Delivering Bioinformatics Training: Bridging the Gaps Between Computer Science and Biomedicine

Christopher Dubay Ph.D., James M. Brundege Ph.D., William Hersh M.D.,
Kent Spackman M.D., Ph.D.

Oregon Health & Science University, Division of Medical Informatics & Outcomes Research
Portland, Oregon 97201-3098, dubayc@ohsu.edu

Abstract

Biomedical researchers have always sought innovative methodologies to elucidate the underlying biology in their experimental models. As the pace of research has increased with new technologies that 'scale-up' these experiments, researchers have developed acute needs for the information technologies which assist them in managing and processing their experiments and results into useful data analyses that support scientific discovery. The application of information technology to support this discovery process is often called bioinformatics. We have observed a 'gap' in the training of those individuals who traditionally aid in the delivery of information technology at the level of the end-user (e.g. a systems analyst working with a biomedical researcher) which can negatively impact the successful application of technological solutions to biomedical research problems. In this paper we describe the roots and branches of bioinformatics to illustrate a range of applications and technologies that it encompasses. We then propose a taxonomy of bioinformatics as a framework for the identification of skills employed in the field. The taxonomy can be used to assess a set of skills required by a student to traverse this hierarchy from one area to another. We then describe a curriculum that attempts to deliver the identified skills to a broad audience of participants, and describe our experiences with the curriculum to show how it can help bridge the 'gap.'

Roots & Branches of Bioinformatics

An assessment of the relatively new field of bioinformatics can benefit from a comparison to the older, more established fields of clinical and medical informatics. The parallels between medical informatics and bioinformatics are striking. Medical informaticists strive to create, evaluate and deliver information technology (IT) for health care in many settings. Bioinformaticists create, evaluate, and deliver IT for biomedical research, and increasingly for industrial applications including agriculture, animal husbandry, drug development, and biomedicine. The customers for research applications of bioinformatics are researchers (from principal investigators to post-docs and students), laboratory technicians, and increasingly domain experts in complementary fields (e.g. biostatisticians, pathologists, physiologists, oncologists, etc.).

In medical informatics many tools and systems with formidable capabilities have gone under-utilized simply because they did not take into account the setting and audience for their use [1, 2]. Medical informaticists need to have an intimate understanding of the culture and environment in which they wish to deploy their systems if they want to have a good probability of success with those systems. The same is true for bioinformatics; to deliver useful tools and systems one needs to have an understanding of one's customers and the 'post-genomic' biomedical research culture in which they work. We emphasize culture because of the very large differences between the engineering culture typical of main-stream information technologists, where the answers can be found in the technical documentation, and the research culture, where there are no existing complete manuals for the biological systems we study. Often the nature of applications is different as well. This can be illustrated by considering the differences between a system to manage existing data (e.g. banking, airline reservation system), and one to help use existing data to create new knowledge and understanding (e.g. cross-species homology searching, biochemical pathway modeling).

The roots of applying computers to biological research are quite deep. Early computers were adapted to store and analyze many types of data including census data for public health analyses (e.g. risk factor identification) [3], storage and analysis of pedigree and genetic marker data for linkage analyses [4], and solutions to results of x-ray crystallographic data for 3D structures of bio-molecules [5]. These roots of storing phenotypes, genotypes, and models have grown large branches.

The storage and analysis of genotypic data, as represented by biosequences (i.e. DNA, RNA, and protein sequences) is a main-stay of bioinformatics. This is because the comparison of biosequences, either between samples or between species, has been highly successful in elucidating the biological relevance of sequence variations. Genotypic bioinformatics systems form the basis for the various

global projects to sequence entire genomes, with a notable example being the completion of the public draft and commercial sequence of the human genome in 2001 [6, 7].

The storage and analysis of phenotypic data is now taking a place in the 'spot light' as expression analysis (i.e. the characterization of mRNA populations in specific cell types). Expression analysis has been scaled-up using DNA chips [8], and is producing large datasets which cannot be analyzed using the traditional methods employed by researchers for interpretation of gene expression levels in small scale experiments. Bioinformatics is helping geneticists tackle the important job of correlating genotypes with phenotypes. By using high-throughput systems to analyze high-density sets of genetic markers (e.g. single nucleotide polymorphisms or SNPs) researchers are beginning to correlate these results with important clinical phenotypes for disease and drug metabolism/action [9-11]. As we begin to deliver biomedicine, the correlation of genotypes with clinical phenotypes will be the initial focus of the intersection between bioinformatics and medical informatics [12].

