Reflections on integrating bioinformatics into the undergraduate curriculum: the Lancaster experience

Derek Gatherer

Division of Biomedical & Life Sciences

Faculty of Health & Medicine

Lancaster University

Lancaster LA1 4YW

d.gatherer@lancaster.ac.uk

Twitter: @DerekGatherer

Abstract:

Bioinformatics is an essential discipline for biologists. It also has a reputation of being difficult for those without a strong quantitative and computer science background. At Lancaster University, we have developed modules for the integration of bioinformatics skills training into our undergraduate biology degree portfolio. This article describes those modules, situating them in the context of the accumulated quarter century of literature on bioinformatics education. The constant evolution of bioinformatics as a discipline, is emphasised, drawing attention to the continual necessity to revise and upgrade those skills being taught, even at undergraduate level. Our overarching aim is to equip students both with a portfolio of skills in the currently most essential bioinformatics tools, and with the confidence to continue their own bioinformatics skills development at postgraduate or professional level.

What is bioinformatics?

Most of the readers of this article will probably know the answer to the above question and, if they read further, may wonder why I feel it necessary to offer a potted history of the field. I do this because it the main contention of this paper that bioinformatics teaching is in a greater state of flux than other branches of biological science education, and that we can only decide what we need to teach now in bioinformatics by considering what was taught in the past. In the light of these issues, I then present the new curriculum for undergraduate bioinformatics at Lancaster University, outlining how it has developed since 2013 and how I think it is likely to develop into the middle of the next decade.

As the name suggests, bioinformatics might be regarded as anything that can be done on a computer that is of relevance to biology. An occasionally undignified scramble for precedence as the inventor of the word "bioinformatics" was ended by the eventual collective acknowledgment that the first usage was by Hogeweg in 1978 [1]. In practice, however, bioinformatics does not have such a wide definition. The first papers to use the word in its modern sense appeared around 1993 or 1994, for instance those by Boguski [2] and Harper [3], and since then there have been several narrower areas where labour in the field known as bioinformatics has been concentrated. These have varied over the years as funding priorities and intellectual fashions have waxed and waned but, despite this, bioinformatics has been accepted for at least the last two decades as an essential discipline within biology. Consequently, the lack of bioinformatics skills among biology graduates is regularly lamented by both the pharmaceutical industry, which has historically been one of the major career destinations for those interested in bioinformatics, and by UK central government as part of a more general anxiety concerning lack of quantitative skills among British graduates. In the words of one report from 2017: "Data analytics, especially bioinformatics, appear to be particularly vulnerable" [4]. National initiatives in the UK to stimulate "Science, Technology, Engineering and Mathematics" (STEM) have regularly included development of bioinformatics skills as one of their key goals [5].

However, due to the rapidity of technological advancement in biology and transformation of the field into a "big data" science [6, 7], it has not always been clear exactly what bioinformatics skills need to be developed among biology graduates. Prior to the launch of the Human Genome Project (HGP) in 1990, bioinformatics was seen very much as an eccentric alternative occupation for those whose careers as laboratory-based researchers had foundered. Despite roots going back to the 1950s, and a modestly thriving literature, bioinformatics was a backwater of science. Suddenly in the mid-1990s, it became hugely in vogue, and the rebranding of Oxford University Press's journal *Computer Applications in the Biosciences* as *Bioinformatics* in 1998, marked a coming-of-age moment. The late 1990s saw the simultaneous mass desertion of academia by bioinformaticians for higher-paid jobs in the pharmaceutical industry – an industry eager to put the data of the HGP to its own use – and the rapid development of one-year masters-level courses in bioinformatics by those who remained. The bioinformatics "gold rush" had arrived. For a flavour of the time, see Brass [8]. For a more detailed account, see Leendert den Besten [9].

The turn of the millennium saw the peak of this first wave of bioinformatics. The bursting of the "dot.com" bubble on 11th March 2000 and the further stock market slump following 11th September 2001, confronted many biotech companies with a withdrawal of investor capital and consequent liquidation or hostile merger. These events occurred just as the HGP was drawing to a close and its results were becoming public domain. A bruised pharmaceutical industry began to move away from the analysis of the genome itself ("target discovery") to specific drug design projects on what had been discovered ("target validation" and "lead discovery") [10]. Those sequence analysts who survived the initial financial crash in industry now found themselves elbowed aside by other bioinformaticians specialising in the analysis of 3-dimensional protein structures and how these interact with drug molecules – the sub-discipline of computer-assisted drug design, or “docking” (since the drugs “dock” into small crevices in the proteins). Crucially, dockers often had more of a background in chemistry than the molecular biology-trained sequence analysts.

Meanwhile, in academic bioinformatics, attention during the first five years of the new millennium turned away from genome sequencing and became oriented towards gene expression analysis using microarrays [11]. Although the major genome projects of the late 90s had been massive undertakings by the standards of previous molecular biology, the advent of microarray genomics and other “-omics” technologies such as proteomics, brought bioinformatics for the first time into the territory of a “big data” science. Omics practitioners, confronted with the problem of making sense of all their data, reached out to the biochemical discipline of metabolic control theory, which had for many years been wrestling with the problems of how to model far smaller-scale biochemical networks. The result was the birth of “systems biology” [12], and an influx of statisticians, mathematicians and computer scientists into biology. For a short while, it seemed as if most academic bioinformaticians were intent on rebranding themselves as systems biologists or systems bioinformaticians. Network analysis tools became the new centre of attention. However, just as this new mainstream in bioinformatics was becoming established, it was once again undermined, not this time by market forces and international politics, but by technological developments.

