4

Computational Tools and Resources in Plant Genome Informatics

Todd J. Vision[1] and Aoife McLysaght[2]

[1] *Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, and* [2] *Ecology and Evolutionary Biology, University of California, Irvine, CA, USA*

INTRODUCTION

Though all biologists deal with information, only recently have the computational challenges of systematically collecting, storing, organising, manipulating visualising and analysing large amounts of biological information come to be widely appreciated. The cause of this is the explosive growth of genomics. The term *bioinformatics* was originally coined for the application of information technology to large volumes of biological, and particularly genomic, data. The field of bioinformatics has come to be intermingled with traditional computational biology and biostatistics, which are strictly concerned not with how to handle the information itself, but rather how to extract biological meaning from it. Thus, bioinformatics, in its broad sense, can be seen as providing both the infrastructure and the scientific framework in which biologists take information and use computers to help convert it into knowledge.

Despite the relative youth of the field as a recognised discipline, there is an impressive diversity of bioinformatics resources currently available. By necessity, we only focus on a small slice of this diversity here. We pay particular attention to sequence analysis because of its centrality to genomics. We also do not attempt to provide specific protocols, as the specific needs of users vary greatly. The resources we describe range drastically in sophistication from little tested programs posted on graduate student web pages to very stable and complex databases maintained by governmental agencies. The better ones typically provide manuals and tutorials, often containing descriptions of the underlying principles. The reader is strongly advised to consult the documentation available for each tool.

Though a wide array of commercial resources exist, some of which are ideally suited to specific tasks, many of the most fundamental and long-lived bioinformatics tools are freely available. For this reason, we describe primarily non-commercial software in this chapter. Many of the databases and analysis tools we describe are hosted by government or academic research centres and can be accessed via user-friendly web interfaces. Tables 4.1 and 4.2 list the Uniform Resource Locators (URLs) for all the online resources that are discussed in the text.

WHAT IS OUT THERE AND HOW TO GET IT

Collectively, online databases allow access to a staggering quantity of data. This partly reflects the way much biological data is now collected. Genome projects popularised the concept of

Handbook of Plant Biotechnology
Edited by Paul Christou and Harry Klee

Table 4.1 URLs for plant genome informatics tools and resources discussed in this chapter

Resource	URL	Description
AtRepBase	nucleus.cshl.org/protarab/AtRepBase.htm	Repeat sequences in Arabidopsis
BioMail	www.biomail.org/	Pubmed search robot
BLOCKS	blocks.fhcrc.org/	Protein family alignments
Comprehensive Microbial Resource	www.tigr.org/tdb/mdb/mdbcomplete.html	Microbial genomes and analysis tools
DDBJ	www.ddbj.nig.ac.jp/	Primary sequence database
Distributed Annotation System	www.biodas.org	Software for managing genome sequence data
EMBL Nucleotide Sequence Database	www.ebi.ac.uk/embl/index.html	Primary sequence database
EMBOSS	www.emboss.org/	Sequence analysis software
Enzyme Commission Database	www.expasy.ch/enzyme/	Enzyme nomenclature
Genbank	www.ncbi.nlm.nih.gov/Genbank/index.html	Primary sequence database
Gene Expression Omnibus (GEO)	www.ncbi.nlm.nih.gov/geo/	Gene expression database
Gene Indices	www.tigr.org/tdb/tgi.shtml	Non-redundant gene sets for many organisms
Gene Ontology Consortium	www.geneontology.org/	Controlled vocabulary for gene function
GeneNet	www.mgs.bionet.nsc.ru/mgs/systems/genenet/	Gene network data and analysis tools.
GOBASE	megasun.bch.umontreal.ca/gobase/gobase.html	Organelle genome database
GOLD	igweb.integratedgenomics.com/GOLD/eukaryagenomes.html	Information about genome projects
Graingenes	wheat.pw.usda.gov/index.shtml	Genomic database for wheat, oat, barley, rye and sugarcane
Gramene	www.gramene.org/	Comparative genome analysis in the grasses
HMMer	hmmer.wustl.edu/	Profile hidden Markov model software
HYPHY	peppercat.stat.ncsu.edu/∼hyphy	Software for analysis of sequence evolution
Indiana University Molecular Biology Software Archive	iubio.bio.indiana.edu/soft/molbio/	Bioinformatics Software
InterPro	www.ebi.ac.uk/interpro/	Integrated database of protein family signatures
Klotho	www.ibc.wustl.edu/klotho/	Biological compound database
MaizedB	www.agron.missouri.edu/	Maize genomics
Mendel Plant Gene Names	genome-www.stanford.edu/Mendel/aboutMendel.html	Nomenclature for sequenced plant genes
MIAME	www.mged.org/Workgroups/MIAME/miame.html	Microarray annotations working group
MMDB	www.ncbi.nlm.nih.gov/Structure/	3D-biomolecular structures
PAML	abacus.gene.ucl.ac.uk/software/paml.html	Software for analysis of sequence evolution
PDB	www.rcsb.org/pdb/	Primary structural database
Pfam	www.sanger.ac.uk/Software/Pfam/	Protein family signatures
PHYLIP	evolution.genetics.washington.edu/phylip.html	Software for phylogenetics
Phylodendron	iubio.bio.indiana.edu/treeapp/	Software for drawing phylogenetic trees
PLACE	www.dna.affrc.go.jp/htdocs/PLACE/	Plant *cis*-acting regulatory elements
Plant Genomes Central	www.ncbi.nlm.nih.gov/PMGifs/Genomes/PlantList.html	Plant genomics database
PlantCARE	sphinx.rug.ac.be:8080/PlantCARE/cgi/index.html	Plant *cis*-acting regulatory elements
PlantGDB	www.zmdb.iastate.edu/PlantGDB/	Plant ESTs and comparative gene modelling
PLANTncRNAs	www.prl.msu.edu/PLANTncRNAs/	Plant non-protein coding RNAs
Primer3	www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi	Primer design

Continued

Table 4.1 *Continued*

Resource	URL	Description
PRINTS	www.bioinf.man.ac.uk/dbbrowser/PRINTS/	Protein family signatures
ProDom	prodes.toulouse.inra.fr/prodom/doc/prodom.html	Protein family signatures
PROSITE	www.expasy.org/prosite/	Protein family signatures
Protein Information Resource (PIR)	pir.georgetown.edu/	Protein sequence database
PubCrawler	www.pubcrawler.ie/	Pubmed and Genbank search robot
PubMed	www4.ncbi.nlm.nih.gov/entrez/query.fcgi	Biological literature
SMART	smart.embl-heidelberg.de/	Protein family signatures
Salk Institute Arabidopsis Gene Mapping Tool	signal.salk.edu/cgi-bin/tdnaexpress	Insertion mutant database for *Arabidopsis*
Solanaceous Genome Network	www.sgn.cornell.edu/	Genomic database for tomato, potato and pepper
Stanford Microarray Database	genome-www5.stanford.edu/MicroArray/SMD/	Microarray data repository
SWISS-PROT/TrEMBL	www.expasy.org/sprot/	Protein sequences
TAIR	www.arabidopsis.org/	*Arabidopsis thaliana* genomic resources
TIGRFams	www.tigr.org/TIGRFAMs/	Protein family signatures
TRANSFAC	transfac.gbf.de/TRANSFAC/index.html	Transcription factors and binding sites
TreeView	taxonomy.zoology.gla.ac.uk/rod/treeview.html	Software for drawing phylogenetic trees
UKCropNet	ukcrop.net	Crop genome databases and comparative mapping tools

high-throughput, highly automated biological data factories, in which data is systematically collected with the express purpose of facilitating as-yet-unknown downstream applications. As a result, the value of such data is only realised when it is made accessible to the research community as a whole.

