BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 2. 88–95
Advance Access publication October 31, 2006
doi:10.1093/bib/bbl035

Evolving research trends in bioinformatics

Carolina Perez-Iratxeta, Miguel A. Andrade-Navarro and Jonathan D. Wren

Submitted: 26th January 2006; Received (in revised form): 7th September 2006

Abstract

The cross-disciplinary nature of bioinformatics entails co-evolution with other biomedical disciplines, whereby some bioinformatics applications become popular in certain disciplines and, in turn, these disciplines influence the focus of future bioinformatics development efforts. We observe here that the growth of computational approaches within various biomedical disciplines is not merely a reflection of a general extended usage of computers and the Internet, but due to the production of useful bioinformatics databases and methods for the rest of the biomedical scientific community. We have used the abstracts stored both in the MEDLINE database of biomedical literature and in NIH-funded project grants, to quantify two effects. First, we examine the biomedical literature as a whole and find that the use of computational methods has become increasingly prevalent across biomedical disciplines over the past three decades, while use of databases and the Internet have been rapidly increasing over the past decade. Second, we study the recent trends in the use of bioinformatics topics. We observe that molecular sequence databases are a widely adopted contribution in biomedicine from the field of bioinformatics, and that microarray analysis is one of the major new topics engaged by the bioinformatics community. Via this analysis, we were able to identify areas of rapid growth in the use of informatics to aid in curriculum planning, development of computational infrastructure and strategies for workforce education and funding.

INTRODUCTION

As biomedical research becomes increasingly data-intensive [1], the use of computational methods have become increasingly common [2]. As such, it is important that biomedical researchers and students are familiar with informatics methods in proportion to the influence these methods have in their fields [3, 4]. In particular, bioinformatics is a scientific discipline that deals with the use of informatics for the analysis and organization of biological data. Here, we estimate how the general community of biomedical researchers uses bioinformatics by analyzing databases that summarize this research.

It has been argued whether current curricula create bioinformatics scientists or just technicians educated in bioinformatics methods [5], but either way we acknowledge that bioinformatics is important to varying degrees for progress in many fields. Most researchers, however, have only a rough feel for the influx of bioinformatics methods into their field, and often such impressions come well after the fact. This results in a slow approach to learning and/or adopting such technologies.

Bioinformatics originated as a cross-disciplinary field as the need for computational solutions to research problems raised in biomedicine [6]. The field evolved as computation became cheaper and widespread during the 80s, as the Internet grew during the 90s, and as high-throughput technologies become common in the 2000s. Ranganathan [7] noted that the boundaries between bioinformatics and biomedical disciplines have become blurred and indeed, recent years have seen the spawning of bioinformatics sub-disciplines such as

Corresponding author. Jonathan D. Wren, PhD, The University of Oklahoma, 101 David L. Boren Blvd., Rm. 2025, Norman, OK 73019, USA. Tel: +1-405-325-3415; Fax: +1-405-325-3442; E-mail: Jonathan.Wren@OU.edu

Carolina Perez-Iratxeta is a Research Associate at the group of **Miguel A. Andrade-Navarro**, who holds a Canadian Research Chair at the University of Ottawa. They both work at the Ottawa Health Research Institute in Ontario (Canada). The focus of the group is mainly on data mining and bioinformatics tools applied to particular biomedical subjects like human inherited diseases.

Jonathan D. Wren is a bioinformatics researcher at the University of Oklahoma, focusing on methods of data integration and knowledge discovery.

cheminformatics [8], neuroinformatics [9] and immunoinformatics [10]. Ostensibly, the only limit to the number of bioinformatics-related sub-disciplines is the number of disciplines themselves. Because of the growing need for an integrative view of biological problems, cross-disciplinary efforts such as these are considered increasingly important to continued scientific progress [11]. Consequently, commercial, government and educational institutions need to make both long-term and short-term strategic decisions about bioinformatics-based resource allocation, training and workforce education based upon their disciplinary foci.