The creation of models for the structure of biomolecules helps give biomedical researchers tools to visualize biological activity. The types and specificity of the models required by researchers is expanding rapidly as they try to 'tell a story' of how large numbers of genes (e.g. an entire genome) are expressed into proteins, and how all these cellular elements function and interact. The description of biological function for genes and their products is called functional genomics [13], and the description of their interactions, often using pathway modeling, is called systems biology [14]. Systems biology is currently concerned with cellular processes, and will grow to model higher level biological systems: organelles, organs, and eventually entire organisms.

The scaling-up of biomedical research experiments has created many new challenges, including the need for methods to share and visualize large data sets. The bioinformatics community has developed controlled vocabularies to support consistency in annotation of biological entities. The Gene Ontology Consortium provides a framework for annotation based on three domains: molecular function, biological process and cellular component [15]. Standards for data interchange employing current IT models (e.g. XML, Web Services, etc.) are being developed and vetted [16]. Methods to assist in the visualization of large data-sets that can help researchers detect underlying biological significance are being developed [17-20]. Giving researchers the ability to 'get their hands around' the large data-sets they create and explore is key to their effectively interpreting the results of an experiment.

As these branches of bioinformatics have grown various IT technologies have been applied to the array of problems in these fields with varying degrees of success: artificial intelligence, machine learning, neural nets, genetic algorithms, etc. [21]. The most accepted and proven methods for applying these technologies and developing information systems for users are found in the field of software engineering; a branch of Computer Science and Engineering (CSE). However these methods have not consistently delivered useful systems in bioinformatics. We propose that this is because of a disconnect between the developer (an engineer) and the customer (a researcher), which is especially acute in the first critical phases of defining the user requirements for the proposed systems. The software development lifecycle is a tool from CSE commonly used to support the process of IT systems development [22]. The successful completion of the first phases of this lifecycle (i.e. problem determination and requirements gathering) have been shown to be directly related to the success of the delivered system.

Thus, the disconnect may come from the combination of: 1) a gap in the skill set of the developer to understand the research lifecycle and that the requirements of the researcher are not fixed, and can change dramatically based of current scientific understanding and technical advances, and 2) a gap in the skill set of the researcher to understand software development lifecycle, and the limitations of information systems in terms of flexibility and development time, costs and quality.

We propose that a reasonable approach to bridging these gaps is to attempt to deliver a skill set to each of these groups which will enhance their ability to work together and understand each other's field and culture. This is opposed to trying to deliver the entire body of knowledge (e.g. a BS in molecular biology and computer science) to each group.

Skills for a Computer Science & Engineering Approach to Bioinformatics

As a framework for describing the skills that we wish to deliver to developers and researchers we have created a working taxonomy for the field of bioinformatics.

The top-level distinction in this taxonomy is between academics and industry. Although many of the tools used in these two branches are identical, it is an important distinction for how we prepare students. The second level is comprised of research, training, and development. For any of these second level activities there are theories, methods and tools delivered from the processes of designing, formalizing, and implementing (respectively).

