Вопросы медицинской химии, 2001, том 47 №, с.353-367

КОМПЬЮТЕРНОЕ КОНСТРУИРОВАНИЕ ЛЕКАРСТВЕННЫХ СРЕДСТВ

УДК 577.152.3.
©Коллектив авторов

КОМПЬЮТЕРНЫЙ ПОИСК НОВЫХ МИШЕНЕЙ ДЛЯ ДЕЙСТВИЯ ПРОТИВОМИКРОБНЫХ СРЕДСТВ НА ОСНОВЕ СРАВНИТЕЛЬНОГО АНАЛИЗА ГЕНОМОВ

А. В. ДУБАНОВ, А. С. ИВАНОВ, А. И. АРЧАКОВ

НИИ Биомедицинской химии им В.Н. Ореховича РАМН
119992, Москва, Погодинская ул., д.10 Факс: (095) 245-0857
Эл. почта:dubanov@ibmh.msk.su

Прогресс в расшифровке различных геномов открывает для исследователей возможность использования геномных баз данных при создании новых лекарственных средств. Большой практический интерес представляет их использование при поиске новых молекулярных мишеней для вновь создаваемых лекарств. Наиболее остро эта проблема стоит в области создания новых противомикробных средств. Применимость геномных подходов для решения этой проблемы была показана рядом авторов в последние годы.

Нами был предложен подход к поиску новых мишеней для действия противомикробных средств, основанный на сравнительном анализе геномов и выборок последовательностей из молекулярно-биологических баз данных. Для каждого из белков, кодируемых геномом целевого микроорганизма, выполняется оценка соответствия ряду медико-биологических и технологических требований. Полученные оценки используются для отбора потенциальных мишеней. Подход был реализован в виде оригинального программного обеспечения GenMesh. Выполненная проверка на примерах ряда известных белков-мишеней для действия противотуберкулезных средств показала корректные оценки. С помощью данного подхода был выполнен успешный поиск новых мишеней для действия противотуберкулезных средств.

Ключевые слова: сравнительный анализ геномов, противомикробные средства, молекулярная мишень, выбор мишени, туберкулез

ВВЕДЕНИЕ. В настоящее время большое число используемых противомикробных средств не удовлетворяет современным медико-биологическим требованиям. Так многие из них оказываются недостаточно эффективными в связи с появлением резистентных штаммов микроорганизмов. Ряд средств не удовлетворяет современным требованиям к безопасности применения. Выходом из сложившейся ситуации является создание новых противомикробных средств, действующих по молекулярным механизмам, отличным от уже известных. Это означает, что они должны воздействовать на новые молекулярные мишени в микробной клетке и, следовательно, поиск новых мишеней для противомикробных средств является актуальной проблемой.

В последние годы наблюдается прогресс по расшифровке геномов различных организмов и наиболее успешно геномов микроорганизмов. По данным Национального центра биотехнологической информации США (NCBI), уже полностью расшифрованы геномы 30 бактерий, в том числе и возбудителей заболеваний человека [1], а число изучаемых геномов микроорганизмов (бактерии, вирусы, грибки) измеряется сотнями. В ближайшее время следует ожидать завершения работ по расшифровке еще порядка 20 бактериальных геномов [2]. Результаты таких исследований, как правило, становятся доступными для научной общественности через публикации в глобальной компьютерной сети Интернет, создавая предпосылки их использования при создании новых противомикробных средств и в первую очередь для поиска новых мишеней. В литературе широко обсуждаются эти возможности и основное внимание сосредоточено на том, какими характеристиками должны обладать новые противомикробные средства, что в свою очередь определяет требования к молекулярным мишеням [3]. В табл. 1 приведены основные характеристики, которыми должны обладать новые противомикробные средства и их молекулярные мишени.

Первая попытка поиска мишени путем проверки соответствия этим требованиям каждого из белков, кодируемых геномом целевого микроорганизма, была осуществлена в программе CATS (Computer-Aided Target Selection and Prioritization) [4], предназначенной для поиска новых белков-мишеней противогрибковых средств. В основу ее алгоритма был положен известный способ полуколичественного сопоставления вариантов выбора с начислением баллов по определенным правилам. В качестве исходных данных выступают белковые последовательности, соответствующие открытым рамкам считывания сравниваемых геномов, а также дополнительная информация, включающая как экспериментальные, так и расчетные данные (необходимость белка для выживания микроорганизма, степень изученности данного белка, наличие в молекуле трансмембранных доменов и др.). Пригодность белка как мишени оценивалась по 4 критериям: (1) "качество" - роль гена для выживания микроорганизма; (2) "распространенность" - наличие близких гомологов данного белка у родственных видов микроорганизмов и у высших эукариот; (3) "специфичность" - сходство данного белка с белками родственных видов микроорганизмов и белками высших эукариот; (4) "разработка метода анализа" - биологические и технологические аспекты выбора мишени.

Значения параметров 2 и 3 рассчитываются путем сравнения геномов с помощью программы парного выравнивания FastA [5], позволяющего получить численную оценку сходства последовательностей. С помощью программы CATS авторы получили корректные оценки для ряда известных мишеней и нашли 25

новых белков в качестве потенциальных мишеней для противогрибковых средств. Хотя описанный подход и позволил сократить число рассматриваемых объектов с нескольких тысяч до десятков, нам представляется не вполне целесообразным использование в качестве критерия отбора простой суммы всех параметров оценки. Так как "удельный вес" каждого из параметров может значительно варьировать, то трудно определить степень важности каждого из них. Мы считаем, что в первую очередь необходимо расширить набор параметров, получаемых на основании сравнительного анализа геномов.

