ПРИЛОЖЕНИЕ

ЧАСТЬ 1. ЧТО ТАКОЕ БИОИНФОРМАТИКА

Биоинформатика, основываясь на признанных научных изданиях. «Биоинформатика состоит из применения математических, статистических и вычислительных методов к данным молекулярной биологии, что позволяет связывать их и выполнять прогнозы и заключения» (цитируется из [1]). Это в основном связано с данными о последовательностях ДНК, РНК и аминокислот, полученными в омике (геномика, протеомика, транскриптомика), особенностями, полученными из этих последовательностей, или всесторонним анализом метаболитов в биологическом образце (например, метаболомика [2]). Целями исследования являются классификация и прогнозирование функций и структур белков, анализ прогнозирования и экспрессии генов, а также путей метаболизма и их регуляции.

Также данные включают клинические отчёты, выявление и анализ механизмов заболеваний. Передовые методы секвенирования привели к значительному увеличению требований к данным и к необходимости применения технологий больших данных в биоинформатике [1]. В области биоинформатики разработаны и применяются многочисленные методы и программное обеспечение для хранения, организации, понимания и интерпретации экспоненциально растущего количества биологических данных, направленных на решение проблем медицины и биологии [3].

В результате огромных усилий по сбору данных исследователи надеются создать концептуальные знания.

ЧАСТЬ 2. КАЧЕСТВО ДАННЫХ

Предлагалось множество различных типов измерений для определения и оценки качества данных [4, 5]. Шесть измерений качества данных, предложенные Международной ассоциацией управления данными (International Data Management Association, DAMA), перечислены ниже.

1. Полнота (Completeness): доля хранимых данных относительно потенциальной «100%-ной полноты»; нет упущенных данных. Это создаёт уверенность в надёжности использования данных. В случае неполных данных корректный результат не может быть получен.

2. Уникальность (Uniqueness): ничто не будет записано более одного раза в зависимости от того, как эта вещь идентифицирована.

3. Своевременность (Timeliness): насколько современны данные. *Данные 1980 г. не так современны, как данные 2021 г.*

4. Действительность (Validity): данные действительны, если они соответствуют синтаксису (формату, типу, диапазону) своего определения.

5. Точность (Accuracy): степень, в которой данные правильно описывают объект или событие «реального мира».

6. Согласованность (Consistency): отсутствие различий при сравнении двух или более представлений объекта со стандартным определением.

Иногда к этому добавляют ещё факторы.

Надёжность (Reliability). Надёжность можно определить как данные, которые достаточно надёжны для использования. Например, данные недостаточно надёжны, если они состоят из двух разных записей под одним и тем же именем в двух разных местах.

Актуальность (Relevance). Релевантность можно определить как уместность данной информации в данных. Например, чтобы иметь возможность голосовать на выборах, возраст гражданина является важной информацией, а его религия или пол – нет.

Термин «Big Data» также часто характеризуется тремя составляющими: объём (размер доступных данных растёт с возрастающей скоростью; петабайтовые

наборы данных являются обычным явлением в наши дни, и экзабайтные байты находятся недалеко), скорость (данные сейчас – потоковая передача на сервер в режиме реального времени непрерывным образом, и результат полезен только в том случае, если задержка очень короткая) и разнообразие (количество типов данных) [6]. Однако многие считают, что эти три составляющие недостаточно характеризуют большие данные, и они часто дополняются буквами «V», такими как ценность, жизнеспособность, изменчивость, визуализация и достоверность (value, vitality, variability, visualization и validity; см. недавние обзоры [1, 7, 8]). Другие используют различные квалификации для характеристики больших данных, такие как исчерпывающая, релятивность, расширяемость и масштабируемость. Несмотря на многочисленные попытки, до сих пор нет единого мнения по поводу термина [7, 8].

ЧАСТЬ 3. НЕКОРРЕКТНОЕ ПРЕУВЕЛИЧЕНИЕ/ПРЕУМЕНЬШЕНИЕ РОЛИ ОДНИХ ФАКТОРОВ ПО СРАВНЕНИЮ С ДРУГИМИ, ОБЪЕКТИВНО РАВНОЦЕННЫМИ (ФАВОРИТИЗМ, BIASES), ВЛИЯЮЩЕЕ НА КАЧЕСТВО ДАННЫХ

Многочисленные искажения приводят к направленному отклонению выборки для анализа от случайно выбранной (см. также [9]).

