БИОФИЗИКА, 2005, том 50, вып.4, с.752–766

ДИСКУССИИ

УДК 577.3

АПОЛОГИЯ БИОИНФОРМАТИКИ

© 2005 г. М.С. Гельфанд

Институт проблем передачи информации РАН, 101447, ГПС–4, Москва, Большой каретный пер., 19

Поступила в редакцию 01.03.04 г.

Науку надо любить: у людей нет силы более мощной и победоносной, чем наука.

Из теста Единого государственного экзамена по русскому языку. «Известия» 10.6.2004 г.

Продолжается дискуссия об адекватном понимании термина «биоинформатика». Рассматриваются взаимоотношения между биоинформатикой и экспериментальной молекулярной биологией. Приводится сводка основных направлений и достижений современной биоинформатики.

Ключевые слова: биоинформатика.

В настоящее время в России существуют два понимания биоинформатики*. Широкое понимание отражено, в частности, в формуле специальности ВАК 03.00.28 («Биоинформатика» – научная специальность, занимающаяся изучением организации и функционирования биологических систем разного уровня (от молекулярного до популяционного) на основе методов и средств информатики. Решение научных проблем данной специальности имеет как фундаментальное, так и прикладное значение»)** и программе курса биоинформатики для кафедры биофизики биологического факультета МГУ [1]. Последний включает в себя различные определения информации, проблему возникновения и эволюции генетического кода, модели онтогенеза, искусственные и природные нейронные сети, задачу распознавания образов и т.д. вплоть до пункта «Возможный механизм пунктурной терапии. Интерпретация понятий восточной медицины на языке фазового пространства математических моделей современных нейропроцессоров».

В узком смысле биоинформатика понимается как область науки, тесно связанная с молекулярной биологией [2–5]. Такое понимание отражено в формирующейся программе факультета биоинженерии и биоинформатики МГУ, цитаты из одного из неопубликованных вариантов которой приведены в [6], а также во многих других источниках, перечисленных ниже. Определение в Оксфордском словаре английского языка (http://www.oed.com) начинается как широкое, но потом почти сразу сводится к узкому. Набрав «define:bioinformatics» в окне запроса поисковой системы Google (http://

* Впрочем, с точки зрения встроенного корректора орфографии программы Microsoft Word «биоинформатики», равно как и «bioinformatics», не существует вовсе.

** Согласно паспорту специальности ВАК 03.00.28, область исследований по биоинформатике – это:

1. Исследование эволюции живой природы с помощью средств информатики и математики.
2. Компьютерное и математическое моделирование информационных процессов в биологических системах.
3. Компьютерная генетика: расшифровка и моделирование структурной организации генов и геномов, а также кодируемых генами белков; корреляционный анализ мутаций и др.
4. Компьютерная нейробиология: моделирование природных нейронных систем, разработка и приложение нейросетей и др.
5. Исследование экологических систем с помощью информационных технологий.
6. Компьютерное моделирование биологического действия ксенобиотиков.
7. Компьютерное моделирование процессов получения, накопления, обработки и систематизации биологических данных.
8. Компьютерное распознавание и синтез изображений биологических объектов.
9. Создание новых информационных технологий на основе результатов исследований живой природы.
10. Организация, ведение и использование автоматизированных банков данных по биологии и медицине, в т.ч. банков междисциплинарных данных.
11. Разработка интеллектуальных систем анализа и прогнозирования свойств биологических объектов на основе специализированных баз и банков данных.
12. Создание систем информационного обеспечения и поддержки биологических и медицинских исследований, включая анализ точек роста и тенденций развития научных направлений.

www.google.com), можно увидеть примерно две дюжины определений из разных страниц интернета, более половины которых понимают «биоинформатику» исключительно в узком смысле, а большая часть других следует пути Оксфордского словаря.

Казалось бы, de verbis non est disputandum: спор о словах бессмыслен. Однако поскольку III Биофизический съезд России и руководство Научного совета по биофизике постановили обсудить вопрос о том, что составляет предмет биоинформатики [6], а журнал «Биофизика» открыл соответствующую дискуссию, я попробую рассмотреть некоторые вопросы, связанные со вторым, узким пониманием. При этом я постараюсь выйти за рамки чисто терминологических споров и посвятить основную часть статьи обсуждению того, как биоинформатика вписывается в современный контекст развития биологии.

В статье Л.М. Чайлахяна [6] приводятся доводы в пользу применения термина «биоинформатика» исключительно в широком смысле. Признавая существование в России соответствующей научной традиции, следует отметить, однако, что это может привести к ряду проблем. Одной из них является неминуемое расхождение в отечественной и международной терминологии. Простой анализ содержания всех международных периодических изданий, имеющих в своем названии слово «биоинформатика» («Bioinformatics», «BMC Bioinformatics», «Briefings in Bioinformatics», «Journal of Bioinformatics and Computational Biology», «Lecture Notes in Bioinformatics»), показывает, что оно употреблено в узком значении. Более того, изменению названия ведущего профессионального журнала (в 1998 году журнал «Computer Applications in the Biosciences» стал называться «Bioinformatics») предшествовало соответствующее изменение содержания. Аналогичные выводы можно сделать из рассмотрения названий конференций, объявлений о вакансиях в исследовательских учреждениях и университетах, материалов учебных курсов по биоинформатике и т.п. Поучительно также просто сопоставить названия ряда научных учреждений мира и область их деятельности: Европейский институт биоинформатики (http://www.ebi.ac.uk/) и сеть биоинформатических центров (http://www.nbcc-online.de/), институты биоинформатики Швейцарии (http://www.isb-sib.ch/), Германии (http://mips.gsf.de/), Южной Африки (http://www.sanbi.ac.za/), Сингапура (http://www.bii.a-star.edu.sg/), Вирджинии (https://www.vbi.vt.edu/), Новой Зеландии (http://www.cebl.auckland.ac.nz/bi/) и др.

Тем самым, есть серьезная опасность оказаться в ситуации, когда «вся рота идет не в ногу, один командир – в ногу»: нас просто не будут понимать. В этой связи замечание о том, что узкое понимание «достаточно активно внедряется во все сферы научно-организационной и научно-педагогической деятельности» [6] представляется не вполне корректным – такое понимание не навязывается кем-либо, а сложилось в мировой практике само собой в результате естественного развития научного дискурса. Посчитано, что около 2% статей в литературной базе данных PubMed, включающей статьи по молекулярной биологии, – это статьи по биоинформатике [3].

С другой стороны, даже если предположить, что можно переломить сложившуюся традицию, не вполне понятно, нужно ли это: в чем необходимость перехода от достаточно устоявшегося понимания биоинформатики в пользу широкого понимания и зачем придумывать для первой новые названия? Как справедливо указывает Л.М. Чайлахян в первом же абзаце своей статьи [6], биоинформатика в широком смысле является синонимом кибернетики – зачем же отказываться от этого вполне ясного и традиционного термина. Есть названия и для области науки, предметом изучения которой является человеческий мозг (она была упомянута как вызывающая наибольший интерес): «neurobiology» и «cognitive science», и проблема тут только в поиске адекватного перевода на русский язык. В журнале «Биофизика» есть специальный раздел «Биофизика сложных систем», где публикуются работы по разного рода моделированию. Поведением животных, в том числе коммуникацией, занимается этология. Взаимодействиями видов и популяций – экология. Наконец, существует достаточно общая дисциплина – теоретическая биология. В то же время предлагаемые варианты названий для «узкой биоинформатики» либо не привились («компьютерная генетика»), либо изобретены заново и еще долго не будут понятны, да и не вполне адекватны («генинформатика»), либо, наконец, просто громоздки («вычислительная молекулярная биология»).

Еще более существенной причиной для отказа от широкого понимания биоинформатики может быть то, что при этом термин «биоинформатика» становится безразмерным. Неясно, каковы точки соприкосновения между специалистами в перечисленных выше областях «широкой биоинформатики», что могло бы служить предметом общего обсуждения, помимо крайне абстрактных и плохо формулируемых проблем «информационных процессов в биологии». В этой связи весьма поучительно сравнение биохимии и биофизики, проведенное Л.М. Чайлахяном [6]. Действительно, вопрос о том, что такое биохимия, не возникает прежде всего в силу того, что это достаточно четко очерченная область науки, в то время как разъяснению

того, что такое биофизика, пришлось посвятить специальный экскурс в «*семантику понятий*». Хотим ли мы, чтобы то же случилось и с биоинформатикой? Можно вспомнить, что судьба наук, претендовавших на роль объединителя различных аспектов изучения информационных процессов, оказалась достаточно печальной. Еще Клод Шеннон, основатель современной теории информации, указывал на опасности, связанные с неумеренно широким пониманием термина «информация» [7]. Похоже, что семиотика, несмотря на важную роль, которую она сыграла в уточнении многих понятий лингвистики, филологии, теории коммуникаций, культурологии и т.п., так и не сложилась в единую научную дисциплину [8].

Разумеется, сказанное выше не означает призыва немедленно «*внедрить*» узкое понимание биоинформатики в существующие нормативные документы. В то же время, по-видимому, было бы разумно дополнить, где это необходимо, устаревшую номенклатуру. Так, например, в Государственном рубрикаторе научно-технической информации «биоинформатика» расшифровывается всего лишь как «информационная деятельность», «биокоммуникация» и «нейрокибернетика», что явно недостаточно (http://www.extech.ru/php/tolko_vak/sootv.php?kod=03.00.28).

Можно выделить три основные области биоинформатики. Во-первых, это относительно прикладные направления, связанные с хранением и обработкой данных. К ним относятся процедуры обработки, применяемые в машинах для автоматического секвенирования, и процедуры выделения сигнала в экспериментах по анализу экспрессии на микрочипах и т.п., в которых интенсивно используются алгоритмы анализа изображений и распознавания образов. Далее, к этой области относятся вопросы разработки баз данных для хранения полученной информации. Хотя на первый взгляд это кажется чисто технической задачей, объемы данных и размеры потоков запросов к таким базам делают их конструирование вполне нетривиальной проблемой. Наконец, сюда же можно причислить создание алгоритмов анализа текстов, например, научных статей, с целью выделения, сбора и систематизации содержащейся в них информации.

