Abstract

From the late 1980s onward, the term “bioinformatics” mostly has been used to refer to computational methods for comparative analysis of genome data. However, the term was originally more widely defined as the study of informatic processes in biotic systems. In this essay, I will trace this early history (from a personal point of view) and I will argue that the original meaning of the term is re-emerging.

Early History: Bioinformatics, a Work Concept

In the beginning of the 1970s, Ben Hesper and I started to use the term “bioinformatics” for the research we wanted to do, defining it as “the study of informatic processes in biotic systems”. (Although several public sources [see below] trace the origin of the term to publications by us that appeared in 1978 [1, 2], in fact we were using it as early as 1970, proposing the definition above in an article in Dutch that is not generally accessible [3].)

It seemed to us that one of the defining properties of life was information processing in its various forms, e.g., information accumulation during evolution, information transmission from DNA to intra- and intercellular processes, and the interpretation of such information at multiple levels. At a minimum, we felt that that information processing could serve as a useful metaphor for...
understanding these processes was the core of bioinformatic research. It itself can be seen as a type of pattern recognition or pattern transformation models; third, to compare the outcome of such models with “real” data; and finally, and most comprised shuffling large tables by hand. At the same time, pattern recognition methods had become with the advent of massive sequencing projects. Indeed, data-driven science was needed in order to begin to understand biological systems as information processing systems. In independent parallel processing, and “turtle geometry” and pattern recognition processing systems, often inspired by biological systems, e.g. neural network models for learning...
Bioinformatics before the Data Deluge

But what could actually been done given the scarcity of data and paucity of computing power?

In fact, many of the basic pattern analysis methods now used in bioinformatics were pioneered in the 1960s (for a nice historical overview see [32]) and further developed in the 1970s. However, with respect to methods and data it was still a matter of everyone for themselves, as no easy exchange was possible. A notable exception was, of course, the work of Dayhoff to make protein sequences available through the yearly printed atlases of protein sequences and structure (from [33] to [34]). Accordingly, we spent much time in developing BIOPAT, an integrated set of supervised and nonsupervised pattern analysis methods, though at the same time we strenuously argued that methods development was NOT what bioinformatics was about.

We used the pattern analysis methods to study both “real” data and data derived from modeling studies. Our questions revolved around relating patterns of variation at different levels of organization. This included a first foray into non-linear genotype/phenotype mapping [35], using the developmental “grammars” introduced by Lindenmayer [36], [37], to demonstrate that the pattern of variation at the level of the genotype (the developmental rules) and at the level of the phenotype (the generated “morphemes”) does not necessarily coincide (as implicitly assumed in phylogenetic studies based on morphological data). We developed cluster analysis methods with iterative character weighting [38] to tease apart intermingled patterns of variation. Thus we could, for example, untangle morphological variation due to lineage differences and due to polyploidy [39]. In hindsight, it is interesting to recall the surprise (and dismay of the editors) when we found that isozyme variation was not correlated with lineage but with climatic conditions [40]. The general expectation was that, the closer to the genome, the closer to the “real” evolutionary relationships.

In the 1970s and 1980s, not only were pattern analysis methods developed, but novel modeling formalisms also were actively explored. Nonlinear systems started to become analyzable due to computer modeling, and new developments, for instance phase plane analysis, bifurcation diagrams, and deterministic chaos, were linked to biological applications (e.g., the logistic growth model is a prototype for deterministic chaos [41]).

Moreover, event-based modeling formalisms were developed; most well-known is the Gillespie algorithm developed for simulating chemical kinetics [41]. Our interests being on information processing and micro-macro transitions (emergent phenomena), we focused on the use and development of modeling formalisms implementing local interactions. Thus, we introduced cellular automata as a modeling formalism in ecology [42] and evolution [43], and developed event-based, individual-oriented (now usually called agent-based) simulation approaches.

