

Articles

Bioinformatics: Current Practice and Future Challenges for Life Science Education

Received for publication, November 11, 2004, and in revised form, December 10, 2004

Catherine Hack^{‡§} and Gary Kendall[¶]

From the [‡]Bioinformatics Research Group and [¶]Faculty of Arts, University of Ulster,
Coleraine BT521SA, United Kingdom

It is widely predicted that the application of high-throughput technologies to the quantification and identification of biological molecules will cause a paradigm shift in the life sciences. However, if the biosciences are to evolve from a predominantly descriptive discipline to an information science, practitioners will require enhanced skills in mathematics, computing, and statistical analysis. Universities have responded to the widely perceived skills gap primarily by developing masters programs in bioinformatics, resulting in a rapid expansion in the provision of postgraduate bioinformatics education. There is, however, a clear need to improve the quantitative and analytical skills of life science undergraduates. This article reviews the response of academia in the United Kingdom and proposes the learning outcomes that graduates should achieve to cope with the new biology. While the analysis discussed here uses the development of bioinformatics education in the United Kingdom as an illustrative example, it is hoped that the issues raised will resonate with all those involved in curriculum development in the life sciences.

Keywords: Bioinformatics, postgraduate, undergraduate, quantitative and analytical skills.

The development of technologies for the large-scale quantification and identification of biological molecules combined with advances in computing technologies and the internet has served to facilitate the delivery of large volumes of biological data to the scientists' desktop. By the time the human genome sequence was published in 2001, the rate of DNA sequencing had increased 2,000-fold since the inception of the technology in 1986. The increased productivity was gained through automation, miniaturization, and integration of technologies; applying this approach to the analyses of other biological molecules including mRNA, proteins, and metabolites (e.g. [1]) has resulted in a massive increase in the generation of biological data. This data has been made easily accessible, in part due to publications such as the Molecular Biology Database Collection [2], an annual listing of the best databases publicly available to the biological community. Analysis of the collection reveals the steady growth in the quality and size of the databases (Fig. 1), with the 2004 edition containing 548 databases classified into 11 categories (Table I).

As the volumes of data increased, the pressing need for practitioners with a good understanding of biology combined with computational and analytical skills became apparent. The first cohort of bioinformaticians were, by necessity, self taught; predominantly biologists who realized they required computational methods to facilitate

the analysis of biological data. These early practitioners were much in demand; often headhunted by companies seeking employees with a sound understanding of biology but also with competency in mathematics, statistics, and computing.

DEVELOPMENT OF MASTERS PROGRAMS IN BIOINFORMATICS

By the late 1990s there was evidently a skills gap, with several European national research organizations calling for the development of postgraduate bioinformatics programs [7–9]. The primary response by Universities in the United Kingdom was to develop masters-level bioinformatics courses, and the past decade has seen a rapid increase in the provision of postgraduate education in bioinformatics (Fig. 2). Course development teams had to face several hurdles in the development of these programs. Bioinformatics was still a poorly defined academic area and faculty staff with specific expertise in bioinformatics were in short supply. Added to this, many of the programs were open to graduates from a diverse range of academic backgrounds.

Undoubtedly, the availability of a wide range of internet resources helped the development of these fledgling course. In 2001, the Education Committee of the International Society for Computational Biologists (ISCB)¹ [10], the professional body for bioinformaticians produced a

[§] To whom correspondence should be addressed: Bioinformatics Research Group, University of Ulster, Coleraine BT521SA, United Kingdom. E-mail: cj.hack@ulster.ac.uk.

¹ The abbreviations used are: ISCB, International Society for Computational Biologists; NCBI, National Center for Biotechnology Information; QAA, Quality Assurance Agency for Higher Education in the United Kingdom.

consultation document on the content of bioinformatics programs, summarized in Table II, while many of the large database curators such as National Center for Biotechnol-

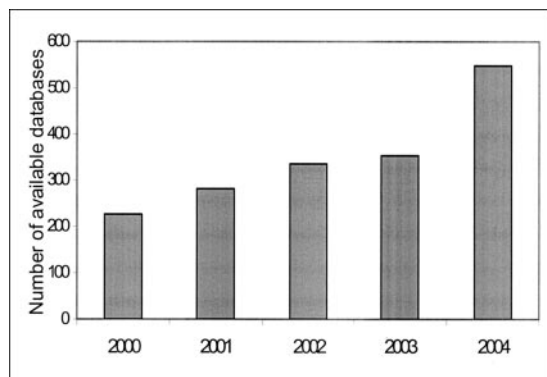


FIG. 1. Growth in number of databases listed in the Molecular Biology Database Collection [2-6].