Bialek and Botstein wisely noted that 'Any attempt to create a multidisciplinary curriculum leads to difficult questions about what will be left out' [12]. These difficult questions can become simpler, however, if one has a better quantitative understanding for each discipline of both the growth and focus in their utilization of bioinformatics methods and their overall reliance upon them. We reasoned that the peer-reviewed scientific literature would be an excellent source of information to reveal this because it reflects worldwide research activities, encompasses all sectors of employment and provides the opportunity for historical analysis of trends. Similarly, successful (i.e. funded) extramural grants from the National Institutes of Health (NIH) offer insight into the growth of informatics methods in biomedical research, as they frequently define the present and future resources necessary to conduct research.

Even though every active area of bioinformatics research interest will likely have a 'half-life' of varying length, this analysis should benefit students by alerting them to the benefit of bioinformatics education and/or training, educators to become aware of the effects of bioinformatics in their fields, and for managers/administrators to better plan computational resource allocations.

RESULTS

We first quantified the expansion of the use of three major informatics topics (computation, the Internet and databases) in research over the past three decades (see Methods for details). According to the scientific bibliography deposited in the MEDLINE database, the use of computational methods in biomedical research has increased steadily and consistently since the 70s (Figure 1A). In 1975, computational terms were present in approximately 1.6% of MEDLINE publications, while in 2004 they were present in 10%. The use of online/internet resources or methods in publications has increased from approximately 0.05% in 1975 to 0.87% in 2004, but the most dramatic burst of growth occurred around 1994—about the time graphical navigation software (i.e. web browsers) for the Internet was developed and used on a widespread scale (Figure 1B). Similarly, an increasing number of publications refer to the use of databases in their abstracts (Figure 1C), which reflects not only the growth in the amount of data gathered to date, but also our increased technological capacity for gathering data. A similar trend is observed in the NIH funded research grants. Analysis reveals an even stronger growth trend in the use of computational keywords (Figure 2A) as well as use of the Internet (Figure 2B; see Methods for details). A similar trend towards the proposed use of the Internet in biomedical research can also be seen in Figure 1B, beginning approximately in 1994. In summary, we observe that the use of computation expanded first, then the Internet became more widely used in research, and then databases appeared when those previous developments made them possible. Our next question is how does the use of bioinformatics, and the development of genomics contrast with a more general growth in the use of computers? For this we need to focus on the latest 15 years, since the expansion of bioinformatics and genomics happened more recently.

We compared the growth of the number of references mentioning computational terms to the term 'bioinformatics' itself. Figure 3 indicates that the use of the term 'bioinformatics' increases steadily since the introduction of the term in 1993. Figure 3A indicates that the term 'comput*' does not increase as rapidly in relative terms, though the number of references mentioning the term is still much larger than the number of references mentioning bioinformatics. Something similar can be observed comparing the use of the term 'bioinformatics' with other computational related terms: 'http', 'internet', and 'database' (Figure 3B).

To complete a study of the context of emergence of bioinformatics, we compared the use of the term with terms related to the production of high-throughput biological data that have driven much bioinformatics research (Figure 3C). The terms 'genomics', 'microarrays' and 'proteomics' increase