Таблица 1. Требования к противомикробному средству и соответствующие требования к белку-мишени

Требования к средству	Требования к мишени
Медико-биологические	
Необходимый спектр противомикробного действия	Мишень должна присутствовать во всех целевых видах микроорганизмов. Видовые отличия мишени должны быть возможно меньшими.
Отсутствие резистентности у целевых видов микроорганизмов	Отсутствие мутаций мишени, отсутствие специфичных механизмов резистентности, ассоциированных с данной мишенью.
Эффективное подавление роста и размножения микроорганизма	Необходимость нормального функционирования белка-мишени для роста и размножения микроорганизма.
Летальность для микроорганизма (желательное свойство)	Необходимость нормального функционирования мишени для выживания микроорганизма.
Возможно меньшая токсичность для макроорганизма	Белок-мишень не должен иметь близких гомологов среди белков макроорганизма.
Практические (технологические)	
"Прозрачный" механизм действия	Мишень должна быть детально изученным объектом или должна иметься возможность ее детального изучения
Возможность компьютерного конструирования прототипов новых лекарственных веществ, воздействующих на ту же мишень	Доступность пространственной структуры белка-мишени

Так как нам представлялась целесообразной дальнейшая разработка методов поиска новых мишеней для противомикробных средств на основе сравнительного анализа геномов, то целью данной работы явилась разработка нового подхода, его программная реализация и тестирование на примерах известных мишеней.

Мы сосредоточили свое внимание на оценках, получаемых путем сравнения последовательностей белков с помощью методов выравнивания. Описанные выше критерии (1) и (2), учитывающие только два медико-биологических требования к новому средству и его мишени, могут быть дополнены еще двумя медико-биологическими критериями: (5) "Отсутствие

мутаций белка в других штаммах целевого вида" - возможность возникновения резистентности к противомикробному средству путем мутации белка-мишени; критерий может быть рассчитан путем сравнения геномов различных штаммов целевого вида и выявления белков, в которых не встречается замен аминокислотных остатков; (6) "Наличие гомологов белка-мишени у микроорганизмов-симбионтов человека" - действие на белки микроорганизмов - симбионтов человека; критерий может быть рассчитан путем сравнения генома целевого микроорганизма с геномами микроорганизмов - симбионтов человека. Кроме того, может быть добавлен еще один технологический критерий: (7) "Доступность трехмерной структуры белка" - доступность трехмерной структуры белка-мишени; критерий может быть рассчитан путем сравнения белков целевого вида с белками, трехмерные структуры которых представлены в белковом банке PDB [6, 7].

Ограничение набора критериев только теми, которые могут быть оценены с помощью методов выравнивания последовательностей, позволит работать с однородными данными, что значительно облегчает их обработку и интерпретацию результатов. В дальнейшем по мере развития метода возможно включение дополнительных менее формализованных данных.

Исходя из изложенных выше соображений и сформулированных в табл. 1 требований к белкам-мишеням, мы решили, что целесообразно использовать следующие параметры:

I. Функция белка (необходимость нормального функционирования белка для выживания микроорганизма).

II. Наличие близких гомологов в геномах родственных видов.

III. Отсутствие мутаций белка в других штаммах целевого вида.

IV. Отсутствие близких гомологов в геноме человека.

V. Доступность пространственной структуры.

Наиболее важным параметром является первый критерий - функция белка. Очевидно, что воздействие на белок-мишень должно приводить к гибели микроорганизма или как минимум к угнетению его роста и размножения. Таким образом, параметр I должен представлять собой численную характеристику важности белка для жизнедеятельности организма. Она должна базироваться как на экспериментальных данных, так и на вычисленной вероятной роли белка с использованием аннотированных баз данных белковых последовательностей и метаболических путей. Эти данные обычно слабо формализованы, что создает значительные затруднения для их автоматизированной обработки, поэтому на данном этапе мы решили отказаться от программной реализации расчета параметра I. Однако это не означает отказ от учета функции белка при его выборе в качестве мишени. Так как использование остальных параметров (II-V) позволит значительно сократить выборку последовательностей, то вопрос об их использовании в качестве мишеней на основе данных об их роли в выживании микроорганизма может быть решен индивидуально для каждого белка.

Параметры II-IV отражают важнейшие медико-биологические требования к белкам-мишеням. Параметр V является "технологическим". Он позволяет оценить возможность применения прямых методов компьютерного конструирования лекарств на основе структуры белка-мишени, что позволит значительно сократить время и затраты на создание прототипа нового лекарственного вещества [8]. Практически этот параметр должен оценивать

возможность компьютерного моделирования трехмерной структуры белка по гомологии.

Для тестирования подхода были выбраны известные белки-мишени для действия противотуберкулезных средств. Выбор был обусловлен тем, что:

1. полностью расшифрован геном хорошо изученного лабораторного штамма *Mycobacterium tuberculosis* H37Rv [9] и эти данные доступны через Интернет [2];

2. большинство белков, кодируемых геномом *Mycobacterium tuberculosis* H37Rv, аннотировано и представлено в специализированных базах данных [10, 11];

3. расшифрован геном высоко контагиозного штамма *M. tuberculosis* CDC1551 [12], информация о котором доступна через Интернет [13]. Сравнение геномов двух штаммов одного вида микроорганизма позволяет выполнить расчет параметра IV;

4. известны последовательности большинства белков *M. leprae* [2], а ряд существующих противотуберкулезных средств активен в отношении этой бактерии, то это создает благоприятные условия для проверки оценки спектра действия с использованием параметра II;

5. механизмы действия ряда противотуберкулезных средств и резистентности к ним детально изучены и подробно описаны в литературе, что позволяет сравнить с ними оценки, получаемые с помощью данного метода.