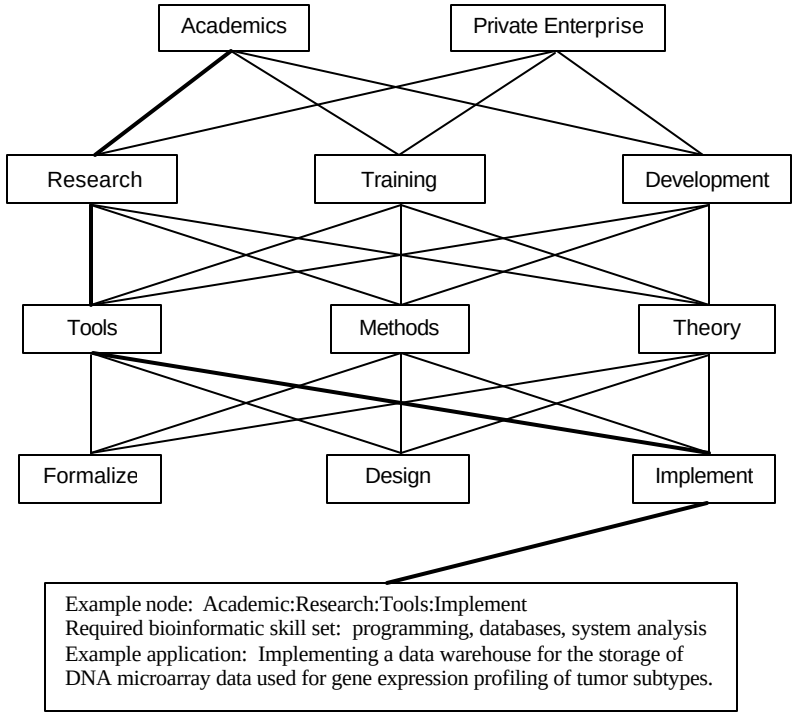

Figure: Taxonomy of career branches in bioinformatics. This framework categorizes the branches of endeavor that are common to all areas of biomedical science. Each node can be comprised of any of the nodes below it. By identifying which nodes a student wishes to position themselves in as part of their career path, we can ensure that they develop the bioinformatic skills needed to excel at that position, regardless of which area of biomedical science they are working in.

We list some of the domains that these branches can be applied to, as a way of acknowledging that there are 'flavors' of these processes appropriate to different fields of application that have different knowledge bases, vocabularies, and cultures.

We have linked specific skills into the various nodes and branches of this taxonomy [figure]. We have then attempted to analyze the current and future locations and directions of our students in this taxonomy as a way of characterizing their expected existing skills, and their desired future skills. In this way we can assess which of the skills inferred from the student's current position are not in place, and add to this the sets of skills they will need to traverse to another location in the hierarchy.

Audience for Training

The courses offered at OHSU are made available to a wide range of students: masters students from our Medical Informatics program, basic science graduate students from our various biomedical doctoral programs, students from both master and Ph.D. programs at the OHSU Oregon Graduate Institute School of Science and Engineering (OGI SOSE) (primarily CSE students), and to CSE and basic science students at the masters level from other local institutions. Additionally, the courses have been attended by various interested people at a number of levels such as OHSU and SOSE faculty, OHSU principal investigators, laboratory personnel, and individuals from private industry.

Students self report their current position and future directions using a survey form. Using this form we have identified two main groups of students: those with a biological science background and those with a computer science background (this is not surprising given audience described above).

OHSU Bioinformatics Curriculum

We have developed a curriculum to deliver the skills identified in our taxonomy, and help to fill the gaps we believe hinder bioinformatic systems development and deployment. The curriculum can be tailored to a student based on their location in, and path through, the taxonomy. The curriculum covers three terms of an academic quarter system. Syllabi of all three terms are available at: www.ohsu.edu/bicc-informatics/ms/coursedesc.shtml.

The first term consists of a survey course which covers a broad range of bioinformatics topics. It begins with two lectures to equilibrate the class: 1) computers explained from a biological perspective, and 2) biology explained from an information science perspective. As the two class populations attend the lectures they get a cross-cultural view of the two domains, and are exposed to the commonality and differences in approaches to problems, professional activities, and vocabulary. Students are surveyed at the beginning of the class to get a feeling for backgrounds and levels of experience which is used by the instructor to help pitch the content at an appropriate level. There is an accompanying laboratory class which is given in a classroom equipped with Macintosh and PC workstations (students are encouraged to use both). In the lab students get hands-on training in UNIX, the GCG software suite (www.gcg.com), and many bioinformatics resources available on the Internet, with a special focus on the tools and databases available at the National Center for Biotechnology Information (NCBI) [23]. Evaluation of performance is based on a mid-term examination and term project. The exam contains questions on information

technology, biology, and bioinformatic analysis. The course project is proposed in the third week of class using a one page abstract with a minimum of three references from the primary-literature in a field of the student's choosing (usually designed to overlap with their current work in other areas). Suggestions on the scope and focus of the project are made individually to students, and accepted projects are delivered at the end of the term as papers or presentations (30 minutes in length, given to the entire class on finals week). To aid students in finding relevant topics, we have created a web distributed database to advertise possible projects which can be submitted by any OHSU personnel.