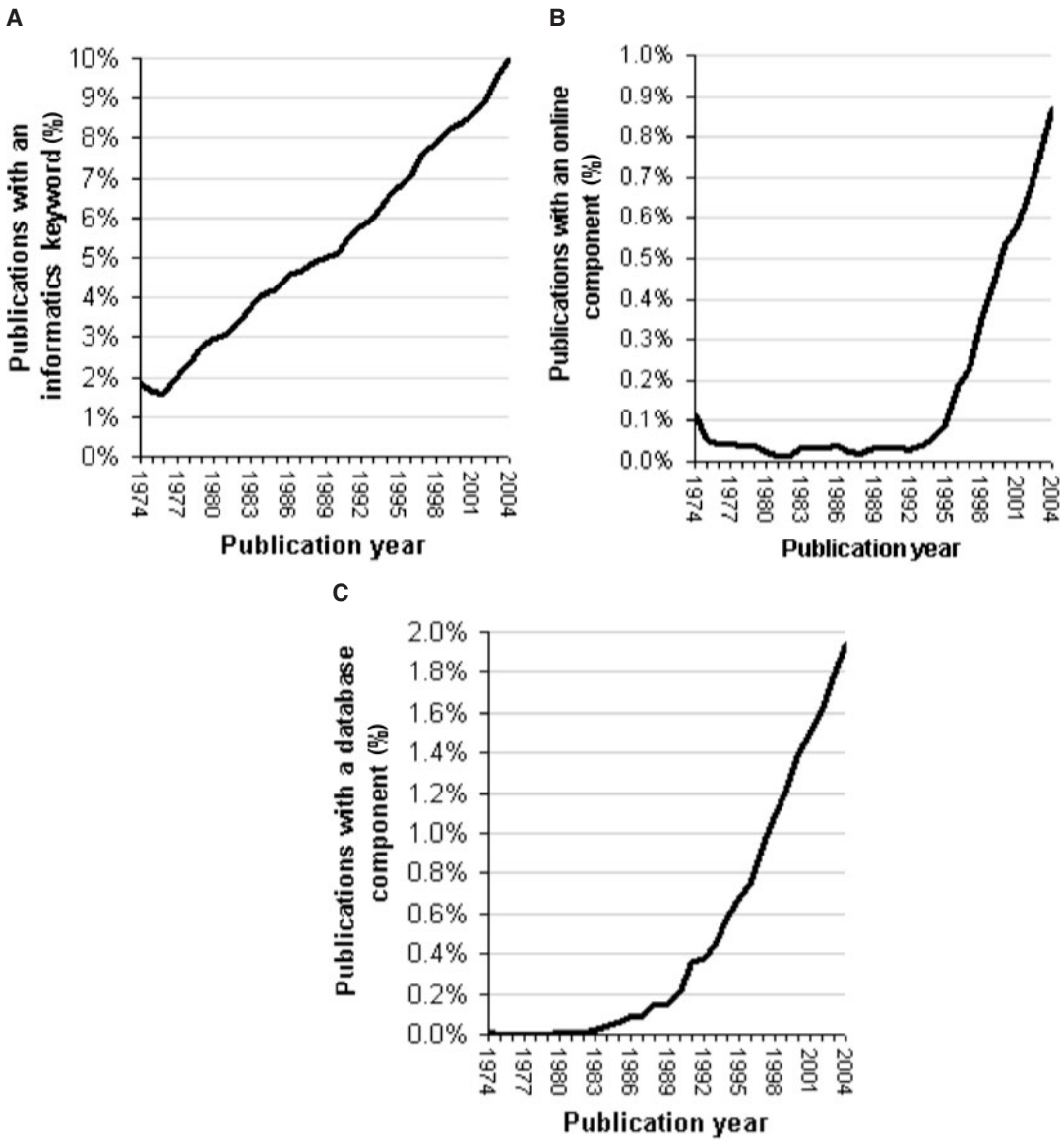

Figure 1: Estimated growth in the use of computational methods by year. (See Methods for details) (**A**) Growth in the percentage of publications containing computational keywords. (**B**) Growth in the percentage of publications that include or use online methods or resources. This trend persists even if only the keywords 'internet' and 'online' are used, but is slightly less pronounced. This consideration is pertinent because one would not expect phrases such as 'world wide web' to occur prior to 1994 and would want to ensure this increase is not merely an artifact of a new vocabulary word entering the literature. (**C**) Growth in the use of databases. Databases are not only useful as a means of organizing data, but are important in data-mining efforts.

continually with a distinct acceleration at 1999, 2000 and 2001, respectively.

The establishment of bioinformatics occurs in a period of already stabilized and accepted use of computation, and is preceded first by the use of the Internet and more immediately by the use of the World Wide Web. A stronger connection seems to happen to the production of massive amounts of genomics data. In summary, we observe that during the last decade the use of genomics and bioinformatics is expanding more rapidly than the use of informatics. The timing of these phenomena suggests that bioinformatics appears as a response to the production of genomics data and not as a byproduct of the evolution of informatics, though it is apparent that the informatics revolution made it possible.

Bioinformatics deals with topics spanning a variety of biological problems and techniques.

