БИОХИМИЯ, 2021, том 86, вып. 12, с. 1826 – 1839

УДК 573

НЕСОСТОЯВШЕЕСЯ ИНТЕРВЬЮ С СИДНЕЕМ БРЕННЕРОМ: ПРЕВРАЩЕНИЕ ДАННЫХ В ЗНАНИЕ, БИОИНФОРМАТИКА, BIG DATA И … «IS WATER H_2O?»

© 2021 Л.Г. Кондратьева[1,2]*, М.В. Патрушев[1], Е.Д. Свердлов[1]*

[1] *НИЦ Курчатовский институт, 123182 Москва, Россия;*
электронная почта: liakondratyeva@yandex.ru, edsverd@gmail.com

[2] *ФГБУН Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН, 117997 Москва, Россия*

Поступила в редакцию 10.11.2021
После доработки 10.11.2021
Принята к публикации 10.11.2021

Обзор представляет собой попытку объяснить некоторые непростые проблемы, связанные с попытками разобраться в механизмах функционирования организмов, в частности, с использованием коллекций Big Data (больших данных). Форма обзора – воображаемое интервью с одной из наиболее ярких фигур эпохи возникновения молекулярной генетики и биологии, уникальным учёным и философом науки Нобелевским лауреатом Сиднеем Бреннером, открывшим для науки замечательный объект исследования – прозрачного 1000-клеточного червя *Caenorhabditis elegans* и многое другое. Его размышления и выводы относительно неизбежного «конфликта» между быстрорастущими массивами данных (Big Data), получаемых с помощью современных технологий секвенирования, и принципиальными ограничениями («запретами»), возникающими потому, что сложные взаимодействующие системы (организмы) вследствие взаимодействий порождают непредсказуемые «возникающие» свойства, абсолютно актуальны для понимания нерешаемых проблем таких современных тенденций, как «системная биология». Помимо принципиальных ограничений, сами по себе Big Data страдают от серьёзных дефектов, среди которых наиболее явными являются скрытые ошибки и принципиально низкая воспроизводимость. Отдельно следует отметить ещё один, возможно, принципиальный барьер – неполноту данных (количество данных $n \neq$ all). Эту проблему демонстрируют два небольших наиболее изученных организма, *Escherichia coli* (1600 генов, то есть 34,6% из 4623 уникальных генов не имеют экспериментальных доказательств функционирования) и *C. elegans* с белками, идентифицированными примерно для 50% генов. Другой яркий пример – «искусственная» бактерия, JCVI-syn3.0, с минимальным набором генов в геноме. Из её 473 генов биологическая функция не приписана 149 (31,5%). Бреннер указывает, что преобразование данных в знания представляет собой серьёзную проблему для будущих биологических исследований. Для этой цели биологии срочно нужна теоретическая база для её унификации. При этом правильным уровнем исследований он считает клетку и предлагает проект CELLMAP, как системы для организации биологической информации. Как абсолютно честный учёный, он говорит: «Если бы я знал [как это делать], я бы делал это, а не писал о проблеме. Понять, как это делать – остаётся главной проблемой биологических наук».

КЛЮЧЕВЫЕ СЛОВА: биоинформатика, Big Data, геном, системная биология, интервью, Сидней Бреннер.
DOI: 10.31857/S0320972521120071

ВВЕДЕНИЕ. КРАТКАЯ БИОГРАФИЯ СИДНЕЯ БРЕННЕРА

Сидней Бреннер родился в Южной Африке в 1927 г. Его отец был сапожником, евреем-иммигрантом из Литвы; его мать эмигрировала из Латвии. Сидней научился читать с помощью старых газет, и в возрасте 14 лет отправился в Университет в Йоханнесбурге, чтобы изучать медицину [1]. С самого начала ему удавалось совмещать формальную программу обучения с работой в лаборатории. Он увлёкся методами окрашивания биологической ткани и наблюдением за её детальной структурой под микроскопом. Он мог окрашивать и проводить диссекцию тканей для визуализации хромосом. Сидней узнал о концепциях теории информации и зарождающейся области информатики. В частности, его очень поразила концепция фон Неймана. В период учёбы он был вовлечён в левую политику, читал Ленина и Троцкого и приобрёл острое понимание социального контекста науки. Это увлечение продолжалось до конца жизни.

В 1952 г. Бреннер получил стипендию на факультете физической химии Оксфорда. Весной 1953 г. группа оксфордских исследователей, включая Бреннера, была приглашена в Кембридж, чтобы увидеть модель двойной спирали

* Адресат для корреспонденции.

Джима Уотсона и Френсиса Крика. В 1956 г. Крик пригласил Бреннера в Кембриджский университет в Великобританию. В Кембридже Бреннер провёл значительную часть своей жизни, тесно сотрудничая с Криком. В начале 1960-х гг., работая с бактериями и бактериофагами, Крик и Бреннер интерпретировали многие важные функции генов. С конца 1950-х гг. и до конца 1970-х гг. Бреннер и Крик занимали одну комнату, сохраняя эту привычку, несмотря на серию переездов возникшей Лаборатории молекулярной биологии (ЛМБ). В процессе этих передвижений ЛМБ постепенно приобретала её нынешнюю форму, становясь Лабораторий № 1 в молекулярной биологии.

Со временем Бреннер стал одним из ведущих биологов 20 и 21 веков. Он был блестящим экспериментатором и одним из самых умных и дальновидных учёных за последние 70 лет. Он постоянно искал следующий большой вопрос. Удивительно, но он поднялся на волнах открытия трёх ключевых моментов биологии 20-го века: подъём молекулярной генетики, поворот к пониманию генетических основ развития и разработка проектов генома.

Бреннер теоретически доказал в середине 1950-х гг., что генетический код «не перекрывается» – каждый нуклеотид является частью только одного триплета (три нуклеотида определяют каждую аминокислоту в белке), а последующие «триплетные кодоны» читаются по порядку. В 1961 г. Бреннер и Крик доказали это экспериментально [2].

В том же году Бреннер вместе с Франсуа Жакобом и Мэтью Мезельсоном (все из поколения научных гигантов, которые формулировали гипотезы и умели их обосновывать экспериментально) опубликовали экспериментальное подтверждение существования информационной РНК, которая переносит информацию от ДНК к белкам [3].

Наступила эпоха центральной догмы молекулярной биологии: генетическая информация передаётся от нуклеиновых кислот к белку, но не в обратном направлении. Правило было сформулировано Френсисом Криком в 1957 г. и опубликовано в 1958 г. [4]. В течение следующих двух лет с помощью Крика Бреннер установил, как завершается синтез белков, кодируемых последовательностями ДНК.

Революционный момент в научной карьере Бреннера наступил, когда вместе с Криком он решил, что интересная часть молекулярной биологии позади. (Заметьте, это произошло в период начала расцвета молекулярной биологии и генетики, который эти два индивидуума в значительной степени выковали!) Когда тысячи исследователей со всего мира начали заниматься молекулярной биологией, Крик и Бреннер начали готовиться к следующему этапу: пытаться с использованием нового молекулярного понимания решить серию биологических проблем, каждая из которых будет основываться на использовании конкретного биологического организма. Таким образом, наметился переход на следующий уровень исследований – молекулярные основы функционирования организма. В 1963 г. Сидней Бреннер писал руководителю ЛМБ, Максу Перуцу: «Сейчас широко известно, что почти все “классические” проблемы молекулярной биологии либо решены, либо будут решены в ближайшее десятилетие... Из-за этого я давно чувствовал, что будущее молекулярной биологии лежит в распространении исследований на другие области биологии, особенно на развитие и нервную систему» [5–7].

Они составили список из четырёх видов, каждый из которых позволял исследовать важную биологическую проблему: фаг лямбда (проект L), мышь (проект M), нематода (проект N) и кишечная палочка (проект K). В то же время и по тем же причинам Сеймур Бензер (Seymour Benzer, 1921–2007) решил изучить нейрогенетику дрозофилы, а Крик в конечном итоге обратился к изучению механизмов сознания. Сидней был в авангарде этого шага: его первоначальная цель состояла в том, чтобы использовать стратегии молекулярной генетики, разработанные на прокариотах. Но в отличие от многих его современников, которые остановились на хорошо зарекомендовавших себя модельных системах, таких как *Drosophila*, Сидней отправился на поиски многоклеточного организма, которого можно было бы поддерживать и размножать на чашках с агаром так же эффективно, как это делалось для изучения бактерий и фагов [1].

