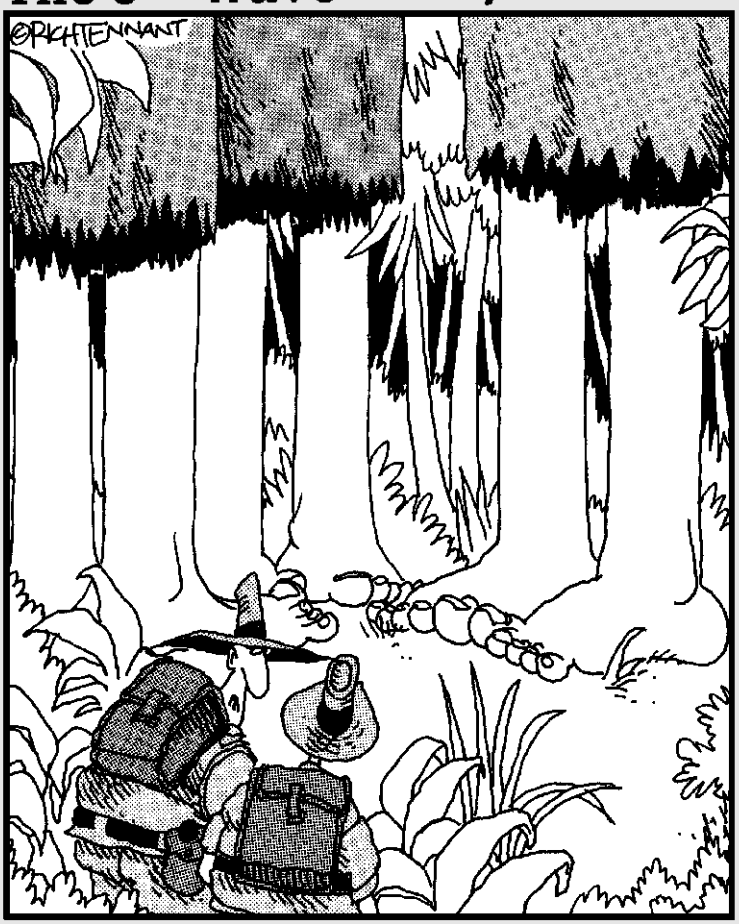
# The 5th Wave

By Rich Tennant




"Okay, it's your turn to ask for a tissue sample for DNA analysis."

## In this part . . .

This part is just what it says on the box: It contains tables of links, resources, and various things useful but difficult to classify. If you haven't yet found what you were looking for anywhere else in this book, you'll probably find it here, or a couple of mouse clicks away from one of the addresses we give you. In this part, we also give you some useful guidelines if you are planning to let your research depend on Internet resources. Better read this before it's too late!

# Commandments for Using Servers

### In This Chapter

- Making sure you don't divulge confidential data
- Ensuring that you're able to reproduce your work
- Using appropriate data
- Making sure you save the right files

*I never did give anybody hell. I just told the truth, and they thought it was hell.*

— Harry S. Truman (1884–1972)

The nicest thing about using Web-based servers is that (from the user's point of view) nothing in bioinformatics is ever a problem on the Internet. Consider: You can run any program you fancy. You don't need to update or maintain any database, and, if the server is broken, you can simply go and use another server. On the Web, you've got it so easy that you might even forget what you're doing and keep pushing buttons until something nice pops onto your screen. It's a bit like what you — and we — do in front of the television!

That's all fine and dandy, but if you need to use these bioinformatics tools for doing some research, you need to be a bit more careful, or this virtual paradise will rapidly turn into real hell. Nothing is more frustrating than a good result you can't reproduce, especially in science.

In this chapter, we give you a few pieces of advice designed to make life easy for you (in the online bioinformatic realm, anyway) if you want to produce results that your work can really depend on.

world. With the exception of a few Java applets that you can use offline, never treat an Internet submission as a safe process.

While the people who run these servers are renowned for their honesty, they rarely have the means to ensure tight security for the data you send them. This means your data may be intercepted during transit.