Because of the often surprising and counterintuitive results of such models, we emphasized a bottom-up modeling methodology. Instead of designing a model to explain a priori well-defined results, in such a bottom-up modeling methodology known (or assumed) basic interactions are implemented, and the resulting dynamics are analyzed in multiple ways and at multiple levels. If and only if various seemingly unrelated and unforeseen consequences of the model correspond to the modeled system, this gives truly novel insight (and confidence in the model) [44], [45]. To analyze such models, pattern analysis methods can be indispensable to relate the outcome of the models to “real” data. For example, this allowed us to demonstrate that the behavioral patterns, division of labor, and adaptation to the environment observed in bumble bee colonies were emergent properties of local interaction of simple entities that “do what there is to do” [46]–[48].

Data-Driven Bioinformatics

I recall the excitement when, in 1982, the first European Molecular Biology Laboratory sequence tape was delivered. Typing in data (on punch cards) from the Dayhoff atlases was cumbersome, even though many aligned sequences were provided. But what to do with this “mess” of data? Just for fun, we clustered species on nucleotide and dinucleotide content. To our surprise (and actually, dismay), a more or less decent classification emerged! This, in spite of our mantra that "synonymous" mutations are therefore not necessarily neutral. Recently [52], it was inferred that conflicting selection pressures on
Propelled by the exponential increase of sequence data, the term bioinformatics became mainstream in the late 1980s, coming to mean the development and use of computational methods for data management and data analysis of sequence data, protein structure determination, homology-based function prediction, and phylogeny. But the rich insights obtained from the massive sequencing projects, and the related bioinformatic analysis to unravel function and evolution, is not really the “roots of bioinformatics”, but rather the “trunk of bioinformatics”, and not the subject of this article.

Back to the Future

In 2002, I received a surprising e-mail from Oxford University Press: “It appears that you may be responsible for the term ‘bioinformatics’. I am preparing an entry for the word in the Oxford English Dictionary, and in this connection am investigating its history...” This led to our 1978 papers on chaotic dynamics in ecological models [11] and genotype-phenotype mapping in growth models [2] being credited as the source of the term (though, as noted, our usage of it dated back to 1970). But was our definition of bioinformatics as the study of informatic processes in biotic systems at multiple levels just an historical quirk, to be superseded by the common meaning of the term as denoting the development and use of computational methods for comparative analysis of genome data?

The set of fully sequenced genomes (including human) was expanding, and high-throughput “omics” data entered the field, adding new dimensions to data-driven comparative research. Organisms were no longer just a “bag of genes or proteins” but also, e.g., a “bag of transcriptomes”, “a bag of interactomes”, and “a bag of metabolomes”. Integrating these various data is a marvelous opportunity and great challenge for bioinformatics in whatever sense of the word!

Indeed, the insight has again taken hold that organisms are not just a bag full of anything, but rather complex dynamical systems, and that an understanding of their functioning requires dynamical modeling. Under the heading “systems biology”, modeling efforts have been revived, and some of these efforts reflect the problems and dilemmas encountered already in the 1970s. How far can models be simplified and still be relevant? (Recall Einstein’s dictum that “models should be as simple as possible but not more so.”) How can models be sensibly scaled up so as to meet the complexity revealed by the genomic data and still be manageable? As was the case in the 1970s with respect to “whole ecosystem” modeling [57], scaling up to the “whole cell” level appears most feasible for energy flow models [58]–[61], while large-scale kinetic models often suffer from the “parameter curse”. (The parameter curse was known in the 1970s as the “Loch Ness monster syndrome” after the existence of the creature was “proven” through population modeling showing that a large super-predator was apparently missing.) One way out of this dilemma might be to use evolutionary models [62].

Individual-based (agent-based) bottom-up modeling is still rare, but the detailed agent-based models of cell division [63] and locomotion [64] of Odell and coworkers are promising examples. The latter paper contains a nice discussion contrasting such detailed modeling with much simpler models that might equally fit the data (even if possibly for the wrong reasons), stressing that the power of such detailed models is to reveal novel counterintuitive consequences of the modeled interactions, as well as the surprising bonus that if detailed local interactions are modeled, robustness with respect to parameter choice often ensues.

