Bioinformatics training in the USA

Marion M. Zatz

Date received (in revised form): 23rd August 2002

Marion M. Zatz, PhD
is a scientist administrator at NIGMS where she administers research grants and predoctoral training grant programmes in several areas, including bioinformatics and computational biology. Dr Zatz was instrumental in the development of the NIGMS training grant programme in this area. Her research background is in immunology and microbiology.

Keywords: *bioinformatics, computational biology, training, PhD programmes, interdisciplinary*

Abstract

This paper provides an overview of the history and funding of bioinformatics training in the USA, and summarises some of the challenges and key features associated with bioinformatics training programmes at PhD level. The paper includes compilations of current PhD bioinformatics training programmes and sources of funding.

INTRODUCTION

The last ten years have seen an explosion of high-throughput technology in the biological and chemical sciences and advances in computational capabilities for large-scale analysis. As a result, enormous quantities of data can quickly be generated and need to be organised, analysed and retrieved.[1] Bioinformatics is the discipline that has emerged to meet this need.

Initially, bioinformatics was driven largely by advances in genomics and sequencing technology. However, most research areas – including structural biology, proteomics, combinatorial chemistry, neurobiology, pharmacokinetics and drug discovery – now rely on bioinformatics. Thus it is necessary to broadly define the 'bio' in bioinformatics before discussing bioinformatics training. Perhaps the simplest definition was stated by David Botstein: 'Bioinformatics is a sort of code word for biologists having to deal with very large amounts of information'.[2] The National Institutes of Health (NIH) offers a more detailed definition of bioinformatics as: 'Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze or visualize such data'.[3] A more complete listing of references that discuss the definition of bioinformatics can be found at the University of Minnesota bioinformatics web site.[4]

A distinction is made between bioinformatics and computational biology, which is defined as: 'The development and application of data – analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral and social systems'.[3] In reality, bioinformatics and computational biology are on a continuum, requiring many of the same skills, and many training programmes encompass both areas. Although for brevity I will mainly use the term bioinformatics in this paper, I intend this term to be construed in its broadest meaning. Medical informatics training, which targets healthcare professionals and addresses issues of medical practice, education and administration, will not be addressed, although some training programmes encompass both bioinformatics and medical informatics.

Many individuals who first applied bioinformatics to biological problems were self-taught. However, the demand for bioinformaticists rapidly outstripped the supply. Recent articles have highlighted the dearth, both in industry and academia, of bioinformaticists.[5,6] Thus, the need to train scientists whose primary professional identification and disciplinary affiliation is in bioinformatics must be addressed effectively at all levels, from college undergraduates to graduate

Marion M. Zatz, PhD,
National Institute of General Medical Sciences,
45 Center Drive MSC6200,
Bethesda,
MD 20892-6200, USA

Tel: +1 301 594 0943
Fax: +1 301 480 2228
E-mail: ZATZM@nigms.nih.gov

students and postdoctoral fellows. In this paper I will focus on bioinformatics training at the PhD level, the level that probably is changing most quickly and that poses unique challenges.

HISTORY OF TRAINING IN BIOINFORMATICS

As the critical need for training in bioinformatics grew in the mid-1990s, the initial efforts were focused on Master of Science (MS) programmes,[7] on certificate programmes that provided short-term training in bioinformatics in conjunction with other degrees, and on postdoctoral fellowships. The MS programmes, generally of one to two years' duration, offered courses and research internships that helped train a cadre of students who could begin to fill the immediate need in industry and function as part of a research team. One funding organisation, the Sloan Foundation, first supported postdoctoral fellows, and then in 1999 began providing funding for professional MS degree programmes in bioinformatics and computational biology. Some institutions, such as the University of Pennsylvania and Boston University, started comprehensive training programmes at multiple levels, including undergraduate, MS, PhD, postdoctoral and certificate, driven in part by a research focus in genomic sciences.