In the late 1990s, while the HGP was still underway, novel sequencing technologies began to be developed, with an eye to faster and cheaper sequence analysis on a grand scale – “deep sequencing”. Many of these technologies were highly innovative and initially beset with multiple technical and engineering problems. However, by the end of the first decade of the 21st century, these difficulties began to be solved and deep sequencing entered the research mainstream [13, 14]. Even microarray analysis, although barely a decade old, began to be edged out by deep sequencing-based transcriptomics as the preferred method for studying gene activity [15]. As the third decade of the present century approaches, another technological shift is underway, as long-read sequencing technologies begin to edge out the short-read technologies of the first wave of deep sequencing [16]. Table 1 summarises the rapid development of bioinformatics during this time, identifying the main trends in molecular biology and how they have impacted bioinformatics. It is evident that anyone trained in bioinformatics in the 1990s or even in the 2000s, will be seriously in need of a refresher course.

Table 1 also demonstrates how bioinformatics has always been both a discipline that creates new software and one in which that software is put to use. Those who wish to have a career as bioinformaticians need to learn how to write computer programmes and, furthermore, to be prepared to learn new computer languages every few years as these are adopted into the field. Bioinformatics has benefitted over the years by influxes of computer science graduates, particularly at times of transition, e.g. when microarrays, systems biology or deep sequencing made their first appearances each with a whole raft of new problems to be solved. Not all bioinformaticians, however, are full-time software developers. Many spend most of their time using existing software tools to analyse data produced in the lab, and need to know only enough programming to be able to organize their data workflows. This distinction between the “pure” bioinformaticians engaged in software development, and the “applied” bioinformaticians engaged in data analysis is often based on undergraduate degree background: computer scientists being the former and biologists the latter. Teaching bioinformatics in a mixed-background Masters-level course often feels like a struggle to explain biology to computer scientists while simultaneously explaining computing to biologists. The focus of this article, however, is on teaching bioinformatics to biology undergraduates. This is a narrower remit, but one which presents its own challenges.

Years	**Pre-1990**	**1990-2001**	**2002-2008**	**2008-2016**	**Post-2016**
Era	*Pre-HGP*	*HGP and pharma/biotech "gold-rush"*	*Early Omics and systems biology*	*Deep sequencing*	*Long-read deep sequencing*
Lab techniques going mainstream	Cloning and Sanger sequencing	Automated Sanger sequencing	Microarrays, proteomics, metabolomics, lipidomics	Short-read deep sequencing, transcriptomics, metagenomics	Nanopore sequencing
Bioinformatics techniques in vogue	Plasmid mapping, base-calling, alignment, search, neighbour-joining	Maximum likelihood, accelerated search, recombination detection, Hidden Markov Modelling	Cluster detection, self-organizing maps, vector support machines, homology modelling	Bayesian phylodynamics, short-read alignment, *de novo* assembly, transcriptome reconstruction	Long-read assembly
Software tools becoming required skills	Clustal [17], FASTA [18], PHYLIP [19],	BLAST [20], Artemis [21], MEGA [22], DAMBE [23],	Jalview [33], BioConductor [34],	BWA [38], Bowtie [39], TopHat [40],	Canu [45], GraphMap [46],

	Wisconsin Package (Genetic Computing Group)	EMBOSS [24], ACEDB [25], DNASP [26], PAML [27], Simplot [28], RasMol [29], HMMER [30], Pfam [31], GeneWise [32]	Cytoscape [35], Chimera [36], Swiss-Model [37]	DataMonkey [41], Galaxy [42], BEAST [43], SPREAD [44]	Minimap2 [47], Nextstrain [48]
Programming languages/platforms entering use	C++, Fortran	Perl, Java, PHP, Javascript	Python, R, SBML, SOAP	Jupyter, Julia, Ruby, Taverna, Bio-Linux	BioCompute, Rosalind

Table 1: The evolution of bioinformatics

The emergence of bioinformatics curricula

Table 1 may also be read as an exercise in the bioinformatics sub-discipline of “workbenching”, the heyday of which happened around the turn of the millennium. Workbenchers focussed on defining a minimum toolkit for bioinformatics, a suite of “must have” programs. For an example of this approach, see Baker et al [49]. Workbenchers saw their contribution as helping other bioinformaticians to adopt common working methods and shared tool sets, to make starting out in the field easier and to encourage reproducibility and sharing of results. The peak of the field was achieved with the release of Bio-Linux [50], which provided in a single download an entire bioinformatics-oriented operating system pre-installed with hundreds of tools. After the appearance of Bio-Linux, workbenching evaporated as an area of research interest. However, since the last update of Bio-Linux was version 8 in 2014, the necessity for workbenching studies is beginning to arise once more. In applying a workbench ethos to bioinformatics curriculum development, I follow in the footsteps of Greene & Donovan [51]. Before describing this in detail, I shall briefly review previous published bioinformatics curricula and discuss the philosophy behind them.

