

## Public Databases

### *Retrieving and Manipulating Sequences for Beginners*

Neil Woodford

#### Summary

This chapter outlines the basic requirements for finding and exploring sequences of interest in public databases, such as GenBank. As such, it is not aimed at experienced sequencers, for whom this will be “second nature,” but at the many clinical bacteriologists who rarely have need of DNA sequences in their usual work, and who would like to develop their interest in what can appear to be a daunting area. The topics discussed include finding and retrieving sequences from GenBank, identifying homologous sequences using BLAST searches, resources for accessing microbial genomes, and the Protein Data Bank. Finally, recommendations are made for useful software (freeware) and online sequence manipulation resources.

**Key Words:** Bioinformatics; GenBank database; BLAST search; homology; Protein Data Bank; TIGR; Sanger Institute; freeware; online resources.

#### 1. Introduction

*Bioinformatics* is a current buzzword, but it means different things to different people. Generally, it is the discipline where maximum biological information is derived from a nucleotide or amino acid sequence. Very often, this requires science to be performed *in silico*—a phrase that emphasizes the fact that many molecular biologists spend increasing amounts of their time in front of a computer screen, generating hypotheses that can subsequently be tested and (hopefully) confirmed in the laboratory. In my very first undergraduate microbiology practical (only 20 yr ago), I was told that there was no such thing as theoretical microbiology. However, this is becoming increasingly questionable and, in a few years time, it may no longer be true! At a fundamental level,

From: *Methods in Molecular Biology*, vol. 266: *Genomics, Proteomics, and Clinical Bacteriology: Methods and Reviews*

Edited by: N. Woodford and A. Johnson © Humana Press Inc., Totowa, NJ

“bioinformaticians” write computer programs to perform searches, align sequences, and so on. Such tasks obviously require detailed understanding of the underlying mathematical algorithms and statistical approaches (computational biology), and are beyond the scope of this chapter (or this book, for that matter). However, far more people use bioinformatics at a more practical level, by accessing and utilizing the vast number of resources available to analyze and derive useful information from nucleic acid or peptide sequences. These sequences may have been generated in their own research, or may have been downloaded from public databases. This chapter provides a beginner’s guide to this second, more pragmatic approach.

## 2. Retrieving DNA Sequences From Public Databases

There are several databases in which DNA sequences may be deposited (e.g., GenBank, EMBL) and there is regular and frequent exchange of new sequences between them. In consequence, researchers deposit sequences with only one database and, similarly, anyone searching for a sequence needs to interrogate only one database. Many bacteriologists use PubMed (a medical literature search tool) regularly and, conveniently, the GenBank search engine (which can be accessed via <http://www.ncbi.nlm.nih.gov/Entrez/index.html>) shares the same front end (**Fig. 1A**). Hence, the GenBank sequence depository is suitable for many beginners because it provides a familiar screen layout. Ultimately, the choice of which database to search is a matter of personal preference.

The easiest way of finding a specific sequence in a database is by using its unique accession number; this is constant in all sequence databases. Most journals require published sequences to be accessible in a public database, unless they are patented. Each paper should include the unique database accession numbers for all deposited sequences (usually found at the end of the **Methods** or **Discussion** sections). Alternatively, if an accession number is not available, searches may be performed on key words or author names. The GenBank search engine will return either a single sequence record (if an accession number was used) (**Fig. 1B**), or a numbered list of sequence records (if key words were used in the search). In the returned listings, the accession numbers are hyperlinks that can be used to open the actual deposition records for each sequence. Within each record there are further hyperlinks to relevant PubMed entries (i.e., corresponding publications) and, if the sequence contains multiple open-reading frames (ORFs), to individual genes and to their translated products. For example, if a search is performed on accession number M97297, the search engine will return the record for the prototype VanA glycopeptide resistance transposon, Tn1546 of *Enterococcus faecium* (**Fig. 1B**) (1). This resistance element is 10,851 bp long and contains nine ORFs. The GenBank record

**A**

**B**

Fig. 1. (A) The front end of the GenBank search engine (which also provides access to PubMed and other resources). To obtain sequences, select “nucleotide” in the “Search” dropdown menu, enter the accession number (or author name, key words, and so on) in the “for” field and click on “Go.” (B) An example of a retrieved entry. The accession number acts as a hyperlink to open the full GenBank record. The check box may be used to store the file on the clipboard or to save it to a file on a local drive. In author or key word searches, multiple sequences are usually retrieved.

gives nucleotide positions of each ORF and the peptide sequence of the predicted product, and hyperlinks to the sequences of individual genes and peptides. The full 10,851-bp sequence appears at the end of the record (Fig. 2).