Принцип Бритвы Оккама гласит: «Не следует привлекать новые сущности без крайней на то необходимости». Бреннер использовал этот принцип, по-видимому, рефлекторно, врождённо.

После нескольких лет изучения различных видов к 1963 г. Бреннер решил изучать прозрачного червя *Caenorhabditis elegans* для решения новой важной проблемы молекулярной биологии (на самом деле, это была не молекулярная биология, а биология организма): генетика и биохимия механизмов контроля клеточного развития [8, 9]. Это была смена парадигм, которая описывалась философом науки Куном [10]. Кун разделял развитие науки на различные периоды: 1) нормальная наука, согласие между учёными, накапливается новое знание; 2) кри-

зисный период, рождаются новые проблемы, которые не находят адекватные решения; 3) период научной революции, когда существующий набор парадигм отбрасывается в пользу других. Появляются новые идеи, новые технологии, которые позволяют отвергнуть отжившие части научной структуры и сделать ещё один шаг к истине. При этом сторонники старой парадигмы оказывают ожесточённое сопротивление апологетам новой. В случае Бреннера – это была счастливая революция без борьбы и сильного сопротивления. Это стало возможным благодаря высокому авторитету, который Бреннер приобрёл благодаря всем своим предыдущим успехам. И, конечно же, благодаря поддержке таких гигантов, как Френсис Крик [11]. Вот оценка самого Бреннера: «Некоторые люди думали, что наш подход слишком “биологический” и уведёт нас от молекулярной биологии, но в любом случае нас попросили сделать официальное предложение, и в октябре 1963 г. соответствующий документ был представлен Совету» [12].

В годы, когда Бреннер работал над проектом нематоды в Кембридже, была особая атмосфера. В любое время дня и ночи там можно было найти Сиднея в генетической лаборатории у компьютера или у электронного микроскопа [9].

Не зря Бреннера называли мятежным лидером золотого века молекулярной биологии (Mischievous steward of molecular biology’s golden age) [11]. Он всегда старался найти более рациональное объяснение имеющимся фактам. Иногда в шутливой форме. Так и для смены парадигм он нашёл своё (почти) объяснение: «Часто новая волна встречает сильное сопротивление, но, как указывал Макс Планк, она побеждает, потому что оппоненты стареют и умирают. Процесс затем повторяется. Радикалы становятся либералами. Либералы становятся консерваторами, консерваторы реакционерами, и реакционеры исчезают» [13].

Иногда его оценки были слишком резкими. Например, такая история: «Однажды меня представили сэру Сиднею Бреннеру, как человека, занимающегося “биоинформатикой”. Доктор Бреннер от души рассмеялся и сказал: “Биоинформатика? Последнее прибежище негодяев (The last refuge of scoundrels)”. Я указал, что работаю над предсказанием структуры белков, и он ответил: “О, это другое. НАМНОГО ЛУЧШЕ”. Это было большим облегчением для меня» [14]. Наверное, автор ошибся и перепутал scoundrel и SPANDREL. Первое действительно означает негодяй, мерзавец, тогда как второе (НАМНОГО ЛУЧШЕ): это неизбежные и бесполезные пространства в архитектуре сложных сооружений. У Бреннера есть широко известная глава «Refuge of spandrels» в книге «Loose Ends and False starts» [15], и её заглавие, скорее всего, должно быть переведено как прибежище бесполезных, но неизбежных. Хотя это всё равно несправедливо по отношению к биоинформатике. Но мы обсудим это в воображаемом интервью.

Решение Бреннера исследовать организм повлекло за собой изменение методологической базы, в частности использование электронной микроскопии для идентификации клеток в червях и применение компьютеров для обработки данных. Это изменило подходы к науке. Из истории исследований *C. elegans* под эгидой Бреннера [16] видно, что Бреннер (сам) начал их примерно в 1966 г. Первая статья по генетике и ряду мутантов *C. elegans*, написанная исключительно Бреннером, появилась только в 1974 г. [17]. Восемь лет он работал без единой публикации. Сегодня невозможно представить, чтобы эффективность учёного оценивали не по количеству опубликованных им статей.

Сидней Бреннер сделал *C. elegans* центром внимания биологов [6, 18].

В 1974 г. Сидней и его сотрудник Джон Салстон (John Sulston, 1942–2018 гг. [19]) подсчитали, что геном *C. elegans* примерно в 20 раз больше, чем геном *E. coli*, и начали думать об установлении последовательности этого генома. Работа по секвенированию полного генома, как и вообще все работы по секвенированию полных геномов, были революционизированы появлением методов клонирования ДНК и разработкой Сэнгером и Максамом с Гилбертом методов быстрого секвенирования ДНК. Полная последовательность генома *C. elegans* была опубликована в 1998 г., это была первая полногеномная последовательность многоклеточного организма [20]. Эта 97-мегабазная геномная последовательность содержит более 19 000 генов.

Бреннер понимал, что возможность сравнивать последовательности генома разных организмов открыла бы целую большую область сравнительной геномики, которая привела бы к более глубокому понимание того, как эволюционные механизмы приводят к видовому разнообразию. Но для секвенирования многих геномов необходимы более быстрые и экономичные методы массового секвенирования, чем те, которые использовались для проектов *C. elegans* и генома человека. И в 1994 г. он предложил технологию для массового параллельного секвенирования, в которой использовались полинуклеотиды, связанные с иммобилизованными микрогранулами (microbeads) [21, 22]. Хотя эта технология не вошла в практику, она стала мощным двигателем в быстроразвивающейся облас-

ти и в конечном итоге вылилась в одну из доминирующих технологией секвенирования фирмы «Illumina».

Здесь хотелось бы упомянуть, что уже в конце 1980-х гг. в Институте молекулярной биологии АН СССР под руководством академика Андрея Мирзабекова активно шли работы по разработке нового метода секвенирования ДНК на микрочипах с использованием близкой идеологии секвенирования, основанного на гибридизации коротких олигонуклеотидов, иммобилизованных на оригинальной российской матрице. Приоритетная российская публикация появилась в печати в журнале «*DNA Sequence*» [23]. Также можно упомянуть, что стратегию быстрого секвенирования ДНК на основе специфического химического расщепления, которая длительно использовалась в методе Максама–Гилберта, была предложена нами в 70-х гг. [24–26].

В области сравнительной геномики Бреннер предложил использовать геном рыбы фугу, *Fugu rubripes*, который содержал в 8 раз меньше ДНК, чем у человека, при примерно таком же количестве генов [27]. Полная последовательность была опубликована в 2002 г. [28].

Бреннер также принимал активное участие в организации науки и формировании направлений её практического применения. В 1975 г. вместе с Полом Бергом и другими он организовал знаменитую встречу в Асиломаре, Калифорния, нацеленную на формирование способов использования новых технологий генной инженерии. Его популярные колонки в Current Biology (названные «Свободные концы», а затем «Ложные начала», Loose Ends and False Starts, собранные в книгу [29]) в середине 1990-х гг. пользовались большим успехом в научной среде.

Почти 10 лет он был директором известной Лаборатории молекулярной биологии. В 1986 г. он перешёл в новый отдел молекулярной генетики Совета медицинских исследований (MRC), где начал исследования в области эволюционной геномики. В начале 1990-х гг. Бреннер также координировал участие Великобритании в зарождающемся проекте «Геном человека».

В последний период своей жизни Бреннер путешествовал по миру. В этот период он был впечатлён динамизмом Сингапура, где он помог создать центр биомедицинских исследований «Биополис» и стал важной фигурой в Агентстве по науке, технологиям и исследованиям. Он также помог реструктурировать молекулярную биологию в Японии. Именно в Сингапуре он провёл свои последние годы, не имея возможности путешествовать по состоянию здоровья. Даже на этом этапе он каждое утро ходил в лабораторию, обсуждал последние результаты своих молодых коллег, вносил предложения по их экспериментам и работал над своим последним неопубликованным проектом: понимание структуры генома.

Бреннер породил поколение выдающихся учёных, в том числе 5 лауреатов Нобелевской премии. В значительной степени молекулярная биология перешла на новый уровень... [11, 30].

