Quantitative Biology 2021, 9(3): 251–254
https://doi.org/10.15302/J-QB-021-0254

FEATURE

Reflections on the quest to obtain biological information from genomic data

Andrew F. Neuwald*

The Institute for Genome Sciences and Department of Biochemistry & Molecular Biology, University of Maryland School of Medicine, Baltimore, MD 21201, USA
* Correspondence: aneuwald@som.umaryland.edu

Received April 16, 2021; Revised April 17, 2021; Accepted April 18, 2021

The 20th anniversary of the completion of the Human Genome Project offers an opportunity to reflect on early efforts to make biological sense out of genomic data. My interest in genomic analysis began at Washington University in 1988 while I was a postdoc in Doug Berg's laboratory, which was in the same building as the laboratory of Maynard Olson, one of the founders of the Human Genome Project. At the time, my focus was on the molecular biology of *E. coli*—an area of research that had appealed to me due to its being information rich. However, other, emerging areas of research began to offer great opportunities going forward. At an inhouse seminar, Maynard Olson made a visionary statement foreseeing the day when each patient's genome would be sequenced as part of a routine medical exam. After having spent years as a graduate student sequencing a mere 5,000 base pairs using the Maxam and Gilbert method, Maynard's statement initially struck me as incredible, but later as eye opening. The whole genome analysis initiative also raised the question of how to best obtain biological information from the vast amount of sequence data becoming available. This directed me toward computer science and statistics, which were soon to become very critical components of biomedical research.

Fortunately for me, I was accepted as a molecular biology postdoc into Washington University's medical informatics program, which previously consisted entirely of physicians and which involved obtaining a master's degree in computer science. My advisor was Phil Green, who had developed two important computer programs facilitating high-throughput genomic sequencing: PHRED, which calls bases from automated sequencer data and assigns error probabilities; and PHRAP, which assembles sequencing reads to construct genomic segments [1]. Phil added a statistical component to my project, which involved detecting significant patterns in unaligned protein sequences. Since my interest was in analyzing protein sequences rather than DNA, I obtained a third postdoc at the National Center for Biotechnology Information (NCBI) under David Lipman, whose FASTP [2] and BLAST [3] programs had dramatically accelerated protein sequence database searches. Prior to FASTP, I recall waiting a day or more to complete such a search, even though the protein database was extremely small by today's standards. David gave me freedom to pursue my own research interests. His occasional wisecracks were amusing: Concerning hidden Markov models, he once remarked that "there is something hidden and we have this Russian name to help us find it". Anecdotally, in 2017 David, whose father ran a butcher shop, stepped down from directing the NCBI to become the chief science officer at Impossible Foods, which applied genomic knowhow to develop the plant-based Impossible Burger.

My time at the NCBI introduced me to several other important players in protein sequence analysis. This included two outstanding statisticians: Jun S. Liu and Chip Lawrence, who introduced me to Bayesian statistics. Together we developed the first Markov Chain Monte Carlo (MCMC) methods to multiply align protein sequences based on subtle, biologically relevant sequence patterns [4]. Applying these methods led, for example, to the discovery that Barth syndrome, a cardiomyopathic disorder, is due to an acyltransferase deficiency [5], and to the detection, within proteins involved in chromosome-related functions, of subtle HEAT repeats, the presence of

This article is dedicated to the feature in 20th anniversary of Human Genome (Eds. Michael Q. Zhang and Xuegong Zhang).

which in a hypothetical protein led to our discovery of a new condensin component in vertebrates [6]. This also led to my initial characterization of the AAA + ATPase class [7] in collaboration with two remarkable NCBI researchers, Eugene Koonin and his student Aravind.

