CHAPTER 1

What I Could Teach Darwin Using "Darwin 2000", an Interactive Web Site for Student Research into the Evolution of Genes and Proteins

Richard P. Hershberger

Carlow College
Division of Natural Sciences and Mathematics
3333 Fifth Avenue
Pittsburgh, PA 15213
Telephone: 412-578-8702; Fax 412-578-8745
e-mail: rickhershberger@bioactivesite.com
web site: http://www.bioactivesite.com

Rick Hershberger is an Assistant Professor of Biology at Carlow College. He received a B.A. in Chemistry from Carnegie Mellon University in 1980 and a Ph.D. in Molecular Biology and Microbiology from Case Western Reserve University in 1991. His postdoctoral work was in the Institute for Molecular Virology, at the University of Wisconsin–Madison. As a molecular biologist, computer analysis of DNA and protein sequences has been an integral part of his research and teaching. His first job in science was to learn, then train his coworkers, to analyze DNA sequences on an IBM AT personal computer. Today he incorporates online investigations into his courses, uses biocomputing in his cancer virus research, and mentors web-based student research projects. He teaches "Introductory Biology and Chemistry" for nurses, "Teaching Elementary Science" for elementary education students, and "Genetics" and "Molecular Biology/Biotechnology" for biology majors. The Bioactive Site (http://www.bioactivesite.com) contains his teaching materials and online exercises.

Reprinted From: Hershberger, R. P. 2000. What I could teach Darwin using "Darwin 2000", an interactive web site for student research into the evolution of genes and proteins. Pages 1–32, *in* Tested studies for laboratory teaching, Volume 21 (S. J. Karcher, Editor). Proceedings of the 21st Workshop/Conference of the Association for Biology Laboratory Education (ABLE), 509 pages.

Contents

Introduction for Instructors

The objective of the Darwin 2000 website is to provide undergraduate students with a cost-effective and accessible biocomputing research environment using today's bioinformatics tools to support student-driven open-ended investigations, and to illustrate the connections between molecular biology and evolution.

The theme "What I could teach Darwin" draws connections between the fields of evolution and genetics by explaining the relationships between what Charles Darwin observed about physical adaptations in species and today's knowledge of the genetic and biochemical processes that produce and select for those adaptations at the molecular level. By studying and comparing the sequences of subunits that comprise genes or proteins, students learn how random genetic change and natural selection for the proper structure and function of macromolecules are the driving forces underlying the evolution of species. The learning objectives for Darwin 2000 are to understand: (1) the "structure-function"-- relationship that the sequence of subunits within a DNA or protein molecule determines its three-dimensional structure, and thus its function, (2) "homology": that similarities in the sequences of different genes or proteins suggest a relatedness in their functions, (3) "conservation": that the need for proteins to function correctly limits the kinds of random changes that occur over evolution, thus functionally important parts of a molecule remain structurally intact, and (4) "molecular phylogenetics": that similarities in molecular sequences among different species reflect evolutionary relationships.

Darwin 2000, developed by Rick Hershberger at Carlow College, is a series of modular activities that lead a student through guided demonstrations and problem-solving activities to an understanding of these concepts. Common to each module is (a) the use of databases, computational tools, and software available free on the Web, (b) the application of molecular biology principles during problem-solving activities, and (c) the study of research problems frequently encountered in today's biological research. The central pedagogical strategy is for students to use the same biocomputing tools, in the same manner and answering the same research questions, as professional researchers. The students first learn about the sequence analysis tool, then work a guided exercise seeking answers to a pre-defined set of research questions, using hemoglobin as a model molecule to study. Students are then challenged to perform their own open-ended investigations, choosing a protein of interest (typically involved in human disease) and predicting what enzymatic functions the protein performs (based on sequence homology to known proteins) and what regions of the protein are responsible for those functions (based on sequence conservation).

Materials

- A computer connected to the internet via a dial-in or direct (Ethernet) connection. Optimally, each student should work at his/her own computer, but teams of two or three students can work collaboratively at a single computer if resources are limited. Because all of the needed software and web sites are free, students can conduct these activities on their home or dorm computers.
- A frames-capable, Java-enabled web browser. Netscape Navigator™ 4.0 or higher and Microsoft Internet Explorer™ 4.0 or higher work well. Use of some parts of the Multiple

Sequence Alignment module requires that a Java "virtual machine" software component be installed. Java software is included with most browser software and is typically installed automatically when browser software is installed. *Note: I have found Netscape's implementation of Java to be more trouble-free than Microsoft's. I have experienced some difficulty working with Java and JavaScript elements in some external web sites when viewed using Microsoft Internet Explorer.*

- Molecular Simulations Inc.'s WebLab ViewerLite molecular graphics software, available for free download by educational users. (http://www.msi.com/solutions/products/weblab/viewer/register/lite/download_lite.html)
- MDL Information Systems' Chemscape Chime molecular graphics browser plug-in, available for free download by educational users. (http://www.mdli.com/support/chime/chimefree.htm)

Notes for the Instructor

Hypertext versions of the Darwin 2000 modules reprinted here are available online at http://www.bioactivesite.com/biocomputing/darwin2000/. The online versions allow the students to access background information, step-by-step instructions, and questions on Darwin 2000 pages, while interacting with, submitting data to, and receiving results from external web sites. To control and guide the student's interaction with all sites, each Darwin 2000 window is divided into two side-by-side frames, one displaying a Darwin 2000 page containing directions, and the other displaying the external web site. Where appropriate, hyperlinks in the Darwin 2000 page displayed in one frame are programmed to produce changes in the external web site displayed in the other frame, thus helping to guide the student through the step-by-step activities.

Each module is written such that it can be used separately from the series of other modules. Thus instructors may choose to use the individual modules that are most applicable to their course (Molecular Graphics in a biochemistry course, BLAST and Multiple Sequence Alignment in a Genetics or Molecular Biology course).

Instructors may choose to use the Darwin 2000 site in several ways, depending on the time available, course objectives, student independence, facilities, etc. In Presentation Mode during lecture periods, the instructor conducts the activities, projecting the computer display for the students to observe and discuss (one-hour for a very brief overview; three-hours for a thorough demonstration with time for student discussion). In Follow-Along Mode in a computer lab setting, the instructor demonstrates and the students then follow each step of the activity on their own computers individually or in small groups, using each module's guided activities focusing on hemoglobin as a target of study (three to four hours, in one or more sessions). In an Investigational Laboratory Mode, students perform the guided and challenge activities at their own pace and discuss or submit answers to questions (approximately two hours per module). In Homework Mode, specific tasks and questions can be assigned to be performed by students independently in the computer lab or at their personal computers as a follow-up to lecture or lab topics. In Research Project Mode, students conducting independent research can conduct the guided and challenge activities to learn the necessary biocomputing skills, then conduct an exploratory activity and write a research report on a molecule of their choice. Because it is entirely online and uses only free resources, Darwin 2000 is available as a solution to the requirement for science research experiences within distributed learning curricula.

If you would like information concerning the scientific issues addressed within each of the questions posed to the students during the activities, contact me directly at rickhershberger@bioactivesite.com. I'd be happy to correspond with you and assist in your use of the Darwin 2000 site.

Student Outline: Overview

Understanding the process of evolution requires an appreciation for its underlying molecular mechanisms. The model of natural selection Darwin proposed can now be explained by, and is entirely consistent with, what we've learned since Darwin's time about how genes work to encode the structure of proteins, and how the structure and function of proteins determine the physiology, and thus the traits, of any organism.

Table 1.1. What Modern Genetics Can Explain About Natural Selection

What Darwin said.	What Geneticists now know.
Individuals have differences in their traits.	Mutation and sexual reproduction produce new genotypes.
Traits are stably inherited by new generations.	Genes are accurately copied and passed on.
Advantageous traits allow the fittest to survive and reproduce.	Phenotype is determined by the form and function of genes and proteins.
Higher reproduction of the fittest lead to adaptations and new species.	Natural selection keeps "fit" or functional genes and eliminates defective genes.

Using the Darwin 2000 series of online exercises, you may explore how genetic variation and selection for protein function represent the forces driving evolution at the molecular level. In addition, you may explore how the proper function of a protein is related to its three-dimensional structure, which in turn is determined by its sequence of amino acids.

The Darwin 2000 site contains a series of modules, each focusing on the use of a specific biocomputing tool. The interactive exercises contained within each module explain how to use the tool, what kinds of scientific questions can be investigated using the tool, and what concepts in molecular biology, biochemistry, and evolution are illustrated through these investigations. Each module is designed to take the investigator through three stages of discovery: A guided activity provides step-by-step instructions for using the biocomputing tool and conducting an investigation using hemoglobin as a model subject of study. A challenge activity gives the investigator a choice of several other molecules to study and provides specific questions to answer. An exploration activity allows the investigator to select his/her own molecule of interest and pose and answer his/her own unique questions. Although the modules are designed to be used in succession to guide an investigator through a comprehensive sequence analysis project, much as professional scientists would begin with newly obtained sequence data and analyze it systematically, each module may be used individually as a tutorial on a specific biocomputing tool.