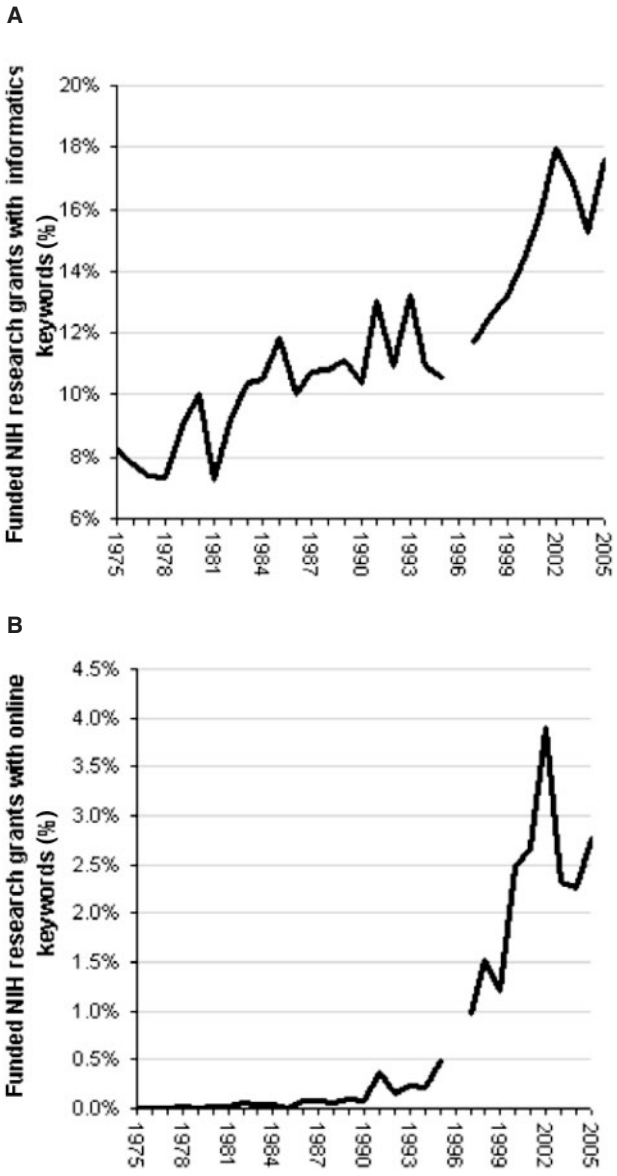

Figure 2: Growth of informatics in NIH funded grants. (**A**) Growth in the use of computational keywords. (**B**) Growth in online keywords. Abstract data for 1996 was not available.

In the following we quantify those topics and the trends in their use by the research community. First, we identified the topics that characterize the field of bioinformatics by examination of the references in MEDLINE to the articles published in the journal *Bioinformatics*, the oldest in the field (Table 1; see Methods for details). Regarding National Library of Medicine (NLM) Medical Subject Headings (MeSH) annotations, three of the top four MeSH terms found in those references are biological databases. The only other high ranked MeSH term describing a particular bioinformatics topic was 'Sequence Analysis, Protein' with a 147-fold over-representation. All other topics, with the exception of 'NLM (U.S.)', were related to computation. There were no biomedical terms unrelated to computation, as might be expected from the methodological focus of the *Bioinformatics* journal. Regarding nouns in titles and abstracts of the references, the results are comparable with those obtained with MeSH. Two terms describing databases ('prodom', a database of protein domains; 'trembl', a database of nucleotide sequences) were at the top of the list. Again, all top terms were related to bioinformatics or computation, with almost all bioinformatics names being related to databases with the exception of two related to sequence formatting and sequence homology searches ('fasta', the popular file format for sequences; and 'waterman' from the Smith & Waterman algorithm).

To identify trends in the use of bioinformatics by the research community, we compared the use of bioinformatics topics (i.e. overrepresented in the *Bioinformatics* journal) in MEDLINE before and after the year 2000 (Table 2). MeSH terms indicate the recent prevalence of genomics and DNA microarray data, databases, and of the computational analysis of protein sequences. The analysis of nouns in titles and abstracts indicates similar trends. It also shows that recently the computational emphasis is more in the operating systems such as 'linux' and 'microsoft' and less in the computer itself such as 'ibm' and 'microcomputer'. This analysis indicates also the decline of some subjects such as 'globin', proteins that are not used as models for the study of protein structure and sequence as much as they used to be.