A search may fail to retrieve a required sequence for several reasons. First, it may not yet have been released for public access. Authors have to deposit

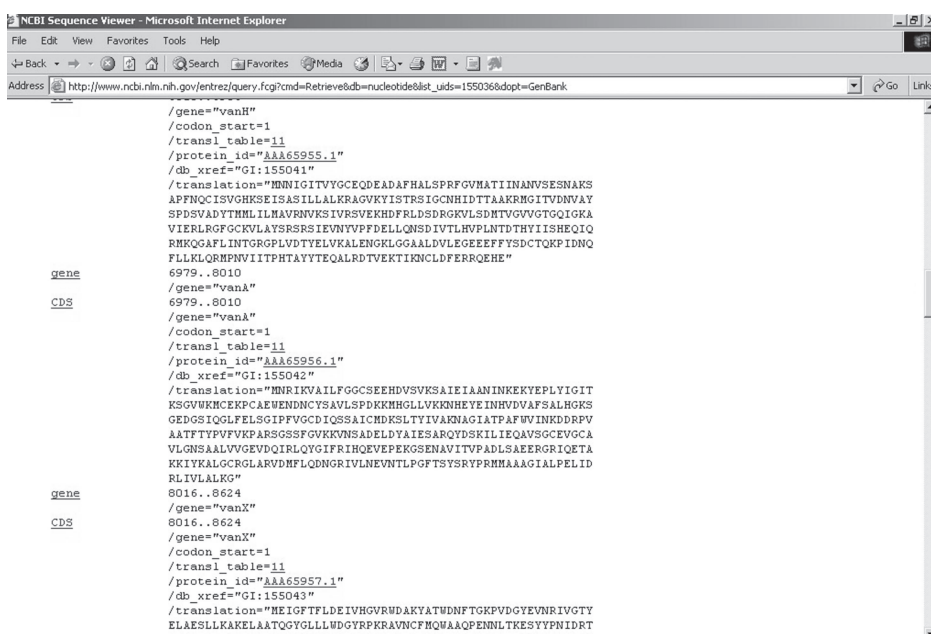


Fig. 2. Part of the GenBank record for accession number M97297 (the 10,851-bp transposon Tn1546 of *E. faecium*). After details of the sequenced organism, PubMed citation etc., there is a full list of ORFs within a record. Nucleotide positions within the sequence are given (e.g., the *vanA* gene spans from nt 6979 to 8010 in the screenshot shown). Adjacent to this information are hyperlinks to the sequence of each specific gene and coding sequence (CDS), which avoids the need to “count and crop” it from the full 10,851-bp sequence at the end of the entry.

sequences with databases prior to submitting a paper so that they can quote accession numbers in manuscripts. However, such unpublished sequences may be withheld from the publicly accessible part of the database until after publication. Once work has been published, the authors have to notify the GenBank administrators so that release may proceed. Authors may forget to do this. Also, the huge explosion in sequencing has caused a corresponding increase in database submissions. Database administrative staff cannot keep pace with requests for sequence modification and release, so there is often a delay (usually only a few days) before instructions are acted upon. Second, there may be a typographical error in the published accession number. To test for this, it is often advisable to try a keyword or author search (when performing such searches in GenBank, remember to keep “Nucleotide” in the search field, so that you query GenBank rather than PubMed). If problems accessing a published sequence

persist, inform the GenBank administrator (gb-admin@ncbi.nlm.nih.gov), quoting the accession number and journal reference in which it appears.

