Journal of Industrial Microbiology & Biotechnology (2000) 24, 314–318

www.nature.com/jim

Bioinformatics in the community college

SG Porter[1] and TM Smith[2]

[1]*Biotechnology Training Program, Science and Math Division, BE5104, Seattle Central Community College, 1700 Broadway, Seattle, WA 98122;* [2]*Geospiza, Inc, 619 N 35th St, Suite 101M, Seattle, WA 98108, USA*

Biotechnology is becoming an information-based field. In this article we describe some resources available to instructors, show how these resources are used in the biotechnology training program, and provide examples of activities used by non-science majors to increase their understanding of biology. We discuss some of the challenges we have encountered using these tools in the classroom. *Journal of Industrial Microbiology & Biotechnology* (2000) **24,** 314–318.

Keywords: bioinformatics; biotechnology; computers; DNA sequence analysis

Introduction

Ten years ago molecular biology research was primarily conducted in small laboratories that concentrated their efforts on understanding one or two genes in great detail. The genome sequencing projects have produced new technology and a tidal wave of data that have changed the scale of biology forever. Commercial products are now available that can be used to analyze thousands of genes in a single experiment. Public data repositories are growing exponentially. GenBank, a national sequence database (http://www.ncbi.nlm.nih.gov), had grown to contain more than 4.6 million individual sequences with over 3.4 billion nucleotides as of August 1999. Although firm conclusions still require experiments, this massive influx of data has triggered a corresponding shift in the field of molecular biology, changing a largely experimental science performed on a small scale to a science based on information.

Bioinformatics, an interdisciplinary field that combines computer science, software engineering, and biology, has emerged to meet the challenges of handling large data sets. Although this article will focus primarily on the application of bioinformatics to the analysis of biological sequences, the field encompasses many areas including the development of algorithms and software for data entry, data storage, analysis, statistics, annotation and tools for linking experimental data with biological sequences, biochemical pathways, and published literature. Researchers in academic laboratories and biotechnology companies spend an increasing amount of time engaged in bioinformatics activities. Routine tasks include setting up databases, storing and processing data, performing multiple calculations on large sets and preparing graphs. Additional tasks in high-throughput genome facilities include the use of bioinformatics to track sequencing efficiency, identify malfunctioning equipment, assemble long contiguous sequences from several shorter sequences, identify clones, and analyze and annotate sequence data.

Correspondence: S Porter, Biotechnology Training Program, Science and Math Division, BE5104, Seattle Central Community College, 1700 Broadway, Seattle, WA 98122, USA. E-mail: sporte@sccd.ctc.edu
Received 8 April 1999; accepted 10 November 1999

Although researchers have used DNA sequence analysis tools for the past decade, courses that teach or incorporate bioinformatics have been limited to upper division or graduate students. Improvements in network technology and the rapid proliferation of free resources on the internet have made it possible to start using these tools with a more diverse group of students, including those with a limited knowledge of biology. For example, open access to scientific literature through databases like PubMed (http://www.ncbi.nlm.nih.gov/) provides new opportunities for students at community colleges and high schools whose libraries can't afford to carry a wide variety of scientific journals. Access to scientific literature is essential to understanding results from bioinformatics experiments.

Students at Seattle Central Community College use bioinformatics as a tool for inquiry-based science. The overall goal for these activities has been to have students discover general principles through the study of biological sequences. Students majoring in biotechnology use sequence analyses to perform original research. On-line analysis tools and national databases such as GenBank [2], Entrez [7], and PubMed are used to identify and study DNA sequences that they've cloned themselves in biotechnology laboratory courses. Non-science majors taking courses such as Biotechnology and Society and students in entry level biology courses use the same tools plus other resources developed by the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/).

An additional benefit is provided to students by helping them become more familiar with information technology. The ability to locate reliable information on the internet concerning health and biological issues has become increasingly important for all members of society. All citizens need to know where to find reliable information in order to make informed decisions when they vote on public health policy, when they discuss health care with their physicians, and when they consider having children.

