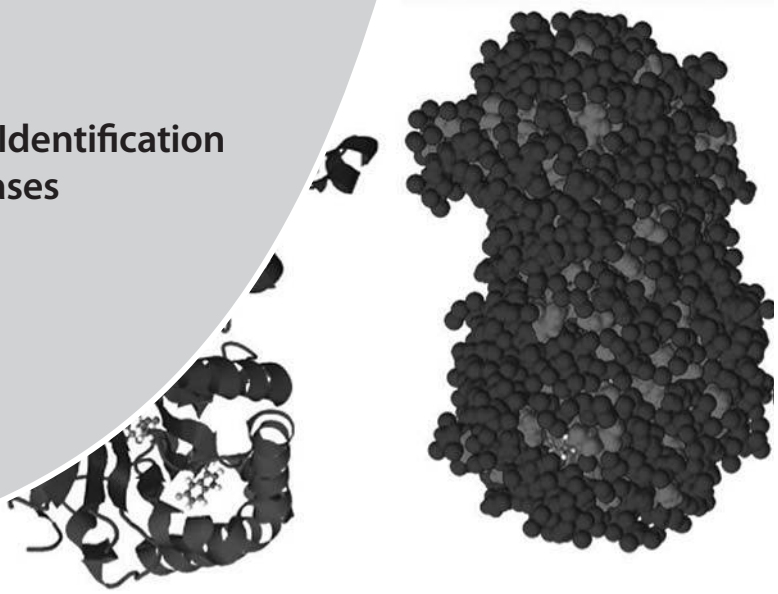
HOW TO DO IT

Proteomics: Protein Identification Using Online Databases



CHRIS EURICH, PETER A. FIELDS, ELIZABETH RICE

Abstract

Proteomics is an emerging area of systems biology that allows simultaneous study of thousands of proteins expressed in cells, tissues, or whole organisms. We have developed this activity to enable high school or college students to explore proteomic databases using mass spectrometry data files generated from yeast proteins in a college laboratory course. Students upload files of "unknown" proteins from our public website, enter them into a proteomics search engine (Mascot), identify the proteins, and use additional proteomics websites to learn about their functions and three-dimensional structures. This activity is suitable for use in units exploring protein structure and function, metabolism, or bioinformatics.

Key Words: *Bioinformatics; mass spectrometry; PAGE; peptides; proteins; proteomics; yeast.*

Bioinformatics, the discipline of biology that employs computerized search algorithms and extensive databases of biological information to investigate biological processes and relationships, has grown exponentially in the past decade. Genomics continues to be the best-known and most data-rich area of bioinformatics; the Human Genome Project, as well as the sequencing of genomes from many other species, has amassed genetic data from laboratories around the world. These data are available in public databases such as the National Center for Biotechnology Information (NCBI; note that the web addresses for websites discussed in this article can be found in Table 1). Many genomics-oriented educational activities have been developed to allow students to use genomic data repositories to study biological questions (BSCS, 2003; Buxeda & Moore-Russo, 2003; Wefer, 2003; Herron et al., 2010).

In addition to genomics, proteomics is a growing discipline of bioinformatics. Like the extensive databases of nucleotide sequences, there are also databases containing amino acid sequences of proteins isolated from species as diverse as bacteria and humans. These databases allow for the discovery and analysis of protein properties in a manner analogous to nucleotides. Thus, proteomics is the area of bioinformatics that makes use of these amino acid sequence databases and allows examination of all the proteins expressed by a cell, tissue, or organism.

Therefore, proteins determine the phenotype of the cell and the organism.

Why put such emphasis on proteins? Proteins are almost exclusively responsible for cellular function and metabolism, as well as for much of cellular structure. Therefore, proteins determine the phenotype of the cell and the organism. Being able to identify and examine the proteins present in cells or tissues, and compare protein expression among groups, can provide important information concerning an organism's physiology, health, or evolutionary history. Thus, proteomics has been the recent focus of much research and technology development (Yates et al., 2009), and the field will grow in importance as scientists explore the links among genes, protein expression, and biological function (Gstaiger & Aebersold, 2009).