Both Koonin and Aravind were quick to apply their encyclopedic knowledge of protein biology in conjunction with programs being developed at the NCBI. At first, Eugene's efforts were not well received. To work around inherent bias against computational research, he often published his findings as letters to journal editors. He once remarked how he was referred to as "Mr. Koonin" by one editor, who apparently assumed that such computationally based nonsense could not have been performed by someone with a Ph.D. However, Eugene prevailed to become one of the most respected computational biologists with over 900 publications [8]. The fact that he is a "hunt-and-peck" typist makes his prolific publication record even more remarkable. The power of statistical and computational discoveries were well illustrated by Koonin, Altschul and Peer Bork in a 1996 publication [9] that corrected an erroneous experimentally-based study indicating that the breast cancer protein BRCA1 is secreted and exhibits the properties of a granin. Unfortunately for the authors, their erroneous results had been highlighted as a major breakthrough in the *New York Times* and on the PBS News hour, which presumably made the revelation by Koonin *et al*. quite humiliating.

In 1994, Roman Tatusov, along with Altschul and Koonin, published an iterative search procedure [10] that Lawrence, Lui and I later integrated into an MCMC search and alignment procedure [11] and that Altschul, Lipman (Fig. 1) and others likewise developed into PSI-BLAST [12], one of the most important breakthroughs in protein sequence analysis. Together, the BLAST and PSI-BLAST papers have been cited over 160,000 times and have inspired similar iterative programs, such as Sean Eddy's JackHMMER. Koonin and Aravind used these programs to make hundreds of biological discoveries and to provide feedback to NCBI researchers working on statistical and algorithmic methods. For these reasons, the NCBI was a very rich environment for computational biology.

These early breakthroughs led to a keen interest in hiring "biologists who compute" by various institutions [13], which led to my recruitment in 1997 to Cold Spring Harbor Laboratory (CSHL) to join an emerging computational biology group. At that time pharmaceutical and other biomedical corporations were aggressively competing with academic and research institutions to recruit computational biologists. One company sought (unsuccessfully) to recruit Stanford mathematician Sam Karlin, who was nearly 80 years old at the time and who, with

Figure 1. Four of the seven coauthors on the PSI-BLAST paper: (clockwise from left) Tom Madden, David Lipman, Stephen Altschul, and Alejandro Schäffer. Initially Stephen thought that David's idea for PSI-BLAST was 'half baked', to which Alejandro responded "Then I will provide the oven." Reprinted from the Spring of 1999 NCBI newsletter.

Altschul, worked out the BLAST statistics [14]. However, corporate enthusiasm later abated, perhaps due to over selling the speed with which the field of computational biology would mature and the speed at which practical benefits would emerge. Also contributing to diminishing interest, was the biomedical naiveté of some researchers coming in from other fields. For example, a major pharmaceutical company hired a well-respected computer scientist to head up their bioinformatics group. This scientist sought to parse the information encoded in DNA just as a compiler parses and translates computer code. This failed to pan out, of course, since the "syntax" of DNA is not based upon a formal language. As the field of sequence analysis matured, the trend has been for computer science and statistical experts either to work more closely with biologists or to develop a deep understanding of biology themselves. This is reminiscent of the cross-disciplinary nature of molecular biology, which combined techniques from microbiology, genetics, and biochemistry, or of computer science, which combined electrical engineering with concepts from information theory, discrete mathematics, logic, and linguistics. Likewise, the rise of computational biology has involved better integration of computer science and statistics with diverse biomedical disciplines.

My more recent research, first at CSHL and subsequently at the University of Maryland Institute for Genome Sciences, was inspired by a 1998 review by Bruce Alberts [15], in which he laid out the goal of

understanding how protein molecular machines work at the atomic level in the context of the living cell. Attaining this goal will require further integrating protein sequence analysis with information from biophysical, biochemical, and structural studies, from molecular dynamics simulations, and perhaps from quantum theory. To quote Richard Feynman: "Nature isn't classical … and if you want to make a simulation of nature, you'd better make it quantum mechanical." So protein molecular mechanisms may well involve some quantum weirdness, as indicated by C_{60} fullerene (a large, ball-shaped, pure-carbon molecule) exhibiting the behavior of a wave that can pass simultaneously through multiple slits in a diffraction grid resulting in constructive and destructive interference patterns [16]. The difficulty of investigating quantum phenomena, however, motivates the use of statistics to glean functional clues from protein sequence patterns reflecting underlying (potentially quantum mechanical) properties. Hence, my current research (performed in collaboration with Altschul and several experimental labs) involves using statistics to identify sequence and structural determinants of protein functional specificity [17].