The growth in the size of Genbank (Benson *et al*., 2002), the DNA and protein sequence repository jointly maintained by the National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL) and the DNA Databank of Japan (DDBJ), is legendary. Genbank contained 14.4 billion base pairs by the end of 2001, 200 times the number of base pairs in the database just ten years earlier. In step with the growth in sequence data, a wide variety of different types of data have become available. These run the gamut from raw sequence data to highly derived computational predictions of protein structure and biomolecular interactions.

Unlike Genbank, which archives sequence data from all organisms, many database resources are organism specific. A variety of crop and model-plant specific genomic databases are accessible through UKCropNet. These include GrainGenes, (which holds molecular and phenotypic information on wheat, barley, oats, rye and sugarcane), and MaizeDB (which performs a similar service for maize). Some databases are specific to somewhat larger taxonomic assemblages. For example, the Gramene database is a recent effort that aims to integrate genomic information from among all grasses using the rice genomic sequence as a focal point (Ware *et al*., 2002).

It can be helpful to recognise a distinction between primary data repositories, on the one hand, and derivative databases that offer a regularly updated analysis of data from primary repositories, on the other. Genbank is an example of a primary repository. Pfam, a protein sequence signature database, is an example of one that is derived. Derived databases in plant genomics frequently only include those plant systems having the most abundant data. One example is the set of Gene Indices at The Institute for Genomic Research (TIGR), which is a collection of very focussed databases, each covering a different plant, animal, protist or fungal species (Quackenbush *et al*., 2001). Each Gene Index computationally assembles the non-redundant set of gene sequences for that organism, with links to expression,

Table 4.2 Internet jump stations for bioinformatics tools

Institution	URL	Some available tools
European Bioinformatics Institute (EBI)	http://www.ebi.ac.uk/Tools/	• SRS—Sequence Retrieval Software • Sequence similarity search tools (including fast true Smith–Waterman searches) • 3D structure analysis tools • Sequence alignment • Protein motif detection • Sequence repeat discovery • Sequence translation • Primer detection • DALI—protein 3D structure similarity search tool
Japanese Genome Net	http://www.genome.ad.jp/	• Sequence similarity search tools • Multiple alignment • Sequence motif search
National Centre for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/Tools/	• PubMed–medline database browser • Entrez–sequence database browser • BLAST–sequence similarity search tool • Electronic PCR • ORF (Open Reading Frame) Finder • VAST–protein 3D structure similarity search tool
Pasteur Institute	http://bioweb.pasteur.fr/intro-uk.html	• BLAST • ClustalW • EMBOSS • PHYLIP • Primer design • RNA analysis
Swiss Institute of Bioinformatics (SIB)	http://www.expasy.org/	• Sequence similarity search tools • Patterns and profile searches • Sequence alignment (including T_COFFEE) • Protein structure prediction • Transmembrane region prediction • Post-translational modification prediction • 2D PAGE analysis

homology and other information. Those plants for which there exist sufficient publicly available sequence data are included. This includes fourteen species at the time of writing. Because it was the first plant nuclear genome to be sequenced in its entirety, *Arabidopsis thaliana* is sometimes the sole plant representative in other genomic databases. An example of this is MODBASE, which contains homology modelled protein structures using predicted amino acid sequences from a variety of completed genomes.

Plant biologists are, of course, also interested in plant symbionts and disease causing organisms. A number of plant pathogenic bacteria and fungi have either been sequenced in their entirety, including *Agrobacterium tumefaciens* (Goodner *et al*., 2001), *Ralstonia solanacearum* (Salanabout *et al*., 2002) and *Xylella fastidiosa* (Simpson *et al*., 2000), or are the subject of ongoing sequencing projects, such as *Magnaporthe grisea* (Zhu *et al*., 1997), *Pseudomonas syringae* pv. *tomato* and *Xanthomonas campestris*. Completed sequence is also available for the legume nodule-associated mutualist *Sinorhizobium meliloti* (Capela *et al*., 2001). In addition, a variety of plant viral genomes have been deposited in Genbank. The Genomes OnLine Database (GOLD) is a regularly updated online listing of prokaryotic and eukaryotic genome projects that have been completed or that are under way. TIGR offers what it calls the Comprehensive Microbial Resource database, which allows exploration and comparison of the annotated microbial sequences. Unfortunately, genomic information for metazoan plant symbionts, such as pathogenic

nematodes and insect herbivores, is much less abundant and likely to remain that way for some time.

An excellent resource to the world of genomic databases is the annual database issue of the journal *Nucleic Acids Research*, published on the 1st of January each year (www3.oup.co.uk/nar/database/c/). In addition to written descriptions of dozens of different databases, a list of links to hundreds of databases, organised by category, is maintained online. Publications describing online databases quickly become obsolete as new databases spring up and old ones change, and no list (online or otherwise) could hope to be comprehensive, but this is a good place to start. Website addresses (URLs) for databases and resources discussed in this chapter are provided in Table 4.1, while major web jump stations for genomics and bioinformatics are given in Table 4.2.

The Growing Role of Standards

The meanings of biological terms are often slippery and operational. For instance, 'gene function' can easily mean different things to different practitioners. Although it may be preferable, in some cases, to allow for ambiguity rather than force misguided precision, computers are not at all adept at handling ambiguity. Thus, there has been much effort expended in adopting standardised terminologies, with clear relationships defined among the terms. Such language standards are referred to as controlled vocabularies, or ontologies. Ontologies provide transparency of meaning to users and greatly facilitate inter-communication among databases.

One of the oldest systematic attempts to standardise plant gene nomenclature is the Mendel Plant Gene Names Database and its derivatives, which provide a useful categorisation of known plant genes and their sequences (Lonsdale *et al*., 2001; Price *et al*., 2001). The Enzyme Commission Database, which is taxonomically broader, offers a heavily used classification system that organises enzymes hierarchically by function. An even more ambitious effort is that of the Gene Ontology (GO) Consortium, which works to produce a dynamic controlled vocabulary, valid across all organisms, that can accommodate accumulating and changing knowledge of gene function (The Gene Ontology Consortium 2001). GO recognises three independent ontologies for genes and gene products:

(i) *Molecular function*, which is specific to an individual gene product (e.g. DNA helicase)
(ii) *Biological process*, which is coordinated by multiple products (e.g. mitosis)
(iii) *Cellular component*, which describes the physical localisation of a gene product (e.g. nucleus)

Controlled vocabularies are not restricted to gene or protein function. A number of plant databases (including TAIR—The Arabidopsis Information Resource, Gramene and MaizeDB) are collaborating to provide a controlled vocabulary for plant-specific terms such as anatomy, morphology and development (The Plant Ontology Consortium, in press).