We personally think that nobody really gives a damn about the sequences you send over the Internet. However, your employers probably think that the risk is real — and if you want to keep your job, you'd better share this fear!

If you need secrecy, you'll just have to install the programs you need on your own computer. No ifs, ands, or buts possible here.

## Remember the Server, the Database, and the Program Version You Used

If you run the same program (such as ClustalW) on two different servers, it does not at all mean you're doing the same thing twice. The setup of the servers — or even the program version — can be different. This is the reason why *"I ran ClustalW over the Internet"* doesn't give other researchers enough information to reproduce your results — and it doesn't sound very professional, either.


To avoid any confusion, write down the name of the server as well as the version of the program that it runs. For instance, you should not expect a server that runs ClustalW 1.77 to return the same results as a server that runs ClustalW 1.81.

If your program (such as BLAST) uses a database (such as Swiss-Prot), you should write down the name and the version of the database you're searching.

Servers also change with time. They get upgraded, improved, and so on. On average, this happens every six months.

on this sequence. Try to keep its ID (identification) and its AC (Accession) numbers, as well as any other information that seems to be relevant. Remember that the ID of an entry can change from one database release to the next, but the AC number never changes.

## Write Down the Program Parameters

When you run a server, you often need to slightly alter the default parameters so the program does exactly what you need it to do. Many programs supplement their output with a file that contains all the parameters you used. This file often has a name that contains the word *log.* For BLAST, these parameters are at the bottom of the result page. If your program outputs such a file, be sure to save it.

As an alternative, you can use the PrntScrn facility of Windows to make a screen shot. (See the next commandment, "Save Your Internet Results the Right Way," for tips on how to do this.) The PrntScrn technique can ease the process of making an exact reproduction of your request.

## Save Your Internet Results the Right Way

Saving Internet results can be trickier than it seems at first glance. This is especially true when you try to save documents that contain images like BLAST results. We cannot give you a foolproof procedure because what works best changes all the time — depending on your Web browser, the way the browser is installed on your computer, and the kind of data you're downloading. When it comes to saving files, Java is a major source of trouble, and often the only solution is to take a screen shot.


Never assume that an important result has been saved correctly; always double-check. The latest version of Internet Explorer and Netscape makes it possible to save complete pages (including pictures).

This puts a copy of your screen onto the Clipboard.

2. **Open Microsoft Paint by choosing Start➪Programs➪Accessories➪Paint.**
3. **Press Ctrl+V and say Yes to the message that proposes that you resize the bitmap.**
4. **Save as usual or print.**

   Save using the `.jpg` format to generate a compressed picture.


Whenever possible, also save a text version of your results! Many programs output a staid text version of the results along with the flashy graphic output. This text version is a much easier way to store and exploit your results. In many cases, it is the only format that other programs can use.

## Use E-Values

E-values (expectation values) tell you how many times a result as good as the one you're looking at could have been reached by chance alone. On its own, this has no special biological meaning. However, the rationale behind the use of E-values in biology is that a good result is something that chance could not have produced without lots of help from nature.


A good E-value is a low E-value. For instance, an E-value of $10^{-32}$ is better than an E-value of $10^{-4}$.

## Make Sure You Can Trust Your Alignments

Just because you have an alignment doesn't mean you can trust it implicitly! Before you build an elaborate theory on your alignment, make sure it's biologically correct. Measure the percentage of identity and look for conserved blocks or important conserved residues. Also, be critical and use the guidelines we give you in Chapter 9!

if you're not sure of your results, try to reproduce them by using different methods. If you can, it probably means that your results are correct and meaningful. If you cannot, it could mean that your results are only a collection of meaningless artifacts. For instance, if you generate a phylogenetic tree with two methods and you get two very different trees, there's a very good chance that none of these trees is, in fact, the correct one. Being able to reproduce the same results with different methods is something biologists sometimes name *robustness,* and it is a notion that applies to most of what we said in this book. (See the identification of transmembrane domains in Chapter 6, for instance.)