Current bioinformatics PHD programmes

At the last count there were approximately 50 programmes in the USA that offer either an MS track or degree in bioinformatics.[8–11] However, it has only been in the last five years that most bioinformatics PhD programmes have appeared and that support from funding agencies has become available. Whereas many MS or postdoctoral fellowship programmes emphasise a curriculum of individual courses in either the biological sciences or math and computer science to complement the undergraduate or graduate background of the trainee, a PhD programme must be more comprehensive. It must go beyond courses or short-term research internships to offer an integrated programme of cross-disciplinary training that will prepare a scientist for independent research in an academic or non-academic setting. A PhD programme must also bridge the culture and the science of biomedical research, bioinformatics and computational biology.

Integrated training

A number of web sites offer compilations of current graduate and postgraduate training programmes.[8–11] An examination of these sites indicates that they rapidly become outdated. Table 1 gives a current listing of PhD training programmes based on these web sites and on a poll I conducted of the Association of American Medical Colleges' listserv of graduate institutions.[12] Only those institutions that have web sites describing training programmes are included, although there are additional programmes that offer a more *ad hoc* training experience, and undoubtedly there are some programmes that I have overlooked. The roots of some of the PhD programmes, such as those at the University of Wisconsin and Stanford University, lie in longer-standing medical informatics programmes that have now expanded or shifted to encompass bioinformatics. Of the 40 programmes listed, almost half have been established within the last two years, or have formally been proposed. Another four institutions that are not listed (University of Buffalo, University of New Mexico, University of Florida and Drexel University) have indicated they have programmes awaiting approval.

SUPPORT FOR BIOINFORMATICS TRAINING

Training grants from the NIH and NSF have influenced the focus and philosophy of bioinformatics training

The availability of training grants from the NIH and the National Science Foundation (NSF) has been an important factor in the development of PhD training programmes in bioinformatics and computational biology. These funding sources have not only provided the resources to subsidise the substantial cost of a training programme that takes five years or longer, they have also helped to

Table 1: Training programmes in bioinformatics and computational biology

Institution	Degree/level	Programme
Boston University http://bioinfo.bu.edu/ with Northeastern University http://www.bioinformatics.neu.edu/index.html	MS PhD Certificate	Bioinformatics
Carnegie Mellon University http://www.cmu.edu/mcs/Merck/ MerckGraduateProgram.html	MS PhD	Computational Biology and Chemistry
Duke University http://www.cbcb.duke.edu/	PhD (proposed) Certificate Postdoctoral	Bioinformatics and Genome Technology
George Mason University http://www.ib3.gmu.edu/	MS PhD	Bioinformatics and Computational Biology
Georgia Institute of Technology http://www.biology.gatech.edu/gt_bioinfo/	MS PhD	Bioinformatics and Computational Biology
Harvard-MIT Division of Health Sciences and Technology http://hst-hu-mit.mit.edu/	PhD	Bioinformatics and Functional Genomics
Iowa State University http://www.bcb.iastate.edu/	MS PhD	Bioinformatics and Computational Biology
Johns Hopkins University http://www.jhu.edu/~ibr/bwf/index.html	PhD Postdoctoral	Computational Biology
Keck Center http://www-bioc.rice.edu/Keck/keck_info/tnsf.html Baylor College of Medicine, Rice University, University of Houston	Undergraduate PhD Postdoctoral	Bioinformatics, Computational Biology and Structural Biology
Montana State University http://cns.montana.edu/academics/	MS PhD	Computational Biology
New Jersey Institute of Technology http://www.njit.edu/New/cb/	MS PhD	Computational Biology
North Carolina State University http://bioinformatics.ncsu.edu/	MS PhD	Statistical Genetics and Bioinformatics
Ohio State University http://www.ibgp.org/over.htm#Track6	PhD (Track)	Interdisciplinary with Bioinformatics Track
Stanford University http://smi-web.stanford.edu/academics/index.html	MS PhD Postdoctoral	Biomedical Informatics
University of Arkansas www.UALR.EDU	PhD	Computational Science
University of California, Irvine http://www.igb.uci.edu/education.htm	MS PhD	Informatics in Biology and Medicine
University of California, Los Angeles http://www.bioinformatics.ucla.edu/	Undergraduate PhD	Bioinformatics
University of California, San Diego http://bioinformatics.ucsd.edu/	PhD	Bioinformatics
University of California, San Francisco http://www.bmi.ucsf.edu/	MS PhD	Biological and Medical Informatics
University of California, Santa Cruz http://www.cse.ucsc.edu/research/compbio/	Undergraduate MS PhD (proposed) Postdoctoral	Bioinformatics and Computational Biology
University of Cincinnati http://www.eng.uc.edu/dept_biomed/grad/	MS PhD (Track)	Biomedical Engineering/Bioinformatics Track
University of Colorado Health sciences Center http://www.uchsc.edu/sm/pmb/bioi/index.htm	MS PhD (Track)	Analytical Health Sciences/ Bioinformatics Track
University of Illinois, Chicago http://www.uic.edu/depts/bioe/main/bioinformatics.htm	MS PhD	Bioinformatics
University of Medicine and Dentistry, New Jersey http://gsbs.umdnj.edu/bioinformatics/ Rutgers University http://cmb.rutgers.edu/	MS PhD or Certificate	Bioinformatics and Computational Biology