In the **Finding Sequences in GenBank** module (http://www.bioactivesite.com/biocomputing/genbank/), you will use the National Center for Biotechnology Information's GenBank genetic sequence database (http://www.ncbi.nlm.nih.gov/Entrez) to access DNA and protein sequence data and read a sequence database record to obtain specific information on mutations, gene structure, functional sites within proteins, etc. This activity will give you an understanding of the importance of the sequence of monomer units within a polymeric biological macromolecule (DNA or protein).

In the **BLAST Homology Searching** module (http://www.bioactivesite.com/biocomputing/blast/), you will use the National Center for Biotechnology Information's BLAST server (http://www.ncbi.nlm.nih.gov/BLAST) to identify DNA and protein sequences within the GenBank sequence database that share regions of homology with your sequence of interest. This activity will allow you to investigate the concept of sequence homology, or similarity within the sequences of different genes or proteins, and examine how genetic sequences diverge over evolutionary time.

In the **Multiple Sequence Alignment** module (http://www.bioactivesite.com/biocomputing/msa/), you will use the European Bioinformatics Institute's ClustalW MSA server (http://www2.ebi.ac.uk/clustalw/) to identify conserved sequences and consensus sequences among genes and proteins. This activity will allow you to investigate the concepts of sequence conservation and consensus sequences, or the patterns of similarity among sequences of genes or proteins that share common three-dimensional structures or functions. It will also let you explore patterns of evolutionary relatedness among species and their molecular sequences.

In the **Molecular Graphics** module (http://www.bioactivesite.com/biocomputing/graphics/), you will use Molecular Simulations Inc.'s WebLab ViewerLite molecular graphics software (http://www.msi.com/solutions/products/weblab/viewer/register/lite/download_lite.html) and MDL Information Systems' Chemscape Chime molecular graphics web browser plug-in (http://www.mdli.com/support/chime/chimefree.htm) to view 3D molecular models obtained from the Research Collaboratory for Structural Bioinformatics' Protein Data Bank (http://www.rcsb.org/pdb/) and the National Center for Biotechnology Information's Molecular Modeling Database (http://www.ncbi.nlm.nih.gov/Structure/). You will download atomic coordinate files from online structure databases, and study 3D representations of macromolecules with the molecular graphics software. This activity will allow you to investigate the relationships between the structure and function of a protein; i.e. how the sequence of monomer units and their folding in three-dimensional space determine the protein's proper function, and how mutational changes can affect protein function.

Student Outline for Module 1: Finding Sequences in GenBank
http://www.bioactivesite.com/biocomputing/genbank/

Introduction: Accessing DNA and Protein Sequence Data from NCBI's GenBank Database

Much of the world's data on the sequences of DNA and protein molecules are available to the global scientific community via several databases available through the World Wide Web (WWW). During this activity you will access the National Center for Biotechnology Information's GenBank genetic sequence database (http://www.ncbi.nlm.nih.gov/Entrez) to obtain and study DNA and protein sequence entries relating to a particular gene, disease, function, or organism of interest. The sequence data obtained can be used as query data in the next activity in this series, BLAST Homology Searching (http://www.bioactivesite.com/biocomputing/blast/).

Learning Objectives: Technical Skills

- You will use browser software to view, download, and save files from the World Wide Web.
- You will use data entry fields on web page forms to input query data for online database searches.

Learning Objectives: Research Skills

- You will find, view, and save copies of sequence data files relating to a specific gene, disease, function, or organism.
- You will locate key molecular features of DNA and RNA sequences, such as promoters, cis-acting regulatory sequences, start codons, stop codons, splice junctions, mutations, etc.
- You will link sequence data records to related records in genetic, bibliographic, taxonomic, and molecular structure databases.

Learning Objectives: Conceptual Knowledge

- You will describe the value to the scientific community of sharing sequence data through global online databases.
- You will determine the kinds of information about a sequence that are available within a sequence database record.
- You will identify examples of structure-function relationships by relating sequence alterations to changes in gene expression or protein primary structure.
- You will interpret evidence of molecular evolution by comparing gene structures.

Guided Activity: Studying Hemoglobin Genes and Proteins

Database Records: Sequences of Hemoglobin Genes and Proteins

Visit the National Center for Biotechnology Information's GenBank online genetic sequence database (http://www.ncbi.nlm.nih.gov/Entrez) and view the molecular sequences of the following hemoglobin beta chain proteins and genes. We'll begin by linking to a nucleotide sequence record, the human hemoglobin beta gene region database entry (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=455025&form=6&db=n&Dopt=g). This is a particularly well-studied gene, thus the database record is unusually lengthy with plenty of information useful to researchers. Acquaint yourself with the parts of the GenBank database record for a nucleotide sequence.

(1) What types of information are contained in the following parts of the record: Locus, Definition, Keywords, Accession and NID, Source, Organism, Reference(s), Medline, Comments, Features, CDS, /translation, /db_xref, mutation, variation, exon, intron, precursor_RNA, mRNA, Base Count, Sequence?

(2) How many proteins are encoded within this region of human chromosomal DNA? (Hint: Read the Comments section.) How are the proteins different from each other?

(3) What happens when you click on the hyperlinks within the Organism, Medline, and CDS/db_xref tags? After clicking on each link, return by using the browser's "Back" button.

(4) Scroll down through the features table until you reach the entries at nucleotide 62137. This is the region of the genome that encodes the beta chain of hemoglobin. Notice that the range of nucleotides corresponding to the precursor_RNA and CDS are different. Explain why. (Note: Use the CDS for beta-globin, not beta-globin thalassemia.)

(5) What cis-acting sequence region would you expect to find somewhere 5' of nucleotide 62137? (Hint: You probably won't find the answer by looking at the sequence.)

(6) Based solely on the features table without referring to the nucleotide sequence below, what should be the sequence of nucleotides beginning at position 62187? What nucleotides should be just before position 63610? Check your answer now within the sequence.

(7) Is nucleotide 62285 present in the mature messenger RNA? Why?

(8) Add up the ranges of nucleotides encoding beta-globin listed under "CDS/join" at nucleotide 62187. How many amino acids would be encoded by that ribo-nucleotide sequence?

(9) What nucleotide position (number) is mutated giving rise to the sickle-cell form of hemoglobin beta chain? What nucleotide at that position encodes the normal form. What nucleotide encodes the sickle-cell form? What amino acid is present in the normal beta chain of hemoglobin. What amino acid is substituted in the mutant form?

(10) What is the difference between the normal beta-globin protein and the beta-globin thalassemia protein listed with its own CDS entry (just above the normal beta chain CDS entry)? How does this difference in protein sequence come about?

(11) Compare the sizes of the first (exon/number=1), second (exon/number=2), and third (exon/number=3) exons of the beta-globin gene (starting at 62187) with those of the epsilon (19541), G-gamma (34531), A-gamma (39467), and delta (54790) globin genes. Which do you think occurred first during globin gene evolution: the divergence of each of these genes from a common ancestor, or the introduction of introns interrupting each globin gene's coding sequence? Why?

(12) Compare the sizes of the first and second introns of the epsilon, G-gamma, A-gamma, delta, and beta globin genes. What is more free to evolve (change) over time: exon size or intron size? Why?

(13) At the end of the Comments section, immediately before the Features table, are listed the presumed sites where pre-mRNAs are cleaved and polyadenylated. What are the nucleotide sequences at these sites at the end of each gene's precursor RNA? Can you derive a consensus sequence from these cis-acting sequences for polyadenylation sites?

Having studied a particularly complete nucleotide sequence record, let's now examine a protein sequence record. Link to the human hemoglobin beta chain protein database entry (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=2144721&form=6&db=p&Dopt=g).

(14) What types of information are contained in the following parts of the record: Region, Site?

(15) How many amino acids make up the hemoglobin beta chain polypeptide? Does this match the answer from (8) above, derived from reading the nucleotide record? Explain any discrepancy.

(16) From the Features table, write down the amino acids numbers corresponding to the two key histidine residues, one of which binds oxygen and one of which binds the heme iron.

Link to the human sickle-cell hemoglobin beta chain protein database entry (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=2392691&form=6&db=p&Dopt=g).

(17) Visually compare the amino acid sequences of the normal and sickle-cell forms of hemoglobin beta chain. What difference(s) do you observe?

Additional hemoglobin gene and protein sequence records are listed in Table 1.2 in the Appendix.

Database Searching: Using Entrez to Find Records for Specific Proteins

Jump to the National Center for Biotechnology Information's Entrez search engine (http://www.ncbi.nlm.nih.gov/Entrez). This system lets you search for genetic sequence database records matching search parameters you select, allowing you to locate specific entries. NCBI offers searching of databases of nucleotide or protein sequences, 3-dimensional structures of macromolecules, and even the MedLine bibliographic database via PubMed, all from this single site. Click on Nucleotides, limiting our search to nucleotide records.