CONCLUSIONS

The use of informatics in biomedical research has been increasing during the last three decades. Computation first, and then the Internet, were accepted by researchers and allowed the generation and popularization of databases, before the genomics revolution. Bioinformatics flourished with the explosion in the production of biological data from genomes. Genomics allowed the emergence of other high-throughput techniques of biological analysis that are of recent research focus in bioinformatics.

In this study, we have generated a list of bioinformatics topics and analyzed the trends in

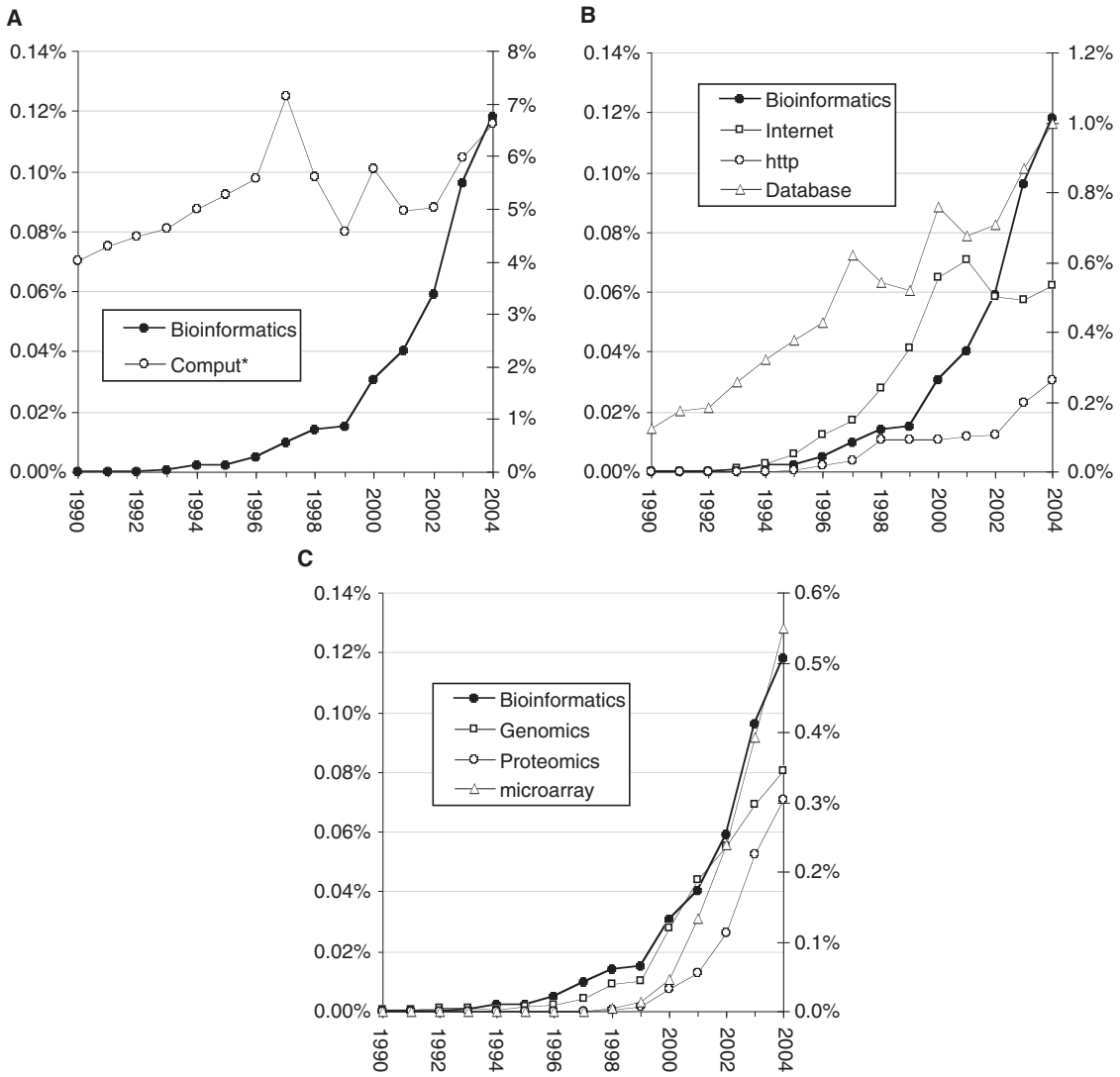

Figure 3: Comparison of the use of the term 'bioinformatics' with other bioinformatics related terms in MEDLINE references. The *y*-axis shows the percentage of entries in MEDLINE in the time period including a given term. The term 'bioinformatics' is always plotted to the scale on the left *y*-axis and other terms are plotted according to the right *y*-axis. (**A**) Comparison with 'comput*'. (**B**) Comparison with other computation related terms: 'database', 'http', and 'internet'. (**C**) Comparison with terms related to high-throughput production of biological data: 'genomics', 'proteomics', and 'microarray'.

their usage. All data are available as supplementary material (see supplementary tables S1–S4; also available at http://www.ogic.ca/projects/BioinfoGrowth06/). Here we summarize these results as a list of topics or aspects of bioinformatics tasks that could be divided into three categories: (1) basic biological knowledge, (2) computational and (3) particular methods and algorithms.

(i) The general thrust of bioinformatics research typically relates to cellular and molecular biology, more specifically, to the mechanisms of encoding and transmitting genomic information, transcriptional control and protein function and structure. Other higher-level concepts such as molecular dynamics, cellular ensembles or interaction between organisms (ecology), belong more to the field of Systems Biology.

(ii) We identified three computational tasks that are part of bioinformatics by examination of the list of words as seen in Table 1. The first is database implementation. The second is basic programming; though we cannot suggest a particular programming language, current trends suggest