Что очень важно, в отличие от громадного большинства современников Сидней не попадал в ловушку мнения большинства и заглядывал в будущее человечества: «Я думаю, что наиболее важные успехи будут достигнуты в понимании биологии наиболее интересного вида – Homo sapiens. Я думаю, что благодаря этому пониманию мы сможем оценить различия между развитыми и разработанными сложными системами. Это если мы вообще выживем... Конечно, даже если произойдёт крупная катастрофа, некоторые из нас выживут. Тогда природа возьмёт верх, и биологическая эволюция начнётся снова, поскольку культурная эволюция потерпит неудачу. Я с уверенностью предсказываю, что будет выбор в пользу маленьких людей с телом, достаточным для поддержки необходимого количества умственных способностей. ...Наши преемники будут поражены количеством обсуждаемого сегодня научного мусора, если у них хватит терпения пролистать электронные архивы устаревших журналов» [31].

Сегодня предсказания Бреннера вот-вот осуществятся. Мы на грани катастрофы [32]: климатической, биосферной и демографической. Учёные бьют тревогу. Правительства практически бездействуют. Народ веселится. Титаник тонет, а оркестр играет бодрые песни.

НЕСОСТОЯВШЕЕСЯ ИНТЕРВЬЮ С СИДНЕЕМ БРЕННЕРОМ

A big computer, a complex algorithm,
and a long time does not equal science.

Robert Gentleman

Поиск истины в условиях предвзятости публикаций

Евгений Свердлов (Е.С.). Существует обычный (и сильный) bias (предвзятость) в публикациях. Др. Бреннер, я в последние годы очень интересуюсь проблемами, связанными с потенциалом науки в расшифровке механизмов функционирования сложных живых систем. Понятно, что большие надежды в этой расшифровке возлагают на компьютерные технологии и, в част-

ности, на биоинформатику. Мы живём в период всеобщих восторгов по поводу биоинформатики (определение см. в Частях 1 и 2 Приложения), и существует обычный (и сильный) bias в публикациях (см. Часть 3 Приложения), который, в частности, выражается в предвзятости ревьюеров и редакторов журналов, которые предпочитают публиковать позитивные оценки теорий, вписывающихся в рамки существующих парадигм, и отвергают всё, что в них не вписывается.

Сидней Бреннер (С.Б., перебивает). Сегодня Бог никогда не получит грант на исследования. Да, я об этом много писал и говорил в своих выступлениях [29]: манускрипты, поданные на публикацию, теперь подвергаются микроскопическому исследованию, но, к сожалению, не их научное содержание. То, что ищется, это с кем ты пишешь статьи, и где они опубликованы. Сегодня Бог никогда не получит грант на исследования. Один член редколлегии будет отрицать это на том основании, что работа была сделана очень давно; второй подтвердит это, отмечая, что это никогда не было воспроизведено. Отказ будет подхвачен третьим членом, указавшим, кроме всего прочего, что работа была опубликована в нерецензируемом журнале.

Прежде чем развивать псевдонауку анализа цитирования, мы должны напоминать себе, что самым главным является научное содержание статьи, и ничто не заменит её знание или чтение. Мы должны также признать, что цитирование часто даёт нам больше информации о социологии науки, чем о самой науке. В быстро развивающихся областях продолжительность жизни средней статьи очень мала, возможно, всего несколько месяцев, прежде чем она полностью исчезает, и о ней больше никогда не будет упоминаний. Мне говорили, что по физике только несколько статей с возрастом более 25 лет все ещё цитируется. Это должно быть очень приятно иметь работу в этом классе, но ещё лучше быть автором работы, которая так хорошо известна, что не требует литературного цитирования. Если при написании теперь процитируют Уотсона и Крика (1953 г.), то это, вероятно, будет рассматриваться как шутка [29].

Самое тревожное развитие состоит в том, что рейтинг цитирования, кажется, принят очень серьёзно. Все мы знаем, что наиболее цитируемые статьи – это те, которые содержат широко используемый рецепт или метод [29].

Е.С. Миражи цитируемости. Др. Бреннер, на минуточку вклинюсь. Я хорошо понимаю Ваше возмущение. В 2006 г. я написал на эту тему статью: «Миражи цитируемости. Библиометрическая оценка значимости научных публикаций отдельных исследователей» [33]. В этой статье, которую, кстати, высоко оценил «отец библиометрии» Евгений Гарфилд, я высказал мнение, что любые библиометрические данные, в том числе цитируемость отдельных статей в качестве независимой меры оценки научной значимости работ учёного, не могут служить критерием эффективности исследований или ценности публикаций. Потом я много раз возвращался к этой проблеме, но тщетно.

С.Б. Не все согласились бы с критериями выбора [статей для публикации]. Авторов бесит бесцеремонность, с которой редакторы и рецензенты относятся к их великим произведениям; редакторы жалуются на огромное количество скучных и повторяющихся манускриптов, которые они получают; и рецензенты жалуются на мусор, на чтение которого им приходится тратить своё драгоценное время. Когда понимаешь, что довольно часто это одни и те же люди, тогда очевидно, что у нас серьёзная проблема.

Все в биологии знают, что необходимо разделить то, что считается важными новостями, и то, что является достойным (но не слишком новым или значительным) дополнением к архиву. Некоторые журналы считают, что это их право и обязанность решать, что доносить до своих аудиторий, оставив большую часть исследований для «более технических журналов». Из-за этой общепризнанной политики журналы имеют высокую видимость, подкреплённую желанием каждого появиться на их страницах. Хотя многие согласны с такой политикой, не все – и особенно те, кто не избран, – согласились бы с критериями выбора.

Это таит в себе опасность того, что все избранные образуют клубы, а те, кто отброшены, сформируют свой клуб, из которого, конечно, они смогут исключить других. Это легко повторить, поэтому количество журналов будет продолжать расти, пока существуют группы, которые чувствует себя исключёнными. Опубликовать свои работы в нужных журналах стало почти так же сложно, как и выполнить само исследование.

Основная проблема, с которой мы сталкиваемся, – что делать с бесконечно растущим архивом полученных научных результатов в массивных томах журналов. Читать статью становится испытанием на физическую силу.

Один из моих самых циничных друзей сказал, что единственный способ быть абсолютно справедливым в принятии решений по заявкам на гранты, – это создать экспертную комиссию, которая совершенно невежественна и не заинтересована, таким образом, гарантируя, что любые предубеждения, которые могут возникнуть из знания предмета, полностью исключены.

Е.С. Есть возможность прибегнуть к помощи блогов. Они показывают резкое изменение отношения к биоинформатике в 2010 г. Спасибо, Доктор Бреннер. Я, с Вашего позволения, продолжу относительно предубеждённости редакторов и рецензентов. В последнее время появилась возможность прибегнуть к помощи блогов, где учёные могут позволить себе высказать своё мнение без опасения быть неопубликованным. Эта идея проникла в научное сообщество (см., например [34, 35]).

Ведение блогов стало широко распространённым социальным явлением. В настоящее время блоги признаны средством беспрецедентной силы для распространения информации. Научное сообщество подхватило эти методы, и в настоящее время существует более 1200 блогов, посвящённых учёным и их беседам [34].

Издатели начали ценить потенциал блогов по более интерактивному взаимодействию со своими читателями, по продвижению обсуждения содержания своих журналов. У многих крупных журналов теперь есть собственные блоги [34]. Я тоже попытался воспользоваться этой возможностью и посмотрел блог Дерека (Derek Lowe – a science writer), дающий оценки биоинформатике в разные годы.

18 июня 2010 г. в его блоге [36] перед читателями был поставлен вопрос: «Что биоинформатика когда-либо сделала для нас (What Has Bioinformatics Ever Done For Us)?» Приведу несколько типичных ответов на этот вопрос.

1) Биоинформатика является основой современной биологии – представьте, насколько бесполезной была бы любая последовательность ДНК (особенно целая) без инструментов для их исследования. ... Находка, что «бактерии» распались на две древние расходящиеся клады (Eubacteria и Archea), произошла исключительно на основе сравнения последовательностей. И открытие того, что Archea действительно больше похожа на эукариот в большинстве своих механизмов ... произошло исключительно благодаря просмотру последовательностей.

2) Очевидный ответ – это BLAST и последующие выводы о сходстве последовательностей.

3) Ни один из противовирусных препаратов, одобренных за последнее десятилетие, не появился бы на рынке без биоинформатики для анализа данных о мутациях резистентности.

В целом практически все ответы весьма позитивно оценивают биоинформатику.

В 2013 г. Дерек опубликовал скетч «Farewell to Bioinformatics» [37], где он привёл высказывание одного из авторов и комментировал его: «Вот разгневанный взгляд, который я не обязательно поддерживаю, но я также не могу сказать, что он полностью ошибочен».

