адресу 127994, Москва, ГСП-4, Большой Каретный переулок, д.19, стр.1.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича РАН

Автореферат разослан «_____»_____2012 г.

Ученый секретарь диссертационного совета Д 002.077.04
доктор биологических наук, профессор Г.И. Рожкова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Выравнивание аминокислотных и нуклеотидных последовательностей является классическим для биоинформатики методом сравнения последовательностей. Выявляемые при этом сходства часто являются следствием функциональных, структурных или эволюционных взаимосвязей между последовательностями. В 70-е годы XX века, когда сравнивались последовательности относительно небольшой длины, выравнивание производилось вручную, затем были предложены алгоритмы построения выравнивания. Актуальность задачи выравнивания обуславливается тем, что для многих белков известны их аминокислотные последовательности, но лишь для малой части из них известны пространственные структуры.

Для приложений важно, насколько алгоритмически полученные выравнивания отражают реальную эволюционную связь между сравниваемыми последовательностями. Количественной мерой этого служит точность выравнивания – доля таких сопоставлений эталонного выравнивания, которые присутствуют в алгоритмическом выравнивании. В качестве эталонных выравниваний аминокислотных последовательностей, например, используются выравнивания, основанные на наложении пространственных структур белков, соответствующих этим последовательностям.

Наиболее точным из известных в настоящее время алгоритмов построения глобального парного выравнивания аминокислотных последовательностей является алгоритм Смита-Ватермана, основанный на построении оптимального выравнивания, то есть выравнивания, для которого достигается максимальное значение весовой функции. Тем не менее, для слабогомологичных последовательностей точность выравниваний Смита-Ватермана невысока.

Существующие на данный момент алгоритмы попарного выравнивания, в частности, алгоритм Смита-Ватермана, требуют задания ряда параметров, выбор значений которых не имеет под собой надежного обоснования. Примером такого параметра является штраф за делецию (удаление) фрагмента; ошибка в выборе его значения приводит к существенному ухудшению точности алгоритмически полученного выравнивания. Одним из путей повышения точности алгоритмически построенных выравниваний является разработка метода выравнивания, который не использует штрафы за делеции, что определяет актуальность темы исследования.

Цели и задачи исследования. Цель исследования состоит в разработке метода глобального выравнивания биологических последовательностей, который позволяет получать выравнивания, более точные, чем выравнивания Смита-Ватермана. Исходя из поставленных целей, были сформулированы и решены следующие задачи:

1. Предложить новую формализацию задачи выравнивания, в которой не используются штрафы за делеции фрагментов.

2. Разработать эффективный алгоритм решения этой задачи.

3. Создать программную реализацию разработанного алгоритма.

4. Провести сравнительное исследование качества выравниваний, полученных с помощью разработанного алгоритма и выравниваний, полученных методом Смита-Ватермана.

Методы исследования. В работе использованы методы теории алгоритмов, теории вероятностей, математической статистики, биоинформатики и вычислительной молекулярной биологии, проведение вычислительных экспериментов с использованием существующих и оригинальных программ.

Достоверность результатов. Достоверность результатов обеспечивается адекватностью используемых методов теории алгоритмов, теории вероятностей, математической статистики и биоинформатики, верификацией математических моделей, а также сравнением аналитических результатов с результатами математического моделирования.

Научная новизна. Научная новизна работы состоит в сформулированной новой постановке задачи построения глобального выравнивания (задача построения упорядоченного множества выравниваний-кандидатов), а также в разработанных оригинальных алгоритмах. Предложен алгоритм построения упорядоченного набора выравниваний-кандидатов с временем работы $O(R \cdot L^2)$, где R – ограничение сверху на количество удаленных фрагментов в рассматриваемых выравниваниях, а L – это средняя длина выравниваемых последовательностей.

Личный вклад. Представленные в диссертационной работе результаты получены лично соискателем.

По материалам диссертационной работы опубликованы 4 научные работы. В работе [1] соискателю принадлежит алгоритм построения набора основных выравниваний, а также сравнительное исследование точности глобальных выравниваний Смита-Ватермана и наборов основных выравниваний, включая разработку методики исследования и подготовку исходных данных. В работе [2] соискателю принадлежит создание программы PARCA, разработка методики проведения компьютерных экспериментов, проведение самих экспериментов и их анализ. В работах [3] и [4] автору принадлежит разработка методики верификации базы данных эталонных выравниваний PREFAB, участие в проведении компьютерных экспериментов (совместно с И. Поверенной) и создание окончательной версии базы данных PREFAB-P.

Теоретическая и практическая значимость. Теоретическая значимость исследования заключается в постановке нового варианта задачи выравнивания – задачи построения упорядоченного набора выравниваний-кандидатов, а также в разработанном алгоритме решения этой задачи.