Table 1: MeSH terms (left) and words (right) most overrepresented in references in the *CABIOS/Bioinformatics* journal *vs* other references. The fold value is the ratio of the frequencies. For example, 'Databases, Protein' is mentioned 958 times more often in the Bioinformatics set than elsewhere. 'pub' refers to a typical name of the location where software will be distributed in a Unix system. In this table and the following, bold terms indicate terms originated in the Bioinformatics community and italicized terms as related to computation and internet. The complete lists are given as supplementary tables S1 and S2, for MeSH terms, and nouns, respectively

MeSH	Fold	Word	Fold
Databases, Protein	958	**prodom**	2028
Computational Biology	872	*parallelization*	1449
Databases, Genetic	761	**trembl**	1092
Databases, Nucleic acid	624	*ftp*	981
Computing Methodologies	539	**waterman**	880
Programming Languages	497	*linux*	875
Software Design	297	*perl*	831
Pattern Recognition	232	*postscript*	795
Database Management Systems	223	**pfam**	790
Principal Component Analysis	218	*unix*	779
Software	200	**fasta**	722
Information Storage and Retrieval	164	**swissprot**	584
Databases, Factual	162	**prosite**	565
National Library of Medicine (U.S.)	159	*workbench*	535
Software Validation	155	*pub*	524
User–Computer Interface	148	*hmms*	496
Sequence Analysis, Protein	147	*http*	492
Mathematical Computing	139	*html*	482
Internet	138	*c++*	454
Computer Graphics	134	*ascii*	430

Table 2: Bioinformatics characteristic MeSH terms and nouns with the highest increase (top) and decrease (bottom) in MEDLINE references in the period 2000–2003 *vs* the period before. For example, 'microarray' was 77 times used more often in the recent period than before. The term is not in the tagger dictionary and its plural (in the 2nd place) is not stemmed. Most of the nouns are related to bioinformatics and computation with others that regard mostly data and, in particular, genomic high-throughput data. In the MeSH study 'Principal Component Analysis' and 'Databases, Proteins' appeared only in the recent period and therefore could not be assigned a fold value. The complete lists are given as supplementary tables S3 and S4 for MeSH terms and nouns, respectively