TABLE I

Classification of databases in the 2004 edition of the Molecular Biology Database Collection [2]

Category	No. of databases
Genomic	164
Protein sequences	87
Human/vertebrate genomes	77
Human genes and diseases	77
Structures	64
Nucleotide sequences	59
Microarray/gene expression	39
Metabolic and signaling pathways	33
RNA sequences	32
Proteomics	6
Other	16

FIG. 2. Growth in postgraduate bioinformatics provision in the United Kingdom. The courses accept either graduates from a life science discipline (black) or from any scientific (including life science), engineering, or computing background (white).

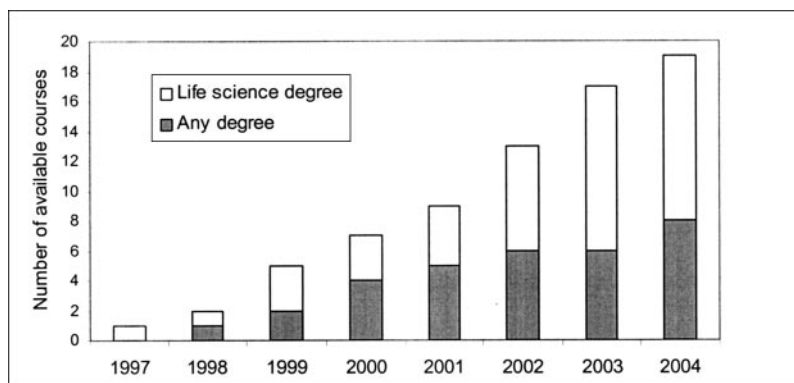


TABLE II

Summary of core content of bioinformatics programs proposed by the Education Committee of the ISCB [10]

Theory and methods	Application areas	Data types
Algorithms	Sequence/structure alignment	Protein and genomic sequences
Mathematical/statistical analysis	Phylogenetics	Gel electrophoresis
Data representation	Fragment/genome assembly	Structures
Knowledge representation	Genome comparison	Expression data
Databases and knowledge bases	Biological databases	Spectroscopic
Programming languages	Expression analysis	Kinetic
Graphics and image analysis	Feature extraction	Thermodynamic
Modeling	Structure prediction	Interaction data
Usability engineering	Docking	Images
Technology support	Knowledge extraction	
	Protein-protein interactions	
	Interaction networks	
	Integrated systems	

ogy Information (NCBI) [11] and the European Bioinformatics Institute [12] provided tutorials on their data analysis tools.

The rapid growth in these courses however raised two important questions:

- Are there enough jobs opportunities for the graduates from these programs?
- Is a 1-year program adequate to produce bioinformaticians or are the graduates from these programs merely “power-users” (see Table III).

Analysis of job listings in scientific journals reveals that there remains a strong demand from industry for biologists with numeracy and computing skills. Fig. 3 shows a snapshot of job advertisements in *Nature* [13] evidencing the requirement for employees with both specialist biological knowledge plus skills in bioinformatics. While there appears to be a continuing and increasing demand for these “numerate” biologists, the question remains of whether a 1-year conversion program is sufficient to develop these skills in young biologists.

UNDERGRADUATE PROGRAMS

The growth in undergraduate bioinformatics courses has been slower than for postgraduate programs; there are only six undergraduate courses in Bioinformatics or Bio-computing currently available in the United Kingdom, with a further two being developed for 2005 entry [14]. Undoubtedly, the problems facing postgraduate course development teams outlined previously are exacerbated for a 3- or 4-year undergraduate program. These, when combined with the promotion problems associated with a new academic discipline, may have constrained demand and resulted in more measured growth. However, many mo-

TABLE III

The terms “super-user” and “power-user” are starting to come into use with respect to the different levels of expertise of bioinformaticians; some popularly conceived skill differentials are described below

Super-user	Power-user	Bioinformatician
Familiar with a range of bioinformatics tools, with some understanding of underlying parameters	Good understanding of underlying parameters and algorithms for a wide range of bioinformatics tools	Develop and implement algorithms to produce new bioinformatics tools
No programming knowledge	Appreciate biological models Write programs to link tools into data pipelines or analyze data	Model and simulate biological data Develop new software suitable for commercial or public use
No knowledge of database development	Develop databases to manage private data and integrate with public data	Use intelligent systems approaches for knowledge extraction
Apply basic statistical tools	Understand a range of statistical software tools and apply them to solve real-world problems in biology	Analyze complex data sets

FIG. 3. Posts advertised in *Nature Jobs* during September 2004 [14]. Posts that included a specific requirement for bioinformatics are indicated (■).