Let's search for the normal human hemoglobin beta-chain mRNA sequence. Type "hemoglobin" in the text entry box (the horizontal rectangle), then click Search. You will be presented a page allowing you to display the records found and add to or modify your search terms. Notice that this search found over 10,000 records, so we'll have to narrow down our search by adding search terms. Type "beta" in the text block under "Add Term(s) to Query". Click Search and we now have fewer entries (but still over a thousand!), each containing both "hemoglobin" and "beta" in some field of the record. Notice that we can restrict the search for any query term to a specific field, such as gene name or author, by choosing a search field name from

the pull-down menu. Let's try this by typing "human" in the text field, and selecting "Organism" as the search field. Click Search. Since we want an mRNA sequence (not the genomic sequence we viewed earlier), lets add "mRNA" as a search query. Remember to reset the "Search Field" pull-down box back to "All Fields". Click Search. We still have over 700 entries! Let's try adding "complete" as a search term, since we don't want to view any partial sequences. Click Search.

That helped! Now we have nine entries. Let's see if we have one we want to work with. Let's look at the results of our search. Click on the button labeled Retrieve _ Documents. Entrez will return a list (or the first page of a long list) of database records containing the word or words you entered. Scroll down the list until you have found entry AF117710: Homo sapiens hemoglobin beta chain (HBB) mRNA, complete cds. Click on the link "GenBank report" to view this database record.

(18) At what nucleotide number of this mRNA does translation start?

FASTA Format: Having Sequence Data Ready for Further Analysis

A number of sequence analysis web servers you may use require that sequence data be in a specific input format called FASTA. You can get your sequence of interest in FASTA format by clicking on the FASTA button at the top of a GenBank database record, or from a list of Entrez matching records by clicking on the FASTA report hyperlink. At the top of the human hemoglobin beta-chain mRNA sequence record (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=AF117710&form=6&db=s&Dopt=g), click on the FASTA button. When you have the FASTA report displayed, use the cursor to click-and-drag to select the text between the > symbol on the top line and the last letter (nucleotide or amino acid) of sequence data. Select Copy from the File menu. Launch a notepad or text editor application on your computer. (On Windows, click on the Start button, select Programs, then Accessories, then Notepad). On a Mac, paste the FASTA data into a blank document, and save the document for future use. You will use this sequence as query data for the next activity in this series, BLAST Homology Searching (http://www.bioactivesite.com/biocomputing/blast/). You can also save this FASTA sequence by selecting your computer type and "text" from the pull-down menus at the bottom of the FASTA page, and clicking Save. Remember where on the hard drive you place this saved file.

Notice that at the top of the human hemoglobin beta-chain mRNA sequence file there is a button that automatically generates a list of related protein sequence files. Download and save the FASTA report for AAD19696: hemoglobin beta chain [Homo sapiens] (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=4378804&form=6&db=p&Dopt=f). You will use this protein sequence as a query in the next activity, so take a moment to save the FASTA version of this file as you did earlier for the nucleotide version.

Challenge Activity: Finding Database Entries for Other Genes and Proteins

Now that you know how to access and interpret sequence database files, use your searching skills to find the following sequence files. Begin at the National Center for Biotechnology Information's Entrez search page (http://www.ncbi.nlm.nih.gov/Entrez). Try to locate the complete gene, mRNA, or protein sequence, not partial sequences. Protein and mRNA sequences tend to be more useful for the subsequent activities in the Darwin 2000 series than genomic DNA sequences.

Each of these sequences was chosen because they yield interesting results upon subsequent analysis during the BLAST Homology Searching activity (http://www.bioactivesite.com/biocomputing/blast/). Make sure that you save FASTA format versions of these sequence records, so that you have the right input for the BLAST searches and multiple sequence alignments. Also record the accession and NID or PID (nucleotide or protein identification) numbers.

- Cystic fibrosis transmembrane conductance regulator (CFTR): mRNA=gi:4502784; protein=gi:625338. What mutational change is associated with the cystic fibrosis-causing allele of this gene? How is the protein sequence altered?
- HER-2/neu, a growth factor receptor and oncogene involved in breast cancer and a target of the Herceptin breast cancer drug: mRNA=gi:183986; protein=gi:306840
- Estrogen receptor beta (human): mRNA=gi:2911151; protein=gi:2911152
- An AZT-resistant variant of HIV Reverse Transcriptase (the enzyme that copies the HIV viral genome): protein=gi:476920

Exploration Activity: Finding Database Entries for Other Genes and Proteins of Your Choice

Begin your own research project by finding a sequence you are interested in studying during subsequent activities in this series. Begin at the National Center for Biotechnology Information's Entrez search page (http://www.ncbi.nlm.nih.gov/Entrez). Select a gene, protein, disease, or biological process you wish to know more about at the molecular level. Or you may wish to study the evolutionary relatedness of a certain set of species by studying the same gene or protein in each of several species. You will need to identify a single specific gene or protein that you will study, record its accession number, and save copies of the database record's URL (web page address) and its GenBank and FASTA format sequence files.

Student Outline for Module 2: BLAST Homology Searching
http://www.bioactivesite.com/biocomputing/blast/

Introduction: Identifying Homologous Sequences using BLAST Servers

Because the global research community shares its sequence data by making it available online via GenBank and other databases, it is possible to compare any newly discovered DNA or protein sequence with all known sequences using the BLAST, or Basic Local Alignment Search Tool, algorithm. Such homology searching is a routine procedure among molecular biologists, as it is a powerful tool for identifying the function of a molecule. Because the function of a protein depends on its three-dimensional structure, which is dependent on its sequence of amino acids (and the nucleotides that encode them), proteins (or genes) with similar sequence can be presumed to have similar or related functions. During this activity you will access the National Center for Biotechnology Information's BLAST server (http://www.ncbi.nlm.nih.gov/BLAST) to search for sequences homologous to your query sequences. Functions of various molecules, and evolutionary relatedness of different species, will be inferred from the data.

Learning Objectives: Technical Skills

- You will use browser software to view, download, and save files from the World Wide Web.
- You will use data entry fields on web page forms to input query data for online database searches.

Learning Objectives: Research Skills

- You will use a BLAST homology search server to identify known sequences homologous to a query sequence.
- You will locate specific regions of homology between protein sequences, and identify their degree of homology.
- You will infer the function of a protein from the function(s) of its homologues.
- You will infer the evolutionary relatedness of species from sequence homology data.

Learning Objectives: Conceptual Knowledge

- You will describe the purpose and value of a BLAST homology search.
- You will describe how and why sequence homology data can be used to measure evolutionary divergence.
- You will identify examples of structure-function relationships among families of homologous proteins.

Guided Activity: Studying Homology Among Hemoglobin Genes and Proteins

Identifying Homologous Sequences using the Basic Local Alignment Search Tool (BLAST)

The BLAST search algorithm takes your input sequence and compares it to all known genetic sequences (DNA or protein), identifying the known molecules that have similar sequences. BLAST is sometimes referred to as a "one-against-all" homology search algorithm, since the input is a single sequence which is compared against all other known sequences. This is

in contrast to the Multiple Sequence Alignment, or MSA homology search, a "many-against-each-other" search in which a small, defined set of sequences are compared only against each other, not against the entire database.

Link to the NCBI's BLAST Sequence Homology Search server (http://www.ncbi.nlm.nih.gov/BLAST). Select Basic BLAST Search. We will demonstrate a BLAST analysis using the human hemoglobin beta chain (HBB) mRNA as a query sequence. The NID number for the human hemoglobin beta-chain mRNA is 4378803. Enter this number into the large text entry box. Select Accession or GI from the pull-down menu. Because we wish to look for nucleotide sequences homologous to the hemoglobin beta chain mRNA, select the blastn program and the nr (non-redundant) database from the pull-down menus, and click on Submit Query. In the span of several seconds, the supercomputer at NCBI will compare the 562 letters in the beta globin mRNA sequence with over 400,000 database entries, comprising more than 1 billion letters, and identify those entries with patterns of sequence similar to the query. After the server computer conducts your analysis, the homology results are presented three ways: as a graphic, a table of "hits" (identified homologies), and a series of sequence alignments.

The graphic has lines showing the positions and ranges of homologous sequence (the location and length of each line) and the extent of homology (how close the match is, shown as the line's color). Moving your mouse over any line displays the name of the entry in the box above the graphic. (Note: the graphic is not always displayed.)

Under the text "Sequences producing significant alignments" is the table of "hits". Each database entry homologous to the query sequence is presented, beginning at the top with the closest match and ending at the bottom with the weakest homology. Clicking on the code on the left of each line (i.e. emb|X02345|PTGLB1) links you to the GenBank entry for the homologous sequence (useful for checking what animal P. troglodytes is!). Clicking on the number at the right end of the line, the score, jumps you downward within the file to the sequence alignment.

Each homologous sequence is presented as a separate alignment with the query sequence. Only the homologous regions of each molecule's sequence are presented here. The numbers after the words Query and Sbjct indicate the position within each database entry to which the nucleotides on that line correspond. This display is where one can analyze in detail the nucleotide differences between the query and its homologue.

Click on the "664" at the right end of the line "gb|M25079|HUMBETGLA Human sickle cell beta-globin mRNA, complet... 664 0.0". This is the sequence alignment between human normal beta globin mRNA and sickle-cell globin mRNAs.

(1) The normal and sickle-cell variants of the beta globin protein differ in only one amino acid. Why are there so many additional differences between their mRNAs?