○ Background

The activity begins with data files generated by undergraduates at Franklin & Marshall College (F&M) during a semester-long proteomics laboratory focused on environmental stress in yeast. Baker's yeast (*Saccharomyces cerevisiae*) is an excellent study organism for bioinformatics laboratories because its genome has been sequenced, there are >58,000 *S. cerevisiae* protein sequences available in the UniProt database, and its complex eukaryotic metabolism is comparable to those of multicellular organisms like plants and animals.

Here, we briefly describe the process by which these data files were produced; detailed descriptions of these lab procedures can be found in the background materials on the Teaching Bioinformatics website (Table 1). Students chose environmental stresses to investigate, including heat, glucose starvation, high ethanol concentration, and hydrogen peroxide (oxidative stress). Differences in protein expression between control and stressed yeast were determined using 2-D polyacrylamide gel electrophoresis (PAGE) (Figure 1). The students located proteins of interest, those that changed in abundance after stress, and cut them from the 2D-PAGE gels. The proteins of interest were digested into smaller peptides using the enzyme trypsin. A liquid chromatography-tandem mass spectrometer (LC-MS/MS)

The American Biology Teacher, Vol. 74, No. 4, pages 250–255. ISSN 0002-7685, electronic ISSN 1938-4211.
DOI: 10.1525/abt.2012.74.4.8

measured the molecular mass of each peptide, as well as the mass of additional "daughter" peptides generated by fragmentation during the analysis.

Students begin this activity with data files generated by F&M undergraduates, available for download from the Teaching Bioinformatics website. They use these peptide mass data in order to identify proteins associated with stress responses in yeast. We provide guidance to additional bioinformatics websites that will allow students to explore the structure and function of these proteins.

○ Teacher Tips

The activity, in different forms, suits undergraduates or advanced high school students. One of us is a high school biology teacher who teaches an elective in biotechnology and another is a college professor who teaches a biochemistry course. In the introductory format, this exercise is appropriate for either a high school unit on protein structure and function or a high school biotechnology unit focused on computer searches of biological databases; the advanced format is intended for upper-level undergraduates. Depending on the amount of time assigned outside the classroom for students to research background information, the introductory-level exercise can be completed in one to two 85-minute blocks or two to three traditional 45-minutes periods. We have had very positive feedback about this activity from high school students. They like using authentic lab data generated by college students, performing the online searches, and viewing the 3D models of the proteins. Students feel as if they are doing what real researchers would do instead of completing a workbook simulation. Not knowing what they will find when completing a query also adds to the authenticity of experimental science.

Before completing this activity, students will need an understanding of the basic structures and properties of amino acids, proteins, and enzymes. They will also need to know how electrophoresis can separate molecules by their size and charge and how a mass spectrometer can be used to determine molecular masses. To help students better understand the proteomic laboratory techniques used to generate the data files presented here, we recommend they view an excellent tutorial

Table 1. Names and web addresses of sites discussed in the text.

Website	Web Address
National Center for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/
Tutorial – Guide to Sequencing & Identifying Proteins	http://www.childrenshospital.org/cfapps/research/data_admin/Site602/mainpageS602P0.html
Mascot Proteomic Search Engine	http://www.matrixscience.com/home.html
F&M Teaching Bioinformatics website	http://teachingbioinformatics.fandm.edu/node/76
UniProt Proteomic Database	http://www.uniprot.org/
Protein Data Bank (PDB)	http://www.pdb.org/pdb/home/home.do