(*continued overleaf*)

Table 1: (*continued*)

Institution	Degree/level	Programme
University of Michigan http://www.bioinformatics.med.umich.edu/	PhD	Bioinformatics
University of Minnesota http://www.binf.umn.edu/	MS PhD (Minor)	Bioinformatics
University of Nebraska http://www.isqa.unomaha.edu/bioinformatics/	MS PhD (Track)	Pathology & Microbiology/ Bioinformatics Track
University of North Carolina, Chapel Hill http://bioinfo.med.unc.edu/Bioinformatics/index.html	PhD	Bioinformatics and Computational Biology
University of Pennsylvania http://www.cbil.upenn.edu/UPCB/	Undergraduate MS PhD Postdoctoral	Computational Biology
University of Pittsburgh http://www.cbmi.upmc.edu/2002-train.htm	MS PhD Postdoctoral	Biomedical Informatics
University of Southern California http://www.usc.edu/dept/LAS/biosci/mcb/graduate_study.shtml	MS PhD	Molecular and Computational Biology
University of Texas, Austin http://www.esb.utexas.edu/molbio/	PhD (Track)	Cellular and Molecular Biology/ Bioinformatics Track
University of Washington http://depts.washington.edu/cmolbiol/	PhD (Track)	Departmental Programmes/ Computational Molecular Biology Track
University of Wisconsin http://www.medsch.wisc.edu/biostat/training/bioinfocert.html	MS PhD Postdoctoral (Certificate Programme)	Biological Science/Bioinformatics Certificate
Vanderbilt University http://www.mc.vanderbilt.edu/dbmi/bmigradprog/index.html	MS PhD	Biomedical Informatics
Virginia Polytechnic Institute and State University http://graduate.bioinformatics.vt.edu/programs.html	MS PhD (Option)	Departmental Programmes/ Bioinformatics Option
Washington University http://dbbs.wustl.edu/Programs/computational.html	PhD	Computational Biology

Sources: http://www.unl.edu/stc-95/ResTools/biotools/biotools4.html, http://biotech.icmb.utexas.edu/pages/bioinform/biprograms_us.html, http://www.bio-itworld.com/careers/biotrain/index.html, personal communications from the American Medical Association GREATMAIL listserv greatmail@aamcinfo.aamc.org

The NIH and NSF as well as private foundations are supporting training programmes at all levels

shape the scientific focus and philosophy of bioinformatics training in the USA.

In 1996, the National Library of Medicine (NLM) was the first NIH component to offer predoctoral and postdoctoral training grant support in the defined area of medical informatics. While focused on medical informatics, these programmes helped to establish courses and an environment in which the broader area of bioinformatics training could develop. Since then, the NLM has expanded its initiative to support predoctoral training programmes in medical informatics and bioinformatics, six of which (Keck Center in Houston Texas; University of Wisconsin; University of California, Irvine; University of California, San Francisco; University of Pittsburgh; and Vanderbilt University) are more focused on bioinformatics. In 1998, the National Human Genome Research Institute (NHGRI) also began supporting predoctoral, postdoctoral and short-term training in genomic analysis and interpretation, geared to providing training at the interface of biological