Although, as mentioned above, bioinformatics in its modern sense was well underway by the mid-90s, it took a while for articles on bioinformatics curriculum development to be written. Altman’s 1998 paper [52] may be the first. Many of these initial efforts were possibly responses to the *ad hoc* nature of the first bioinformatics Masters courses during the 90s “gold rush” era, and the need to inform universities where there were no actual bioinformaticians among the staff, about what was needed if their graduate product was to be fit for purpose in industry. One early influential paper by

Hughey & Karplus [53] reviewed the experience of the first five years of undergraduate bioinformatics teaching at University of California, Santa Cruz, culminating in a degree major in the subject. Dubay et al [54] were the first to describe a Masters curriculum. One of the most striking things in these pioneering papers is their description of the heavy mathematics and engineering pre-requisites for entry to the final year of the course, which would exclude most prospective bioinformatics students in the UK. Some curricula were specifically aimed at computer science students [55, 56] or emphasised the need for a strong computer science grounding [57]. A second surprising feature is how theoretical the courses are, but it must be remembered that they were constructed in an era when far less bioinformatics software had been written, and the emphasis was on teaching students to program new tools rather than master existing ones. The next few years after Hughey & Karplus's seminal 2001 paper saw a huge surge in similar descriptive and discursive considerations of bioinformatics teaching [e.g. 58]. Zatz [59] produced something almost equivalent to a "which guide" to bioinformatics courses. A workbenching perspective was represented by Green & Donovan [51], and Rustad [60] explored if special tools are needed for bioinformatics education. Tusch et al [61] were the first to discuss the technical infrastructure needed to run such a course. Most papers were written from a US perspective, but bioinformatics education became a global phenomenon and Shamsir et al [62], Tastan Bishop et al [63] and Richard et al [64] provided views from other continents.

The precursors of today's mixed "Bioinformatics & …." courses also began to appear in the five years after the turn of the millennium, and these also became subjects for discussion in the burgeoning bioinfo-curricular literature. For instance, see LeBlanc & Dyer [65] on the "Genomics" course at Wheaton College, and Pham et al [66] on the University of Wisconsin-Parkside's "Molecular Biology & Bioinformatics" undergraduate course. Governmental bodies and professional societies also began to take an interest [67, 68] and as early as 2003, discussions began to appear of how to do it all online [69-72], and for those with no prior experience [73]. One interesting trend [74-77] is to choose to emphasize structural bioinformatics, perhaps with an eye to continued demand for drug development "dockers" within the pharmaceutical industry. At the other end of the spectrum, Wightman & Hark [78] emphasise the positive impact bioinformatics education has on the mathematical skills of biologists otherwise disinclined to numeracy.

Debate concerning which methods really are the best has had to wait for more recent publications, where a variety of education research perspectives have been presented, such as: the core competencies approach [79, 80], case study based learning [81], peer-assisted and team-based learning [82-84] and the use of the popular hobbyist 4273pi hardware system [85]. Now bioinformatics education has sufficient scholarly groundwork to be considered a field in its own right and reviews have begun to appear [86].

The Lancaster undergraduate bioinformatics curriculum

The scarcity of bioinformatics provision in the undergraduate curriculum was lamented in 2005 by Hack & Kendal [87]. At Lancaster University, bioinformatics only began to appear in the undergraduate biology curriculum in academic year 2013-2014. In writing about the integration of bioinformatics into the undergraduate curriculum, I follow in the footsteps of various authors [55, 57, 58, 74, 76, 78, 81-84, 88-93]

My own efforts to stand on the shoulders of these giants began initially in a single module, BIOL273 *DNA Technology*. This module had been running for several years and was a techniques-based course focussed on teaching second-year undergraduates the basic skills required in gene cloning, polymerase chain reaction (PCR) and DNA sequencing. To introduce bioinformatics, two of the laboratory sessions were replaced with bioinformatics computer workshops. In the following academic year, bioinformatics content was added to BIOL113 *Genetics* and BIOL313 *Protein Biochemistry*, again by removing some of the existing material to make space for bioinformatics workshops. These module contributions constituted the undergraduate bioinformatics component for the academic years 2014-2015 to 2016-2017 inclusive. In academic year 2017-2018, two major changes were introduced: BIOL313 was redesigned and rebranded as *Proteins: Structure, Function & Evolution*, removing the remnants of classical protein biochemistry from the course to make way for greater bioinformatics content, and a fourth-year course BIOL445 *Bioinformatics* was initiated. This latter course was the first module at Lancaster devoted entirely to bioinformatics. Lancaster University fourth-year modules have a very mixed group of students, divided approximately equally into undergraduates on 4-year extended undergraduate degrees (MSci), postgraduates on a taught Masters degrees (MSc) and postgraduates in the first year of a 4-year joint PhD programme with the Liverpool School of Tropical Medicine (LSTM). Many of the last category are medical or veterinary graduates with several years of professional experience. Those in the second category are divided fairly equally between overseas students, often from China, and our own undergraduates who have opted to stay for an MSc after graduation. BIOL445 is also unusual in that the entire content is delivered in a single week, rather than the 5 or 10 week courses normal at Lancaster. The compression is designed to minimise student travel between Lancaster and Liverpool for the joint LSTM PhD students.

Finally in academic year 2018-2019, bioinformatics content was withdrawn from BIOL273 *DNA Technology*, replaced by material on CRISPR and synthetic biology. A new module BIOL275 *Bioinformatics* was introduced. Just as BIOL445 was the first Lancaster course dedicated entirely to bioinformatics, BIOL275 was the first offered at exclusively undergraduate level. Table 2 summarises the bioinformatics content of the modules mentioned above.