Вторая область – это функциональный анализ молекулярных данных. Сюда относятся задачи разметки генома, в частности, распознавания генов и регуляторных сигналов, предсказания структуры и функции белка по последовательности, а также разработка и применение методов сравнительной геномики для функциональной аннотации генов и предсказания их регуляции. Пожалуй, первым примером нетривиального утверждения, сделанного в этой области, было установление близкого родства онкогена *v-sis* и фактора роста тромбоцитов [9]. С началом массового секвенирования геномов эта область прибрела особое значение. Ставший уже общим местом простой подсчет показывает, что экспериментальное исследование тысяч генов не представляется возможным [4]. Более того, поток геномов продолжает нарастать, так что уже и ручная компьютерная обработка всех новых геномов становится невозможной, и возникла необходимость в создании программ автоматической аннотации генов и метаболической реконструкции (GenQuiz, PathoLogic, Pedant) либо систем для облегчения ручного аннотирования (ERGO, Artemis, SEED, Genome-Explorer). Совершенно новые методы необходимы для анализа метагеномов – фрагментов, которые являются результатом секвенирования целых бактериальных сообществ. Так, в результате секвенирования бактерий из Саргассова моря были определены последовательности около 70 тысяч новых генов, образующих более 15 тысяч семейств [10]. Обсуждается проект секвенирования бактериального сообщества парижской канализации (Cloaca maxima, http://www.cns.fr/externe/English/corps_anglais-old.html).

В то же время секвенирование и анализ геномов показали, насколько наши представления о количестве и функциях генов и о физиологии даже таких хорошо изученных («модельных») организмов, как кишечная палочка, дрожжи и дрозофила, далеки от полноты. Доля генов с неизвестной функцией составляет обычно немногим более половины генома свободноживущей бактерии [11]. При этом большую часть таких генов (20–40% генома) составляют консервативные гены, присутствующие в большом количестве геномов (так называемые универсальные гипотетические гены) [12,13], в то время как около 15% генов в бактериальных геномах не имеют гомологов, т.е. специфичны для данного генома или узкого таксона [14]. В то же время экспериментальное изучение кишечной палочки позволяет описывать лишь 20–30 новых генов в год [15]. Ситуация с генами эукариот еще хуже. Недавно опубликованные результаты секвенирования генов 30 различных нематод позволили описать около ста тысяч генов и показали, что примерно половина из генов (более 4 тысяч белковых семейств) специфичны для этого типа, а около четверти генов каждого организма видоспецифичны [16]. В то же время оценка числа белок-кодирующих генов человека снизилась в результате анализа полного генома человека, последовательностей EST и геномов других позвоночных с 120–140 тысяч [17,18] до 30–35 тысяч [19–21] и далее до 20–25 тысяч [22,23].

С другой стороны, для более 40% биохимических реакций, катализируемых белковыми ферментами, не известно ни одного белка с данной активностью [24,25]; даже среди реакций центрального метаболизма, общих для подавляющего большинства организмов, доля не имеющих известных представителей составляет 10% [13]. Существенно больше ферментов, неизвестных в пределах крупных таксонов. Тем самым, возникают две задачи: частной метаболической реконструкции, т.е. идентификации пробелов в метаболической карте данного организма и поиске генов в геноме, продукты которых могли бы заполнить эти пробелы, и глобального пополнения метаболической карты, что сводится к установлению соответствий между множествами «отсутствующих» ферментов и универсальных гипотетических генов.

Компьютерный анализ играет при этом критическую роль [26]. Функции десятков генов были предсказаны путем сравнительно-геномного анализа и уже затем проверены экспериментально; примеры можно найти, в частности, в обзорах [13,27,28], а также в книге [29]. В то же время иногда высказывается опасение, что компьютерный анализ сводится лишь к переносу известных фактов на все новые объекты, а потому не может привести к существенно новым открытиям [А.В. Финкельштейн, частное сообщение]. Как показывают примеры в перечисленных обзорах, на самом деле многие предсказания, полученные путем анализа последовательностей, являются совершенно нетривиальными*.

Помимо анализа молекулярных данных (последовательности, концентрации мРНК и белков, белок-белковые взаимодействия), к этой области функционального анализа можно отнести и работы по моделированию клеточного метаболизма. При этом существуют два класса таких моделей: потоковые, в которых рассматриваются скорости реакций в стационарном состоянии, и динамические, в которых исследуется ответ системы на воздействия [43]. Потоковые модели используют вычислительно эффективный алгоритм линейного программирования и могут применяться для анализа метаболизма всего организма (лучше всего изучена с этой точки зрения кишечная палочка). В динамических моделях строятся системы дифференциальных уравнений и изучаются лишь отдельные метаболические пути, поскольку применимость таких моделей ограничена как вычислительными сложностями, так и необходимостью знать большое число экспериментально измеренных кинетических параметров, которые редко бывают доступны.

Потоковые модели успешно воспроизводят экспериментально измеренные параметры метаболизма при росте на обычных источниках углерода, таких как глюкоза. При росте на необычных источниках, например на глицероле, метаболизм моделируется плохо [44], однако после отбора популяции на рост в данной среде измеренные потоки приходят в соответствие с предсказанными. Аналогично, потоки в мутантных штаммах, вообще говоря, восстанавливаются плохо, однако их можно предсказать, если искать не оптимальное распределение потоков, а субоптимальное, ближайшее к первоначальному [45]. Сами же потоки в мутантах приходят к оптимальному распределению в течение нескольких сотен поколений [46]. Тем самым, похоже, что потоковые модели адекватно воспроизводят стационарные состояния, в том числе

* В дополнение приведем два случая из нашей практики.

Анализ регуляторных областей бактериальных генов биосинтеза рибофлавина и тиамина позволил предсказать механизм регуляции, основанный на образовании альтернативных вторичных структур РНК [30–32], причем в различных таксонах эта регуляция осуществляется через преждевременную терминацию транскрипции либо через ингибирование инициации трансляции. Сопоставление с опубликованными экспериментальными данными, из которых следовало отсутствие потенциального белка-регулятора, позволило предположить, что регуляция зависит от непосредственного связывания малых молекул с РНК. Чуть позднее группой Брекера были опубликованы независимые экспериментальные данные, подтвердившие эти предсказания [33,34], а регуляторные структуры получили название РНК-переключателей («riboswitches»). Еще в одной работе наши предсказания были проверены непосредственно [35]. Тиаминовые РНК-переключатели были найдены группой Брекера и в геномах эукариот [36], и этот результат также получил экспериментальное подтверждение – оказалось, в геномах растений тиаминовый РНК-переключатель регулирует сплайсинг [37]. Для другой разновидности РНК-переключателей, S-боксов, нами было показано, что они являются предковой системой регуляции синтеза метионина у бактерий [38]. Тем самым, есть основания думать, что РНК-переключатели являются древнейшей регуляторной системой: они встречаются у бактерий, архей и эукариот, регулируют транскрипцию, трансляцию и сплайсинг и не зависят от каких-либо посредников [39,40]. При этом одну из основных ролей в открытии РНК-переключателей сыграл компьютерный анализ регуляторных последовательностей, опиравшийся лишь на косвенные экспериментальные данные.

При анализе регуляторных систем гомеостаза цинка в геномах бактерий было показано, что среди генов, регулируемых цинк-зависимыми репрессорами, часто встречаются гомологи рибосомных белков, включающих в себя структурный элемент «цинковая лента», при этом в регулируемых белках этот мотив разрушен [41]. На основании этого наблюдения была построена модель, согласно которой эти белки экспрессируются в условиях цинкового голодания и встраиваются в состав небольшой фракции рибосом, замещая основные цинксодержащие рибосомные белки и тем самым высвобождая часть связанного цинка для использования в других важных белках, таких как протеазы и ДНК-полимеразы. В дальнейшем предсказания этой модели были подтверждены в эксперименте [42]. И здесь надо отметить, что предсказанный эффект был совершенно неожиданным и никак не следовал из заранее известных экспериментальных данных, несмотря на долгую историю изучения структуры и функции рибосом.

и такие, к которым приходит клетка после изменения среды или структуры метаболической карты. В отличие от потоковых, кинетические модели относительно удовлетворительно предсказывают рост мутантов на различных средах [43,47], однако, как уже упоминалось выше, моделированию поддаются лишь отдельные метаболические пути.

Наконец, третья область – это, если можно так сказать, теоретическая биоинформатика. Сюда относятся работы по молекулярной эволюции, эволюции геномов и т.п., а также разного рода глобальные статистические исследования: структуры геномов, статистики экзон-интронной структуры, распределения размеров белковых семейств и типов укладки белков, свойств метаболических и регуляторных сетей, белок-белковых взаимодействий и т.п. Отличительной чертой этих работ является то, что они не направлены на решение конкретных задач, а описывают либо историю геномов, либо самые общие их свойства.

Использование белковых последовательностей в эволюционных исследованиях было предложено Цукеркандлем и Полингом еще в 1965 г. [48]. Сейчас исследования по молекулярной эволюции позволяют существенно дополнить, а иногда пересмотреть существующие таксономические представления. Например, было показано, что такие события, как приспособление к эхолокации у летучих мышей [49] и появление (точнее, восстановление) крыльев у палочников [50], происходили неоднократно в истории соответствующих отрядов.