Table 2: Support for training programmes in bioinformatics and computational biology

Organisation	Type	Degree/level	Areas
Burroughs Wellcome Fund http://www.bwfund.org/interfaces_in_science.htm	Private foundation	Postdoctoral	Computational biology
Howard Hughes Medical Institute http://hhmi.org/news/070902.html	Private foundation	Undergraduate	Genomics and computational biology
National Institutes of Health http://www.nibib1.nih.gov/training/NIHbioinfo.html	US Government	PhD, postdoctoral and short term	Genomic analysis and interpretation (NHGRI), bioinformatics and computational biology (NIGMS) and medical informatics and bioinformatics (NLM)
National Science Foundation www.nsf.gov/igert	US Government	PhD	Interdisciplinary
PhRMA Foundation http://www.phrmafoundation.org/awards/informatics/	Private foundation	Postdoctoral starter grants	Informatics, biological, chemical and pharmacological sciences
Sloan Foundation http://www.sloan.org/programs/supresearch_cmb.shtml	Private foundation	MS Postdoctoral	Bioinformatics and computational biology

disciplines and non-biological disciplines relevant to genomic sciences, including computational biology, bioinformatics and computer sciences. The NHGRI supports at least two training programmes with an emphasis on bioinformatics, at the Massachusetts Institute of Technology and the University of Pennsylvania.

While programmes were developing in the specialised areas of genomic sciences and medical informatics, there was a clear need for broader training at the predoctoral level that would include the use of theory, computer implementation and application to the full spectrum of basic research in the biomedical sciences. In 1998 and 1999, the National Institute of General Medical Sciences (NIGMS), which supports nearly half of all institutional predoctoral training programmes funded by the NIH, convened a series of workshops to consider the future needs of biomedical research training in the USA[13] and focused on bioinformatics and computational biology.[14] The result was a new grant programme to support PhD training in Bioinformatics and Computational Biology. The purpose of this programme was to encourage an interdisciplinary training environment for students coming from either a biological science background (with strong undergraduate training in quantitative skills) or a math/ computer science background, with the goal of training a cadre of scientists whose primary disciplinary identification is bioinformatics. The NIGMS has funded three predoctoral training grants under this programme, at Washington University in St Louis; the University of California, San Diego; and Stanford University.

The NSF also provides support for cross-disciplinary predoctoral training in bioinformatics through its Integrative Graduate Education and Research Training (IGERT) programme. The bioinformatics training programme at the University of California, Los Angeles is an example of one such programme. Table 2 lists funding agencies or organisations that

support bioinformatics training at all levels, from undergraduate to early career development. While an effort was made to provide a current listing, these resources may rapidly change and become outdated. A comprehensive listing of NIH and NSF research and training funding opportunities, including individual fellowship and career awards, can be found on the NIH Bioinformatics Training webpage.[15]

CHALLENGES FOR PHD BIOINFORMATICS TRAINING PROGRAMMES

There are inherent challenges and obstacles to establishing an integrated and cross-disciplinary PhD training programme in bioinformatics. The principal challenges are as follows.

Faculty

The curriculum needs to be flexible and integrated

There is a dearth of bioinformatics' faculties to train students

The basic conundrum is that there is a dearth of research scientists, including faculty members, who are trained in bioinformatics, which is why there is a critical need for these training programmes. Thus the number of faculties who can assemble and teach a bioinformatics curriculum, and can mentor trainees, is sparse, and generally such faculty members are scattered among departments or programmes that traditionally have not interacted. This problem is exacerbated at medical schools that lack departments in math and the computer sciences. A strong institutional commitment is necessary to bring together a core faculty from disparate departments and to encourage an environment in which cross-disciplinary research and training can take place.

Students

Biological and computational scientists must learn to communicate with each other

Strong institutional commitment is needed to build and coalesce a bioinformatics training faculty

Two scenarios are required to train students from different backgrounds

The challenge is to channel bioinformaticists to academia as well industry

The undergraduate students who enter bioinformatics programmes are likely to have diverse backgrounds. Training programmes need to have two scenarios for training students, one for students whose principal undergraduate background is in the biological sciences, and a second for students whose background is in computer science and mathematics. Students whose primary training has been in the biological sciences will require a strong set of quantitative skills in order to master a bioinformatics curriculum.

