MICROBIOLOGY EDUCATION, May 2004, p. 30-35 VOLUME FIVE

A Web-Based Comparative Genomics Tutorial for Investigating Microbial Genomes

MICHAEL STRONG*, DUILIO CASCIO, AND DAVID EISENBERG
Howard Hughes Medical Institute, UCLA-DOE Institute of Genomics and Proteomics, Molecular Biology Institute, University of California, Los Angeles, Box 951570, Los Angeles, California 90095-1570

As the number of completely sequenced microbial genomes continues to rise at an impressive rate, it is important to prepare students with the skills necessary to investigate microorganisms at the genomic level. As a part of the core curriculum for first-year graduate students in the biological sciences, we have implemented a web-based tutorial to introduce students to the fields of comparative and functional genomics. The tutorial focuses on recent computational methods for identifying functionally linked genes and proteins on a genome-wide scale and was used to introduce students to the Rosetta Stone, Phylogenetic Profile, conserved Gene Neighbor, and Operon computational methods. Students learned to use a number of publicly available web servers and databases to identify functionally linked genes in the *Escherichia coli* genome, with emphasis on genome organization and operon structure. The overall effectiveness of the tutorial was assessed based on student evaluations and homework assignments. The tutorial is available to other educators at http://www.doe-mbi.ucla.edu/~strong/m253.php.

With the emergence of high-throughput DNA sequencing, the availability of complete microbial genomes has increased at an accelerated pace. Research institutions such as the Sanger Center, Pasteur Institute, and The Institute for Genomic Research have sequenced and catalogued over 100 microbial genomes, many of which are publicly available via web-based servers.

As educators, it has become increasingly important to train students in classical methods of microbial analysis as well as introduce them to the emerging field of comparative genomics. Complementing traditional methods of microbial analysis and education, the field of comparative genomics introduces students to topics ranging from prokaryotic genome organization to complex metabolic networks.

The availability of completely sequenced genomes has led to the development of a number of computational methods to identify functionally linked genes and proteins on a genome-wide scale. Among these are the Rosetta Stone (2), Phylogenetic Profile (5), conserved Gene Neighbor (1, 3), and Operon (8) computational methods.

The Rosetta Stone method identifies individual genes that occur as a single fusion gene in another organism. For example the *Escherichia coli* gyraseA and gyraseB genes (both involved in DNA replication) occur as a single fusion gene in yeast, topoisomerase II (2). The Phylogenetic Profile method links genes that have a correlated presence or absence in multiple genomes. For example the *E. coli* flagellar genes flgL and flgG are both present in a number of motile bacterial species but are absent in nonflagellar microorganisms (5). The conserved Gene Neighbor method identifies genes that occur in close chromosomal proximity in multiple genomes, such as the GroEL and GroES chaperone genes. This conserved organization often reflects the clustering of genes of related function as well as bacterial operon organization. Lastly, the Operon method identifies genes likely to belong to a common operon based on the nucleotide distance between adjacent genes in the same genomic orientation (6, 8).

All four computational methods can be applied to identify functionally linked proteins on a genome-wide scale (9). These methods can also be used to aid in the inference of protein function for previously uncharacterized proteins. Functional linkages among proteins may indicate proteins that participate in a common biochemical pathway, proteins that physically interact via protein-protein interactions, or proteins that serve related functions within the cell.

In addition to advancements in comparative genomics, a number of web-based servers have been implemented to aid in the investigation of microbial genomes. From genome browsers such as the Pasteur Institute GenoList to comprehensive databases of raw genome sequences such as those at the National Center for Biotechnology Information website, it has become important to expose students to the wide availability of genomic databases and web servers.

We have developed and implemented a web-based comparative genomics tutorial that introduces students to the concepts of the Rosetta Stone, Phylogenetic Profile, conserved Gene Neighbor, and Operon computational methods. Throughout the tutorial, students are exposed to a number of genome databases and web servers that are used to investigate microbial genome organization as well as to identify functional linkages among microbial proteins. The goals of the comparative genomics tutorial are two-fold. First, we have attempted to provide students a strong foundation in the computational concepts and terminology, and secondly, we have tried to expose students to a number of web-based genome resources and databases that they may find useful in future research activities.