MeSH term	Fold	Noun	Fold
Genomics	112.333	microarray	77.340
Databases, Genetic	99.000	microarrays	54.403
Gene Expression Profiling	27.196	**pfam**	20.411
Oligonucleotide Array Sequence Analysis	21.669	**xml**	18.699
		genomics	17.715
Amino Acid Motifs	12.768	*linux*	17.495
Sequence Analysis, Protein	8.617	**bioinformatics**	15.034
Databases, Nucleic Acid	5.465	**trembl**	11.664
Computational Biology	2.953	*http*	8.855
Expressed Sequence Tags	2.630	**annotation**	7.013
Internet	2.428	throughput	6.582
Protein Structure, Tertiary	2.370	**prodom**	6.152
Gene Duplication	2.138	*internet*	5.643
Contig Mapping	1.838	ests	5.286
Amino Acid Substitution	1.653	*datasets*	4.818
Databases	1.224	*graphics*	0.619
Genes	0.056	globin	0.549
Restriction Mapping	0.048	*fortran*	0.546
Genes, Structural	0.040	*ibm*	0.530
DNA Restriction Enzymes	0.030	*microcomputer*	0.240
Methods	0.025		

that a bioinformaticist should be able to use and install software on a computer running a Unix/Linux operating system. The third is web server implementation, given the fact that bioinformatics tools and databases are increasingly being provided through web servers.

(iii) The third emphasis of bioinformatics is in methods of data analysis. These deal with genomics data (gene expression, proteomics, gene prediction), sequence similarity searches (Smith–Waterman, BLAST, Hidden Markov Models), protein evolution (multiple alignment, phylogenetics), protein structure (secondary and tertiary structure prediction, binding site analysis), and data and text mining (database parsing, natural language processing). We observed also mathematical/statistics-based simulations (protein folding, molecular kinetics, molecular interaction simulations), but it can be argued that this is 'computational biology' rather than bioinformatics (e.g. see http://www.bisti.nih.gov/CompuBioDef.pdf).

Regarding the analysis of funded grants, it is unfortunate that unfunded grant abstracts are not included in the CRISP database. So an analysis of whether or not the use of informatics components confers an advantage upon grant applications could be made directly. Rather, we just note that abstracts from NIH funded grants display a much larger use of computational or online terms than the references from published research in general during the same period. Notwithstanding many possible explanations for this (such as possible idiosyncrasies of researchers eligible to apply for NIH grants), we note that the

differences are large. For example, regarding the 2003–2004 period, percentages of abstracts with informatics keywords of approximately 9% in MEDLINE against 16% in grants, and with online keywords of 0.8% in MEDLINE against 2.5% in grants (compare Figures 1 and 2). One possible explanation for this is that the inclusion of such components in a research project confers an advantage regarding NIH funding, but other interpretations are possible as well.

Our study shows that bioinformatics is making an impact in the community of biomedical researchers in general. This happens through a crosstalk of topics between communities, which evolves as new types of biological data are produced, and bioinformatics produces databases and tools for its analysis. We hope that this report may guide persons interested in bioinformatics—employers, students, researchers, funding policy makers—to a better understanding of this relation between bioinformatics and other fields of biology.

METHODS

The MEDLINE database was used for analysis, and contains references to biomedical literature, many of them including an abstract. It would have been optimal to include computer science abstracts in our analysis as well, such as those indexed in the CiteSeer database, but the abstracts are not freely available for downloading and analysis. References in MEDLINE are annotated by the NLM with keywords from a controlled vocabulary thesaurus describing the contents of the article (MeSH; http://www.nlm.nih.gov/mesh/).

For the study of the evolution of the use of informatics in research, we scanned all electronically available MEDLINE abstracts and MeSH headings between 1966 and June 2005 for the presence of keywords suggesting computational methods: 'comput*', '*informatic*', 'algorithm*', 'software' or 'database' (where * represents a wildcard match). Separately, keywords indicating online components employed as part of the study were also flagged: 'internet', 'online', 'world wide web', 'web-based', 'http:*' and 'ftp:*'. These keyword sets were neither intended nor expected to encompass all possible words used to describe informatics methods, but rather were expected to be broadly reflective of their utilization and at the same time minimize false positives. For example, the word 'program' can refer to either computer programs or organizational programs, the word 'interface' could be computational or personal, etc. We performed a similar analysis on a total of 181 436 abstracts of newly funded NIH research grants (the largest category of NIH research funding) that we downloaded from the CRISP database (http://crisp.cit.nih.gov/), for the years 1975–2005.