Практическая значимость состоит в разработанных программах[1,2], а

[1] http://server2.lpm.org.ru/bio/online/pareto/

[2] http://server2.lpm.org.ru/static/parca/

также в подготовленной базе эталонных выравниваний[3], которая может быть использована при оценке точности других алгоритмов выравнивания. Реализация предложенного метода используется при создании опытного образца программного комплекса «СИМВОЛ», разрабатываемого в рамках государственного контракта № 07.514.11.4004. Результаты работы использовались при выполнении работ по темам «Сравнительный анализ структур белков и нуклеиновых кислот» (номер государственной регистрации 01.2.00409635), «Математические методы анализа белков и нуклеиновых кислот: связь между последовательностями, структурой и функцией» (номер государственной регистрации 01.2.00952309), а также при выполнении проекта РФФИ 09-04-01053-а «Достоверность и полнота результатов при компьютерном анализе последовательностей биополимеров». Результаты работы можно рекомендовать к использованию для получения выравниваний слабогомологичных последовательностей – такие выравнивания необходимы при решении многих задач биоинформатики.

Апробация результатов. Результаты работы были представлены на международном рабочем совещании RECESS (Мюнхен, декабрь, 2010 г.), международной конференции MCCMB'2011 (Москва, июль, 2011 г.), на семинарах в Институте математических проблем биологии РАН, Московском Государственном Университете, Институте белка РАН, Институте теоретической и экспериментальной биофизики РАН.

Публикации. Основные материалы диссертации изложены в четырех работах общим объемом 47 страниц, из них – три статьи, написанных в соавторстве (общий объем 46 страниц), в том числе две статьи, опубликованных в журналах из списка ВАК (общий объем 32 страницы). Кроме того, опубликованы тезисы сообщения на международной конференции (объем – 1 страница).

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения и списка литературы (85 наименований). Полный объем диссертации составляет 98 страниц, количество рисунков – 16, количество таблиц – 19.

Основные положения, выносимые на защиту:

1. Предложенная новая постановка задачи глобального выравнивания биологических последовательностей – задача построения заданного количества ранжированных выравниваний-кандидатов – позволяет получать более точные выравнивания слабогомологичных белков, чем это делают существующие методы.

2. Предложенный алгоритм решения задачи построения упорядоченного набора глобальных выравниваний двух заданных символьных последовательностей имеет время работы $O(R \cdot L^2)$, где R – ограничение сверху на количество удаленных фрагментов в рассматриваемых вырав-

[3] http://server2.lpm.org.ru/static/prefab-p/

ниваниях, L – ограничение сверху на длину сравниваемых последовательностей.

3. При сравнении слабогомологичных последовательностей, точность лучшего из предлагаемых алгоритмом выравниваний-кандидатов в среднем на 5 процентных пунктов выше точности выравниваний Смита-Ватермана; этот результат достигается при использовании не более шести выравниваний-кандидатов.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы диссертации, определены цели и задачи исследования, обоснована его научная новизна и практическая ценность. Приведены основные используемые определения, в частности, ключевые для диссертации определения выравнивания и его веса.

Определение 1. *Сопоставлением позиций* в последовательностях S_1 и S_2 называется пара целых чисел $\tau=<\lambda_1, \lambda_2>$, таких что $1 \le \lambda_1 \le |S_1|$ и $1 \le \lambda_2 \le |S_2|$.

Определение 2. *Выравниванием* последовательностей S_1 и S_2 называется тройка $<S_1, S_2, A>$, где $A=\{<i_1, j_1>,...,<i_n, j_n>\}$ – последовательность сопоставлений, таких что

$$1 \le i_1 \le ... i_n \le |S_1|,\ 1 \le j_1 \le .. j_n \le |S_2|.$$

Это определение соответствует тому, что i_k-й символ последовательности S_1 сопоставлен j_k -му символу в S_2 (k=1, ..., n), а остальные символы в последовательностях S_1 и S_2 удалены.

Алгоритм Смита-Ватермана использует следующее определение веса $W_{SW}(A)$ выравнивания A:

$$W_{SW}(A)=M(A) - GEP \cdot d(A) - GOP \cdot g(A).$$

Здесь $M(A)$ – суммарный вес сопоставлений относительно заданной матрицы весов сопоставлений символов; $d(A)$ – число удаленных символов; $g(A)$ – число удаленных фрагментов; *GEP* (*Gap Elongation Penalty*) и *GOP* (*Gap Opening Penalty*) – параметры алгоритма (соответственно, штраф за удаление символа и штраф за удаление фрагмента).

Первая глава посвящена обзору литературы. В обзоре продемонстрирована актуальность выбранной темы исследования и показана связь проанализированных работ с предметом исследования диссертации.

Во второй главе описаны тестовые данные, с помощью которых проводились компьютерные экспериментов. Тестовые данные состоят из двух частей:

1. Набор реально существующих аминокислотных последовательностей, для которых известны эталонные выравнивания (*эмпирические данные*). Эталонные выравнивания для этих последовательностей получены с помощью программы оптимального совмещения пространственных структур.

2. Набор псевдослучайных аминокислотных последовательностей,

имеющих статистическое сходство с реально существующими последовательностями (*модельные данные*). Эталонные выравнивания для этих последовательностей определяются их построением.

Задача построения корпуса модельных последовательностей возникает из-за ограниченности числа реальных последовательностей, на которых можно проводить испытания. В частности, корпус эталонных выравниваний, полученных из базы данных PREFAB содержит всего 581 пару последовательностей, а корпус модельных последовательностей – 12 000 последовательностей. Результаты, полученные на обоих наборах данных, в целом, совпадают.

