Tina O'Grady

Bioinformatics

A brief review of resources on the Web

Information sources in the biological sciences have grown far beyond journals and books. Sequencing projects and new molecular biology tools are creating vast amounts of new data and information in new formats. The manipulation, storage, and analysis of these new data constitute the growing field of bioinformatics. This field is diverse and interdisciplinary, drawing necessary expertise from information science and computer science, as well as biology. With our keen interest in the organization, provision, and management of information, it is natural for librarians to be intrigued by bioinformatics and play a role in bioinformatics teaching and research.

From finding out which molecules make up the sequence of a tiny stretch of DNA to teasing out the entire family tree of life on earth, the tools listed below span a wide swath of research efforts. There are many more tools and resources available than can be listed here, with new ones appearing every day; this guide attempts to give a representative sample of the kinds of tools that are available and refers the reader to other sources to look for more.

Background information

A basic understanding of nucleic acids (DNA and RNA), proteins, and their relationship to one another is necessary for an understanding of bioinformatics tools.

A Science Primer. This site from the National Center for Biotechnology Information (NCBI) at the National Library of Medicine provides basic information about the science behind bioinformatics tools available from NCBI and elsewhere. It defines terms while introducing specific tools. *Access:* http://www.ncbi.nlm.nih.gov/About/primer.

Genome News Network. From the J. Craig Venter Institute, this is an online news magazine about genomics. It also features a basic introduction to the science behind bioinformatics and genomics called "What's a Genome?" as well as "A Quick Guide to Sequenced Genomes," which provides brief information about organisms whose genomes have been sequenced, including photographs, literature references, and links to the sequencing organization. *Access:* http://www.genomenewsnetwork.org.

McGraw Hill's AccessScience. Available by subscription, this site has articles that provide much useful background information. The encyclopedia articles "Deoxyribonucleic Acid (DNA)," "Ribonucleic Acid (RNA)," and "Protein" are helpful in understanding key concepts. *Access:*http://www.accessscience.com.

Finding sequences

Probably the most broadly used bioinformatics tools are databases consisting of gene and protein sequences. An unknown sequence can be identified if the same sequence is found in a database, or relatives (genes, proteins, or organisms) of known sequences can be found. Sequence databases can generally be searched by keyword or by inputting a query sequence.

Tina O'Grady is science librarian at the University of New Orleans, e-mail: tmogrady@gmail.com

National Center for Biotechnology Information (NCBI). NCBI houses GenBank, the international depository of publicly available DNA sequences. Subsets of GenBank, including databases such as Nucleotide, CoreNucleotide and Protein, can be searched by keyword or sequence. Keyword queries search sequence annotations (including the GenBank accession numbers researchers are required to include in scienti c papers), while the BLAST search algorithms allow searchers to use sequence data as a query to retrieve sequences of signi cant similarity. *Access:* http://www.ncbi.nlm.nih.gov.

UniProt. The Universal Protein Resource (UniProt) contains information aggregated from several major protein databases. Included are protein sequences with extensive annotation, including names, functions, structures, literature references, and more. UniProt can be searched by keyword or using BLAST search algorithms for sequence searches. *Access:* http://www.ebi.ac.uk/uniprot.

Sequence analysis

Much information is contained in the sequences of DNA and proteins. The order of nucleotides in DNA determines the order of amino acids in proteins, and the amino acids in proteins determine shape, size, function, and other characteristics. Many tools exist to use sequence data to help determine physicochemical properties of proteins (such as weight, solubility, electrical charge, etc.) and to nd telltale patterns of DNA or protein sequence that correspond with particular functions.

ExPASy (Expert Protein Analysis System) Proteomics Server. This site contains many tools to analyze protein sequence data obtained from sequence databases or experiments. Some tools included are *ProtParam,* which calculates molecular weight, amino acid composition, atomic charge, and other parameters; *Radar* and *REP,* which search for repeated sections of amino acid sequence in the protein; and *Protein Colourer,* which allows you to color-code your protein sequence. *Access:* http://www.expasy.org, mirror site at http://ca.expasy.org/tools.