Анализ полных геномов позволил оценить частоту горизонтального переноса генов в бактериальных геномах (минимум 15% семейств генов испытали события горизонтального переноса внутри семейства) [51]; описать горизонтальный перенос от архей к термофильным бактериям (*Aquifex aeolicus* [52], *Thermotoga maritima* [53]), от бактерий к археям (до трети и как минимум одна шестая генов археи *Methanosarcina mazei* имеют бактериальное происхождение [54]), между фотосинтезирующими бактериями [55] и между внутриклеточными паразитами [56], и даже от бактерий к эукариотам (в дополнение к генам митохондриального происхождения) [57] и от эукариот к бактериям [58]; доказать события полногеномной дупликации в истории дрожжей рода *Saccharomyces* [59] и рыб [60].

Полногеномный анализ позволил по-новому взглянуть и на относительную важность процессов, происходящих в клетке. Так, было показано, что минимум треть, а по современным представлениям более половины генов млекопитающих подвержены альтернативному сплайсингу [61,62]; существенная доля альтернативных изоформ уничтожается механизмом проверки целостности открытой рамки считывания (nonsense-mediated decay) [63], более половины альтернативно сплайсируемых пар генов человека и мыши имеют геном-специфичные изоформы [64,65]; минимум треть генов человека регулируются микроРНК [66]. Следует отметить, что исследования регуляторных микроРНК – это, пожалуй, первый пример полноценной интеграции экспериментальных и вычислительных подходов. В лучших работах в этой области (например, [67]) уже невозможно разделить эксперимент и компьютерную обработку результатов – или предсказание и экспериментальную проверку, – поскольку оба подхода полноценно комбинируются с самого начала. Результаты получаются весьма впечатляющими: от открытия совершенно нового механизма регуляции экспрессии генов высших эукариот до осознания того, что он играет не меньшую роль, чем регуляция транскрипции и альтернативный сплайсинг, прошло всего несколько лет.

Наконец, к этой области можно отнести работы, в которых описываются общеструктурные свойства геномов, протеомов, регуляторных и белок-белковых взаимодействий. Изучается распределение количества генов в мультигенных семействах, доменов в структурных семействах белков и т.п. [68–74]. В последнее время большую популярность приобрели работы по изучению свойств различных графов, описывающих белок-белковые и регуляторные взаимодействия, метаболические пути и т.п.. В этих работах описываются фрактальные свойства этих графов (так называемая гипотеза «малого мира») [75–78], выделяются структурные модули [79–83] и мотивы [84–86] и т.п.

Разумеется, границы между описанными выше областями весьма условны, и возможны другие деления. Так, в [3] выделяются проблемы сбора и хранения информации, разработка методов и алгоритмов анализа, применение этих методов. Модный термин «системная биология» объединяет исследования в масштабах целого генома или организма: анализ данных по экспрессии, метаболическое моделирование, исследование различных графов, описывающих метаболические пути, регуляторные сети, ДНК-белковые и белок-белковые взаимодействиям. Независимо от деления, велико взаимное влияние и взаимопроникновение работ в различных областях. Разработка методов анализа нуклеотидных и аминокислотных последовательностей существенно опирается на обратную связь с экспериментом. С другой стороны, теоретические исследования генома часто приводят к

разработке новых методов анализа, которые в дальнейшем применяются в конкретных исследованиях.

Например, одним из основных способов анализа баз данных нуклеотидных и белковых последовательностей является поиск наиболее сходных последовательностей. Формально эта задача решается алгоритмом динамического программирования, однако поскольку скорость работы этого алгоритма является произведением длин выравниваемых последовательностей, он не может широко применяться при современных объемах баз данных. Другая важная задача – оценка статистической значимости наблюдаемого сходства. Обе эти задачи решаются программами из семейства BLAST [87,88]. Разработка этих программ потребовала существенного продвижения не только в теории алгоритмов и теории вероятности, но и в понимании того, какие виды сходства последовательностей представляют биологический интерес.

Изучение нуклеотидного состава последовательностей ДНК позволило создать методы для определения участков начала (oriR) и конца (terR) репликации бактерий [89–92], а обнаруженное неравномерное распределение ряда олигонуклеотидов на бактериальной хромосоме [93] было впоследствии связано с заданием направления для таких белков, как участвующая в репарации хеликаза RecBCD [94] и транслоказа FtsK [95].

Анализ взаимного расположения генов на хромосомах [96] и детальное изучение корреляции между степенью позиционной сцепленности генов и положением кодируемых ими ферментов на метаболической карте [97] позволили разработать методы для предсказания функции генов на основе анализа позиционной информации [98].

Даже такая на первый взгляд абстрактная область как молекулярная эволюция имеет неожиданные экспериментальные и практические приложения. Так, анализ эволюции штаммов ВИЧ имел существенное значение при рассмотрении уголовного дела американского дантиста, заразившего СПИДом ряд пациентов [99]. Реконструкция предковых белков позволяет делать заключения об условиях, в которых существовали предковые организмы. Так, было показано, что фактор элонгации трансляции EF-Tu из общего предка бактерий имеет температурный оптимум 55–65°C и, тем самым, что этот предок был умеренным термофилом [100], а реконструкция зрительного пигмента родопсина из общего предка крокодилов и птиц – архозавра – показала, что он имеет максимальное поглощение в красной области спектра, что соответствует приспособлению к ночному зрению [101]; еще ряд примеров приведен в обзорах [102] и [103].

Хорошо известна задача реконструкции происхождения и истории современного человека на основе молекулярных данных [104,105]. Менее тривиальный пример – использование этих данных в качестве косвенных маркеров доисторических событий. Так, существование двух современных линий вши *Pediculus humanus*, разошедшихся более миллиона лет назад, послужило косвенным доказательством существования контактов между *Homo sapiens* и *Homo erectus* [106] (эту теория подтверждает также существование парных видов других паразитов [107]), расхождение линий головной (*P. humanus capitis*) и платяной (*P. humanus corporis*) вши позволяет датировать возникновение одежды, совпадающее по времени с радиацией современного человека из Африки [108], а эволюция вирусов герпеса использовалась для датировки возникновения современного полового поведения [109] (впрочем, по поводу последней работы есть некоторые сомнения, поскольку она была напечатана в номере, датированном первым апреля). Интересно также совпадение эпохи возникновения современного поливного земледелия и времени начала радиации малярийного плазмодия *Plasmodium falciparum* [110] и недавнее (менее 20 тысяч лет) происхождение чумной бациллы *Yersinia pestis* из относительно слабого патогена *Y. pseudotuberculosis* [111].

Несмотря на все эти достижения, до сих пор многие биологи испытывают недоверие к результатам компьютерного анализа и к биоинформатике в целом. Например, документ «Результаты дополнительного конкурса по программе фундаментальных исследований Президиума РАН «Молекулярная и клеточная биология»» (http://www.molbiol.edu.ru/data/grant04r.doc) начинается словами: «Отношение к проектам по биоинформатике у большинства членов комиссии было неоднозначным и осторожным», – хотя в конечном счете из семнадцати проектов, поддержанных в 2004 году по разделу «Фундаментальные исследования», четыре – это именно биоинформатические проекты (из них

три заняли места в начале списка, упорядоченного по убыванию рейтинга заявки, еще один лидер, также биоинформатический проект, не был поддержан по формальным причинам).

В значительной степени это, по-видимому, обусловлено психологическими причинами: используется чужой язык, из-за чего непонятна система аргументации и нет критериев оценки достоверности результатов; даже близкие понятия часто имеют существенно различное наполнение (например, статистическая значимость в 1% в экспериментальной работе вполне приемлема, тогда как значимость 0,01 при поиске в базе данных в очень многих случаях не значит практически ничего); многие экспериментаторы слабо владеют современным компьютерным инструментарием и потому плохо представляют, какого рода проблемы поддаются такого рода анализу*; с другой стороны, многие утверждения становятся очевидными после того, как они сделаны (и потому непонятно, в чем, собственно говоря, состоит достижение)**; наконец, есть ощущение тривиальности значительной части работ, усугубленное плодовитостью ряда ведущих биоинформатических групп и исследователей. Эти проблемы усугубляются тем, что, как всегда бывает при становлении новой области, действительно публикуется заметное число спекулятивных, тривиальных, да и попросту неверных работ по биоинформатике.

Впрочем, нет оснований полагать, что ситуация в биоинформатике сильно отличается в этом смысле от ситуации в экспериментальной молекулярной биологии. В большинстве случаев, экспериментатор наблюдает не сам молекулярный механизм, а лишь его косвенные проявления (грубо говоря, полоски на геле – это не сами молекулы)***. Поэтому, как правило, сам по себе единичный эксперимент не дает ответа на поставленный вопрос, а имеет смысл лишь при правильно поставленных контролях, в контексте других экспериментов, использующих другую технику и т.п. Даже при технически безупречных данных все равно остается проблема правильного истолкования сделанных наблюдений, что зависит от теории, в рамках которой производится интерпретация и где нередки разногласия. Свидетельством этому служат, например, обсуждения в разделах писем и технических комментариев в таких журналах, как «Nature» и «Science», не говоря уже о регулярно появляющихся письмах авторов с отзывами или значительными поправками к уже опубликованным статьям. Уровень информационного шума в менее престижных журналах, по-видимому, выше, хотя прямая полемика там встречается реже просто в силу иных редакционных традиций: считается, что неверные результаты не находят подтверждения и постепенно сами собой исчезают из поля зрения****.

По-видимому, таков же естественный путь развития в любой науке: важные неверные утверждения опровергаются, а неважные – забываются. В этом смысле поучительно сравнение с такими полярными областями, как матема-

* В качестве анекдотического примера приведем использование редактора Microsoft Word для поиска сайтов связывания регуляторов транскрипции [112], впрочем, следует признать, что даже такой анализ позволил получить результаты, которые послужили поводом для совершенно разумной экспериментальной работы.