Существуют различные причины предвзятости в опубликованных данных. Способы получения образцов приводят к большому количеству систематических ошибок [10].

Пристрастия авторов. «Ловля значимости» (Fishing for significance). Проблема «ловли значимости» усугубляется так называемым «предвзятым отношением к публикации»: у положительных результатов гораздо больше шансов быть опубликованными, чем у отрицательных [11, 12, 13, 14]. В результате некоторые авторы сообщают только о некоторых результатах, обычно наиболее благоприятных. Следовательно, опубликованные результаты могут систематически отличаться от тех, которые остаются неопубликованными [15–17]. «Сложность публикации отрицательных результатов побуждает авторов находить что-то положительное в своих исследованиях, выполняя многочисленные анализы, пока один из них случайно не даст положительные результаты» [18–20]. «Нас привлекает, иногда непреодолимо, принятие сенсационных положительных результатов и склонность отвергать отрицательные, которые могут быть столь же важными… исследования со статистически значимыми результатами не только с большей вероятностью будут опубликованы, но и с большей вероятностью будут цитироваться и продвигаться, … статистикой можно злоупотреблять, манипулировать данными и анализировать до тех пор, пока не будут извлечены значимые эффекты. Эта форма неправомерного поведения в науке на удивление распространена и известна как p-hacking, или рыбалка с *p*-значением» [20].

Предубеждения экспертов. Давление существующих парадигм. В 1960-х гг. Томас Кун описал, как учёные предпочитают работать в рамках существующей доминирующей парадигмы [21]. Существует сильная оппозиция идеям, которые не

согласуются с традиционными догмами. Только когда придёт время, когда неопровержимые доказательства в поддержку новых идей больше нельзя будет игнорировать, у них появится шанс быть опубликованными. Кун описал этот критический момент времени как «смену парадигмы». Примеры, подтверждающие такое поведение, описаны в литературе [22]. Выдающиеся примеры влияния парадигмы на публикации были описаны Campanario [23]: 19 будущих лауреатов Нобелевской премии столкнулись с сопротивлением со стороны научного сообщества своим открытиям, и случаи, когда 24 будущих лауреата Нобелевской премии столкнулись с сопротивлением со стороны редакторов научных журналов или рецензентов рукописей, посвящённых открытиям, которые впоследствии принесли им Нобелевскую премию. В ответ на эти утверждения журнал «*Nature*» признаёт, что «… в нашей истории есть неоспоримые оплошности. К ним относятся отказ публикациям излучения Черенкова, мезона Хидеки Юкавы, работа Иоганна Дайзенхофера, Роберта Хубера и Хартмута Мишеля по фотосинтезу, а также первоначальное отклонение (но в конечном итоге принятие) излучения черной дыры Стивена Хокинга [24].

Адаптерную гипотезу выдвинул Френсис Крик. Он предположил с очень небольшим количеством доказательств, что существуют короткие РНК. Гипотеза адаптера, о которой он говорил в Лондоне в 1958 г., была отвергнута всеми биохимиками. Они сказали, что, если бы было 20 РНК и 20 ферментов для связывания аминокислоты с каждым адаптером, мы, биохимики, уже должны были бы это открыть. Но поскольку мы их не открыли, их не существует. Однако не потребовалось много времени, чтобы показать, что они ошибались. Эти ферменты и все тРНК были вскоре открыты [25].

Собственные предубеждения редакторов и рецензентов. Эксперты могут быть предвзяты по разным причинам. Рецензенты могут предвзято относиться к своей оценке, если содержание рукописи противоречит их собственному или общепринятому мнению [26]. Ряд международных комитетов созданы для продвижения передовой практики в публикации медицинских журналов; в их число входят Комитет по этике публикаций (COPE), Совет научных редакторов (CSE), Международное общество профессионалов в области медицинских

публикаций (ISMPP) и другие. Однако международный опрос редакторов научных журналов показал, что многие редакторы не знакомы с имеющимися руководящими принципами [22, 26]. Ряд мотивов, которые могут способствовать предвзятости экспертов, подробно рассмотрен в [26]. Предвзятость, внесённая рецензентами, кажется, получает меньше внимания, чем заслуживает [22, 26].

