

EMBER: a European Multimedia Bioinformatics Educational Resource

Jane E Mabey and Teresa K Attwood

School of Biological Sciences, Stopford Building,
The University of Manchester, Oxford Rd,
Manchester M13 9PT, United Kingdom

A report compiled on behalf of The EMBER Consortium.

Introduction

Bioinformatics has taken centre stage in the post-genomic era. The data over-load arising from the many now-fruitful genome projects has created an insatiable demand for suitably qualified people to build and maintain databases, to design more incisive analysis software, to use disparate databases and software tools, and to understand both the statistical and biological significance of results generated *in silico*. It is rare to find individuals with such a range of skills, yet such scientists are now needed urgently in sequencing centres, research/academic institutes, pharmaceutical/agrochemical companies, software houses and start-up companies. But the rate of growth of this field, and its cross-disciplinary nature, has created a problem: while there are many trained biologists and computer scientists, there are few computer-literate biologists or biology-literate computer scientists. Consequently, there is a dearth of skilled staff in bioinformatics. This is especially problematic for universities, which are less able than large multinational companies to compete for the small numbers of trained individuals emerging from current MSc, MRes or PhD courses.

In an attempt to address the current European skills shortage in bioinformatics, the European Commission has recently funded an innovative new educational project that aims to develop a suite of multimedia bioinformatics educational tools (collectively termed EMBER). EMBER will provide teaching materials for undergraduate and early postgraduate studies; it will comprise a self-contained, interactive Web tutorial in bioinformatics, the equivalent stand-alone course on CD-ROM, and an accompanying introductory textbook. The use of conventional text, coupled with Web- and CD-based media, will ensure that students for whom Internet access is not optimal also have access to the same fundamental level of bioinformatics education.

Ten institutes from around the world are participating in the EMBER project: University of Manchester (United Kingdom); Swiss Institute of Bioinformatics (Switzerland); University of Nijmegen (The Netherlands); University of the Western Cape (South Africa); European Bioinformatics Institute (United Kingdom); Instituto Gulbenkian de Ciência (Portugal); University of Bruxelles (Belgium); Canada Institute for Marine Biosciences (Canada); Research Institute for Genetic Engineering and Biotechnology (Turkey); and Expert Centre for Taxonomic Identification (The Netherlands). All participants are members of the European Molecular Biology Network (EMBNET), with expertise in different areas of biology, bioinformatics and computer science. EMBER will therefore draw on the complementary strengths of its participants, both to develop the package and test its effectiveness as a learning instrument.

The project contains several inter-related workpackages, which encompass five main stages. The first of these address general course design and the necessary collation, revision and unification of existing materials that will form the educational core of EMBER; the last involve course evaluation and standardisation. Specifically, the project will: (1) survey training needs and desired learning outcomes; (2) collate and revise existing teaching materials; (3) unify these text- and Web-based materials to form a new integrated course; (4) create and use formal assessment tools to trial the course throughout Europe; and (5) offer a standardised course in Web-based and CD-ROM formats.

The initial phases of this project aim to tailor the proposed courses to the requirements of industrial and academic employers, specifically by identifying precisely the nature of the current skills shortfall, and defining a minimum standard of required knowledge. We report here the results of these early stages of EMBER.

Evaluating the skills shortage in bioinformatics

To delineate the skills portfolio sought by potential employers of today's graduates in bioinformatics, a questionnaire was composed. Based on the collective skills of the consortium, the questions mapped onto tentative areas to be covered in EMBER. We envisaged these areas as falling into discrete themes - core and advanced bioinformatics, and supplementary material, as shown in Figure 1.