** «Ты вот полазил бы на карачках под паровозом, а то велика штука – перышком чирикать! » (Л. Кассиль, «Кондуит и Швамбрания»)

*** Вот пример того, как непоправимо неправильная первоначальная интерпретация привела к отзыву статьи: «Заключения в нашей статье … основывались на интерпретации плоского пилообразного паттерна на кривых растяжения как постепенного распутывания компактных филаментов MukBEF/ДНК. Однако последующие эксперименты, проведенные после публикации статьи, показали, что пилообразный паттерн соответствует расхождению двух цепей ДНК» (в оригинале: «The conclusions in our paper … were based on the interpretation of a flat sawtooth pattern in the force-extension curves as a progressive unraveling of compact MukBEF/DNA filaments. However, subsequent experiments done after the paper apperaed suggested that the sawtooth pattern corresponds to the unzipping of the two strands of DNA») [113].

**** Приведем лишь один пример [А.Б. Рахманинова, частное собщение]. Запись DSDX_ECOLI базы данных SwissProt, содержащей курируемые вручную сведения о белковых последовательностях, содержит следующее предостережение: «Открытая рамка считывания *dsdC* была первоначально (Ref.3) отнесена к неправильной цепи и описана как активатор D-серин деаминазы. Затем она была повторно секвенирована в Ref.2, где по-прежнему считалось, что это *dsdC*, хотя ей была приписана функция транспортера D-серина. Наконец, в Ref.1 было показано, что *dsdC* – это другой ген, а эта последовательность должна называться *dsdX*. Следует отметить также, что С-концевая часть *dsdX* (начиная с позиции 338) также была секвенирована (Ref.6 и Ref.7) и выделена как отдельная рамка считывания (не беспокойтесь, нам тоже было трудно понять, что случилось!)» (в оригинале: «An ORF called *dsdC* was originally (Ref.3) assigned to the wrong DNA strand and thought to be a D-serine deaminase activator, it was then resequenced by Ref.2 and still thought to be «*dsdC*», but this time to function as a D-serine permease. It is Ref.1 that showed that *dsdC* is another gene and that this sequence should be called *dsdX*. It should also be noted that the C-terminal part of *dsdX* (from 338 onward) was also sequenced (Ref.6 and Ref.7) and was thought to be a separate ORF (don't worry, we also had difficulties understanding what happened!)»). Следует отметить, что ссылки, обозначенные в этом комментарии Ref.1, Ref.3, Ref.6 – это работы одной группы экспериментаторов, опубликованные в одном и том же, вполне авторитетном, журнале «Journal of Bacteriology».

тика и сравнительная лингвистика. В явной форме социальная составляющая доказательства обсуждалась, в частности, замечательными математиками Ю.И. Маниным: «Доказательство становится таковым только в результате социального акта «принятия доказательства». Это относится к математике в той же мере, что и к физике, лингвистике или биологии. Эволюция признанных критериев доказательности – почти не исследованная тема в истории науки» [114] и В.А. Успенским: «Откуда же в математике берется убеждение, что доказанные теоремы, доказательства которых он так никогда и не узнает, действительно являются доказанными, т.е. располагают доказательствами? Видимо, такое убеждение основано не на чем ином, как на доверии ... Но если современное доказательство основано на доверии к авторитету, то в чем же его принципиальное отличие от древнеегипетского? Ответ на этот непростой вопрос заключается, возможно, в том, что доказательства постепенно переходят из разряда явлений индивидуального опыта в разряд явлений опыта коллективного» [115].

Индийские математики считали убедительным геометрическим доказательством чертеж с подписью «Смотри!» [115]; кажется, маркизу де Лопиталю (XVII век) приписывается высказывание: «Даю Вам честное слово дворянина, что эта теорема верна» ; а на Первом международном химическом конгрессе в Карлсруэ (1860 г.) научные вопросы решались голосованием. Ю.И. Манин проводит различие между индуктивными естественными науками, которые опираются на обширные исходные данные и потому могут позволить себе не очень жесткие «правила вывода», и дедуктивными науками, к которым относится математика, и которые предъявляют повышенные требования к соблюдению «правил гигиены в длинных выводах». Однако длина математических доказательств часто бывает такова, что человек в принципе не в состоянии проверить их целиком: «Отсутствие ошибок в математической работе (если они не обнаружены), как и в других естественных науках, часто устанавливается по косвенным данным: имеет значение соответствие с общими ожиданиями, использование аналогичных аргументов в других работах, разглядывание «под микроскопом» отдельных участков доказательства, даже репутация автора; словом, воспроизводимость в широком смысле слова» [114]. В последнее время такой проверке было подвергнуто доказательство теоремы Ферма, предложенное Э. Вайлсом, причем в первом варианте доказательства был найден пробел, заполненный через год в результате дополнительной напряженной работы [116]. Сейчас столь же внимательно проверяется опубликованное в 2003 году Г. Перельманом доказательство гипотезы Пуанкаре*. Особое место занимают доказательства, полученные при помощи компьютера, например, доказательство теоремы четырех красок, принадлежащее К. Аппелю и В. Хакену [118], – строго говоря, оно может быть перепроверено только с помощью независимого воспроизведения программы, желательно на другом языке и для другой операционной системы. Тем самым, тут уже имеется прямая аналогия с воспроизводимостью в смысле экспериментальных наук (впрочем, сейчас сомнению подвергается как раз «домашинная» часть доказательства [115], и ее также проверяют при помощи компьютера [119]).

Аналогична, хотя существенно более зыбка, ситуация в сравнительно-исторической лингвистике (компаративистике): «реконструкция фонетики праязыка недоступна непосредственной проверке, поэтому при ее оценке применяются критериии, использующиеся для оценки научной теории вообще – экономность описания, объяснительная сила и т.п.» [120]. В процитированном учебнике С.А. Бурлак и С.А. Старостина приведены большое количество практических рекомендаций («правил гигиены»), иногда практически дословно совпадающих с аналогичными рекомендациями в исследованиях по сравнительной геномике (см. ниже). В то же время в большинстве случаев спорные вопросы решаются голосованием. Редкой попыткой применить в компаративистике методологию естественных наук является работа Е. Хелимского [121]. В ней критика некоторой фонологической реконструкции была проведена следующим образом. В исходной работе приводились примеры на предлагаемые фонологические соотношения между прауральскими языками, однако этимологические ряды объединяли слова с достаточно далеким смыслом (хотя в принципе подобного рода семантические сдвиги возможны). Количество рядов также было невелико. Для опровержения был поставлен «контрольный опыт»: были заданы произвольные, хотя и типологически правдоподобные, фонологические соотношения и были построены подтверждающие их этимологические ряды,

* В этом контексте крайне поучительна цитата из интервью с лауреатом премии Абеля за 2004 год И. Зингером. Вопрос: «Что Вам известно о состоянии гипотезы Пуанкаре?», – ответ: «Пока все сходится, насколько я слышал от участников семинара Лотта в Университете Мичигана и семинара Тиана в Принстоне. Хотя пока никто не готов ручаться за все детали, похоже, доказательство Перельмана будет подтверждено» (В оригинале: «What do you know about the status of the Poincaré conjecture? – To date, everything is working out as Perelman says. So I learn from Lott's seminar at the University of Michigan and Tian's seminar at Princeton. Although no one vouches for the final details, it appears that Perelman's proof will be validated») [117].

причем количество и качество этих рядов было таким же, как в критикуемой работе. Тем самым, была продемонстрирована недостаточная доказательность исходных построений.

Биоинформатика занимает в каком-то смысле промежуточное положение. Многие предсказания поддаются непосредственной экспериментальной проверке. В то же время работы в области молекулярной эволюции принципиально непроверяемы, и здесь применяются специальные приемы, как статистические (бутстреппинг), так и биологические (построение эволюционных деревьев по многим различным семействам белков). Тем самым, производится контроль статистической значимости и воспроизводимости результатов. В работах по моделированию проверяются следствия из построенных моделей. При этом проводится сравнение с доступными экспериментальными данными, не использовавшимися при подгонке параметров в самой модели [47]. Наконец, при глобальном статистическом анализе осмысленность результатов критически зависит от чистоты исходных выборок (см., например, полемику [122–124]).

Для функциональной аннотации по сходству требуется:

тщательный контроль статистической значимости уровня сходства последовательностей (в частности, особого внимания требует сходство областей с аномальным распределением аминокислот, таких как неглобулярные домены и трансмембранные сегменты);

выяснение всего геномного и эволюционного контекста (в особенности различный подход к ортологичным и паралогичным генам);

осторожное отношение к возможности предсказания не только общей функции белка, но и его клеточной роли (например, специфичности ферментов и транспортеров, где до 30% существующих аннотаций могут быть ошибочными [125]);

контроль за доменной структурой белков (сходства отдельных доменов недостаточно для сохранения функции);

проверка исходной информации о функции родственного белка: по возможности она должна быть установлена экспериментально (в противном случае может происходить размножение ошибочных компьютерных аннотаций [126]);

при анализе далеких гомологов – проверка существования остатков, консервативных во всем белковом семействе (функциональной подписи, соответствующей каталитическому центру), поскольку «диффузное» сходство в отдельных позициях в такой ситуации не является значимым.

Сравнительно-геномный анализ должен опираться на множественные наблюдения, поскольку каждое конкретное наблюдение обычно является слабым. При этом относительно достоверными могут считаться предсказания, полученные различными методами (анализ позиционных кластеров, доменных перестроек, филогенетических паттернов, регуляторных сигналов), с использование разных типов данных (включая данные о экспрессии генов, белок-белковых и белок-ДНКовых взаимодействиях и т.п.), а также на многих различных геномах. В то же время пока не существует никаких способов формальной оценки статистической значимости таких предсказаний, и в этом смысле интерпретация сделанных наблюдений является таким же искусством, как интерпретация экспериментальных результатов*.

