Briefings in Bioinformatics Advance Access published September 5, 2013

doi:10.1093/bib/bbt065

# Teaching the ABCs of bioinformatics: a brief introduction to the Applied Bioinformatics Course

*Jingchu Luo*

Submitted: 20th May 2013; Received (in revised form): 15th August 2013

## Abstract

With the development of the Internet and the growth of online resources, bioinformatics training for wet-lab biologists became necessary as a part of their education. This article describes a one-semester course 'Applied Bioinformatics Course' (ABC, http://abc.cbi.pku.edu.cn/) that the author has been teaching to biological graduate students at the Peking University and the Chinese Academy of Agricultural Sciences for the past 13 years. ABC is a hands-on practical course to teach students to use online bioinformatics resources to solve biological problems related to their ongoing research projects in molecular biology. With a brief introduction to the background of the course, detailed information about the teaching strategies of the course are outlined in the 'How to teach' section. The contents of the course are briefly described in the 'What to teach' section with some real examples. The author wishes to share his teaching experiences and the online teaching materials with colleagues working in bioinformatics education both in local and international universities.

***Keywords:*** *bioinformatics education; introductory course; hands-on course; project-based learning; on-site teaching*

## INTRODUCTION

The international Human Genome Project started in the early 1990s, and the rapid development of the World Wide Web at the same time marked the beginning of a new era of biological research, foreseen by Walter Gilbert in his perspective article published on Nature 'News and Views' in January 1991 [1]. Under the title 'Towards a paradigm shift in biology', he pointed out that 'We must hook our individual computers to the worldwide network that gives us access to daily changes in the database and also makes immediate our communication with each other'.

With a massive explosion of nucleotide and protein sequence data, other types of information stored in primary and secondary biological databases, experimental wet-lab biologists started to use computers to obtain and analyze biological data in their daily research with online or desktop bioinformatics tools. To help these biologists access the databases effectively and use the analysis tools efficiently, bioinformatics training became a vital part of a biological education.

As we all know, nationally and internationally renowned bioinformatics centers such as the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) have been playing a leading role in the provision, not only of bioinformatics resources and services but also online training materials [2, 3]. The NCBI Education web portal (http://www.ncbi.nlm.nih.gov/education/) is an entry point for either novice or advanced users to find various information including how-to guides, handbooks, frequently asked

Corresponding author. Jingchu Luo, State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences and Center for Bioinformatics, Peking University, Beijing 100871, P.R. China. Tel.: +86 10 62756590; Fax: +86 10 6275 9001; E-mail: luojc@pku.edu.cn

**Jingchu Luo** is a Professor in the College of Life Sciences at Peking University and the China Node manager of the European Molecular Biology Network. His main areas of research are development of bioinformatics platforms; construction of bioinformatics databases; and computational comparative genomics. He has been teaching the Applied Bioinformatics Course for graduate students of biology for 13 years.

questions, help manuals and tutorials. The online training courses developed and maintained at EBI (http://www.ebi.ac.uk/training/online/) are especially helpful for users looking for eLearning opportunities. On the other hand, traditional, on-site, face-to-face teaching incorporating online teaching materials also plays a significant role in both lecture-based and hands-on bioinformatics education and training, either as semester courses for undergraduate students [4–7] or as short-term training courses [8]. Community efforts, especially in Europe, such as the Bioinformatics Training Network (BTN, http://www.biotnet.org/) have provided a forum for bioinformatics trainers to share their experiences [9]. For a broader view and detailed description of the current situation in bioinformatics education and training, readers are encouraged to refer to the two recent reviews [10, 11] and the dedicated articles published in this special issue of Briefings in Bioinformatics.

In this article, the Applied Bioinformatics Course (ABC), a face-to-face on-site semester-long course for biological graduate students is introduced. ABC is a project-based and problem-oriented hands-on course for the graduate students enrolling in molecular biology research projects. Benefited from various resources in bioinformatics education mentioned above and tailored for special requirement from biological graduates, the goal of this course is to help students to solve practical biological problems related to their ongoing research projects using bioinformatics resources and analysis tools. The outline of the course, the main databases, analysis tools, problem examples and the demonstration projects are narrated, and the teaching strategy and organization of the course are described. Thus, the aim of this article is to give the reader an overview of the course and offer an understanding of the online materials dedicated to this course. The author wishes to share experiences accumulated over the past 13 years and to stimulate discussion with both local and international colleagues working on bioinformatics education.