МЕТОДИКА.Исходные данные. В качестве целевого генома был выбран геном наиболее изученного лабораторного штамма *M. tuberculosis* H37Rv [2, 10]. В работе использованы 3918 последовательностей белков, которые соответствуют открытым рамкам считывания ДНК бактериальной хромосомы, полученные из банка данных Entrez Genomes [1]. Для сравнения были использованы геномы штамма *M. tuberculosis* CDC1551, *M. leprae* и человека. Геном штамма CDC1551 [12] был получен с Web-сайта TIGR [13]. Были использованы 4269 пептидные последовательности, полученные путем трансляции соответствующих нуклеотидных последовательностей с использованием таблицы трансляции #11 [1]. В работе также использованы видовые выборки белковых последовательностей из базы данных GenBank [1], соответствующие геномам *M. leprae* и человека. Их объем составил соответственно порядка 2600 и 84000 последовательностей. Для определения возможности моделирования пространственной структуры белков использована выборка из PDB [6, 7], включающая все пептидные цепи, представленные в этом банке (около 25000 последовательностей). Мы приводим примерные количества последовательностей, поскольку объемы соответствующих баз данных постоянно растут и точное число белков в выборках изменялось даже во время выполнения работы.

Программное обеспечение. Сравнение геномов выполнялось с помощью пакета программ локального парного выравнивания BLAST 2.0.9 (Basic Local Alignment Search Tool, программы BLASTP и TBLASTN) [14]. Программа TBLASTN была использована для сравнения геномов *M. tuberculosis* H37Rv и CDC1551. В остальных случаях была использована программа BLASTP. Выравнивания были выполнены с использованием матрицы BLOSUM62. В выходную выборку BLAST включались последовательности, для которых значение ожидания присутствия в выборке сравнения такого белка, для которого значение оценочной функции выравнивания было бы большим или равным данному ("Expectation value"), составляло 0,001 и менее.

Выходные данные BLAST анализировались с помощью оригинального программного пакета GenMesh 0.2. Пакет включает в себя: программу расчета параметров отбора мишеней, конвертер выходных данных BLAST в оригинальный компактный входной формат GenMesh (текстовый файл, размеченный с помощью тегов) и ряд утилит для подготовки исходных данных для BLAST. Исходный текст пакета написан на языке C++ с использованием стандартной библиотеки шаблонов STL, что обеспечивает относительно высокую переносимость пакета между вычислительными платформами. Анализ выходных данных GenMesh был выполнен с помощью электронной таблицы Microsoft Excel 97 SR2.

Все вычисления были выполнены на компьютере типа Pentium MMX (233 МГц, 32 МБ RAM) с операционной системой Windows 95.

Тестирование. Проверка корректности оценок GenMesh была выполнена на примерах известных мишеней ряда противотуберкулезных лекарственных веществ: изониазида, этионамида, этамбутола, циклосерина, рифампицина, аминогликозидов и фторхинолонов. Были рассмотрены только те противотуберкулезные средства, которые действуют на строго определенные белки (табл. 2).

Для проверки корректности расчета параметра IV была использована информация из базы данных PDB Neighbors [1].

Таблица 2. Известные белки-мишени для действия некоторых противотуберкулезных средств.

№	Код[1]	Название[1]	Лекарственные вещества
1	Rv1908c, *katG*	Каталаза/пероксидаза G [2]	Изониазид, этионамид
2	Rv0824c, *desA1*	*ACP*-зависимые десатуразы [3]	
3	Rv1094, *desA2*		
4	Rv3229c, *desA3*		
5	Rv1484, *inhA*	Еноил-*ACP*-редуктаза	
6	Rv0667, *rpoB*	ДНК-полимераза, цепь *B*	Рифампицин
7	Rv0682, *rpsL*	Белок S12 30S-субъединицы рибосомы	Аминогликозиды, виомицин
8	Rv2981c, *ddlA*	D-аланил-D-аланинлигаза	*D*-Циклосерин
9	Rv3423c, *alr*	Аланинрацемаза	
10	Rv0006, *gyrA*	ДНК-гираза, цепь *A*	Фторхинолоны
11	Rv3794, *embA*	Арабинозилтрансферазы (вероятные)	Этамбутол
12	Rv3795, *embB*		
13	Rv3793, *embC*		

Примечание: [1]По базе данных TubercuList [11]. [2]Фермент не является мишенью, однако необходим для перехода лекарственных веществ в активную форму. [3]*ACP* (Acyl Carrier Protein) - белок-переносчик ацильных групп.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ. Условные обозначения и оценка сходства последовательностей. Для формального описания правил расчета был введен ряд условных обозначений.

Г - набор белковых последовательностей, кодируемый определенным геномом, или выборка последовательностей из базы данных, сформированная по определенному признаку:

$$\Gamma = [\pi_1, \pi_2, \ldots \pi_n],$$

где π_1, π_2, ... π_n – белковые последовательности, входящих в состав генома или выборки;

n - число последовательностей в составе генома (выборки).

Далее в тексте целевой геном и любая последовательность, принадлежащая ему, обозначается как Γ^0 и π^0 соответственно (верхний индекс "0" обозначает принадлежность объекта к целевому геному).