In the second term a topics course provides more in-depth coverage of a sub-set of the topics from the previous term. Students are again surveyed at the beginning of the class to determine their interests, and the set of topics to be covered is determined from this list. The two main themes underlying the delivery of content on these topics are: databases and the bioinformatics published literature. We have collaborated with faculty from the OGI SOSE to deliver a series of lectures on database technologies, including the application of traditional data storage and mining techniques (e.g. those used in retail forecasting) to bioinformatics. To cover the bioinformatics literature we have an in-class journal club. Students propose three papers each, drawn from the bioinformatics literature, and the entire class votes on which paper the student will give in a 30 minute presentation (~10 minutes of which is Q&A). To augment the in-depth coverage of the class's chosen topics we have guest lectures, who are currently working in the topic field, present their current research and future directions, with an emphasis on what the un-met and under-met information technology needs they face are. Evaluation is based on performance in the journal club, and a term project (using the same project vetting and formats from the previous term). Students are told that the project they choose should have un-met and under-met information technology needs which they should identify as a starting point for their work in the third term.

The third term is a bioinformatic systems development course which extends the in-depth coverage of topics from the previous term by having students develop information systems based on needs identified from their projects in previous terms. The course delivers an overview of software development best-practices and methodologies with an emphasis on development of the functionality prevalent in bioinformatic tools (e.g. database interoperability, client/server and distributed computing designs, visual user interfaces, etc.). The paradigm for the course is that of a software development project, with students working alone or in groups (which we attempt to populate with both CSE and biomedical students). Students complete the system requirements gathering and high level design tasks for the topic area based on interactions with the targeted user base for the software, and implement prototype systems based on commercial software and/or software developed by the group. We hold weekly in-class project status meetings with a rapid delivery format (i.e. 'what I did last week, what project barriers were, what will be done next week') starting the third week of class to allow all students to see the progress and pitfalls faced in all projects, and give their advice and encouragement to each other. We have collaborated with faculty from the OGI SOSE to deliver a series of lectures on software engineering with a focus on the software development lifecycle. Evaluation is based on a mid-term examination on software engineering, and the results of the systems development projects which are delivered as presentations during finals week.

Since OHSU and its OGI campus are separated by a 20 minute commute, we use H.323 broadcasting devices (www.polycom.com) to deliver the course to the OGI campus with live two-way voice and video. We have captured the course lectures on digital video, and have produced and edited them for delivery over the web. We use the Blackboard platform (www.blackboard.com) to deliver web content consisting of video of the live lectures, and to support instructor-led asynchronous electronic discussions of course topics. We have delivered a web based version of the first term three times so far, and will deliver the second and third terms at distance in Summer and Fall 2002 respectively.

Experiences

We have given the first term curriculum for four years now, to over 100 people, and have found it to be of great utility to a wide variety of participants. Term projects form the course have become chapters in both M.S. and Ph.D. theses. The course has been successful because it exposes the class to a wide range of applications and gives a cross-cultural perspective. Although in some courses difficulties arise when the background of participants is unequal in terms of knowledge, we have benefited from this diversity because it forces explanations on all topics to be generally understandable by the entire audience. By explaining a topic from a variety of perspectives, it is made clear; in the same way that it is important when a developer talks with their customer (e.g. a researcher) to clarify the requirements for and the deliverables of, a bioinformatic application or analysis.

The process of evaluating student term project proposals early in the term has helped students focus on real-world problems, and identified real 'customers' for the project deliverables. We have been able to 'plug-in' students to a number of biomedical research projects on campus, giving

them invaluable experience (and important lines for their resumes).

The students who have completed the full year curriculum (so far eight in our first year), and graduated from the masters program in medical informatics, receive an 'emphasis in Bioinformatics' on their degrees. The feedback to delivering a range of bioinformatics skills, culminating in the delivery of 'something double-clickable', has been very well received by this first class. We currently have 15 students set to complete the entire three terms this year.