## HOW TO TEACH

The name of the course ABC implies that this is a practical course to teach the ABCs of bioinformatics. A Web site dedicated to this course (http://abc.cbi.pku.edu.cn/) was created in 2004 and is maintained and updated regularly. Most of the materials mentioned in this article can be found in greater detail on the ABC Web site.

### Prerequisites

ABC is aimed to teach biological graduate students how to access online bioinformatics resources and how to use bioinformatics tools to solve practical problems related to their research projects.

The prerequisites for the students to take the course are as follows:

- A good background in biological sciences.
- An ability to read in English.
- Availability for at least 3 h every week outside the class for group discussion.

A set of self-test exercises is provided for the students to check if they are qualified to take the course. These include basic taxonomy and molecular evolution, the genetic code, the concept of the central dogma, the eukaryotic gene structure, the principle of protein structure and conformation, as well as sequence–structure–function relationship of biomolecules. A survey form is designed for each student to fill in before taking the course. In addition to the general information such as their educational background and ongoing research project, the students are encouraged to give their suggestions and to state any special requirements based on their work or interest. Finally, a good command of the English language is also required to take the course, as most of the materials and reference books we use in the course are in English.

Owing to the increasing demand from the students, the nature of the on-site hands-on course as well as the limited availability of computer facilities in the training room, the students are divided into several classes in each semester. For example, in the 2013 spring semester ending in June this year, 168 students were divided into three classes.

### Format

ABC is essentially a hands-on practical course. It usually takes 15 weeks, with 3 h each week used for teaching and a further 3 h/week for group discussion. Except for the introductory session at the beginning of the semester and the final group report with presentations at the end, most sessions of the course are run in a computer training room. Each student has a desktop computer connected to the Internet. Special teaching software with a broadcasting function is installed on each computer so that the contents on the teacher's screen are simultaneously displayed on the students' screen.

This makes it much easier to present step-by-step demonstrations.

The students are divided into working groups, usually consisting of four members each, with one volunteer as the group head. The members of each individual group often come from the same laboratories and have similar research projects. They work together during the class. Organized by the group head, each group also works together outside the class at least 3 h each week to do homework assignments or to discuss and solve practical problems related to their own research projects. Reports of group discussion with achievements and questions are submitted and some of the questions are discussed in the class.

Dedicated email accounts for each class are created to communicate with students. For example, the email account pku13s@gmail.com is used to exchange messages with Peking University (PKU) students taking the ABC course of the 2013 spring semester. Students send their homework assignments, reports of group discussion as well as any comments and suggestions to this email address.

## Exercises

A set of simple exercises has been created for the students to do either in or outside the class. They have been designed to familiarize the student with various bioinformatics tools. The topics of the exercises include PubMed searches, UniProt database queries, sequence alignments, dot plots, BLAST searches, DNA and protein sequence analysis, motif identification, gene structure prediction, tree construction and protein structure visualization and analysis.

## Projects

As a project-based and problem-orientated course, the emphasis has been put on two main areas. First, a set of predefined projects are used in the course. These projects are mainly selected from the real examples that the author has worked on, either in his own research projects or collaboratively with wet-lab colleagues in the college. Students from individual groups are required to bring their own projects closely related to the research they are already enrolled in or plan to start.

Second, biological problems are focused on from the beginning to the end of the course. Not only the comprehensive problems are embedded in the predefined projects but also general biological questions are asked throughout the teaching of the course. For example, in the UniProt database query, students are asked to find the list of species that has an alpha hemoglobin (HBA) sequence identical to human HBA. During Ensembl database browsing, students are asked to find the difference in chromosome numbers between human and chimpanzee, and the syntenic regions between human chromosome 2 and chimpanzee chromosome 2A and 2B.