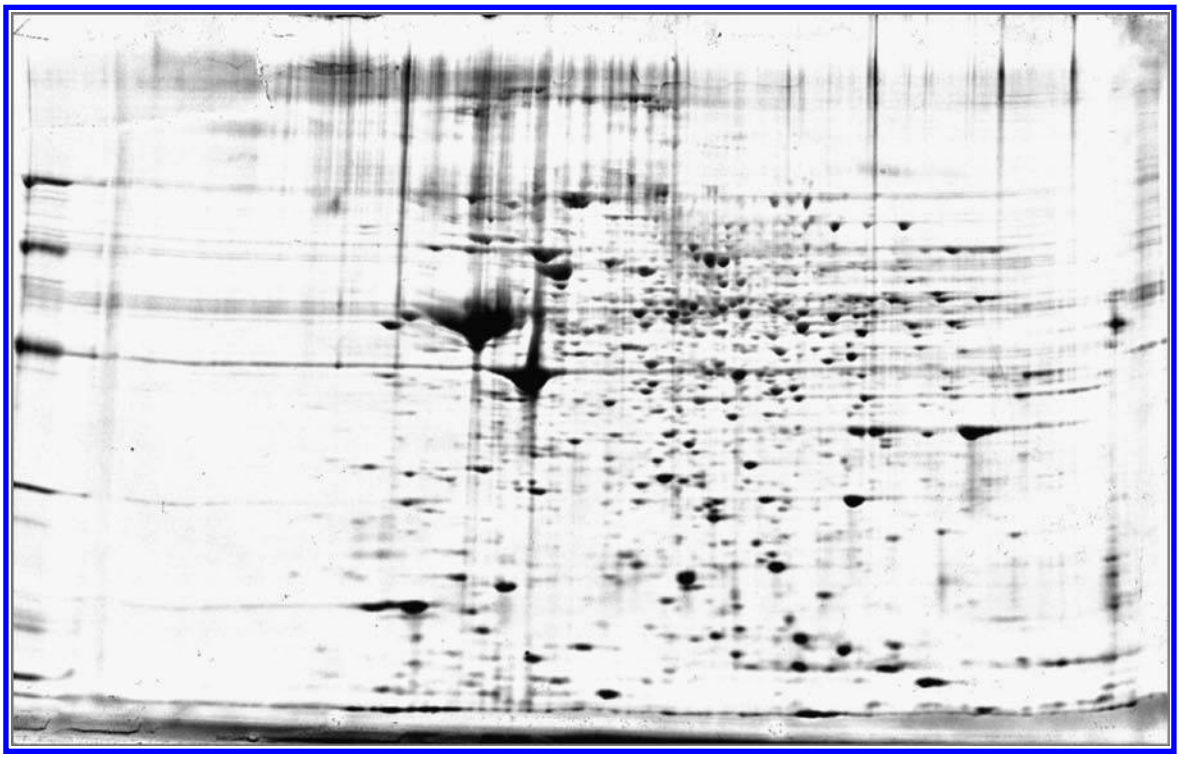


Figure 1. Two-dimensional polyacrylamide gel in which yeast proteins are separated by isoelectric point (horizontal axis) and molecular mass (vertical axis). Each spot on the gel represents an individual protein; specialized gel analysis software can detect approximately 800–1,200 protein spots on a gel like this.

entitled "Guide to Sequencing & Identifying Proteins" (Table 1), produced by Children's Hospital, Boston.

A simplified student procedure for this activity, along with a summary worksheet, is included in the file entitled "Student Worksheet—Introductory" on the Teaching Bioinformatics website. We highly recommend that teachers also view the tutorial and complete several practice proteomic searches (described below) prior to implementing this activity in the classroom. For college classes, a "Student Worksheet—Advanced" is provided on the website, where additional background information concerning the procedures can also be found.

