entropy

MDPI

Editorial

Progress in and Opportunities for Applying Information Theory to Computational Biology and Bioinformatics

Alon Bartal *,† **and Kathleen M. Jagodnik** †

The School of Business Administration, Bar-Ilan University, Ramat Gan 5290002, Israel; kathleen.jagodnik@biu.ac.il

* Correspondence: alon.bartal@biu.ac.il

† These authors contributed equally to this work.

This editorial is intended to provide a brief history of the application of Information Theory to the fields of Computational Biology and Bioinformatics; to succinctly summarize the current state of associated research, and open challenges; and to describe the scope of the invited content for this Special Issue of the journal *Entropy* with the theme of "Information Theory in Computational Biology".

Information Theory as a field of research was established with the publication of Claude Shannon's seminal monograph "A Mathematical Theory of Communication" in 1948 [1]. This work introduced concepts including information entropy, mutual information (a term that was later coined by Roberto M. Fano [2]), and the representation of information as binary digits (bits, a term that is credited to John Tukey) [3]. Progressing beyond earlier related work by Harry Nyquist and Ralph Hartley in the 1920s, and by Alan Turing and Norbert Wiener in the 1940s [4,5], Shannon's work describes the fundamental laws of data transmission and compression [6] and the theoretical limits on the efficiency of communicating over noisy channels [7]. As a unifying theory that intersects with many disciplines including Probability, Statistics, and Computer Science [6], Information Theory is applied to study the extraction, transmission, processing, and use of information in a variety of systems. Shannon's concepts, and those inspired by them, underlie modern digital information technology [5].

In the 1960s, improvements in experimental methods, including crystallography, and the rapid expansion of molecular biology methods across the biological subdisciplines, permitted biologists to advance our understanding of a variety of phenomena [8] including the characteristics of the RNA code [9], the structures of proteins [10,11], and the evolution of genes and proteins [10,12–14]. The central dogma of molecular biology [15] was developed following the foundational discoveries of the processes of RNA transcription and translation. With the advent of Computer Science theory and the era of modern computation starting in the 1960s, the application of computational strategies to address biological questions introduced the field of Computational Biology [16]. The early achievements in the application of computational methods to biological questions include computational studies of evolution [17] and protein structures [18], and the development of the first sequence alignment algorithms [19,20].

We note that Computational Biology is sometimes referenced interchangeably with Bioinformatics [21–23], although these disciplines are also often differentiated in various ways. We make the following distinction: Bioinformatics seeks to develop algorithms, databases, software tools, and other computational resources that permit the insightful analysis of biological data, including its acquisition, storage, quantification, annotation, visual exploration, and other forms of processing [23]. A single software-based product of a Bioinformatics project can often be widely applied to address a variety of biological questions. Complementing the scope of Bioinformatics, Computational Biology seeks to

Citation: Bartal, A.; Jagodnik, K.M. Progress in and Opportunities for Applying Information Theory to Computational Biology and Bioinformatics. *Entropy* **2022**, *24*, 925. https://doi.org/10.3390/e24070925

Received: 26 May 2022
Accepted: 30 June 2022
Published: 3 July 2022

answer specific biological questions using computational strategies [23]. Some Computational Biology projects develop algorithms and computational tools to analyze biological data for addressing the question of interest, and many Computational Biology analyses use the tools created by bioinformaticians. The work of many researchers spans both domains. Both of these disciplines have benefited from the application of Information Theory, and accordingly, this Special Issue welcomes submissions involving the application of Information Theory to both Computational Biology and Bioinformatics.

Information Theory has been among the strategies used to advance Computational Biology from the earliest period of the latter discipline's development onward. Initial studies included reports on the informational properties of DNA [24] and protein sequences [11]. Information Theory has been applied at the molecular level to quantify information of DNA binding sites [25], to understand gene regulation and metabolic networks [26], and to study protein–DNA interactions [25] and protein–protein interactions [27]. It has also been used to elucidate biological sequences [28]. Information Theoretical concepts such as Mutual Information have been employed for protein structure prediction, including the identification of co-evolving amino acid residues [29–34]. Signal transmission in cellular systems, which are inherently noisy, has been quantified using Information Theoretical concepts including entropy [27,35,36]. Information Theory has facilitated the effective modeling of non-linear relationships involving biological entities and has contributed to representing biological systems as stochastic processes [37]. This brief editorial only touches on the numerous applications of Information Theory to the fields of Computational Biology and Bioinformatics.