Для расчета параметров II и IV могут быть использованы группы из *N* числа геномов (выборок) - Γ^*:

$$\Gamma^* = [\Gamma_1, \Gamma_2, \ldots \Gamma_N]$$

Результатом сравнения пары последовательностей с помощью программы BLAST являются выравнивания фрагментов (φ) последовательностей, отвечающие критериям локального сходства, используемых этой программой. Далее такие фрагменты будут обозначаться как, а Абсолютное значение оценочной функции выравнивания фрагмента φ последовательности из целевого генома с гомологичным фрагментом последовательности из генома сравнения обозначаются как $s(\varphi^0, \varphi)$. При локальном выравнивании двух последовательностей может быть найдено несколько пар гомологичных фрагментов. Поэтому для локального выравнивания каждой пары последовательностей рассчитывалась сумма значений оценочной функции выравнивания для всех найденных пар гомологичных фрагментов:

$$S(\pi^0, \pi) = \sum_{i=1}^{n} s(\varphi_i^0, \varphi_i) \qquad (1),$$

где *n* - число локальных выравниваний (т.е. число пар гомологичных фрагментов) сравниваемых последовательностей.

Для общей оценки результата выравнивания двух белков был введен коэффициент сходства:

$$q(\pi^0, \pi) = \frac{S(\pi^0, \pi)}{S(\pi^0)} \qquad (2),$$

где $S(\pi^0)$ - собственное значение оценочной функции выравнивания для последовательности белка из целевого генома. Такое значение рассчитывалось с теми же параметрами локального выравнивания (матрица сравнения, фильтры и пр.), которые были использованы при выравнивании белка из целевого генома и белка из генома сравнения. Т.е. $S(\pi^0)$ является результатом выравнивания π^0 к самому себе и отражает степень сходства двух последовательностей, так как автоматически учитывает длину последовательности белка из целевого генома. Несмотря на то, что такой способ оценки степени сходства является упрощенным, ранее была показана его пригодность для поиска новых мишеней [4].

Обычно $0 \le q \le 1$. При этом значение $q = 1$ указывает на полную идентичность последовательностей (за исключением участков кода пониженной сложности, если был использован обнаруживающий их фильтр) и $q = 0$, если

гомологов при установленном значении ожидания найдено не было. Ситуация, когда $0 \le q \le 1$, является типичной при невысоком пороге значения ожидания, который был использован в данной работе. Однако, в некоторых случаях возможно ошибочное вычисление (завышение) коэффициента сходства, когда один фрагмент последовательности из целевого генома имеет сходство с двумя фрагментами одной последовательности из генома сравнения ($0 \le q \le 2$). Так как такие случаи достаточно редки, то в данной работе они не рассматривались.

При скрининге генома может быть обнаружено несколько гомологов белка из целевого генома. При этом может быть получен ряд (массив) значений q (Q):

$$Q(\pi^0, \Gamma) = [q(\pi^0, \pi_1), q(\pi^0, \pi_2), \ldots q(\pi^0, \pi_n)],$$

где n - число гомологов белка π^0 из целевого генома, обнаруженных в геноме сравнения Г.

Расчет параметров оценки. Для каждого из белков целевого генома были рассчитаны параметры II-V. Способы расчета этих параметров излагаются в порядке усложнения их методов и интерпретации результатов.

"Наличие близких гомологов в геноме человека" (IV). Наличие гомологов потенциального белка-мишени среди белков, кодируемых геномом человека, следует рассматривать как неблагоприятный фактор. Соответствующий параметр частично отражает потенциальный спектр побочных эффектов вновь создаваемого противомикробного средства. Параметр рассчитывался следующим образом:

$$X_{IV}(\pi^0) = \max Q(\pi^0, \Gamma) \qquad (3),$$

где $X_{IV}(\pi^0)$ - значение параметра для белка π^0, Г - геном человека (видовая выборка белковых последовательностей из GenBank), max Q - наибольшее значение коэффициента сходства в ряду найденных гомологов.

Следует отметить, что многие белки представлены в GenBank как в виде полных последовательностей, так и в виде их фрагментов, также дублированы с различными идентификаторами. Данный способ расчета учитывает только наибольшее выявленное сходство между двумя последовательностями, что позволяет исключить многократный учет повторяющихся последовательностей. Очевидно, что использование этого параметра не сможет исключить всех нежелательных эффектов нового противомикробного средства на макроорганизм, однако вероятность воздействия побочных эффектов несомненно может быть снижена.

"Наличие близких гомологов в геномах родственных видов" (II). Этот параметр отражает возможный спектр действия создаваемого противомикробного средства. Наиболее благоприятным является наличие близких гомологов белка из целевого генома во всех геномах сравнения. Поэтому расчет параметра был ориентирован на поиск наиболее гомологичных белков во всех геномах, соответствующих желаемому спектру действия. Для каждого белка из целевого генома учитывались как степень сходства с найденными гомологами из геномов сравнения, так и число геномов, в которых эти гомологи были найдены:

$$X_{II}(\pi^0) = \frac{1}{N}\sum_{i=1}^{N} \max Q_i(\pi^0, \Gamma_i) \qquad (4),$$

где $X_{II}(\pi^0)$ - значение параметра для белка π^0 из целевого генома, Γ_i - геном сравнения, max Q_i - наибольшее значение коэффициента сходства в ряду найденных гомологов в геноме Γ_i, N - число геномов сравнения.

В случае единственного генома сравнения формула упрощается до вида уравнения (3).

"Отсутствие мутаций белка в других штаммах целевого вида" (III). Одним из наиболее распространенных механизмов лекарственной резистентности микроорганизмов является мутация белка-мишени [15]. Отсутствие различий в последовательности целевого белка в геномах других штаммов целевого вида следует рассматривать как благоприятный фактор. Поэтому расчет параметра был ориентирован на поиск одинаковых белков, а не на выявление межгеномных различий:

$$X_{III}(\pi^0) = \frac{1}{N}\sum_{i=1}^{N} C\left[\max Q_i\,(\pi^0, \Gamma_i) = 1 \quad (5),\right.$$

где $X_{III}(\pi^0)$ - значение параметра для белка π^0 из целевого генома, C - число геномов в наборе сравнения Γ^*, для которых соблюдается условие max Q_i (π^0, Γ_i)= 1, т.е. белков, полностью идентичных белку π^0.