Этот автор написал: «Биоинформатика – это попытка сделать молекулярную биологию актуальной. Все молекулярные биологи, лишённые навыков, превышающих навыки лаборанта, взывали к математикам и программистам, чтобы они волшебным образом извлекли науку из их горы дерьмовых результатов.

И вот программисты спустились и построили гигантские базы данных, в которых можно было быстро искать огромное количество дерьмовых результатов. Они написали алгоритмы, чтобы систематизировать дерьмовые результаты в виде деревьев и построить из них красивые графики, а молекулярные биологи старательно избегали сообщать программистам фактическое качество результатов. Когда для всех участников стало очевидно, что какой-то массив результатов, например данные, полученные с помощью микрочипов, бесполезен, последовала волна разговоров о том, что “эти данные не совсем количественные, но мы можем сделать качественные выводы”, после чего последовал поспешный переход на новую технику, бесполезность которой ещё не была доказана.

И базы данных росли, и все аннотировали свои данные путём поиска в базах данных, а затем отправляли их обратно в базы данных. Кажется, никто не указал, что это делает вашу базу данных отражением вашей базы данных, а не реальности. Вытащите какую-нибудь аннотацию из GenBank сегодня, и не так уж маловероятно, что она полностью неверна».

Далее следовали более 60 комментариев в целом согласных с этой точкой зрения.

Что же изменилось с 2010 г.? Рисунок показывает скорость роста сиквенсовой информации по годам, и можно видеть, что 2010 г. в какой-то степени переломный: примерно с этого времени начинается взрывной рост информации о последовательностях нуклеиновых кислот и белков. Это связано с новыми технологиями секвенирования (next-generation sequencing, NGS) [38, 39]. Общее количество данных о последовательностях удваивается примерно каждые 7 месяцев. Необработанные показания секвенирования, используемые в большинстве опубликованных исследований, архивируются либо в Архиве данных секвенирования (Sequence Read Archive, SRA), который поддерживается Национальным центром биотехнологической информации Национального института здравоохранения США (NIH/NCBI), либо в одном из его международных партнеров. В настоящее время SRA содержит более 3,6 петабайт (ПБ) данных и, по прогнозам, вырастет до 43 ПБ к 2023 г. В частности,

там есть последовательности геномов растений и животных и ~250 000 индивидуальных геномов человека, которые секвенированы или находятся в стадии разработки. Объём хранилища становится всё дороже в обслуживании, а данные труднее масштабировать. В настоящее время мировая производственная мощность по секвенированию, вероятно, превышает 35 петабаз (petabases) в год [40]. Если рост продолжится с теми же темпами, то к 2025 г. эта цифра приблизится к одной зеттабазе (zettabase) последовательностей в год. В общей сложности к 2025 г. будет расшифровано не менее 2,5 млн последовательностей геномов растений и животных. Таким образом, проблемы перед вычислительными мощностями невероятно возрастут (Computational challenges will thus incredibly increase [40]).

21 октября 2016 г. всё тот же Дерек опубликовал блог: «Ограничения Big Data» [42].

Из введения Дерека: «Данные помогут вам только постольку, поскольку они ведут к большему пониманию. Усилия по работе с Big Data помогут, но они не сразу откроют руководство к действию».

Эта статья в блоге привлекла 40 комментариев, некоторые из которых заслуживают внимания, потому что исследователи, написавшие комментарии, работают с биологическими и биомедицинскими базами данных, содержащими соответствующие Big Data. Ниже я привожу несколько типичных комментариев.

1) Большие данные = большой шум, слабый сигнал. «Получение большого количества неверных данных не помогает».

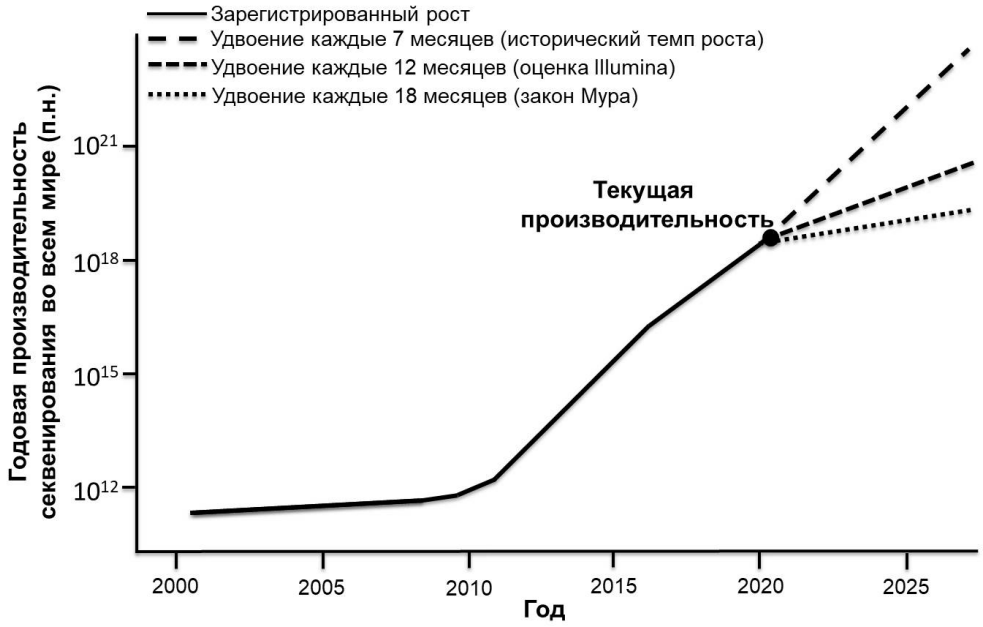
2) Со всеми деньгами, временем, презентациями, публикациями и общей суетой, затраченными для секвенирования ДНК больных раком, узнали ли мы что-нибудь действенное? И можно утверждать, что мы никогда не узнаем, потому что рак по определению имеет сотни зависимых мутаций. ...Big Data, небольшие данные, любые данные – все это бесполезно, если вы не измеряете что-то реальное и повторяемое.

3) Какой смысл подгонять всё больше и больше переменных ко всё большему количеству данных, чтобы проверить всё больше и больше потенциальных корреляций, когда половина данных всё равно не может быть воспроизведена? (E.C. см. мою статью [43].)

4) Раньше я работал с большими данными и обнаружил, что это в основном GIGO (Garbage In, Garbage Out, мусор на входе, мусор на выходе). Хуже того, «мусор» – это, по сути, шум, который заглушает любые полезные данные.

Sic transit gloria mundi. Биоинформатика вступила в сферу Big Data. И это оказалось выше уровня её компетенции.

E.C. «Big Data» и проблема полноты данных. n (количество данных) = all (все)? Др. Бреннер, я читал, что однажды Вы сказали, что хотели бы построить Грифона [14], мифологическое существо с головой, крыльями и когтями орла и те-



Рост общемировой производительности секвенирования ДНК. Значения после 2020 г. (пунктирные линии) представляют собой прогноз по трём различным сценариям (изменено из [40, 41])

лом льва, поскольку, только построив такой организм, Вы могли бы утверждать, что понимаете до малейших деталей, как работает развитие организма. Вы и искали в конце 1960-х гг. подходящий организм для изучения развития, содержащий относительно мало клеток, так чтобы можно было получить исчерпывающие данные: n = all. И нашли *C. elegans*. Меньше 1000 клеток. Но позже, если позволите, я вернусь к этой проблеме, потому что, как я понимаю, даже в этом случае полноты данных достичь не удалось.

Что касается «Big Data», то, несмотря на широкое использование термина, согласованного или единого определения «больших данных» не было и нет. И здесь в полный рост возникает проблема полноты данных, т.е. насколько n (количество данных) = all (все)? Многие авторы обсуждали двусмысленность, связанную с термином «Big Data». В глобальном интернет-словаре IGI можно найти 217 его определений [44]. Среди них 16 содержат утверждения, что эти данные превышают возможности обычно используемого оборудования и программного обеспечения. Более того, иногда авторы прямо указывают, что данные становятся большими данными, когда их объём, разнообразие и скорость превышают возможности архитектуры и алгоритма нашей системы (см., например [45, 46]). Некоторые авторы считают, что можно определить Big Data, как полную коллекцию данных, которая содержит все данные, существующие в исследуемой области, и может быть описана просто как n = all. Полнота является важным аспектом крупномасштабных биологических исследований [47].