Bioinformatics resources

A multitude of on-line resources is available to instructors who wish to use bioinformatics in the classroom (http://genetics.nature.com/web_specials/gazing). We will

not attempt to provide a comprehensive review in this article but will limit our discussion to the resources that we have used most extensively in the classroom. The most useful resource has been NCBI. NCBI was established by the late Senator Claude Pepper in 1988 as part of the National Library of Medicine. NCBI's mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease.

A central aspect of NCBI's program is managing GenBank, an annotated collection of all publicly available DNA and protein sequences. NCBI also supports and distributes a wide variety of databases for the medical and scientific communities. These include databases of molecular sequences, inheritance of genetic information, scientific literature, and more specialized resources. In our courses, we have used Online Mendelian Inheritance in Man (OMIM), a Gene Map of the Human Genome, Genes and Disease, the Taxonomy Browser, GenBank, Entrez, and PubMed. NCBI also provides programs to query these databases and review the scientific literature. One tool that is heavily used is BLAST. BLAST (Basic Local Alignment Search Tool) is a set of programs designed to search databases of protein or nucleotide sequences with a query sequence [1,5]. Researchers use BLAST to ask if any sequences in the database are similar to their query sequence. This activity is central to predicting gene function from sequence data.

A BLAST search returns a set of sequences from the database that are similar to the query sequence along with accession numbers (or links) that can be used to retrieve an annotated record for each sequence. Annotations often contain cross references to other sources of information such as additional sequence records, specific databases or scientific literature.

BLAST and other search tools are used when a sequence is in hand. Questions about sequences in GenBank, based on their annotations, can be addressed by using Entrez. Entrez has a web-based interface that provides users with access to sequence, mapping, taxonomy, and structural data. It also provides graphical views of sequences and chromosome maps. Sequences can be retrieved by a number of criteria including gene names, author names, and keywords that describe gene function.

We have also made extensive use of the Baylor College of Medicine Search Launcher (http://www.hgsc.bcm.tmc.edu/SearchLauncher). Search Launcher contains a well-organized collection of links to a wide variety of programs for analysis of nucleotide and protein sequences.

Text editors and word processing programs such as MS Word can be used for bioinformatics, also. These programs can be used to determine the length of DNA sequences and to locate specific strings of characters, such as restriction sites, within a larger sequence.

Overview of the biotechnology training program

Our biotechnology students make extensive use of bioinformatics, therefore we will present the most detailed description of classroom bioinformatics activities in relationship to the biotechnology training program. Seattle Central Community College (SCCC) offers a 2-year program in biotechnology. Graduates of the program currently work at full-time positions in more than 36 companies and 13 nonprofit institutions in the Puget Sound area. In the first year of the program students take academic courses that serve as prerequisites for a wide variety of scientific fields. These include chemistry (inorganic, organic, and biochemistry), general biology, microbiology, composition, precalculus, and computer applications. The only course specific to biotechnology in the first year of the program is a two-quarter seminar series that introduces students to the local industry and equips students with job hunting skills. Biotechnology skills are emphasized in the second year. Students take a year-long laboratory course in biotechnology in addition to courses in media and solution preparation, scientific computing, technical writing, genetics, quantitative analytical chemistry, and immunology.

SCCC's biotechnology laboratory course was designed with two goals in mind. The first was to provide our students with marketable skills. The second was to remain flexible in order to respond to a rapidly changing industry. Using these goals as a framework, we designed a year-long, project-based, course that provides extensive hands-on experience with laboratory techniques commonly used in local industry.

Students begin the project in the Fall quarter by cloning restriction fragments from *E. coli* that increase production of β-galactosidase. They characterize the cloned fragments through restriction mapping, PCR, Southern blots, DNA sequencing, and bioinformatics. In the second quarter, students use enzyme assays and SDS-PAGE to identify clones that produce a large quantity of β-galactosidase, then they purify β-galactosidase from selected strains. In the last quarter, students purify antibodies of β-galactosidase and develop an ELISA that they use to quantify β-galactosidase in unknown samples. Over time the laboratory course has been updated to incorporate material on quality systems regulations [current good manufacturing practices (cGMPs) and good laboratory practices (GLPs)] and place an increased emphasis on the development of computer skills.