Раздел 2.1 посвящен созданной базе эталонных выравниваний PREFAB-P, которая является уточненным вариантом разработанной Р. Эдгаром свободно распространяемой базы данных PREFAB.

Современные базы данных эталонных выравниваний реальных аминокислотных последовательностей, как правило, построены на основе структурных выравниваний белков, то есть выравниваний, основанных на совмещении пространственных структур. Различные базы данных отличаются выбором семейств белков, использованными алгоритмами наложения структур и методикой уточнения алгоритмических структурных выравниваний, которое обычно проводится экспертами. При создании корпуса эталонных выравниваний реальных последовательностей за основу была принята свободно распространяемая база данных PREFAB, поскольку данная база использовалась для тестирования таких известных продуктов, как MUSCLE и CLUSTALW.

Как показал проведенный анализ, база данных PREFAB не свободна от недостатков. В ней есть опечатки и отсутствуют данные о степени родства выравниваемых белков. Анализ базы данных состоял из двух этапов: верификации последовательностей и верификации выравниваний. В первом случае имеется в виду сравнение между собой последовательности из PREFAB-выравнивания и соответствующей ей последовательности, полученной из банка PDB. Под верификацией эталонных выравниваний мы подразумеваем сравнение типов структур сравниваемых доменов по классификации SCOP: выравнивания последовательностей, не принадлежащих к одному семейству в смысле SCOP исключались из рассмотрения.

В результате проведенных уточнений, на основе оригинальной базы данных PREFAB, состоящей из 1682 выравниваний, был создан корпус, состоящий из 581 выравнивания. В дополнение к сведениям, содержащимся в базе PREFAB, база данных PREFAB-P содержит сведения о типе пространственной структуры белков по классификации SCOP, наличии не выровненных фрагментов на концах и другие данные. База эталонных выравниваний PREFAB-P доступна[4] в виде реляционной базы данных.

[4] http://server2.lpm.org.ru/static/prefab-p/

Раздел 2.2 посвящен подготовке модельных данных, эти данные сгруппированы по сериям. Каждая серия состоит из 200 пар последовательностей и характеризуется следующими параметрами:

– средняя длина последовательностей *LSeq*;

– процент совпадений в эталонном выравнивании сравниваемых последовательностей *%id*;

– суммарное количество вставленных в последовательности фрагментов *NGap*.

При построении корпуса модельных тестовых данных для этих характеристик были выбраны следующие значения:

– *%id* = 10, 15, 20, 30, 50 и 90% (6 возможных значений);

– *LSeq* = 300, 600 и 1000 (3 возможных значения);

– *NGap* = 5, 10, 15 и 20 (4 возможных значения) при *LSeq*=300, и 10, 15, 20 (3 возможных значения) при *LSeq*=600 и *LSeq*=1000;

Эти значения варьировались независимо друг от друга. Таким образом, всего было построено 6·4+6·3+6·3 = 60 серий, что составляет 12 000 выравниваний.

Модельные данные готовились в два этапа.

1 этап – предварительный. На этом этапе производится анализ набора выравниваний реальных белков. В качестве такового использовалась база данных PREFAB; она была разбита на секции в соответствии с уровнем сходства сравниваемых последовательностей от 5 до 95% с шагом в 10%: от 5% до 15%, от 15% до 25% и т. д. В результате анализа для каждой секции была вычислена «вероятность появления удаленного фрагмента данной длины». Говоря более формально, вычислялась следующая величина:

$$p(l)=N_l / N,$$

где l – длина удаленного фрагмента, N_l – суммарное число удаленных фрагментов длины l в выравниваниях данной секции базы данных, N – суммарное число удаленных фрагментов во всех выравниваниях, входящих в секцию. При этом удаленные фрагменты, находящиеся на концах последовательностей, не учитывались.

2 этап. Построение нужного количества тестовых аминокислотных последовательностей с заданными параметрами:

– средняя длина последовательности *Lseq*;

– процент совпадения в сравниваемых последовательностях *%id*;

– количество вставленных фрагментов *Ngap*;

– матрица вероятностей замен *MutationProbMatrix*;

– статистическая плотность распределения длин вставок $p(l)$.

Построение псевдослучайных последовательностей и их эталонных выравниваний заключается в построении *исходной* последовательности, и преобразовании ее к *мутантной*, после чего – внесении в последовательности *вставок*.

Исходная последовательность генерируется в соответствии с Бернуллиевской моделью, параметры которой взяты из работы[5]. Вероятность наличия мутации в каждой позиции определялась в соответствии со значением параметра *%id*. В качестве матрицы вероятностей замен *MutationProbMatrix* использовалась одна из матриц семейства PAM[6]. Основываясь на результатах работы[7], использовались следующие матрицы:

– PAM-480 для $\%id \leq 10\%$;

– PAM-360 для $10\% < \%id \leq 15\%$;

– PAM-240 для $15\% < \%id \leq 20\%$;

– PAM-120 для $\%id > 20\%$.