○ Procedure

The data produced by the LC-MS/MS are saved electronically in a format – the "Mascot generic file," or mgf – that can be uploaded into the Mascot online proteomic search engine (Table 1). Mascot examines the masses of the peptides and peptide fragments and compares them to the predicted masses of known amino acid sequences from an online database such as UniProt (Table 1). Mascot will return a list of probable protein identifications, along with a measure of statistical confidence for each, based on the number of "hits" (peptide matches) that the software finds between the LC-MS/MS data and amino acid sequences in known proteins. The more hits registered, the greater the confidence one has in the identity of a particular protein.

1. Downloading Files

To complete the activity, students will need to access the Teaching Bioinformatics website, which includes

a. 40 mgf data files, each representing a single protein whose abundance increased after yeast were exposed to one of the four

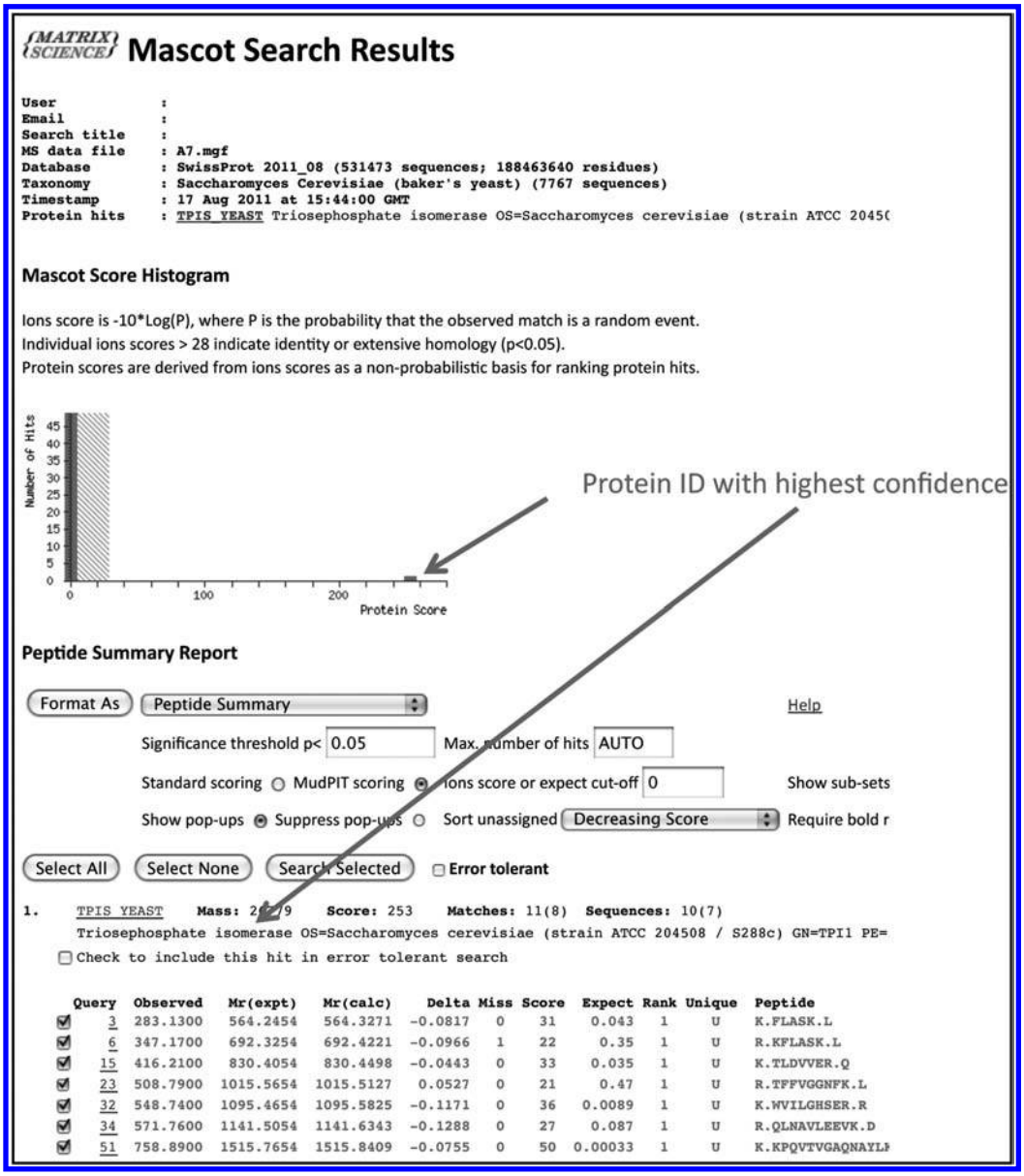


Figure 2. Mascot summary report for the file A7.mgf. The summary report indicates the level of statistical confidence that Mascot assigns an identification (protein score) and, if there is a significant match, the identity of the protein. In this case, Mascot identified yeast triosephosphate isomerase (TPI) with high confidence, based on matches between the TPI sequence in the SwissProt database and the sequences of seven peptides determined from MS data (bottom of figure).

stress treatments (heat, ethanol, H_2O_2, or acetate as the sole carbon source);

b. an Excel spreadsheet (intended for the teacher only) that identifies the 40 proteins and provides basic descriptive information for each, including the stress that induced its overexpression;

c. two background documents that describe the principles of proteomics and how these mgf data files were produced (one of these documents is introductory; the second is advanced and provides a more in-depth, college-level explanation of the procedures used in the F&M proteomics lab to produce the data); and