In addition to controlled vocabularies, there is an important role for standards that define the salient features of particular kinds of data. For example, a group has been working to develop a standard for the minimum information about microarray experiments (MIAME). The diversity of experimental and analytical approaches to microarray expression data could potentially be a major barrier to the verification and integration of such data by the research community as a whole. MIAME is a set of evolving guidelines designed to 'facilitate the establishment of databases and public repositories and enable the development of data analysis tools' (Brazma *et al*., 2001).

Each of these approaches at facilitating transparent communication among multiple users and databases has slightly different goals and guiding philosophies. Some of the earliest and most successful initiatives to date in this area have tackled the practical, and limited, goal of establishing concrete relationships among the entities in a small number of related databases. The InterPro database, for example, provides a single point of entry for searching a large number of different protein signature (motif and domain) databases, including PROSITE, PRINTS, ProDom and Pfam, SMART, and TIGRFams (Apweiler *et al*., 2001).

Interface Issues

The interface one uses to interact with a database is partly a decision of the database developers.

But frequently, multiple options are available to the user. In such cases the choice of interface may determine the complexity of the queries that are possible. In principle, interfaces can be designed to be largely independent of the database implementation. Here we discuss the SRS system provided by EMBL, and the Entrez system provided by the NCBI.

SRS (Sequence Retrieval System) is an integrated system providing an interface to multiple databases, including sequence databases, OMIM mutations database, protein structure databases, gene family databases, metabolic pathway databases, Medline and many more. The complete list of databases can be seen on the starting page. The basic procedure is to select a database (or databases), and then perform a query. A basic query can be performed in the 'Quick Search' box on the top page, or through the standard query form. It is possible to perform complicated database queries with relative ease through the SRS query forms. For example, requesting all the annotated introns within *Arabidopsis thaliana* genes using the 'extended query form'. SRS will automatically save your performed queries and results (available through the 'results' tab on the navigation bar) which you can later exploit by asking for entries that are shared or unique between several sets of runs. One of the other powerful aspects of SRS is its ability to link data from different databases. So, for example, if you are looking at an EMBL (Genbank) entry (or group of entries) which codes for one or more proteins, you may link to the SWISS-PROT protein database (described below) to retrieve the protein sequences and the excellent annotation that goes along with them. You may also launch BLAST or Fasta similarity searches, and ClustalW alignments (all described below) from within SRS. A useful user's guide to SRS can be found under the 'information' link on the top SRS page.

One can also access the data at NCBI by multiple routes, collectively referred to as the Entrez system (Schuler *et al*., 1996): through a simple keyword search using the text box on the NCBI homepage, through an advanced search from on the NCBI webpage that allows the user to enter a Boolean statement (combinations of logical AND, OR and NOT operators with search terms that are specific to individual data fields), through Network Entrez (a Graphical User Interface, or GUI, program on the user's computer that queries the database remotely), and through Batch Entrez (a process allowing the user to save multiple database records on a local computer—these records may be specified in a file containing a list of accession numbers uploaded from the user's computer, or by a normal database search). In addition, one can do a similarity search of the sequences in the database (i.e. BLAST–described below) and then access the individual data records that are returned via direct HTML-formatted URL links. Using the same HTML-formatted query syntax as found in the BLAST output, one can access individual records over the network using homespun programs. Alternatively, one can download the regularly updated sequence databases from the FTP site to perform local searches. Most of the same access routes can be used to obtain data from any of the other databases hosted at NCBI, such as PubMed (literature database), and MMDB (protein and nucleic acid 3D structure database; Wang *et al*., 2002). However, NCBI is exceptional in this regard, and most database developers typically offer a much more limited range of access points to the data.

At least in principle, submission of data to a remote database can also be accomplished in any number of ways. In the major primary sequence databases, users submit all new records. NCBI offers a web-based submission tool for this purpose called Bankit, and also a GUI application called Sequin which runs on the user's computer, allows error-checking, long submissions and, most importantly, batch submissions. The entries in the PubMed literature database are supplied by the publishers. The MMDB database, by contrast, derives all of its records from another protein structure database (see below), but excludes certain records based on *a priori* criteria (those based on theoretical models). Some databases display the results of an analysis pipeline that operates on data derived from elsewhere; this is the case for the TIGR Gene Indices (described above). As discussed below, the route(s) of submission and the amount of scrutiny each record receives are critical features in judging the potential utility and limitations of a database.

Databases are often classified as object-oriented, relational or hybrid. The definition of these terms is beyond the scope of this volume, but it is helpful to recognise that object-oriented and relational

databases require different technologies to be accessed. The major object-oriented databases in the genomics community use AceDB, a software system devised initially for the *Caenorhabdites elegans* genome project by Durbin and Thierry-Mieg (1991). Many of the USDA-ARS and UKCropNet-sponsored databases employ AceDB. The interface to these databases is through AceDB Query Language, or AQL, a language which provides the basic tools needed to load, retrieve, filter, summarise and sort data in an AceDB database. By contrast, many recently developed, large-scale databases are relational, which means that they are structured as multiple tables (similar to spreadsheets) that are linked by the entries in particular columns (so-called keys). These databases use some variant of Standard (or Structured) Query Language (SQL). Although the syntax is not identical, both AQL and SQL use similar keywords, allow analogous operations and are useful additions to the arsenal of the bioinformaticist. Many database interfaces are simply restrictive wrappers for AQL or SQL queries; being able to formulate such a query directly allows much greater flexibility and power.

Curation, Updates and Reliability

Just as in any other scientific endeavour, it is important to understand the limitations of the data when undertaking a bioinformatic analysis. Databases vary in their submission standards, the level of curation, update policies and procedures for detecting and resolving inconsistencies or redundancies. An excellent example is the problem of assigning tentative functions to proteins on the basis of DNA or amino acid homology. The biological role of only a very small number of proteins has been experimentally determined. It is frequently the case that assignments are made based on the annotations of closely related sequences. These annotations can themselves be indirect. Thus, functional annotation may be propagated through a chain of sequence relationships resulting in the erroneous assignment of function to a protein that is only distantly related to that for which function has been experimentally determined.

The route by which data enters the database is critical. For Genbank, the initial depositor of a record is the only party with permission to update the record. Though a versioning system exists to allow updates to records, this is generally only used by high-throughput sequencing facilities. Sequence corrections and updated annotations are otherwise exceedingly rare. Although Genbank has mechanisms for screening submissions to ensure that all records are complete and self-consistent, they do not exercise editorial control over the accuracy of sequences or annotations. Thus, it is often useful to retain a healthy degree of scepticism when analysing these data. By contrast, the protein database at SWISS-PROT is highly curated, with efforts to keep current with the literature on each protein and to incorporate the knowledge of outside expert referees. With the exception of a few resources such as SWISS-PROT, there is generally an inverse relationship between the size of the database and the amount of human oversight. However, one cannot conclude that a small database is necessarily more a reliable one; a problem with some of the small, highly curated databases is that they sometimes do not have well-defined or rigorously enforced curatorial policies.