### **2.1. Saving a Sequence**

When the required sequence has been retrieved from the database it may be viewed online or, for subsequent manipulation, a copy may be saved to a local drive. When using the GenBank database, the procedure for saving sequence files follows exactly the same steps as for saving literature citations from PubMed. Thus, to save a sequence, (1) check the box next to its accession number, (2) choose “GenBank” or “FASTA” format in the dropdown “Display” option box, (3) choose “File” in the dropdown “Send to” option box, and (4) click “Send to.” Follow the standard Windows instructions to save to the required place on a local drive.

For more complex analyses, it is necessary to retrieve and save multiple sequences, often as a single file. Usually subsequent manipulation is intended (e.g., multiple alignments), and saving the sequences in FASTA (rather than GenBank) format is most appropriate. If the required sequences are retrieved in a single search (e.g., on a particular author or key word) they can be saved simply by selecting the box next to each required accession number. However, saving multiple sequences from separate searches is not a problem (again, it’s the same as saving different author citations from PubMed): (1) Perform the first search, select the boxes for all required sequences, and store the sequence on the search engine’s clipboard (this is achieved by choosing “Clipboard” in the dropdown “Send to” option box, and then clicking “Send to”); (2) perform the next search, check the boxes for all required sequences, and store them on the search engine’s clipboard; (3) repeat the previous step as many times as necessary (the clipboard will store up to 500 items); (4) when all searches are complete, click “Clipboard” to reveal all of your stored sequences; (5) select the boxes again, choose “FASTA” format in the dropdown “Display” option box; (6) choose “File” in the dropdown “Send to” option box; and (7) click “Send to.” Follow standard Windows instructions to save all to a single file in the required place on a local drive.

### **2.2. Viewing a Downloaded Sequence File**

Even with no bioinformatics software on a PC, saved GenBank files can be opened using WordPad (a standard Windows program) or a similar text viewer. Opening the saved file as text in Microsoft Word or a similar processing program is not recommended, as such packages will try to reformat the file on closing. If the sequence was saved in GenBank format, WordPad shows the entire record (including genes, peptides, Medline citations and so on), although the hyperlinks present in the original record no longer function. If the file was

saved in FASTA format (the usual option for many subsequent manipulations), WordPad reveals a leading descriptive line (in the format ">sequence title") followed by the nucleotide sequence only.

Although using WordPad is arguably one of the best ways of simply viewing sequence files, any manipulation of the sequences necessitates the use of bioinformatics software. It is not essential, but it makes life a whole lot simpler (an analogy would be creating a Web page using a specific Web design package, compared with using just a text editor and knowledge of HTML; not impossible, but unnecessarily time-consuming). Many commercial bioinformatics packages are available, but there are many freeware programs available and they are often just as good. Indeed, the freeware can frequently do everything one is likely to need day to day. The exception to this seems to be the absence of primer design freeware for local installation; there are DOS-based programs, but very little that is Windows based, as far as I know (*see Subheading 6.8.* for an online solution to this problem). BioEdit (*see Subheading 6.2.*) is an excellent, easy to use program. It is simple to install and goes a long way toward providing all required bioinformatics applications in a single package.

### 3. Basic Local Alignment Search Tool (BLAST)

BLAST (2) is an online program that searches for significant homology between query sequences and those in the GenBank database. These searches can be initiated from within some bioinformatics packages, such as BioEdit (*see Subheading 6.2.*), or directly from the main BLAST site (<http://www.ncbi.nlm.nih.gov/BLAST/>). The BLAST tool is actually a group of programs that can be used to compare nucleotide sequences against other nucleotide sequences (blastn), peptide sequences against peptide sequences (blastp), or peptide sequences against translated nucleotide sequences (tblastn). Using the tool is simply a matter of cutting and pasting the required query sequence into the appropriate box, selecting the most appropriate program, and clicking on "Blast."