(2) What sequence comparisons do you think will typically indicate more variability between similar sequences in species: protein-protein comparisons or DNA-DNA comparisons? Why?

(3) Study the table of hits and look for general patterns. Remember that the hits are ranked from most similar to least similar to the query, which was normal human beta globin mRNA. Rank the following as to which sequences tend to display the greatest similarities.

- different variant alleles of human beta globin (normal, sickle, thalassemias, haplotypes, etc.)
- human beta globin and other human globins (epsilon, G-gamma, A-gamma, delta)
- human beta globin and primate beta globins
- human beta globin and non-primate mammal beta globins

(4) Given your analysis in question (3), in what order do you think the following happened: the divergence of different beta globin genes (epsilon, G-gamma, A-gamma, delta, and beta), the divergent of different human beta globin alleles, or the divergence of mammalian species (humans, primates, rodents, etc.)? Why? What additional homology searches would you conduct to test your hypothesis (i.e. what query sequences would you use for additional BLAST searches)? What would be your predicted results?

Return to the NCBI's BLAST Sequence Homology Search server (http://www.ncbi.nlm.nih.gov/BLAST). Select Advanced BLAST Search. Once again, let's search for homologues of the human beta globin mRNA sequence, but let's restrict our search to human sequences. Enter NID 4378803, select Accession or GI, the blastn program, and the nr (non-redundant) database from the pull-down menus. Scroll down and select Homo sapiens from the Organism pull-down menu. Click on Submit Query. Once the results appear, scroll down to the entry "emb|X02133|HSBGLOP Human beta-type globin pseudogene".

(5) What is a "pseudogene"? How do you think a pseudogene is identified?

Now we will do a BLAST search for homologous protein sequences. Return to the NCBI's BLAST Sequence Homology Search server (http://www.ncbi.nlm.nih.gov/BLAST). Select Basic BLAST Search. To see what sequences are similar to the amino acid sequence of human hemoglobin beta chain protein, enter its accession or PID number, 231023, in the large box. Select Accession or GI from the input data pull-down menu. Because your query is a protein sequence, select the blastp program and the swissprot database from the pull-down menus, and click on Submit Query. If you wish to search a larger protein database, select the nr (non-redundant) database. If you wish to search only among proteins for which three-dimensional structural data is available, select the "pdb" for the Protein Data Bank database.

(6) Over what length of sequence did the computer find homologous sequence between HBB_HUMAN (human) and HBB_GORGO (gorilla)? How many non-identical amino acids were there among the two proteins? At those sites where they differ, how chemically significant is the difference in amino acid(s) between the two proteins.

(7) Over what length of sequence did the computer find homologous sequence between HBB_HUMAN (human) and HBB_CANFA (coyote)? How many non-identical amino acids were there among the two proteins? At those sites where they differ, how chemically significant is the difference in amino acid(s) between the two proteins?

(8) Identify three different species of animal whose hemoglobin beta-chain sequences (not delta chains!) are most similar to the human sequence (High Score greater than 740).

(9) Identify three different species of animal whose hemoglobin beta-chain sequences are mildly similar to the human sequence (High Score between 710 and 690).

(10) Identify three different species of animal whose hemoglobin beta-chain sequences are least similar to the human sequence (High Score less than 650).

(11) How might the process of evolution explain your results from (8), (9), and (10)?

(12) What does this suggest to you about how sequence analysis may be used to study the evolutionary relatedness of species?

A list of hemoglobin gene and protein database record accession numbers is provided in Table 1.2 of the Appendix.

Challenge Activity: Inferring Protein Function From Homology Data

Do your own experiment! You can now conduct BLAST sequence homology searches on one or more of the recommended proteins listed below. When you have identified known proteins with regions of sequence homologous to your protein's sequence, answer the following questions.

(13) What proteins are most homologous to your protein: the same protein from different species, or different proteins from the same species? Recognize that not all proteins in all species have been sequenced. Some proteins, and some species, have been studied far more extensively than others, and thus have many more database entries than less-studied species or proteins.

(14) Do your protein and its homologues share the same extent of homology as the known hemoglobin beta chains?

(15) Is your protein homologous, in whole or in part, to other proteins with different names or functions? What might that tell you about the structure and function of your protein?

(16) How might a BLAST sequence analysis be used to predict the possible cellular functions of newly discovered and sequenced molecules whose functions may not yet be characterized?

Recommended Sequences to Analyze

- Cystic fibrosis transmembrane conductance regulator (CFTR) (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=625338&form=6&db=p&Dopt=g)

(17) Based on the types of proteins that are homologous to the CFTR, what do you think the CFTR protein does? Why?

(18) The CFTR is homologous to several multidrug resistance proteins. These proteins are overexpressed on cancer cells that have become resistant to the cytotoxic effects of various chemotherapy drugs. How do you think multidrug resistance proteins may work?

- HER-2/neu (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=306840&form=6&db=p&Dopt=g), a protein overexpressed on metastatic breast cancer cells and a target of the Herceptin breast cancer drug. (Read an abstract about HER-2/neu.)

(19) What proteins share sequence homology with HER-2/neu? Where on or in a cell would you expect to find the HER-2/neu protein? How does it work? What might bind to HER-2/neu? What enzymatic activity might HER-2/neu have?

(20) The homologous proteins are listed under three major categories of terms. Are these three categories mutually exclusive or do they overlap? Explain how these terms are interrelated.

- Estrogen receptor beta (human) (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=2911152&form=6&db=p&Dopt=g)

(21) What other receptor proteins share sequence homology with the estrogen receptor beta? Why aren't receptors for insulin, growth hormone, or epidermal growth factor listed?

- Endostatin (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=4558170&form=6&db=p&Dopt=g), an inhibitor of angiogenesis (formation of blood vessels) and a candidate for an anti- cancer drug.

(22) Study the sequence alignment of the first homologous sequence. How do you think endostatin is formed?

Exploration Activity: Conduct Your Own Structure-Function or Evolution Investigation

If you conducted the previous activity in this series, Finding Sequences in GenBank (http://www.bioactivesite.com/biocomputing/genbank/), you may have identified the NID or PID number, or copied the FASTA sequence, of a gene or protein you would like to study. If not, return to the Finding Sequences in GenBank page to locate a gene or protein to study. Use the National Center for Biotechnology Information's Entrez search page (http://www.ncbi.nlm.nih.gov/Entrez) to find a protein related to a gene, disease, or biological process you wish to know more about at the molecular level. Or you may wish to study the evolutionary relatedness of a certain set of species by studying the same gene or protein in each of several species. You will need to identify a single specific gene or protein that you will study,

record its accession number, and save copies of the database record's URL (web page address) and its GenBank and FASTA format sequence files.

To begin a new BLAST search, return to the NCBI's BLAST Sequence Homology Search server (http://www.ncbi.nlm.nih.gov/BLAST). Select Basic BLAST Search.

1. Enter your FASTA sequence: Within any gene or protein database record, click the FASTA button, and copy the FASTA sequence from the web page. Paste it into the form, selecting Sequence in FASTA format from the pull-down menu.

or

2. Enter your accession number: Find the number after the PID or NID tag (omit the "g") from the database record. Type it into the form, and select Accession or GI from the pull-down menu.

(23) What do you hypothesize is the function or enzymatic activity of your protein? Why?

(24) Are there specific portions of your protein that are present in other proteins? What function(s) might these shared protein domains serve?

Record the accession numbers of between five and ten of the proteins homologous to your protein of interest. Obtain their FASTA format sequences, creating a text document with all FASTA reports listed one after the other. You will use this as input for the Multiple Sequence Alignment, or MSA (http://www.bioactivesite.com/biocomputing/msa/) type of sequence analysis.

I am very interested in the results of your own investigations regarding other genes and proteins. Please take a moment to e-mail me at rickhershberger@bioactivesite.com to tell me about the gene or protein you studied, and what you found out about it. If you wish, I'd like to share your work on the Darwin 2000 web site.

Student Outline for Module 3: Multiple Sequence Alignment

http://www.bioactivesite.com/biocomputing/msa/

Introduction: Identifying Conserved Sequences using Multiple Sequence Alignment (MSA) Servers

The Multiple Sequence Alignment algorithm ClustalW takes a set of input sequences and aligns them so that the features that are common to the entire set of sequences are highlighted. This serves to identify the nucleotides or amino acids within the sequences that have been conserved during their evolutionary divergence. Natural selection tends to select against changes that result in loss of molecular function, thus conserved residues identified in an MSA are presumed to be important for the structure and function of the molecule. During this activity you will access the European Bioinformatics Institute's ClustalW MSA server (http://www2.ebi.ac.uk/clustalw/) to conduct multiple sequence alignments and identify conserved residues and consensus sequences among families of homologous sequences.

Learning Objectives: Technical Skills

- You will use browser software to view files from the World Wide Web.
- You will use data entry fields on web page forms to input query data for online database searches.

Learning Objectives: Research Skills

- You will obtain and format sequences for use as query data for an MSA analysis.
- You will use a ClustalW server to conduct a Multiple Sequence Alignment.
- You will identify conserved residues within sets of aligned sequences.
- You will identify patterns of evolutionary relatedness among species by interpreting sequence homologies.