So what about the long-term goals we set for bioinformatics in the 1970s, i.e., what of the “horse part” and the “elephant”? Some progress has been made in modeling morphogenesis in a strict sense (the “horse part”), through the use of cell-based models that incorporate some of the physical properties of cells [65]. In particular, the simple but biophysically reasonable representation of a cell in the CPM modeling formalism [66], [67] allows the scaling up to “computing an organism” [68] (e.g., the life cycle of Dicystostelium [69], [70]). But, as Segel emphasized, “the importance of linking changing gene expression with cell movement means that this achievement (i.e., computing an organism) is not the beginning of the end but rather the end of the beginning” [68]. Indeed, there lies the current challenge.

Constructive models of evolution (“the elephant”) have progressed from studies on the evolutionary consequences of non-linear “physical” genotype-phenotype mapping as exemplified by RNA folding [71]–[74] to the evolved genotype-phenotype mapping in the form of metabolic networks [75], [76], regulatory networks [77]–[80], and chromosome organization [81]–[83], and in “virtual cells” [84], [85]. These models shed light on the evolution of robustness and evolvability, and the interplay between neutrality and selection. Interestingly, the surprisingly large gene content of common ancestors as inferred from phylogenetic analysis of fully sequenced genomes and the major role of gene loss in the differentiation of lineages (cf. [86]) appear to be “normal” features in constructive models of evolution (T. Cuypers and P. Hogeweg, unpublished data, [87]). A general conclusion that can be drawn from these studies is that the multi-level nature of biological systems makes the evolutionary process through mutation and selection “easier” because of self-organization at many levels. However, here again the outstanding challenge is the closer integration of what does evolve in the models to what did
evolve in nature, as gleaned from the bioinformatic analysis of genomic data.

As I am writing this, a video of Nobel laureate Paul Nurse has been posted in the science supplement of the Guardian newspaper [88]. Emphasizing self-organization and the resulting counterintuitive results, he argues that the next “quantum leap” in biology will come through studying information processing in biological systems. I conclude by asserting that, whether bioinformatics in the wider sense of studying information processing in biotic systems is a quirk or a quantum leap, it is certainly a mighty interesting quest!

Acknowledgments

Foremost I thank Ben Hesper for conceiving and developing with me the concept “bioinformatics”. I thank Jaap Heringa for his courage in becoming the first graduate in “bioinformatics” in 1984. I thank Rob de Boer for tackling the challenging complexity of immune systems as information processing systems, as well as all others who helped me develop bioinformatics in whatever sense of the word.

References

1. Hogeweg P, Hesper B (1978) Interactive instruction on population interactions. Comput Biol Med 8: 319–327. View Article • Google Scholar

2. Hogeweg P (1978) Simulating the growth of cellular forms. Simulation 31: 90–96. View Article • Google Scholar

3. Hesper B, Hogeweg P (1970) Bioinformatica: een werkconcept. Kameleon 1(6): 28–29. (In Dutch.) Leiden: Leidse Biologen Club. View Article • Google Scholar

4. Nurse P (2008) Life, logic and information. Nature 454: 424–426. View Article • Google Scholar

5. Szekely M (1980) From deoxyribonucleic acid to protein: transfer of genetic information. Wiley. View Article • Google Scholar

6. Wagner A (2007) From bit to it: How a complex metabolic network transforms information into living matter. BMC Sys Biol 1: 33. View Article • Google Scholar

7. Thomson M, Gunawardena J (2009) Unlimited multistability in multisite phosphorylation systems. Nature 460: 274–277. View Article • Google Scholar

8. Turner B (2002) Cellular memory and the histone code. Cell 111: 285–291. View Article • Google Scholar

9. Von Bertalanffy L (1950) An outline of general system theory. Br J Philos Sci 1: 134–165. View Article • Google Scholar

10. Von Bertalanffy L (1973) General system theory. New York: George Braziller.

11. Waddington CH (1968–1972) Towards a theoretical biology. Volumes 1–4. Edinburgh: Edinburgh University Press.

12. Kauffman S (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22: 437–467. View Article • Google Scholar

13. Huang S, Ingber D (2000) Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. Exp Cell Res 261: 91–103. View Article • Google Scholar