d. two sample evaluation worksheets – one introductory (high school level) and the second more advanced (college level) – that can be used by students as they perform the activity.

The mgf files are stored in a zipped folder and can be taken directly from the website by students, or the teacher can download them in advance of the activity and make them available for students during class. Teachers may wish to assign proteins randomly or based on function, stress, 3D structure, or other factors.

2. Mascot Search & Protein Identification

Once the students have downloaded their mgf files, they log on to the Mascot search website. There they click on the "Mascot" link at the top of the page and choose the "MS/MS Ions Search" option, which will take them to the search form. To initiate the search, the students type in a user name, e-mail address, and search title that Mascot will use to e-mail their search results if they are disconnected during analysis. The e-mail addresses are used for no other purpose.

The program allows the user to vary numerous search parameters. Although these can be helpful during advanced searches, they are unnecessary and potentially confusing for first-time users. For the introductory-level activity, students will only need to verify or select the following parameters from the menu choices on the Mascot MS/MS Ions Search data entry page:

- Database: SwissProt
- Enzyme: Trypsin
- Missed Cleavages: 1
- Taxonomy: Yeast (*Saccharomyces cerevisiae* – baker's yeast)
- Data file: Students should use the "Browse" button to locate the mgf file they wish to analyze.
- All other fields should be left as the default settings.

Once the students have chosen the parameters, selecting the "Start Search" button at the bottom of the page initiates identification of the unknown protein. Figure 2 illustrates the initial Mascot "Peptide Summary Report" of a search based on the file A7.mgf, one of the files included on the Teaching Bioinformatics website.

As shown in the histogram of Figure 2, one protein from the SwissProt protein database matches the peptides of the mgf file with higher confidence than any other – in this case, triosephosphate isomerase (TPI) has a protein score of 253, well above the probability

