Ten Thousand Views of Bioinformatics: A Bibliome Perspective

I. Kohane
Center for Biomedical Informatics, Harvard Medical School, Boston, Massachussetts, USA

Summary

Objective: Summarize the current state bioinformatics research from the published literature in 2008.

Methods: The entire corpus of publications indexed by the National Library of Medicine in the PubMed repository was reviewed for articles tagged as belonging to the discipline of bioinformatics by Medical Subject Heading or by term in the title or abstract of the article. Selected summary statistics of this corpus were then used to motivate additional exploration.

Results: Over ten thousand articles published in 2008 populated the bioinformatics corpus. Significantly, there were at least as many publications in genomics and genetics that used computational techniques but that were not identified as bioinformatics research. Genomics and proteomics continued to be the leading application domains of bioinformatics research but despite the proliferation of human studies, the genes most studied in the corpus were from yeast rather than the human organism. The growth in the genomic studies of human disease was accompanied by a growing critical literature regarding the methods, results and impact of these studies. Concurrently, the availability of full genome sequences at commodity prices has increased the computational challenges of human studies by several orders of magnitude. Further concerns were raised about the consequences of public disclosure of comprehensive or even aggregate genomic data.

Conclusion: The impressive size of the bioinformatics bibliome is easily dwarfed by the challenges generated by the continued increased growth of high-throughput biological data sets. The demand for bioinformatics expertise and tools is therefore likely to continue to increase, at least in the near term.

Keywords

Bioinformatics, publications, genomics, proteomics, computational biology

Yearb Med Inform 2009:113-6

Evidence based medicine is shifting from being primarily based on the synthesis of expert opinions of medical practitioners and increasingly towards the meta-analysis of primary data. In that spirit of evidence based analysis, this review takes its cue from the raw numbers of the bibliome, the collective output of peer review by medical literature, for the year 2008. In that perspective, bioinformatics is an extremely vibrant and active discipline. Although the numbers will change as late arriving publications are compiled into the master lists of the National Library of Medicine, the current totals for publications of 2008 that included a mention of bioinformatics in their abstract or title or medical subject heading sum to 10,169. This prodigious output included publications in 1,478 journals by 39,003 authors. The journals that hosted the most of the publications were Bioinformatics with 751 articles on bioinformatics and Nucleic Acid Research with 353 articles on bioinformatics. The top most topics cooccurring with bioinformatics in these ten thousand publications were computational biology, genomics and proteomics. Algorithms, proteins, and software were the next 3 common occurring subjects.

Not surprisingly then, it appears that genomics and proteomics are the chief application domains of bioinformatics and that these application domains take precedence, as measured by publication volume, over algorithms and software. Much like the Sherlock Holmes story “Silver Blaze” by Sir Arthur Conan Doyle, in which the fact that a dog did not bark provided the detective with his insights to solve a case, it is most revealing that there are thousands of publications in genomics, proteomics and metabolomics, for example that employ a variety of computational techniques but do not mention bioinformatics in either of their title or abstract. Nor are they codified by the expert bibliome taxonomists at the National Library of Medicine with the Medical Subject Heading (MeSH) of “bioinformatics”. Specifically there are 6,786 publications on genomics that do not fall within the 2008 corpus of publications in bioinformatics and likewise there are 5,370 articles on proteomics that are similarly not classified as belonging to that 10,169. From an optimistic vantage point, this is the reflection of the outstanding success of bioinformatics. The techniques of computer science and biostatistics as applied to the fields of biological research may be so deeply ingrained into the culture of these domains of investigation that the use of such techniques are not seen as relevant tags for by the MeSH coders of the National Library of Medicine nor by the authors writing the abstracts and providing key words associated with their publication. It should nonetheless lend us pause as to whether we are sufficiently supporting the discipline of bioinformatics because it has now become so routine in a broad array of biological investigations that it frequently goes unacknowledged as a component of those efforts. If we were not still grossly lacking in bioinformatics enabled investigators [1] this would be a relatively minor point but if we are making the case to the private and public funders of the value of the bioinformatics community, this lack of acknowledgement may slow down the institutionalization of the discipline (e.g. the creation of department of bioinformatics, training programs in bioinformatics, graduate programs in bioinformatics). Indeed, it is this perspective that may have motivated Lincoln Stein to write an opinion piece [2] in follow-up to his famous pronouncement in 2003 that "predicted that bioinformatics as a discipline separate from mainstream biology would be gone in ten years..." He reviews his original pronouncement, often reported as "bioinformatics is dead," and argues that he was only half right. For example, he notes, in 1998 there were only 10 bioinformatics degree granting programs listed for the entire USA whereas 10 years later, there are "at least 74 such programs in the United States and Canada, and 150 worldwide." He does report that, as expected, part of the methodological armamentarium of bioinformatics has been adopted by biologists comfortable with both bench and computing and as such that they do not even report or perceive their investigations as being part of the bioinformatics discipline. This may indeed be why, as documented above, so many publications in genomics and proteomics, which do use bioinformatics techniques, are not included in the 2008 bioinformatics bibliome corpus. Nonetheless, a query of GoogleTrends (http://www.google.com/trends?q=genomics%2C+bioinformatics) suggests that even as interest in bioinformatics (as measured however crudely by searches for the term) appears to be waning relative to overall search, so is the interest in related application domains such as genomics. It remains to be seen if this trend represents a merging of what used to be regarded as separate disciplines into a single blended, multidisciplinary toolkit for state-of-the-art, well-trained biologists. The same chart also reveals the continued globalization of bioinformatics. The language most used to search for bioinformatics topics was Korean (English came in second) and the country from which most of the queries came from was India (followed by South Korea).