Кроме того, существует острая проблема, связанная с использованием организации и способов обработки данных особенно в эпоху Big Data. Марк Твен предложил: «Сначала соберите факты, а потом вы сможете их сколько угодно искажать». Действительно, как заметил Gaudet [27] в отношении Gene Ontology (GO), центральной базы данных для функциональной генетики: «Неправильные представления и вводящие в заблуждения, которые обычно делаются в отношении GO, включая эффект неполноты данных, важность квалификаторов аннотаций и транзитивность или её отсутствие, связанные с другими онтологиями».

ЧАСТЬ 4. ЦИТАТЫ

Actually, the orgy of fact extraction in which everybody is currently engaged has, like most consumer economies, accumulated a vast debt. This is a debt of theory, and some of us are soon going to have an exciting time paying it back – with interest, I hope.

I was asked by a student what ethical standards should be adopted by life scientists. I could immediately think of two prescriptions. The first, common to all scientists, is to tell the truth. The second is to stand up for all humanity.

The attitude of my generation that all problems can be solved in the next decade, and should be solved in the next decade – these expectations are changed. Maybe science should be done better, but more slowly. I think a large number of mediocre people are in science today, and carried along by the system. General concepts are rare. Nobody publishes theory in biology – with few exceptions. Instead they get out the structure of still another protein. I'm not saying it's mindless. But the mind only acts on the day-to-day.

There is a strong and widely held belief that all organisms are perfect and that everything within them is there for a function. Believers ascribe to the Darwinian natural selection process a fastidious prescience that it cannot possible have and some go so far as to think that patently useless features of existing organisms are there as an investment for the future ...

Even today, long after the discovery of repetitive sequences and introns, pointing out that 25% of our genome consists of millions of copies of one boring sequence, fails to move audiences. They are all convinced by the argument that if this DNA were totally useless, natural selection would already have removed it. Consequently, it must have a function that still remains to be discovered. Some think that it could even be there for evolution of the future – that is, to allow the creation of new genes. As this was done in the past, they argue, why not in the future? ...

Some years ago I noticed that there are two kinds of rubbish in the world and that most languages have different words to distinguish them. There is the rubbish we keep, which is junk, and the rubbish we throw away, which is garbage. The excess DNA in our genomes is junk, and it is there because it is harmless, as well as being useless, and because the molecular processes generating extra DNA outpace those getting rid of it. Were the

extra DNA to become disadvantageous, it would become subject to selection, just as junk that takes up too much space, or is beginning to smell, is instantly converted to garbage.

...we need to put everything into an evolutionary framework, simply because complexity arises in biological systems by accretion and modification and not by reinvention. Thus, the properties of many of the components in our cells, whether these are mRNAs or proteins, will be conditioned not only by processes of selection for specified activities and levels because these are positively required but may also take up any value because there are no negative consequences for the organism. This 'don't care' condition will almost certainly be present because it is a cheap solution to the regulation problem of complex systems. Thus a 20% or a twofold increase, or indeed the very presence, of a protein may be very significant or totally irrelevant depending on whether it is following a 'don't care' condition. Only experiment can decide that.

I once made the remark that two things disappeared in 1990: one was communism, the other was biochemistry and that only one of these should be allowed to come back. Of course, biochemistry never really went away but continued to flourish in the thousands of unread pages of biochemical journals. Protein interactions will not be solved by proteomics or protein chips but by protein biochemistry. The genome sequences tell us about the proteins we can expect to find in cells and give us the tools to make large amounts of the proteins for reconstitution studies and for detailed structural analysis. We do not have to resurrect biochemistry, and it will flourish because it provides the only experimental basis for causal understanding of biological mechanisms. That is why this article is not called 'The return of biochemistry.' (from *"Biochemistry Strikes Back"* [28]) 25 сентября 2008 года «По словам Синднея Бреннера» из блога Ларри Морана «Песчаная дорожка». Прогуливаясь со скептически настроенным биохимиком. (Larry Moran, Sandwalk. Strolling with a skeptical biochemist «In the Words of Sydney Brenner») [29].