In Ronald Owston's article "The World Wide Web: A Technology to Enhance Teaching and Learning," he discusses three specific advantages the web provides that can be uti-

*Corresponding author. Mailing address: Howard Hughes Medical Institute, UCLA-DOE Institute of Genomics and Proteomics, Molecular Biology Institute, University of California, Los Angeles, Box 951570, Los Angeles, CA 90095-1570. Phone: 310-206-3642. Fax: 310-206-3914. E-mail: strong@mbi.ucla.edu.

lized by instructors to "promote improved [student] learning" (4). He states that the first advantage is that the "web appeals to the mode of student learning" since many students are accustomed to working with computers in their everyday life. Owston comments that the "computer has become an integral part of [the students'] world…and that they thrive on interacting with [the computer]." The second advantage detailed by Owston is that the "web provides for a flexible learning environment…that enables students to take advantage of the wealth of learning opportunities available through the Internet." Owston notes that this is often utilized by instructors to create a more "project-based" learning environment. The third advantage discussed by Owston is that "the web enables new kinds of learning…that can promote critical thinking and problem solving skills" since projects involving the web often require students to evaluate a variety of data from a variety of sources (4).

In our functional genomics tutorial, we have employed a combined strategy that includes web-based instruction in conjunction with traditional teacher-based instruction. Bruce Tuckman described a related model Active Discovery and Participation through Technology (ADAPT) (10), in which a hybrid method of web-based instruction and traditional teacher-based instruction was used to teach a study skills course at Ohio State University. Tuckman demonstrated that the ADAPT model, which employed both web-based and traditional instructional methods, increased student learning as compared to traditional lecture style instruction alone. Tuckman noted that this hybrid ADAPT model allowed students to become more "actively involved in the learning process," through a variety of computer-based instructional activities (10).

Since the field of comparative genomics relies heavily on the use of computers and web-based resources, we hypothesized that a web-based tutorial would be an effective method to introduce students to the field of comparative genomics, as well as to teach them how to use a variety of web-based resources and databases to investigate functional linkages among prokaryotic genes. We assess our hypothesis using both student evaluations and student homework assignments, which suggest that our web-based tutorial is an effective method to introduce students to the field of comparative genomics.

METHODS

Figure 1 shows the introductory screen of our web-based comparative genomics tutorial. This site can be accessed at http://www.doe-mbi.ucla.edu/~strong/m253.php. The tutorial consists of 25 web pages partitioned into five sections.

We have implemented this tutorial in conjunction with the core curriculum for first-year graduate students in the biological sciences at the University of California, Los Angeles. Each student attended one of five computer-based laboratory sessions in which the instructor presented the material in the tutorial. The average enrollment for each computer-based laboratory was approximately 27 students. The students were each assigned their own computer terminal and proceeded through the initial tutorial along with the instructor. This gave students the chance to ask questions along the way and helped emphasize the concepts covered in the tutorial.

The first section of the comparative genomics tutorial introduced students to the concepts of the Rosetta Stone, Phylogenetic Profile, conserved Gene Neighbor, and Operon methods, while the subsequent four sections involved an introduction and demonstration of the four databases shown in Fig. 2. The first database discussed in the tutorial was the European Molecular Biology Laboratory Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) server (12). The STRING server is used to identify genes linked by either the Rosetta Stone, Phylogenetic Profile, or conserved Gene Neighbor Method. The STRING server can be accessed at http://www.bork.embl-heidelberg.de/STRING, or as a link from our tutorial.

In order to demonstrate the applications of each of the four web servers we chose the *E. coli otsA* gene for demonstration purposes. The *E. coli* otsA protein is involved in the first step of trehalose biosynthesis and catalyzes the biosynthesis of trehalose-6-phosphate from UDP-glucose and glucose-6-phosphate (11). Although the *E. coli otsA* gene was used for demonstration purposes, it was emphasized to students that the methods applied in this tutorial could be applied to any gene of interest.

Using the comparative genomics tutorial, students proceeded through a step-by-step introduction to the STRING server. Each student submitted the query gene (*otsA*) to the STRING server on his or her own computer terminal. The *E. coli otsA* gene was chosen for demonstration purposes because it is linked to a single gene by the Rosetta Stone, Phylogenetic Profile, and conserved Gene Neighbor computational methods. While the *otsA* gene demonstrates a simplified example, it enabled students to become familiar with both the computational concepts and the introduced databases. Student homework assignments involved protein linkages of higher complexity.