Процесс внесения вставок был организован следующим образом:

1) генерируется заданное число *Ngap* вставок, длины l_i которых определяются случайным образом в соответствии с вероятностью $p(l_i)$, а их символьные последовательности определяются таким же образом, как и при построении исходной последовательности;

2) полученные фрагменты вставляются в каждую из последовательностей выравнивания равномерно и равновероятно.

Здесь следует отметить, что процесс порождения нужного количества вставленных фрагментов выполняется до построения исходной последовательности. Это необходимо для определения длины исходной последовательности *Lbase*, равной:

$$Lbase = Lseq - \sum l_i.$$

Описанная процедура построения пары псевдослучайных последовательностей однозначно определяет их эталонное выравнивание, основываясь на истории внесенных в исходную последовательность изменений.

Третья глава посвящена разработке новых алгоритмов выравнивания и реализации этих алгоритмов. В разделе 3.1. описано исследование зависимости точности выравниваний Смита-Ватермана от значений параметров *GOP* и *GEP*. Далее описана новая постановка задачи выравнивания (раздел 3.2), сформулированная в результате этого анализа, и разработанный алгоритм ее решения (раздел 3.3). Программная реализация разработанных алгоритмов в виде программного комплекса PARCA описана в разделе 3.4. Раздел 3.5 посвящен построению выравниваний Смита-Ватермана с помощью программы PARCA, этот раздел носит технический характер, поэтому его содержание не разбирается в автореферате подробно.

---

[5] Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) "A model of evolutionary change in proteins." In "Atlas of Protein Sequence and Structure, vol. 5, suppl. 3." M.O. Dayhoff (ed.), pp. 345-352, Natl. Biomed. Res. Found., Washington, DC.

[6] http://www.biorecipes.com/Dayhoff/code.html

[7] Поляновский В. О., Ройтберг М. А., Туманян В. Г. Новый подход к оценке достоверности выявления вставок-делеций в парном выравнивании. // Биофизика. 2008. Т. 53, № 4. С. 533–537

На предварительном этапе работы (см. раздел 3.1 диссертации) с помощью компьютерных экспериментов была исследована зависимость точности выравниваний Смита-Ватермана от параметров *GEP* (штраф за удаление символа) и *GOP* (штраф за удаление фрагмента) для различных уровней эволюционного расстояния между сравниваемыми последовательностями. Эксперименты были проведены на всех сериях модельных данных по описанной ниже схеме.

1. Строятся выравнивания Смита-Ватермана для каждой пары последовательностей из данной серии и каждой пары значений параметров *GEP* и *GOP;* значения параметров меняются независимо друг от друга в диапазонах от 0 до 3 с шагом 0.5 (параметр *GEP*) и от 0 до 100 с шагом 1 (параметр *GOP*).

2. Определяется среднее значение точности выравниваний для каждой серии при каждой паре значений *GEP* и *GOP.*

3. Для каждой серии модельных данных определяется пара значений параметров *GEP* и *GOP,* обеспечивающая для данной серии максимальную среднюю точность.

4. В качестве весовых матриц замен использовались матрицы семейства *PAM.* Выбор конкретной матрицы определялся по уровню сходства между последовательностями в соответствии с работой[12] (см. выше).

Эксперименты показали, что значение *GEP* можно брать равным 1.0 в широком диапазоне эволюционных расстояний (при проценте сходства *%id < 50%)*, см. рис. 1 и табл. 1. Для высокогомологичных последовательностей несколько большая точность достигается при меньших значениях параметра *GEP,* однако разница в точностях невелика. Значения в таблице 1 были получены со следующими матрицами весов замен: PAM120 (для последовательностей с %id=30 и выше), PAM240 (при %id=20), PAM360 (при %id=15), PAM480 (при %id=10).

Таблица 1

Значения параметров GOP и GEP, обеспечивающие максимальную среднюю точность выравниваний Смита-Ватермана в различных сериях экспериментов.

%ID	5 удаленных фрагментов		10 удаленных фрагментов		15 удаленных фрагментов		20 удаленных фрагментов	
	GOP	GEP	GOP	GEP	GOP	GEP	GOP	GEP
10	37	1.0	29	1.0	21	1.0	16	1.0
15	27	1.0	20	1.0	18	1.0	16	1.0
20	16	1.0	16	1.0	14	1.0	11	1.0
30	17	1.0	13	1.0	11	1.0	10	1.0
50	12	1.0	11	1.0	9	1.0	10	0.5
70	12	0.5	9	0.5	9	0.5	10	0.5
90	7	0.0	6	0.0	5	0.0	6	0.0

Типичные кривые точности при различных значениях *GEP* показаны на рисунке 1, где отдельные графики (а) – (б) соответствуют различным количествам удаленных фрагментов $g(A)$. Различным линиям на графиках соответствуют различные значения штрафа за удаление символа *GEP* (0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0). Полужирная линия на каждом графике соответствует значению GEP=1.0. Кривая на графиках, которая имеет ярко выраженную низкую точность, соответствует значению GEP=0.0.