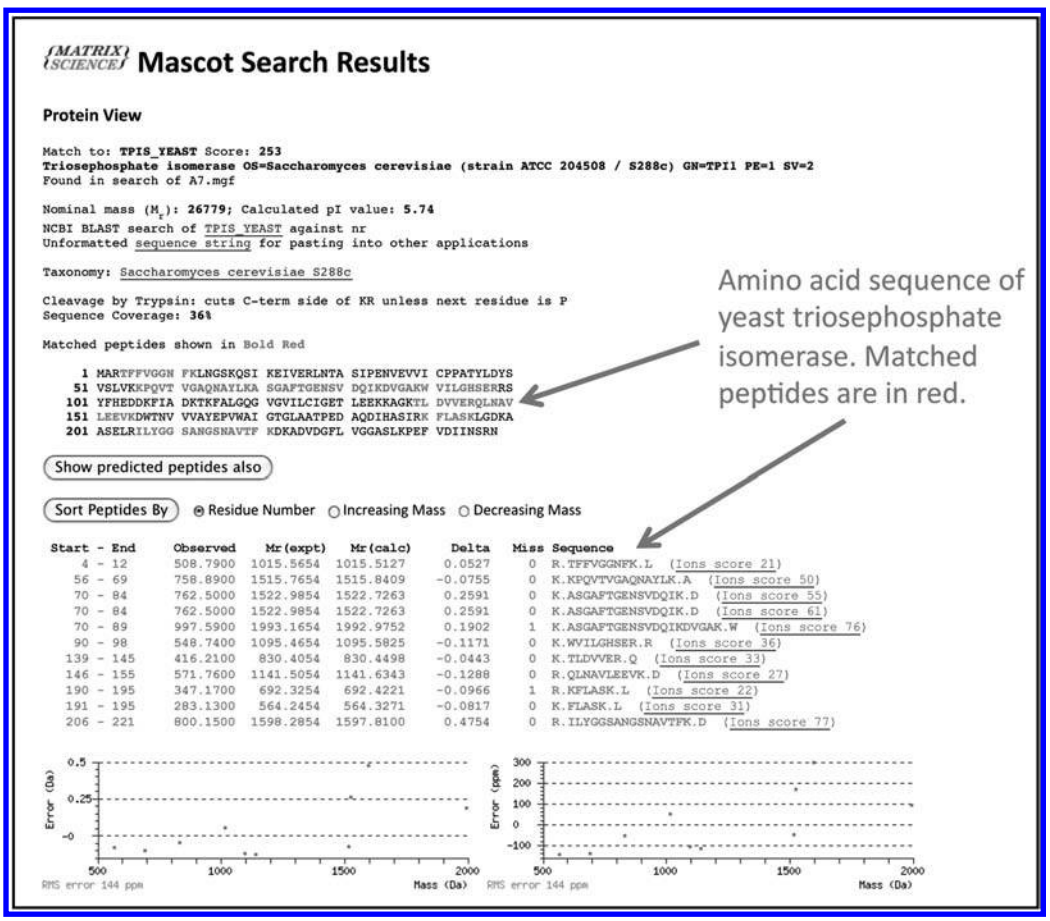


Figure 3. Mascot protein view report for the file A7.mgf, identified as yeast triosephosphate isomerase (see Figure 2). The protein view shows the complete amino acid sequence of TPI, and the location of identified peptides within the sequence.

cut-off (green shaded box) of 28 ($P < 0.05$). As listed in the spreadsheet for teachers on the Teaching Bioinformatics website, TPI was up-regulated in response to both oxidative stress and growth on acetate medium. Further down the page is the protein identification, including its accession tag ("TPIS_Yeast"), name, Mowse score, and a list of the peptides detected by the mass spectrometer that are found in the amino acid sequence of yeast TPI.

By clicking on the accession tag, students will access the Mascot "Protein View" page (Figure 3), which shows the primary sequence (248 amino acids indicated by their one-letter abbreviations) of the entire protein, with peptides identified by mass spectrometry highlighted in red. In this example, the peptides detected by the mass spectrometer cover 36% of the amino acid sequence, which is more than enough to ensure unambiguous identification of the protein.

3. Exploration of Protein Function & Structure

The identification of unknown proteins using mass spectrometry data to query amino acid databases, as described above, is a hallmark of proteomics. However, in many cases the researcher wishes to find out more about the protein identified. For example, many students participating in this activity will not know what role TPI plays in the cell. Students can use the Mascot ID they obtain to quickly determine the function of their protein and to visualize the three-dimensional protein structure.