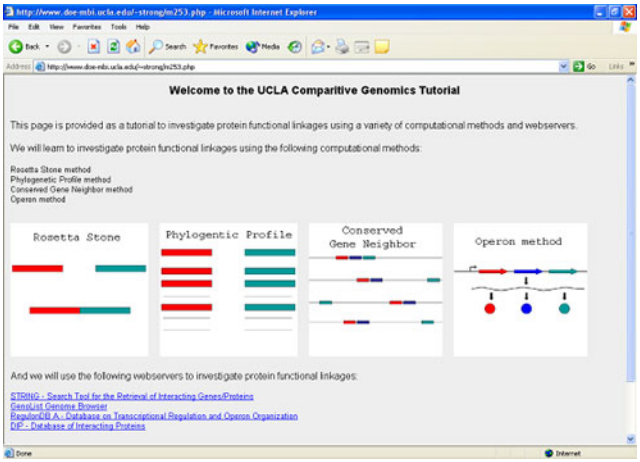

FIG. 1. Introductory page for the web-based comparative genomics tutorial. This tutorial is available at http://www.doe-mbi.ucla.edu/~strong/m253.php.

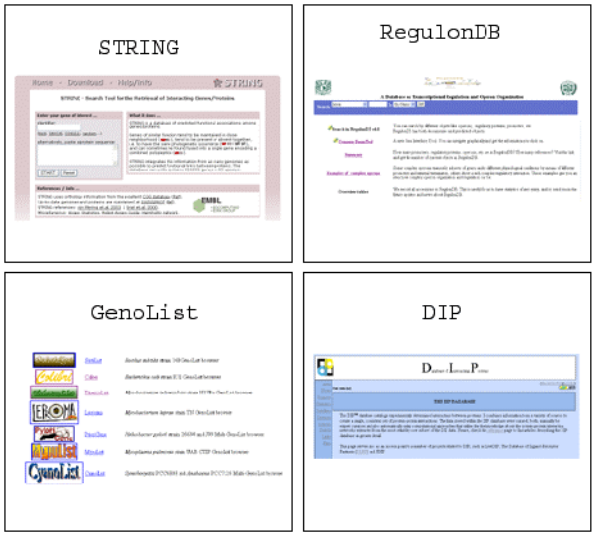

FIG. 2. Overview of the four databases and web servers used in the web-based tutorial: European Molecular Biology Laboratory Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), Pasteur Institute GenoList, RegulonDB, and Database of Interacting Proteins.

Using the STRING server the students found that the *otsA* gene was linked to a single gene (*otsB*) by the Rosetta Stone, Phylogenetic Profile, and conserved Gene Neighbor computational methods. The functionally linked gene, *otsB*, is involved in the second step of trehalose biosynthesis. OtsB is a phosphatase that dephosphorylates trehalose-6-phosphate to yield trehalose. Figure 3a-c summarizes the results of the *otsA* STRING query. The *otsA* and *otsB* genes occur in close chromosomal proximity in multiple genomes (Fig. 3a) linking them by the conserved Gene Neighbor method. The *otsA* and *otsB* genes also have a correlated presence or absence in a number of genomes (Fig. 3b) linking them by the Phylogenetic Profile method, and *otsA* and *otsB* occur as a single fusion gene in *Pyrobaculum aerophilum* (Fig. 3c) linking them by the Rosetta Stone method.

The second database introduced was the Pasteur Institute GenoList web server (http://genolist.pasteur.fr). GenoList was used to examine prokaryotic genome organization in *E. coli*. Continuing with the analysis of the *E. coli otsA* gene, students examined the genome organization of this gene in the *E. coli* K-12 genome. Students learned to navigate the GenoList Colibri (*E. coli*) web server by following the examples illustrated in the tutorial. The genome organization of the *otsA* gene revealed that this gene overlaps another gene (*otsB*) by 25 bp. The overlap of adjacent genes in the same orientation is a common feature of prokaryotic operon organization (6), and we therefore link these two genes by the Operon method (Fig. 3d).