WebMOTIFS. This site allows you to choose from many different algorithms to search DNA sequences for biologically signi cant patterns, or motifs. *Access:* http://fraenkel.mit.edu/webmotifs.

Sequence alignment

Aligning sequences, whether two (pairwise) or more (multiple), is the rst step in comparing them. Once they are lined up, amino acids and nucleotides that have been deleted, mutated, or repeated in different sequences become apparent. This information can be used to determine how sequences are related, and can be the raw data for further research, such as determining evolutionary relationships.

Blast2Sequences. This site uses the BLAST algorithm (also used for performing sequence searches of GenBank and other sequence databases) to align two sequences. Pairwise alignments can be performed on nucleic acids or proteins, and algorithm parameters can be changed to customize the alignment. *Access:* http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi.

ClustalW. ClustalW, from the European Bioinformatics Institute, can be downloaded or used on the Web to align multiple DNA or protein sequences. It uses a hierarchical alignment method; it aligns the two most similar sequences rst, then adds sequences to the alignment in order of similarity. *Access:* http://www.ebi.ac.uk/Tools/clustalw2.

T-Coffee. From the Swiss Institute of Bioinformatics, T-Coffee is a package of related programs for alignment of multiple DNA or protein sequences. Its alignment method is rigorous as opposed to hierarchical; it seeks the mathematically optimal alignment. *Access:* http://www.tcoffee.org.

Phylogenetics

Phylogenetics is the study of the family tree of life, using biological information to infer evolutionary relationships. It is related to taxonomy, the practice of naming and classifying organisms into species, genus, family, etc. Evolutionary relationships have long been inferred using the physical characteristics of organisms and with the growing availability of sequence data, many tools have been developed to analyze how mutations and deletions in sequences may reveal evolutionary relationships.

PHYLIP. PHYLIP is a widely used package of programs available for free download. Nucleotide or protein sequences, or other data, are input to create evolutionary trees using several different computational methods. Extensive documentation and help files are available online. *Access:* http://evolution.genetics.washington.edu/phylip.html.

Phylogeny Programs. This site, maintained by Joe Felsenstein at the University of Washington, is an extensive list of phylogeny tools; both downloadable software and Web servers. It can be browsed by several criteria, including data type and phylogenetic method. *Access:* http://evolution.genetics.washington.edu/phylip/software.html.

The Tree of Life Web Project. This site is the product of hundreds of contributors from around the world. It allows browsing up and down the tree of life, with information about branches at different levels and how they relate to one another. Information about both living and extinct species is included, and annotation varies, but generally includes pictures and literature references. *Access:* http://tolweb.org.

Structure prediction

A protein's sequence determines its three-dimensional structure, but how? Visualizing a protein's structure in the lab with X-ray crystallography or NMR spectroscopy is considered the most accurate method, but this is not always practical or even possible. Obtaining a protein's sequence is much easier, and so a great deal of research has focused on elucidating structure from sequence data. This can work quite well for proteins that have sequences similar to proteins for which structures are known. Prediction methods are much more complex, and less reliable, for proteins that are very different from known structures.

JPred. This site is produced and maintained by the University of Dundee, and can be used for protein sequences for which no related structure is known. Its artificial neural networks threads a sequence through the standard protein structures of helices, sheets and coils, and calculate how well the sequence fits those shapes. The result is a prediction, for each amino acid in the sequence, of the structure it forms. Predicted solubility information is also included. *Access:* http://www.compbio.dundee.ac.uk/~www-jpred.

LiveBench. On the BioInfoBank metaserver, LiveBench evaluates different protein prediction servers by comparing predicted structures with actual structures that have been newly determined in the laboratory. Results are collated and posted weekly. The site also contains a metaserver that will accept sequence submissions and submit them to several structure prediction servers. *Access:* http://meta.bioinfo.pl/livebench.pl.