Students should copy or record the accession tag provided by Mascot (in this case, TPIS_Yeast) and open the UniProt database (Table 1) in a second browser window. After ensuring that the search will be conducted in the "Protein Knowledgebase" (UniProtKB) in the menu at the upper left, the accession tag can be entered into the search window. The search results provide a succinct description of the protein's function, as well as links to additional information about its role in metabolism and literature citations. Many proteins, but not all, will also have 3D structural information available further down the UniProt information page, under the heading "Cross-references – 3D structure databases." If the protein has a link to the Protein Data Bank (PDB), clicking one of the entries will open a new window at the PDB site (Table 1). (Note that the spreadsheet for teachers on the Teaching Bioinformatics website indicates whether there is a corresponding PDB for each mgf file.) If there are multiple PDB entries on the UniProt page, it is best to select the one with the highest resolution – that is, the lowest value in angstroms.

At the PDB site, the name of the protein will be listed and an image will be shown on the right side of the page. Students can click the "view in Jmol" button below the image to view and export a larger, rotatable image (Figure 4). Controls below the image allow highlighting of different structures, and display of the protein using ribbons, a backbone trace, or as ball-and-stick. For college-level courses, the advanced worksheet available on the Teaching Bioinformatics website leads students through an exercise to visualize the secondary structural elements of the protein and the distribution of hydrophilic versus hydrophobic residues within the 3D structure. When ligands are present in the structure, the activity asks students to examine the geometry of the binding site by determining which amino acids are most closely associated with the ligand. These are only a few of the many possible structural explorations that students can undertake using Jmol. If the instructor wishes, the PDB files can be exported to more advanced visualization and modeling software packages such as PyMOL and Swiss-Model.