Learning Objectives: Conceptual Knowledge

- You will identify examples of structure-function relationships by examining patterns of sequence conservation.
- You will interpret evidence of molecular evolution by comparing protein sequences.
- You will interpret the role of natural selection in constraining the evolutionary divergence of sequences.

Guided Activity: Aligning Multiple Hemoglobin Sequences

The Multiple Sequence Alignment, or MSA, homology search algorithm is sometimes called a "many-against-each-other" search because the input is a small, defined set of sequences which are compared only against each other, not against an entire database. This is in contrast to the BLAST homology search algorithm (http://www.bioactivesite.com/biocomputing/blast/), a "one-against-all" homology search, in which the input is a single sequence that is compared against all other known sequences listed in the database. Thus the starting point for an MSA is a set of sequences that are already presumed to be homologous.

Link to the HbB_FASTAs.txt file (http://www.bioactivesite.com/biocomputing/msa/HbB_FASTAs.txt). This is a collection of hemoglobin beta chain protein sequences from a variety of species, formatted in the FASTA format required by many biocomputing servers. Select all of these sequences and copy the text from this web page.

Note: Full use of the features of the EBI site requires a Java-enabled web browser. If you experience difficulty accomplishing the tasks described below, you may suspect that your web browser is not configured with Java support. Visit www.netscape.com to download Navigator or Communicator, or visit www.microsoft.com to download Internet Explorer. Check during the installation process that Java is installed with the web browser software.

Link to the European Bioinformatics Institute's ClustalW MSA server (http://www2.ebi.ac.uk/clustalw/). You will use a supercomputer in Hinxton, near Cambridge, England for your sequence alignment. Paste your selected hemoglobin sequences, in FASTA format, into the large text entry field. You can add to and delete sequences within the entry field. Click on Run ClustalW to run your MSA analysis.

The results of the MSA are a series of stacked lines, each line representing one of the sequences in the query set. Gaps (dashes) are introduced as necessary to maximize the alignment of identical or similar residues among the set of sequences. This reflects insertion or deletion events during evolution. The ClustalW algorithm takes into account the chemical properties of each amino acid, thus "conservative substitutions" (substitution of an amino acid by another with similar chemical properties: acidic replaces acidic, hydrophobic replaces hydrophobic, etc.) are penalized less by the algorithm than are "non-conservative substitutions" (substitution of a polar amino acid for a hydrophobic one, for instance). This simply reflects how natural selection is more likely to allow amino acid substitutions that have a smaller impact on the chemical or physical properties of that portion of the molecule. At the bottom of each stack of aligned sequences are symbols that summarize the alignment at that position in the sequence. An asterisk denotes a position at which all query sequences have the exact same amino acid. Dots indicate the degree of homology when there is not complete sequence conservation.

For a more graphical view of the alignment, click on the gray button labeled "JalView". (For instruction on all of JalView's features, click on the text link "Use JalView".) A new browser window will open (don't close the old one!). In the JalView window colored boxes group homologous residues. The darker the color, the greater percentage of sequences within the set that have the same residue at that position. Notice that this view lets you quickly spot broad regions of high homology, and note individual sequences that are non-homologous at a given position.

Inferring Structure-Function Relationships by Identifying Conserved Residues and Regions

Identify the amino acid residues that are absolutely conserved among the complete set of sequences (marked by asterisks; there are eight). Switch to the JalView window. Record the position number within the human sequence for each conserved residue by clicking on that residue. The sequence entry and position will appear in the lower left corner of the applet window (Sequence ID: Human (4) Residue=P (37)). This proline at position 37 of the human sequence is the first of the eight conserved residues. Do not use the numbers at the top of the alignment as these numbers indicate positions within the alignment, not within any single sequence entry. Note that amino acid number 1 in the human sequence is position number 10 within the alignment. Also note that each gap introduced in the human sequence further changes the numbering of residues within the human sequence relative to the number scale at the top of the alignment.

(1) What amino acid residues within the human beta globin protein sequence appear to be conserved among vertebrate (and one invertebrate) hemoglobins? List amino acid and sequence position for each. How might these individual amino acids be important for protein structure or function?

(2) Are there any broad regions of consistent homology (as opposed to absolute conservation) among the sequences? How might these series of amino acids be important for protein structure or function?

Now view the GenBank sequence database entry for the human hemoglobin beta chain protein (www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=2144721&form=6&db=p&Dopt=g).

(3) Are any of the amino acids or regions you identified as conserved known to be important for hemoglobin structure or function as indicated in the Features table of the database record?

Inferring Evolutionary Relationships by Interpreting Patterns of Sequence Homology

Examine the sequence alignments for patterns of sequence homology that let you group some sequences together as distinguished from one or more other sequence(s). JalView is probably best for this task. For instance, the gap at alignment positions 97-99 distinguishes lampreys and sea cucumbers from the remaining species. The valine at position 148 is present in all mammals, but is different in bird, fish, or other sequences. The valine at position 149 groups the gull sequence with the mammal sequences. Study a number of positions for homologies that allow you to group some species away from others, or allow you to create other groupings (both trout and rockcod have glycine at position 157).

(4) Which mammal is most closely related to humans? Which mammal is most distantly related to humans?

(5) Which are more closely related to mammals: fish or birds? What data would you wish to have, or what sequences would you wish to analyze, to confirm or refute your hypothesis?

(6) Do lampreys show more homology with trout and rockcod (jawed, bony fishes, thus vertebrates) or with sea cucumber (an echinoderm, thus an invertebrate)? What does this tell you about lampreys? How are lampreys different than trout (refer to a Zoology text, if you wish).

(7) Draw an evolutionary tree with main branches separating groups of species, then smaller branches ending with individual species. Try to make the lengths and separation of branches roughly consistent with the sequence divergence you observe in the alignment.

Challenge Activity: Aligning Sequences Within Other Protein Families

Cystic Fibrosis Transmembrane Conductance Regulator

Open the file CFTR_FASTAs.txt

(http://www.bioactivesite.com/biocomputing/msa/CFTR_FASTAs.txt). This is a select set of FASTA format sequences that were collected from a BLAST homology search using the cystic fibrosis transmembrane conductance regulator (CFTR) as a query sequence. The set of homologous proteins include:

- CFTR (http://www.ncbi.nlm.nih.gov/htbinpost/Entrez/query?uid=625338&form=6&db=p&Dopt=g), or cystic fibrosis transmembrane conductance regulator, a cAMP-dependent chloride channel. Patients homozygous for mutant alleles in the CFTR gene (thus they make only defective CFTR

protein) misregulate chlorine ion transport and hence do not regulate water levels correctly, resulting in salty sweat and thick, viscous mucus in the lungs.

- MDR (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=2506118&form=6&db=p&Dopt=g), or multidrug resistance proteins. These proteins tend to be overexpressed on cancer cells that are resistant to multiple chemotherapeutic drugs.
- STE6 (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=134967&form=6&db=p&Dopt=g), the yeast protein responsible for secreting the mating type a factor.
- YCFI (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=731347&form=6&db=p&Dopt=g), a protein involved in heavy metal (cadmium) resistance in yeast.
- YOR1 (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=1730876&form=6&db=p&Dopt=g), an ATP-dependent permease involved in yeast resistance to the drug oligomycin.

Select all of the text in this file and select your browser's Copy command. Return to the European Bioinformatics Institute's ClustalW MSA server (http://www2.ebi.ac.uk/clustalw/). Paste the CFTR-like FASTA sequences into the large text entry block, and run ClustalW. When the results are displayed, click on the JalView button.

(8) Take a quick look at the sequence database entries for these proteins using the links above. Examine the Keywords, Titles, and Comments fields. What properties, functions, cellular locations, or enzymatic activities might you suspect that these proteins share in common? What might make each class of proteins unique?

(9) Can you locate within the sequence alignment any regions that might correspond to distinct or shared functions or properties?

(10) Can you locate any conserved regions that correspond to sites listed in the features tables of the database records?