When you use different programs, make sure that you double-check with a method that relies on a different principle. For instance, there's no point in comparing an NJ tree from ClustalW with another NJ tree from Phylip. What you want to do is confront your NJ tree with a Fitsch or ML tree.

As an alternative, you can also change some of the key parameters of your program and check how they affect the results you're interested in. In a multiple alignment, for instance, you could change the gap penalties.

## Stay Away from Unpublished Methods!

This isn't to say that you should never try something new, but using well-established methods is the best way to produce results that people trust. Never use an unknown program that's not published and whose principle you don't understand. In brief, stay away from any student's midterm-project pages!

## Databases Are Not Like Good Wine

Databases are not like good wine: They do not age well! Avoid using data that a colleague downloaded last year and left to rot away on a PC.

Any time you need a database or some sequence data, always make sure that you're using fresh data.

you start using some data on a server, make sure that you know whom this data belongs to. If you work in a company, you may have to purchase a license before using it.


Many companies advertise themselves by providing free data or services. This lasts only as long as the company's management types see fit. If your work depends on such resources, an abrupt discontinuation can mean disaster! We aren't saying that you should never use private resources, but you should always know how long those resources are going to last.

## Biting the Bullet at the Right Time

The reason we've dedicated this entire book to online resources is because online resources are easy to use and don't require any kind of complicated setup. They let you concentrate on biology, not on your computer! This is exactly what you need . . . if you use bioinformatics only occasionally. Don't expect to be able to assemble and annotate your own viral or bacterial genome, however, without some serious investment in bioinformatics — and access to knowledgeable people!

If you find that you start relying heavily on some servers, you'll have to bite the bullet and install these programs on your own computer. Installing and maintaining software is easier than it seems — such software, however, works best on Linux or Unix systems. If you start from scratch, learning all you need to know to install and use Linux would probably take you a couple of weeks. Even if your time is tight, you can consider that a safe investment.

How will you know when that time has come? Well, if you find yourself doing 20 BLASTs in a row using an online server, it's a sure sign that you need the independence and flexibility that comes with having the software installed on your own computer!

# Some Useful Bioinformatics Resources

### In This Chapter

- Ten important databases
- Some very important programs
- Our favorite link collections and generic resources
- Finding out about the latest techniques

*"How did you know that I knew you knew?"*

— Typical example of knowledge-based programming

One of the very nice and unique things about bioinformatics is that (almost) everything you need is available online. In this chapter, we show you around the main resources. If you have a fairly good idea of what's behind every link we list here, consider yourself fairly well-trained already in the field of bioinformatics.

Remember that this field is moving rapidly. Being operational in bioinformatics is not so much about knowing things but about knowing where to find the knowledge you need. This book is not exhaustive at all, but it offers you many gateways. Do not hesitate to go further and explore. While on this fascinating journey, keep new things in the corner of your mind until you need them.

## Ten Major Databases

Bioinformatics is about exploring biological information. This information is safely kept in databases. These databases are our modern museums, cyber-reference collections where all the knowledge of an era is carefully stored and classified. Table 15-1 lists ten of these databases. It isn't at all an exhaustive list, but a biased vision that this chapter title limits to just ten items!

| | | |
|---|---|---|
| | db=Nucleotide | |
| Ensembl | www.ensembl.org | Human/mouse genome |
| PubMed | www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed | Literature references |
| NR | www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein | Non-redundant protein sequences |
| Swiss-Prot | www.expasy.ch | Protein sequences |
| InterPro | www.ebi.ac.uk | Protein domains |
| OMIM | www.ncbi.nlm.nih.gov | Genetic diseases |
| Enzymes | www.chem.qmul.ac.uk | Enzymes |
| PDB | www.rcsb.org/pdb/ | Protein structures |
| KEGG | www.genome.ad.jp | Metabolic pathways |

# Ten Major Bioinformatics Software Programs

Data is nothing if you don't know how to analyze it. This is what bioinformatics software programs are made for. You can use these programs to search for your data, to analyze it, and to display it. Table 15-2 lists the five main categories of programs that you may come across in bioinformatics, along with what we believe to be the most characteristic pieces of software in each of these categories. Each of these programs is introduced somewhere else in this book, as indicated in the table.