A related issue is the extent of documentation and the transparency of curatorial policies. Many sites have minimal documentation online, though some of these are more fully described in published papers. One consequence is that it is sometimes not obvious to the casual user that a site has a strong taxonomic (or other) bias. For many genomic databases, the hidden taxonomic bias is at the expense of plant data. Even when present, database documentation is not always upfront about problems of representation. This problem is particularly acute in smaller molecular biology databases, which are often focussed on mammalian, or other animal, systems.

Despite the best efforts of database developers, biological knowledge is inherently dispersed. In addition, there are often conflicting ideas about the same biological entity, such as a gene model. A partial solution to this has been adopted by the creators of the Distributed Annotation System (DAS). With DAS, a single 'client' collects genome annotation information from multiple remote 'servers' (including potentially the user's own local database) and displays it to the user in an integrated fashion. This frees the user from reliance on a single, possibly static, version of the annotation. DAS achieves cross-platform interoperability by

using the eXtensible Markup Language (XML), a very flexible and widely used web protocol for the exchange of data. TIGR manages a DAS server for the *Arabidopsis* genome data.

Data Manipulation

Increasingly, molecular biology studies use databases not just to retrieve individual records, but as a source of large datasets that can be explored to test a hypothesis or search for a pattern. The data often need to be post-processed, and sometimes integrated among different sources, before they can be effectively run through an analysis pipeline. One may need to convert DNA sequence files into a format that can be read by BLAST (see below), or transfer a data file designed for input into one application into a different format. In some cases, general purpose tools have been written to accomplish this. For example, NCBI distributes FORMATDB, a program which reads FASTA formatted input files and converts them to a set of binary files that can be read by BLAST. Readseq, available from the Indiana University Molecular Biology Software Archive (Table 4.1) is a handy tool that can inter-convert among a variety of nucleic acid and protein sequence formats. There are several web servers that run Readseq remotely. Many simple sequence analysis utilities that do such things as format conversion, plus also a number of more sophisticated tools, have been incorporated into the excellent open source EMBOSS package (Rice *et al*., 2000).

If a tool cannot be found for post-processing, or format conversion, an enormous amount of manual time can be wasted on the task. A better solution is to write a program using a scripting language such as Perl, which has very rich text-processing capabilities (including the ability to formulate very general *regular expressions*, described below). It takes only rudimentary skill in this language to write a program that can perform sophisticated text manipulations. The small investment of effort required to write a script is quickly repaid the next time the same problem is encountered, so Perl, or something like it, is well worth learning. Perl has the advantage that it is in widespread use by bioinformaticians, giving rise to open-source projects such as bioperl**b** where many common biologically relevant tasks have already been coded as Perl modules, or add-ons. As more people use and contribute to this project, the toolsets becomes richer and more versatile. Basic Perl tutorials aimed at novice-programming biologists are included in some general bioinformatics texts (e.g. the bioinformatics textbooks by Baxevanis and Ouellette, and by Gibas and Jambeck, see bibliography).

QA1

PISE (Letondal, 2001) is a software tool that can be used to increase the user friendliness of the many bioinformatics software tools that presuppose a certain level of familiarity with command windows and text processing. PISE provides an intuitive web interface for input to and output from any standard text-driven software programs, has the flexibility to handle messy format conversions behind-the-scenes and puts the program parameter options directly in front of the user. It includes an interface to PHYLIP, and EMBOSS programs. The capabilities of PISE are on display at the Institut Pasteur Bioweb (Table 4.2).

A TOUR OF SOME ONLINE DATABASES

Literature

Due to the proliferation of the scientific literature, citation and abstract databases have become indispensable to the scientific enterprise. PubMed, the NCBI literature database for molecular biology and biomedicine, has, at the time of writing, records for over 27 000 journals. Citation and abstract databases represent an under-utilised resource for data-mining (Iliopoulos *et al*., 2001). Tools such as Pubcrawler (Hokamp and Wolfe, 1999) and Biomail (Mozzherin, Herrera, and Miller, unpublished) will perform predefined searches on the PubMed database at regular intervals and mail the resulting citation URLs to the user. Pubcrawler will also perform regular searches of Genbank (see below), so you know when any new genes from your favourite organism, or pathway, are submitted.

DNA Sequence

The premier DNA sequence databases are maintained by the members of the International

Nucleotide Sequence Database Collaboration (INSDC), which includes the DNA Data Bank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database and Genbank, operated by the United States NCBI. The three databases use somewhat different formats, but are identical in content; submissions to any one of the members are recorded in all three databases. Each sequence record contains information concerning the submission, the source of the sequence, and varying degrees of feature annotation. Several specialised categories of DNA sequence, including Genome Survey Sequence, High-Throughput Genomic, Expressed Sequence Tag and others, are now recognised. In addition to the sequence data itself, these sites offer many analysis tools, including sophisticated similarity searches. They also provide a network of links between related records in different databases (such as literature, protein, etc.). DNA polymorphism data are growing in quantity and importance. Currently, the most extensive compiled dataset is for *Arabidopsis*, and can be accessed through dbSNP at NCBI. A curated set of organelle genome maps and sequences, with extensive annotation, are available through GOBASE at the University of Montreal. There exist a variety of other niche sequence databases that are curated by specialists in particular areas, such as AtRepBase for repetitive sequences in *Arabidopsis*, and PLANTncRNAs for expressed, non-protein coding RNAs in plants. The PlantGDB database offers annotated species-specific EST databases for plants from mosses to angiosperms and also hosts sophisticated comparative sequence analysis tools for gene identification and characterisation.

Protein Sequence

Primary amino acid sequences, mostly derived from translations of presumptive coding sequences, are available through the member databases of the INSDC. These data repositories tend to have many errors, a good deal of redundancy, and little functional annotation. To improve upon these resources, more intensively curated collections are provided by Protein Information Resource (PIR) and SWISS-PROT (Bairoch and Apweiler, 2000). SWISS-PROT includes comprehensive links to other databases and should be one of the first places to look when dealing with protein sequences.

There are a variety of databases that capture information about protein sequence features that are conserved across multiple proteins, and thus likely to be of functional significance. These include such loosely defined terms as protein motifs and domains, and can be generally referred to as protein signatures. The BLOCKS (Henikoff *et al*., 1999), PROSITE (Falquet *et al*., 2002), Pfam (Bateman *et al*., 2002) and ProDom (Corpet *et al*., 2000) databases use different but related technologies to generate and search collections of protein motifs, domains, or conserved regions of a multiple alignment among members of a protein family. PRINTS (Attwood *et al*., 2002) and SMART (Letunic *et al*., 2002) can also identify protein families by fingerprints, or combinations of different motifs. InterPro (Apweiler *et al*., 2001) provides a single interface for text or sequence-based searching of most of the major protein signature databases.