<p><i>Core Bioinformatics</i></p> <ul style="list-style-type: none"> • Biological databases (e.g., sequence and family databases, database technologies) • Principles of sequence analysis (e.g., pairwise sequence analysis, scoring matrices) • Protein structure (e.g., structure classification databases, visualisation) • The genome (e.g., gene prediction, genome annotation, technology platforms) • The transcriptome (e.g., EST data, EST clustering and assembly, microarrays) • The proteome (e.g., 2D gel data, mass spectrometry data, image analysis) <p><i>Advanced Bioinformatics</i></p> <ul style="list-style-type: none"> • Molecular evolution and phylogeny (e.g., biological foundations, terminology, methodologies) • Ontologies in bioinformatics (e.g., Gene ontology, EcoCyc) • Principles of protein structure prediction (e.g., homology modelling, threading) <p><i>Supplementary Material</i></p> <ul style="list-style-type: none"> • Information theory • Basic statistics
--

Figure 1: The provisional syllabus. This figure outlines the basic scheme proposed by EMBER; it comprises several topics that fall into three main categories: (1) core bioinformatics, (2) advanced bioinformatics, and (3) supplementary material.

Within the basic scheme shown in Figure 1, core topics were those that tend to dominate day-to-day analytical and experimental bioinformatics approaches; advanced topics were those requiring, perhaps, a greater degree of analytical and/or mathematical experience; and supplementary topics were those that, while desirable, were regarded as being more challenging to include within a basic bioinformatics course. The idea of the questionnaire was to establish if the scheme was appropriate, and whether any topics had been missed. In an attempt to arouse interest in the survey, the questionnaire was designed to be simple, requiring short easy answers (*i.e.*, predominantly 'yes/no' answers) to questions. However, sections for specific comments were also included for those wishing to express concerns outside the scope of the given questions.

On the basis of their involvement in bioinformatics, biotechnology or drug discovery, a number of European companies and academic institutes were selected as targets for feedback; EMBER participants and other EMBnet nodes were also included in the survey. The total number of contacts made via e-mail during the survey was 188 (30 EMBnet nodes, 145 corporate and 13 academic).

Survey results

A disappointingly small proportion (16%) of the contacts made during the survey responded, details of which are provided in Table 1. Of the companies/institutes ('Academic' and 'Corporate' in Table 1) that responded, eight did not complete the questionnaire because they did not use bioinformatics and, hence, did not seek such graduates. For those that did complete the questionnaire, Table 2 shows the coverage of responses, which are illustrated graphically in Figure 2.

Table 1: Types of contact involved in the survey and the number of responses

Type	No. of Contacts	Response (%)	No response (%)
EMBNET node	30	17	83
Academic	13	15	85
Corporate	145	17	83

From the results provided in Figure 2, it can be seen that a highly desirable skill sought by all contacts is sequence analysis: 100% of the contacts expected a graduate to have working knowledge of sequence similarity and multiple sequence alignment tools. In conjunction with this, 96% expected a graduate to be familiar with different types of databases. Other areas considered to be of primary importance (*i.e.*, where more than 75% of the contacts responded 'yes') were: basic knowledge of genomic and protein structure; gene prediction techniques; basic statistics and information theory; and alignment algorithms. Knowledge of underlying experimental techniques was given less emphasis (*i.e.*, where 50% to 75% of the contacts responded 'yes'), alongside protein structure prediction, EST analysis, molecular evolution and programming languages. Database management systems, biological ontologies and image analysis were generally considered to be more advanced subjects.

Table 2: The number of responses to each question posed in the survey

Question	Response (%)	
	Yes	No
1A Sequence, structure and specialized databases	96	4
1B Database management systems	61	39
1C Database development strategies	43	52
2A Molecular evolution	70	30
2B Sequence similarity tools	100	0
2C Multiple sequence alignment tools	100	0
3A Secondary, super-secondary and tertiary structure	91	8
3B Visualisation tools	87	8
3C Protein structure prediction	70	22
4A Genomic structure	91	8
4B Sequencing techniques	74	26
4C Gene prediction methods	91	8
5A EST clustering and assembly techniques	74	22
5B High-throughput gene expression techniques	74	26
5C Microarray design and analysis	70	30
6A Experimental techniques in protein identification and characterization	61	35
6B Proteomic prediction techniques	74	17
6C Image analysis	52	43
7A Knowledge of procedural and object-oriented programming languages	70	26
7B Basic statistics and information theory	91	4
7C Alignment algorithms	78	17
7D Biological ontologies	57	39
8 Are there any obvious deficiencies in graduates from bioinformatics courses	43	17

Only a minority of contacts (43%) felt that current bioinformatics graduates are lacking obvious skills that are of interest to companies and institutes. In the case where additional comments were supplied, the need to attract more computer scientists to the field was expressed (9%), as well as the need for different skill levels (13%), *i.e.*, those with general knowledge and those with specialised knowledge. Two contacts felt that the problem was not so much the lack of any particular skills *per se*, but rather the current shortage of trained individuals.