Modules	BIOL113 *Genetics*	BIOL275 *Bioinformatics*	BIOL313 *Proteins: Structure, Function & Evolution*	BIOL445 *Bioinformatics*
Length of course	10 weeks	10 weeks	5 weeks	1 week
Hours of Bioinformatics Lectures	1 (of 12 total)	2 (of 2 total)	6 (of 10 total)	10 (of 10 total)
Hours of Bioinformatics Labs	1.5 (of 4.5 total)	15 (of 15 total)	12 (of 15 total)	15 (of 15 total)
Lecture content	NCBI website and how to search for resources	Introductory and revision lectures, "book-ending" the practical bloc	Selection (dN/dS); phylogenetics; structural bioinformatics	Algorithmic foundations of: a) cluster detection b) alignment c) phylogenetics d) motif detection
Bioinformatics lab content	Systematic literature search technique	Extensive suite of tools (Windows); Introduction to Bio-Linux	Tools (Windows) related to lectures	Extensive suite of tools (Windows and Bio-Linux)

Coursework	None	Report demonstrating competence in techniques	Report analysing protein for selection, and its Bayesian phylogenetics	Report demonstrating competence in techniques
Exam	MCQ	MCQ	Essays (2 from 4 options)	Essays (2 from 3 options)

Table 2: Summary of the Lancaster bioinformatics curriculum

Table 2 illustrates how the bulk of the bioinformatics delivery at Lancaster takes place in 2nd and 4th years. For the majority of undergraduates who are only on three-year degrees, bioinformatics is introduced in 1st year, studied intensively in 2nd year, and then applied to the subject of protein evolution in 3rd year. Those staying for the 4th year receive the same experience as the Masters students. The first three years are designed to develop progression from point-and-click internet-focussed bioinformatics in 1st year, through advanced internet-focussed bioinformatics and basic Windows stand-alone tool use in 2nd year, to a more advanced command of the tools and their application to a specific problem in protein evolution in the 3rd year. For Biochemistry undergraduates, all levels are compulsory. Students from other degree programmes are only compelled to enrol for BIOL113 *Genetics*. This can mean that occasionally students may appear in the 3rd year class without the 2nd year grounding. However, since the tools used within BIOL313 *Proteins: Structure, Function & Evolution* are a subset focussed on protein evolution, the time required to catch up with the rest of the class is limited. The 4th year partly sits within this learning arc insofar as, for the undergraduates on 4-year degrees, it represents a return from the narrower focus of the 3rd year bioinformatics teaching to the general scope and emphasis on mastery of tools introduced in 2nd year. However, since postgraduate students of various types must also be catered for in 4th year, some of whom will be complete beginners, a certain amount of crash course introduction must also be delivered in that module. Whether 4th year undergraduates find this a welcome refresher or an annoying distraction largely depends on the extent to which they absorbed their 2nd year course.

We therefore deliver bioinformatics across our degree programmes as an almost equal mixture of dedicated modules (2nd and 4th year) and integration (1st and 3rd year). Our general trajectory has been away from integration towards dedicated modules, with the removal of bioinformatics from BIOL273 *DNA Technology* in 2018-2019, and the transformation in 2017-2018 of BIOL313 *Protein Biochemistry* into a strongly bioinformatics-oriented *Proteins: Structure, Function & Evolution*. We therefore do not follow the trend of integrating bioinformatics teaching as a minor component of several modules (e.g. Furge et al [94], or for an extreme example the integration of bioinformatics into 10 courses at University of Wisconsin – La Crosse [95]).

Table 3 summarises the software training in our two applications based modules.

Modules	BIOL275 *Bioinformatics* Basic software training	BIOL445 *Bioinformatics* Building on BIOL275 + extra training as indicated
Genome structure viewing	Artemis [21]	
Sequence alignment	Clustal [17] EMBOSS (needle, water) [24], Muscle [96], MAFFT [97]	
Sequence search	BLAST [20], Pfam [31], Prosite [98]	

General tools	EMBOSS (seqret, getorf) [24], Primer-BLAST [99]	
Protein structure	Swiss-Model [37], GOR [100], Coils [101], FPROM [102]	+ Chimera [36]
Phylogenetics/phylodynamics	MEGA [22]	+ Simplot [28], BEAST [43], SPREAD [44]
Evolution	Not covered	DNASP [26], DataMonkey [41]
Next Generation Sequencing	Not covered	BWA [38], Velvet [103]

Table 3: Software training in the Lancaster bioinformatics curriculum, grouped by sub-discipline

Technical delivery of teaching and learning

Likić [91] emphasized the introduction of programming skills and the need to go beyond "internet bioinformatics". My own experience at Lancaster (and in previous bioinformatics teaching in Glasgow) is that teaching biology students a programming language from scratch, requires more time than is available. Within a dedicated Masters course on bioinformatics, programming is of course essential, and several languages need to be mastered (Table 1), even if only those currently in vogue are chosen. However, a decision not to include programming skills in undergraduate bioinformatics need not confine us to internet-focussed techniques. The large quantity of open-source or closed-but-free tools in the field means that there is ample scope for developing expertise that goes beyond simple knowledge of the best bioinformatics websites (although that is important and is included in 1st and 2nd year teaching). Lancaster University deploys AppsAnywhere (https://www.appsanywhere.com) as an interface to deliver a large range of software to all Windows PCs fully connected to the university network, including both computer lab PCs, staff offices and the personal devices of students. Lancaster University is a Windows-only desktop environment, which precludes the deployment of some popular classic Macintosh applications such as MacClade [104]. We use VMWare Horizon (https://www.vmware.com/uk/products/horizon.html) to deliver a virtual Bio-Linux server. The Bio-Linux file system is shared with Windows, allowing students to work on the same files within both Windows and Bio-Linux (c.f. Floriano [105]).

Evolution of learning objectives and assessment methods over time

The extensive changes to course content and delivery described above, have also necessitated change in the learning objectives over the years. At Lancaster, a cascade system of learning objectives is used, starting with over-arching objectives for degree programmes, then devolving more specific learning objectives to each module, with the bottom level consisting of detailed objectives for each teaching session. Approval of new teaching, or of changes to existing teaching, is governed at the module level. Consideration of learning objectives for bioinformatics teaching at Lancaster must therefore take account of the fact that first and third year teaching are embedded within modules - BIOL113 *Genetics* and BIOL313 *Proteins: Structure, Function & Evolution* – where most or some of the content, respectively, is not bioinformatics, and therefore the learning objectives must be congruent with the broader aims of the module. With the modules entitled *Bioinformatics* – BIOL275 and BIOL445 – there is considerably more room to specify relevant learning objectives in more detail.