In order to define an initial set of terms characteristic of bioinformatics as a field, we studied the set of abstracts from the journal *Bioinformatics* (together with its predecessor, *Computer Applications in the Biological Sciences*, which was started in 1985) in the MEDLINE version distributed by the NLM in November 2004. This set contains 2675 references.

Regarding the analysis of titles and abstracts in MEDLINE references, we focused only on nouns, because we have observed that they are the best discriminator for a general subject (such as 'stem cells') [13]. Here, we studied whether nouns from titles and abstracts suggested a similar trend to the MeSH analysis. Nouns were obtained by parsing the text of titles and abstracts using a part of speech tagger (TreeTagger; H. Schmidt, Institute for Computational Linguistics of the University of Stuttgart; http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/). Then, for all the references in the version of MEDLINE used, we registered all nouns used in title and abstract.

We analyzed the trends in the use of bioinformatics terms in MEDLINE by comparing the frequency of references with those in a recent period, 2000–2003 (978 632 references), and before (4 891 837 references). We did not consider references after 2003 because they were not thoroughly annotated with MeSH terms yet.

Key Point

- As the amount of available data grows in various biomedical fields, the need for computational approaches to manage, analyze, predict and model grows as well. Bioinformatics is a cross-disciplinary field that may have been greatly expanded with the genomics revolution, but is now entering other biomedical disciplines and, in turn, co-evolving with them. We use the biomedical literature to study the generation, growth and evolution of bioinformatics methods over time, and find that reliance upon computational methods has been growing steadily over the past three decades across all fields, but at various rates within almost every field. Furthermore, as technology advances, so does the focus of bioinformatics research expand to meet the new challenges.

Acknowledgements
This work was funded by a grant # EPS-0447262 from NSF (JDW). The NLM graciously provided electronic MEDLINE records in XML format. MAA is a recipient of a Canada Research Chair.

References

1. Reichhardt T. It's sink or swim as a tidal wave of data approaches. *Nature* 1999;**399**:517–20.
2. National_Research_Council, BIO 2010: Transforming Undergraduate Education for Future Research Biologists. In: *Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century, Board on Life Sciences, Division of Earth and Life Sciences. 2003*. Washington, DC, USA: National Academies Press.
3. Honts JE. Evolving strategies for the incorporation of bioinformatics within the undergraduate cell biology curriculum. *Cell Biol Educ* 2003;**2**:233–47.
4. Altman RB. A curriculum for bioinformatics: the time is ripe. *Bioinformatics* 1998;**14**:549–50.
5. Pevzner PA. Educating biologists in the 21st century: bioinformatics scientists *vs* bioinformatics technicians. *Bioinformatics* 2004.
6. Roos DS. Computational biology. Bioinformatics—trying to swim in a sea of data. *Science* 2001;**291**:1260–1.
7. Ranganathan S. Bioinformatics education—perspectives and challenges. *PLoS Comput Biol* 2005;**1**:e52.
8. Olsson T, Oprea TI. Cheminformatics: a tool for decision-makers in drug discovery. *Curr Opin Drug Discov Devel* 2001; **4**:308–13.
9. Amari S, *et al*. Neuroinformatics: the integration of shared databases and tools towards integrative neuroscience. *J Integr Neurosci* 2002;**1**:117–28.
10. Brusic V, Petrovsky N. Immunoinformatics—the new kid in town. *Novartis Foundation Symposium* 2003;**254**:3–13; discussion 13–22.
11. Sung NS, *et al*. Science education. Educating future scientists. *Science* 2003;**301**:1485.
12. Bialek W, Botstein D. Introductory science and mathematics education for 21st-Century biologists. *Science* 2004; **303**:788–90.
13. Suomela BP, Andrade MA. Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics* 2005;**6**:75.