14. Huang S, Eichler G, Bar-Yam Y, Ingber D (2005) Cell fates as high-dimensional attractor states of a complex gene regulatory network. Phys Rev Lett 94: 128701. View Article • Google Scholar
15. Boiteux A, Goldbeter A, Hess B (1975) Control of oscillating glycolysis of yeast by stochastic, periodic, and steady source of substrate: a model and experimental study. Proc Natl Acad Sci U S A 72: 3829–3833. View Article • Google Scholar

16. Goodwin B (1963) Temporal organization in cells: a dynamic theory of cellular control processes. London: Academic Press.

17. Wolpert L (1969) Positional information and the spatial pattern of cellular differentiation. J Theor Biol 25: 1–47. View Article • Google Scholar

18. Griffith J (1968) Mathematics of cellular control processes II. Positive feedback to one gene. J Theor Biol 20: 209–216. View Article • Google Scholar

19. Turing A (1952) The chemical basis of morphogenesis. Philos Trans R Soc Lond B Biol Sci 237: 37. View Article • Google Scholar

20. Crick F (1970) Diffusion in embryogenesis. Nature 225: 420–422. View Article • Google Scholar

21. Rosenblatt F (1962) Principles of neurodynamics: perceptrons and the theory of brain mechanisms. Washington (D.C.): Spartan Books.

22. Minsky M, Papert S (1969) Perceptrons. Cambridge (Massachusetts): MIT Press.

23. Holland J (1975) Adaptation in natural and artificial system: an introduction with application to biology, control and artificial intelligence. Ann Arbor (Michigan): University of Michigan Press.

24. Hewitt C (1977) Viewing control structures as patterns of passing messages. Artificial Intelligence 8: 323–364. View Article • Google Scholar

25. Abelson H, DiSessa A (1986) Turtle geometry: the computer as a medium for exploring mathematics. Cambridge (Massachusetts): MIT Press.

26. Papert S (1993) Mindstorms: children, computers, and powerful ideas. New York: Basic Books.

27. Sneath P, Sokal R (1972) Numerical taxonomy: the principles and practice of numerical classification. San Francisco: Freeman.

28. Lance G, Williams W (1966) A generalized sorting strategy for computer classifications. Nature 212: 218. View Article • Google Scholar

29. Macnaughton-Smith P, Williams W, Dale M, Mockett L (1964) Dissimilarity analysis: a new technique of hierarchical sub-division. Nature 202: 1034–1035. View Article • Google Scholar

30. Hogeweg P (1976) Topics in biological pattern analysis [PhD thesis]. Faculty of Science, University of Utrecht.

31. Rosen R (1983) Dynamical modelling of genetic and epigenetic control. In: Bellmann K, editor. Modelling and simulation of molecular genetic information systems. Berlin: Akademie Verlag. pp. 17–30.

32. Hagen J (2000) The origins of bioinformatics. Nat Rev Genet 1: 231–236. View Article • Google Scholar

33. Dayhoff M, Eck R (1968) Atlas of protein sequence and structure 1967–1968. Maryland (Silver Spring): National Biomedical Research Foundation.

34. Dayhoff M (1978) Atlas of protein sequence and structure. Volume 5. Washington (D.C.): National Biomedical Research Foundation.

35. Hogeweg P, Hesper B (1974) A model study on biomorphological description. Pattern Recognit 6: 165–179. View Article • Google Scholar
36. Lindenmayer A (1968) Mathematical models for cellular interactions in development I. Filaments with one-sided inputs. J Theor Biol 18: 280–299.

37. Lindenmayer A (1968) Mathematical models for cellular interactions in development II. Simple and branching filaments with two-sided inputs. J Theor Biol 18: 300–315.

38. Hogeweg P (1976) Iterative character weighing in numerical taxonomy. Comput Biol Med 6: 199–211.

39. Mastenbroek O, Hogeweg P, Heringa J, Niemann G, van Nigtevecht G, et al. (1984) Isozyme variation in Silene pratensis: a response to different environments. Biochem Syst Ecol 12: 29–36.