Bioinformatics for biotechnology majors

Biotechnology students learn how to use bioinformatics during their course in scientific computing. They apply their skills in the biotechnology lab course to study a DNA sequence that they obtain from a fragment of DNA that they've cloned from the *E. coli* genome. During this project students complete the following steps: (1) Determine the length of their DNA sequence; (2) Use their DNA sequence to identify the restriction fragment that was cloned from the *E. coli* genome; (3) Create a computer-generated restriction map from the *E. coli* sequence and compare it to restriction maps obtained experimentally by restriction mapping and Southern hybridization; (4) Use annotations to determine if their cloned fragment contains previously identified promoter sequences and if those promoters might be used for driving expression of *lacZ* in pMC1403; (5) Determine if their cloned fragment contains any open reading frames (ORFs), whether those ORFs correspond to previously identified genes, whether a translational fusion might have been created between a coding sequence and β-galacto-

sidase and if so, to predict the size of the fusion protein; and (6) Search the literature available through PubMed and discuss the possible function of any genes located within their cloned fragment.

Materials and methods

Plasmid libraries are constructed from *E. coli* genomic DNA by digesting it with *Eco*RI and *Bam*HI and ligating it to pMC1403 [4], also digested with *Eco*RI and *Bam*HI (Figure 1). pMC1403 contains a modified *lacZ* gene that is missing a promoter and a translational start site. Clones that contain potential promoter sequences are identified by screening them on L agar containing 100 μg ml^{-1} carbenicillin and spread with 40 μl of X-gal (20 mg ml^{-1}) [4]. Plasmid DNA is isolated from *lac*+ clones. The cloned fragments of *E. coli* DNA are further characterized by restriction mapping and Southern hybridization. Afterwards, the *E. coli* DNA fragments are subcloned into pBC SK+ (Stratagene, La Jolla, CA, USA), amplified by polymerase chain reaction (PCR) and used as templates for DNA sequencing. Sanger dideoxy sequencing is carried out using a biotin-labeled primer complementary to the T3 promoter. Products of the sequencing reactions are separated by electrophoresis and transferred to a nitrocellulose membrane. Newly synthesized DNA fragments are identified using streptavidin and biotinylated alkaline phosphatase as a detection system (personal communication, Maureen Munn, University of Washington High School Human Genome Program, (http://hshgp.genome.washington.edu).