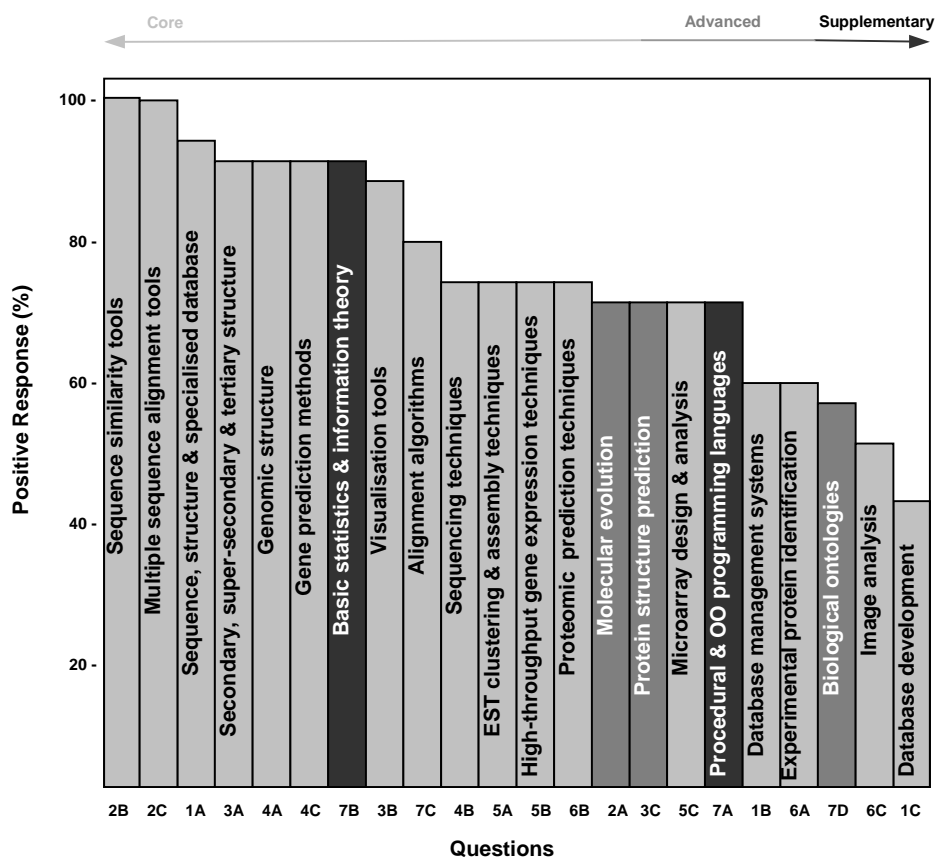


Figure 2: The coverage of positive responses. For all questions posed in this survey (with the exception of question 8 regarding graduate deficiencies), this bar chart shows the spread of positive responses. The lower x-axis correlates with the questions listed in Table 2: these are colour-coded according to their suggested classification in the syllabus in Figure 1: core (light grey), advanced (grey), supplementary (dark grey). The upper x-axis indicates topic priority in relation to the original syllabus (Figure 1); the bar chart illustrates a slightly modified priority with respect to contact responses.

Discussion

By mapping the results of the survey (Table 2) onto the basic syllabus proposed in Figure 1, the bar chart shown in Figure 2 suggests that areas we intended to cover within EMBER were generally appropriate: it appears that most employers expect graduates to possess general knowledge of all bioinformatics techniques. However, as shown in Figure 2, it is clear that some elements we considered to be supplementary (*i.e.*, basic statistics and information theory) were given higher priority by employers. It is also evident that some core elements (*e.g.*, database management/development, microarray design/analysis, experimental protein identification, *etc.*) were given less priority by comparison with molecular evolution and protein structure prediction, which we had considered to be more advanced topics.

