Commentary

Translational bioinformatics applications in genome medicine

Atul J Butte

Addresses: Stanford Center for Biomedical Informatics, Department of Medicine and Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305, USA, and Lucile Packard Children's Hospital, Palo Alto, CA 94304, USA. Email: abutte@stanford.edu

Published: 29 June 2009

Genome Medicine 2009, **1**:64 (doi:10.1186/gm64)

The electronic version of this article is the complete one and can be found online at http://genomemedicine.com/content/1/6/64

Abstract

Although investigators using methodologies in bioinformatics have always been useful in genomic experimentation in analytic, engineering, and infrastructure support roles, only recently have bioinformaticians been able to have a primary scientific role in asking and answering questions on human health and disease. Here, I argue that this shift in role towards asking questions in medicine is now the next step needed for the field of bioinformatics. I outline four reasons why bioinformaticians are newly enabled to drive the questions in primary medical discovery: public availability of data, intersection of data across experiments, commoditization of methods, and streamlined validation. I also list four recommendations for bioinformaticians wishing to get more involved in translational research.

Introduction

Over the past decade, a large amount of individual-level molecular data has come from the use of gene expression microarrays [1,2], proteomics [3], and DNA sequencing [4,5]. Although high-throughput measurement modalities such as these have been used in biomedical research for over a decade, the role of the bioinformatician has often been relegated to that of data analyst, librarian, database manager, distribution specialist, or software engineer. Occasionally, with introductions made early enough, bioinformaticians have been included in the early design phases of experiments, and their role noted as such on manuscripts and publications. These engineering and infrastructure roles, although important, evolved under the assumption that the scientists making these measurements already know good questions to ask but lack the specific skills to analyze, store, retrieve, and disseminate their data. Engineering roles in bioinformatics are important and are reasonably well funded today (such as in the Cancer Bioinformatics Grid (caBIG), Bioinformatics Research Network (BIRN), and the National Centers for Biomedical Computing (NCBC), all in the United States).

But considering and funding solely the engineering roles in bioinformatics understates the potential function of bioinformaticians as scientists - here defined as those who come up with questions - and, even more importantly, it limits the vision for bioinformaticians to ask questions that no other scientists can ask or answer today. It has become increasingly rare for the bioinformatician to take the role of questioner, especially with regard to research that has an impact on medical care or research that yields tools for clinicians or patients. Here, I argue that the next steps needed for the field of bioinformatics are a shift in role towards asking questions and a shift in focus to medicine. The field of translational bioinformatics, defined as '...the development of storage, analytic and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into proactive, predictive, preventative, and participatory health' [6], is the mechanism for this shift. I outline below four reasons why bioinformaticians are newly enabled to drive the questions in primary medical discovery, and provide four recommendations for bioinformaticians who would like to get more involved in translational research.

Four enabling opportunities

The most revolutionary force in translational bioinformatics is the public availability of molecular data. Sharing data is not new; large epidemiological datasets and DNA sequences have been shared in various forms for several decades, even before the internet era. In addition, the use of previously published data is not new; the biostatistics literature is full of novel methodology applied to well known datasets. But instead of using public data to just improve one's methodology (for example, to build yet another classifier on Todd Golub's leukemia data [7]), or in basic science (for example, to build yet another predictor for transcription factor binding sites), such data can now be used to enable new questions in applied sciences.

Coupled with the public availability of molecular measurement data is the promising capability of intersecting across multiple experiments. At the time of writing, the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) contains data from over 307,000 microarrays, from 12,100 independent experiments [8]. Although the growth rate has been exponential, GEO can be currently described as having made available roughly 100 new microarrays each day since its launch in January 2001. Imagine: a high school student today who needs to run a science fair project can type 'breast cancer' at the NCBI GEO home page to find data from nearly 400 experiments totaling 24,200 samples, as easily as she can find songs on iTunes. With the right tools, she could even discover the 'common denominator' across tens or hundreds of models of breast cancer. Rhodes *et al.* [9] used this approach to compile publicly available published microarray datasets in which cancer samples were compared with appropriate normal samples to find common changes in gene expression across cancers, such as cell cycle genes involved in metastasis, and my colleague and I [10] used 49 publicly available gene expression, proteomics, and RNA interference datasets to predict novel variants associated with obesity. Although there are challenges in using this approach [11], with over 30% of the human-disease morbidity already represented in GEO [12] there is clearly power in large numbers.