If searching for homologs in diverse species, where genes may not show high levels of nucleotide similarity, the tblastn option can be particularly useful. This is because proteins with similar functions often retain conserved domains and motifs (i.e., groups of conserved amino acids) (*see Chapter 3*) even though variation in codon usage, for example, means that the corresponding genes are less similar. These characteristic protein family "signatures" can serve to identify potential homologs. Hence, unless alleles of particular genes are sought, it is usually good practice first to perform a tblastn search with a peptide sequence; the results are often more fruitful than those of a direct blastn search with the nucleotide sequence.

The results of BLAST searches are usually available in a short time (seconds to a few minutes), but can be returned by e-mail at particularly busy periods. The results indicate the degree of similarity between your sequence and those in the database. The results include score ( $S$ ) and probability ( $E$ ) values. The score for each sequence is determined by the algorithms used in the program. The  $E$  value indicates how likely an equal or greater score is to have arisen by chance (i.e., is the homology likely to be a real or chance occurrence). Results are returned in order of decreasing  $E$  value, so the most significant hits appear first.

The list of sequences returned in a BLAST report will consist of true-positives and false-positives. True-positives are sequences that share an evolutionary origin with the query sequence (are true homologs), or are examples of evolutionary convergence. False-positives show sequence similarity owing to chance. There are no mathematical algorithms that can distinguish absolutely between these possibilities. Negatives are homologous sequences that were not flagged by the search (3). All algorithms that assess sequence similarity must reach a compromise between two conflicting goals: (1) to produce a list that includes all significant hits (i.e., avoiding negatives); (2) to avoid generating long lists of meaningless chance homologies (i.e., avoiding false-positives).

Thus, in a BLAST report, only some of the returned sequences are genuine homologs. For many searches, common sense can indicate which results should be pursued and which discarded. For example, a returned sequence showing 100% identity to the query sequence over 15 nt is unlikely to be meaningful if that query sequence was 800-bp long!  $E$  values  $<10^{-50}$  are commonly generated by BLAST searches and indicate likely close relationships; these result from good homology and a long overlap between the subject and query sequences.

The returned BLAST report includes a graphical representation of homologous sequences (**Fig. 3A**) in which each line is hyperlinked to the alignment between the query and a particular retrieved sequence. The GenBank accession numbers of the retrieved sequences are given next to each alignment (**Fig. 3B**), and these are hyperlinked to the corresponding GenBank entries. There is also an option to select sequence boxes and save retrieved sequence files.

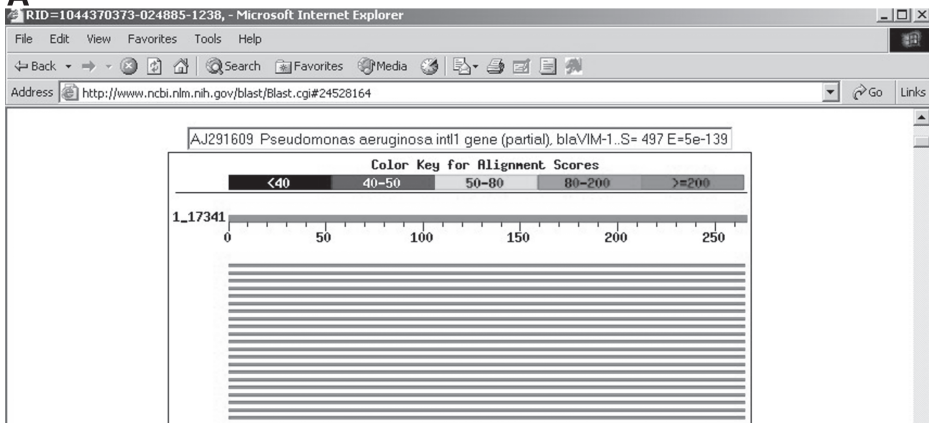
#### 4. Sources of Microbial Genome Sequences

Many chapters in this volume describe the extensive and ever-increasing number of complete microbial genome sequences. Many of these are freely available, and there are appropriate resources to facilitate their use. Indeed, much of the drive toward improved bioinformatics software comes from the large genome-sequencing corporations.