Table 15-2 also contains the address of an online server for each program. In most cases, you can find alternative addresses in the corresponding chapter.

| | | | |
|---|---|---|---|
| | Entrez | www.ncbi.nih.gov/Entrez | Database search (Chapter 3) |
| | BLAST | www.ncbi.nlm.nih.gov/blast | Homology search (Chapter 7) |
| | DALI | www.ebi.ac.uk/dali | Structure database search (Chapter 11) |
| **Multiple alignment** | ClustalW | www.ebi.ac.uk/clustalw | Multiple sequence alignment (Chapter 9) |
| | MUSCLE | phylogenomics.berkeley.edu/muscle/ | Multiple sequence alignment (Chapter 9) |
| | Tcoffee | www.tcoffee.org | Multiple sequence alignment (Chapter 9) |
| **Prediction** | GenScan | genes.mit.edu | Gene prediction (Chapter 5) |
| | PsiPred | bioinf.cs.ucl.ac.uk/psipred/ | Protein structure prediction (Chapter 11) |
| | Mfold | www.bioinfo.rpi.edu/applications/mfold/ | RNA structure prediction (Chapter 12) |
| **Phylogenetics** | Phylip | bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html | Tree reconstruction (Chapter 13) |
| | PhyML | atgc.lirmm.fr/phyml/ | Tree reconstruction (Chapter 13) |
| **Edition/ Visualization** | Jalview | www.jalview.org | Alignment editor (Chapter 10) |
| | Logos | weblogo.berkeley.edu | A MSA Visualization Tool (Chapter 10) |
| | Trees | iubio.bio.indiana.edu/treeapp/treeprint-form.html | Tree Visualization (Chapter 13) |
| | Rasmol | www.umass.edu/microbio/rasmol/ | Structure visualization (Chapter 11) |

unsettling, here's some reassuring news: Anything new you might be looking for is probably only one or two clicks away from the generic resources we list in Table 15-3. In this list, the last four links take you to some very good online resources where you can find almost anything you want on bioinformatics-related topics.

**Table 15-3 Ten Bioinformatics Resource Locators**

| *Name* | *Address* | *Description* |
|---|---|---|
| ExPASy | www.expasy.ch | Dedicated to proteins |
| ArrayExpress | www.ebi.ac.uk/microarray/ | DNA chips |
| Swbic | www.swbic.org/ | Miscellaneous links |
| Pasteur | bioweb.pasteur.fr/intro-uk.html | Miscellaneous links; many online tools |
| RNA World | www.imb-jena.de/RNA.html | RNA-related links |
| miRNAs | microrna.sanger.ac.uk/sequences/index.shtml | Extensive Resources on miRNA |
| Phylip | evolution.genetics.washington.edu/phylip/software.html | Everything on phylogeny |
| NCBI primers | www.ncbi.nlm.nih.gov/education | Very good primers on many subjects |
| Bielefeld | bibiserv.techfak.uni-bielefeld.de/intro/dist.html | Awesome online course |
| Bio-informer | www.ebi.ac.uk/Information/News/ | The EBI online news |
| Coffee Corner | www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=coffeebrk | NCBI Online News |

if you are always looking for fresh meat and just out of the oven cakes, the resources in Table 15-4 are for you. Better take a look if you want to turn into a guru yourself.

**Table 15-4 Ten Places to Go Farther**

| *Name* | *Address* | *Description* |
|---|---|---|
| Nucleic Acid Research | `nar.oxfordjournals.org/` | Once a year, NAR publishes both a database issue and Web-server issue. These are available for free — and contain the state of the art in bioinformatics. |
| Bioinformatics | `bioinformatics.oxfordjournals.org/` | Bioinformatics contains articles describing the most recent methods in bioinformatics. |
| Conferences | `www.iscb.org/events/event_board.php` | An exhaustive list of major conferences in the field of bioinformatics, provided by the International Society For Computational Biology. |