Известно, что для возникновения резистентности бывает достаточно замены небольшого числа аминокислотных остатков. Формула (5) исключает из списка потенциальных мишеней все белки, в которых имеется хотя бы одна замена вне зависимости от ее местоположения. Очевидно, что данная оценка зависит от числа анализируемых геномов. К сожалению, в настоящее время для большинства микроорганизмов изучены геномы только одного, реже – двух штаммов [1, 2]. Поэтому в данной работе этот критерий не является характеристикой частот возникновения замен в белках различных штаммов *M. tuberulosis*.

"Доступность пространственной структуры" (V). В настоящее время широко применяются компьютерные методы конструирования лекарств на основе структуры белка-мишени. Поэтому доступность пространственной структуры мишени является не только благоприятным фактором, но часто и необходимым условием для успешного создания нового лекарственного вещества. Однако к настоящему моменту известны пространственные структуры лишь небольшого числа всех известных белков. Поэтому часто прибегают к компьютерному моделированию пространственной структуры белков-мишеней по гомологии. Следовательно, наличие близких гомологов с известной пространственной структурой, следует рассматривать как благоприятный фактор. Необходимо отметить, что наличие такого гомолога позволяет более точно охарактеризовать и функцию целевого белка. Расчет параметра был ориентирован на поиск наиболее гомологичных белков и получение результатов в форме, удобной для оценки возможности моделирования пространственной структуры белка по гомологии. Данный параметр может быть рассчитан по формуле (3), однако более информативной и привычной характеристикой является доля идентичных аминокислотных остатков при выравнивании целевого белка и белка-шаблона. Согласно результатам экспериментов CASP [16], для успешного моделирования необходимая доля таких остатков должна составлять не менее 0,4. Доля идентичных остатков в локальном парном выравнивании вычислялась аналогично коэффициенту сходства:

$$ID(\pi^0, \pi) = \sum_{i=1}^{n} id(\varphi_i^0, \varphi_i),$$

$$q_{ID}(\pi^0, \pi) = \frac{ID(\pi^0, \pi)}{ID(\pi^0)},$$

$$Q_{ID}(\pi^0, \Gamma) = [\, q_{ID}(\pi^0, \pi_1), q_{ID}(\pi^0, \pi_2), \ldots q_{ID}(\pi^0, \pi_n)],$$

$$X_V(\pi^0) = \max Q_{ID}\ (\pi^0,\ \Gamma),$$

где: $id(\varphi^0_i,\ \varphi_i)$ - количество идентичных остатков в выравнивании гомологичных фрагментов последовательностей белка из целевого генома и белка с известной пространственной структурой; $ID(\pi^0,\ \pi)$ - количество идентичных остатков в выравнивании последовательностей данных белков; $ID(\pi^0)$ - количество идентичных остатков при выравнивании целевой последовательности к самой себе (численно соответствует длине последовательности, но только в том случае, если в программе локального выравнивания отключен фильтр кода низкой сложности, в противном случае она меньше длины последовательности); $q_{ID}(\pi^0,\ \pi)$ - доля идентичных остатков; $Q_{ID}(\pi^0,\ \Gamma)$ - массив значений $q_{ID}(\pi^0,\ \pi)$ для белка π^0, полученной в результате сравнения с выборкой белков из PDB Г; $X_V(\pi^0)$ - значение параметра.

Следует отметить, что данная оценка не является точной и окончательной. Такой упрощенный способ был выбран, исходя из необходимости быстрого получения оценок для большого числа последовательностей. После сокращения числа выборки до десятков белков для оценки возможности моделирования их пространственной структуры могут быть применены более точные и сложные методы. В выходных данных BLAST для целевого белка фактически содержится список его ближайших гомологов с известной пространственной структурой и это может быть использовано при моделировании.

В качестве выборки сравнения использовалась выборка белковых последовательностей из PDB.

Результаты тестирования. Полученные оценки для мишеней известных противотуберкулезных средств приведены на рис. Рассмотрим некоторые результаты более подробно, учитывая при этом что все перечисленные в табл. 3 противотуберкулезные средства активны в отношении штамма *M. tuberculosis* H37Rv [2]. Мы также считали, что эти средства эффективны также и против штамма CDC-1551, так как данные о резистентности этого штамма отсутствуют. Кроме того, необходимо помнить, что видовые выборки из GenBank включают в себя только часть открытых рамок чтения (белков) соответствующих геномов.

При тестировании мы не принимали во внимание нулевые значения параметров II и IV. Следует отметить, что неполные данные о белках человека могут привести к включению в конечную выборку белков микроорганизмов, которые могут иметь гомологичные фрагменты с белками человека. Этот недостаток информации может быть частично компенсирован включением в выборку сравнения белков других высших млекопитающих, например, мыши [4].

Изониазид и этионамид. Эти вещества воздействуют преимущественно на *ACP*-десатуразы (*desA1*, *desA2*, *desA3*) и еноил-*ACP*-редуктазу (*inhA*) [23]. Для того чтобы эти лекарственные вещества в бактериальной клетке перешли в активную форму, необходим фермент каталаза-пероксидаза (*katG*) [17].
Для всех этих белков были получены нулевые значения параметра III.