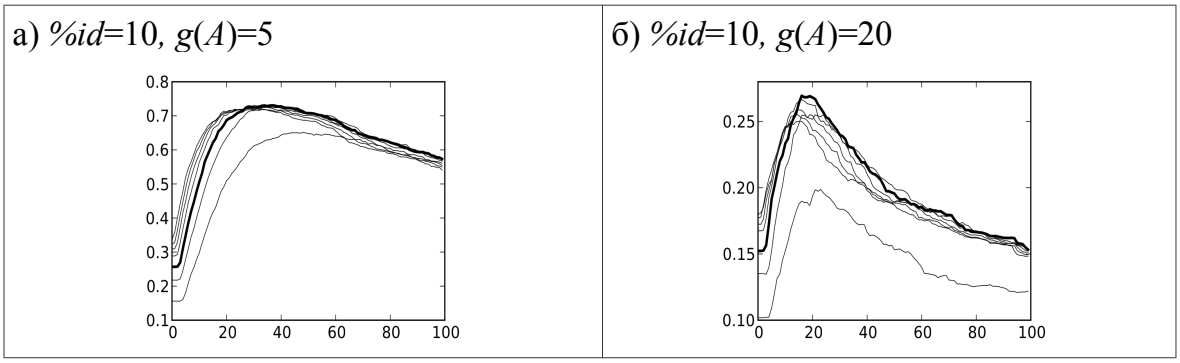

Рис. 1. Типичные кривые зависимости точности выравнивания Смита-Ватермана от штрафа за удаление фрагмента GOP для модельных данных.

Результаты проведенных экспериментов позволили упростить задачу, считая штраф за удаление символов *GEP* фиксированным. Как известно, за счет модификации весовой матрицы, задача построения глобального оптимального выравнивания Смита-Ватермана может быть сведена к решению этой же задачи при $GEP=0$, при этом значение параметра *GOP* не меняется[8]. Поэтому каждое выравнивание можно характеризовать двумя величинами:

1) суммарным весом замен (относительно модифицированной матрицы замен) и

2) количеством удаленных фрагментов.

Определения 3–5 и утверждение 1 получены на основе результатов из работы[9].

Определение 3. Пусть заданы последовательности S_1 и S_2, матрица весов замен W и значение параметра *GEP*. Векторным весом $V(A)$ выравнивания A называется величина

$$V(A)=<M(A), -g(A)>, \qquad (1)$$

где $k=2$, $M(A)$ – суммарный вес сопоставлений в соответствии с матрицей весов W, модифицированной относительно значения *GEP*, а $g(A)$ – количество удаленных фрагментов.

[8] Waterman M.S.(ed.) Mathematical methods for DNA sequences. // CRC Press, Boca Raton, FL. 1989. 293 p.

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

Определение 4. Выравнивание A_1 *доминирует* над выравниванием A_2, если выполнено

$$M(A_1) \geq M(A_2),\ g(A_1) \leq g(A_2),$$

причем хотя бы одно из неравенств является строгим.

Выравнивание A называется *Парето-оптимальным,* если никакое выравнивание не доминирует над ним. Вес $<M, g>$ называется *Парето-оптимальным,* если существует такое выравнивание A, что $V(A)=<M, g>$.

Определение 5. Набор выравниваний называется *Парето-оптимальным,* если в этом наборе для каждого Парето-оптмального веса $<M, g>$ существует такое выравнивание A, что $V(A)=<M, g>$.

Утверждение 1. Пусть выравнивание A является оптимальным выравниванием Смита-Ватермана относительно весовой матрицы W, штрафа за удаление символа *GEP* и некоторого значения штрафа за удаление фрагмента. Тогда выравнивание A является Парето-оптимальным.

В цитированной выше работе[9], описан алгоритм, который строит набор глобальных Парето-оптимальных выравниваний двух заданных последовательностей относительно весовой функции (1). Ввиду утверждения 1, этот набор выравниваний можно рассматривать, как набор выравниваний-кандидатов. Однако на практике работать с таким набором неудобно – он может содержать десятки выравниваний и при этом полученные выравнивания «равноправны».

Таким образом, задача глобального выравнивания биологических последовательностей может быть сформулирована следующим образом.

Дано:

– две биологические последовательности;

– матрица весов замен (говоря неформально, одна из стандартных весовых матриц, модифицированная в соответствии с заданным значением параметра *GEP*).

Требуется: построить упорядоченный набор из заданного количества глобальных выравниваний-кандидатов так, чтобы среди построенных выравниваний было выравнивание как можно более высокой точности.

Определение 6. Точностью набора выравниваний-кандидатов считается точность наилучшего из выравниваний этого набора.

Приведенная постановка задачи является обобщением задачи построения одного оптимального выравнивания, которая решается, например, алгоритмом Смита-Ватермана. С другой стороны, она является обобщением цитированной выше работы[9], а также работ М. Ватермана, Д. Гасфилда и других авторов, в которых рассматривается задача построения множества выравниваний. Отличие предложенной задачи от задач, рассмотренных в цитированных работах, состоит, во-первых, в том, что размер множества выравниваний-кандидатов невелик и заранее определяется пользователем, и, во-вторых, что набор выравниваний-кандидатов упорядочен. Таким образом, пользователь может перебрать все

предложенные выравнивания и попробовать использовать каждое из них при решении своей задачи. Это оправдывает определение точности набора выравниваний, данное выше.