The third database discussed in the web-based tutorial was RegulonDB (7) (http://www.cifn.unam.mx/Computational_Genomics/regulondb). RegulonDB contains information regarding a large number of experimentally documented *E. coli* operons, as well as computationally inferred *E. coli* operons. The comparative genomics tutorial details the navigation of this site.

The final database in the web-based tutorial was the Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu) (13). This database contains a record of thousands of published protein-protein interactions, with the majority of interactions identified in yeast. Protein interactions involving the yeast *otsA* homologue, TPS2, were demonstrated using the Database of Interacting Proteins.

RESULTS

The comparative genomics tutorial was taught during the second week of the University of California, Los Angeles, graduate course M253 (Macromolecular Structure). A total of five tutorial sessions were given throughout the week to accommodate all students. In order to evaluate the effectiveness of the tutorial, we administered a student evaluation where students ranked various aspects of the tutorial using a 0 to 9 scale, where 9 indicated strong agreement with the statement (or a positive response) and 0 indicated strong disagreement with the statement (or a negative response). The paper-based, voluntary evaluation was handed out in class to all students at the end of each tutorial session. A total of 136 students turned in completed evaluation forms. The anonymous evaluation form included a question regarding the student's current educational year and intended major, and six questions regarding their personal assessment of the tutorial (Table 1).

The current educational year of the 136 respondents was as follows: 71% of respondents were first-year graduate students, 14% were second-year graduate students, 5% were third-year graduate students, 7% were undergraduates, and 3% did not specify their year.

The undergraduate and graduate disciplines of the respondents varied dramatically and included: biology, mo-

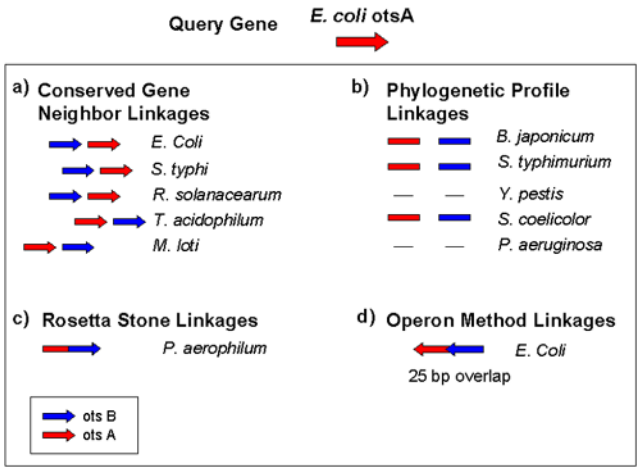

FIG. 3. Summary of functional linkages involving the *E. coli otsA* gene, which is involved in the first step of trehalose biosynthesis. This gene is linked to *otsB*, which is involved in the second step of trehalose biosynthesis, by the conserved Gene Neighbor, Phylogenetic Profile, Rosetta Stone, and Operon computational methods.

TABLE 1. Results of student evaluations

Question	Mean[a]	Median[a]	Standard deviation
Value of the tutorial justified time and effort	7.6	8	1.6
Your overall rating of the tutorial	7.6	8	1.5
The examples in the tutorial were clear and concise	7.6	8	1.7
Would you recommend this tutorial to other students interested in Comparative Genomics	7.9	8	1.6
The combination of the web-based tutorial and in-class/lab demonstrations were helpful	7.8	8	1.5
After completing the tutorial, do you feel comfortable using the web servers presented	7.3	8	1.6

[a] Students ranked various aspects of the web-based comparative genomics tutorial on a scale of 0 to 9, where 9 indicated strong agreement with the statement or a positive response and 0 indicated strong disagreement with the statement or a negative response.

lecular biology, biochemistry, cell biology, pharmacology, physiology, bioengineering, computer science, biomathematics, public health, environmental sciences, chemistry, chemical engineering, psychology, neuroscience, cybernetics, and mathematics. Since the educational backgrounds of the students varied dramatically, we had a broad audience in which to assess our web-based tutorial.

The results of the student evaluations are presented in Table 1. Although the students had a broad array of educational backgrounds, with most students in disciplines outside the fields of genomics or bioinformatics, the majority of student responses suggested an overall positive reaction to the tutorial.