При визуальном анализе выравниваний, полученных с помощью программы BLAST, было обнаружено, что последовательности *katG* в обоих штаммах различаются только первым N-концевым остатком. Последовательности *desA2* идентичны в обоих штаммах. Таким образом, возможно действие изониазида и этионамида по крайней мере на одну из ACP-десатураз. Этот результат согласуется с имеющимися данными о чувствительности штамма CDC-

1551 к противомикробным средствам [13] Таким образом для мишеней изониазида и этионамида были получены корректные значения параметра III.

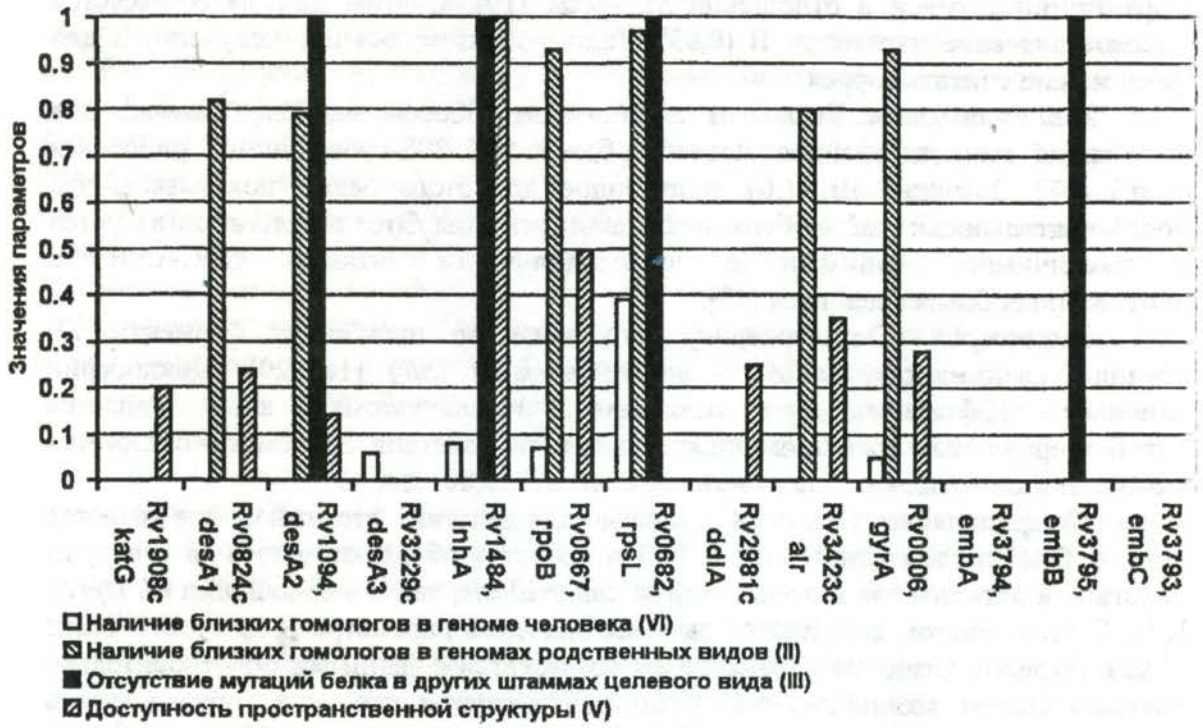

Таблица 3. Правила выбора новых мишеней и результаты их применения к геному *M. tuberculosis* H37Rv

Параметр отбора	Формулировка правила	Значения параметра (X)	Число белков, удовлетворяющих правилу
"Наличие близких гомологов в геноме человека"	Не должно быть гомологов белка-мишени среди белков, кодируемых геномом человека	$X_{IV} = 0$	2882
"Отсутствие мутаций белка в других штаммах целевого вида"	Последовательности белка-мишени должны быть идентичны в различных штаммах целевого вида	$X_{III} = 1$	1880
"Наличие близких гомологов в геномах родственных видов"	Должны быть близкие гомологи белка-мишени у родственных видов	$X_{II} >> 0$	-
"Доступность пространственной структуры"	Должны быть гомологи с известной пространственной структурой с идентичностью с белком-мишенью 0,4 и выше	$X_V \geq 0,4$	144
Число последовательностей, удовлетворяющих всем правилам			13

Рифампицин. Мишенью для действия этого средства и его аналогов является цепь *B* ДНК-полимеразы (*rpoB*) [18]. Помимо *M. tuberculosis*, рифампицин активен в отношении *M. leprae* [19]. С этим фактом согласуется высокое значение параметра II (0,93). Таким образом, оценку, полученную для *rpoB*, можно считать корректной.

Аминогликозиды. Белковым компонентом рибосом, наиболее важным для связывания аминогликозидов, является белок S12 30S-субъединицы рибосомы (*rpsL*) [17]. Значение III (1,0), полученное для этого белка, показывает, что последовательности *rpsL* в обоих штаммах идентичны. Этот результат согласуется с имеющимися данными о чувствительности штамма CDC-1551 к противомикробным средствам [13].

Циклосерин (*D*-циклосерин). Это вещество ингибирует ферменты *D*-аланил-*D*-аланинлигазу (*ddlA*) и аланинрацемазу (*alr*) [18, 20]. Циклосерин наиболее эффективен в отношении *M. tuberculosis*, хотя является антибактериальным средством широкого спектра действия. Это также согласуется с относительно высоким значением параметра II (0,80) для *alr*.

Фторхинолоны. Основной мишенью для действия фторхинолонов является цепь *A* ДНК-гиразы (*gyrA*) [15, 17]. Эти средства обладают широким спектром действия и активны как в отношении *M. tuberculosis*, так и в отношении *M. leprae* [21]. С этим фактом согласуется высокое значение параметра II для этого белка (0,93). Нулевое значение параметра III согласуется с данными об относительно высокой частоте возникновения мутаций ДНК-гиразы, что может обусловливать резистентность к фторхинолонам [15]. Таким образом, для *gyrA* было получено корректное значение параметров II и III.