40. May R (1974) Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. Science 186: 645–647.

41. Gillespie D (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81: 2340–2361.

42. Hogeweg P (1988) Cellular automata as a paradigm for ecological modeling. Appl Math Comput 27: 81–100.

43. Boerlijst M, Hogeweg P (1991) Spiral wave structure in pre-biotic evolution: hypercycles stable against parasites. Physica D: Nonlinear Phenomena 48: 17–28.

44. Hogeweg P, Hesper B (1986) Knowledge seeking in variable structure models. In: Elzas MS, Oren TI, Zeigler P, editors. Simulation in the artificial intelligence era. Amsterdam: North Holland. pp. 227–243.

45. Hogeweg P, Hesper B (1989) An adaptive, selfmodifying, non goal directed modelling methodology. In: Elzas MS, Oren TI, Zeigler BP, editors. Knowledge systems paradigms. Amsterdam: North Holland. pp. 77–92.

46. Honk C, Hogeweg P (1981) The ontogeny of the social structure in a captive Bombus terrestris colony. Behav Ecol Sociobiol 9: 111–119.

47. Hogeweg P, Hesper B (1983) The ontogeny of the interaction structure in bumble bee colonies: a MIRROR model. Behav Ecol Sociobiol 12: 271–283.

48. Hogeweg P, Hesper B (1985) Socioinformatic processes: MIRROR modelling methodology. J Theor Biol 113: 311–330.

49. Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J Mol Evol 20: 175–186.

50. Konings D, Hogeweg P, Hesper B (1987) Evolution of the primary and secondary structures of the E1a mRNAs of the adenovirus. Mol Biol Evol 4: 300–314.

51. Huynen M, Konings D, Hogeweg P (1992) Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. J Mol Evol 34: 280–291.

52. Stoletzki N (2008) Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. BMC Evol Biol 8: 224.
53. Konings D, Van Duijn L, Voorma H, Hogeweg P (1987) Minimal energy foldings of eukaryotic mRNAs form a separate leader domain. J Theor Biol 127: 63–78.
View Article • Google Scholar

54. Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361: 13–37.
View Article • Google Scholar

55. Gu W, Zhou T, Wilke C (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol 6: e1000664.
View Article • Google Scholar

56. Kertesz M, Wan Y, Mazor E, Rinn J, Nutter R, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature 467: 103–107.
View Article • Google Scholar

57. Odum EP (1968) Energy flow in ecosystems: a historical review. Integr Comp Biol 8: 11–18.
View Article • Google Scholar

58. Varma A, Palsson B (1994) Metabolic flux balancing: basic concepts, scientific and practical use. Nat Biotechnol 12: 994–998.
View Article • Google Scholar

59. Covert M, Knight E, Reed J, Herrgard M, Palsson B (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429: 92–96.
View Article • Google Scholar

60. Pál C, Papp B, Lercher M, Csermely P, Oliver S, et al. (2006) Chance and necessity in the evolution of minimal metabolic networks. Nature 440: 667–670.
View Article • Google Scholar

61. Freilich S, Kreimer A, Borenstein E, Gophna U, Sharan R, et al. (2010) Decoupling environment-dependent and independent genetic robustness across bacterial species. PLoS Comp Biol 6: e1000690.
View Article • Google Scholar

62. Van Hoek M, Hogeweg P (2006) In silico evolved lac operons exhibit bistability for artificial inducers, but not for lactose. Biophys J 91: 2833–2843.
View Article • Google Scholar

63. Odell G, Foe V (2008) An agent-based model contrasts opposite effects of dynamic and stable microtubules on cleavage furrow positioning. J Cell Biol 183: 471–483.
View Article • Google Scholar

64. Rafelski S, Alberts J, Odell G, Goodson H (2009) An experimental and computational study of the effect of ActA polarity on the speed of Listeria monocytogenes actin-based motility. PLoS Comput Biol 5: e1000434.
View Article • Google Scholar

65. Anderson A, Chaplain M, Rejniak K, Fozard J (2008) Single-cell-based models in biology and medicine. Basel: Birkhauser Verlag.