The Institute for Genomic Research (TIGR) has compiled and maintains a Comprehensive Microbial Resource (CMR) tool that allows researchers to



A



B

Sequences producing significant alignments:

Score E  
(bits) Value

gi 5420397 emb Y18050.1 PAE18050	Pseudomonas aeruginosa int...	497	e-139
gi 26514696 gb AY152821.1	Escherichia coli class I integro...	497	e-139
gi 24528164 emb AJ439689.1 PPU439689	Pseudomonas putida par...	497	e-139
gi 13508543 emb AJ278514.1 XY278514	Achromobacter xylosoxi...	497	e-139
gi 24421057 emb AJ291609.1 PAE291609	Pseudomonas aeruginosa...	497	e-139
gi 24636971 gb AF317511.1	Pseudomonas aeruginosa integron ...	497	e-139
gi 22758830 gb AY135661.1	Pseudomonas aeruginosa class I i...	495	e-138
gi 22347613 gb AF531419.1	Pseudomonas aeruginosa class I i...	495	e-138
gi 24415594 gb AY144612.1	Klebsiella pneumoniae metallo-be...	484	e-135
gi 14017420 gb AY029772.1	Pseudomonas aeruginosa isolate Y...	457	e-127
gi 7288608 gb AF227532.1 AF227532	Pseudomonas aeruginosa in...	457	e-127
gi 11559847 gb AF317681.1 AF317681	Serratia marcescens meta...	457	e-127
gi 10717156 gb AF305559.1 AF305559	Enterobacter cloacae met...	457	e-127
gi 9965931 gb AF291700.1 AF291700	Pseudomonas aeruginosa me...	457	e-127
gi 9937490 gb AF291438.1 AF291438	Pseudomonas putida metall...	457	e-127

C

Alignments

Get selected sequences

Select all

Deselect all

☐ >gi|5420397|emb|Y18050.1|PAE18050 Pseudomonas aeruginosa intI1, blaVIM and aacA4 (partial) genes  
Length = 2310

Score = 497 bits (1280), Expect = e-139  
Identities = 254/266 (95%), Positives = 254/266 (95%)  
Frame = +1

Query: 1 MLKVISSLLVYNTASVMAVASPLAHSGEPGSEYPTVNEIPVGEVRLYQIADGVVSHIATQ 60  
MLKVISSLLVYNTASVMAVASPLAHSGEPGSEYPTVNEIPVGEVRLYQIADGVVSHIATQ  
Sbjct: 1252 MLKVISSLLVYNTASVMAVASPLAHSGEPGSEYPTVNEIPVGEVRLYQIADGVVSHIATQ 1431

Query: 61 SFDGAVYPSNGLIVRDGDELLIDTAWGAKNATAALLAEIEKQIGLPVTRAVSTHFHDDR 120  
SFDGAVYPSNGLIVRDGDELLIDTAWGAKNATAALLAEIEKQIGLPVTRAVSTHFHDDR  
Sbjct: 1432 SFDGAVYPSNGLIVRDGDELLIDTAWGAKNATAALLAEIEKQIGLPVTRAVSTHFHDDR 1611

Query: 121 GGVDVLRAGVATYASPSRRLAEAEAGNEIPTHSLEGLSSGDAVRFGPVELFYPGAHS 180  
GGVDVLRAGVATYASPSRRLAEAEAGNEIPTHSLEGLSSGDAVRFGPVELFYPGAHS  
Sbjct: 1612 GGVDVLRAGVATYASPSRRLAEAEAGNEIPTHSLEGLSSGDAVRFGPVELFYPGAHS 1791

Fig. 3. Three components of a BLAST search report. A tblastn search was performed with the peptide sequence of VIM-1 carbapenemase. (A) Graphic representation of returned sequences. Each line in the graphic is hyperlinked to specific query sequence–retrieved sequence alignment further down the report. As the mouse moves over each line, details of the retrieved sequence appear in the title bar. (B) List of



access all bacterial genome sequences completed to date (<http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl>) (4). The Sanger Institute also has extensive interests in genome sequencing, and its data are also freely available (<http://www.sanger.ac.uk/Projects/Microbes/>). Data from the Sanger Institute and TIGR can be accessed in various ways. Importantly, there are BLAST facilities incorporated within these websites. Rather than searching the whole of GenBank, BLAST searches from the TIGR or Sanger sites allow specified genomes to be interrogated using a query sequence (remember that `tblastn` is often the best first search option; *see Subheading 3.*). The matching contigs (or even complete genomes) can be retrieved and downloaded to local drives for further manipulation (often via FTP).