The journal club and project management segments of the second and third terms are so popular that we will move them to separate one credit courses next year. We plan to deliver the bioinformatics journal club by web broadcast over our intranet.

Conclusions

We have developed a curriculum in bioinformatics that attempts to deliver a wide range of skills mapped to a taxonomy of bioinformatics, with a focus to cross-culturizing participants between the fields of biomedical research and CSE. We believe that this curriculum helps bridge the 'gap' between these fields which is a barrier to evolving bioinformatic applications for use by biomedical researchers and industry in the post-genomic era.

Acknowledgements

This work was supported in part by the National Library of Medicine through a Biomedical Information Science and Technology Initiative (BISTI) Administrative Supplement to the OHSU Fellowship program in Medical Informatics (Grant #:T15 LM07088).

References

1. Kaplan, B., Addressing organizational issues into the evaluation of medical systems. J Am Med Inform Assoc, 1997. 4(2): p. 94-101.
2. Sittig, D.F., Grand challenges in medical informatics? J Am Med Inform Assoc, 1994. 1(5): p. 412-3.
3. Vansteenkiste, G.C., The use of computers in biostatistics. Pharmacol Ther [B], 1975. 1(2): p. 311-56.
4. Conneally, P.M. and M.L. Rivas, Linkage analysis in man. Adv Hum Genet, 1980. 10: p. 209-66.
5. Bernstein, F.C., et al., The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol, 1977. 112(3): p. 535-42.
6. Lander, E.S., et al., Initial sequencing and analysis of the human genome. Nature, 2001. 409(6822): p. 860-921.
7. Venter, J.C., et al., The sequence of the human genome. Science, 2001. 291(5507): p. 1304-51.
8. Lander, E.S., Array of hope. Nat Genet, 1999. 21(1 Suppl): p. 3-4.
9. Mackay, T.F., The genetic architecture of quantitative traits. Annu Rev Genet, 2001. 35: p. 303-39.
10. Riley, J.H., et al., The use of single nucleotide polymorphisms in the isolation of common disease genes. Pharmacogenomics, 2000. 1(1): p. 39-47.
11. Schmitz, G., C. Aslanidis, and K.J. Lackner, Pharmacogenomics: implications for laboratory medicine. Clin Chim Acta, 2001. 308(1-2): p. 43-53.
12. Altman, R.B. and T.E. Klein, Challenges for biomedical informatics and pharmacogenomics. Annu Rev Pharmacol Toxicol, 2002. 42: p. 113-33.
13. Yaspo, M.L., Taking a functional genomics approach in molecular medicine. Trends Mol Med, 2001. 7(11): p. 494-501.
14. Ideker, T., T. Galitski, and L. Hood, A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet, 2001. 2: p. 343-72.
15. Ashburner, M., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000. 25(1): p. 25-9.
16. Stein, L.D. Bioinformatics - Building a Nation from a Land of City States. in O'Reilly Bioinformatics Technology Conference. 2002.
17. Apweiler, R., et al., InterPro--an integrated documentation resource for protein families, domains and functional sites. Bioinformatics, 2000. 16(12): p. 1145-50.
18. Bassett, D.E., Jr., M.B. Eisen, and M.S. Boguski, Gene expression informatics--it's all in your mine. Nat Genet, 1999. 21(1 Suppl): p. 51-5.
19. Kanehisa, M., et al., The KEGG databases at GenomeNet. Nucleic Acids Res, 2002. 30(1): p. 42-6.
20. Ruths, D.A., E.S. Chen, and L. Ellis, Arbor 3D: an interactive environment for examining phylogenetic and taxonomic trees in multiple dimensions. Bioinformatics, 2000. 16(11): p. 1003-9.
21. Searls, D.B., Bioinformatics tools for whole genomes. Annu Rev Genomics Hum Genet, 2000. 1: p. 251-79.
22. McConnell, S., Rapid Development: Taming Wild Software Schedules. 1996: Microsoft Press.
23. Wheeler, D.L., et al., Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res, 2002. 30(1): p. 13-6.