Публикуемые во всех областях биоинформатики результаты часто становятся предметом полемики, иногда весьма жесткой (см., например, дискуссии о датировке эволюционных деревьев [133–136] и о доле генов бактериального происхождения в геноме человека [137–140]. Даже конкретные предсказания, для которых, вообще говоря, существует возможность экспериментальной проверки, часто подвергаются де-

* Опять приведем пример из нашей практики. Ортологи гена кишечной палочки *ybaD* существуют в большинстве бактериальных геномов (COG1327). На основании наличия в последовательности белка YbaD потенциального ДНК или РНК-связывающего мотива и АТФ-связывающего регуляторного домена был сделан вывод о том, что он может кодировать фактор транскрипции [127,128], а наблюдение, что ген *ybaD* во многих геномах расположен в непосредственной близости от генов биосинтеза рибофлавина, позволило предположить для него функцию регулятора рибофлавиновых генов и предложить название *ribX* [129,130]. В то же время сравнение генов рибонуклеотид-редуктаз (ферментов, катализирующих превращение рибонуклеотидов в деоксирибонуклеотиды и, тем самым, поставляющих исходный материал для репликации ДНК) позволило нам выделить потенциальный регуляторный сигнал, присутствующий во многих геномах. Филогенетический паттерн этого сигнала, т.е. набор геномов, в которых он встретился, в точности совпал с филогенетическим паттерном гена *ybaD*. Более того, в ряде геномов ген *ybaD* расположен рядом с генами рибонуклеотид-редуктаз и генами системы репликации (*dnaB, dnaI, polA*). Наконец, в некоторых геномах консервативные регуляторные сигналы были обнаружены перед генами, продукты которых осуществляют ре-утилизацию деоксирибонуклеотидов и участвуют в репликации (ДНК-лигаза LigA, ДНК-хеликаза II DR1775, ДНК-топоизомераза I TopA, инициатор репликации DnaA и др.). Тем самым, основная функция YbaD – регуляция рибонуклеотид-редуктаз, тесно связанная с репликацией. В соответствии с этим, для данного белка было предложено название NrdR [131], а уже когда исследование было завершено и результаты направлены в печать, появилось экспериментальное подтверждение этого предсказания в стрептомицетах [132]. При этом потенциальных сигналов связывания YbaD/NrdR перед рибофлавиновыми генами не было найдено ни в одном из геномов, что позволяет отклонить гипотезу о роли NrdR в регуляции биосинтеза рибофлавина и оставляет открытым вопрос о причинах позиционной сцепленности *nrdR* и рибофлавиновых генов.

тальному обсуждению [141], иногда на грани фола [142,143].

Следует отметить, что, несмотря на заметный уровень шума в геномных аннотациях [125,144], появление новых геномов, новых методов анализа и новых экспериментальных данных приводит к тому, что качество аннотаций, особенно полученных в результате аккуратного ручного анализа, постоянно улучшается [145]. Проблемой остается, однако, внесение своевременных изменений в базы данных.

Как и при интерпретации экспериментальных данных, иногда ошибки предсказания неминуемы, поскольку вызываются принципиальной неполнотой доступных знаний. Так, предсказание, что ген *MJ0539* архебактерии *Methanococcus jannaschii* кодирует цистеинил-тРНК-синтетазу [146] опиралось на считавшееся установленным распределение аминоацил-тРНК-синтетаз с различной специфичностью по структурным классам I и II. Впоследствии оказалось, что это – уникальный пример лизил-тРНК-синтетазы, принадлежащей к классу I [147]. Роль цистеинил-тРНК-синтетазы, тем самым, оставалась свободной то тех пор, пока не было показано, что ее выполняет пролил-тРНК-синтетаза, обладающая двойной активностью [148]. Другой предложенный кандидат, *MJ1477* [149], по-видимому, является внеклеточной гидролазой полисахаридов [iyer].

Последний случай представляет особый интерес, поскольку иллюстрирует не только существенные противоречия между опубликованными практически одновременно экспериментальными работами [148] и [149], но и между экспериментом и компьютерной аннотацией. Еще ряд аналогичных примеров приведен в [150]. В рассмотренных в этой работе ситуациях экспериментальные данные о белках, опубликованные в престижных журналах, таких как «Nature» и «Science», совершенно противоречат результатам компьютерного анализа, которые являются весьма убедительными. Авторы обсуждают возможные причины экспериментальных ошибок.

Итак, «биоинформатика в узком смысле» является важной и динамично развивающейся областью молекулярной биологии. Помимо задач, мотивировка которых в значительной степени состоит в облегчении экспериментальной работы, определении наиболее обещающих направлений для эксперимента и т.п. – таковыми являются, в частности, задачи функциональной аннотации белков и предсказания регуляторных сигналов, – она имеет свои собственные задачи, такие как глобальное описание общих свойств геномов и реконструкция эволюции по молекулярным данным. По мере взросления, биоинформатика вырабатывает критерии доказательности, позволяющие оценивать ее результаты как сами по себе, так и в контексте других методов исследования биологических объектов, а биоинформатические методы входят в необходимый арсенал средств, доступных молекулярным биологам.

Автор выражает благодарность И.И. Артамоновой, М.Ю. Гальперину, Е.В. Кунину, В.Ю. Макееву, А.А. Миронову, А. Остерману, А.Б. Рахманиновой за многочисленные обсуждения и некоторые примеры, использованные в этом тексте. Разумеется, вся ответственность за интерпретацию этих примеров и другие сделанные утверждения лежит на авторе. Обсуждаемые результаты были получены в ходе совместной работы с А.Г. Витрещаком, А.А. Мироновым, Е.М. Паниной и Д.А. Родионовым. Эти исследования были частично поддержаны Медицинским институтом Ховарда Хьюза, Российским фондом фундаментальных исследований, Российской академией наук (в рамках программы «Молекулярная и клеточная биология») и Фондом поддержки отечественной науки.

СПИСОК ЛИТЕРАТУРЫ

1. *Аксенов С.И., Гуляев Б.А., Шайтан К.В., Чернавский Д.С.* // Биоинформатика. Программы спецкурсов. Московский государственный университет им. М.В. Ломоносова, биологический факультет, специальность 01.22 – биофизика. М.: МГУ им. М.В. Ломоносова, 2000. С. 9–11.
2. *Гельфанд М.С., Миронов А.А.* Вычислительная биология на рубеже десятилетий // Молекуляр. биология. 1999. Т. 33. С. 969–984.
3. *Luscombe N.M., Greenbaum D., Gerstein M.* What is bioinformatics? A proposed definition and overview of the field // Method. Inform. Med. 2001. V. 40. P. 346–358.
4. *Гельфанд М.С., Любецкий В.А.* Биоинформатика: от экспермента к компьютерному анализу и снова к эксперименту // Вестн. РАН. 2003. Т. 73. С. 987–994.
5. *Гельфанд М.С.* Вычислительная геномика: от пробирки к компьютеру и обратно // Biomediale: Современное общество и геномная культура / Ред. Д. Булатова. Государственный центр современного искусства, Калининграский филиал, 2004. С. 28–39. (Gelfand M.S. Computational genomics: from the wet lab to computer and back // Biomediale: Contemporary Society and Genomic Culture. National Centre for Contemporary Art, Kaliningrad brach, 2004. P. 28–39.)
6. *Чайлахян Л.М.* Что является предметом науки «биоинформатика» // Биофизика. 2005. Т. 50, вып. 1. С. 152–155.
7. *Шеннон К.* Бандвагон. Работы по теории информации и кибернетике. Пер. с англ. М.: Изд-во иностранной литературы, 1963. С. 667–668.
8. *Степанов Ю.С.* Семиотика. Лингвистический энциклопедический словарь. М.: Сов. энциклопедия, 1990. С. 440–442.

9. *Doolittle R.F., Hunkapiller M.W., Hood L.E. et al.* Simian sarcoma virus oncogene, *v-sis*, is derived from the gene (or genes) encoding a platelet-derived growth factor // Science. 1983. V. 221. P. 275–277.

10. *Venter J.C., Remington K., Heidelberg J.F., Halpern A.L., Rusch D., Eisen J.A., Wu D., Paulsen I., Nelson K.E., Nelson W., Fouts D.E., Levy S., Knap A.H., Lomas M.W., Nealson K., White O., Peterson J., Hoffman J., Parsons R., Baden-Tillson H., Pfannkoch C., Rogers Y.H., Smith H.O.* Environmental genome shotgun sequencing of the Sargasso Sea // Science. 2004. V. 304. P. 66–74.

11. *Bork P.* Powers and pitfalls in sequence analysis: the 70% hurdle // Genome Res. 2000. V. 10. P. 398–400.

12. *Galperin M.Y.* Conserved 'hypothetical' proteins: new hints and new puzzles // Comp. Funct. Genom. 2001. V. 2. P. 14–18.

13. *Osterman A., Overbeek R.* Missing genes in metabolic pathways: a comparative genomics approach // Curr. Opin. Chem. Biol. 2003. V. 7. P. 238–251.

14. *Siew N., Fischer D.* Analysis of singleton ORFans in fully sequenced microbial genomes // Proteins. 2003. V. 53. P. 241–251.

15. *Thomas G.H.* Completing the *E. coli* proteome: a database of gene products characterised since the completion of the genome sequence // Bioinformatics. 1999. V. 15. P. 860–861.

16. *Parkinson J., Mitreva M., Whitton C., Thomson M., Daub J., Martin J., Schmid R., Hall N., Barrell B., Waterston R.H., McCarter J.P., Blaxter M.L.* A transcriptomic analysis of the phylum Nematoda // Nat. Genet. 2004. V. 36. P. 1259–1267.

17. *Liang F., Holt I., Pertea G., Karamycheva S., Salzberg S.L., Quackenbush J.* Gene index analysis of the human genome estimates approximately 120,000 genes // Nat. Genet. 2000. V. 25. P. 239–240.

18. *Scott R.* The future in understanding the molecular basis of life // 11th Int. Genome Sequencing and Analysis Conference. Miami, 1999.

19. *Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C. et al.* Initial sequencing and analysis of the human genome // Nature. 2001. V. 409. P. 860–921.