## Homework assignments

One of the most successful aspects of the course is the design of a set of homework assignments for the students to work on outside of the class. Different from the simple exercises described above, the questions in homework assignments are much more comprehensive. The topics of the assignments include literature searches, database queries, sequence alignments, BLAST searches, as well as the comprehensive predefined projects introduced throughout the course. Currently, 10 homework assignments are designed and being updated or revised over the time. Students usually discuss the questions in the homework assignments within each group outside the class, exchange ideas and experiences first and then work on them independently. Undoubtedly, the homework assignments strengthen the students' understanding about the tools that can be used for their own projects.

For PKU students, teaching assistants are assigned. They help with checking the homework assignments and join in discussions with certain groups every week.

## Presentations

All the students are encouraged to make informal presentations and answer questions raised by other students throughout the classes. For some of the more theoretical topics, such as the algorithm of the BLAST similarity search, the principles of molecular evolution and phylogenetics, the methods of homology modeling, special lectures are given either by the instructor or by invited senior students working on research projects of bioinformatics and molecular evolution.

At the end of the course, two workshops are organized, a small one and a large one. The small one is a class-based group report. Representatives of each group chosen by group members make presentation to summarize what they have learned and applied to their own projects. One or two good presentations are selected to make further refinement and to

present in the larger workshop with students from all classes taking the course in the same semester. The presentation slides are made available online after the workshop.

The large workshop is also open to other students who took the course previously or will take the course in the next semester. In addition to the refined representative talks from each class, invited lectures are also given by well-established scientists from our university or external institutions. The topics of these lectures are more advanced and specific: such as the methods and applications of next-generation sequencing (NGS), the principle and application of molecular evolution, the applications of comparative genomics and structure-based drug design.

### Exams and credits

As a hand-on practical course, there are no conventional examinations at the end of it. Instead, the students are required to make a summary of the course and work on this during the summer or winter vacation. The topics of the summary are flexible. Students are encouraged to apply the approaches implemented in the predefined projects and the techniques they learn during the course to their own ongoing research projects.

Course credits are valued based on several factors: (i) the performance during the classes; (ii) the homework assignments; (iii) the report of group discussions; (iv) the final group presentations; (v) the final summary.

### Reference books

Several reference books including the two introductory English books 'Essential Bioinformatics' [12] and 'Bioinformatics for Dummies' [13] are recommended to all students taking the course. Students are also encouraged to read a more advanced and comprehensive one 'Bioinformatics: Sequence and Genome Analysis' written by David Mount and published in China as reprints [14].

In recent years, various bioinformatics books with diverse contents and different levels either in English or in Chinese became available. And students are advised to find reference books that are most suitable for them based on their background.

A Chinese textbook 'Bioinformatics Techniques and Applications' is being written by the author and will be published in the near future. Some of the materials retrieved from this book are tailored as Power Point slides and presented in the course.

## WHAT TO TEACH

With the teaching strategies and focuses described above, the following section tries to give a general picture of the course contents.

### Analysis tools

Different from short-term training courses, which might focus on certain packages owing to a limited time, this course does not take the tool-centric approach. Instead, different tools are introduced in several stages when they are needed to solve certain problems. Also, students are advised to find suitable tools by themselves and read the documentation and help materials carefully to use the tools correctly and efficiently. The ABC Tools page (http://abc.cbi.pku.edu.cn/tools.php) collects groups of popular tools used in the course. The most commonly used tools are described as follows.

Both web-based platforms and desktop packages are used in this course (Table 1). WebLab is the main online analysis platform used in the course. It is an open-source package developed and maintained by a team from our center over the past 10 years and contains >200 individual programs mainly for DNA and protein sequence analyses [15]. The built-in user data space can be used to store the sequence data and analysis results, and the data-sharing mechanism is well suited for students to share the data among members of a group. For desktop programs, we use

**Table 1:** The main analysis tools used in the course

| Name | Type | URL |
|---|---|---|
| WebLab | Online sequence analysis platform | http://weblab.cbi.pku.edu.cn/ |
| Jemboss | Desktop sequence analysis interface | http://emboss.sourceforge.net/Jemboss/ |
| Blast | Database similarity search | http://blast.ncbi.nlm.nih.gov/ |
| MEGA | Phylogenetic analysis | http://www.megasoftware.net/ |
| SPDBV | Structure analysis | http://www.expasy.org/spdbv/ |

Jemboss [16], which offers a Java interface to the well-known EMBOSS package [17].