A negative disruptive factor, potentially steering bioinformaticians away from staid approaches, has been the increasing commoditization of bioinformatics methodology. Over 1,100 databases are now listed in the Annual Database issue of *Nucleic Acids Research* [13], with another hundred web-servers listed separately [14]. Approximately 60 manuscripts are published each month describing software or methodology in bioinformatics in the journals *Genome Biology*, *BMC Bioinformatics*, *BMC Genomics*, and *Bioinformatics*. Even sophisticated choices on the best machine-learning algorithm to use in a particular context have been made trivial by free tools such as Weka [15], which essentially abstract away the need to know specific methodology. It is getting progressively harder to argue that increasing sophistication and knowledge of this type of methodology significantly improves one's results.

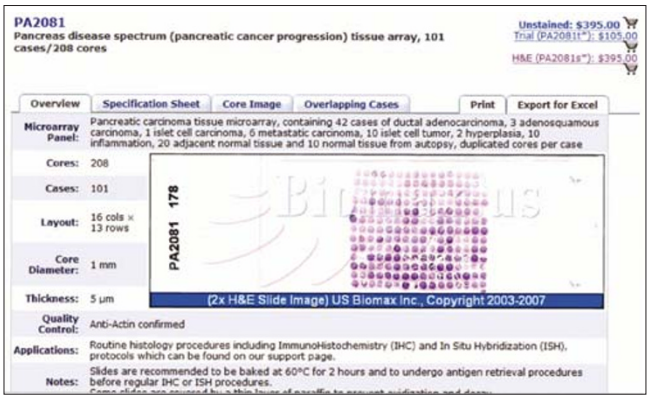

Figure 1
Screenshot from US Biomax [19] showing a tissue microarray for sale with 101 cases of pancreatic cancer or adjacent normal tissues. This figure is representative of many other available companies offering products and services for validation. Other such services can be found by searching the internet with terms such as 'tissue microarray', 'tissue samples', and 'serum samples'.

With the availability of enormous sets of data and the commoditization of methodology, merely making lists of potential biomarkers and causal factors will eventually lose value and significance. Although much additional value comes from validation in real human samples, these samples have typically been difficult to obtain, until now. Figure 1 shows one example out of many websites that now offer human samples, antibodies that can be used to stain those samples, and pathology services that can be used to read the results. One can always question the reliability and quality of these samples and services, as one can question samples and services within one's own institution. However, it is difficult to ignore the importance of having these facilities available to the bioinformatician. Although caveats must be acknowledged, in many ways all that is now left to do is to ask the interesting question.

Four recommendations

How can the field of bioinformatics successfully adapt to the translational movement? First, if the hardest part to scientific endeavors in biomedical informatics is to ask the right question, then investigators in biomedical informatics need to learn more about open problems in medicine. Some of this learning will come from non-traditional sources, such as medical or surgical grand rounds (regular conferences discussing the science around particularly challenging or instructive cases) in a medical center. Often, 'domain-specific learning' is viewed as a slippery slope; informaticians sometimes retort that it is not possible to gain competence across all areas of medicine while retaining

expertise in a computational discipline. But learning about the unaddressed challenges even in one particular area of medicine is still better than knowing little or nothing about any area of medicine; as most physician scientists know, focus in one particular medical area of interest provides more than enough challenges for a career. As informatics tools become more easily accessible, understood, and used without assistance by medical researchers, the reverse also has to occur, with medical problems becoming understood and addressed by computational investigators.

The corollary to this point is a second recommendation directed towards bioinformaticians: with the commoditization of bioinformatics methodologies, researchers in informatics should not just build tools, they should be the first to use them, even on publicly available data. Indeed, no other investigator knows those tools better than the inventor. Those who build tools to address a specific medical question can and should report on both their tool and their findings. After tools and methods have been shown to answer one question particularly well, they can then be generalized for additional questions. This recommendation is contrary to the usual practice of building tools in bioinformatics to enable others. In general, this will mean that tools that have successfully enabled their creator to discover an important finding should be viewed with higher regard, as opposed to tools presenting a fancier user interface or marginal gains in performance.

It is often easiest to criticize the quality of publicly available resources, whether these resources are data or tools. Many initiatives within the community of biomedical informatics have tried to add value to these public resources by creating standardized annotations (and metadata), catalogs, structured vocabularies, and ontologies, which can be used to store, index, and retrieve them more efficiently and effectively [16,17]. Although these efforts have the best of intentions, we have to ensure that, in the push to improve the quality of metadata, we do not inadvertently cause a delay in the release of data or tools.