В диссертации предложен алгоритм решения сформулированной выше задачи; этот алгоритм состоит из двух этапов. На первом этапе строится набор Парето-оптимальных выравниваний. Алгоритм выполнения этого этапа, в целом, следует цитированной выше работе, некоторые уточнения описаны в диссертации. Время выполнения этого этапа – $O(R \cdot L^2)$ операций, используемая память – $O(R \cdot L^2)$ байт. Здесь R – априорное ограничение на максимальное количество удаленных фрагментов в просматриваемых выравниваниях.

На втором этапе алгоритма (см. раздел 3.3 диссертации) производится выделение в полученном множестве Парето-оптимальных выравниваний подмножества т. н. *основных* выравниваний и упорядочивание этого подмножества. В качестве искомого набора выравниваний-кандидатов, таким образом, предъявляется нужное число первых (в смысле проведенного ранжирования) основных выравниваний. Работа алгоритма на втором этапе основана на анализе графика зависимости суммарного веса за сопоставление $M(g)$ от числа удаленных фрагментов g.

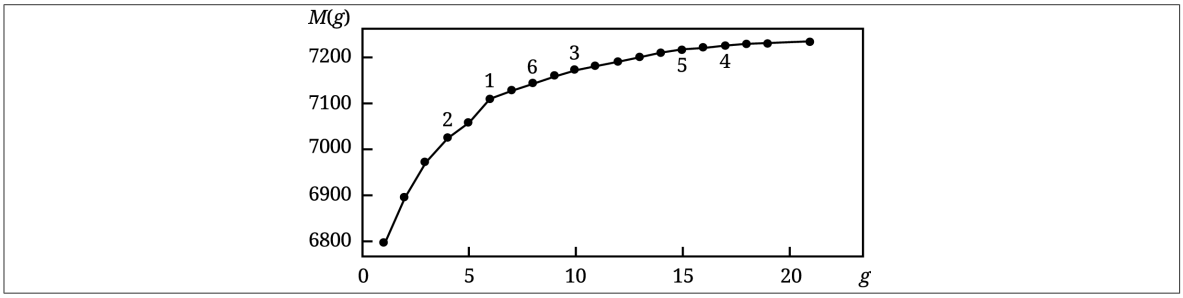

Рис. 2. Пример графика зависимости суммарного веса сопоставлений $M(g)$ от числа удаленных фрагментов g.

Пусть $A(g)$ – Парето-оптимальное выравнивание, содержащее g удаленных фрагментов, $M(g)$ – суммарный вес сопоставлений в выравнивании $A(g)$; Положим (для $g>1$):

$$dM(g) = M(g) - M(g\text{-}1).$$

Тангенсы углов наклона отрезков, примыкающих к точке $T(g)$ будут равны соответственно $tg_{left} = dM(g)/a$ и $tg_{right} = dM(g+1)/a$, здесь a – масштабный коэффициент. В настоящей работе он принят равным 20 (это число подобрано эмпирически, результаты не зависят от изменения коэффициента в очень широких пределах). Тангенс угла между отрезками, примыкающими к точке $T(g)$, равен

$$tg(g) = (tg_{left} - tg_{right})/(1+tg_{left} \cdot tg_{right}) =$$
$$= (dM(g) - dM(g+1)) / (a^2 + dM(g) \cdot dM(g+1))$$

Парето-оптимальное выравнивание $A(g)$ будем называть *основным,* если g – точка локального максимума в последовательности $\{ tg(g) \}$.

Основные выравнивания будем считать упорядоченными по убыванию величины tg(g), через $E(n)$ будем обозначать набор из n первых основных выравниваний. На рис. 2 числами обозначены точки, соответствующие выравниваниям-кандидатам, вошедшим в набор $E(6)$, каждое число соответствует порядковому номеру выравнивания в наборе.

Время работы алгоритма на втором этапе и используемая память зависит только от размера полученного множества Парето-оптимальных выравниваний и существенно меньше аналогичных характеристик для первого этапа. Таким образом, верно следующее

Утверждение 2. Предложен алгоритм построения набора выравниваний-кандидатов с временем работы $O(R \cdot L^2)$ и использующий память $O(R \cdot L^2)$.

Раздел 3.4 посвящен реализации предложенных алгоритмов – программному комплексу PARCA (от *PAReto CAndidate*). Программный комплекс реализован частично на языке C++ (построение Парето-оптимального множества выравниваний), и частично – на языке Python (все остальные алгоритмы). Такое решение, с одной стороны, эффективно использует вычислительные ресурсы, а с другой – позволяет добиться гибкости как при проведении вычислительных экспериментов, так и при интеграции с другими программными пакетами.

Исходные тексты программного комплекса распространяются под открытой лицензией MIT и являются общедоступными[10]. Комплекс может использоваться как в UNIX-подобных системах, так и на компьютере под управлением операционных систем Windows XP/7. Помимо исходных текстов программ, поставка комплекса включает в себя готовые к работе пакеты для openSUSE и Fedora Linux, а также программу для использования в среде Windows. Возможно использование комплекса с другими UNIX-подобными системами, например MacOS X или FreeBSD, в поставке исходных текстов прилагается инструкция по сборке программы для этих систем.