For database similarity searching, we use the NCBI BLAST web server as the main online search tool, with the emphasis on selecting individual programs, choosing different databases, setting specific parameters, changing output formats and analyzing search results. We also install the command line BLAST package locally under both Windows and Linux systems to make demonstrations.

The MEGA package [18] is chosen for initial phylogeny analysis, as it has a graphical user interface offering different methods in tree construction. Other phylogenetic packages such as Phylip (http://evolution.genetics.washington.edu/phylip.html) and PAML (http://abacus.gene.ucl.ac.uk/software/paml.html) are also introduced, though not in detail.

For the protein three-dimensional structure analysis and visualization, Swiss-PDBViewer [19] is selected as the main tool because it has a rich functionality with well-written help documentation and online tutorials. Other visualization tools such as the off-line package PyMol and online web browser plug-in JMol are also introduced.

## Bioinformatics resources

We start with the introduction to bioinformatics resources maintained by the two most popular bioinformatics centers in the world, NCBI and EBI. It is impossible to browse all the web pages embedded in these two web portals. Instead, we encourage students to use the NCBI online educational materials and the EBI online training courses to find useful resources by themselves. Advanced search of PubMed, together with the useful tool MyNCBI, is introduced at the beginning of the semester. Other literature resources are also introduced in the first session, such as the scientific stories in 'Molecule of the Month' and 'Protein Spotlight' publications. Lists of online databases published in the Nucleic Acid Research (NAR) special issues and web-based online analysis tools collected by Swiss Institute of Bioinformatics (SIB) are introduced in the middle of the semester with the emphasis on various bioinformatics resources that are available on the Internet (Table 2).

## Database query

After a brief introduction to the bioinformatics resources, we move on to introduce the most popular sequence databases including the UniProt protein sequence database, the NCBI RefSeq database, the Ensembl genome database, the protein data bank (PDB) protein structure database, as well as other databases such as the gene structure database Gene, the gene expression database Expression Atlas, the protein–protein interaction database STRING, the metabolic pathway database KEGG, the protein domain database PFam, the proteomic database Pride, the phylogenetic database OrthoDB, the disease-related database OMIM.

Rather than going through all these databases one by one using Power Point slides, we use a different approach to introduce these databases. First, we take the human hemoglobin alpha subunit (HBA_HUMAN) as an example to show the rich annotations of a UniProt/Swiss-Prot database entry, and further explore the specific annotations provided by these databases by clicking the links in the Cross-Reference section of the entry. After browsing the entries of these databases, a brief introduction to the commonly used databases and the differences among these databases is given, so that the students can have a general overview of various databases in bioinformatics.

The advanced search functionality is another focus in this session and uses the powerful functions in the database query systems of international bioinformatics centers such as NCBI, EBI and the SIB. For

**Table 2:** Online bioinformatics resources

| Resource | URL | Developer |
|---|---|---|
| Educational | http://www.ncbi.nlm.nih.gov/education/ | NCBI |
| Training | http://www.ebi.ac.uk/training/online/ | EBI |
| Molecular of the month | http://www.rcsb.org/pdb/101/motm.archive.do | David Goodsell |
| Protein Spotlight | http://web.expasy.org/spotlight/ | Vivienne Gerritsen |
| List of databases | http://www.oxfordjournals.org/nar/database/c/ | NAR |
| List of tools | http://www.expasy.org/ | SIB |

**Table 3:** Example of UniProt advanced search

| Query field | Query text and topic, term | Hits |
|---|---|---|
| Organism | *Arabidopsis thaliana* [3702] | 53 755 |
| Reviewed | Yes | 12 019 |
| Protein existence | Evidence at protein level | 4086 |
| General annotation | Subcellular localization: membrane; confidence: experimental | 1243 |
| Sequence annotation | Signal peptide; confidence: any | 234 |

example, a simple search in UniProt using text 'Human hemoglobin' returns not only different subunits of human hemoglobin but also a 100 entries from other species. Using the 'Advanced Search' function developed by the UniProt team, we may easily eliminate these false positives. By selecting 'hemoglobin' as protein name, 'human [9606]' as organism and 'HB*' as gene name, we can obtain all nine human hemoglobin subunits in UniProt/Swiss-Prot without false positives or false negatives. Several more comprehensive examples are also given as demonstrations in the class. With the rich annotation embedded in the Swiss-Prot entries, we can find all experimentally verified Arabidopsis membrane proteins with potential signal peptides using the step-by-step 'Advanced Search' function (Table 3).