66. Graner F, Glazier J (1992) Simulation of biological cell sorting using a two-dimensional extended Potts model. Phys Rev Lett 69: 2013–2016.
View Article • Google Scholar

67. Marée A, Grieneisen V, Hogeweg P (2007) The Cellular Potts Model and biophysical properties of cells, tissues and morphogenesis. In: Anderson A, Rejniak K, editors. Single-cell-based models in biology and medicine. Basel: Birkhauser Verlag. pp. 107–136.

68. Segel L (2001) Computing an organism. Proc Natl Acad Sci U S A 98: 3639–3640.
View Article • Google Scholar

69. Savill N, Hogeweg P (1997) Modelling morphogenesis: from single cells to crawling slugs. J Theor Biol 184: 229–235.
View Article • Google Scholar
70. Marée A, Hogeweg P (2001) How amoeboids self-organize into a fruiting body: multicellular coordination in Dictyostelium discoideum. Proc Natl Acad Sci U S A 98: 3879–3883. View Article • Google Scholar

71. Schuster P, Fontana W, Stadler P, Hofacker I (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci 255: 279–284. View Article • Google Scholar

72. Huynen M, Stadler P, Fontana W (1996) Smoothness within ruggedness: the role of neutrality in adaptation. Proc Natl Acad Sci U S A 93: 397–401. View Article • Google Scholar

73. van Nimwegen E, Crutchfield J, Huynen M (1999) Neutral evolution of mutational robustness. Proc Natl Acad Sci U S A 96: 9716–9720. View Article • Google Scholar

74. Huynen M (1996) Exploring phenotype space through neutral evolution. J Mol Evol 43: 165–169. View Article • Google Scholar

75. Kacser H, Beeby R (1984) Evolution of catalytic proteins or on the origin of enzyme species by means of natural selection. J Mol Evol 20: 38–51. View Article • Google Scholar

76. Soyer O, Pfeiffer T (2010) Evolution under fluctuating environments explains observed robustness in metabolic networks. PLoS Comput Biol 6: e1000907. View Article • Google Scholar

77. Crombach A, Hogeweg P (2008) Evolution of evolvability in gene regulatory networks. PLoS Comput Biol 4: e1000112. View Article • Google Scholar

78. Draghi J, Wagner G (2009) The evolutionary dynamics of evolvability in a gene network model. J Evol Biol 22: 599–611. View Article • Google Scholar

79. Wagner A (2008) Robustness and evolvability: a paradox resolved. Proc Biol Sci 275: 91–100. View Article • Google Scholar

80. Draghi J, Parsons T, Wagner G, Plotkin J (2010) Mutational robustness can facilitate adaptation. Nature 463: 353–355. View Article • Google Scholar

81. Crombach A, Hogeweg P (2007) Chromosome rearrangements and the evolution of genome structuring and adaptability. Mol Biol Evol 24: 1130–1139. View Article • Google Scholar

82. Hurst L, Pál C, Lercher M (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5: 299–310. View Article • Google Scholar

83. Batada N, Hurst L (2007) Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet 39: 945–949. View Article • Google Scholar

84. Neyfakh A, Baranova N, Mizrokhi L (2006) A system for studying evolution of life-like virtual organisms. Biol Direct 1: 23. View Article • Google Scholar

85. Goldstein R, Soyer O (2008) Evolution of taxis responses in virtual bacteria: non-adaptive dynamics. PLoS Comput Biol 4: e1000084. View Article • Google Scholar

86. Koonin E (2007) The Biological Big Bang model for the major transitions in evolution. Biol Direct 2: 21. View Article • Google Scholar
87. de Boer FK, Hogeweg P (2010) Eco-evolutionary dynamics, coding structure and the information threshold. BMC Evol Biol 10: 361.

88. The Guardian (12 November 2010) Sir Paul Nurse: organisms are information networks http://www.guardian.co.uk/science/video/2010/nov/05/paul-nurse-life-information-networks[video]. Accessed 28 February 2011.