It is naive to think that if a species' environment changes the species must adapt or else become extinct.... Just as a changed environment need not set in motion selection for new adaptations, new adaptations may evolve in an unchanging environment if new mutations arise that are superior to any pre-existing. Сидней Бреннер «Биохимия наносит ответный удар» [28].

An atheist before Darwin could have said, following Hume: I have no explanation for complex biological design. All I know is that God isn't a good explanation, so we must wait and hope that somebody comes up with a better one. I can't help feeling that such a position, though logically sound, would have left one feeling pretty unsatisfied, and that although atheism might have been logically tenable before Darwin, Darwin made it possible to be an intellectually fulfilled atheist. Ричард Докинз «Слепой часовщик» [30].

As was predicted at the beginning of the Human Genome Project, getting the sequence will be the easy part as only technical issues are involved. The hard part will be finding out what it means, because this poses intellectual problems of how to understand the participation of the genes in the functions of living. Сидней Бреннер «Loose ends» [31, 32].

The modern computer hovers between the obsolescent and the nonexistent. «Цитаты Сиднея Бреннера» [32].

Progress in science depends on new techniques, new discoveries and new ideas, probably in that order. «Цитаты Сиднея Бреннера» [32, 33].

I think one of the things about creativity is not to be afraid of saying the wrong thing. «Цитаты Сиднея Бреннера» [32].

Whereas Mathematics is the art of the perfect. Physics is the art of the optimal. Biology, because of evolution, is only the art of the satisfactory. Сидней Бреннер «Последовательности и последствия» [34].

A lot of the things that have been accomplished in science have been accomplished on the basis of ignorance … in the sense that you import into the science people from outside. Because once you have an established science, it has got its high priests – the guys who know everything that will work or won't work. And they don't want to be bothered. So you have to have a challenge. And the great thing is that young people are ignorant, and we should catch them before they turn into the priesthood. So I think that science should have a much more daring approach. Из интервью Сиднея Бреннера к 25-летию GenBank [35].

Even God wouldn’t get a grant today because somebody on the committee would say, oh those were very interesting experiments (creating the universe), but they’ve never been repeated. And then someone else would say, yes and he did it a long time ago, what’s

he done recently? And a third would say, to top it all, he published it all in an un-refereed journal. Из интервью Сидней Бреннера для King's Review [36].

I think for the first time we can attack the fundamental biology of man. «Цитаты Сидней Бреннера» [32].

Data should be a means to knowledge, not an end in themselves.

To wit, it's important to get the facts right, but new ideas are useful, as long as they are based on reasonable evidence and are amenable to correction. We need data produced from new technologies to advance understanding. The importance of 'hypothesis-free research' is well established: the philosopher Francis Bacon proposed it as part of his 'empirical method' in 1620. In his book Novum Organum, he argued that the first step in establishing scientific truth should be the description of facts through systematic observations. But this is only the first step. For example, it would have been rather a pity if Darwin had stopped thinking after he had described the shapes and sizes of finch beaks, and had not gone on to propose the idea of evolution by natural selection. The next step is to extract knowledge from the data. To refocus on that goal, we must improve our working processes, placing a greater emphasis on theory and shifting our research culture.

More theory is needed. My exemplars for this include the evolutionary biologists Bill Hamilton and John Maynard Smith, and the geneticists Barbara McClintock and Francis Crick. Their papers are permeated with richly informed biological intuition, which makes them a delight to read. This sort of thinking will accelerate a shift from description to knowledge. Theorists can find fertile ground in considering the flow of information through living systems, which can help them to make better sense of the flood of biological data.

Seeking to be led by theory and knowledge will probably require shifts in research culture. Theorizing should be encouraged, and theories should be included in experimental papers to put data in context. Attempts to do this should not be dismissed by editorial and funding processes as idle speculation. As Darwin said, it allows ideas to be attacked and either dismissed or modified. A sort of 'tyranny of the field' sometimes inhibits the generation of explanations different from the current consensus, but this is a mistake. If the new ideas are not satisfactory, then they will soon be eliminated and progress will be made.