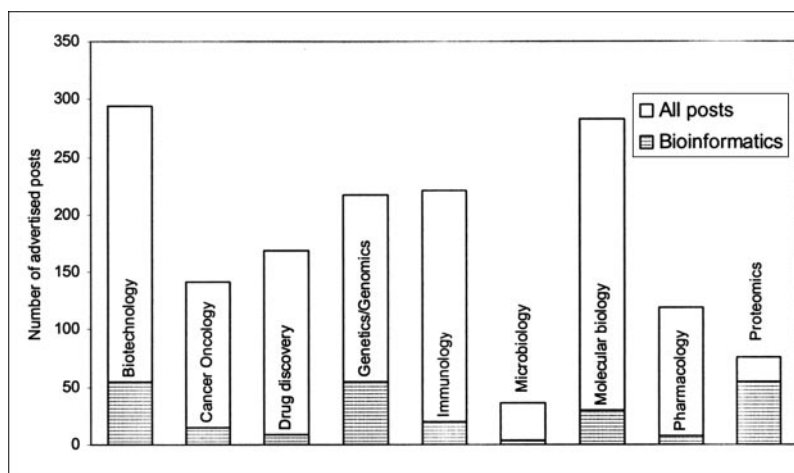


TABLE IV

Proposed competencies in mathematics, statistics, and information technology for life science graduates, indicating the expected “threshold” (or minimum) and “good” level of attainment

	Threshold	Good
Models	An understanding of simple biological models	An ability to use mathematical techniques and analysis to model simple biological systems
Problem solving	Solve biological problems using appropriate mathematical tools	Solve biological problems using appropriate mathematical tools Understand and incorporate approximations where necessary to obtain solutions
Tools and algorithms	Competent use of popular bioinformatics tools for the analysis of data, requiring some understanding of underlying parameters and algorithms	Effective use of popular bioinformatics tools for the analysis of data, requiring a good understanding of underlying parameters and algorithms
Statistics	Use appropriate statistical and analytical methods to analyze and present data, and evaluate uncertainty and significance of results	Use appropriate statistical and analytical methods to analyze and present data, and evaluate uncertainty and significance of results Apply these methods to solve real-world problems in biology
Data resources	Identify and use appropriate resources to find information Understand requirement to manage and integrate data	Identify and use appropriate resources to find information Use databases to manage and integrate data

lecular bioscience programs include the use of information technology and software packages to retrieve and analyze biological data, [15–19], yet graduates from these programs are seldom provided with sufficient training in the underlying algorithms to meet the demands of academia and industry.

PROPOSALS AND RECOMMENDATIONS

In 2002, the Quality Assurance Agency for Higher Education in the United Kingdom (QAA) published the bench-

mark statement for the biosciences [20]. The benchmark statements are part of a major project coordinated by the QAA to define the general academic characteristics and standards of honors degrees for each academic discipline in the United Kingdom. For the biosciences, the graduate and key skills related to numeracy and information technology that should be achieved are:

- preparing, processing, interpreting, and presenting data, using appropriate qualitative and quantitative

techniques, statistical programs, spreadsheets, and programs for presenting data visually;

- solving problems by a variety of methods including the use of computers;
- using the internet and other electronic sources critically as a means of communication and a source of information.

As part of the benchmark process, students can achieve either the threshold *i.e.* minimum standard or a good standard of competency. For example, in regard to numerical analysis of data a student attaining the threshold level would be able to record data accurately and to carry out basic manipulation of data (including qualitative data and some statistical analysis when appropriate), while a good graduate would be able to apply relevant advanced numerical skills (including statistical analysis where appropriate) to biological data. Many graduates from biological science degree programs will not achieve the level of competence in numeracy, statistics, and information technology to allow them to succeed in the new data-driven environment of the life sciences.