Protein Structure

The primary database for experimentally (and some theoretically) determined 3D structures is the Protein DataBank or PDB (Berman *et al*., 2000). The data come primarily from X-ray crystallography and NMR spectroscopy. Though it has a focus on proteins, PDB also contains information about the structures of nucleic acids, carbohydrates, and other biomolecules. Another important resource for structural biology is the Molecular Modeling DataBase (MMDB) at NCBI (Wang *et al*., 2002). MMDB builds upon the data in PDB by trying to reconcile apparent discrepancies between the sequence and structure and allowing a deterministic reconstruction of the chemical bonds in the molecule from the coordinate data. Through the Entrez system, NCBI also provides links between MMDB and its putative structural neighbours in the protein sequence database. These are identified using an all-by-all search with the VAST algorithm (Gibrat *et al*., 1996). A number of databases, such as CATH (Pearl *et al*., 2000) and SCOP (Murzin *et al*., 1995), provide classifications of protein structures or substructures, and many of these also provide

access to software tools for viewing and comparing structures.

Genome Maps

Genome maps can be classified as marker maps or trait maps. A marker map is a model of the arrangement among physical features of the genome. The distances between features may be represented in base pairs, or centiRays (physical map units of a radiation hybrid map), or centiMorgans (recombinational map units), or some other measure. Trait maps, which represent the contribution made by different genomic regions to phenotypic variation, are not yet routinely integrated into genomic databases, although some effort has been made within certain organism-specific databases such as MaizeDB. Marker maps are also best explored at organism-specific databases, although Plant Genomes Central at NCBI now provides a view of the genetic and/or physical map data for a small number of well-studied plants. One of the more important tasks of a genome map database is to cross-link different maps of the same genome. TAIR, for example, has an excellent interface that allows one to browse the *Arabidopsis* genome while keeping track of the position in multiple maps (including the assembled sequence, a tiling path of large-insert clones, two genetic linkage maps of molecular markers and a genetic linkage map of visible mutations).

Gene Expression

A large number of high-throughput gene expression studies have been and are being carried out in plants using a variety of competing technologies. Collectively, these data are a treasure trove for comparative studies of gene function. Unfortunately, there is currently no universal site from which to access results of these studies. In part, this is due to uncertainty concerning the best way to organise and disseminate microarray data electronically. Only the results from a small number of *Arabidopsis* experiments are currently available in NCBI's Gene Expression Omnibus (Edgar *et al*., 2002) and in the Stanford Microarray Database (Sherlock *et al*., 2001). Many efforts are under way to remedy the lack of standards and the paucity of centralised repositories for microarray data, such as the MIAME project discussed above.

Other Databases

In addition, many specialty resources have been developed that do not fit into the above categories. There are a number of databases that attempt to curate particular functional groups of genes or sequence elements. PLACE (Higo *et al*., 1999) and PlantCARE (Lescot *et al*., 2002) both specialise in plant *cis*-acting regulatory elements, while TRANSFAC (Wingender *et al*., 2000) covers both trancription factors and binding sites in all eukaryotes. The Database of Interacting Proteins (Xenarios *et al*., 2002), which contains a limited number of plant records, keeps track of experimentally determined protein–protein interactions. GeneNet (Kolpakov *et al*., 1998) provides information on a small number of select gene networks with tools for visualisation and dynamic simulation; some important plant processes are included. The widely used Enzyme Commission system of nomenclature is available through the ExPASy (Expert Protein Analysis System) server of the Swiss Institute of Bioinformatics. Klotho collects and categorises information on many biological compounds. Stock centre databases can be incredibly valuable resources, as well. For instance, collections of transposon insertion mutant flanking sequences, a key tool in reverse genetics, can be searched for matches to a gene of interest. A number of such databases have been established for *Arabidopsis*, such as FLAGdb and the Salk Institute *Arabidopsis* Gene Mapping Tool. Some of these provide a useful service that will automatically notify the user of new submissions matching a specified gene.

ANALYSIS TOOLS

Bioinformatics is a discipline that, by its nature, has more occasional users than specialists. Occasional users are typically not too familiar with the algorithms and statistics underlying the analytical tools they use (any more than most biologists understand the workings of their thermocyclers). While knowledge of these details can be immensely helpful, it is unrealistic

to assume that every bench scientist who may want to perform an alignment, or construct a phylogenetic tree, would first engage in a thorough study of the literature. So, for the impatient bench scientist, we provide here a condensed introduction to some of the most commonly used (and misused) tools in sequence analysis.

As a general rule, most parameters for the algorithms implemented in bioinformatics applications have default settings in the program (i.e. the value that is used unless the user specifies otherwise). However, there is rarely, if ever, an *a priori* best value for a particular parameter; they are almost invariably dataset dependent. Thus, while it is possible to obtain a multiple sequence alignment by simply feeding your sequences into an alignment program, there is no guarantee that the resulting alignment is the best (i.e. closest to the true alignment) that computational methods can offer. It is wise to both experiment with program parameters, and to apply biological judgement in deciding upon the optimal parameters for a particular dataset.

Another useful thing to remember is that, if you are faced with a biological question requiring some computational application, the chances are that someone before you has been in the same spot. Rather than spend valuable time reinventing the wheel (although doubtless a great learning experience), a web search will often be fruitful, and may even leave you with a choice of programs. New programs are arriving on the web all the time. In the case of most standard sequence analysis methods, the programs one is likely to need are already freely available on the Web and do not require the user to instal any software. This is particularly important for the bench scientist who may not be using these tools frequently enough to justify investment in a commercial software package. For example, there are many websites with free primer design software (e.g. Primer 3, Table 4.1). To verify that the software is of high quality, one should examine the associated publication and determine whether it is being used by reputable research groups. Several bioinformatics research institutions provide well-maintained interfaces to the most commonly used (and generally respected) tools. Some of these jump stations are listed in Table 4.2 and are useful starting points when looking for a new tool.

Some of the most commonly asked questions in bioinformatics are essentially evolutionary in nature. One dogma of bioinformatics is that closely related protein sequences should have similar structures, and thus have similar functions. There are vast amounts of sequence data available and very less data on protein structure and function. Therefore, identification of similar sequences can be one of the quickest routes to understanding the function of a protein. This is accomplished by pairwise local alignment between a single query sequence and a large sequence database (e.g. SWISS-PROT/TrEMBL). Multiple sequence alignment and phylogenetic tree construction can throw additional light on conserved DNA or amino acid sequence signatures and on the pattern of functional divergence in the gene family. We describe each of these analyses below. In addition, we briefly mention some of the tools available for protein structural analysis and gene expression data, as these are increasingly vital areas of genome informatics. Areas of sequence analysis of relevance to genomics that we do not cover include sequence assembly, gene prediction and sequence annotation. Space limitations also require us to neglect the considerable computational issues in genetic and physical mapping.

Homology Detection

Homologous sequences are related by descent from a common ancestor. Highly similar sequences are often, but not always, homologous: and homologues are often, but not always, highly similar (Bork *et al*., 1992). Thus, sequence similarity is an indirect indicator of homology, which is itself an indirect indicator of shared functionality. There are no degrees of homology; sequences are either related by descent, or they are not (Reeck *et al*., 1987; Fitch, 2000). In contrast, similarity comes in shades of grey, as two sequences may have diverged to a greater or lesser extent from their common ancestor. It would thus be correct to say that two proteins are x% identical at the amino acid level, though incorrect to say that they are x% homologous.