False facts should not be tolerated, but journals and research funders should be open to reasonable new ideas and interpretations, particularly if they differ from the current consensus. Evaluation committees should be tolerant when some of the ideas of people they are considering for promotion or funding are shown to be incorrect. Пол Нерс «Биология должна генерировать идеи, так же как и данные» [37].

The following text is an edited version of a recent interview with Sydney Brenner who has been at the forefront of many developments in molecular biology since the 1950s. It provides a participant's view on current issues in the history and epistemology of molecular biology. The main issue raised by Brenner regards the relation of molecular biology to the new field of systems biology. Brenner defends the original programme of molecular biology – the molecular explanation of living processes – that in his view has yet to be completed. The programme of systems biology in contrast he views as either trivial or as not achievable since it purports to deal with inverse problems that are impossible to solve in complex living systems. Other issues covered in the conversation concern the impact of the human genome sequencing project, the commercial turn in molecular biology and the contested disciplinary status of the science. Из интервью Сорайи де Чадаревян с Сиднеем Бреннером [38].

СПИСОК ЛИТЕРАТУРЫ

1. Navarro, F. C. P., Mohsen, H., Yan, C., Li, S., Gu, M., et al. (2019) Genomics and data science: an application within an umbrella, *Gen. Biol*., **20**, 109, doi: 10.1186/s13059-019-1724-1.
2. 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.
3. Ideker, T., Winslow, L. R., and Lauffenburger, D. A. (2006) Bioengineering and systems biology, *Ann. Biomed. Eng*., **34**, 1226-1233, doi: 10.1007/s10439-006-9119-3.

4. Ramasamy, A., and Chowdhury, S. (2020) Big Data quality dimensions: a systematic literature review, *J. Inform. Systems Technol. Manag.*, **17**, doi: 10.4301/S1807-1775202017003.
5. Soni, S., and Singh, A. (2021) *Improving Data Quality using Big Data Framework: A Proposed Approach*, IOP Publishing.
6. Mayer-Schonberger, V., and Cukier, K. (2014) *Big Data: A Revolution that will Transform How We Live, Work, and Think*, Mariner Books, Houghton Mifflin Harcourt., Boston, MA.
7. Stevens, M., Wehrens, R., and de Bont, A. (2018) Conceptualizations of Big Data and their epistemological claims in healthcare: a discourse analysis, *Big Data and Society*, 1-21, doi: 10.1177/2053951718816727.
8. Helzlsouer, K., Meerzaman, D., Taplin, S., and Dunn, B. K. (2020) Humanizing Big Data: recognizing the human aspect of big data, *Front. Oncol.*, **10**, 186, doi: 10.3389/fonc.2020.00186.
9. Kiran, R. (2020) *Big Data Characteristics: Know the 5'Vs of Big Data*, *edureka*! URL: www.edureka.co/blog/big-data-characteristics/.
10. Danchin, A., Ouzounis, C., Tokuyasu, T., and Zucker, J. D. (2018) No wisdom in the crowd: genome annotation in the era of big data – current status and future prospects, *Microb. Biotechnol.*, **11**, 588-605, doi: 10.1111/1751-7915.13284.
11. Boulesteix, A. L. (2010) Over-optimism in bioinformatics research, *Bioinformatics*, **26**, 437-439, doi: 10.1093/bioinformatics/btp648.
12. Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., and Boulesteix, A. L. (2010) Over-optimism in bioinformatics: an illustration, *Bioinformatics*, **26**, 1990-1998, doi: 10.1093/bioinformatics/btq323.
13. Boulesteix, A. L. (2015) Ten simple rules for reducing overoptimistic reporting in methodological computational research, *PLoS Computat. Biol.*, **11**, e1004191, doi: 10.1371/journal.pcbi.1004191.
14. Jones, G., and Millard, M. (2017) *Big Data = Big Bias? The Fallibility of Big Data*. in *Use Cases in Big Data Software and Analytics* (von Laszewski, G., ed.) Bloomington, Indiana, pp. 189-193.