20. *Ewing B., Green P.* Analysis of expressed sequence tags indicates 35,000 human genes // Nat. Genet. 2000. V. 25. P. 232–234.

21. *Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J. et al.* The sequence of the human genome. // Science. 2001. V. 291. P. 1304–1351.

22. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome // Nature. 2004. V. 431. P. 931–945.

23. *Southan C.* Has the yo-yo stopped? An assessment of human protein-coding gene number // Proteomics. 2004. V. 4. P. 1712–1726.

24. *Karp P.D.* Call for an enzyme genomics initiative // Genome Biol. 2004. V. 5. P. 401.

25. *Lespinet O., Labedan B.* Orphan Enzymes // Science. 2005. V. 307. P. 42.

26. *Galperin M.Y., Koonin E.V.* 'Conserved hypothetical' proteins: prioritization of targets for experimental study // Nucleic Acids Rec. 2004. V. 32. P. 5452–5463.

27. *Koonin E.V.* Comparative analysis of biopolymer sequences: Reflections on the validity of the methodology and the underlying general principles // Mathematical Methods of Analysis of Biopolymer Sequences / Ed. Gindikin S. (DIMACS Series in Discrete Mathematics and Theoretical Computer Science. V. 8). Providence, RI: American Mathematical Society, 1992. P. 63–74.

28. *Makarova K.S., Koonin E.V.* Comparative genomics of archae: how much have we learned in six years, and what's next // Genome Biol. 2003. V. 4. P. 115.

29. *Koonin E.V., Galperin M.Y.* Sequence–Evolition–Function. Computational Approaches in Comparative Genomics. Kluwer Academic Publishers, 2003.

30. *Gelfand M.S., Mironov A.A., Jomantas J., Kozlov Y.I., Perumov D.A.* A conserved RNA structure element involved in regulation of bacterial riboflavin synthesis genes // Trends Genet. 1999. V. 15. P. 439–442.

31. *Vitreschak A.A., Rodionov D.A., Mironov A.A., Gelfand M.S.* Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation // Nucleic Acids Res. 2002. V. 30. P. 3141–3151.

32. *Rodionov D.A., Vitreschak A.A., Mironov A.A., Gelfand M.S.* Computational analysis of thiamin regulation in bacteria: Possible mechanisms and new THI-element-regulated genes // J. Biol. Chem. 2002. V. 277. P. 48949–48959.

33. *Winkler W.C., Cohen-Chalamish S., Breaker R.R.* An mRNA structure that controls gene expression by binding FMN // Proc. Natl. Acad. Sci. USA. 2002. V. 99. P. 15908–15913.

34. *Winkler W., Nahvi A., Breaker R.R.* Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression // Nature. 2002. V. 419. P. 952–956.

35. *Mironov A.S., Gusarov I., Rafikov R., Lopez L.E., Shatalin K., Kreneva R.A., Perumov D.A., Nudler E.* Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria // Cell. 2002. V. 111. P. 747–756.

36. *Sudarsan N., Barrick J.E., Breaker R.R.* Metabolite-binding RNA domains are present in the genes of eukaryotes // RNA. 2003. V. 9. P. 644–647.

37. *Kubodera T., Watanabe M., Yoshiuchi K., Yamashita N., Nishimura A., Nakai S., Gomi K., Hanamoto H.* Thiamine-regulated gene expression of *Aspergillus oryzae thiA* requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR // FEBS Lett. 2003. V. 555. P. 516–520.

38. *Rodionov D.A., Vitreschak A.G., Mironov A.A., Gelfand M.S.* Comparative genomics of the regulation of methionine metabolism in Gram-positive bacteria // Nucleic Acids Res. 2004. V. 32. P. 3340–3353.

39. *Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S.* Riboswitches: the oldest mechanism for the regulation of gene expression? // Trends Genet. 2004. V. 20. P. 44–50.

40. *Mandal M., Breaker R.R.* Gene regulation by riboswitches // Nat. Rev. Mol. Cell. Biol. 2004. V. 5. P. 451–463.

41. *Panina E.M., Mironov A.A., Gelfand M.S.* Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins // Proc. Natl. Acad. Sci. USA. 2003. P. 100. V. 17. P. 9912–9917.

42. *Nanamiya H., Akanuma G., Natori Y., Murayama R., Kosono S., Kudo T., Kobayashi K., Ogasawara N., Park S.M., Ochi K., Kawamura F.* Zinc is a key factor in controlling alternation of two types of L31 protein in the *Bacillus subtilis* ribosome // Mol. Microbiol. 2004. V. 52. P. 273–283.

43. Metabolic Engineering in the Post Genomic Era /Eds. B.N. Kholodenko, H.V. Westerhoff. Wymondham (England): Horizon Biosciences, 2004.

44. *Ibarra R.U., Edwards J.S, Palsson B.O. Escherichia coli* K–12 undergoes adaptive evolution to achieve in silico predicted optimal growth // Nature. 2002. V. 420. P. 186–189.

45. *Segre D., Vitkup D., Church G.M.* Analysis of optimality in natural and perturbed metabolic networks // Proc. Natl. Acad. Sci. USA. 2002. V. 99. P. 15112–15117.

46. *Fong S.S., Palsson B.O.* Metabolic gene deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes // Nature Genet. 2004. V. 36. P. 1056–1058.

47. *Goltsov A.N., Lebedeva G.V., Lavrova A.I., Demin O.V.* Modeling of purine nucleotides biosynthesis in *E. coli* // 5th Int. Conf. on Computational Systems Biology ICSB'2004. Heidelberg, Germany, 2004. P. 95.

48. *Zuckerkandl E., Pauling L.* Molecules as documents of evolutionary history // J. Theor. Biol. 1965. V. 8, № 2. P. 357–366.

49. *Teeling E.C., Madsen O., Van den Bussche R.A., de Jong W.W., Stanhope M.J., Springer M.S.* Microbat paraphyly and the convergent evolution of a key innovation in Old World rhinolophoid microbats // Proc. Natl. Acad. Sci. USA. 2002. V. 99. P. 1431–1436.

50. *Whiting M.F., Bradler S., Maxwell T.* Loss and recovery of wings in stick insects // Nature. 2003. V. 421. P. 264–267.

51. *Novichkov P.S., Omelchenko M.V., Gelfand M.S., Mironov A.A., Wolf Y.I., Koonin E.V.* Genome-wide molecular clock and horizontal gene transfer in bacterial evolution // J. Bacteriol. 2004. V. 186. P. 6575–6585.

52. *Aravind L., Tatusov R.L., Wolf Y.I., Walker D.R., Koonin E.V.* Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles // Trends Genet. 1998. V. 14. P. 442–444.

53. *Nelson K.E., Clayton R.A., Gill S.R., Gwinn M.L., Dodson R.J., Haft D.H., Hickey E.K., Peterson J.D., Nelson W.C., Ketchum K.A., McDonald L., Utterback T.R., Malek J.A., Linher K.D., Garrett M.M., Stewart A.M., Cotton M.D., Pratt M.S., Phillips C.A., Richardson D., Heidelberg J., Sutton G.G., Fleischmann R.D., Eisen J.A., Fraser C.M. et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima* // Nature. 1999. V. 399. P. 323–329.

54. *Deppenmeier U., Johann A., Hartsch T., Merkl R., Schmitz R.A., Martinez-Arias R., Henne A., Wiezer A., Baumer S., Jacobi C., Bruggemann H., Lienard T., Christmann A., Bomeke M., Steckel S., Bhattacharyya A., Lykidis A., Overbeek R., Klenk H.P., Gunsalus R.P., Fritz H.J., Gottschalk G.* The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea // J. Mol. Microbiol. Biotechnol. 2002. V. 4. P. 453–461.

55. *Raymond J., Zhaxybayeva O., Gogarten J.P., Gerdes S.Y., Blankenship R.F.* Whole-genome analysis of photosynthetic prokaryotes // Science. 2002. V. 298. P. 1616–1620.

56. *Wolf Y.I., Aravind L., Koonin E.V.* Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange // Trends Genet. 1999. V. 15. P. 173–175.

57. *Koonin E.V., Aravind L.* Origin and evolution of eukaryotic apoptosis: the bacterial connection // Cell Death Differ. 2002. V. 9. P. 394–404.

58. *Ponting C.P., Aravind L., Schultz J., Bork P., Koonin E.V.* Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer // J. Mol. Biol. 1999. V. 289. P. 729–745.

59. *Kellis M., Birren B.W., Lander E.S.* Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae* // Nature. 2004. V. 428. P. 617–624.

60. *Van de Peer Y. Tetraodon* genome confirms *Takifugu* findings: most fish are ancient polyploids // Genome Biol. 2004. V. 5. P. 250.

61. *Mironov A.A., Fickett J.W., Gelfand M.S.* Frequent alternative splicing of human genes // Genome Res. 1999. V. 9. P. 1288–1293.

62. *Brett D., Hanke J., Lehmann G., Haase S., Delbruck S., Krueger S., Reich J., Bork P.* EST comparison indicates 38% of human mRNAs contain possible alternative splice forms // FEBS Lett. 2000. V. 474. P. 83–86.

63. *Lewis B.P., Green R.E., Brenner S.E.* Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans // Proc. Natl. Acad. Sci. USA. 2003. V. 100. P. 189–192.

64. *Нуртдинов Р.Н., Миронов А.А., Гельфанд М.С.* Консервативен ли альтернативный сплайсинг генов млекопитающих? // Биофизика. 2002. Т. 47, вып. 2. С. 197–203.

65. *Modrek B., Lee C.J.* Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss // Nat. Genet. 2003. V. 34. P. 177–180.

66. *Lewis B.P., Burge C.B., Bartel D.P.* Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets // Cell. 2005. V. 120. P. 15–20.

67. *Lim L.P., Lau N.C., Garrett-Engele P., Grimson A., Schelter J.M., Castle J., Bartel D.P., Linsley P.S., Johnson J.M.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs // Nature. 2005. V. 433. P. 769–773.