Однако, кажется, сложно (если возможно) даже дать определение полноты. Достижение n = all для многих биологических данных может быть недостижимой или нереальной целью. Вопрос в том, есть ли момент, когда сбор данных «достаточно хорош», даже если он не является исчерпывающим.

В общем, доступные знания обычно фрагментарны, разрознены и разбросаны по многочисленным базам данных и бесчисленным информативным статьям. Не менее важно преобразовать полученные знания в полезную информацию. Однако реализовать это по-прежнему сложно [34]. Big Data не могут установить причинно-следственную связь [47].

В заключение: пока информация полная и стоит задача её систематизировать, биоинформатика находится на своём уровне компетенции. Но она не может дать положительный достоверный результат при анализе неполных или ошибочных или неполных и ошибочных данных. Произошёл классический переход от количества к качеству. Но крайне редко можно найти критическую статью, призывающую к точной оценке творческого потенциала биоинформатики больших данных.

Обычно там, где новые технологии обещают «большие успехи», как и в случае с большими данными, серьезные проблемы остаются или возникают [48]. Настоящая полезность больших данных может заключаться в их сочетании с классическими методами, основанными на гипотезах. При использовании обоих может повыситься клиническая эффективность и уменьшиться эффект «Garbage in» и, как следствие, «Garbage out», Gigo.

С.Б. Попытки получить концептуальные знания из огромных коллекций Big Data. Компьютеры могут извлекать биологическое значение непосредственно из последовательностей ДНК?

Биоинформатика имеет своё место. Её основная деятельность была полезной в том, что большие массивы данных теперь можно легко использовать для исследований. Однако идея о том, что данные последовательности могут добавить дополнительную информацию к той, которую дадут нам знания о функциях, неуместна. Для этого мы должны сделать больше, чем перегруппировать то, что известно; и нужна теория, которую мы можем проверить. Предмет, который будет развиваться, следует назвать «Теоретической Биологией», но так как это название имеет плохую репутацию, мы назовём это «Компьютационной биологией». Последовательность станет инструментом исследования, как от неё ожидали, а не самоцелью. Вероятно, что проекты генома для *C. elegans* и *Drosophila* окажут такое же влияние на эти области исследований в основном из-за большого числа исследователей, которые могут немедленно использовать этот продукт [29].

С геномами позвоночных мы обнаруживаем, что на первый план выходит новая идея. Грубо говоря, сторонники этого явления пришли к выводу, что компьютеры могут извлекать биологическое значение непосредственно из последовательностей ДНК. Этот подход породил два новых направления деятельности. Одно – биоинформатика, которая просто претенциозно пытается произвести впечатление, что имеет большее значение или заслуги, чем имеет на самом деле; другое – функциональная геномика – нелепа. Последнее использует первое, чтобы попытаться найти функцию из последовательности генов. Я не думаю, что есть какие-нибудь факультеты университета, посвящённые этим предметам, но, безусловно, растёт число компаний, занятых одной или двумя этими проблемами [29].

Я начал свою Нобелевскую лекцию первыми словами из моей статьи по нематоде в 1974 г. «The genetics of *Caenorhabditis elegans*» [17]: «Как гены могут определять сложные структуры, обнаруженные у высших организмов, – главная нерешённая проблема биологии».

Выступая в 2008 г. на симпозиуме в Индии, я сказал [49]: «Существует кризис во всех науках в наши дни. Мы тонем в море данных, и всё-таки мы томимся от жажды. Сегодня наука вознаграждает только тех, кто коллекционирует и распределяет данные... Данные являются заменителями мышления».

«Самое время остановиться и спросить самих себя: что мы ожидаем найти в конце этой громадной головоломки омиков» (omic brainbow) [50].

С.Б. Превращение данных в знание – большая проблема. «Превращение данных в знание представляет собой большую проблему для будущих биологических исследований. Новая наука, системная биология, заявляет, что она способна решить эту проблему, но я утверждаю, что этот подход потерпит неудачу» [51].

Идея, если её можно так назвать, состоит в том, чтобы взять миллионы точек данных и пропустить их через некую компьютерную программу, и посмотреть, какие ассоциации могут быть обнаружены. Некоторые даже намекнули на анализ данных Фурье, но, что наиболее несправедливо, один из миссионеров новой области заявил, что это освободит нас от «оков биологических исследований, в которых доминируют гипотезы». Проще говоря, вам не нужно больше думать, чтобы проводить исследования. Неужели мы действительно собираемся вступить в фазу упадка биологии, в которой учёные не смогут увидеть, в чём заключаются проблемы, или, если они это сделают, не смогут сформулировать вопросы, на которые можно было бы ответить либо путём наблюдения и измерения, либо путём вмешательства и эксперимента [29]?

Биология отличается от физики тем, что организмы появились путём естественного отбора, а не как решения математических уравнений. Много лет назад я слышал, как великий теоретический физик, Юджин Вигнер, выступал с докладом про нефизические или «чудесные» свойства биологической системы. Он утверждал, что невозможно получить достаточное количество уравнений для определения квантовых состояний и что что-то ещё должно было быть вовлечённым – возможно, сознание.

Я указал, что если я возьму профессора Вигнера и разложу его на ансамбль элементарных частиц, шансы на то, что они снова соберутся в одного и того же профессора Вигнера с акцентом, будут равны нулю и действительно потребуют чуда.

Но профессор Вигнер и другие биологические организмы не производятся путём конденсации в пакете элементарных частиц, а «формируются» некоторыми очень особенными процессами, которые, конечно, происходят в соответствии с законами физики, но не могут быть напрямую выведены из них.

Проблема с физикой состоит в том, что её самые глубокие заявления совершенно непонятны почти всем, кроме самых глубоких физиков, и, хотя заявления вполне могут быть абсолютно правдивыми, все они довольно бесполезны, если моя цель – понять *E. coli*.

Бесполезно сожалеть о кончине золотого века молекулярной генетики, когда многое было достигнуто путём объединения мысли с несколькими хорошо подобранными экспериментами с простыми вирусными и бактериальными системами. Также бесполезно осуждать нынешний подход биологии «низкие затраты, высокая производительность, отсутствие результатов», который доминирует на страницах наших остро конкурирующих научных журналов. Мы должны с распростёртыми объятиями приветствовать всё, что могут предложить нам современные технологии, но мы должны научиться использовать их по-новому. Биологии срочно нужна теоретическая база для её унификации, и только теория позволит нам преобразовать данные в знания [51].

С.Б. Геном должен лежать в основе любой теории, которую мы строим. Но нет простого способа картировать организм на его геном. «Геном должен лежать в основе любой теории, которую мы строим, но так как преобразование информации в геноме в конечный живой организм включает в себя множество сложных процессов, опосредованных молекулами, запрограммированными в геноме, всё это нужно будет изучить довольно подробно, прежде чем мы сможем читать и понимать геномы. Нет простого способа “картировать” организмы на их геномы, если они достигли определённого уровня сложности. Таким образом, хотя последовательность геномов является центральной, она представляет собой уровень абстракции, который является слишком загадочным, чтобы использоваться как таковой для организации данных и построения теоретических моделей. Предложения основывать всё на последовательности генома, аннотируя его дополнительными данными (Прямая задача. – *Е.С.*), будут приводить только к увеличению его непонятности» [51].

Е.С. Сложные системы. Возникающие в сложных системах свойства непредсказуемы. Is water H_2O? В этой связи я хотел бы ещё раз подчеркнуть то, что уже давно присутствует в нашей бе-

седе, но не получило чёткого определения. Это то, что называется «сложной системой». Сложная система – многокомпонентная система, состоящая из взаимодействующих субъединиц, результатом взаимодействия которых являются так называемые возникающие (emergent) свойства, присущие целой системе и не предсказуемые на основании свойств исходных субъединиц.

В своё время я нашёл в Интернете замечательный и наглядный пример сложной системы. Заметка называлась: «Is water H_2O?» Ответ на заданный в заголовке вопрос кажется очевидным: «Конечно». На самом деле H_2O – это только структура молекул, из которых состоит вода. Но формула мало говорит нам о свойствах воды, например таких, как температура кипения и замерзания, поверхностное натяжение, свойства льда и др. Это очень простой и поучительный пример. Он полностью относится к биологии. Из структур молекул их функции прямо не следуют. В результате взаимодействий молекул в веществе, образуемом ими, появляются новые непредсказуемые свойства – возникающие свойства, по-английски – emergent properties. Три взаимодействующих атома – это уже достаточно сложная система, чтобы стать непредсказуемой для нас. А представьте себе мозг!