The two questions receiving the highest number of positive responses were of great interest to us. First, we wanted to know if students would recommend this tutorial to other students interested in comparative genomics, and second we were interested to see if the students liked the combination of a web-based tutorial and in-class demonstrations. Both of these items received high scores from respondents with a mean of 7.9 and 7.8 and a median of 8 and 8 respectively. This indicated to us that the majority of the students felt this tutorial would be helpful for new students interested in the field of comparative genomics. It also suggested that the combination of a web-based tutorial and in class demonstrations were well received by the students.

We were also interested to see if the students thought the tutorial examples were clear and concise. This question yielded the greatest standard deviation, and probably reflected the diversity of student educational backgrounds. While the majority of students agreed that the tutorial was clear and concise (median = 8), a small minority of students did have some difficulty following the tutorial. The overall assessment of the tutorial yielded a mean of 7.6. This is quite a positive response, since only a small fraction of the students were specifically focused on a discipline related to that of the tutorial. For the most part, students also felt comfortable navigating the databases and web servers following the tutorial (mean 7.3, median 8), and felt that the tutorial was worth the time and effort they put into it (mean 7.6, median 8).

A 25-point homework assignment was also given to students to complete over a 1 -week period. The complete homework assignment, with answers, is shown in Fig. 4. The homework assignment assessed the students' understanding of the material covered in the tutorial, as well as required students to apply their newly acquired skills to investigate a new set of genes using the various databases and web servers. Since our web-based tutorial was available online, students were encouraged to go back to the web tutorial to reinforce any concepts they may have had difficulty with during the class. The home page of the web tutorial also had links to all of the web servers that the students needed to complete the homework assignment.

Of the 134 students that turned in the homework assignment, 78 students received a perfect score or only missed a single point, 54 students received 23 points, and two students received 21 points. Most students did quite well on the homework, emphasizing that the methods we employed were effective in teaching the students not only the general concepts covered in the tutorial but also enabled them to independently identify functionally linked genes and proteins using the discussed web servers and databases.

CONCLUSIONS

Here we have described an interactive web-based tutorial that we designed to introduce students to the field of comparative microbial genomics. This tutorial complements traditional lessons in microbiology and helps students con-

Homework Questions: (25 points total)

1. Using the Colibri webserver at the Pasteur Institute, do you think panD might be in an operon with any other genes? Why or Why not. (2pts)

Answer: No, panD is flanked by two genes in the opposite orientation

2. Do you see any other genes in the panD region that may participate in a similar pathway as panD (pantothenate biosyntheis). If so, what can you say about the genome organization these genes. (2pts)

Answer: Yes, panC and panB also participate in pantothenate biosynthesis and are organized in a potential operon since they are in the same orientation and are separated by minimal distance (~12bp).

3. Using the EMBL STRING webserver, answer the following questions.

a) Does panD occur as a conserved Gene Neighbor (Neighborhood) with any other proteins. (List COG and E.coli K12 gene name. Use confidence cutoff of 0.4) (2pts)

Answer: Yes, 2 genes. COG0414 (panC) and COG0413 (panB)

b) Does panD occur as a fusion protein (Rosetta Stone) with any other proteins. (List COG and E.coli K12 gene name.) (2pts)

Answer: No

c) Does panD share a similar Phylogenetic Profile (Phylogeny) with any other proteins. (List COG and E.coli K12 gene name.) (2pts)

Answer: Yes, 2 genes. COG0414 (panC) and COG0413 (panB)

d) Do these computationally inferred functional linkages make sense in respect to the proteins biochemical functions. (2pts)

Answer: Yes, they all participate in a common biochemical pathway.