68. *Wolf Y.I., Grishin N.V., Koonin E.V.* Estimating the number of protein folds and families from complete genome data // J. Mol. Biol. 2000. V. 299. P. 897–905.

69. *Yanai I., Camacho C.J., DeLisi C.* Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification // Phys. Rev. Lett. 2000. V. 85. P. 2641–2644.

70. *Qian J., Luscombe N.M., Gerstein M.* Protein family and fold occurrence in genomes: power-law behaviour

and evolutionary model // J. Mol. Biol. 2001. V. 313. P. 673–681.

71. *Rzhetsky A., Gomez S.M.* Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome // Bioinformatics. 2001. V. 17. P. 988–996.

72. *Luscombe N.M., Qian J., Zhang Z., Johnson T., Gerstein M.* The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties // Genome Biol. 2002. V. 3. P. RESEARCH0040.

73. *Karev G.P., Wolf Y.I., Koonin E.V.* Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? // Bioinformatics. 2003. V. 19. P. 1889–1900.

74. *Unger R., Uliel S., Havlin S.* Scaling law in sizes of protein sequence families: from super-families to orphan genes // Proteins. 2003. V. 51. P. 569–576.

75. *Jeong H., Tombor B., Albert R., Oltvai Z.N., Barabasi A.L.* The large-scale organization of metabolic networks. // Nature. 2000. V. 407. P. 651–654.

76. *Wagner A., Fell D.A.* The small world inside large metabolic networks // Proc. R. Soc. Lond. B Biol. Sci. 2001. V. 268. P. 1803–1810.

77. *Yook S.H., Oltvai Z.N., Barabasi A.L.* Functional and topological characterization of protein interaction networks // Proteomics. 2004. V. 4. P. 928–942.

78. *Arita M.* The metabolic world of *Escherichia coli* is not small // Proc. Natl. Acad. Sci. USA. 2004. V. 101. P. 1543–1547.

79. *Ravasz E., Somera A.L., Mongru D.A., Oltvai Z.N., Barabasi A.L.* Hierarchical organization of modularity in metabolic networks // Science. 2002. V. 297. P. 1551–1555.

80. *Spirin V., Mirny L.A.* Protein complexes and functional modules in molecular networks // Proc. Natl. Acad. Sci. USA. 2003. V. 100. P. 12123–12128.

81. *Rives A.W, Galitski T.* Modular organization of cellular networks // Proc. Natl. Acad. Sci. USA. 2003. V. 100. P. 1128–1133.

82. *Wuchty S., Oltvai Z.N., Barabasi A.L.* Evolutionary conservation of motif constituents in the yeast protein interaction network // Nat. Genet. 2003. V. 35. P. 176–179.

83. *Han J.D., Bertin N., Hao T., Goldberg D.S., Berriz G.F., Zhang L.V., Dupuy D., Walhout A.J., Cusick M.E., Roth F.P., Vidal M.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network // Nature. 2004. 430. P. 88–93.

84. *Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U.* Network motifs: simple building blocks of complex networks // Science. 2002. V. 298. P. 824–827.

85. *Shen-Orr S.S., Milo R., Mangan S., Alon U.* Network motifs in the transcriptional regulation network of *Escherichia coli* // Nat. Genet. 2002. V. 31. P. 64–68.

86. *Yeger-Lotem E., Sattath S., Kashtan N., Itzkovitz S., Milo R., Pinter R.Y., Alon U., Margalit H.* Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction // Proc. Natl. Acad. Sci. USA. 2004. V. 101. P. 5934–5939.

87. *Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J.* Basic local alignment search tool // J. Mol. Biol. 1990. V. 215. P. 403–410.

88. *Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Res. 1997. V. 25. P. 3389–3402.

89. *Lobry J.R.* Asymmetric substitution patterns in the two DNA strands of bacteria // J. Mol. Biol. Evol. 1996. V. 13. P. 660–665.

90. *McLean M.J., Wolfe K.H., Devine K.M.* Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes // J. Mol. Evol. 1998. V. 47. P. 691–696.

91. *Grigoriev A.* Analyzing genomes with cumulative skew diagrams // Nucleic Acids Res. 1998. V. 26. P. 2286–2290.

92. *Mackiewicz P., Zakrzewska-Czerwinska J., Zawilak A., Dudek M.R., Cebrat S.* Where does bacterial replication start? Rules for predicting the *oriC* region // Nucl. Acids Res. 2004. V. 32. P. 3781–3791.

93. *Salzberg S.L., Salzberg A.J., Kerlavage A.R., Tomb J.F.* Skewed oligomers and origins of replication // Gene. 1998. V. 217. P. 57–67.

94. *Spies M., Bianco P.R., Dillingham M.S., Handa N., Baskin R.J., Kowalczykowski S.C.* A molecular throttle: the recombination hotspot chi controls DNA translocation by the RecBCD helicase // Cell. 2003. V. 114. P. 647–654.

95. *Pease P.J., Levy O., Cost G.J., Gore J., Ptacin J.L., Sherratt D., Bustamante C., Cozzarelli N.R.* Sequence-directed DNA translocation by purified FtsK // Science. V. 307. P. 586–590.

96. *Overbeek R., Fonstein M., D'Souza M., Pusch G.D., Maltsev N.* The use of gene clusters to infer functional clustering // Proc. Natl. Acad. Sci. USA. 1999. V. 96. P. 2896–2901.

97. *von Mering C., Zdobnov E.M., Tsoka S., Ciccarelli F.D., Pereira-Leal J.B., Ouzounis C.A., Bork P.* Genome evolution reveals biochemical networks and functional modules // Proc. Natl. Acad. Sci. USA. 2003. V. 100. P. 15428–15433.

98. *von Mering C., Jensen L.J., Snel B., Hooper S.D., Krupp M., Foglierini M., Jouffre N., Huynen M.A., Bork P.* STRING: known and predicted protein-protein associations, integrated and transferred across organisms // Nucleic Acids Res. 2005. V. 33. P. D433-D437.

99. *Ou C.Y., Ciesielski C.A., Myers G., Bandea C.I., Luo C.C., Korber B.T., Mullins J.I., Schochetman G., Berkelman R.L., Economou A.N. et al.* Molecular epidemiology of HIV transmission in a dental practice // Science. 1992. V. 256. P. 1165–1171.

100. *Gaucher E.A., Thomson J.M., Burgan M.F., Benner S.A.* Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins // Nature. 2003. V. 425. P. 285–288.

101. *Chang B.S.W., Jönsson K., Kazmi M.A., Donoghue M.J., Sakmar T.P.* Recreating a functional ancestral visual pigment // Mol. Biol. Evol. 2002. V. 19. P. 1483–1489.

102. *Chang B.S.W., Donoghue M.J.* Recreating ancestral proteins // Trends Ecol. Evol. 2000. V. 15. P. 109–114.

103. *Brookfield J.F.* The ecology of the genome – mobile DNA elements and their hosts // Nat. Rev. Genet. 2005. V. 6. P. 128–136.

104. *Laan M., Paabo S.* Demographic history and linkage disequilibrium in human populations // Nat. Genet. 1997. V. 17. P. 435–438.

105. *Боринская С.А., Хуснутдинова Э.К.* Этногеномика: история с географией // Человек. 2002. № 1. С. 19–30.

106. *Reed D.L., Smith V.S., Hammond S.L., Rogers A.R., Clayton D.H.* Genetic analysis of lice supports direct contact between modern and archaic humans // PloS Biology. 2004. V. 2. P. 1972–1983.

107. *Ashford R.W.* Parasites as indicators of human biology and evolution // J. Med. Microbiol. 2000. V. 49. P. 771–772.

108. *Kittler R., Kayser M., Stoneking M.* Molecular evolution of *Pediculus humanus* and the origin of clothing // Curr. Biol. 2003. V. 13. P. 1414–1417.

109. *Gentry G.A., Lowe M., Alford G., Nevins R.* Sequence analyses of herpesviral enzymes suggest an ancient origin for human sexual behavior // Proc. Natl. Acad. Sci. USA. 1988. V. 85. P. 2658–2661.

110. *Joy D.A., Feng X., Mu J., Furuya T., Chotivanich K., Krettli A.U., Ho M., Wang A., White N.J., Suh E., Beerli P., Su X.Z.* Early origin and recent expansion of *Plasmodium falciparum* // Science. 2003. V. 300. P. 318–321.

111. *Achtman M., Morelli G., Zhu P., Wirth T., Diehl I., Kusecek B., Vogler A.J., Wagner D.M., Allender C.J., Easterday W.R., Chenal-Francisque V., Worsham P., Thomson N.R., Parkhill J., Lindler L.E., Carniel E., Keim P.* Microevolution and history of the plague bacillus, *Yersinia pestis* // Proc. Natl. Acad. Sci. USA. 2004. V. 101. P. 17837–17842.

112. *Horsburgh M.J., Ingham E., Foster S.J.* In Staphylococcus aureus, Fur is an interactive regulator with PerR, contributes to virulence, and is necessary for oxidative stress resistance through positive regulation of catalase and iron homeostasis // J. Bacteriol. 2001. V. 183. P. 468–475.

113. *Case R.B., Chang Y.-P., Smith S.B., Gore J., Cozzarelli N.R., Bustamante C.* Retraction // Science. 2005. V. 307. P. 1409.

114. *Манин Ю.И.* // Доказуемое и недоказуемое. М.: Советское радио, 1979.

115. *Успенский В.А.* Семь размышлений на темы философии математики // Закономерности развития современной математики. М.: Наука, 1987. С. 106–155.

116. *Singh S.* Fermat's Enigma: The Epic Quest to Solve the World's Greatest Mathematical Problem. N.Y.: Walker, 1997.