Любая живая система или даже её отдельный модуль – это система сложная. Такие системы характеризуются большим числом гетерогенных компонентов, будь то гены, белки или клетки. Эти компоненты взаимодействуют мириадами способов во временно́м (от микросекунд до лет) и пространственном (от нанометров до метров) режимах. Полное понимание этих систем требует, чтобы большая часть этих взаимодействий была экспериментально или компьютационно изучена. Это очень трудно [52].

Такие разные области, как нейробиология и биология рака, не поддаются лёгким предсказаниям относительно неминуемых практических приложений. Улучшенные технологии наблюдения и тестирования биологических систем привели только к дальнейшим уровням сложности, с которой нужно работать... Мы очень далеки от понимания клеточной биологии, геномов или мозгов и от превращения этого понимания в практически полезное знание [53]!

С.Б. Френсис Крик в раю. Сам Бог не знает, как работают сложные системы. Я делил офис с Френсисом Криком 20 лет в Кембридже. Одно время его интересовала эмбриология, и он потратил много времени, думая о имагинальных дисках дрозофилы. Однажды он бросил книгу, которую он читал, на свой стол с раздражённым криком: «Бог знает, как эти воображаемые диски работают». В мгновение ока я увидел историю о прибытии Френсиса на небеса, и апостол Пётр приветствует его словами: «О, Доктор Крик, вы, должно быть, устали после долгого путешествия. Садитесь, есть выпить, и расслабьтесь». «Нет», – говорит Френсис: «Я должен увидеть этого парня, Бога. Я должен задать ему вопрос». После некоторых уговоров ангел соглашается привести Френсиса к Богу. Они пересекают среднюю часть неба и подходят к сараю с крышей из гофрированного железа. А в задней части сидит человечек в комбинезоне с большим гаечным ключом в заднем кармане. «Бог», – говорит ангел: «Это доктор Крик; Доктор Крик, – это Бог». «Я очень рад познакомиться с вами», – говорит Френсис: «Я должен задать Вам один вопрос. Как имагинальные диски работают?» «Хорошо», – отвечает Бог: «Мы взяли немного этого материала, и мы добавили к нему кое-что и..., на самом деле, мы не знаем, но я могу сказать вам, что мы строили эту муху, которая летает здесь в течение 200 млн лет, и жалоб нам не поступало» [54].... Бог никогда не патентовал эволюцию. Он держал это, как производственный секрет [54].

С.Б. Биологические науки должны уделять внимание деталям. Биологические науки должны уделять внимание деталям, потому что живые организмы являются продуктами эволюционировавших геномов и не могут быть представлены в виде решений дифференциальных уравнений. Будет важно найти ***все*** факторы транскрипции и ***все*** последовательности, с которыми они связываются, и мы не должны рассматривать первый случай как революционную новость, а последующие случаи – как повторяющиеся [54].

Хорошие теории молекулярных или клеточных сетей потребуют знания ***всех*** взаимодействий [54].

Е.С. Даже наиболее изученные простые живые организмы далеки от n = all. Это относится также к широко обсуждаемым сейчас так называемым ген-регуляторным сетям (ГРС, см. последние обзоры [55–57]). ГРС представляет собой систему молекулярных взаимодействий, в которой внутренние сигналы (например, в процессе эмбрионального развития) или сигналы из окружающей среды преобразуются в дифференциальную экспрессию генов, иными словами, в экспрессию, различающуюся для разных клеток или для одних и тех же клеток, но в разное время. Регуляция транскрипции опосредована комбинаторными взаимодействиями между *цис*-регуляторными элементами ДНК и *транс*-действующими факторами транскрипции и является, по-видимому, самым важным механиз-

мом контроля экспрессии генов. Необходимость полноты данных для создания ГРС обсуждал ещё один из пионеров этого подхода, Эрик Дэвидсон [58].

Но даже в самых простых системах мы не можем достичь этой полноты.

Первая геномная последовательность *E. coli* была определена более 20 лет назад. Однако молекулярные и физиологические функции для 1600 (около 35%) из 4623 генов остаются неизвестными [59].

Более того, у Вашей любимой и наиболее изученной *C. elegans* 23% генов, кодирующих белки, остаются функционально неясными [47].

И ещё хуже: недавно Вентер и его коллеги синтезировали геном *Mycoplasma genitalium*, названный JCVI-syn3.0. Из 473 генов в сокращённом наборе 149 генам (31,5%) никакая конкретная биологическая функция не может быть приписана. Вопрос в том, что делают эти гены и почему они необходимы. Это «известные неизвестные» [60, 61].

Поистине, мир легче создать, чем понять. По-видимому, *n* никогда не будет = all.

С.Б. Сложность биологических систем возникает в эволюции в результате приобретений и модификаций. Нам нужно поместить всё в эволюционные рамки просто потому, что сложность биологических систем возникает в результате приобретений и модификаций, а не в результате повторного изобретения. Свойства многих компонентов в наших клетках, будь то мРНК или белки, будут обусловлены не только процессами отбора на определённые активности и уровни, потому что они дают позитивный эффект, но также и теми, которые не вызывают негативных последствий для организма и могут принимать любое значение. Это условие «безразличия» почти наверняка будет присутствовать, потому что это дешёвое решение проблемы регулирования сложных систем. Таким образом, 20% или двукратное увеличение, или даже само присутствие белка может быть очень значительным или совершенно не относящимся к делу в зависимости от того, соответствует ли он условию «безразличия». Только эксперимент может решить, что происходило [62, 63].

В 1990 г. я заметил, что биохимия и коммунизм, казалось, исчезли в этом году (оригинальные высказывания Сиднея Бреннера на английском языке приведены в Части 4 Приложения). Большинство людей думали, что я это сказал с ликованием, но, на самом деле, – с сожалением, по крайней мере для биохимии. Есть ещё одна тема, которая исчезла несколько десятилетий назад, которую нам тоже нужно изобретать заново – физиология. Классическая физиология интересовалась функциями организмов. Я часто слышу, как говорят, что то, что нам сейчас нужно – это интегративная (системная) биология; что у нас очень хорошо получается выяснить, как работают простые системы с небольшим количеством компонентов, но мы очень плохо складываем детали многокомпонентных систем вместе. Что касается последнего, я считаю, что нам понадобятся две вещи. В конце концов, клетка выполняет интеграцию функций всех своих компонентов, так что клетки могли бы выявить, что такое интегрированное поведение.

Итак, первое требование будет разработать теоретическую модель (framework), в которую можно встроить все подробные знания, которые мы накопили, чтобы позволить нам вычислять результаты сложных взаимодействий и начать понимать динамику системы. Во-вторых, возможность проводить параллельные измерения поведения многих компонентов во время выполнения клеткой интегрированного действия с целью проверки, верна ли теория. Есть ли другие подходы? Если бы я знал это, я бы делал это, а не писал о проблеме [29].

На самом деле, оргия добычи фактов, в которой все в настоящее время участвуют, накопила огромный долг. Это долг создания теории, и некоторые из нас скоро будут иметь захватывающее время, оплачивая его обратно – надеюсь, с интересом.

С.Б. Правильный уровень абстракции – это клетка. CELLMAP. Правильный уровень абстракции – это клетка. Клетка является фундаментальной единицей структуры, функции и организации живых систем. Это ключевая особенность того, что я назвал CELLMAP и что является основой для биологической информационной системы, которая позволит нам не только обрабатывать огромное количество данных, но и генерировать и проверять гипотезы. CELLMAP представляет собой карту молекул в клетках и карту клеток в организме. Для микробов клетка также является организмом. Все мы начали своё существование как одна клетка, которая умножалась, производя больше клеток. Эти клетки далее дифференцировались во многие разные типы клеток, образуя ткани и органы, ответственные за наши физиологические функции. При выборе уровня клетки мы избегаем вопроса о том, должны ли наши анализы быть сверху вниз (top-down) или снизу вверх (bottom-up); вместо этого наш подход является средним, потому что, с точки зрения клетки, мы можем смотреть вниз на молекулы, которые составляют её, и смотреть вверх на организм, который её содержит. Кроме того, мы можем принять единую концептуальную архитектуру

для всех уровней, рассматривая организм, как сеть взаимодействующих клеток так же, как мы рассматриваем клетку в виде сети взаимодействующих молекул [51].