SWISS-MODEL. From the Swiss Institute of Bioinformatics, SWISS-MODEL uses homology modeling to predict protein structures. Users submit a protein sequence, and SWISS-MODEL searches sequence databases to find a similar sequence with a known structure. The similar sequence acts as a template and the program generates a structure prediction, which is sent to the user by e-mail. http://swissmodel.expasy.org/SWISS-MODEL.html.

Function prediction

Like its structure, a protein's function is also largely determined by its amino acid sequence. Protein function can only be confirmed by laboratory experimentation and observation, but can often be inferred based on sequence and structure; proteins with very similar sequences or structures tend to have similar functions. Certain characteristic sequence patterns, or

motifs, tend to form particular structures and perform particular functions. Motifs and larger functional regions, called domains, appear in proteins in different numbers and combinations, leading to a diverse range of functions for proteins.

GeneTrail. GeneTrail, from the Universit t des Saarlandes, can analyze DNA or protein sequences in sets (e.g., genomic data sets). Groups of sequences are analyzed for the presence and numbers of genes belonging to a particular functional category, as de ned by the Gene Ontology or other controlled vocabularies. *Access:* http://genetrail.bioinf.uni-sb.de.

InterProScan. This tool from the European Bioinformatics Institute searches the InterPro collection of databases. These databases contain records for functional domains of well-characterized proteins, and match the user-submitted sequence with similar sequences found in these domains. Database records contain information about the detected domains, including function, proteins in which it is present, literature references, and controlled vocabulary from the Gene Ontology to aid subsequent information seeking. *Access:* http://www.ebi.ac.uk/Tools/InterProScan/.

Genome/model organism databases

In many cases, the research communities focusing on particular organisms have developed databases speci c to their organisms. The features of each database vary with the needs of the community and the particulars of the organism, but may include literature references, advanced genome mapping features, new sequences not yet submitted to GenBank, and more.

Biocurator.org. This Web site is maintained by biocurators, the scientists who create and maintain biological databases, frequently built around model organisms. This page contains information for and about biocurators, and links to many model organism databases. *Access:* http://biocurator.org.

FlyBase and **ZFIN.** These are two examples of model organism databases, for the fruit y and zebra sh, respectively. Both contain records for gene and protein sequences for their organisms, and also integrate literature references, images, anatomical information and many other tools deemed useful to their research communities. *Access:* http:// ybase.bio.indiana.edu and http://z n.org.

UCSC Genome Bioinformatics. This site contains the genomes of many organisms, which can be searched, browsed, and compared. *Access:* http://genome.ucsc.edu.

Information and tutorials

Bioinformatics is a very swiftly moving eld, and tools and resources are updated, appear, and disappear frequently. The level of user support and documentation varies widely, as well. Below are some links to assist with resource use and discovery.

Bioinformatics Links Directory. From the University of British Columbia, Bioinformatics Links Directory contains hundreds of well-organized, annotated links to information pages, databases, programs, and other tools. All of the links from *Nucleic Acids Research* Web server issues since 2003 are included, as well as many other resources recommended by users. *Access:* http://bioinformatics.ca/links_directory.

Bioinformatics Tutorial Series (BITS). BITS is being produced by MIT s Engineering and Science Libraries and Harvard s Countway Library. They are brief streaming video tutorials for specific bioinformatics tools. Several are on the Web now, with more to come. *Access:* http://countway.harvard.edu/video_tutorials and http://libraries.mit.edu/video.

Molecular Biology and Genomics SIG of the Medical Library Association. This site provides links to various bioinformatics tools, and resources for teaching them. New

(continues on page 421)

("Bioinformatics" continued from page 407)
developments in the eld are also discussed on their electronic list. *Access:* http://medicine.wustl.edu/%7Emolbio/index.html.

Nucleic Acids Research. This open access journal publishes yearly database and Web server issues, with short articles about hundreds of bioinformatics tools. These issues provide useful overviews of new and updated tools, and the articles sometimes provide more information about a tool or resource than can be found elsewhere on the Web (including the resource s own site). *Access:* http://nar.oxfordjournals.org.