After doing several exercises likes this, the students start to search their own proteins, browse the annotations in these entries and further explore information in other databases by clicking the cross-links.

## Sequence alignment

Sequence alignment is one of the most commonly used approaches to study the relationship among different molecules and organisms. It seems easy to run the global alignment program Needle implemented in the WebLab and Jemboss packages. However, it is not trivial to obtain and analyze the output results without solid biological knowledge.

We start with pairwise alignment of the amino acid sequence of the alpha subunit of human, mouse and rat hemoglobin. Surprisingly, the output of the alignment is different to that which would be expected—the sequence identity between mouse and rat is less than that of mouse and human. This triggers the interests of the students with enthusiastic discussion. Finally, they find the answer to this question by retrieving and comparing the nucleotide coding sequence (CDS) of the hemoglobin alpha gene of these three mammals. By using the correct scoring matrix and setting the appropriate gap penalties, they can obtain an output where results are more convincing. Indeed, the sequence identity of CDSs between mouse and rat is higher than that of mouse and human, which indicates that mouse and rat are closer relatives.

## Database search

Running BLAST seems an easy job for most of the students. However, running a 'good' BLAST is another story. Literature searches tell us that neuroglobin is a member of the human globin family of which nine hemoglobin subunits (alpha, beta, gamma, delta, etc) as well as myoglobin and cytoglobin belong to.

Taking this as an example for BLAST search, we then ask, 'can we find a human neuroglobin through BLAST search using the amino acid sequence of the alpha subunit [Swiss-Prot: HBA_HUMAN] as a query?' The answer is 'No' if we use the default parameters to run BLASTP through the NCBI BLAST server. Nevertheless, we can obtain a good match by using PSI-BLAST and choosing Swiss-Prot as the preferred database, selecting Human as the organism and setting E-value to 0.001. The first run returns 11 hits including nine hemoglobin subunits as well as myoglobin and cytoglobin, but not neuroglobin. However, with the position-specific scoring matrix that is automatically constructed based on these 11 sequences obtained in the first run, the output of the second PSI-Blast search does return the neuroglobin (NGB_HUMAN) that has only 21.9% identical residues with HBA_HUMAN.

Other BLAST programs are also introduced with real examples to show the students how to choose the different programs and how to set better parameters to obtain biologically meaningful results. At the end, we give a brief introduction to the algorithm behind BLAST so that the students have a better understanding of the parameters such as the scoring matrix, the E-value and the word size. By doing so, the students become aware of the importance of knowing the general principles and biological background behind this and other sequence analysis programs.

## Analysis of a plant mRNA sequence

After several sessions of general introduction to the online resources, database query and analysis

platform, we start to work on more comprehensive predefined projects to solve real biological problems. In 1997, a *Pisum sativum* post-floral-specific gene (PPF-1) was identified from a cDNA library of short-day grown G2 pea tissue by a colleague in our college [20]. BLAST search hits with the query sequence of PPF-1 protein contains several bacterial inner membrane proteins, which demonstrate hydrophobic regions in sequences. This suggests that the PPF-1 protein may have transmembrane (TM) helices and could be located in the chloroplast membrane of pea leaves.

Taking the 1523 bp mRNA sequence as an example, we show the students how to use different tools to display and retrieve the CDS using several tools such as PlotORF, ShowORF and GetORF, and make analysis such as GC content, codon usage and restriction enzyme cleavage. By translating the CDS to an amino acid sequence, we can make further analysis at the protein level, such as the statistics of amino acids, the identification of signal peptide, the prediction of potential TM regions, helical wheel display of these potential TM helices and generally gain an overview of the protein.