The supplementary files (see "Availability of teaching materials" below) contain the hand-outs for the various courses on which lists of learning objectives may be found. These have varied from year-to-year as the emphasis of teaching has evolved. To give one particular example, in BIOL113 *Genetics* the 2014-2015 bioinformatics content covered recognition of common sequence formats, retrieval of sequences from GenBank, BLAST searching, multiple alignment and phylogenetic tree building in MEGA. These session-specific detailed learning objectives report upwards to the module learning objectives for BIOL113, among which are two bioinformatics-focussed objectives: 1) to become aware of bioinformatics as a discipline and 2) to be able to perform a set of basic bioinformatics techniques. The specific bioinformatics workshop content in BIOL113 changed on two occasions since 2014-2015, requiring adjustments to the detailed sessional learning objectives but without any need to change the overarching module-level objective pertinent to the bioinformatics content. Similar adjustments have been made to BIOL313 over the years, changing sessional learning objectives while maintaining relevance to those of the module as a whole. In the dedicated bioinformatics modules, by contrast, module-level learning objectives often appear directly at sessional level, sharpened or elaborated as appropriate.

Assessment is also governed at the module level (Table 2). BIOL275 *Bioinformatics* is part of a series of techniques-focussed second year modules, which includes BIOL273 *DNA Technology* in which bioinformatics was previously taught, that are all assessed via equally weighted multiple-choice test and practical report. BIOL313 *Proteins: Structure, Function & Evolution* is assessed via an exam in which two out of the four essay choices will be on bioinformatics – and the students must write one bioinformatics essay – and a practical report, weighted 60:40 respectively. A similar 60:40 exam/report structure is used for BIOL445 *Bioinformatics*. In the first run of BIOL445, the exam was a mixture of problem solving questions and essays, but in subsequent runs only essay questions have been used. This change resulted from an observation in the first run of BIOL445, that there was a very bipolar marks distribution for problem solving questions which skewed the overall exam marks distribution from the bell-curve ideal.

The future of bioinformatics teaching

The future of bioinformatics teaching is difficult to predict. The only things that can confidently be said are that bioinformatics will continue to be of central importance to biology education in general, and that bioinformatics teaching a decade from now will look very different to that of today. Table 1 provides a guide to what would have been taught in each of what I conjecture to be the five eras of the discipline. Many of the earlier era columns of Table 1 contain software of continued usefulness in the present day, whereas other mentioned software has reached obsolescence (compare Table 1 to Table 3). A particularly rapid turnover is evident in the field of sequencing assembly. The decade spent developing tools for short read deep sequencing assembly, and the corresponding time spent teaching those tools, may soon seem an archaic epoch if the latest long read sequencing technologies fulfil their initial promise. A movement away from the recent years of intense focus on sequence assembly may produce a situation reminiscent of the early 2000s, with systems biology and the omics field beginning to figure once more as a main research orientation of bioinformatics. What is new now in 2020 that was not around in 2005 is the potential for bringing virtual reality, artificial intelligence and the internet-of-things approaches into bioinformatics. I speculate that the first of these, especially as applied to protein structure and electron microscopy, would seem to be the most likely to break through soon into the mainstream. Perhaps bioinformatics classes in the year 2030 will be delivered to students encased in headsets, spinning detailed simulations of proteins and cells before their virtual eyes.

In the meantime, students need to have certain fundamental skills, and they need to have skills that are in demand. Some of those skills are challenging to acquire, especially for those who have not had much previous experience of thinking abstractly, or of thinking quantitatively. There are several places where “threshold concepts”, as defined by Meyer & Land [106], need to be grasped. Given the fickle nature of the employment market in bioinformatics, students also need to have a foundation that will enable them to build new bioinformatics skills once graduated and in the workplace. As with so much in higher education, it is the ability to learn to the highest level, rather than what is actually learned, that is the key.

Availability of teaching materials

Selected bioinformatics laboratory class protocols and instructional videos from the courses mentioned are available under CC-BY at https://doi.org/10.17635/lancaster/researchdata/308.

Acknowledgements

I thank all the students, undergraduate and postgraduate, who have participated in my bioinformatics classes at Lancaster University since 2013, and who have been so forthcoming with their feedback. I also thank the Lancaster University Organizational & Educational Development (OED) team for indirectly prompting me to write this article.