4. Using the EMBL STRING webserver, answer the following questions.

a) Does leuC (COG0065) occur as a conserved Gene Neighbor (Neighborhood) with any other proteins. (List COG only. Use confidence cutoff of 0.4) (2pts)

Answer: Yes, 4 Genes, COG0066, COG0473, COG0002, COG0140.

b) Does leuC occur as a fusion protein (Rosetta Stone) with any other proteins. (List COG and E.coli K12 gene name.) (2pts)

Answer: Yes, 1 gene COG0066 (LeuD).

c) What organisms do the leuC fusion proteins occur in? Are these prokaryotic or eukaryotic organisms? (2pts)

Answer: S. cerevisiae and S. pombe. They are both eukaryotic organisms.

d) Can you hypothesize how a fusion protein may arise in an organism that has two genes in a common operon? (2pts)

Answer: A fusion protein may arise from a mutation (nucleotide insertion or deletion) in the stop codon of GeneA. If GeneB is in-frame with GeneA then a fusion protein may result.

e) Does leuC share a similar Phylogenetic Profile (Phylogeny) with any other proteins. (List COG only) (2pts)

Answer: Yes, 10 genes. COG0066, COG0106, COG0107, COG0118, COG0139, COG0141, COG0473, COG0040, COG0140, COG0002.

f) Based on the Phylogenetic Profiles, why do you think some organisms have all of these linked genes while others have none? (3pts)

Answer: The organisms that have all of these genes may need to actively synthesize certain precursors via a common pathway involving these genes, while the organisms that do not have these genes may either acquire the precursors from the environment or may not need the precursors. (Note, it is likely that all organisms will need the precursors since some of them are essential amino acids, ie Leucine).

FIG. 4. Functional genomics homework problem set.

sider microbial organisms from a genomic perspective. In addition to providing a foundation in computational concepts and terminology, this tutorial introduces students to a variety of web servers and genomic databases. It is our hope that these exercises will promote critical thinking and independent learning skills, since the resources presented in the tutorial can be applied to investigate a diversity of research projects.

Both the student evaluations and the homework assignments suggest that the web-based comparative genomics tutorial was an effective teaching tool that provided a clear introduction to the field of comparative genomics, as well as taught students the skills necessary to navigate a variety of web-based servers and databases in order to identify functionally linked genes and proteins in prokaryotic organisms. These results further support the use of a hybrid instructional model (10) that incorporates web-based instruction in conjunction with traditional teacher-based instructional methods. The assessment of the homework assignments also supports the notion that the web can promote "critical thinking and problem solving skills" (4) since the homework assignments required students to creatively apply their knowledge to solve a series of problems using a variety of databases and web servers.

It is likely that it will become increasingly important to train students in the field of comparative genomics since the number of sequenced genomes will continue to rise at an accelerated pace. The comparative genomics tutorial presented here is available at http://www.doe-mbi.ucla.edu/~strong/m253.php.

REFERENCES

1. **Dandekar, T., B. Snel, M. Huynen, and P. Bork**. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. **23:**324–328.
2. **Marcotte, E. M., M. Pellegrini, N. Ho-Leung, D. Rice, T. Yeates, and D. Eisenberg.** 1999. Detecting protein function and protein-protein interactions from genome sequences. Science **285:**751–753.
3. **Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev.** 1999. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA **96:**2896–2901.
4. **Owston, R. D.** 1997. The world wide web: a technology to enhance teaching and learning? Educ. Researcher **26:**27–33.
5. **Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates.** 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA **96:**4285–4288.
6. **Salgado, H., G. Moreno-Haelsieb, T. Smith, and J. Collado-Vides.** 2000. Operons in *Escherichia coli*: genomic analysis and predictions. Proc. Natl. Acad. Sci. USA **97:**6652–6657.
7. **Salgado, H., A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez, and J. Collado-Vides.** 2001. RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. Nucleic Acids Res. **29:**72–74.
8. **Strong, M., P. Mallick, M. Pellegrini, M. J. Thompson, and D. Eisenberg.** 2003. Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. Genome Biol. **4:**R59.
9. **Strong, M., T. G. Graeber, M. Beeby, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg**. 2003. Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps. Nucleic Acids Res. **31:**7099–7109.
10. **Tuckman, B. W.** 2002. Evaluating ADAPT: a hybrid instructional model combining web-based and classroom components. Computers Educ. **39:**261–269.
11. **Tzvetkov, M., C. Klopprogge, O. Zelder, and W. Liebl.** 2003. Genetic dissection of trehalose biosynthesis in *Corynebacterium glutamicum*: inactivation of trehalose production leads to impaired growth and an altered cell wall lipid composition. Microbiology **149:**1659–1673.
12. **von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel.** 2003. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. **31:**258–261.
13. **Xenarios, I., L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg.** 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. **30:**303–305.