HER-2/neu

Open the file HER2_FASTAs.txt

(http://www.bioactivesite.com/biocomputing/msa/HER2_FASTAs.txt). This is a select set of FASTA format sequences that were collected from a BLAST homology search using the HER-2/neu (http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=306840&form=6&db=p&Dopt=g) sequence as a query sequence. HER-2 is a protein overexpressed on metastatic breast cancer cells and a target of the Herceptin breast cancer drug. (Read an abstract about HER-2/neu.). The set of homologous proteins include:

- gi|306840 (M11730) HER2 receptor
- sp|P04626|ERB2_HUMAN ERBB-2 RECEPTOR PROTEIN-TYROSINE KINASE PRECURSOR
- sp|P21860|ERB3_HUMAN ERBB-3 RECEPTOR PROTEIN-TYROSINE KINASE PRECURSOR
- sp|Q15303|ERB4_HUMAN ERBB-4 RECEPTOR PROTEIN-TYROSINE KINASE PRECURSOR
- sp|P00533|EGFR_HUMAN EPIDERMAL GROWTH FACTOR RECEPTOR PRECURSOR
- sp|Q05397|FAK1_HUMAN FOCAL ADHESION KINASE 1 (FADK 1)
- sp|P00519|ABL1_HUMAN PROTO-ONCOGENE TYROSINE-PROTEIN KINASE ABL
- sp|P12931|SRC_HUMAN PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC
- sp|P11362|FGR1_HUMAN BASIC FIBROBLAST GROWTH FACTOR RECEPTOR 1
- sp|P06213|INSR_HUMAN INSULIN RECEPTOR PRECURSOR (IR)
- sp|P17948|VGR1_HUMAN VASCULAR ENDOTHELIAL GROWTH FACTOR RECEPTOR
- sp|P16234|PGDS_HUMAN ALPHA PLATELET-DERIVED GROWTH FACTOR RECEPTOR
- sp|P23443|KS6_HUMAN RIBOSOMAL PROTEIN S6 KINASE (S6K)
- sp|P17252|KPCA_HUMAN PROTEIN KINASE C, ALPHA TYPE
- sp|Q14012|KCC1_HUMAN CALCIUM/CALMODULIN-DEPENDENT PROTEIN KINASE
- sp|P24941|CDK2_HUMAN CELL DIVISION PROTEIN KINASE 2
- sp|Q00526|CDK3_HUMAN CELL DIVISION PROTEIN KINASE 3
- sp|P11802|CDK4_HUMAN CELL DIVISION PROTEIN KINASE 4
- sp|Q00535|CDK5_HUMAN CELL DIVISION PROTEIN KINASE 5
- sp|P53355|DAPK_HUMAN DEATH-ASSOCIATED PROTEIN KINASE 1
- sp|Q16539|MP38_HUMAN MITOGEN-ACTIVATED PROTEIN KINASE P38
- sp|P37173|TGR2_HUMAN TGF-BETA RECEPTOR TYPE II PRECURSOR
- sp|P25098|ARK1_HUMAN BETA-ADRENERGIC RECEPTOR KINASE 1

Select all of the text in the HER2_FASTAs.txt file and select your browser's Copy command. Return to the European Bioinformatics Institute's ClustalW MSA server (http://www2.ebi.ac.uk/clustalw/). Paste the HER2-like FASTA sequences into the large text entry block, and Run ClustalW. When the results are displayed, click on the JalView button.

(11) Protein kinases add a phosphate group to another protein, often modifying the activity or function of the target protein. Can you identify sequence regions that you would hypothesize correspond to the protein kinase domain of this set of proteins? Check your hypothesis by clicking on the link for one of the specific proteins above, and checking the Features table in the database record.

(12) Are there domains present among any set of these proteins that might correspond to other specific functions: Cell cycle regulation, hormone or growth factor binding, oncogenicity?

Exploration Activity: Conducting Your Own Structure-Function or Evolution Investigation

Conduct your own MSA analysis of a set of protein or DNA sequences of your choice. To locate a sequence to research, use the Finding Sequences in GenBank page (http://www.bioactivesite.com/biocomputing/genbank/). Once you decide on a DNA or protein to research, use the BLAST Homology Searching page (http://www.bioactivesite.com/biocomputing/blast/) to identify other sequences homologous to your query sequence. Using a text editor (WordPad, Notepad, SimpleText) copy the FASTA

format sequence for each of your family of homologous sequences. Copy the group of FASTA sequences into the European Bioinformatics Institute's ClustalW MSA server page (http://www2.ebi.ac.uk/clustalw/). Click on Run ClustalW. When the results are displayed, click on the JalView button.

Research investigations on protein structure-function often focus on several different sequences drawn from closely related species, such as our HER2 example using only human sequences or the CFTR example using a select set of proteins with common functions. Alternatively, you can investigate evolutionary relationships by analyzing a sequence common to many species (hemoglobins, cytochrome, ribosomal RNAs, etc.) using GenBank database entries for the identical molecule in a wide variety of species.

I am very interested in the results of your own investigations regarding other genes and proteins. Please take a moment to e-mail me at rickhershberger@bioactivesite.com to tell me about the gene or protein you studied, and what you found out about it. If you wish, I'd like to share your work on the Darwin 2000 web site.

Student Outline for Module 4: Molecular Graphics

http://www.bioactivesite.com/biocomputing/graphics/

Introduction: Studying the 3D Structure of Macromolecules using Molecular Graphics Software

The proper function of a biological macromolecule is dependent on its three-dimensional shape. Its shape is determined in a large part by its sequence of subunits, but also by its proper folding, particularly in the case of proteins. In this exercise you will be able to view and manipulate 3D representations of macromolecules, exploring the roles sequence and structure play in macromolecular function.

Learning Objectives: Technical Skills

- You will use browser software to view files from the World Wide Web.
- You will use data entry fields on web page forms to input query data for online database searches.

Learning Objectives: Research Skills

- You will obtain atomic coordinate files from online structure databases.
- You will use molecular graphics software to view 3D models of macromolecules.
- You will identify and highlight specific atoms, amino acids, regions, and structures within a macromolecular model.

Learning Objectives: Conceptual Knowledge

- You will identify relationships between the structure and function of macromolecules by examining 3D structures.

Downloading and Using Molecular Graphics Software

You will use two forms of molecular graphics software in this module: Chemscape Chime and WebLab Viewer Lite. Chemscape Chime functions as a browser plug-in. This means that the software file is placed in the "Plug-Ins" folder within the folder containing your web browser program (typically Netscape Navigator or Microsoft Internet Explorer). Link to MDL Information Systems' Chemscape Chime download page (http://www.mdli.com/support/chime/chimefree.htm). Follow the links and instructions for downloading and installing the software. The license allows free use by students and educators, so e-mail them to thank them! On Macintosh computers, typically the downloaded installer program is decompressed and launched automatically, thus installing the software in the appropriate directory on your hard drive. On Windows computers, select "Run this program from its current location" to have the installation take place automatically.

After installing the Chime plug-in, link to http://www.bioactivesite.com/biocomputing/graphics/. Spinning models of a DNA molecule and a hemoglobin protein molecule should be displayed. If not, it may be necessary to exit your web browser software and relaunch it to get the browser to recognize the plug-in, then return to this page. Please be patient as the models load. The 1hbs.pdb file is 770K, so it takes a minute or so to download and be displayed.

You may now explore the use of the Chime plug-in. Place your cursor over one of the models, then click the right mouse button (Windows) or click and hold the mouse button still (Macintosh). A menu appears, allowing you to stop rotation (uncheck "Rotate") and to change how the model is presented. Experiment with this menu, manipulating both the DNA and protein models. Try different Display styles (Spacefill vs. Ball & Stick vs. Wireframe; Ribbons, Strands, etc.). To move the molecule, click on either of the models (Windows: use the left mouse button) and drag your mouse. View the DNA molecule from the side and top.

A standard color code is used to identify different atoms:

gray = carbon
blue = nitrogen
orange = phosphorus
orange = iron
white = hydrogen (when displayed)
red = oxygen
yellow = sulfur

(1) Looking down along the long axis of the DNA model, where are the negatively charged phosphate groups positioned: within the inside of the helix or around the outside?

(2) Looking at the side of the DNA molecule, what can you say about how rings of the nitrogenous bases are positioned relative to each other? What is the chemical explanation for the way these rings are positioned?

On the hemoglobin model, right-click, then in the "Color" menu select "Chain". Right-click again, move to the "Select" menu, choose "Chain" and select the B chain. Then select "Display", "Spacefill", and "Van der Waals". Repeat these steps (Select -> Chain -> X; Display -> Spacefill -> Van der Waals) for the D, F, and H chains. You have now highlighted the beta-globin polypeptides. The remaining polypeptides are alpha-globin chains.

(3) The sickle-cell hemoglobin structure you are viewing represents a clump of two hemoglobin tetramers. Normal hemoglobin is a single tetrameric complex. How many alpha and beta chains are part of a normal hemoglobin tetramer? What two polypeptides interact at the point where two sickle-cell tetramers clump together: two alpha chains, two beta chains, or one of each?

Because it functions as a browser plug-in, Chime can be used to view molecular structure files embedded within a web page, as you have just done. Another molecular graphics software package is WebLab ViewerLite. This software works differently than Chime, in that it is a stand-alone application, not a browser plug-in. Thus WebLab ViewerLite can be used independently of a browser, and it opens atomic coordinate files in its own application window. Link to Molecular Simulations Inc.'s WebLab ViewerLite download page (http://www.msi.com/solutions/products/weblab/viewer/register/lite/download_lite.html) and follow the instructions for registering, downloading, and installing the software (most of which is done automatically, or at least with plenty of prompting). Please e-mail and thank them for making WebLab ViewerLite available free to academic users!

After WebLab ViewerLite is downloaded and installed, you will do a quick test to confirm that the software is working. For this you will need an atomic coordinates file to open. A quick way to obtain a structure file is to right-click (Windows) or click-and-hold (Macintosh) on the DNA model on the http://www.bioactivesite.com/biocomputing/graphics/ page. Select "File" and "Save Molecule As..." from the menu. Keep the default filename, 265d.pdb. (Always keep the ".pdb" extension on the filename; it's what tells the software what kind of file it's dealing with!) Save the file to your desktop. Launch WebLab Viewer from your Start menu (Windows) or find the icon for the program and launch it (Macintosh). Under the "File" menu, select "Open", choose "Brookhaven PDB Files (.pdb)" from the "Files of Type" pull-down menu, and open the file you saved to the desktop, 265d.pdb.