Within the ten thousand publications, who are the most prolific authors? They are Matthias Mann, Helmut Mayer, and Ruedi Aebersold. Beyond the human interest story in their remarkable productivity, it is also quite revealing that they all largely publish in the area of proteomics. Whether it is an intrinsic property of these authors, biases with which different disciplines identify with bioinformatics or a characteristic of proteomic research remains to be determined. Although 2008 is still too recent to determine which papers have had the most impact, we can determine in this short interval, which have already had the most citations. The top 10 most cited publications [3-12] out of the ten thousand have already been cited 1,899 times as of March 1st, 2009. Two of these describe important and popular bioinformatics knowledge sources [4, 6], two describe particular analytics or methods [5, 8], and three address the applications of molecular and bioinformatics techniques to human disease [3, 11, 12].

As bioinformatics is often associated with the study of genetics, it is revealing that out of the top 120 genes that are studied in these ten thousand bioinformatics publications only 4 of 120 are human genes with the remainder of the 120 consisting of yeast genes. The top two human genes are BRBB2 and TP53, both implicated in a variety of transcriptional control processes but often studied in the context of malignancy. That is, from the very narrow perspective of publications in which there was an explicit identification of bioinformatics as an important component and in which individual genes were studied, genetics looks like mostly like yeast genetics. There are certainly thousands of publications that have studied human disease using genomics or proteomic tools, but the essential requirement for bioinformatics techniques for each of those studies apparently does not rise to the level in which it is explicitly called out in MeSH, title or abstract. What about methodological rather than domain focus? The topmost specific bioinformatics technique listed is "Sequence Alignment" (568 publications) suggesting the continued primacy of genetic sequence in bioinformatics research, certainly reinforced by the massively increasing deluge of sequence data that so-called "next generation sequencing" is generating.

Now that we have gleaned some of what the bibliome tells us about bioinformatics, what can bioinformatics tell us about the bibliome? The number of investigators using natural language processing (NLP) techniques to plumb the literature to cull knowledge about genes, proteins, and their relationship to disease remains small but is growing (55 publications in 2008). Most obviously reflecting these efforts is the BioCreative workshop whose attendees have demonstrated substantially improved performance over prior years [13]. It remains to be seen just how accurate these NLP-driven efforts can be relative to trained human curators [14]. Nonetheless, as the corpus of available biomedical publications increases, accelerated by the NIH-mandated deposit of publications of NIH-funded research and the European mandate for EU-funded research into open access repositories [15, 16] (even as the publishers argue that such mandates are contrary to the best interests of biomedical science [17]) it is likely that these efforts will result in increasingly accurate annotations.

It is not surprising that in the era of Internet-borne social networking, and community-edited encyclopedias, there has arisen an active debate in the bioinformatics community about decentralization of information resources [18]. Of the much larger number of sites that cropped up in 2008, those that were published in the peer-reviewed literature included GENESTAT and the RNA Wiki Project [19, 20]. Just what funding models will be developed to make these sites sustainable in the long term, if funding is required, will be of significance when these are compared to more centralized resources.

The information science around privacy, disclosure and genomics continued to make waves in 2008. Another hole in our sorely misplaced comfort in privacy was found by Homer et al. [21] who shed new light on the disclosure risks of the public posting of even aggregated genetic data. They did so by showing that reporting even of mixtures at enough loci allowed one to identify which arm of a study (e.g. case vs. control) a specific individual had participated it. Shortly after the publication of this "PLoS bomb," multiple sites around the world withdrew their data from the publicly accessible Internet. Cassa et al. [22], added to our sense of collective responsibility by quantifying just how much information regarding your siblings could be gleaned by the disclosure of your genomic data.