15. Echevarria, L., Malerba, A., and Arechavala-Gomeza, V. (2020) Researcher's perceptions on publishing "negative" results and open access, *Nucleic Acid Ther*., doi: 10.1089/nat.2020.0865.
16. Ayorinde, A. A., Williams, I., Mannion, R., Song, F., Skrybant, M., et al. (2020) Assessment of publication bias and outcome reporting bias in systematic reviews of health services and delivery research: a meta-epidemiological study, *PLoS One*, **15**, e0227580, doi: 10.1371/journal.pone.0227580.
17. Fanelli, D., Costas, R., and Ioannidis, J. P. (2017) Meta-assessment of bias in science, *Proc. Natl. Acad. Sci. USA*, **114**, 3714-3719, doi: 10.1073/pnas.1618569114.
18. Chase, J. M. (2013) The shadow of bias, *PLoS Biol*., **11**, e1001608, doi: 10.1371/journal.pbio.1001608.
19. Chan, M. E., and Arvey, R. D. (2012) Meta-analysis and the development of knowledge, *Perspect. Psychol. Sci*., **7**, 79-92, doi: 10.1177/1745691611429355.
20. Marin-Franch, I. (2018) Publication bias and the chase for statistical significance, *J. Optometry*, **11**, 67-68, doi: 10.1016/j.optom.2018.03.001.
21. Kuhn, T. (1962) *The Structure of Scientific Revolutions*, University of Chicago Press Chicago.
22. Chalmers, T. C., Frank, C. S., and Reitman, D. (1990) Minimizing the three stages of publication bias, *JAMA*, **263**, 1392-1395.
23. Campanario, J. (2009) Rejecting and resisting Nobel class discoveries: accounts by Nobel Laureates, *Scientometrics*, **81**, 549-565.
24. Editorial (2003) Coping with peer rejection, *Nature*, **425**, 645, doi: 10.1038/425645a.
25. Brenner, S., and Sejnowski, T. (2018) *In the Spirit of Science: Lectures by Sydney Brenner on DNA, worms and brains*, World Scientific.
26. Phillips, J. S. (2011) Expert bias in peer review, *Curr. Med. Res. Opin*., **27**, 2229-2233, doi: 10.1185/03007995.2011.624090.
27. Gaudet, P., and Dessimoz, C. (2017) Gene Ontology: Pitfalls, Biases, and Remedies, in *The Gene Ontology Handbook, Methods in Molecular Biology*

(Dessimoz, C., and Škunca, N., eds.) Springer Open Humana Press, **1446**, 189-205, doi: 10.1007/978-1-4939-3743-1_14.

28. Brenner, S. (2000) Biochemistry strikes back, *Trends Biochem. Sci*., **25**, 584.
29. 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.
30. Dawkins, R. (1996) *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe without Design*, WW Norton & Company.
31. Brenner, S. (2019) *Loose Ends... False Starts*, World Scientific.
32. *Sydney Brenner Quotes*, AZ Quotes, URL: https://www.azquotes.com/author/24376-Sydney_Brenner.
33. Brenner, S. (2002) Life sentences: detective rummage investigates, *Genome biology*, **3**, 1-2,
34. Brenner, S. (2010) Sequences and consequences, *Philos. Trans. R. Soc. Lond. Ser. B. Biol. Sci*., **365**, 207-212, doi: 10.1098/rstb.2009.0221.
35. Brenner, S. (2010) *Sydney Interview on the Genbank 25th Anniversary*, URL: https://www.youtube.com/watch?v=bDm7i3Rc8wU.
36. Brenner, S. (2014) *How academia and publishing are destroying scientific innovation: a conversation with Sydney Brenner.* Kings Review, https://elizabethdzeng.com/2014/02/26/how-academia-and-publishing-are-destroying-scientific-innovation-a-conversation-with-sydney-brenner/.
37. Nurse, P. (2021) Biology must generate ideas as well as data, *Nature*, **597**, 305-305.
38. Brenner, S. (2009) Interview with Sydney Brenner by Soraya de Chadarevian, *Stud. Hist. Philos. Biol. Biomed. Sci*., **40**, 65-71, doi: 10.1016/j.shpsc.2008.12.008.