## 5. The Protein Data Bank (PDB)

The Protein Data Bank (<http://www.rcsb.org/pdb/>) is a publicly accessible repository for proteins of known 3D structure. The structures were determined experimentally, by X-ray crystallography or solution nuclear magnetic resonance. As with DNA sequences in GenBank, each deposited structure has a unique accession number, which typically consists of a number followed by three characters (e.g., 1FOF is the accession number of the class D  $\beta$ -lactamase OXA-10 [PSE-2]; 5). The homepage has search and retrieval facilities using either accession numbers or keywords. As with GenBank, searches performed on accession numbers will return the specific record, while those using key words may return multiple records. Each record has hyperlinks to relevant PubMed citations. The returned PDB files can be downloaded to a local drive and viewed or manipulated with suitable software, such as Deep View (*see Subheading 6.5.*) or RasMol (<http://www.bernstein-plus-sons.com/software/rasmol/>).

For researchers studying protein structure, the value of this database is obvious, but it can be of great use to bacteriologists generally. First, it can be used simply as a resource to discover more about proteins of interest. Second, known protein structures can be used as templates to perform comparative modeling of homologs whose 3D structures have not been determined (6).

## 6. Recommended Downloadable Freeware and Online Resources

The software listed below represents a very small fraction of the available resources. The author has found the selected programs useful, and they will

---

Fig. 3. (*continued*) returned sequences, with hyperlinks to the full GenBank record. (C) An example of a query sequence–retrieved sequence alignment. The accession number acts as a hyperlink to open the full GenBank record. Note also the check box for saving retrieved sequence files.

provide a good starting point for any new bioinformatics user. All URLs were accurate at the time of writing, but they are prone to change. If a link gets broken, keyword searches on the particular program name using a search engine such as Google should identify the new homepage rapidly.

### **6.1. Artemis**

Artemis (7) is a genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence and its six-frame translation. Available from: <http://www.sanger.ac.uk/Software/Artemis/>

### **6.2. BioEdit**

BioEdit is an excellent general bioinformatics program that can handle most of the manipulations that researchers need regularly. It can be used simply to view downloaded GenBank files, or for a variety of more complex procedures. Available from: <http://www.mbio.ncsu.edu/BioEdit/bioedit.html>

### **6.3. BLAST**

BLAST (2) is likely the most commonly used bioinformatics tool. It performs online searches for sequences in GenBank that show significant homology to a query sequence. Cut and paste your sequence (DNA or peptide), check relevant option boxes, and away you go. BLAST is frequently incorporated into genome websites to allow specific genomes to be interrogated. Available at: <http://www.ncbi.nlm.nih.gov/BLAST/>

### **6.4. ClustalX**

ClustalX (8,9) is a Windows interface for the ClustalW multiple sequence alignment program (10,11). It provides an integrated environment for performing multiple sequence and profile alignments and analyzing the results. Alignments can be generated in various formats, and can be imported into GeneDoc (see **Subheading 6.6.**) or PHYLIP (see **Subheading 6.7.**). The tree files may be viewed with TreeView (see **Subheading 6.10.**). Available from: <ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/>

### **6.5. Deep View (Swiss-Pdb Viewer)**

Swiss-PdbViewer (12) is a powerful application with a fairly user-friendly interface that allows analysis of PDB files (3D protein structures), including several proteins at the same time. The proteins can be superimposed in order to deduce structural alignments and compare their active sites, and so on. The program is linked to Swiss-Model (<http://swissmodel.expasy.org/>), an automated homology modeling server. There is an excellent tutorial (<http://>

[www.usm.maine.edu/~rhodes/SPVTut/index.html](http://www.usm.maine.edu/~rhodes/SPVTut/index.html)) to get you started with Deep View. Available from: <http://ca.expasy.org/spdbv/>