The Smith–Waterman Algorithm

The best mathematical solution to the not inconsiderable problem of identifying the most similar sequence in a database of over 14 million

a

	C	A	G	C	C	U	C	G	C	U
A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
U	0.0	0.0	0.7	0.4	0.0	1.0	0.0	0.0	0.0	1.0
G	0.0	0.0	1.0	0.4	0.1	0.0	0.7	1.0	0.0	0.0
C	1.0	0.0	0.0	2.0	?					
C										
A										
U										
C										

b

	C	A	G	C	C	U	C	G	C	U
A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
U	0.0	0.0	0.7	0.4	0.0	1.0	0.0	0.0	0.0	1.0
G	0.0	0.0	1.0	0.4	0.1	0.0	0.7	1.0	0.0	0.0
C	1.0	0.0	0.0	2.0	1.4	0.4	1.0	0.4	2.0	0.7
C	1.0	0.7	0.0	1.0	3.0	1.7	1.4	1.1	1.4	1.7
A	0.0	2.0	0.7	0.4	1.7	2.7	1.4	1.1	0.8	1.1
U	0.0	0.7	1.7	0.4	1.4	2.7	1.4	1.1	0.8	1.8
C	0.0	0.4	1.7	2.7	1.4	1.4	3.7	2.4	2.1	1.8

c

```
GCC-UC
GCCAUC
```

Fig. 4.1 The Smith–Waterman algorithm for pairwise local alignment. Matches are scored +1, mismatches –0.3 and gaps $-(1+0.3k)$, where k is the length of the gap. (a) Each sequence is represented along a side of a matrix. The values of the matrix indicate the similarity score for the residues. (b) The alignment scores are calculated for each position in the table, as described in the text. In the case of the element marked with a question mark '?', the score resulting from adding this aligned pair (C aligned with C in this case) to an alignment ending at any of the pairs shaded grey is calculated by adding the score of this pair to the existing alignment, and deducting a gap penalty where required. The highest score is retained, or the score is zero (whichever is higher). (c) The highest value in the completed table marks the right-hand terminus of the highest-scoring alignment. The alignment itself is obtained by tracing back from this position to the first position encountered that has a value of zero (shaded elements). (d) The resulting alignment

sequences (e.g. Genbank) is the Smith–Waterman algorithm (1981), illustrated in Figure 4.1. This algorithm (and its fast approximate derivatives BLAST and Fasta) can be used to compare a query sequence to each sequence in a database, constructing an optimal pairwise alignment for each one and generating a *score* that can be used to rank the alignments. The total score for a pairwise alignment is simply the sum of the individual scores for each position in the alignment. The individual scores are positive whenever two identical or similar residues are aligned and negative when the residues are dissimilar or when a gap is introduced or extended. We discuss below how these individual scores are derived.

The Smith–Waterman algorithm differs from its predecessor, the Needleman–Wunsch algorithm (1970), in that, instead of aligning two sequences in their entirety (i.e. a *global* alignment), it efficiently compares segments of all possible lengths and chooses the *local* alignment that optimises the similarity score. Local alignment is more appropriate for database searches because conservation between long-diverged homologues is often restricted to specific regions (such as key structural domains) and a reliable global alignment would require at least some conservation throughout.

The fundamental principle of the Smith–Waterman algorithm is that, to calculate the alignment score, $S(i, j)$, (where i and j are the endpoints of the alignment in sequence 1 and 2, respectively) one need only enumerate and score the limited set of possible ways of generating this alignment by extending a subalignment.

The basis for the algorithm is a recursion equation, applied for all possible i and j, than can be described in words as follows. There are four possible ways of extending a subalignment: (i) align the next residue of sequence 1 with the next residue of sequence 2 and increase the score by the similarity score for that pair of residues ($score(a_i, b_j)$), (ii) align the next residue of sequence 1 with a gap (i.e. a deletion sequence 2) and deduct a gap penalty proportional to the length of the gap (W_k for gap of length k), (iii) insert a gap into sequence 1 and deduct a gap penalty as above or (iv) stop the alignment (and set the score to zero). The method is initialised with a null alignment, proceeds by accepting the highest scoring of the four options at every point and uses the resultant scores to fill in a table such as that in Figure 4.1. The optimal local alignment is the path through this diagram that leads to the cell having the highest score. Gaps are inserted at points where the path moves vertically or horizontally.

Local Alignment Statistics

After a database search, one wishes to obtain those sequences (and alignments) in descending order of score (normalised to correct for differences in

sequence length). A key breakthrough of recent years is the widespread use of rigorous statistics for pairwise alignment. The expectation value (*E*-value) is the number of alignments of the same or higher score that would be expected in a database search of random sequences (Altschul and Gish, 1996). As *E*-values approach zero they approximate *P*-values (i.e. the *probability* that the observed similarity would be observed in a random search). *E*-statistics are reported by the major database search algorithms, including MPsrch (a Smith–Waterman implementation described below), BLAST and Fasta. It is important to recognise that the ranking of alignments by *E*-value may well differ from the ranking according to time since divergence or per cent identity (Koski and Golding, 2001). Furthermore, although these statistics allow one to judge the most similar sequences in a given database according to a defined scoring function, it requires human judgement to interpret the biological meaning of the alignments and rankings. For example, it may be the case that even the most similar sequence in a given database is not a homologue of the query sequence.

Protein Substitution Matrices

The highest scoring alignment will depend on the method of scoring residue matches and mismatches. The simplest method is to assign a score of +1 to a match, and −1 or 0 to a mismatch. However, when dealing with protein sequences, this is overly naive. For example, replacing a leucine with isoleucine might not have a major effect on the function of the protein, whereas replacing that same leucine with the biochemically and physically dissimilar phenylalanine may have a considerable effect. Therefore, we would like to penalise a mismatch between a leucine and a phenylalanine more heavily than a mismatch between a leucine and an isoleucine.

These biological properties are (indirectly) taken into account by the PAM (Point Accepted Mutation; Dayhoff *et al*., 1978) and BLOSUM (BLOcks SUbsitution Matrix; Henikoff and Henikoff, 1992) scoring matrices. These matrices are actually based on the observed patterns of substitutions in carefully curated sets of sequence alignments. Positive scores in these scoring matrices indicate common replacements, whereas negative numbers indicate uncommon replacements. There is a good correspondence between observed rates of substitution between particular amino acid pairs and those expected based upon their physicochemical similarities.

The numbering system of the matrices refers to the evolutionary distance for which they are calibrated. For example, the PAM 250 matrix (Table 4.3) is optimised for sequences with an average of 250 substitutions per 100 amino acids. The BLOSUM 62 matrix is optimised for sequences with approximately 62% identity. (Note the difference in the meaning of the numbers for the two families of matrices.) Because of the calibration, the default scoring matrix used by a particular piece of software may be inappropriate for many sequence comparisons. In particular, if one is trying to detect highly similar, or highly diverged sequences, then it would be wise to pick a substitution matrix accordingly. Unfortunately, since there is no universal 'molecular clock' (Zuckerkandl and Pauling, 1965; Li, 1993), different matrices will be appropriate for different genes and proteins.