Этамбутол. Мишенями для действия этамбутола являются арабинозилтрансферазы (*embA*, *embB*, *embC*) [17]. Достоверно не установлено, действует ли это вещество на все эти ферменты, или на один из них. Кроме того, в базе данных TubercuList [11] эти объекты обозначены как "вероятные". Одна из этих мишеней имеет идентичные последовательности в обоих штаммах (по результатам расчета параметра III). Этот результат согласуется с имеющимися данными о чувствительности штамма CDC-1551 к противомикробным средствам. Значения параметра IV для большинства мишеней не превышают 0,1, что представляется вполне корректным исходя из таксономической удаленности организмов, геномы которых сравнивались. Исключение составляет рибосомальный белок *rpsL*, который имеет относительно высокий уровень сходства с соответствующим белком рибосом человека (значение параметра равно 0,39). Это согласуется с данными о том, что возможно взаимодействие аминогликозидов с рибосомами человека с нарушением их функции [22].

Параметр V принимает значения в широком диапазоне. Для *inhA* значение этого парметра равно 1, так как его пространственная структура установлена экспериментально (код PDB - 1bvr) [23]. Большинство белков (*katG*, *desA1*, *desA2*, *rpoB*, *ddlA*, *alr*, *gyrA*), согласно данным PDB Neighbors [1], имеют сходство с белками с известной пространственной структурой. Значения параметра для этих белков находятся в интервале от 0,14 до 0,5, что говорит об их относительно низком сходстве с белками из PDB [1]. Для остальных мишеней близких гомологов с известной пространственной структурой нет.

Таким образом, тестирование подхода показало, что полученные оценки для известных мишеней ряда противотуберкулезных средств согласуются с информацией, имеющейся в литературе и базах данных и, следовательно, пакет

GenMesh дает корректные оценки и может быть использован для поиска новых белков-мишеней для противомикробных средств.

Выбор новых мишеней. Нами был осуществлен поиск новых мишеней для создания новых противотуберкулезных средств на основании выполненного анализа. Использованные правила и результаты их применения приведены в табл. 3. В результате были выбраны 13 белков из генома *M. tuberculosis* H37Rv. Аннотации этих белков были получены из баз данных TubercuList и Swiss-Prot. Верификация результатов отбора при поиске гомологов с доступной пространственной структурой выполнялась с использованием базы данных PDB Neighbors. Семь из 13 белков участвуют в матричном биосинтезе, один - в биосинтезе компонентов клеточной стенки, четыре - в различных промежуточных метаболических реакциях и один является компонентом клеточной мембраны. По своим функциям пять из этих белков являются ферментами или субъединицами ферментов, один - сигнальной молекулой, один - транспортным белком, один - каналом, три - рибосомными белками и один - гистоноподобным белком.

Все 13 последовательностей имеют гомологию с рядом белков с известной пространственной структурой и для девяти белков (в том числе четырех ферментов) компьютерное моделирование их пространственной структуры по гомологии может быть выполнено без особых затруднений и с высокой степенью достоверности.

По крайней мере четыре белка входят в состав комплексов, в которых присутствуют другие белки известные как мишени применяемых в настоящее время противотуберкулезных средств. Таким образом, с помощью GenMesh была получена относительно небольшая выборка белков, удовлетворяющих важнейшим требованиям, предъявляемым к мишеням для действия противомикробных средств. Из баз данных получены подробные аннотации для каждого из белков. Аннотации содержат сведения о функции белков, что может быть использовано при окончательном выборе белков-мишеней.

Благоприятным фактом является присутствие в выборке белков, участвующих в матричном биосинтезе и синтезе компонентов клеточной стенки, а также четырех белков, входящих в состав комплексов, другие белки которых являются известными мишенями для действия противотуберкулезных средств и следовательно имеется возможность распространить известные механизмы действия на новые белки-мишени, что позволит создать средства не менее эффективные, чем уже известные, но более безопасные.

Следует отметить наличие в выборке потенциальных мишеней четырех ферментов. Использование фермента в качестве мишени для нового лекарства, являющегося его ингибитором, позволяет добиться высокой специфичности действия. Кроме того, методы компьютерного конструирования специфичных лигандов по структуре макромолекулы-мишени наиболее разработаны применительно к ферментам.

При поиске мишеней в геноме *M. tuberculosis* параметром, обусловливающим наибольший отсев вариантов, является параметр V. Сохраняется вероятность того, что оптимальная по другим параметрам мишень не попадает в конечную выборку белков с известной трехмерной структурой или для которых имеется близкий гомолог с известной трехмерной структурой. В то же время отказ от использования этого параметра приводит к значительному увеличению объема выборки. В случае *M. tuberculosis* число белков, удовлетворяющих правилам выбора для параметров III и IV (табл. 3), составляет 1398. При

этом только белков, имеющих значение параметра $\Pi \geq 0{,}9$ обнаруживается около 40. Окончательное решение о выборе каждого из этих белков в качестве мишени может быть сделано на основании данных о важности его функции для выживания микроорганизма. Как было отмечено, такие данные в настоящее время плохо формализованы, что препятствует их автоматическому анализу и вопрос о разработке простого и эффективного способа получения такой оценки в виде удобной для интерпретации числовой величины остается открытым.