Bibliography

1. Hogeweg, P. (1978) Simulating the growth of cellular forms *Simulation.* **31**, 90-96.
2. Boguski, M. S. (1994) Bioinformatics, *Current opinion in genetics & development.* **4**, 383-8.
3. Harper, R. (1994) Access to DNA and protein databases on the Internet, *Current opinion in biotechnology.* **5**, 4-18.
4. BBSRC & MRC (2017) BBSRC and MRC review of vulnerable skills and capabilities. Executive summary in *http://wwwbbsrcacuk/documents/1501-vulnerable-capabilties-report-pdf/*
5. Pain, E. (2015) Report: U.K. postdocs need more skills, *Science.* **http://www.sciencemag.org/careers/2015/05/report-uk-postdocs-need-more-skills**.
6. Leonelli, S. (2019) The challenges of big data biology, *Elife.* **8**.
7. Marx, V. (2013) Biology: The big challenges of big data, *Nature.* **498**, 255-60.
8. Brass, A. (2000) Bioinformatics education--a UK perspective, *Bioinformatics.* **16**, 77-8.
9. Leendert den Besten, M. (2003) *The Rise of Bioinformatics. An historical approach to the emergence of a new scientific discipline*, University of Oxford, Oxford.
10. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. (2011) Principles of early drug discovery, *British journal of pharmacology.* **162**, 1239-49.
11. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science.* **270**, 467-70.
12. Wolkenhauer, O. (2001) Systems biology: the reincarnation of systems theory applied in biology?, *Briefings in bioinformatics.* **2**, 258-70.
13. Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B. & Stein, N. (2006) 454 sequencing put to the test using the complex genome of barley, *BMC Genomics.* **7**, 275.
14. Rougemont, J., Amzallag, A., Iseli, C., Farinelli, L., Xenarios, I. & Naef, F. (2008) Probabilistic base calling of Solexa sequencing data, *BMC bioinformatics.* **9**, 431.
15. Tariq, M. A., Kim, H. J., Jejelowo, O. & Pourmand, N. (2011) Whole-transcriptome RNAseq analysis from minute amount of total RNA, *Nucleic Acids Res.* **39**, e120.
16. Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman,

N. J. & Loose, M. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads, *Nature biotechnology.* **36**, 338-345.
17. Higgins, D. G. & Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer, *Gene.* **73**, 237-44.
18. Pearson, W. R. & Lipman, D. J. (1988) Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A.* **85**, 2444-8.
19. Fink, W. L. (1986) Microcomputers and phylogenetic analysis, *Science.* **234**, 1135-9.
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) Basic local alignment search tool, *J Mol Biol.* **215**, 403-10.
21. Mural, R. J. (2000) ARTEMIS: a tool for displaying and annotating DNA sequence, *Briefings in bioinformatics.* **1**, 199-200.
22. Kumar, S., Tamura, K. & Nei, M. (1994) MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers, *Comput Appl Biosci.* **10**, 189-91.
23. Xia, X. & Xie, Z. (2001) DAMBE: software package for data analysis in molecular biology and evolution, *J Hered.* **92**, 371-3.
24. Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet.* **16**, 276-7.
25. Walsh, S., Anderson, M. & Cartinhour, S. W. (1998) ACEDB: a database for genome information, *Methods Biochem Anal.* **39**, 299-318.
26. Rozas, J. & Rozas, R. (1995) DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data, *Comput Appl Biosci.* **11**, 621-5.
27. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood, *Comput Appl Biosci.* **13**, 555-6.
28. Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., Sheppard, H. W. & Ray, S. C. (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination, *J Virol.* **73**, 152-60.
29. Sayle, R. A. & Milner-White, E. J. (1995) RASMOL: biomolecular graphics for all, *Trends Biochem Sci.* **20**, 374.
30. McClure, M. A., Smith, C. & Elton, P. (1996) Parameterization studies for the SAM and HMMER methods of hidden Markov model generation, *Proc Int Conf Intell Syst Mol Biol.* **4**, 155-64.
31. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins, *Nucleic Acids Res.* **27**, 260-2.
32. Birney, E. & Durbin, R. (2000) Using GeneWise in the Drosophila annotation experiment, *Genome Res.* **10**, 547-8.
33. Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. (2004) The Jalview Java alignment editor, *Bioinformatics.* **20**, 426-7.
34. Dudoit, S., Gentleman, R. C. & Quackenbush, J. (2003) Open source software for the analysis of microarray data, *Biotechniques.* **Suppl**, 45-51.
35. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.* **13**, 2498-504.
36. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis, *J Comput Chem.* **25**, 1605-12.
37. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. (2003) SWISS-MODEL: An automated protein homology-modeling server, *Nucleic Acids Res.* **31**, 3381-5.
38. Li, H. & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform, *Bioinformatics.* **26**, 589-95.

39. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol.* **10**, R25.
40. Trapnell, C., Pachter, L. & Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics.* **25**, 1105-11.
41. Poon, A. F., Frost, S. D. & Pond, S. L. (2009) Detecting signatures of selection from DNA sequences using Datamonkey, *Methods Mol Biol.* **537**, 163-83.
42. Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Gruning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A. & Goecks, J. (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update, *Nucleic Acids Res.* **44**, W3-W10.
43. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7, *Mol Biol Evol.* **29**, 1969-73.
44. Bielejec, F., Rambaut, A., Suchard, M. A. & Lemey, P. (2011) SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics, *Bioinformatics.* **27**, 2910-2.
45. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, *Genome Res.* **27**, 722-736.
46. Sovic, I., Sikic, M., Wilm, A., Fenlon, S. N., Chen, S. & Nagarajan, N. (2016) Fast and sensitive mapping of nanopore sequencing reads with GraphMap, *Nat Commun.* **7**, 11307.
47. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics.* **34**, 3094-3100.
48. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T. & Neher, R. A. (2018) Nextstrain: real-time tracking of pathogen evolution, *Bioinformatics.* **34**, 4121-4123.
49. Baker, P. G., Goble, C. A., Bechhofer, S., Paton, N. W., Stevens, R. & Brass, A. (1999) An ontology for bioinformatics applications, *Bioinformatics.* **15**, 510-20.
50. Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N. & Thurston, M. (2006) Open software for biologists: from famine to feast, *Nature biotechnology.* **24**, 801-3.
51. Greene, K. & Donovan, S. (2005) Ramping up to the Biology Workbench: A Multi-Stage Approach to Bioinformatics Education, *Bioscene: Journal of College Biology Teaching.* **31**, 3-11.
52. Altman, R. B. (1998) A curriculum for bioinformatics: the time is ripe, *Bioinformatics.* **14**, 549-50.
53. Hughey, R. & Karplus, K. (2001) Bioinformatics: a new field in engineering education in *31st Annual Frontiers in Education Conference*, IEEE, Reno, NV.
54. Dubay, C., Brundege, J. M., Hersh, W. & Spackman, K. (2002) Delivering bioinformatics training: bridging the gaps between computer science and biomedicine." in *Proceedings of the AMIA Symposium 2002*, American Medical Informatics Association,
55. Doom, T., Raymer, M., Krane, D. & Garcia, O. (2002) A proposed undergraduate bioinformatics curriculum for computer scientists, *SIGCSE Bull* **34**, 78-81.
56. Morrow, C. & Wilkins, D. (2004) A bioinformatics course in the computer science curriculum. in *Proceedings of the 2nd annual conference on Mid-South College computing* Mid-South College.
57. Burhans, D. T. & Skuse, G. R. (2004) The role of computer science in undergraduate bioinformatics education, *ACM SIGCSE Bulletin.* **36**, 417-421.
58. Zadeh, J. (2006) An undergraduate program in bioinformatics, *Potentials, IEEE.* **25**, 43-46.
59. Zatz, M. M. (2002) Bioinformatics training in the USA, *Briefings in bioinformatics.* **3**, 353-60.
60. Rustad, D. L. (2005) *Developing an interactive web-based learning environment for bioinformatics*, University of Oslo, Oslo, Norway.
61. Tusch, G., Leidig, P., Wolffe, G., Elrod, D. & Strebel, C. (2004) Technology infrastructure supporting a medical & bioinformatics masters degree, *ACM SIGCSE Bulletin.* **36**.
62. Shamsir, M. d. S., Hussein, H., Hashim, S. Z. M. d. & Salim, N. (2006) Educating the educators: Incorporating bioinformatics into biological science education in Malaysia in *National Biology*

Conference pp. https://core.ac.uk/download/files/392/11778523.pdf, UTM, Universiti Teknologi Malaysia.

63. Tastan Bishop, O., Adebiyi, E. F., Alzohairy, A. M., Everett, D., Ghedira, K., Ghouila, A., Kumuthini, J., Mulder, N. J., Panji, S. & Patterton, H. G. (2015) Bioinformatics education--perspectives and challenges out of Africa, *Briefings in bioinformatics.* **16**, 355-64.

64. Richard, R. J. A. & Sriraam, N. (2005) A feasibility study of challenges and opportunities in computational biology: A Malaysian perspective, *American Journal of Applied Sciences* **2**, 1296-1300.

65. LeBlanc, M. D. & Dyer, B. D. (2003) Teaching together: a three-year case study in genomics, *J Comput Sci Coll.* **18**, 85-95.

66. Pham, D. Q. D., Higgs, D. C., Statham, A. & Schleiter, M. K. (2008) Implementation and assessment of a molecular biology and bioinformatics undergraduate degree program, *Biochemistry and Molecular Biology Education.* **36**, 106-115.

67. Canadian Genetic Diseases Network (2002) White Paper: Bioinformatics Curriculum. Recommendations for Undergraduate, Graduate and Professional programs. in

68. Welch, L. R., Schwartz, R. & Lewitter, F. (2012) A report of the Curriculum Task Force of the ISCB Education Committee, *PLoS computational biology.* **8**, e1002570.

69. Searls, D. B. (2012) An online bioinformatics curriculum, *PLoS computational biology.* **8**, e1002632.

70. Searls, D. B. (2014) A New Online Computational Biology Curriculum, *PLoS computational biology.* **10**, e1003662.

71. Tolvanen , M. & Vihinen, M. (2004) Virtual bioinformatics distance learning suite, *Biochemistry and Molecular Biology Education.* **32**, 156-160.

72. Lim, Y. P., Höög, J. O., Gardner, P., Ranganathan, S., Andersson, S., Subbiah, S., Tan, T. W., Hide, W. & Weiss, A. S. (2003) The S-star trial bioinformatics course: An on-line learning success, *Biochemistry and Molecular Biology Education.* **31**, 20-23.

73. Vincent, A. T., Bourbonnais, Y., Brouard, J. S., Deveau, H., Droit, A., Gagné, S. M., Guertin, M., Lemieux, C., Rathier, L., Charette, S. J. & Lagüe, P. (2018) Implementing a web-based introductory bioinformatics course for non-bioinformaticians that incorporates practical exercises, *Biochemistry and Molecular Biology Education.* **46**, 31–38.

74. Centeno, N. B., Villà-Freixa, J. & Oliva, B. (2003) Teaching structural bioinformatics at the undergraduate level, *Biochemistry and Molecular Biology Education.* **31**, 386-391.

75. Badotti, F., Barbosa, A. S., Reis, A. L. M., do Valle, Í. F., Ambrósio, L. & Bitar, M. (2014) Comparative modeling of proteins: A method for engaging students' interest in bioinformatics tools, *Biochemistry and Molecular Biology Education.* **42**, 68–78.

76. Oke, M., Agbalajobi, R., Osifeso, M., Muhammad, B., Lawal, H., Mai, M. & Adegunle , Q. (2018) Design and implementation of structural bioinformatics projects for biological sciences undergraduate students, *Biochemistry and Molecular Biology Education.* **46**, 547–554.

77. Inlow, J. K., Miller, P. & Pittman, B. (2007) Introductory bioinformatics exercises utilizing hemoglobin and chymotrypsin to reinforce the protein sequence-structure–function relationship, *Biochemistry and Molecular Biology Education.* **35**, 119-124.