Since the 1990s, significant improvements in sequencing technology, and steady increases in computing power and reductions in the costs of computing, have led to an exponential increase in the generation of biological data [8]. The present era of 'Big Data' requires innovative strategies for data mining, exploration, and management [8], which can be addressed with Information Theory-based strategies. For example, while dimensionality reduction for omics datasets has often used Principal Components Analysis (PCA) [38], an alternative, Information Theory-based method, Independent Component Analysis (ICA), enables the identification of meaningful content using the measure of negentropy [37,39]. The ICA method has the advantage, relative to PCA, that it does not require latent factors to be orthogonal [37]. Additionally, managing massive quantities of biological data has been proposed via the entropy-scaling search, which was shown to dramatically accelerate searches of protein, metagenomic, and chemical data [40,41]. Many opportunities remain to apply Information Theory principles to further improve the management and effective use of biological data.

Recent innovations in Computational Biology and Bioinformatics invite new applications of Information Theory to these disciplines. For example, advances in analyses at the level of single cells have revolutionized many biological fields, as these techniques permit the high-resolution discovery of characteristics that are masked by bulk sampling strategies. Recent developments have started to apply Information Theory to the analysis of single-cell data, including gene expression data. Since single-cell gene expression data is characterized by distinctly different data distribution patterns and other properties [37], compared with bulk samples, new statistical methods are needed, for which Information Theory can be useful. As an example, Chan et al. [42] presented an approach to analyze single-cell gene expression data based on multivariate Information Theory, a strategy that is reported to be more reliable than classical approaches [43]. Additionally, entropy has been used to measure variability and other properties including stemness in single-cell transcriptomic data [44]. However, a significant need remains for more accurate and optimized analysis methods that are specific to single-cell data [37].

A variety of other promising potential applications of Information Theory to Computational Biology and Bioinformatics remain. Multi-omics integration is a strategy that continues to evolve, as new experimental platforms and data types emerge. Information Theory has previously been demonstrated as a useful approach for such an integration [45],

and many opportunities remain to apply these concepts to yield more informative integrative analyses. More broadly, high-dimensional statistical theory for biological applications remains to be advanced, with a need for unifying definitions and interpretations of statistical interactions [37]. Information Theoretical concepts including entropy have facilitated the analysis of complex networks [46], which characterize a variety of biological systems [47,48]. The origins of life remain uncertain, and Information Theory has been demonstrated as a useful tool to address this problem [49]. The complex dynamics of neural information processing remain to be fully elucidated, and Information Theory has previously been applied to address these questions [50,51]. Assessing human physiological and emotional states in more informative and accurate ways can be facilitated with Information Theory concepts, including mutual information [52]. Shannon's classical Information Theory is being advanced toward the use of quantum Information Theory [37] for applications including studying quantum information transfer from DNA to proteins [53], quantum-mechanical modeling of mutations in cancer [54], and error-correction coding in genetics [55]. This editorial provides a brief overview of some key opportunities for advancing Computational Biology and Bioinformatics by applying Information Theory; a more comprehensive review of the progress and open challenges is available in [37].

The goal of this *Entropy* Special Issue is to present a curated collection of expert perspectives on applying Information Theory to Computational Biology and Bioinformatics in diverse contexts. Its areas of research may include, but are not limited to, sequencing, sequence comparison, and error correction; gene expression and transcriptomics; biological networks; omics analyses; genome-wide disease-gene association mapping; and protein sequence, structure, and interaction analysis. Original research manuscripts, review papers, and Perspective/Commentary articles are welcome. The fields of Computational Biology and Bioinformatics, facilitated by continuing improvements in technology, can be advanced in exciting new directions by the thoughtful application of Information Theory principles.

Funding: This research received no external funding.