В целом использованный в данной работе набор параметров отбора можно считать достаточно эффективным для поиска новых белков-мишеней для создаваемых противомикробных средств. Предложенные принципы расчета параметров могут быть использованы для оценки соответствия белков и другим требованиям, которые могут быть сформулированы для мишеней противомикробных средств. Так, при необходимости исключения воздействия создаваемого противомикробного средства на микроорганизмы-симбионты человека, можно осуществить оценку наличия гомологов белка-мишени у соответствующих микроорганизмов. При этом к набору геномов сравнения должны быть добавлены известные геномы микроорганизмов-симбионтов, а соответствующий параметр может быть рассчитан по уравнению (4).

ЗАКЛЮЧЕНИЕ. 1. Предложен подход к поиску новых мишеней для действия противомикробных средств на основе сравнительного анализа геномов. По сравнению с существующим аналогом значительно увеличен объем информации, получаемый путем сравнения непосредственно последовательностей белков, кодируемых геномами. Предложенный подход реализован в виде оригинального программного пакета GenMesh.

2. На примере известных мишеней ряда противотуберкулезных средств показана корректность оценок, получаемых с помощью пакета GenMesh.

3. Поиск новых белков-мишеней в геноме *M. tuberculosis* показал, что, несмотря на ряд недостатков, использование пакета GenMesh позволяет сократить число рассматриваемых объектов с нескольких тысяч до десятков и единиц на начальном этапе и предоставляет в распоряжение исследователей информацию, необходимую для правильного выбора белка-мишени.

4. Использование пакета GenMesh позволяет выбрать целевые белки микроорганизмов, на которые должно воздействовать создаваемое лекарственное вещество, с учетом необходимого спектра противомикробного действия и уменьшения риска воздействия на белки макроорганизма. Кроме того, GenMesh позволяет ограничить круг целевых белков теми, для которых прототипы лекарственных веществ могут быть найдены с использованием компьютерных методов конструирования по трехмерной структуре белка-мишени.

Работа была выполнена при частичной поддержке Минпромнауки РФ (ФЦНТП "Исследования и разработки по приоритетным направлениям развития науки и техники гражданского назначения", подпрограмма "Создание новых лекарственных средств методами химического и биологического синтеза", направление "Компьютерное конструирование лекарств").

ЛИТЕРАТУРА

1. National Center for Biotechnology Information. *http://www4.ncbi.nlm.nih.gov/*
2. The Sanger Centre. *http://www.sanger.ac.uk/*
3. *Allsop A. E.* (1998) Current Opinion in Microbiology, **1**, 530–534.

4. *Spaltmann F. et al.* (1999) Drug Discovery Today, **4**(1), 17-26.
5. *Pearson W. R. et al.* (1998) Proc. Natl. Acad. Sci USA, **85**, 2444-2448.
6. *Berman H. M. et al.* (2000) Nucleic Acids Research, **28**, 235-242.
7. Research Collaboratory for Structural Bioinformatics (RCSB). Protein Data Bank (PDB). *http://www.rcsb.org/pdb/*
8. *Арчаков А. И., Иванов А. С.* (1996) Вестник РАМН, **1**, 60-63.
9. *Cole S. T., et al.* (1998) Nature, **393**, 537-544.
10. Expert Protein Analysis System (ExPASy). Seuqence Retrieval System (SRS). *http://www.expasy.ch/srs5/*
11. TubercuList. Institut Pasteur. *http://genolist.pasteur.fr/TubercuList/*
12. *Fleischmann R. D. et al.* (1995) Science, **269**, 496-512.
13. The Institute for Genomic Research (TIGR). *http://www.tigr.org*
14. *Altschul, S. F. et al.* (1997) Nucleic Acids Research, **25**, 3389-3402.
15. *Drilka K.* (1999) Current Opinion in Microbiology, **2**, 504–508.
16. *Sanchez R. et al.* (1997) Curr. Opinion Struct. Biol., **7**, 206-214.
17. *Piddock L. J. V.* (1998) Current Opinion in Microbiology, **1**, 502-508.
18. *Chopra I. et al.* (1997) Tuber. Lung. Dis., **78**(2), 89-98.
19. *Машковский М. Д.* (1993) Лекарственные средства. Москва: Медицина.
20. *Гейл Э. и др.* (1975) Молекулярные основы действия антибиотиков. Москва: Мир.
21. *Jacobs M. R.* (1999) Drugs, **58**, 19-22.
22. *Manuvakhova M. et al.* (2000) RNA, **6**(7), 1044-55.
23. *Rozwarski D. J. et al.* (1999) J. Biol. Chem., **274**(22), 15582-15589.

Поступила 03.01.01.

COMPUTER SEARCHING OF NEW TARGETS FOR ANTIMICROBIAL DRUGS BASED ON COMPARATIVE ANALYSIS OF GENOMES

A. V. DUBANOV, A. S. IVANOV, A. I. ARCHAKOV

Orekhovich institute of Biomedical Chemistry, RAMS
Pogodinskaya str. 10, Moscow, 119992, Russia;
Fax: (095) 245-0857, E-mail: dubanov@ibmh.msk.su

The progress in genome research allows to use genomic databases for drug discovery. The major interest consists in their usage for searching of new molecular targets for new drugs. It is especially important in the area of new antimicrobial drugs creation. In recent years, the applicability of genome analysis for solution of this problem was shown by different authors. We propose an approach for searching new targets for antimicrobial drugs based upon comparative analysis of genomes and samples from molecular databases. For each protein encoded by the target microorganism genome the conformance to a number of medico-biological and technological requirements was tested. The obtained evaluations were used for selection of potential targets. The approach was implemented in the original software GenMesh. It was successfully tested with known targets of drugs against tuberculosis when correct valuations were obtained. The attempt of new targets selection for design of new drugs against tuberculosis was done with reliable results.

Keywords: comparative analysis of genomes, antimicrobial drugs, molecular target, target selection, tuberculosis