The disclosure directly to patients of the potential clinical implications of their genome-wide ascertained genetic variants became much more prevalent in the commercial realm with the increased activity of companies such as Navigenics, 23andme, and DecodeMe. This led to a very vigorous debate in the genomics and bioinformatics, medical and ethics communities regarding the appropriate conduct of such disclosures [23]. It became even more vivid with the announcement of the Personal Genome Project's public disclosure of a percentage of the coding genome (the "exome") as they ramp up to full exome sequencing. Even while first full genome by massively parallel next generation sequencing was published in 2008 [24], Complete Genomics delivered the first full commercially available genome sequence (on a terabyte disk drive delivered in the mail) and is poised to deliver 20,000 more at a cost of $5000 per genome within 18 months. The data analytic challenges of such a pipeline can be glimpsed by the specifications of their data center for 2010: 60,000 processors with 30 petabytes of storage. Pop et al. [25] summarize the current perspective of the particular methodological challenges that will result from analyzing this tsunami of sequence data parceled in very short reads. The maturation of these technologies suggests that the hybridization-based microarrays that were so successfully used for expression profiling will now be used for comprehensive measurements of the transcriptome as well as scans for common and rare genetic variants.

Bioinformaticians have been notoriously efficient at promoting public dissemination of experimental data sets, re-interpreting these and then arguing against the conclusions asserted by original authors. In that spirit: On the one hand, the tide of Genome Wide Association Studies (GWAS) that either replicated findings of prior studies (e.g. in Crohn's disease), implicated new variants in disease or even of non-disease traits (e.g. eye and hair color) continued to rise [26-34]. On the other hand, there were continued vigorous debates regarding the validity of many prior GWAS [35-37], and their clinical relevance (e.g. in pharmacogenomics [38]. With the aforementioned commoditization of genome-wide sequencing we will potentially be able to truly quantify the relative contribution of common and rare variants [39], a necessary step if we are truly to understand the dependence of phenotype and genotype.

With regard to phenotype, there also was increasing awareness that while our methods in analyzing biological data were steadily improving in throughput and quality, there were relatively few such comparable developments in accurately and rapidly characterizing the medical history, current physiological state, environmental exposures, and family history of each subject. The consequences of such deficits have become increasingly clear [40] and the medical informatics community joined forces with the bioinformatics community to address this challenge by, for example, using electronic health records, dissected using NLP techniques, to create efficiently and accurately phenotyped patient populations from the informational byproducts of the healthcare process [41, 42]. The extension of phenotyping to the agents who are most likely to know the subjects' phenotypic details, the subjects themselves, is a likely consequence of the emergence of personally controlled health records as a source of clinical data outside the traditional boundaries of the healthcare establishment [43].

In summary, the bibliome of 2008 appears to disclose what those of us engaged in bioinformatics research have experienced viscerally. With so many new forms of biomedical data generated in ever larger, exponentially growing, quantities the playground for researchers in this discipline has grown richer, more exciting and more challenging. Let's see if we can play some really enjoyable and productive games there in the coming years.

References

1. Donovan S. Big data: teaching must evolve to keep up with advances. Nature 2008;455(7212):461.
2. Stein LD. Bioinformatics: alive and kicking. Genome Biol 2008;9(12):114.
3. Asangani IA, Rasheed SA, Nikolova DA, Leupold JH, Colburn NH, Post S et al. MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer. Oncogene 2008;27(15), 2128-36.
4. Bairoch A, Bougueleret L, Altairac S, Amendolia V, Auchincloss A, Puy GA et al. The Universal Protein Resource (UniProt). Nucleic Acids Research. 2008;36:D190-D5.
5. Bennett-Lovsey RM, Herbert AD, Sternberg MJE, Kelley LA. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins-Structure Function and Bioinformatics 2008;70:611-25.
6. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA; Mouse Genome Database Group. The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res 2008;36:D724-8.
7. De Vos RC, Moco S, Lommen A, Keurentjes JJ, Bino RJ, Hall RD. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nature Protocols 2007;2(4): 778-91.
8. Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. Proteomics 2008;8(4):750-78.

9. Graumann J, Hubner NC, Kim JB, Ko K, Moser M, Kumar C, et al. Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. Mol Cell Proteomics 2008;7(4):672-83.

10. Gwinn DM, Shackelford DB, Egan DF, Mihaylova MM, Mery A, Vasquez DS, et al. AMPK phosphorylation of raptor mediates a metabolic checkpoint. Mol Cell 2008;30(3):214-26.

11. Kawamata N, Ogawa S, Zimmermann M, Kato M, Sanada M, Hemminki K, et al. Molecular allelokaryotyping of pediatric acute lymphoblastic leukemias by high-resolution single nucleotide polymorphism oligonucleotide genomic microarray. Blood 2008;111(2):776-84.

12. Wang WX, Rajeev BW, Stromberg AJ, Ren N, Tang G, Huang Q, et al. The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1. J Neurosci 2008;28(5):1213-23.