Помимо использования программы на локальном компьютере, комплекс PARCA может использоваться с помощью WEB-интерфейса, доступного на сервере лаборатории прикладной математики ИМПБ РАН[11]. Это WEB-приложение поддерживает как построение списка выравниваний кандидатов, так и сервисные функции – просмотр построенных выравниваний и их хранение.

Архитектурно комплекс PARCA состоит из следующих компонент.

1. Библиотека, оформленная в виде Python-модуля, которая может быть использована в качестве компоненты более крупного программного комплекса. Эта библиотека является платформозависимой и поставляется в вариантах для Windows и Linux.

[10] http://server2.lpm.org.ru/static/parca/

[11] http://server2.lpm.org.ru/bio/online/pareto/

2. Консольная программа `parca`, которая является программной оболочкой над библиотекой комплекса, предназначенной для использования из командной строки UNIX-подобной системы или Windows.

3. Веб-приложение, работающее на Linux-сервере, и предоставляющее возможность удаленной работы с программой, используя браузер. Данное веб-приложение реализовано с использованием инструментальных средств Django.

В четвертой главе описаны методика проведения численных экспериментов и их результаты. Все эксперименты производились на сервере лаборатории прикладной математики ИМПБ РАН под управлением openSUSE Linux 64-bit, с 4-х ядерным процессором Intel Core i5 и 8 гигабайтами оперативной памяти. Поскольку предложенный алгоритм не был спроектирован с учетом возможности параллельного вычисления, эффективное использование всех ядер процессора обеспечивалось одновременным запуском нескольких различных серий экспериментов.

Эксперименты состояли из 60 серий, две из которых были посвящены работе с реальными последовательностями (581 пара последовательностей из базы PREFAB-P) с использованием одной из матриц семейства PAM, и с использованием матрицы BLOSUM-62. Остальные серии экспериментов проводились с модельными последовательностями. В каждой такой серии анализировались 200 пар последовательностей, серии отличались друг от друга процентом начального сходства, числом удаленных фрагментов и длиной сравниваемых последовательностей.

Методика проведения вычислительных экспериментов с каждой серии (в том числе – в экспериментах с реальными последовательностями) состояла в следующем:

1. Для каждой пары последовательностей строится набор Парето-оптимальных выравниваний. Значение параметра R принималось равным 40. В качестве матриц весов замен использовались матрицы семейства PAM с эволюционным числом: 480 – для последовательностей с уровнем начального сходства (*%id*) от 7 до 12%, 360 – от 13 до 17%, 240 – от 17 до 28%, и 120 при уровне начального сходства более 28%. Для последовательностей из базы данных PREFAB-P были также проведены отдельные эксперименты с матрицей весов BLOSUM-62. Поскольку данный шаг является вычислительно трудоемким, то все внутреннее представление этого этапа сохраняется на диске (dump состояния). Это неоднократно использовалось позже, поскольку позволяло проводить различные эксперименты, не выполняя заново длительную процедуру получения множества Парето-оптимальных выравниваний. Вычисляются характеристики каждого Парето-оптимального выравнивания, в том числе – точность и достоверность (доля сопоставлений алгоритмического выравнивания, которые присутствуют в эталонном выравнивании). Как показали проведенные предварительные эксперименты, достовер-

ность глобальных алгоритмических выравниваний коррелирует с их точностью, поэтому данные о достоверности сохраняются только на данном этапе и в дальнейшем не используются. Также на данном этапе определяются характеристики выравнивания, которое имеет наибольшую точность. Это выравнивание далее называется *субэталонным*.

2. Среди Парето-оптимальных выравниваний для каждой пары последовательностей выделяются основные выравнивания, вычисляются сводные характеристики построенных выравниваний-кандидатов. Этап является несложным с вычислительной точки зрения, поэтому может повторяться многократно при рассмотрении различных гипотез о различных способах выбора выравниваний-кандидатов.

В ходе компьютерных экспериментов сравнивались полученные точности наборов основных выравниваний, состоящих из трех элементов (набор $E(3)$) и шести элементов (набор $E(6)$), с точностью соответствующих выравниваний Смита-Ватермана при значениях штрафов, приведенных в табл. 1. Данные сравнения точности наборов выравниваний-кандидатов с точностями выравниваний Смита-Ватермана, которые были определены на корпусе модельных данных со средней длиной последовательности 300, приведены в табл. 2. Эти данные согласуются с результатами выравниваний последовательностей с длинами 600 и 1000.

Результаты аналогичных экспериментов для последовательностей из базы данных PREFAB-P приведены в табл. 3. Здесь были проведены эксперименты как с одной из матриц семейства PAM (конкретная матрица выбиралась в соответствии со степенью сходства сравниваемых последовательностей, см. раздел 4.2 диссертации), так и с использованием матрицы BLOSUM-62 – лучшей «универсальной» матрицей замен.