Feel free to click-and-drag with the mouse to move the model into different orientations, or to experiment with different display settings available within the menus. Once you learn to search for and download a Protein Data Bank (PDB) atomic coordinate file from an online database, you will return to WebLab ViewerLite to examine in detail the sickle-cell form of hemoglobin.

Guided Activity: Studying the 3D Structure of Hemoglobin

Obtaining Atomic Coordinate Files from Online Structure Databases

Just as the entire scientific community posts sequence data to one or more global databases (GenBank and others) when new sequences are obtained, researchers share atomic coordinate files through online databases such as the Protein Data Bank as new structures are determined. When scientists determine the three-dimensional structure of molecules using X-ray crystallography and nuclear magnetic resonance (NMR) techniques, atomic coordinate files are created that list the position of each atom within the molecule along the x, y, and z axes. These long text files list each atom, its x-y-z coordinates, the other atoms it is bound to, and what chemical subunit (nucleotides, amino acids) and polymeric chain to which it belongs. Molecular graphics software such as Chemscape Chime and WebLab ViewerLite read these atomic

coordinate files (ending in the ".pdb" or ".ent" filename extension) and display models that can be edited and manipulated.

Note: Full use of the features of the NCBI Structure Group and Protein Data Bank sites requires a Java-enabled web browser. If you experience difficulty accomplishing the tasks described below, you may suspect that your web browser is not configured with Java support. Visit www.netscape.com to download Navigator or Communicator, or visit www.microsoft.com to download Internet Explorer. Check during the installation process that Java is installed with the web browser software.

Just as genetic sequence databases provide tools to search the entire database for entries containing specific keywords, organisms, diseases, etc., structure databases are searchable. Link to the National Center for Biotechnology Information's Entrez search engine (http://www.ncbi.nlm.nih.gov/Entrez/). Entrez is the search engine you used during the GenBank module to find gene and protein sequences, so it should look familiar. (It also cross-links between sequence, structure, and bibliographic databases, so it is a wonderful biocomputing starting point!) Click on 3D Structures, then type human sickle hemoglobin in the text entry box. Click Search. Then click on Retrieve 3 Documents to see the individual database records that match your query. We want the 1HBS entry, so click on the Structure Summary link under 1HBS.

To download the atomic coordinate file in PDB format, scroll down to the Options section and select the Save File radio button. Click on the Rasmol (PDB) radio button under Viewers. Then click on View/Save Structure. Rename the file 1hbs.pdb and save it to your desktop. This is a rather large file (>700K) containing >9000 atoms, so it may take a few minutes to download. If you wish to have Chime view the file within the right browser window, change the Options setting to Launch Viewer, then click the View/Save Structure button. Click the browser's Back button to return to the list of entries matching your original search keywords. Repeat this download procedure to obtain a copy of the 6hbw.pdb file.

Access to information on using 3D structures is also available through the NCBI Structure Group (http://www.ncbi.nlm.nih.gov/Structure/). This page links to the same database as Entrez.

Another site that can be used to obtain atomic coordinates files is the SearchLite page at the Research Collaboratory for Structural Bioinformatics' (RCSB) Protein Data Bank (http://www.rcsb.org/pdb/searchlite.html/). Enter human and sickle and hemoglobin into the text entry block. This search engine supports Boolean searching, so if you want an entry that contains all of your keywords, remember to connect your keywords with "and". Click on the Search button. You will be presented with a list of database records that match your search criteria. Entries 1HBS and 2HBS differ in their resolution, or how accurately the position of each atom has been measured within the crystal. Notice that entry 6HBW contains a genetically engineered form of the beta-chain, with the sickle-cell mutation (beta6 Glu->Val) replaced by another modification (beta6 Glu->Trp).

To download the 1HBS entry, click on Explore to the right of the entry name. Click on Download/Display File. Scroll down to the area under "Download the Structure File". Click on the X under "PDB" and to the right of "none". Accept the default filename, which should be

1HBS.pdb, and save the document to your desktop. Since you downloaded a file named 1hbs.pdb to the desktop a moment ago, either replace that file with this one, or cancel the download and leave the original copy in place. Use the browser's Back button several times to return to the original list of entries matching your search (the page entitled "Query Result Browser"). Repeat this download procedure to obtain the 6HDW file, if you didn't download it already.

The Protein Data Bank also provides a Java interface to call up Chime scripts displaying customized variations of the model. Click on the View Structure link, then click on the Chime link. Use the square buttons to change the way the molecule is displayed. You can turn off Ligand to hide the heme groups. Right-click (Windows) or click-and-hold (Macintosh) on the model to open the Chime menu to change the model display manually.

If for any reason you have difficulty getting these files from NCBI or PDB (perhaps because your browser is not configured for Java), open the 1hbs.pdb file directly into a new browser window using the URLs http://www.bioactivesite.com/biocomputing/graphics/1hbs.pdb. Since your browser is now configured to open the Chemscape Chime plug-in to view PDB files, right-click (Windows) or click-and-hold (Macintosh) on the model and select Save Molecule As under the File menu.

Viewing, Manipulating, and Studying 3D Models of Hemoglobin

You now have downloaded atomic coordinate files describing the three-dimensional structure of the sickle-cell form of hemoglobin (1hbs.pdb), and a genetically engineered form (6hbw.pdb) in which the sickle-cell mutation (beta6 glu->val, or beta E6V) is replaced by another mutation (beta6 glu->trp, or beta E6W). Let's first examine the overall structure of the hemoglobin molecule, and get used to working with WebLab ViewerLite. Launch WebLab ViewerLite (Use the Start menu in Windows.), then under the File menu choose Open and select the 1hbs.pdb file you downloaded and saved earlier. Another way to open this file is to select File, Open Location, and then type http://www.bioactivesite.com/biocomputing/graphics/1hbs.pdb to open the copy of the file on this web site. Resize the WebLab ViewerLite window so that it covers the frame to the right, so that you can read this text in the background when the WebLab ViewerLite window is in the front. Remember that you can jump between open programs using the TaskBar on the bottom of the Windows screen or the Application Icon/Menu at the upper right corner of the Macintosh screen.

WebLab ViewerLite lets you select individual chains, subunits, and atoms and change how they are displayed. This is done by toggling back and forth between the window displaying the model (3D Window) and a window listing all of the chains, subunits, and atoms (Hierarchy Window). Under the Window menu, select New Hierarchy Window. Chains are labeled with an icon that looks like some chain links. Amino acids have an icon that looks like a line diagram of an amino acid. At the left side of each icon is a box with a plus or minus sign. Clicking in this box expands or collapses the items contained within that group. Collapse all eight chains to save yourself a lot of scrolling on the next steps. A, C, E, and G chains are alpha-globin polypeptides, and B, D, F, and H chains are beta-globin polypeptides.

Let's remove the solvent water molecules from the display. Click on the chain icon next to the word Water. It's now highlighted yellow. Click your delete key to remove these molecules from the file. (Re-collapse the chains, because they expand automatically after items are deleted.) At the bottom of the Window menu you will see 1hbs.pdb:1 (this is the 3D Window) and 1hbs.pdb:2 (this is the Hierarchy Window). Select 1hbs.pdb:1 to toggle back to the graphical display. You see that the model no longer shows the red dots representing the oxygens in the water molecules. You should realize that models automatically open with the hydrogens hidden. If you would want to display the hydrogens, select Hydrogens, Add under the Tools menu. But for the time being let's leave the hydrogens hidden. Return to the Hierarchy window.

Let's now highlight the different polypeptide chains that form the entire hemoglobin complex. Click on the icon next to the top chain A (the polypeptide, not the heme!) to highlight it yellow. Now select the View, Display Style menu, or instead click on the icon of the water molecule (one red ball with two white balls). This is the menu you will use a lot to change how different portions of any model are displayed. This will help you see specific chains, subunits, or atoms within the overall model. Let's leave the atoms in the A chain displayed in Line format, but let's display the entire A chain in yellow. Under the Coloring section click Custom, then click on the color swatch and select the yellow block. Click OK twice, then toggle to the 3D window. Because the A chain is still selected in the Hierarchy menu, its atoms will be marked with yellow selection blocks in the 3D window. To delete these atoms, click somewhere in the black space around the molecule. Just as you can select and deselect atoms or groups in the hierarchy window, you can manually select or deselect atoms or regions in the 3D window. When you've deselected the A chain, it should be in Line display format, but entirely in yellow. Toggle back to the Hierarchy menu, and click on the top C chain icon. Hold the Control button and click on the top E and G chain icons. Holding the Control key while clicking adds or subtracts the clicked item from your selection, so you should now have the remaining alpha polypeptides - C, E, and G - selected. Use the Display Style box to color these chains yellow, just as you did with the A chain. Now deselect the C, E, and G chains, select the four beta polypeptides - B, D, F, and H - and color them blue. Each of the eight polypeptides contains a heme group, or porphyrin ring that binds oxygen and contains the iron atom. Select the heme groups (the bottom of each pair of groups labeled with each letter) and color each of the heme groups red. When you toggle back to the 3D window you should easily see the polypeptide subunits that comprise each hemoglobin tetramer, and the locations of the heme groups.