The final recommendation is for informaticians to broadly consider their sources for molecular data. A tertiary care academic medical center might see tens to hundreds of thousands of patients with injuries and diseases each year. In modern hospitals, nearly every intervention applied to these patients is electronically recorded, and hundreds of thousands of blood measurements are made yearly, along with high-resolution images and tissue pathology. The scale of the clinical enterprise easily dwarfs the abilities of most typical animal model facilities, and the requirements for quality assurance for medical measurements greatly exceeds the typical levels of rigor applied in model experimentation. Put another way, the typical clinical laboratory measurement is much more believable than the typical spot on a microarray. There are barriers to accessing clinical data, but as these can be overcome, bioinformaticians should start considering humans as the ultimate model organism [18].

Conclusions

It is remarkable that in the decade or two since their creation, high-throughput molecular measurements, such as microarrays, have already been used to study so many human diseases, and that data from these experiments are publicly available. Representing so many diseases by molecular measurements in gene expression (and other measurement modalities in the future) brings us closer to a consideration of the nature of disease itself. As the community of biomedical informaticians is increasingly involved (and funded) in the construction of infrastructure and policies to gather and consolidate clinical and experimental data, we have to consider that this community will also be the prime user of these tools and techniques. Those who apply their research to publicly available data, commoditized tools, and streamlined paths through validation will be able to create novel diagnostics and discover fundamental causes of disease as targets for therapies. Investigators empowered by methodologies in bioinformatics have never been so well positioned to take on the role of translational scientist, to build the tools to ask the questions that yield discoveries to improve human health.

Abbreviations

GEO, Gene Expression Omnibus; NCBI, National Center for Biotechnology Information.

Competing interests

AB is or has served as a scientific advisor and/or consultant to NuMedii, Genstruct, Prevendia, Tercica, Eli Lilly and Company, and Johnson and Johnson.

Acknowledgements

The work was supported by grants from the Lucile Packard Foundation for Children's Health, National Library of Medicine (K22 LM008261 and R01 LM009719), National Institute of General Medical Sciences (R01 GM079719), Howard Hughes Medical Institute, and the Pharmaceutical Research and Manufacturers of America Foundation.

References

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
2. Lipshutz RJ, Morris D, Chee M, Hubbell E, Kozal MJ, Shah N, Shen N, Yang R, Fodor SP: **Using oligonucleotide probe arrays to access genetic diversity.** *Biotechniques* 1995, **19:**442-447.
3. Mann M: **Quantitative proteomics?** *Nat Biotechnol* 1999, **17:**954-955.
4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, *et al.*:

Initial sequencing and analysis of the human genome. *Nature* 2001, **409:**860-921.

5. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
6. **American Medical Informatics Association Strategic Plan** [http://www.amia.org/inside/stratplan/]
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
8. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles - database and tools update.** *Nucleic Acids Res* 2007, **35:**D760-D765.
9. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci USA* 2004, **101:**9309-9314.
10. English SB, Butte AJ: **Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes.** *Bioinformatics* 2007, **23:**2910-2917.
11. Ramasamy A, Mondry A, Holmes CC, Altman DG: **Key issues in conducting a meta-analysis of gene expression microarray datasets.** *PLoS Med* 2008, **5:**e184.
12. Dudley J, Butte AJ: **Enabling integrative genomic analysis of high-impact human diseases through text mining.** *Pac Symp Biocomput* 2008:580-591.
13. Galperin MY, Cochrane GR: **Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009.** *Nucleic Acids Res* 2009, **37:**D1-D4.
14. Brazas MD, Fox JA, Brown T, McMillan S, Ouellette BF: **Keeping pace with the data: 2008 update on the Bioinformatics Links Directory.** *Nucleic Acids Res* 2008, **36:**W2-W4.
15. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bio informatics using Weka.** *Bioinformatics* 2004, **20:**2479-2481.
16. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: **Minimum information about a micro array experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29:**365-371.
17. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nat Biotechnol* 2007, **25:**1251-1255.
18. Butte AJ: **Medicine. The ultimate model organism.** *Science* 2008, **320:**325-327.
19. **US Biomax** [http://www.biomax.us/tissue-arrays/Pancreas/PA2081].