## Analysis of a fugu genomic sequence

In 1990s, Sydney Brenner and his colleagues initiated the fugu genome project (http://www.fugu-sg.org/). The genome size of this model organism is only 390 Mb, yet it contains most of the human homolog genes and can be used as a model system to study the function of human genes. For example, the human multidrug resistance gene (MDR) family has several members that belong to the ABC transporter superfamily. To investigate the mechanism of drug resistance, a PhD student at the Chinese Academy of Medical Science obtained a 39 kb genomic sequence from a Fugu cosmid using the human MDR gene as a probe. Analysis of the sequence data confirmed that this genomic sequence contained two MDR genes.

The approach and tools used in this project are rather different from those used in the analysis of the pea mRNA sequence as discussed in the previous section. First, a repeat region can be identified using a dot plot program to compare the genomic sequence with itself. Several gene identification programs can then be used to predict the exon/intron structure of the different fragments of the genomic sequence. Finally, confirmation that this genomic sequence does contain two multidrug resistance genes is obtained by running a BLASTX search against the UniProt/Swiss-Prot database.

## Analysis of the bar-headed goose hemoglobin

More comprehensive projects are introduced during the course. One of them is the analysis of the sequence, structure, function and evolution of the bar-headed goose hemoglobin. Hemoglobin is one of the most well-studied proteins of the past century. Hundreds of hemoglobin protein sequences have been deposited into the Swiss-Prot database. Three-dimensional structures of wild type and mutants from dozens of species have been solved. This provides us with a good opportunity to study the structure–function relationship of hemoglobin molecules.

The bar-headed goose is a special species of migratory bird. They live in the Qinghai lake during summer time, fly to India over the Tibetan plateau in autumn and return to the lake in spring. Interestingly, the graylag goose, which is a close relative of the bar-headed goose, lives in the lowland all year round. Sequence alignment of bar-headed goose hemoglobin with that of graylag goose shows only four substitutions. The most critical one is the Pro119 to Ala119 substitution located at the surface of the alpha/beta interface. In 1983, Max Perutz proposed that this substitution reduces the contact between the alpha and beta subunit and increases the oxygen affinity, owing to the relation of the tension status in the deoxy form [21].

During the past decade, a research group at our university has solved the crystal structure of both the deoxy (PDB ID: 1HV4) and the oxy (PDB ID: 1A4F) form of the bar-headed goose hemoglobin, as well as the oxy form (PDB ID: 1FAW) of the graylag goose hemoglobin. Using the powerful free Swiss-PDBViewer, the students create a Magic Fit to superimpose the alpha/beta heterodimer of the oxy form of the two goose hemoglobins (1A4F and 1FAW) on each other. They are excited to see the difference of the side chains between Pro119 in graylag goose hemoglobin and Ala119 in bar-headed goose hemoglobin, and make measurements of the distance between the atoms of these two side chains and the atoms in the side chain of the corresponding residue in the beta subunit.

Several other projects are also discussed throughout the course, such as analyzing and modeling of the human carcinoembryonic antigen, predicting the

three-dimensional structure of an antifungal peptide from pokeweed, engineering the metallothionein, proposing the evolutionary mechanism of a plant-specific transcription factor family.

In addition to the predefined projects, students are also encouraged to bring their own problems to solve either in or outside the class. For example, the following list is taken from some group presentations of the 2013 spring semester course:

- The sequence analysis of EIN3 in *Arabidopsis thaliana*.
- Analysis of the cold-response receptor kinase OsRLK1 in rice.
- Analysis of the PDF1 gene in *Brassica napus*.
- Identification of odorant binding protein7 in *Apolygus lucorum*.
- Sequence analysis and structure prediction of SIR2 in *leishmania*.
- Structural and functional analysis of structural proteins of foot-and-mouth disease virus.
- Identification of the antigenic sites in the hemagglutinin of avian influenza virus.
- Analysis of structure and function of DNA binding protein HU in *Synechococcus* PCC7002.

Using the techniques on sequences they are personally interested in enhances the learning experience. And fundamentals of the programs used become clearer to the student as they can relate the computational findings to their expectations based on what they already know about their own projects.