### **6.6. GeneDoc**

GeneDoc (**13**) is a full-featured multiple-sequence alignment editor, analyzer, and shading utility for Windows. Excellent for formatting ClustalX-generated alignments into publication-quality figures. Available from: <http://www.psc.edu/biomed/genedoc/>

### **6.7. PHYLIP**

PHYLIP is a powerful package of programs for inferring phylogenies (evolutionary trees). PHYLIP is the most widely distributed phylogeny package, but it does not have a Windows interface and can appear daunting. Trees can be calculated using parsimony, distance matrix, and likelihood methods. Bootstrapping and preparation of consensus trees is also available. An excellent overview of the package is recommended (**14**). Available from: <http://evolution.genetics.washington.edu/phylip.html>

### **6.8. Primer 3**

An online resource for designing PCR and sequencing primers. Cut and paste your DNA sequence, check relevant option boxes, and away you go. Available at: [http://www-genome.wi.mit.edu/cgi-bin/primer/primer3\\_www.cgi](http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi)

### **6.9. SeqVISTA**

SeqVISTA (**15**) enables annotated nucleotide or protein sequences to be viewed graphically. Once installed, it may be launched from your Web browser's toolbar to visualize GenBank records. SeqVISTA aims to display results from diverse sequence analysis tools in an integrated fashion. Available from: <http://zlab.bu.edu/SeqVISTA/>

### **6.10. TreeView**

TreeView (**16**) is a simple program for displaying phylogenies. It can read many different tree file formats (including those generated by ClustalX and PHYLIP) and can produce publication-quality trees simply. Available from: <http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>

### **6.11. WebCutter**

An online resource for identifying restriction endonuclease cutting sites within a sequence. Cut and paste your DNA sequence, check relevant option boxes, and away you go. Available at: <http://www.firstmarket.com/cutter/cut2.html>

## References

1. Arthur, M., Molinas, C., Depardieu, F., and Courvalin, P. (1993) Characterization of Tn1546, a Tn3-related transposon conferring glycopeptide resistance by synthesis of depsipeptide peptidoglycan precursors in *Enterococcus faecium* BM4147. *J. Bacteriol.* **175**, 117–127.
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
3. Pagni, M. and Jongeneel, C. V. (2001) Making sense of score statistics for sequence alignments. *Brief. Bioinform.* **2**, 51–67.
4. Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K., and White, O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res.* **29**, 123–125.
5. Paetzel, M., Danel, F., de Castro, L., Mosimann, S. C., Page, M. G., and Strynadka, N. C. (2000) Crystal structure of the class D  $\beta$ -lactamase OXA-10. *Nat. Struct. Biol.* **7**, 918–925.
6. Mullan, L. J. (2002) Protein 3D structural data—where it is, and why we need it. *Brief. Bioinform.* **3**, 410–412.
7. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945.
8. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998) Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**, 403–405.
9. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882.
10. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
11. Higgins, D. G. and Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244.
12. Guex, N. and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723.
13. Nicholas, K. B., Nicholas, H. B. Jr., and Deerfield, D. W. II. (1997) GeneDoc: analysis and visualization of genetic variation. *EMBNEW News* **4**, 14.
14. Retief, J. D. (2000) Phylogenetic analysis using PHYLIP. *Methods Mol. Biol.* **132**, 243–258.
15. Hu, Z., Frith, M. C., Niu, T., and Weng, Z. (2003) SeqVISTA: a graphical tool for sequence feature visualization and comparison. *BMC. Bioinformatics* **4**, 1.
16. Page, R. D. M. (1996) TREEVIEW: an application to display phylogenetic trees on personal computers. *Comp. Appl. Biosci.* **12**, 357–358.



<http://www.springer.com/978-1-58829-218-6>

Genomics, Proteomics, and Clinical Bacteriology  
Methods and Reviews

Woodford, N.; Johnson, A.P. (Eds.)

2004, X, 398 p., Hardcover

ISBN: 978-1-58829-218-6

A product of Humana Press