(4) What two polypeptides interact at the point where two sickle-cell tetramers clump together: two alpha chains, two beta chains, or one of each?

(5) Where are the heme groups located: near the surfaces of the complex or embedded in the interior of the complex where alpha and beta polypeptides interact? How might their location be explained, considering the function of hemoglobin and the role heme plays in that function?

Now let's study the sickle-cell mutation - the Glu->Val amino acid substitution at position 6 of each beta chain - that causes hemoglobin tetramers to stick to each other, forming large aggregates and distorting the shape of red blood cells. Return to the Hierarchy window, and expand the four beta chains (B, D, F, H) to show the individual amino acids within each chain. Select Val6 within each beta chain, and click on the Display Style icon. Under Display Style click on CPK, under Coloring click on By Element, and click OK.

(6) Before toggling back to the 3D window, predict where in the model the mutant amino acids may be located. Consider that the mutation leads to clumping of tetramers into aggregates. Where might an amino acid responsible for that biochemical phenotype by located within the molecule?

Now toggle to the 3D window. The valine amino acid is now displayed in "space filling" style, showing the approximate size of the electron cloud around each of its atoms, and with each atom displayed in its coded color. Changing the style and/or color of specific portions of a model is a handy way of finding points of interest with a large, complicated model. You should now be able to see the mutant amino acid clearly within each of the four beta polypeptides in this model.

(7) At the point where sickle-cell hemoglobin tetramers clump together, what two polypeptides are involved: two alpha, two beta, or one of each?

(8) Was your prediction correct about where you would find the mutant amino acid with the aggregate? Would the hemoglobin tetramers in sickle-cell patients form only clumps of two tetramers (as depicted in the model), or could tetramers clump into larger aggregates?

(9) HbS is the name given to tetramers containing two sickle-cell beta chains and two normal alpha chains. HbA is the normal adult hemoglobin tetramer of two alpha and two beta chains. Predict whether HbS tetramers would stick to HbA tetramers. What hemoglobin aggregates would you predict to be present in patients homozygous or heterozygous for the sickle-cell allele of the beta chain gene?

You've now seen how a view of the three-dimensional structure of a macromolecule can give us insights into why it behaves the way it does. This is of key interest to biochemists and molecular biologists. Now let's return to the theme of evolution, and see how sequence conservation through natural selection points out to us those parts of a molecule that are critical to its function, and are thus invariant across species. First, let's focus only on the B chain, or a single beta polypeptide. In the Hierarchy window, select every chain except the B chain and its heme (labeled B). Press the Delete key. Under the Edit menu choose Select All to select every atom in the structure. Select Display Style - Line and Color By Element. To make the heme group stand out a bit, select it and display it as Ball and Stick, with the Custom Color of yellow. Now is a good time to click on the Fit to Screen icon, the one that looks like a black square with four arrows around it. You're getting better at this, so you know how to display the sickle-cell Valine #6 in CPK (space filling) style.

If you are working through the Darwin 2000 modules in succession, during the Multiple Sequence Alignment (MSA) module (http://www.bioactivesite.com/biocomputing/msa/) you aligned the amino acid sequences of hemoglobin beta-chains from mammal, bird, fish, and even invertebrate species. Link to the results of an MSA analysis conducted on 9 beta-chain sequences from human to echinoderm (sea cucumber). You should see that several amino acid residues are conserved across this entire evolutionary span. The conserved amino acids (with their position number within the PDB file, not the aligned sequence) are:

P, Pro, proline #36
Q, Gln, glutamine #39
F, Phe, phenylalanine #42
H, His, histidine #63
L, leu, leucine #88
H, His, histidine #92
V, Val, valine #98
K, Lys, lysine #132

Select these 8 amino acids (Use the Control key to select multiple amino acids.) and display each in CPK style. While they are still selected, under the Tools menu select Labels, Add, then under Object choose Amino Acid and click OK. Now view your masterpiece in the 3D window!

(10) Why might evolution and natural selection have maintained some or all of these amino acids? How might they be important for the function of the hemoglobin molecule?

One of the WebLab ViewerLite button bars has buttons with icons showing a yellow dot and arrows of different styles. These icons change the effect clicking-and-dragging has on the molecule. Use these buttons to zoom in or out, or to move the model from side to side without rotating it. Get a close-up view of the region of the molecule immediately adjacent to the heme group. In the Hierarchy window, change every amino acid in the beta chain except for His63 and His92 back to Line style. Keep the heme in Ball-and-Stick style, except for its iron (Fe) atom, which you should now make CPK and Color by Element.

Link to the GenBank database entry for human hemoglobin beta chain protein (www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=2144721&form=6&db=p&Dopt=g). Look through the features table in the database record and find the functionally important amino acids. Note: the numbering in the GenBank record is one digit off from the numbering in the PDB record, because the N-terminal methionine amino acid is removed from the polypeptide post-translationally.

(11) Which of the functionally important amino acids listed in the database record did you predict to be important because of their sequence conservation through evolution?

(12) What purpose do the two conserved histidines serve in the hemoglobin beta chain? Might one find conserved histidines in the alpha polypeptide? Why? Where might you expect histidines to be located in the alpha chain? Would they be conserved across species? Design your own investigation to test your hypothesis!

Challenge Activity: Studying the 3D Structure of Other Molecules

Cystic Fibrosis Transmembrane Conductance Regulator
HER-2/neu

During the GenBank, BLAST, and MSA modules you may have studied the cystic fibrosis transmembrane receptor or the HER-2/neu breast cancer gene. Locate these or related structures within the Protein Data Bank and relate any sequence conservation you determined through MSA to the structure you see using the molecular graphics software.

Exploration Activity: Conducting Your Own Structure-Function or Evolution Investigation

If you have worked through the Darwin 2000 series of exercises, you are now well versed in sequence and structure databases (sources of raw data), sequence homology search and alignment servers (ways to analyze the raw data) and molecular graphics software. You are ready to choose any gene or protein and study its structural biochemistry, molecular biology, and evolution using these bioinformatics tools. If you wish to end with a structural modeling study like you have just done within this module, you should begin your research project in the structure database. This is because the sequence of many macromolecules are available, but far fewer have their structure determined. Pick a molecule that interests you out of the PDB structure database using a keyword search, then get its matching sequence from GenBank and begin your BLAST and MSA analyses. If you wish, return to each module (listed below) and follow the step-by-step instructions using your sequence as the query. Happy researching!

Now go out there and discover something!

I am very interested in the results of your own investigations regarding other genes and proteins. Please take a moment to e-mail me at rickhershberger@bioactivesite.com to tell me about the gene or protein you studied, and what you found out about it. If you wish, I'd like to share your work on this web site.

Acknowledgements

The author thanks Dr. Craig Johnson, Associate Professor of Chemistry, Carlow College for insights into the biochemistry of hemoglobins and the use of, and choices between, molecular modeling software.

Bibliography

Hardison, Ross. 1999. The Evolution of Hemoglobin. American Scientist, 87 (2) (http://www.amsci.org/amsci/articles/99articles/Hardison.html)
Globin Gene Server (http://globin.cse.psu.edu/)
Bridges, K. R. An Overview of Hemoglobin (http://www-rics.bwh.harvard.edu/sickle/hemoglobin.html)
Hemoglobin (molecular graphics) (http://www.umass.edu/microbio/chime/hemoglob/2frmcont.htm)

Appendix: Some Hemoglobin Database Records

The following is a list of hemoglobin gene and protein sequences for which 3D structures are available. Gene and protein sequence records can be accessed in GenBank by searching for the Accession or NID number (gi: ####### or sp:X#####) and 3D molecular structures can be obtained from the Protein Data Bank by searching for its accession number (pdb: #XXX).

Table 1.2: Some Hemoglobin Sequence Entries with GenBank and PDB Accession Numbers

Species	Gene	Protein
Primate Mammals		
Human *Homo sapiens*	normal - gi:455025	normal form gi:231023; db:4HHB sickle form- gi:2392691; pdb:1HBS
Lowland Gorilla *Gorilla gorilla*		sp:P02024
Long-haired Spider Monkey *Ateles belzebuth*		sp:P02034
Non-primate Mammals		
Horse *Equus caballus*		gi:230633; pdb:2MHB
White-tailed Deer *Odocoileus virginianus*		sickle form - gi:229975; pdb:1HDS
Cow *Bos taurus*		gi:576143; pdb:1HDA
Pig *Sus scrofa*		gi:809283; pdb:2PGH
Birds		
Black-Headed Gull *Larus ridibundus*		sp:P08261
Fish		
Rainbow Trout *Oncorhynchus mykiss*		gi:1942655; pdb:1OUT
Emerald Rockcod *Pagothenia bernacchii*		gi:1065040; pdb:1HBH
Sea Lamprey *Petromyzon marinus*		gi:230607; pdb:2LHB
Echinoderms (Invertebrates)		
Sea Cucumber *Caudina arenicola*		gi:576151; pdb:1HLB