## DISCUSSIONS AND FUTURE DIRECTIONS

As indicated by the name of the course and the title of this article, ABC is an entry-level introductory course for biological graduate students and aimed to teach them how to use the software tools available to solve their biological problems. The methods and algorithms of these tools, as well as the statistical aspects behind them, are taught in another course, Methods in Bioinformatics (MiB), by colleagues in our college. Analyses of high-throughput Omics data generated by NGS are briefly introduced in invited lectures, not as hands-on exercises in either course, owing to limited hardware resources. An Omics data analysis forum was organized in the past year together with colleagues from two other institutions, and a seminar for a small class of about 20 students working on NGS data analysis was run in the past semester.

With a background in biology and self-taught computer scientist, the author would like to emphasize that the strategy used in teaching the ABC course is more biological than computational. This means that a bottom-up but not top-down approach is used. The course evolved from an original one under the name 'Computer Applications to Molecular Biology' (CAMB), which he taught during the 1990s. Whereas CAMB was a lecture-based course, which was taught in a conventional classroom with presentations using transparencies in the early days and Power Point slides later on, ABC is a hands-on course, which is arranged in a computer training room with desktop computers for each student to access the Internet and to make analyses with bioinformatics tools. The transit from CAMB to ABC was inspired by the success of several training courses that the author organized during the late 1990s. A dozen colleagues and training officers from the European Molecular biology Network (EMBnet) and institutions around the world, including Europe and the United States, came to our university to teach several courses such as an EMBnet Bioinformatics Course, a WhatIf Protein Structure Course, a Genome Data Analysis Course and a Database Query Course with the Sequence Retrial System. The experiences gained from these courses show that a hands-on practical course is necessary for wet-lab biologists working in molecular biology.

The 'evolution' keeps going throughout the teaching of the course during the past 13 years, and is still underway, mainly propelled with the comments and suggestions from the students' feedback delivered at the end of each semester. For example, the idea of dividing the students into working group was suggested by several students based on their experience of benchwork laboratory experiments >10 years ago and has been proved to be a good strategy. The small workshop for group report at the end of semester was started at the same time, but the large workshop with talks by representative groups to stimulate students' further learning, and with invited lectures by senior scientists to expand bioinformatics knowledge, was initiated in the recent years. On the other hand, some impractical settings were eliminated. For example, Linux system and command-line user environment were introduced in the early years, which took a great amount of time for most of the students to get familiar with

the commands and parameters. However, a survey shows that only a few students use Linux system after the course. Therefore, we changed the command-line–based environment to web-based one for the course, and organize seminars for students working on Omics data analysis.

As application of bioinformatics techniques become more and more important in biological research, as well as the success of the ABC course, the number of the students enrolling for the course grows every semester, from an average of 100 each year during the early 2000s to >250 in the past year. Currently, ABC is only for biological graduate students. In recent years, many senior undergraduate students in our college have shown great interest in taking the ABC course. Some of them have already been engaging in wet-lab experiments in the laboratory and working in certain projects. Clearly, it is necessary for them to learn the essential bioinformatics tools to help with their ongoing projects. A plan to run ABC as a selective course for some undergraduate students is being proposed. It is challenging to run the course by a single person for both graduate students with increasing number and for undergraduate students in the same semester. A plan to design some modules of the course and make them available as network-accessible open course is underway. At the same time, improvements of the ABC Web site with more detailed materials and step-by-step tutorials may help students with self-learning.

The main pages of the ABC Web site are in English, though some of the WORD and PDF files such as exams and homework assignments are in mixture of English and Chinese. We are currently considering whether it is necessary to create a mirror site and translate all the pages into Chinese. With new technologies in web design, dynamic pages such as discussion groups are also being planned.

## CONCLUSION

ABC is an entry-level introductory course with the aim of teaching the ABCs of bioinformatics. We hope that, by taking the course, students will be in a position to efficiently access the online bioinformatics resources and to use the bioinformatics tools to solve their own practical biological problems. The goal of the ABC course is best summed up using a phrase coined by Dr Alan Bleasby (coauthor of the EMBOSS sequence analysis package): 'half a day on the web, saves you half a month in the lab!' (personal communication)—if you do your software analysis in the correct manner!