Programme

The scientific breadth of the programme must be defined. Scientific strengths vary at different institutions, and a decision will have to be made as to what areas of biomedical science (such as genomics, proteomics or structural biology) can successfully be integrated into a bioinformatics training programme. Any programme will need to address both the science and the culture of the disparate fields, and immerse students in both the theory and the experimental aspects of bioinformatics research. A flexible curriculum will be needed to ensure that students from each background can complete the PhD in a reasonable period of time (five years), while acquiring core competency in the biological and computer sciences.

Culture

The different cultures of the biological sciences and the computer sciences must be bridged to establish a bioinformatics identity that students and faculty can embrace, and to ensure that biological scientists, computational scientists and bioinformaticists can communicate with each other.

Career pathways

There are career opportunities in both the academic and non-academic sectors, and students should be advised of the rewards of each. Because of the attractive job opportunities in the private sector, another challenge will be to avoid the attrition of students who are lured to non-academic positions prior to completing the PhD.

FEATURES OF PHD BIOINFORMATICS TRAINING PROGRAMMES

There must be a programme rather than a set of courses to train bioinformaticists

Many of the important features of an integrated training programme in bioinformatics and computational biology are a logical consequence of the challenges noted above. Clearly there must be an emphasis on a programme of study, rather than simply a set of courses. The Report of the NIGMS *Ad Hoc* Bioinformatics Training committee[14] and two recent editorials[16,17] summarise many of these features, including:

- a collaborative infrastructure between biologists, computer scientists and mathematicians that involves faculty from a broad range of departments;
- a flexible curriculum and research programme that is tailored to the different backgrounds of the incoming students, yet can be accomplished in five years;
- rotations that expose students early in their training to the science and culture of the relevant disciplines that they did not study as undergraduates;
- a course curriculum that confers core competency in biomedical science, computer science, statistics and bioinformatics;
- research projects that address important biological problems;
- joint mentors from biological and computational disciplines;
- participation in forums, such as journal clubs, seminar series and scientific meetings, that foster student interactions and a common identity grounded in bioinformatics, while bridging the cultural and communication barriers between disciplines;
- internships that expose students to non-academic research environments;
- career advising to ensure that students fully consider academic and non-academic options;
- teaching opportunities and instruction in pedagogy to promote the training of future faculty members who can educate the next generation of bioinformaticists; and
- instruction in ethical issues that addresses the impact of computer technology on science and society.

The question of what a bioinformatics curriculum should cover is a matter of debate, but the basic elements should include instruction in algorithms, computer programming, database management, sequence alignment and analysis, and molecular dynamics, as well as training in statistics theory and methods.[18] A more detailed listing of potential core and elective topics can be found in an editorial by Russ Altman.[16]

FUTURE OPPORTUNITIES

The NIH Bioinformatics Training webpage lists a multitude of institutional and individual opportunities currently available at the NIH, the NSF and elsewhere, for training at the predoctoral, postdoctoral and more advanced career development levels.[15] Inevitably, the development of PhD training programmes will have an impact on training and curriculum development in undergraduate and master's degree programmes. In 1999, in a broad effort to develop biomedical computing research and training resources for the scientific community, the NIH launched the Biomedical Information Science and Technology Initiative (BISTI) and reported a series of recommendations to implement the BISTI goals.[19–21] The most recent report[21] focuses on bioengineering and bioinformatics research training, education and career development. This report identifies gaps in research training and career development at all levels, identifies obstacles that need to be

Further development of cross-disciplinary training with a focus on quantitative biology, computational biology and integrative systems modelling will be needed

addressed, and encourages collaborative efforts between the NIH, the NSF and industry. The report recommends further development of opportunities for cross-disciplinary training focused on quantitative biology, computational biology and integrative systems modelling. Undoubtedly, training in bioinformatics will continue to evolve, and it is hoped that this training will eventually meet the growing need for bioinformatics teachers and researchers in industry and academia.