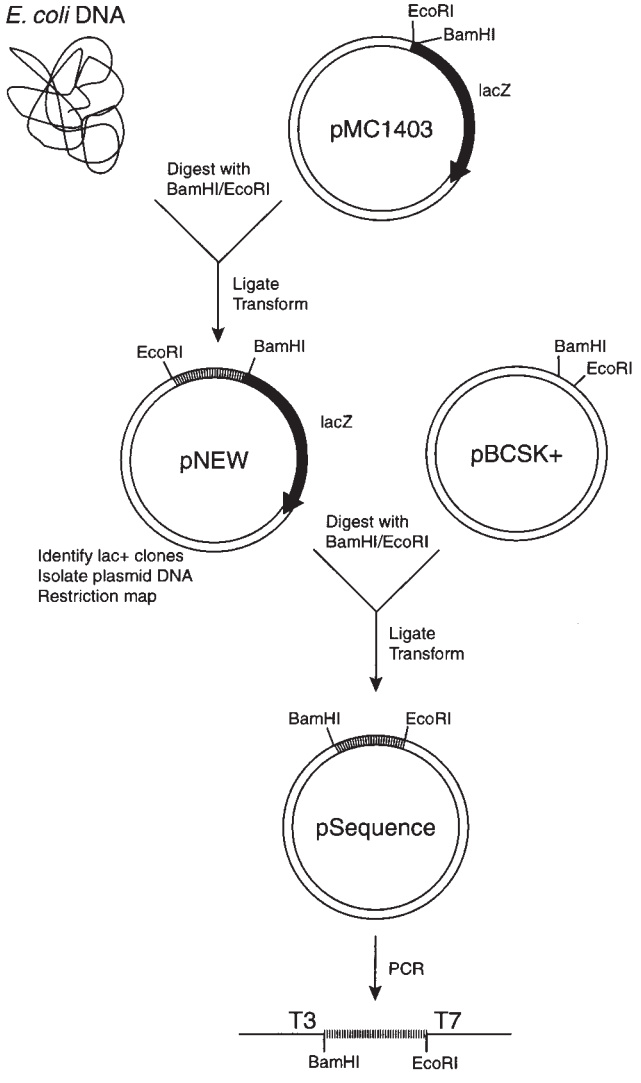

Figure 1 Isolation and subcloning of *E. coli* genomic DNA. *E. coli* DNA was digested with *Eco*RI and *Bam*HI and ligated to pMC1403. Each group of students selected a *lac*+ clone to characterize. Plasmid DNA was isolated and mapped with restriction enzymes, then the fragment of *E. coli* DNA was subcloned as an *Eco*RI/*Bam*HI fragment into pBC SK+. The region between the M13 Universal forward and reverse primer binding sites, including the *Eco*RI/*Bam*HI fragment with the cloned *E. coli* DNA was amplified by PCR and sequenced using a biotinylated T3 primer to direct DNA synthesis.

Students read the DNA sequences from the filters and store them as MS Word files. Sequences are analyzed using a variety of different programs, including MS Word, BLAST [1], Entrez [7], the programs accessed through BCM Search Launcher, and Webcutter 2.0 (http://www.firstmarket.com/utter/ut2.html).

Sequence analysis results

The following results are from a representative project carried out by students in the biotechnology class of 1998. A 159-base sequence was obtained from pSJ12, a plasmid containing a 3.1-kb *Eco*RI/*Bam*HI fragment cloned from the *E. coli* genome. The number of nucleotides in this sequence was determined using 'character count' tool from MS Word.

The 159-base sequence was converted to a FASTA format by two different methods, either by using ReadSeq from BCM Search Launcher or directly in MS Word by typing a '>' and a character return. The FASTA-formatted sequence was used to query the GenBank *E. coli* database using the blastn program from BLAST 2.0 [1,2,5]. The best score (E value = 1×10^{-44}) was obtained for a region of the *E. coli* genome between nucleotides 2301701 and 2314975 containing the *yojI* gene (Accession number AE000310). The E value corresponds to the number of matches that would be expected to be seen by chance if one were to search a database containing a certain number of sequences [1]. The nucleotide alignment between the 159-base query sequence and the section of the *E. coli* genome containing *yojI* showed 91% identity (146/159 bases matched). A blastn search also performed against the nonredundant database (all of the GenBank sequences) identified three separate entries for the same region of the *E. coli* genome (E value = 1×10^{-42}). No other scores with a significant E value were obtained.

Next, the accession number (AE000310) was used to retrieve the complete nucleotide sequence of the 13,275 contiguous sequence containing *yojI*. The *Eco*RI and *Bam*HI sites in this sequence were located by two methods. In the first method, the 'Find' tool in MS Word was used to search for the sequences 'GAATTC' and 'GGATCC'. The 'character count' tool was used to determine the distance between these restriction sites by counting the number of letters. Word is used as an analysis tool, in part, to reinforce the notion that sequence information is contained in strings of text. The second method relied on Webcutter

to locate restriction sites for enzymes that recognize a six-base sequence. The *Bam*HI site most likely to border the cloned fragment was found near the 159-base sequence as would be expected since that region is close to the binding site for the T3 primer used for sequencing (Figure 2). The *Eco*RI site was located 3140 nucleotides away from the *Bam*HI site, consistent with experimental data from restriction mapping. The location and presence (or absence) of the other mapped sites (*Pst*I and *Sal*I) were also consistent with the student-generated restriction map (data not shown).