AlphaFold's recent success in predicting protein structures [18] suggests that deep neural networks (DNNs) trained on vast amounts of sequence and other big data may play a major role in furthering our understanding of protein molecular machines. The Universal Approximation Theorem [19] explains why neural networks work: It states that a neural network with a sufficient number of neurons can approximate any continuous mathematical function. Due to such versatility, DNNs should be applicable to a wide range of biological problems—though the complexity of DNNs often obscures precisely what it is that is being modeled, making them difficult to interpret and understand. Perhaps the metabolic, signaling, and regulatory networks within living cells are analogous to neural networks. If so, might the Universal Approximation Theorem explain the ability of biological systems to exhibit the wide range of functional forms observed in nature? And, as for DNNs, will our models of biological systems, even when quite accurate, nevertheless be very difficult to interpret and understand mechanistically?

In the quest to obtain biological information from genomic data, there are, of course, many other key players with whom I have had little interaction and thus have been left out of this narrative. Nevertheless, I hope that this brief article will give readers a sense of the development of this field from my own perspective.

ACKNOWLEDGEMENTS

This work was supported by National Institute of General Medical Sciences grant R01 GM125878.

COMPLIANCE WITH ETHICS GUIDELINES

The author Andrew F. Neuwald declare that he has no competing interests.

REFERENCES

1. Brownlee, C. (2004) Biography of Phil Green. Proc. Natl. Acad. Sci. USA., 101, 13991–13993
2. Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. Science, 227, 1435–1441
3. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410
4. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science, 262, 208–214
5. Neuwald, A. F. (1997) Barth syndrome may be due to an acyltransferase deficiency. Curr. Biol., 7, R462–R466
6. Ono, T., Losada, A., Hirano, M., Myers, M. P., Neuwald, A. F. and Hirano, T. (2003) Differential contributions of condensin I and condensin II to mitotic chromosome architecture in vertebrate cells. Cell, 115, 109–121
7. Neuwald, A. F., Aravind, L., Spouge, J. L. and Koonin, E. V. (1999) AAA +: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. Genome Res, 9, 27–43
8. Gabrielsen, P. (2017) Profile of Eugene V. Koonin. Proc. Natl. Acad. Sci. USA, 114, 793–796
9. Koonin, E. V., Altschul, S. F. and Bork, P. (1996) ... Functional motifs.... Nat. Genet., 13, 266–268
10. Tatusov, R. L., Altschul, S. F. and Koonin, E. V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. USA, 91, 12091–12095
11. Neuwald, A. F., Liu, J. S., Lipman, D. J. and Lawrence, C. E. (1997) Extracting protein alignment models from the sequence database. Nucleic Acids Res., 25, 1665–1677
12. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

Nucleic Acids Res., 25, 3389–3402
13. Marshall, E. (1996) Hot property: biologists who compute. Science, 272, 1730–1732
14. Karlin, S. and Altschul, S. F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. USA, 90, 5873–5877
15. Alberts, B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell, 92, 291–294
16. Arndt, M., Nairz, O., Vos-Andreae, J., Keller, C., van der Zouw, G. and Zeilinger, A. (1999) Wave-particle duality of C_{60} molecules. Nature, 401, 680–682
17. Tondnevis, F., Dudenhausen, E. E., Miller, A. M., McKenna, R., Altschul, S. F., Bloom, L. B. and Neuwald, A. F. (2020) Deep Analysis of Residue Constraints (DARC): identifying determinants of protein functional specificity. Sci. Rep., 10, 1691
18. AlQuraishi, M. (2019) AlphaFold at CASP13. Bioinformatics, 35, 4862–4865
19. Hornik, K. (1991) Approximation capabilities of multilayer feedforward networks. Neural Netw., 4, 251–257