Acknowledgments
I wish to acknowledge the GREATMAIL listserv members of the AAMC who responded to my request for information on bioinformatics and provided important updates that have been included in this article. I also wish to thank Ann Dieffenbach for her helpful comments.

References

1. Roos, D. S. (2001), 'Bioinformatics – trying to swim in a sea of data', *Science*, Vol. 291, pp. 1260–1261.
2. Botstein, D. (2000), 'Chat Transcript: David Botstein on bioinformatics. . .funding, education, jobs, and the future', *Science Next Wave*, 15th September 2000 (URL: http://nextwave.sciencemag.org/cgi/content/full/2000/09/14/18).
3. Bioinformatics at the NIH webpage, 'Working definition of bioinformatics and computational biology', 17th July 2000 (URL: http://grants2.nih.gov/grants/bistic/bistic.cfm).
4. University of Minnesota webpage, 'What is bioinformatics?', 9th July 2002 (URL: http://www.binf.umn.edu/whatsbinf.html).
5. Taylor, C. (2000), 'Jobs galore', *Science Next Wave*, Bioinformatics Feature Index, 1st September 2000 (URL: http://nextwave.sciencemag.org/cgi/content/full/2000/08/23/1).
6. Schachter, B. (2002), 'Informatics moves to the head of the class', *Bio-IT World*, 12th June (URL: http://www.bio-itworld.com/archive/061202/class.html).
7. Diehl, B. and Cummings, C. (2000), 'A Master's Degree in biotechnology: your bridge to a bioinformatics future?', *Science Next Wave*, Bioinformatics Feature Index, 1st September 2000 (URL: http://nextwave.sciencemag.org/cgi/content/full/2000/08/30/9).
8. Biotech web site (1998), University Bioinformatics Programs (URL: http://biotech.icmb.utexas.edu/pages/bioinform/biprograms_us.html).
9. Smith, C.M. (2001), 'Bioinformatics centers & institutions', CMS Molecular Biology Resource (URL: http://www.unl.edu/stc-95/ResTools/biotools/biotools4.html).
10. Bio-IT World, 12th June 2002, Degreed Programs (URL: http://www.bio-itworld.com/careers/biotrain/index.html).
11. International Society for Computational Biology webpage, 'Degree programs in bioinformatics and computational biology', 2002 (URL: http://www.iscb.org/univ.html).
12. Association of American Medical Colleges (AAMC), Group on Graduate Research and Education, and Training (GREAT) (URL: http://www.aamc.org/members/great/start.htm).
13. Report on the NIGMS Workshop on Future Training of Biomedical Research Scientists, 12th November 1998 (URL: http://www.nigms.nih.gov/news/meetings/training.html).
14. Report of the NIGMS *Ad Hoc* Bioinformatics Training Committee, 22nd March 1999 (URL: http://www.nigms.nih.gov/news/reports/bioinformatics.html).
15. NIH Training Opportunities in Bioinformatics (URL: http://www.nibib1.nih.gov/training/NIHbioinfo.html).
16. Altman, R. B. (1998), 'A curriculum for bioinformatics: the time is ripe', *Bioinformatics*, Vol. 14(7), pp. 549–550.
17. Pearson, W. R. (2001), 'Training for bioinformatics and computational biology', *Bioinformatics*, Vol. 17(9), pp. 761–762.
18. Vingron, M. (2001), 'Bioinformatics needs to adopt statistical thinking', *Bioinformatics*, Vol. 17(5), pp. 389–390.
19. The Biomedical Information Science and Technology Initiative, 1999 (URL: http://www.nih.gov/about/director/060399.htm).
20. Recommendations of the Biomedical Information Science and Technology Initiative Implementation Group, 2000 (URL: http://grants2.nih.gov/grants/bistic/bisti_recommendations.cfm).
21. Assessing Bioengineering and Bioinformatics Research Training, Education and Career Development: Opportunities for NIH and NSF Collaboration, 2001 (URL: http://grants2.nih.gov/grants/bistic/NSFNIHFinalReport824.pdf).