Acknowledgments: This work was conducted as part of K.M.J.'s post-doctoral fellowship in the School of Business Administration at Bar-Ilan University. We thank the reviewers for their helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gleick, J. *The Information: A History, A Theory, A Flood*; Vintage: New York, NY, USA, 2011.
2. Kreer, J. A question of terminology. *IRE Trans. Inf. Theory* **1957**, *3*, 208. [CrossRef]
3. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
4. Geoghegan, B.D. Historiographic conceptualization of information: A critical survey. *IEEE Ann. Hist. Comput.* **2008**, *30*, 66–81. [CrossRef]
5. Guizzo, E.M. The Essential Message: Claude Shannon and the Making of Information Theory. Ph.D. Thesis, Massachusetts Institute of Technology, Singapore, 2003.
6. Verdu, S. Fifty years of Shannon theory. *IEEE Trans. Inf. Theory* **1998**, *44*, 2057–2078. [CrossRef]
7. Calderbank, A.R. The art of signaling: Fifty years of coding theory. *IEEE Trans. Inf. Theory* **1998**, *44*, 2561–2595. [CrossRef]
8. Gauthier, J.; Vincent, A.T.; Charette, S.J.; Derome, N. A brief history of bioinformatics. *Brief. Bioinform.* **2019**, *20*, 1981–1996. [CrossRef] [PubMed]
9. Nirenberg, M.; Leder, P.; Bernfield, M.; Brimacombe, R.; Trupin, J.; Rottman, F.; O'Neal, C. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc. Natl. Acad. Sci. USA* **1965**, *53*, 1161. [CrossRef]
10. Margoliash, E. Primary structure and evolution of cytochrome C. *Proc. Natl. Acad. Sci. USA* **1963**, *50*, 672–679. [CrossRef] [PubMed]
11. Nolan, C.; Margoliash, E. Comparative aspects of primary structures of proteins. *Annu. Rev. Biochem.* **1968**, *37*, 727–791. [CrossRef]
12. Crick, F.H. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38*, 367–379. [CrossRef]
13. Woese, C.R. On the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* **1965**, *54*, 1546. [CrossRef] [PubMed]
14. Zuckerkandl, E.; Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **1965**, *8*, 357–366. [CrossRef]
15. Crick, F.H.C. Central dogma of molecular biology. *Nature* **1970**, *227*, 561–563. [CrossRef] [PubMed]
16. Ouzounis, C.A.; Valencia, A. Early bioinformatics: The birth of a discipline—A personal view. *Bioinformatics* **2003**, *19*, 2176–2190. [CrossRef]
17. Fitch, W.M.; Margoliash, E. Usefulness of amino acid and nucleotide sequences in evolutionary studies. *Evol. Biol.* **1970**, *4*, 67–109.