78. Wightman, B. & Hark, A. T. (2012) Integration of bioinformatics into an undergraduate biology curriculum and the impact on development of mathematical skills, *Biochemistry and Molecular Biology Education.* **40**, 310-319.

79. Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B. & Schneider, M. V. (2014) Bioinformatics curriculum guidelines: toward a definition of core competencies, *PLoS computational biology.* **10**, e1003496.

80. Wu, H. & Palani, A. (2015) Bioinformatics curriculum development and skill sets for bioinformaticians in *Frontiers in Education* pp. 1-7, IEEE, El Paso, TX.

81. Serve, K. M., Clayton, N. & Thomas, M. A. (2013) Using an on-line case study to introduce undergraduate students to bioinformatics, *Journal of the Idaho Academy of Science.* **49**, 35-36.

82. Brown, J. A. L. (2016) Evaluating the effectiveness of a practical inquiry-based learning bioinformatics module on undergraduate student engagement and applied skills, *Biochemistry and Molecular Biology Education.* **44**, 304–313.
83. Smith, J. T., Harris, J. C., Lopez, O. J., Valverde, L. & Borchert, G. M. (2015) "On the job" learning: A bioinformatics course incorporating undergraduates in actual research projects and manuscript submissions, *Biochemistry and Molecular Biology Education.* **43**, 154–161.
84. Shapiro, C., Ayon, C., Moberg-Parker, J., Levis-Fitzgerald, M. & Sanders , E. R. (2013) Strategies for Using Peer-assisted Learning Effectively in an Undergraduate Bioinformatics Course, *Biochemistry and Molecular Biology Education.* **41**, 24–33.
85. Barker, D., Ferrier, D. E., Holland, P. W., Mitchell, J. B., Plaisier, H., Ritchie, M. G. & Smart, S. D. (2013) 4273pi: bioinformatics education on low cost ARM hardware, *BMC bioinformatics.* **14**, 243.
86. Magana, A. J., Taleyarkhan, M., Alvarado, D. R., Kane, M., Springer, J. & Clase, K. (2014) A survey of scholarly literature describing the field of bioinformatics education and bioinformatics educational research, *CBE life sciences education.* **13**, 607-23.
87. Hack , C. & Kendall, G. (2005) Bioinformatics: Current practice and future challenges for life science education, *Biochemistry and Molecular Biology Education.* **33**, 82-85.
88. Zhang, X. Exploring cystic fibrosis using bioinformatics tools: A module designed for the freshman biology course, *Biochemistry and Molecular Biology Education.* **39**, 17–20.
89. Weisman, D. (2010) Incorporating a collaborative web-based virtual laboratory in an undergraduate bioinformatics course. **38**, 4-9.
90. Chapman, B. S., Christmann, J. L. & Thatcher, E. F. (2006) Bioinformatics for undergraduates: Steps toward a quantitative bioscience curriculum, *Biochemistry and Molecular Biology Education.* **34**, 180-186.
91. Likić, V. A. (2006) Computer programming and biomolecular structure studies: A step beyond internet bioinformatics, *Biochemistry and Molecular Biology Education.* **34**, 1-4.
92. Boyle, J. A. (2004) Bioinformatics in undergraduate education: Practical examples, *Biochemistry and Molecular Biology Education.* **32**, 236-238.
93. Feig, A. L. & Jabri, E. (2002) Incorporation of bioinformatics exercises into the undergraduate biochemistry curriculum, *Biochemistry and Molecular Biology Education.* **30**, 224-231.
94. Furge , L. L., Stevens-Truss, R., Moore, D. B. & Langeland, J. A. (2009) Vertical and horizontal integration of bioinformatics education, *Biochemistry and Molecular Biology Education.* **37**, 26-36.
95. Miskowski, J. A., Howard, D. R., Abler, M. L. & Grunwald, S. K. (2007) Design and implementation of an interdepartmental bioinformatics program across life science curricula, *Biochemistry and Molecular Biology Education.* **35**, 9-15.
96. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.* **32**, 1792-7.
97. Katoh, K., Asimenos, G. & Toh, H. (2009) Multiple alignment of DNA sequences with MAFFT, *Methods Mol Biol.* **537**, 39-64.
98. Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. & Bucher, P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors, *Briefings in bioinformatics.* **3**, 265-74.
99. Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. & Madden, T. L. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction, *BMC bioinformatics.* **13**, 134.
100. Garnier, J., Osguthorpe, D. J. & Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *J Mol Biol.* **120**, 97-120.
101. Lupas, A., Van Dyke, M. & Stock, J. (1991) Predicting coiled coils from protein sequences, *Science.* **252**, 1162-4.
102. Solovyev, V. V. & Shahmuradov, I. A. (2003) PromH: Promoters identification using orthologous genomic sequences, *Nucleic Acids Res.* **31**, 3540-5.
103. Zerbino, D. R. (2010) Using the Velvet de novo assembler for short-read sequencing technologies, *Curr Protoc Bioinformatics.* **Chapter 11**, Unit 11 5.

104. Maddison, W. P. & Maddison, D. R. (1989) Interactive analysis of phylogeny and character evolution using the computer program MacClade, *Folia Primatol (Basel).* **53**, 190-202.
105. Floriano, W. B. (2008) A portable bioinformatics course for upper-division undergraduate curriculum in sciences, *Biochemistry and Molecular Biology Education.* **36**, 325-335.
106. Meyer, J. & Land, R. (2003) Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practising within the disciplines in *Enhancing Teaching-Learning Envoronments in Undergraduate Courses*, School of Education, University of Edinburgh, Edinburgh.