Entrez was used to determine the orientation of *yojI* relative to the *lacZ* coding sequence in pMC1043. The cloned *Eco*RI/*Bam*HI fragment was predicted to contain either a promoter or a translational start site because it increased the production of β-galactosidase protein when inserted upstream of *lacZ* in pMC1403 (BIO 285–287 student data, 1997–98). An examination of the coding sequence for *yojI* and its orientation relative to the *Bam*HI site showed that the position of the insert allowed the correct translational reading frame to be maintained, creating a fusion between the *yojI* protein and β-galactosidase. Thus, transcription of *yojI-lacZ* fusion protein would be driven by the *yojI* promoter and use the ribosome binding site and translational start codon from *yojI*.

The students were unable to locate experimental data for the function of *yojI* from PubMed. However they did find that a putative function for *yojI* as an ATP-binding component of an ABC transport system had been assigned based on sequence homology [3]. ABC transport systems are defined as a group of proteins that contain three conserved functional domains, a nucleotide-binding domain, a membrane-spanning domain, and a solute-binding protein [6]. Entrez was used to find genes related to *yojI*. Several genes were listed, all coded for proteins involved in transporting small molecules across membranes, consistent with the putative function for *yojI*.

Bioinformatics activities for biology students and non-science majors

Learning activities based on bioinformatics can be useful for all students with an interest in biology. We will describe two types of activities in this section: translation of DNA sequences and the identification of 'unknown' DNA sequences.

The learning objectives for the translation activity are to have students discover that DNA contains coded information and begin to construct a model of gene anatomy that goes beyond the protein coding sequence to include the idea of regulatory sequences for transcription and translation. To accomplish these goals, two different DNA sequences were placed at a class web site. Students were told to copy the sequences, paste them into a form at the BCM Search Launcher site, and choose the 'translate' function. Students were also told to record their observations and write down any questions they had about the results. The first sequence contained a hidden message that the students would recognize. The second sequence consisted of a protein coding sequence flanked by additional nucleotides at the 5′ and 3′ ends. Even though many students thought they understood the relationship between RNA and DNA, they were astonished to find that each sequence could be translated in six different ways. Some of the questions they asked were, why are there stars in the sequence? The translation program uses stars (*) to represent stop codons. Which of the six sequences is correct? How does the cell know which sequence to use? These questions typically lead to a discussion of sequences that punctuate the genome, controlling where the processes of transcription and translation start and stop.

A second activity we've used is to have students use BLAST to obtain information about an 'unknown' DNA sequence. The objectives for this activity are: (1) to understand how the information stored in a DNA sequence can be used to uncover the function of a gene; (2) to discover that different organisms contain similar genes; and (3) to learn how to use databases to obtain information about genes. Entrez was used by the instructor to retrieve cDNA sequences for a variety of genes, from diverse organisms, whose functions are known. These sequences were placed on a web page and assigned to small groups of students. Students were asked to determine the organism that their sequence probably came from, describe the possible function for the protein encoded by their gene, determine if any other organisms have closely related genes, and use PubMed to obtain additional information about the function of their gene.

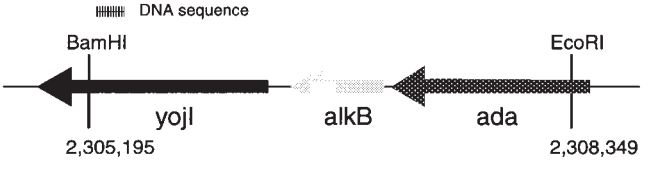

Figure 2 Alignment of the 159-base sequence (dark bar) with the corresponding region of the *E. coli* genome. The *Eco*RI and *Bam*HI sites bordering the cloned sequence are identified. Numbers shown below the line correspond to the nucleotide position in the *E. coli* genome. Also shown are the genes in the cloned fragment and the direction of transcription.

Discussion

Seattle Central Community College students have been using bioinformatics for the past 3 years as a tool for studying biology in many types of courses. For the most part, students have found the research projects and assignments described here to be useful learning activities. In this section, we will discuss some of our experiences using bioinformatics in the classroom.