Programs for Database Search Using Local Pairwise Alignment

While most similarity search programs are available in standalone versions that can be run on a local computer, there are advantages to implementing a search online. Most importantly, there is no need to download any database. EBI and NCBI, plus many mirror sites, have local up-to-date versions of the centralised DNA and protein databases that can be searched from their websites. However, if a custom database of sequence data is to be searched, a local installation will be necessary. NCBI currently implements a queuing system in which multiple searches from the same network domain are given low priority. Since it generally the case that all computers at a particular research site are in the same domain, it is advisable to instal NCBI's BLAST locally when conducting a large number of searches.

MPsrch

Until recently the only software implementations of the Smith–Waterman algorithm were very

Table 4.3 A PAM 250 scoring matrix for amino acid substitutions. The row and column headings are the IUPAC single-letter codes for each residue. The matrix is symmetric. Each number is the logarithm of the ratio of (i) the probability of substitution between the row and column residues based on empirical data relative to (ii) the same probability derived from amino acid frequencies alone. Thus, the more positive the numbers, the higher the probability of that particular substitution. Numbers along the diagonal are related to the probability that the residue is identical after 250 PAM units (250 substitutions per 100 amino acids)—note that this does not require that the site has not undergone any substitution events

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	−2	0	0	−2	0	0	1	−1	−1	−2	−1	−1	−3	1	1	1	−6	−3	0	0	0	0	−8
R	−2	6	0	−1	−4	1	−1	−3	2	−2	−3	3	0	−4	0	0	−1	2	−4	−2	−1	0	−1	−8
N	0	0	2	2	−4	1	1	0	2	−2	−3	1	−2	−3	0	1	0	−4	−2	−2	2	1	0	−8
D	0	−1	2	4	−5	2	3	1	1	−2	−4	0	−3	−6	−1	0	0	−7	−4	−2	3	3	−1	−8
C	−2	−4	−4	−5	12	−5	−5	−3	−3	−2	−6	−5	−5	−4	−3	0	−2	−8	0	−2	−4	−5	−3	−8
Q	0	1	1	2	−5	4	2	−1	3	−2	−2	1	−1	−5	0	−1	−1	−5	−4	−2	1	3	−1	−8
E	0	−1	1	3	−5	2	4	0	1	−2	−3	0	−2	−5	−1	0	0	−7	−4	−2	3	3	−1	−8
G	1	−3	0	1	−3	−1	0	5	−2	−3	−4	−2	−3	−5	0	1	0	−7	−5	−1	0	0	−1	−8
H	−1	2	2	1	−3	3	1	−2	6	−2	−2	0	−2	−2	0	−1	−1	−3	0	−2	1	2	−1	−8
I	−1	−2	−2	−2	−2	−2	−2	−3	−2	5	2	−2	2	1	−2	−1	0	−5	−1	4	−2	−2	−1	−8
L	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	6	−3	4	2	−3	−3	−2	−2	−1	2	−3	−3	−1	−8
K	−1	3	1	0	−5	1	0	−2	0	−2	−3	5	0	−5	−1	0	0	−3	−4	−2	1	0	−1	−8
M	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6	0	−2	−2	−1	−4	−2	2	−2	−2	−1	−8
F	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9	−5	−3	−3	0	7	−1	−4	−5	−2	−8
P	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6	1	0	−6	−5	−1	−1	0	−1	−8
S	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2	1	−2	−3	−1	0	0	0	−8
T	1	−1	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3	−5	−3	0	0	−1	0	−8
W	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17	0	−6	−5	−6	−4	−8
Y	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	−2	−3	−4	−2	−8
V	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4	−2	−2	−1	−8
B	0	−1	2	3	−4	1	3	0	1	−2	−3	1	−2	−4	−1	0	0	−5	−3	−2	3	2	−1	−8
Z	0	0	1	3	−5	3	3	0	2	−2	−3	0	−2	−5	0	0	−1	−6	−4	−2	2	3	−1	−8
X	0	−1	0	−1	−3	−1	−1	−1	−1	−1	−1	−1	−1	−2	−1	0	0	−4	−2	−1	−1	−1	−1	−8
*	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	−8	1

time consuming and, as a result, were feasible only on very powerful computers searching small databases. However, recently there have been advances both in the speed of the average computer and in the implementation of the algorithm. The fastest running implementation of the true Smith–Waterman algorithm is MPsrch. Though about 10 times slower than the BLAST algorithm (described below), it is acceptable for smaller databases. The software is available free to academic users from the Edinburgh Biocomputing Systems website (www.edinburgh-biocomputing.com/) for local use on a UNIX/Linux operating system. MPsrch is also available online at the EBI website (see Table 4.2).

Unlike other similarity search programs (such as BLAST), MPsrch is currently only available for querying protein sequences against a protein database. If you have a DNA sequence, you must perform the translation yourself (which may be done using the transeq program on the EBI website). Thus, in the case where you do not know the reading frame of the encoded protein, it is necessary to translate in all six possible reading frames, run MPsrch six times independently, and manually combine the output.

BLAST

The most commonly used similarity search method is the Basic Local Alignment Search Tool (BLAST; Altschul *et al*., 1997). BLAST is a heuristic modification of the Smith–Waterman (1981) algorithm (i.e. it is not guaranteed to produce an optimal pairwise alignment and score); in practice, however, it performs very well. More importantly, it is orders of magnitude faster than any true Smith–Waterman implementation. The most popular online interface to BLAST is available at NCBI, where a standalone version is also available for download.

There are several parameters controlling the behaviour of the BLAST algorithm, and these

need to be carefully considered. Apart from selecting the appropriate program to suit the data (described below), the user should also explore different substitution matrices and gap penalties. In these days of burgeoning databases, the problem is not to identify a database hit, but more to limit these hits to a manageable number. To this end, NCBI BLAST allows the user to limit the results by taxonomic groups.

BLAST includes a low complexity filter, called SEG for protein sequences (Wootton and Federhen, 1996) and DUST for nucleotide sequences, that excludes some repetitive or simple portions of the query sequence from consideration. In the output, the appropriate letter code for these residues will be replaced with an 'X'. Filtering may be important if the query protein sequence contains, for example, a proline-rich domain. If this domain is included in the query then BLAST may just return any other proline-rich proteins. To avoid such artefacts, it is generally advisable to turn this filter on. Similarly, to force BLAST to ignore a particular part of a query sequence, manually replace it with 'X's. NCBI BLAST also allows one to do this by writing the residues to be masked in lower case and selecting the appropriate filter.

BLAST includes several parameters controlling the opening and extension of gaps into the alignment. The user should investigate the consequences of changing these parameter settings since the optimal settings cannot be known beforehand and are very unlikely to correspond to the defaults. The gap and extension parameters have an even greater effect on the quality of multiple sequence alignments, and are discussed more thoroughly in that section (see below).

The standard BLAST algorithm has been implemented for nucleotide–nucleotide similarity searches (i.e. comparison of a nucleotide query to a nucleotide database without translation—BLASTN), protein–protein searches (BLASTP), protein–nucleotide comparisons with the query conceptually translated in all six reading frames (BLASTX), nucleotide–protein comparisons with each database sequence conceptually translated in all six reading frames (TBLASTN), and nucleotide–nucleotide comparisons conducted at the protein level, with both query and database translated in all six reading frames (TBLASTX). Obviously, the searches requiring translation of nucleotide sequences are multiplying the work and so take longer than a standard protein–protein search. However, BLAST produces much more reliable alignments (and thus has greater sensitivity to detect homologues) when protein sequences (known, or conceptually translated) rather than DNA sequences are used.