13. Smith L, Tanabe LK, Ando RJ, Kuo CJ, Chung IF, Hsu CN, et al. Overview of BioCreative II gene mention recognition. Genome Biol 2008;9 Suppl 2:S2.

14. Alex B, Grover C, Haddow B, Kabadjov M, Klein E, Matthews M, et al. Assisted curation: does text mining really help? Pacific Symposium on Biocomputing Pac Symp Biocomput 2008:556-67.

15. Bloom T, Ferguson C, Gross L, Maccallum CJ, Milton J, Shields R, et al. PLoS Biology at 5: the future is open access. PLoS Biol 2008;6:e267.

16. Cockerill MJ, Norton M. Open-access journals are delivering high impact, and more. Lancet 2008;371: 2084.

17. McMullan, E. Open access mandate threatens dissemination of scientific information. J Neuroophthalmol 2008;28:72-4.

18. Hu JC, Aramayo R, Bolser D, Conway T, Elsik CG, Gribskov M, et al. The emerging world of wikis. Science 2008;320 (5881):1289-90.

19. Ripatti S, Becker T, Bickeböller H, Dominicus A, Fischer C, Humphreys K, et al. GENESTAT: an information portal for design and analysis of genetic association studies. Eur J Hum Genet 2009 Apr;17(4):533-6

20. Daub J, Gardner PP, Tate J, Ramsköld D, Manske M, Scott WG, et al. The RNA WikiProject: community annotation of RNA families. RNA 2008;14:2462-4.

21. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008;4:e1000167.

22. Cassa CA, Schmidt B, Kohane IS, Mandl KD. My sister's keeper?: genomic research and the identifiability of siblings. BMC Med Genomics 2008;1:32.

23. Foster M, Sharp R. Out of sequence: how consumer genomics could displace clinical genetics. Nat Rev Genet2008;9(6):419.

24. Wheeler D, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452(7189):872-6.

25. Pop M, Salzberg S. Bioinformatics challenges of new sequencing technology. Trends in Genetics 2008;24:142-9.

26. Weidinger S, Gieger C, Rodriguez E, Baurecht H, Mempel M, Klopp N, et al. Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus. PLoS Genet 2008;4:e1000166.

27. O'Donovan MC, Norton N, Williams H, Peirce T, Moskvina V, Nikolov I, et al. Analysis of 10 independent samples provides evidence for association between schizophrenia and a SNP flanking fibroblast growth factor receptor 2. Mol Psychiatry 2009;14(1):30-6.

28. van den Oord EJ, Kuo PH, Hartmann AM, Webb BT, Möller HJ, Hettema JM, et al. Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. Arch Gen Psychiatry 2008;65(9):1062-71.

29. Hofmann S, Franke A, Fischer A, Jacobs G, Nothnagel M, Gaede KI, et al. Genome-wide association study identifies ANXA11 as a new susceptibility locus for sarcoidosis. Nat Genet 2008.

30. Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet 2008;4(5):e1000074.

31. Franke A, Balschun T, Karlsen TH, Hedderich J, May S, Lu T, et al. Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. Nat Genet 2008;40:713-5.

32. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci U S A2008; 105:4340-5.

33. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, et al. Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet 2008;40:310-5.

34. van Ommen GJ. Popper revisited: GWAS here, last year. Eur J Hum Genet 2008;16:1-2.

35. Klupa T, Malecki MT. All we need is GWAS: Genome-Wide Association Studies in Type 2 Diabetes Mellitus presented on the 2008 EASD Meeting in Rome. Rev Diabet Stud 2008;5:175-9.

36. Need AC, Attix DK, McEvoy JM, Cirulli ET, Linney KN, Wagoner AP, et al. Failure to replicate effect of Kibra on human memory in two large cohorts of European origin. Am J Med Genet B Neuropsychiatr Genet 2008;147B:667-8.

37. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356-69.

38. Grossman I, Sullivan PF, Walley N, Liu Y, Dawson JR, Gumbs C, et al. Genetic determinants of variable metabolism have little impact on the clinical use of leading antipsychotics in the CATIE study. Genet Med 2008;10:720-9.

39. Burnett JR, Hooper AJ. Common and Rare Gene Variants Affecting Plasma LDL Cholesterol. Clin Biochem Rev 2008;29:11-26.

40. Wojczynski MK, Tiwari HK. Definition of phenotype. Adv Genet 2008;60:75-105.

41. Uzuner O, Goldstein ., Luo Y, Kohane IS. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15:14-24.

42. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008;84:362-9.

43. Mandl KD, Kohane IS. Tectonic shifts in the health information economy. N Engl J Med 2008; 358:1732-7.

Correspondence to:

Isaac S. Kohane, MD, PhD
Professor of Pediatrics and Health Sciences Technology
Center for Biomedical Informatics
Harvard Medical School
10 Shattuck Street
Boston, MA 02115, USA
E-mail: Isaac_kohane@harvard.edu