ПОЛУСЕРЬЕЗНОЕ ЗАКЛЮЧЕНИЕ. ГЕНЕТИКА ГЕНИАЛЬНОСТИ УДАЧИ. ВОСПРОИЗВОДИМОЕ И УНИКАЛЬНОЕ

Е.С. Др. Бреннер, в заключение позвольте задать такой специфический вопрос: «Вы очень рано научились читать, причём росли, как можно понять, в семье не очень интеллектуальной. Это очень напоминает биографии по крайней мере некоторых гениальных людей, Гаусса (Иоганн Карл Гаусс (1777–1855)), например. Он также родился и рос в семье далёкой от образования и считается одним из величайших математиков всех времен, «королём математики». Кстати, иностранный почётный член Петербургской АН. Так же, как и Вы, он сам научился и в 3 года уже умел читать и писать. Как и Вы, в 15 лет поступил в колледж. Есть по крайней мере несколько подобных примеров. Есть ли гены гениальности? Удачи?

В России есть прочно устоявшийся термин «фартовый» (удачливый). Во время Отечественной войны – это, например люди, которые прошли всю войну, не получив ни одного ранения, при этом не отсиживались в тылу. С удачливыми командирами солдаты охотно шли в разведку, будучи уверенными, что вернуться живыми с этим командиром более вероятно, чем с другим. Можно привести много других примеров: есть люди, которые открывали древние клады и т.д.

С.Б. Не буду говорить о гениальности. Но несколько лет назад я обнаружил, что в каждой семье есть дядя Фрэнк. Он тот, кто выкуривал 60 сигарет и выпивал 2 бутылки водки каждый день своей жизни с четырёхлетнего возраста, у него было 6 жён и бесчисленное количество подруг, и он участвовал в гонках на Феррари.

К сожалению, он погиб в результате несчастного случая во время альпинизма в возрасте 92 лет в Гималаях. Я тщетно пытался заинтересовать учёных и политиков запуском проекта «Геном дяди Фрэнка», чтобы мы могли заполучить все эти хорошие гены [29].

Мне сказали, что дяди Фрэнки в этом мире единственные, кто удачлив; но я не считаю это удовлетворительным ответом. Генетика удачи кажется мне хорошей темой, намного лучше, чем генетика алкоголизма или гомосексуализма.

Е.С. Спасибо, Доктор Бреннер. Я, в связи с уникальностью, вспомнил слова Вольфганга Паули (Wolfgang Pauli), одного из основателей квантовой физики: «Учёный-естествоиспытатель интересуется определёнными явлениями, ...он должен ограничиваться воспроизводимым... Я не утверждаю, что воспроизводимое, само по себе, более важно, чем уникальное. Но я утверждаю, что уникальное выше исследования научным методом».

Финансирование. Работа выполнена при поддержке гранта Министерства науки и высшего образования Российской Федерации, выделенного Курчатовскому геномному центру (грант № 075-15-2019-1659).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Соблюдение этических норм. Настоящая статья не содержит описания каких-либо исследований с участием людей или животных в качестве объектов.

Дополнительные материалы. Приложение к статье опубликовано на сайте журнала «Биохимия» (https://biochemistrymoscow.com) том 86, вып. 12, 2021.

СПИСОК ЛИТЕРАТУРЫ

1. White, J., and Bretscher, M. S. (2020) *Sydney Brenner. 13 January 1927–5 April 2019*, The Royal Society Publishing, doi: 10.1098/rsbm.2020.0022.
2. Crick, F. H., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961) General nature of the genetic code for proteins, *Nature*, **192**, 1227-1232, doi: 10.1038/1921227a0.
3. Brenner, S., Jacob, F., and Meselson, M. (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis, *Nature*, **190**, 576-581, doi: 10.1038/190576a0.
4. Crick, F. H. C. (1957) Nucleic acids, *Sci. Am.*, **197**, 188-203.
5. CSHL Archives repository. Preserving and promoting the history of molecular biology. Letter from Sydney Brenner to Max Perutz, URL: http://libgallery.cshl.edu/items/show/64089.
6. Sydney Brenner moves *C. elegans* into the limelight, *Worm History*, URL: https://www.hobertlab.org/how-the-worm-got-started/.
7. Brenner, S. (1963) A letter to Max Perutz, 5, June, URL: http://nemaplex.ucdavis.edu/General/Biographies/SBrenner.htm.
8. Kenyon, C. (2019) Sydney Brenner (1927-2019), *Science*, **364**, 638, doi: 10.1126/Science.Aax8563.
9. Grens, K. (2019) *Sydney Brenner, mRNA Discoverer, Dies*, TheScientist, URL: https://www.the-scientist.com/news-opinion/sydney-brenner--mrna-discoverer--dies-65708.8.
10. Kuhn, T. S. (1962) The Structure of Scientific Revolutions: University of Chicago Press, Original Edition.
11. Friedberg, E. (2019) Sydney Brenner (1927-2019), *Nature*, **568**, 459-460.