It is often stated that the biosciences will become an information science akin to physics and chemistry, with practitioners modeling systems and predicting outcomes prior to experimental work and spending more time on data management and analysis. For graduates to succeed in this environment, they will require a more robust training in numeracy and information technology skills. It was therefore interesting to investigate the learning outcomes produced by the physics subject benchmarking group [21]. These were used to inform the proposed competencies in quantitative analysis described in Table IV.

CONCLUSION

The growth in the volume of biological data is transforming biology into an information science, requiring practitioners to have similar levels of quantitative and analytical skills as physicists; this has important implications for curriculum design in the biosciences. The primary response by academia in the United Kingdom has been the development of postgraduate bioinformatics programs, and the past 5 years has seen a rapid increase in provision at this level. However, the growing skills gap in the life sciences will not be breached by masters programs alone. Teaching of the life sciences at undergraduate level has not yet adapted to this change, and graduates with good first degrees often lack the skills required to succeed in the new data-driven environment. In this article we propose

that the expected learning outcomes for life science graduates are revised, and the standards currently in place for physicists used as a starting point for the development of a curriculum more suited to modern biology. For students to cope with this more robust approach, they will need to enter the university environment with a sound education in mathematics; this message has to be fed into schools for the predicted paradigm shift in the life sciences to be realized.

REFERENCES

- [1] D. Delneri, F. L. Brancia, S. G. Oliver (2001) Towards a truly integrative biology through the functional genomics of yeast, *Curr. Opin. Biotechnol.* **12**, 87–91.
- [2] M. Y. Galperin (2004) The Molecular Biology Database Collection: 2004 update, *Nucleic Acids Res.* **32**, D3–D22.
- [3] A. D. Baxevanis (2003) The Molecular Biology Database Collection: 2003 update, *Nucleic Acids Res.* **31**, 1–12.
- [4] A. D. Baxevanis (2002) The Molecular Biology Database Collection: 2002 update, *Nucleic Acids Res.* **30**, 1–12.
- [5] A. D. Baxevanis (2001) The Molecular Biology Database Collection: an updated compilation of biological database resources, *Nucleic Acids Res.* **29**, 1–10.
- [6] A. D. Baxevanis (2000) The Molecular Biology Database Collection: an on-line compilation of relevant database resources, *Nucleic Acids Res.* **28**, 1–7.
- [7] M. MacLean, C. Miles (1999) Swift action required to close the skills gap in bioinformatics, *Nature* **401**, 10.
- [8] H. McCabe (1999) France losing genome race says report, *Nature* **400**, 199.
- [9] A. Abbott (1999) German agency to boost bioinformatics, *Nature* **401**, 102.
- [10] Education Working Group, International Society for Computational Biologists (update April 2001) www.iscb.org/iscb-edu.shtml.
- [11] National Center for Biotechnology Information (update 2004) www.ncbi.nlm.nih.gov/Education.
- [12] European Bioinformatics Institute (update 2004) www.ebi.ac.uk/2can/tutorials/index.html.
- [13] Nature Jobs (September 2004) www.nature.org/jobs.
- [14] Universities and Colleges Admission Service (September 2004) www.ucas.ac.uk.
- [15] J. A. Boyle (2004) Bioinformatics in undergraduate education: Practical examples, *Biochem. Mol. Biol. Educ.* **32**, 236–238.
- [16] N. B. Centeno, J. Villà-Freixa, B. Oliva (2003) Teaching structural bioinformatics at the undergraduate level, *Biochem. Mol. Biol. Educ.* **31**, 386–391.
- [17] A. L. Feig, E. Jabri (2002) Incorporation of bioinformatics exercises into the undergraduate biochemistry curriculum, *Biochem. Mol. Biol. Educ.* **30**, 224–231.
- [18] J. L. Brewster, K. B. Beason, T. T. Eckdahl, I. M. Evans (2004) The microarray revolution: Perspectives from educators, *Biochem. Mol. Biol. Educ.* **32**, 217–227.
- [19] J. K. Tillotson (2002) Strategies for introducing computer technologies into a biology laboratory program, *Biochem. Mol. Biol. Educ.* **30**, 232–234.
- [20] Quality Assurance Agency for Higher Education in the UK (2002) Subject benchmark statements academic standards—Biosciences, www.qaa.ac.uk/crntwork/benchmark/phase2/biosciences.htm.
- [21] Quality Assurance Agency for Higher Education in the UK (2002) Subject benchmark statements academic standards—Physics, astronomy and astrophysics, www.qaa.ac.uk/crntwork/benchmark/phase2/physics.htm.