With regard to advanced topics, opinions differed slightly depending on the interest of the contact. For example, one bioinformatics contact commented that an appreciation of experimental techniques, such as microarrays, is required in order to communicate effectively to biologists and to help solve problems. Another pharmaceutical contact stressed the growing need for graduates to acquire more knowledge of computer science and statistics in order to analyse

experimental data effectively. Others highlighted the need to understand the success/reliability of various techniques and to appreciate the significance of results generated. For example, one observation was that few graduates fully appreciate the difference between BLAST and PSI-BLAST, or the meaning of E-values reported in BLAST results. Such awareness is considered to be essential; this is highlighted in Figure 2, where questions concerning alignment algorithms and statistics are clustered with those considered to be core material. Another contact (pharmaceutical) suggested that many courses are preoccupied with sequence analysis techniques (despite all contacts confirming sequence analysis as a core element of bioinformatics), leading to a lack of knowledge in other areas: *e.g.*, enzymology, signal transduction, gene regulation, histology databases, and so on.

Feedback from the survey has highlighted the tension in current MSc courses between trying to introduce a broad spectrum of subjects and trying to add adequate depth to any of them. It also suggests that a modular structure would be desirable, allowing students from different backgrounds to take only those modules that are relevant to their needs. This leads to the concept of an "ideal" syllabus. Such a syllabus would aim to provide a course suitable for those with either a biological or a computational background. It would, for example, consist of four main categories (Figure 3): molecular biology, core bioinformatics, advanced bioinformatics and informatics.

1. Molecular Biology
 - 1.1. Central dogma of molecular biology
 - 1.2. Genomic structure
 - 1.3. Protein structure
2. Core Bioinformatics
 - 2.1. Biological databases
 - 2.2. Principles of sequence analysis
 - 2.3. Functional genomics I - The genome
3. Advanced Bioinformatics
 - 3.1. Molecular evolution and phylogeny
 - 3.2. Protein structure prediction
 - 3.3. Functional genomics II - The transcriptome
 - 3.4. Functional genomics III - The proteome
4. Informatics
 - 4.1. Information theory
 - 4.2. Basic statistics
 - 4.3. Database technologies
 - 4.4. Knowledge representation
 - 4.5. Biocomputing

Figure 3: The "ideal" syllabus. This syllabus comprises of four modular components: (1) molecular biology, which would provide computer scientists with adequate biological knowledge; (2) core bioinformatics, which would provide postgraduates from all backgrounds with a basic knowledge of bioinformatics techniques; (3) advanced bioinformatics, which would introduce more complex topics, as well as increasing awareness of new technologies; and (4) informatics, which would strengthen the computer skills of biologists.

By splitting the course into four modules, postgraduates could embark on studies more suited to their needs: *i.e.*, a biologist could focus on informatics, core and advanced bioinformatics, while a computer scientist, following a similar strategy, could study molecular biology instead of informatics. However, not all the above modules are within the scope of EMBER: this project was devised to address the shortage of trained individuals by providing a standard package of basic teaching materials in bioinformatics, within a fixed timeframe. Specifically, by providing a stand-alone tutorial in bioinformatics with a supplementary text book, EMBER will offer an educational package that can be adopted as a core

component of undergraduate/Masters courses, as a supplement to PhD research programmes, or as a tool for personalised learning or for in-house corporate training.

Conclusions

Overall, only a small number of contacts responded in this survey; nevertheless, the results provided valuable insight into the expectations of potential employers and the observed deficiencies of current bioinformatics graduates. The results of the survey indicate that topics we originally planned to cover in EMBER are generally in agreement with the skills sought by employers of graduates in bioinformatics; this has allowed us to refine a general syllabus to underpin EMBER course material.

Acknowledgements

The EMBER consortium wishes to thank the European Commission for supporting this project. JEM is funded by the European Commission and TKA is a Royal Society University Research Fellow.