A successful experience with bioinformatics will depend on the quality of computer support available at an individual school. If the school server and student e-mail accounts are unreliable and the internet connection is inadequate for the number of students in a course, then students can easily become frustrated. Instructors must also learn to manage new classroom challenges, such as preventing students from printing 30 pages of a BLAST report. Students can become frustrated when servers are down or they follow links that no longer exist. A BLAST search from the West Coast in the middle of the afternoon can be excruciatingly slow. Although BLAST will return search results by e-mail, students who lack experience with e-mail are reluctant to

make use of this option. These challenges are magnified in entry-level courses or courses with non-science majors because these students are not as comfortable using computers and saving files to a disc.

Effective methods for finding biological information are not intuitively obvious. Often students don't look carefully at their results or even read the entire page on a computer screen. Thus in the classroom it is important to focus on strategies, and to provide practical examples, for how to use and access these growing information resources.

The dynamic environment of the internet presents challenges to designing learning materials and assignments that will work from one quarter to the next. Although large volumes of information are available, links to information change on a regular basis, and the databases themselves change composition daily. Instructors who include pictures of computer screens in handouts must update their written materials every few months to keep up with new versions of software and changing interfaces. Database changes occur frequently as well and can take instructors by surprise. During the 1998 winter quarter, one of us collected DNA sequences and posted them on a web page for students to identify. Two weeks later the composition of the database had changed. Instead of finding the viral DNA polymerase chosen by the instructor, the students found at least 100 complete viral genomes.

An important goal in biotechnology education should focus on demonstrating the utility of databases for rejecting or further researching a hypothesis based on information from a homology search. For example, do matching sequences, with annotations describing a particular function, have links to literature describing biochemical experiments? For a number of reasons, annotations can be incorrect. And as the databases grow, fewer annotations are verified experimentally. Although two sequences might be similar, they might not share the same function.

In this last section, we described some challenges to using bioinformatics effectively in the classroom. As we have described, some of the problems are related to the instructor's experience and to the computing resources at the individual school. Some of the problems are inherent with using a rapidly changing resource like the internet as a classroom tool. And a few problems arise from pioneering resources developed for scientists in classrooms with students who possess little background knowledge of biology. Nevertheless, bioinformatics has an untapped potential as a learning tool. We forsee the day when biological discoveries can be made by amateurs connected to the internet in the same way that amateur astronomers race to find stars and comets.

Acknowledgements

We thank the hard-working biotechnology students from the classes of 1997–1999. Seattle Central Community College, for their willingness to experiment. We also thank our biology students and non-science majors and appreciate their enthusiasm for trying new things. SGP was supported in part by Bio-Link, a National Science Foundation Advanced Technology Education center, NSF Award No. 9850325.

References

1 Altschul SF, TL Madden, A Schäffer, J Zhang, Z Zhang, W Miller and DJ Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
2 Benson DA, MS Boguski, DJ Lipman, J Ostell, BF Ouellette, BA Rapp and DL Wheeler. 1999. GenBank. Nucleic Acids Res. 27: 12–17.
3 Blattner FR, G Plunkett III, CA Bloch, NT Perna, V Burland, M Riley, J Collado-Vides, JD Glasner, C Rode, G Mayhew, J Gregor, N Davis, H Kirkpatrick, M Goeden, D Rose, B Mau and Y Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1474.
4 Casadaban MJ, J Chou and S Cohen. 1980. *In vitro* gene fusions that join an enzymatically active beta-galactosidase segment to amino-terminal fragments of exogenous proteins: *Escherichia coli* plasmid vectors for the detection and cloning of translational initiation signals. J Bacteriol 143: 971–980.
5 Ouellette BF and MS Boguski. 1997. Database divisions and homology search files: a guide for the perplexed. Genome Res 7: 952–955.
6 Quentin Y, G Fichant and F Denizot. 1999. Inventory, assembly and analysis of *Bacillus subtilis* ABC transport systems. J Mol Biol 287: 467–484.
7 Schuler GD, JA Epstein, H Ohkawa and JA Kans. 1996. Entrez: molecular biology database and retrieval system. Meth Enzymol 266: 141–162.