### Key Points

- ABC is an on-site, practical, hands-on course for graduate students of biology.
- ABC is a project-based and problem-orientated course to teach students how to access the online bioinformatics resources and how to use analysis tools to solve problems related to research projects.
- Students are divided into groups to work together in and outside the class collaboratively.
- Special exercises and homework assignments are designed for the course.
- Discussion and presentations are encouraged throughout the course, and workshops are organized at the end of the course.

**Acknowledgements**

The author thanks all the students who have taken the ABC course and made suggestions for its further development. The author also thanks the critical comments from the anonymous reviewers and helpful suggestions from colleagues and students in the revision of the manuscript. Special thanks to Dr Lisa Mullan for her help in critical reading and language editing of the manuscript.

## FUNDING

College of Life Sciences and Computing Center of PKU; Graduate School of the Chinese Academy of Agricultural Sciences; State Key Laboratory of Protein and Plant Gene Research.

## *References*

1. Gilbert W. Towards a paradigm shift in biology. *Nature* 1991;**349**:99.
2. Cooper PS, Lipshultz D, Matten WT, *et al*. Education resources of the National Center for Biotechnology Information. *Brief Bioinform* 2010;**11**:563–9.
3. Wright VA, Vaughan BW, Laurent T, *et al*. Bioinformatics training: selecting an appropriate learning content management system—an example from the European Bioinformatics Institute. *Brief Bioinform* 2010;**11**:552–62.
4. Chapman BS, Christmann JL, Thatcher EF. Bioinformatics for undergraduates: steps toward a quantitative bioscience curriculum. *Biochem Mol Biol Educ* 2006;**34**:180–6.
5. Cummings MP, Temple GG. Broader incorporation of bioinformatics in education: opportunities and challenges. *Brief Bioinform* 2010;**11**:537–43.
6. Weisman D. Incorporating a collaborative web-based virtual laboratory in an undergraduate bioinformatics course. *Biochem Mol Biol Educ* 2010;**38**:4–9.
7. Furge LL, Stevens-Truss R, Moore DB, *et al*. Vertical and horizontal integration of bioinformatics education: a

modular, interdisciplinary approach. *Biochem Mol Biol Educ* 2009;**37**:26–36.

8. Fernandes PL. The GTPB training programme in Portugal. *Brief Bioinform* 2010;**11**:626–34.
9. Schneider MV, Walter P, Blatter MC, *et al*. Bioinformatics Training Network (BTN): a community resource for bioinformatics trainers. *Brief Bioinform* 2012;**13**:383–9.
10. Via A, Blicher T, Bongcam-Rudloff E, *et al*. Best practices in bioinformatics training for life scientists. *Brief Bioinform* 2013. (Advance Access publication 25 June 2013).
11. Shapiro C, Ayon C, Moberg-Parker J, *et al*. Strategies for using peer-assisted learning effectively in an undergraduate bioinformatics course. *Biochem Mol Biol Educ* 2013;**41**:24–33.
12. Xiong J. *Essential Bioinformatics*. Cambridge, UK: Cambridge University Press, 2006.
13. Claverie J-M, Notredame C. *Bioinformatics for Dummies*. 2nd edn. Hoboken, USA: Wiley Publishing Inc., 2007.
14. Mount D. *Bioinformatics: Sequence and Genome Analysis*. 2nd edn. New York, USA: Cold Spring Laboratory Press, 2004.
15. Liu X, Wu J, Wang J, *et al*. WebLab: a data-centric, knowledge-sharing bioinformatic platform. *Nucleic Acids Res* 2009;**37**:W33–9.
16. Mullan L. Jemboss reloaded. *Brief Bioinform* 2004;**5**: 193–5.
17. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000;**16**:276–7.
18. Kumar S, Nei M, Dudley J, *et al*. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform* 2008;**9**:299–306.
19. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 1997;**18**:2714–23.
20. Zhu Y, Zhang Y, Luo J, *et al*. PPF-1, a post-floral-specific gene expressed in short-day-grown G2 pea, may be important for its never-senescing phenotype. *Gene* 1998; **208**:1–6.
21. Perutz MF. Species adaptation in a protein molecule. *Mol Biol Evol* 1983;**1**:1–28.