В таблицах 2 и 3 столбец SW соответствуют точностям выравниваний Смита-Ватермана; эволюционные значения используемых в ходе экспериментов матриц семейства PAM приведены в квадратных скобках после соответствующих уровней сходства (столбец %id). Точностям выравниваний с использованием матрицы BLOSUM-62 в табл. 3 соответствует столбец BS-62.

Как видно из результатов, предложенный в диссертации метод имеет выигрыш в точности по сравнению с методом Смита-Ватермана в тех случаях, когда выравниваются слабогомологичные последовательности (процент сходства %id < 30). Для высокогомологичных последовательностей, использование метода построения выравниваний-кандидатов не имеет превосходства в точности. Эта закономерность характерна как для модельных последовательностей, так и для реальных исходных данных из базы данных PREFAB-P.

Таблица 2

Сравнение средней точности основных наборов $E(3)$, $E(6)$ со средними точностями субэталонных выравниваний и выравниваний Смита-Ватермана для модельных данных.

%id	SW	$E(3)$	$E(6)$	Субэтал. выр.
10 [480]	0.38	0.42	0.43	0.46
15 [360]	0.54	0.57	0.59	0.61
20 [240]	0.75	0.76	0.77	0.79
30 [120]	0.88	0.88	0.89	0.90

Таблица 3

Сравнение средней точности основных наборов $E(3)$, $E(6)$ со средними точностями субэталонных выравниваний и выравниваний Смита-Ватермана для последовательностей из базы данных PREFAB-P.

%id	SW		$E(3)$		$E(6)$		Субэтал. выр.	
	PAM	BS-62	PAM	BS-62	PAM	BS-62	PAM	BS-62
7..12 [480]	0.17	0.21	0.20	0.23	0.22	0.26	0.25	0.32
12..17 [360]	0.28	0.33	0.33	0.41	0.35	0.43	0.38	0.46
17..28 [240]	0.58	0.58	0.63	0.64	0.64	0.65	0.67	0.67
>28 [120]	0.90	0.90	0.90	0.90	0.90	0.91	0.91	0.92

Различия в абсолютных значениях точности выравниваний в таблицах 2 и 3 связаны с тем, что в модельных выравниваниях отсутствует невыровненные фрагменты на концах выравниваний. Тем не менее, из обеих таблиц видно, что средняя точность наборов основных выравниваний превосходит точность выравниваний Смита-Ватермана.

Важным практическим результатом, полученным в ходе компьютерных экспериментов, является то, что точность выравнивания набора из шести элементов не сильно отличается от точности набора всех Парето-оптимальных выравниваний (в среднем – на 3%). Это позволяет утверждать о том, что предложенный в работе метод ранжирования набора выравниваний-кандидатов с большой долей вероятности обеспечивает выбор небольшого набора кандидатов, в который попадает либо субэталонное выравнивание, то есть имеющее максимально возможную точность, либо выравнивание, близкое к субэталонному.

В **заключении** приведены **основные результаты,** полученные в диссертации:

1. Предложена новая постановка задачи глобального выравнивания биологических последовательностей – задача построения заданного количества ранжированных выравниваний-кандидатов.

2. Предложен алгоритм построения упорядоченного набора выравниваний-кандидатов, который не использует штрафы за удаления фраг-

ментов. Время работы алгоритма составляет $O(R \cdot L^2)$, где R – ограничение сверху на количество удаленных фрагментов в рассматриваемых выравниваниях, L – ограничение сверху на длину сравниваемых последовательностей.

3. Предложенный алгоритм реализован в виде общедоступного программного комплекса PARCA и соответствующего веб-сервиса, доступного по адресу: http://server2.lpm.org.ru/bio/online/pareto

4. Проведены вычислительные эксперименты на модельных и реальных тестовых данных, позволяющие сравнить точность наборов выравниваний, получаемых с помощью предложенного алгоритма и точность выравниваний Смита-Ватермана.

5. Показано, что наилучшее по точности выравнивание Смита-Ватермана в случае слабогомологичных последовательностей и использования матриц семейства PAM получается при значении параметра $GEP = 1.0$.

6. Построена база данных эталонных выравниваний PREFAB-P, которая может быть использована для оценки точности различных алгоритмов парного глобального выравнивания последовательностей.

РАБОТЫ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

1. В. В. Яковлев, М. А. Ройтберг. Увеличение точности глобального выравнивания аминокислотных последовательностей с помощью построения набора выравниваний-кандидатов. // Биофизика, 2010. Т. 55, N 6. – С. 965–975.

2. И. В. Поверенная, М. А. Ройтберг, В. В. Яковлев. Использование программы PARCA для выравнивания аминокислотных последовательностей. // Информационные процессы, 2011. Т. 11, N 4. – С. 510–519.

3. Т. В. Астахова, И. В. Поверенная, М. А. Ройтберг, В. В. Яковлев. Верификация базы эталонных выравниваний PREFAB. // Биофизика, 2012. Т. 57, N 2. – С. 205–211.

4. Poverennaya I., Lobanov M., Yacovlev V., Roytberg M.. Using of PREFAB for analysis of amino-acid sequence alignment algorithms. Proc, MCCMB'11. P. 327.