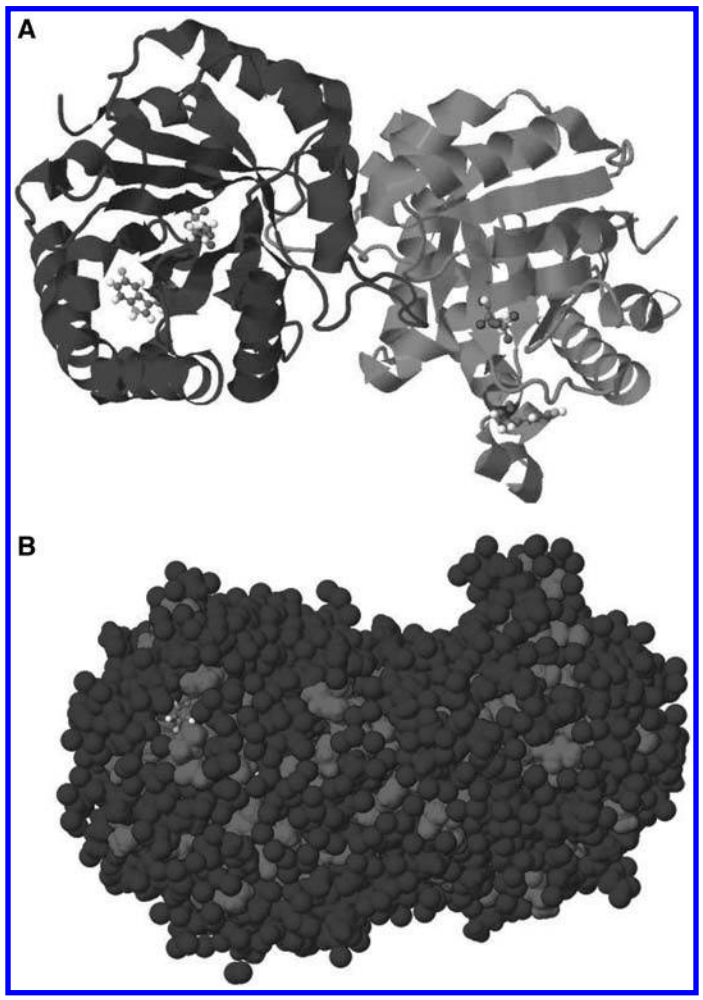


Figure 4. Three-dimensional images of the homodimer of TPI (PDB 1NEY) generated by Jmol at the PDB website. (A) The TPI backbone is illustrated with ribbons while ligands are shown in ball-and-stick; the two identical monomers are colored separately. (B) The same dimer shown in spacefill, illustrating hydrophilic residues (blue) on the exterior of the molecule and hydrophobic residues (red) buried in the interior.

○ Evaluation

Student work in this activity can be evaluated by the use of a worksheet containing a series of questions about the activity; two sample worksheets, an introductory one used in a high school class and an advanced one from a college course, are included on the activity website. Students can also submit printed pictures or PowerPoint files of the 3D structure of the protein. A class discussion of the activity, along with a sharing of student findings, can serve to reinforce the variety of protein functions and structures the students have discovered. More advanced classes can examine the relationship between protein function and protection from the effects of environmental stress.

○ Summary

Though the field of proteomics is rapidly becoming an essential part of biological inquiry, the equipment, time, and resources needed to analyze proteins, as described in the Background section of this article, are far beyond the scope of most high school students and

teachers. We hope to overcome that hurdle by sharing information between the college laboratory and the high school classroom. Both undergraduates and high school students can use the same data files (accessible on the Teaching Bioinformatics website) to see the power of proteomics databases, to identify proteins of interest and investigate their structure and function.

○ Acknowledgments

This work was supported by an award to Franklin & Marshall College (F&M) from the Howard Hughes Medical Institute's (HHMI) Undergraduate Science Education Program, and by a grant from the National Science Foundation to P.A.F. (IOS-0920103). It grew out of a week-long, HHMI-funded bioinformatics summer workshop for high school teachers at F&M. Other bioinformatics activities developed by high school teachers are available at the Teaching Bioinformatics website.

References

BSCS. (2003). *Bioinformatics and the Human Genome Project: A Curriculum Supplement for High School Biology*. Colorado Springs, CO: BSCS.

Buxeda, R.J. & Moore-Russo, D.A. (2003). Enhancing biology instruction with the Human Genome Project. *American Biology Teacher*, *65*, 664–668.

Gstaiger, M. & Aebersold, R. (2009). Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nature Reviews Genetics*, *10*, 617–627.

Herron, S.S., Parr, J., Davis, B. & Nelson, P. (2010). Theme-based instruction: making conceptual ties with the sickle cell story. *American Biology Teacher*, *72*, 422–426.

Jogl, G., Rozovsky, S., McDermott, A.E. & Tong, L. (2003). Optimal alignment for enzymatic proton transfer: structure of the Michaelis complex of triosephosphate isomerase at 1.2-Å resolution. *Proceedings of the National Academy of Sciences, USA*, *100*, 50–55.

Perkins, D.N., Pappin, D.J.C., Creasy, D.M. & Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, *20*, 3551–3567.

Wefer, S.H. (2003). Name that gene: an authentic classroom activity incorporating bioinformatics. *American Biology Teacher*, *65*, 610–613.

Yates, J.R., Ruse, C.I. & Nakorchevsky, A. (2009). Proteomics by mass spectrometry: approaches, advances, and applications. *Annual Reviews of Biomedical Engineering*, *11*, 49–79.

CHRIS EURICH (chris_eurich@etownschools.org) is Chair of the Science Department, Elizabethtown Area High School, 600 East High Street, Elizabethtown, PA 17022. PETER A. FIELDS (peter.fields@fandm.edu) is Associate Professor of Biology, Franklin & Marshall College, PO Box 3003, Lancaster, PA, 17603, where ELIZABETH RICE (erice@fandm.edu) is Assistant Director of the Bioinformatics Program.