18. Krzywicki, A.; Slonimski, P.P. Formal analysis of protein sequences: I. Specific long-range constraints in pair associations of amino acids. *J. Theor. Biol.* **1967**, *17*, 136–158. [CrossRef]
19. Gibbs, A.J.; McIntyre, G.A. The diagram, a method for comparing sequences: Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* **1970**, *16*, 1–11. [CrossRef] [PubMed]
20. Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443–453. [CrossRef]
21. Diniz, W.D.S.; Canduri, F. Bioinformatics: An overview and its applications. *Gen. Mol. Res.* **2017**, *16*. [CrossRef] [PubMed]
22. Tang, B.; Pan, Z.; Yin, K.; Khateeb, A. Recent advances of deep learning in bioinformatics and computational biology. *Front. Genet.* **2019**, *10*, 214. [CrossRef] [PubMed]
23. Tiwary, B.K. Introduction to Bioinformatics and Computational Biology. In *Bioinformatics and Computational Biology*; Springer: Singapore, 2022; pp. 1–9.
24. Gatlin, L.L. The information content of DNA. *J. Theor. Biol.* **1966**, *10*, 281–300. [CrossRef]
25. Schneider, T.D. A brief review of molecular information theory. *Nano Commun. Netw.* **2010**, *1*, 173–180. [CrossRef] [PubMed]
26. Mousavian, Z.; Kavousi, K.; Masoudi-Nejad, A. Information theory in systems biology. Part I: Gene Regulatory and Metabolic Networks. *Semin. Cell Dev. Biol.* **2016**, *51*, 3–13. [CrossRef] [PubMed]
27. Mousavian, Z.; Díaz, J.; Masoudi-Nejad, A. Information Theory in Systems Biology. Part II: Protein–Protein Interaction and Signaling Networks. *Semin. Cell Dev. Biol.* **2016**, *51*, 14–23. [CrossRef] [PubMed]
28. Vinga, S. Information theory applications for biological sequence analysis. *Brief. Bioinform.* **2014**, *15*, 376–389. [CrossRef] [PubMed]
29. Little, D.Y.J. Application of Information Theory to Modeling Exploration and Detecting Protein Coevolution. Ph.D. Thesis, University of California, Berkeley, CA, USA, 2013.
30. Simonetti, F.L.; Teppa, E.; Chernomoretz, A.; Nielsen, M.; Marino Buslje, C. MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Res.* **2013**, *41*, W8–W14. [CrossRef] [PubMed]
31. Carbone, A.; Dib, L. Co-evolution and information signals in biological sequences. *Theor. Comput. Sci.* **2011**, *412*, 2486–2495. [CrossRef]
32. Dunn, S.D.; Wahl, L.M.; Gloor, G.B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **2008**, *24*, 333–340. [CrossRef] [PubMed]
33. Gloor, G.B.; Martin, L.C.; Wahl, L.M.; Dunn, S.D. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* **2005**, *44*, 7156–7165. [CrossRef]
34. Martin, L.C.; Gloor, G.B.; Dunn, S.D.; Wahl, L.M. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **2005**, *21*, 4116–4124. [CrossRef]
35. Uda, S. Application of information theory in systems biology. *Biophys. Rev.* **2020**, *12*, 377–384. [CrossRef] [PubMed]
36. Waltermann, C.; Klipp, E. Information theory based approaches to cellular signaling. *Biochim. Biophys. Acta (BBA)-Gen. Subj.* **2011**, *1810*, 924–932. [CrossRef]
37. Chanda, P.; Costa, E.; Hu, J.; Sukumar, S.; Van Hemert, J.; Walia, R. Information theory in computational biology: Where we stand today. *Entropy* **2020**, *22*, 627. [CrossRef]
38. Pearson, K. Principal components analysis. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *6*, 559. [CrossRef]
39. Comon, P. Independent component analysis, a new concept? *Signal Processing* **1994**, *36*, 287–314. [CrossRef]
40. Ishaq, N.; Student, G.; Daniels, N.M. Clustered hierarchical entropy-scaling search of astronomical and biological data. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 780–789.
41. Yu, Y.W.; Daniels, N.M.; Danko, D.C.; Berger, B. Entropy-scaling search of massive biological data. *Cell Syst.* **2015**, *1*, 130–140. [CrossRef]
42. Chan, T.E.; Stumpf, M.P.; Babtie, A.C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* **2017**, *5*, 251–267. [CrossRef]
43. Stumpf, M.P. Inferring better gene regulation networks from single-cell data. *Curr. Opin. Syst. Biol.* **2021**, *27*, 100342. [CrossRef]
44. Gandrillon, O.; Gaillard, M.; Espinasse, T.; Garnier, N.B.; Dussiau, C.; Kosmider, O.; Sujobert, P. Entropy as a measure of variability and stemness in single-cell transcriptomics. *Curr. Opin. Syst. Biol.* **2021**, *27*, 100348. [CrossRef]
45. Lovino, M.; Randazzo, V.; Ciravegna, G.; Barbiero, P.; Ficarra, E.; Cirrincione, G. A survey on data integration for multi-omics sample clustering. *Neurocomputing* **2022**, *488*, 494–508. [CrossRef]
46. Bersanelli, M.; Mosca, E.; Remondini, D.; Giampieri, E.; Sala, C.; Castellani, G.; Milanesi, L. Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinform.* **2016**, *17*, S15. [CrossRef]
47. Costa, L.D.F.; Rodrigues, F.A.; Cristino, A.S. Complex networks: The key to systems biology. *Genet. Mol. Biol.* **2008**, *31*, 591–601. [CrossRef]
48. Lopes, F.M.; Cesar, R.M., Jr.; Costa, L.D.F. Gene expression complex networks: Synthesis, identification, and analysis. *J. Comput. Biol.* **2011**, *18*, 1353–1367. [CrossRef]
49. Yockey, H.P. Information theory, evolution and the origin of life. *Inf. Sci.* **2002**, *141*, 219–225. [CrossRef]
50. Ball, K.R.; Grant, C.; Mundy, W.R.; Shafer, T.J. A multivariate extension of mutual information for growing neural networks. *Neural Netw.* **2017**, *95*, 29–43. [CrossRef]

51. Coolen, A.C.; Kühn, R.; Sollich, P. *Theory of Neural Information Processing Systems*; Oxford University Press: Oxford, UK, 2005.
52. Li, X.; Song, D.; Zhang, P.; Zhang, Y.; Hou, Y.; Hu, B. Exploring EEG features in cross-subject emotion recognition. *Front. Neurosci.* **2018**, *12*, 162. [CrossRef]
53. Djordjevic, I.B. Quantum Information Theory and Quantum Mechanics-Based Biological Modeling and Biological Channel Capacity Calculation. In *Quantum Biological Information Theory*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 143–195.
54. Djordjevic, I.B. Quantum-Mechanical Modeling of Mutations, Aging, Evolution, Tumor, and Cancer Development. In *Quantum Biological Information Theory*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 197–236.
55. Djordjevic, I.B. Classical and quantum error-correction coding in genetics. In *Quantum Biological Information Theory*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 237–269.