12. From the *C. elegans* server: Sydney Brenner, January 1, 2020, URL: http://nemaplex.ucdavis.edu/General/Biographies/SBrenner.htm.
13. Brenner, S. (2012) The revolution in the life sciences, *Science*, **338**, 1427-1428.
14. Dunbrack, R. L. (2003) A scoundrel's refuge, *Nat. Struct. Mol. Biol.*, **10**, 590-590, doi: 10.1038/nsb0803-590.
15. Brenner, S. (1998) Refuge of spandrels, *Curr. Biol.*, **8**, R669.
16. Ankeny, R. A. (2001) The natural history of *Caenorhabditis elegans* research, *Nat. Rev. Genet.*, **2**, 474-479.
17. Brenner, S. (1974) The genetics of *Caenorhabditis elegans*, *Genetics*, **77**, 71-94.
18. Brenner, S. (2009) In the beginning was the worm, *Genetics*, **182**, 413-415, doi: 10.1534/genetics.109.104976.
19. Sulston, J. (2002) A conversation with John Sulston, *Yale J. Biol. Med.*, **75**, 299-306.
20. *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology, *Science*, **282**, 2012-2018, doi: 10.1126/science.282.5396.2012.
21. Brenner, S. (1997) *Massively Parallel Sequencing of Sorted Polynucleotides*, Google Patents.
22. Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays, *Nat. Biotechnol.*, **18**, 630-634, doi: 10.1038/76469.
23. Khrapko, K., Lysov, Y. P., Khorlin, A., Ivanov, I., Yershov, G., et al. (1991) A method for DNA sequencing by hybridization with oligonucleotide matrix, *DNA Sequence*, **1**, 375-388.
24. Sverdlov, E., Monastyrskaya, G., Chestukhin, A., and Budowsky, E. (1973) The primary structure of oligonucleotides. Partial apurinization as a method to determine the positions of purine and pyrimidine residues, *FEBS Lett.*, **33**, 15-17.
25. Sverdlov, E., Monastyrskaya, G., Budowsky, E., and Grachev, M. (1972) A novel approach to structural analysis of oligonucleotides, *FEBS Lett.*, **28**, 231-235.
26. Müller-Hill, B. (1996) *The Lac Operon: A Short History of a Genetic Paradigm*, Berlin, New York, Walter de Gruyter, doi: 10.1515/9783110879476.
27. Brenner, S., Elgar, G., Sanford, R., Macrae, A., Venkatesh, B., et al. (1993) Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome, *Nature*, **366**, 265-268.
28. Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes, *Science*, **297**, 1301-1310, doi: 10.1126/science.1072104.
29. Brenner, S. (2019) *Loose Ends... False Starts*, World Scientific.
30. Cobb, M. (2019) Sydney Brenner (1927-2019), *Dev. Cell*, **49**, 493-495.
31. Brenner, S. (2006) Sydney Brenner forecasts the future, *NewScientist*, 15 November 2006, URL: https://www.newscientist.com/article/mg19225780-079-sydney-brenner-forecasts-the-future.
32. Buzdin, A. V., Patrushev, M. V., and Sverdlov, E. D. (2021) Will plant genome editing play a decisive role in "Quantum-Leap" improvements in crop yield to feed an increasing global human population? *Plants*, **10**, 1667.
33. Свердлов, Е. (2006) Миражи цитируемости. Библиометрическая оценка значимости научных публикаций отдельных исследователей, *Вестник Российской академии наук*, **76**, 1073-1085.
34. Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., et al. (2009) Calling International Rescue: knowledge lost in literature and data landslide! *Biochem. J.*, **424**, 317-333, doi: 10.1042/BJ20091474.
35. Batts, S. A., Anthis, N. J., and Smith, T. C. (2008) Advancing science through conversations: bridging the gap between blogs and the academy, *PLoS Biol.*, **6**, e240, doi: 10.1371/journal.pbio.0060240.
36. Lowe, D. (2010) What has bioinformatics ever done for us? *Science*, URL: https://www.science.org/content/blogpost/has-bioinformatics-ever-done-us.
37. Lowe, D. (2013) Farewell to bioinformatics, *Science*, URL: https://blogs.sciencemag.org/pipeline/archives/2013/01/30/farewell_to_bioinformatics.
38. Maljkovic Berry, I., Melendrez, M. C., Bishop-Lilly, K. A., Rutvisuttinunt, W., Pollett, S., et al. (2020) Next Generation sequencing and bioinformatics methodologies for infectious disease research and public health: approaches, applications, and considerations for development of laboratory capacity, *J. Infect. Dis*, **221**, S292-S307, doi: 10.1093/infdis/jiz286.
39. Sakr, S., and Zomaya, A. Y. (2019) *Encyclopedia of Big Data Technologies*, Springer International Publishing, doi: 10.1007/978-3-319-77525-8.
40. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., et al. (2015) Big Data: astronomical or genomical? *PLoS Biol.*, **13**, e1002195, doi: 10.1371/journal.pbio.1002195.
41. Improving our understanding of genome structure and function is central to biology and medicine, URL: https://www.tugraz.at/tu-graz/services/news-stories/planet-research/einzelansicht/article/cracking-the-code-within-us-bioinformatics-of-the-human-genome/.
42. Lowe, D. (2016) *The Limits of Big Data*, URL: https://blogs.sciencemag.org/pipeline/archives/2016/10/21/the-limits-of-big-data.
43. Alekseenko, I. V., Pleshkan, V. V., Monastyrskaya, G. S., Kuzmich, A. I., Snezhkov, E. V., et al. (2016) Fundamentally low reproducibility in molecular genetic cancer research, *Genetika*, **52**, 745-760.
44. IGI-Global. *What is Big Data*, URL: https://www.igi-global.com/dictionary/data-knowledge-and-intelligence/39008.
45. Moseley, E. T., Hsu, D. J., Stone, D. J., and Celi, L. A. (2014) Beyond open big data: addressing unreliable research, *J. Med. Int. Res.*, **16**, e259, doi: 10.2196/jmir.3871.
46. Ware, A., Janvale, G., Shaikh, F., and Harke, S. (2017) HADOOP: Solution for Big Data challenges in bioinformatics and its prospective in India, *IOSR J. Comp. Eng.*, 51-54.
47. Hutter, H., and Moerman, D. (2015) Big Data in *Caenorhabditis elegans*: *quo vadis*? *Mol. Biol. Cell*, **26**, 3909-3914, doi: 10.1091/mbc.E15-05-0312.
48. Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., et al. (2019) From Big Data to precision medicine, *Front. Med. (Lausanne)*, **6**, 34, doi: 10.3389/fmed.2019.00034.
49. Brenner, S. (2008) *Data Is a "Substitute for Thinking"*, URL: http://www.genomeweb.com/blog/data-substitute-thinking.
50. Brenner, S., and Sejnowski, T. J. (2011) Understanding the human brain, *Science*, **334**, 567, doi: 10.1126/science.1215674.
51. Brenner, S. (2010) Sequences and consequences, *Phil. Trans. R. Soc. Lond. Ser. B Biol. Sci.*, **365**, 207-212, doi: 10.1098/rstb.2009.0221.
52. Koch, C. (2012) Systems biology. Modular biological complexity, *Science*, **337**, 531-532, doi: 10.1126/science.1218616.
53. Greek, R., and Hansen, L. A. (2013) Questions regarding the predictive value of one evolved complex adaptive system for a second: exemplified by the SOD1 mouse, *Prog. Biophys. Mol. Biol.*, **113**, 231-253, doi: 10.1016/j.pbiomolbio.2013.06.002.

54. Brenner, S. (1995) Loose ends, *Curr. Biol.*, **5**, 1328.
55. Ewe, C. K., Cleuren, Y. N. T., and Rothman, J. H. (2020) Evolution and developmental system drift in the endoderm gene regulatory network of Caenorhabditis and other nematodes, *Front. Cell. Dev. Biol.*, **8**, 170, doi: 10.3389/fcell.2020.00170.
56. Halfon, M. S. (2017) Perspectives on gene regulatory network evolution, *Trends Genet.*, **33**, 436-447, doi: 10.1016/j.tig.2017.04.005.
57. Lavin, D. P., and Tiwari, V. K. (2020) Unresolved complexity in the gene regulatory network underlying EMT, *Front. Oncol.*, **10**, 554, doi: 10.3389/fonc.2020.00554.
58. Peter, I. S., and Davidson, E. H. (2009) Modularity and design principles in the sea urchin embryo gene regulatory network, *FEBS Lett.*, **583**, 3948-3958, doi: 10.1016/j.febslet.2009.11.060.
59. Ghatak, S., King, Z. A., Sastry, A., and Palsson, B. O. (2019) The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function, *Nucleic Acids Res.*, **47**, 2446-2454, doi: 10.1093/nar/gkz030.
60. Hutchison, C. A., 3rd, Chuang, R. Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., et al. (2016) Design and synthesis of a minimal bacterial genome, *Science*, **351**, aad6253, doi: 10.1126/science.aad6253.
61. Coyle, M., Hu, J., and Gartner, Z. (2016) Mysteries in a minimal genome, *ACS Cent. Sci.*, **2**, 274-277, doi: 10.1021/acscentsci.6b00110.
62. Moran, L. (2008) *In the Words of Sydney Brenner*, Sandwalk: Strolling with a skeptical biochemist, URL: https://sandwalk.blogspot.com/2008/09/in-words-of-sydney-brenner.html.
63. Brenner, S. (2000) Biochemistry strikes back, *Trends Biochem. Sci.*, **25**, 584.

NON-HAPPENED INTERVIEW WITH SYDNEY BRENNER: TRANSFORMING DATA INTO KNOWLEDGE, BIOINFORMATICS, BIG DATA, AND … "IS WATER H_2O?"

L. G. Kondratyeva[1,2]*, M. V. Patrushev[1], and E. D. Sverdlov[1]*

[1] *National Research Center Kurchatov Institute, 123182 Moscow, Russia; E-mail: liakondratyeva@yandex.ru, edsverd@gmail.com*

[2] *Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, 117997 Moscow, Russia*

The review is an attempt to explain some of the challenges associated with the efforts to understand the mechanisms of organisms functioning, in particular, using Big Data collections. This review is a fictional interview with one of the brightest figures of the golden era of modern molecular genetics and biology, a unique scientist and a philosopher, the Nobel prize winner Sydney Brenner, who among other things introduced a remarkable organism – a transparent roundworm *C. elegans*. His reflections and conclusions regarding the inevitable "conflict" between rapidly growing data sets (Big Data) accumulated with use of the next generation sequencing technologies, and fundamental "taboos" arising due to complex interactions in organisms generating unpredictable "emerging" properties explain unsolvable problems faced by such modern trends as "systems biology". On the other hand, Big Data itself suffers from serious shortcomings such as hidden errors and fundamentally low reproducibility. Another possibly insurmountable barrier facing Big Data is data incompleteness ($n \neq$ all). An example is two small best-studied organisms, *E. coli* (1600 genes, that is, 34.6% of 4623 unique genes have unknown functions) and *C. elegans*, with proteins identified for only about 50% of genes. Another striking example is an "artificial" bacterium, JCVI-syn3.0, with a minimal set of genes in its genome. Out of its 473 genes, biological function could not be assigned to 149 (31.5%). Brenner points out that converting data into knowledge is a major challenge for future biological research and that biology urgently needs a strong theorethical basis. He considers the cell to be the correct level of research and proposes the CELLMAP project as a system for organizing biological information. As a completely honest scientist, he says: If I knew [how to do it], I would do it, and not write about the problem. Understanding how to do this transformation remains the main problem of the biological sciences.

Keywords: bioinformatics, Big Data, genome, system biology, interview, Sydney Brenner