117. *Raussen M., Skau C.* Interview with Michael Atiyah and Isadore Singer // Notices of the American Mathematical Society. 2005. V. 52. P. 225–233.

118. *Appel K., Haken W.* The solution of the Four-Color-Map problem // Sci. American. 1977. V. 237, № 4.

119. *Mackenzie D.* What in the name of Euclid is going on here? // Science. 2005. V. 307. P. 1402–1403.

120. *Бурлак С.А., Старостин С.А.* Введение в лингвистическую компаративистику. М.: Эдиториал УРСС, 2001.

121. *Хелимский Е.* О прауральском происхождении чередования ступеней согласных (Helimski E. The Proto-Uralic origin of consonant gradation) // Московский лингв. журн. 1995. Т. 1. С. 34–40.

122. *Wang Z., Lo H.S., Yang H., Gere S., Hu Y., Buetlow K.H., Lee M.P.* Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer // Cancer Res. 2003. V. 63. P. 655–657.

123. *Sorek R., Basechess O., Safer H.M.* Expressed sequence tags: clean before using // Cancer Res. 2003. V. 63. P. 6996.

124. *Wang Z., Lo H.S., Yang H., Gere S., Hu Y., Buetow K.H., Lee M.P.* Reply to Sorek et al. // Cancer Res. 2003. V. 63. P. 6996–6997.

125. *Devos D., Valencia A.* Instrinsic errors in genome annotation // Trends Genet. 2001. V. 17. P. 429–431.

126. *Gilks W.R., Audit B., De Angelis D., Tsoka S., Ouzounis C.A.* Modeling the percolation of annotation errors in a database of protein sequences // Bioinformatics. 2002. V. 18. P. 1641–1649.

127. *Aravind L., Koonin E.V.* DNA-binding proteins and evolution of transcription regulation in the archaea // Nucl. Acids Res. 1999. V. 27. P. 4658–4670.

128. *Aravind L., Wolf Y.I., Koonin E.V.* The ATP-cone: an evolutionarily mobile, ATP-binding regulatory domain // J. Mol. Microbiol. Biotechnol. 2000. V. 2. P. 191–194.

129. *Wolf Y.I., Rogozin I.B., Kondrashov A.S., Koonin E.V.* Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genome context // Genome Res. 2001. V. 11. P. 356–372.

130. *Doerks T., Andrade M.A., Lathe W. 3rd, von Mering C., Bork P.* Global analysis of bacterial transcription factors to predict cellular target processes // Trends Genet. 2004. V. 20. P. 126–131.

131. *Rodionov D.A., Gelfand M.S.* Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling // Trends Genet. (in press).

132. *Borovok I., Gorovitz B., Yanku M., Schreiber R., Gust B., Chater K., Aharonowitz Y., Cohen G.* Alternative oxygen-dependent and oxygen-independent ribonucleotide reductases in Streptomyces: cross-regulation and physiological role in response to oxygen limitation // Mol. Microbiol. 2004. V. 54. P. 1022–1035.

133. *Graur D., Martin W.* Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision // Trends Genet. 2004. V. 20. P. 80–86.

134. *Hedges S.B., Kumar S.* Precision of molecular time estimates // Trends Genet. 2004. V. 20. P. 242–247.

135. *Reisz R.R., Muller J.* The comparative method for evaluating fossil calibration dates: a reply to Hedges and Kumar // Trends Genet. 2004. V. 20. P. 596–597.

136. *Glazko G.V., Koonin E.V., Rogozin I.B.* Molecular dating: ape bones agree with chicken entrails // Trends Genet. 2005. V. 21. P. 89–92.

137. *Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas P., Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T., Humphray S.,*

Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne S., Mullikin J.C., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston R.H., Wilson R.K., Hillier L.W., McPherson J.D., Marra M.A., Mardis E.R., Fulton L.A., Chinwalla A.T., Pepin K.H., Gish W.R., Chissoe S.L., Wendl M.C., Delehaunty K.D., Miner T.L., Delehaunty A., Kramer J.B., Cook L.L., Fulton R.S., Johnson D.L., Minx P.J., Clifton S.W., Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N., Cheng J.F., Olsen A., Lucas S., Elkin C., Uberbacher E., Frazier M., Gibbs R.A., Muzny D.M., Scherer S.E., Bouck J.B., Sodergren E.J., Worley K.C., Rives C.M., Gorrell J.H., Metzker M.L., Naylor S.L., Kucherlapati R.S., Nelson D.L., Weinstock G.M., Sakaki Y., Fujiyama A., Hattori M., Yada T., Toyoda A., Itoh T., Kawagoe C., Watanabe H., Totoki Y., Taylor T., Weissenbach J., Heilig R., Saurin W., Artiguenave F., Brottier P., Bruls T., Pelletier E., Robert C., Wincker P., Smith D.R., Doucette-Stamm L., Rubenfield M., Weinstock K., Lee H.M., Dubois J., Rosenthal A., Platzer M., Nyakatura G., Taudien S., Rump A., Yang H., Yu J., Wang J., Huang G., Gu J., Hood L., Rowen L., Madan A., Qin S., Davis R.W., Federspiel N.A., Abola A.P., Proctor M.J., Myers R.M., Schmutz J., Dickson M., Grimwood J., Cox D.R., Olson M.V., Kaul R., Raymond C., Shimizu N., Kawasaki K., Minoshima S., Evans G.A., Athanasiou M., Schultz R., Roe B.A., Chen F., Pan H., Ramser J., Lehrach H., Reinhardt R., McCombie W.R., de la Bastide M., Dedhia N., Blocker H., Hornischer K., Nordsiek G., Agarwala R., Aravind L., Bailey J.A., Bateman A., Batzoglou S., Birney E., Bork P., Brown D.G., Burge C.B., Cerutti L., Chen H.C., Church D., Clamp M., Copley R.R., Doerks T., Eddy S.R., Eichler E.E., Furey T.S., Galagan J., Gilbert J.G., Harmon C., Hayashizaki Y., Haussler D., Hermjakob H., Hokamp K., Jang W., Johnson L.S., Jones T.A., Kasif S., Kaspryzk A., Kennedy S., Kent W.J., Kitts P., Koonin E.V., Korf I., Kulp D., Lancet D., Lowe T.M., McLysaght A., Mikkelsen T., Moran J.V., Mulder N., Pollara V.J., Ponting C.P., Schuler G., Schultz J., Slater G., Smit A.F., Stupka E., Szustakowski J., Thierry-Mieg D., Thierry-Mieg J., Wagner L., Wallis J., Wheeler R., Williams A., Wolf Y.I., Wolfe K.H., Yang S.P., Yeh R.F., Collins F., Guyer M.S., Peterson J., Felsenfeld A., Wetterstrand K.A., Patrinos A., Morgan M.J., de Jong P., Catanese J.J., Osoegawa K., Shizuya H., Choi S.,Chen Y.J. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome // Nature. 2001. V. 409. P. 860–921.

138. *Salzberg S.L., White O., Peterson J., Eisen J.A.* Microbial genes in the human genome: lateral transfer or gene loss? // Science. 2001. V. 292. P. 1903–1906.

139. *Stanhope M.J., Lupas A., Italia M.J., Koretke K.K., Volker C., Brown J.R.* Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates // Nature. 2001. V. 411. P. 940–944.

140. *Genereux D.P., Logsdon J.M. Jr.* Much ado about bacteria-to-vertebrate lateral gene transfer // Trends Genet. 2003. V. 19. P. 191–195.

141. *Galperin M.Y., Koonin E.V.* Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption // In Silico Biol. 1998. V. 1. P. 55–67.

142. *Kyrpides N.C., Ouzounis C.A.* Whole-genome sequence annotation: «Going wrong with confidence» // Mol. Microbiol. 1999. V. 32. P. 886–887.

143. *Mushegian A.R.* Annotations of biochemically uncharacterized open reading frames (ORFs). Reply to Kyrpides and Ouzounis // Mol. Microbiol. 2000. V. 35. P. 697–698.

144. *Brenner S.E.* Errors in genome annotation // Trends Genet. 1999. V. 15. P. 132–133.

145. *Ouzounis C.A., Karp P.D.* The past, present and future of genome-wide re-annotation // Genome Biology. 2002. V. 2. P. COMMENT2001.

146. *Koonin E.V., Mushegian A.R., Galperin M.Y., Walker D.R.* Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea // Mol. Microbiol. 1997. V. 25. P. 619–637.

147. *Ibba M., Morgan S., Curnow A.W., Pridmore D.R., Vothknecht U.C., Gardner W., Lin W., Woese C.R., Soll D.* A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases // Science. 1997. V. 278. P. 1119–1122.

148. *Stathopoulos C., Jacquin-Becker C., Becker H.D., Li T., Ambrogelly A., Longman R., Soll D. Methanococcus jannaschii* prolyl-cysteinil-tRNA synthetase possesses overlapping amino acid binding sites // Biochemistry. 2001. V. 40. P. 46–52.

149. *Fabrega C., Farrow M.A., Mukhopadhyay B., de Crecy-Lagard V., Ortiz A.R., Schimmel P.* An aminoacyl tRNA synthetase whise sequence fits neither of the two known classes // Nature. 2001. V. 411. P. 110–114.

150. *Iyer L.M., Aravind L., Bork P., Hofmann K., Mushegian A.R., Zhulin I.B., Koonin E.V. Quod erat demonstrandum*? The mystery of experimental calidation of apparently erroneous computational analyses of protein sequences // Genome Biology. 2001. V. 2. P. RESEARCH0051.

The Apology of Bioinformatics

M.S. Gel'fand

Institute of Problems of Information Transmission, Russian Academy of Sciences, Bol'shoi Karetnyi per. 19, Moscow GSP–4, 101447 Russia

The discussion about adequate understanding of the term «bioinformatics» is continued. The relationships between bioinformatics and experimental molecular biology are considered. The list of the main branches and achievements of modern bioinformatics is presented.

Key words: bioinformatics