Fasta

Fasta (Fast Alignment; Pearson, 2000) was the first widely used heuristic algorithm for very rapid local sequence alignment against a database. It works somewhat differently than BLAST, although this is not apparent to the user, and the results of the two programs often coincide quite closely. Fasta can be freely downloaded as a standalone program and is also available online at EBI. Pearson (2000) provides a useful comparison of the programs mentioned here along with practical guidelines on how to implement sequence similarity searches and interpret the results. Fasta also lends its name to a sequence format that is readable by many different software packages. A Fasta formatted sequence has a *definition line* that begins with a '>' character followed by one or more lines of raw sequence. No sequence should be included on the definition line because it will be ignored. A Fasta file may have multiple sequences as long as each is preceded by its own definition line. BLAST expects both the query sequence(s) and the unprocessed sequence database file(s) to be stored in Fasta format, or in the ASN.1 format (they should be processed using FORMATDB of the standalone BLAST package).

Motif Search

In these days of abundant sequence data, the majority of proteins have multiple homologues in the public databases; thus, the programs discussed above will usually return many related sequences. Still, it is often necessary to use more sensitive methods in order to locate homologues for which there are experimentally determined structural or functional data.

Position Specific Iterated (PSI)-BLAST

PSI-BLAST is an implementation of a *profile search* methodology that facilitates the retrieval

of distant homologues of a protein (Altschul and Koonin, 1998). It is currently only available for protein searches. It is easy to use, but it is also important to pay attention to the parameter settings and not blithely accept the defaults. PSI-BLAST works iteratively. It first conducts a regular protein–protein BLAST search using one of the standard substitution matrices applied equally to all positions. A multiple alignment is constructed from the set of sequences obtained and a position-specific substitution matrix (PSSM) is calculated. Using this PSSM, which is often called a *profile*, the database is searched again for more distant homologues. After several iterations, the process should converge (i.e. no more significant hits are discovered). The accidental inclusion of non-homologous sequences can have disastrous consequences for the rate of convergence and the reliability of the results. For this reason, between each iteration, the user is required to select which new sequences to include in the updated PSSM. These searches are typically much more sensitive than a standard BLAST search, which uses the same substitution penalties for each site. The PSSM, by contrast, is calculated from the frequencies of residues observed at each site in the alignment. Other motif-related variants of BLAST at NCBI include RPS-BLAST, which searches a database of PSSMs for domain matches to a query sequence and PHI-BLAST, which combines matching of a *regular expression* (see below) with a local alignment search for a query sequence.

QA2

Regular expressions are a flexible way to specify sequence motifs and other linear text patterns. The notation for regular expressions varies, but a simple convention for protein sequence motifs is to use square brackets indicating that a pattern may contain any of the enclosed amino acids, and curly brackets indicating that the pattern may have any residue except those enclosed at that position. For example, EL[GV]I{AN} would match ELVIS, but not ELGIN. Wildcards, repeats and variable spacing between residues can also be included in regular expressions.

Hidden Markov models (HMM) can also be used to describe protein sequence motifs (Eddy, 1998). Similar to a PSSM, but unlike a regular expression, the residues at each site may have varying probabilities of occurrence. Similar to a PSSM, the residues before and after the position in question will generally alter that probability, as would often be the case biologically. Another advantage of profile HMMs is that they allow for variable spacing between the residues of a motif. The HMMer program (Table 4.1) can be used to derive a profile HMM from a multiple sequence alignment (see below).

Multiple Sequence Alignment

Once a family of homologous sequences is in hand, the next step is typically to construct a multiple sequence alignment (MSA). MSA identifies equivalent positions in homologous sequences and is a powerful way to identify conserved (and thus likely functionally important) motifs in sequences. MSA is also a prerequisite to phylogenetic tree construction.

MSA is a far more difficult computational problem than pairwise alignment. Though methods for optimal MSA have been developed, they are too slow and memory intensive to be used on real data and do not allow for biologically realistic models of sequence evolution. Thus, all of the MSA algorithms used in practice are heuristic (none can guarantee an optimal alignment in the sense that the Smith–Waterman algorithm can). In addition, although local MSA algorithms exist, the most commonly used methods perform global alignments. This is often desirable and possible, since sets of closely related sequences are typically being compared. But, as a consequence, many MSA algorithms perform very poorly (i.e. return nonsense) when given distantly related sequences (especially in the so-called *twilight zone* of sequence identity below approximately 25%) and those of differing lengths.

The most commonly used heuristic for MSA is called *progressive alignment*. The first step in progressive alignment is to use distances obtained from pairwise Needleman–Wunsch alignments to construct a (very approximate) phylogeny known as a *guide tree*. The algorithm continues to completion by performing pairwise alignments between individual sequences or sets of aligned sequences—starting with the most closely related pairs and working down to the interior branches of the tree (Feng and Doolittle, 1987; Taylor, 1988; Thompson *et al*., 1994). Note that the guide tree generated by this part of the algorithm should not

be taken to be an estimate of the true phylogeny, it is merely a tool to facilitate sequence alignment (phylogenetic tree construction is discussed below). The most widely used implementation of a pure progressive alignment approach is ClustalW (Thompson *et al*., 1994).

Progressive alignment works fairly well and, of equal importance, it is relatively rapid. One flaw, however, is that mistakes made in the early stages of alignment cannot be rectified later after new sequences are added. By contrast, *iterative alignment* starts with a best-guess of the full MSA and then iteratively optimises it. Some implementations of iterative alignment (e.g. PPRP, Gotoh, 1996) perform well, in that they generally find alignments at or close to the optimum, but convergence tends to be very slow. T-Coffee (Notredame *et al*., 2000), a relatively recent method for MSA, incorporates features of both progressive and iterative alignment. T-Coffee appears to perform at least as well as (and often better than) the major progressive and iterative software packages in a variety of different situations.

The chief parameters to consider for MSA are the scoring matrix used and the gap penalty settings. The scoring matrix has already been discussed. The most common way to parameterise the gap penalties is to specify a gap opening penalty and a gap extension penalty. The gap opening penalty is the basic cost of any gap (in terms of a deduction from the alignment score) and the gap extension penalty is multiplied by the length of the gap. By adjusting these two parameters, one can steer the alignment towards many small gaps, a few large ones or somewhere in between. Normally, the gap extension penalty is a value much lower than the gap opening penalty.

While default values provide a useful starting point, one needs to carefully explore the sensitivity of the results to changes in these values. Figure 4.2 shows the differences between alignments for a set of *Arabidopsis* IAA proteins obtained with two software packages.

Phylogenetic Tree Construction

A phylogenetic tree is a model of evolutionary divergence events. These trees contain two types of information. The tree topology illustrates the order of divergence events and the branch lengths indicate the extent of sequence divergence (which is sometimes, but not always, proportional to time). Phylogenetic trees may be constructed using any type of biological data, but molecular sequence data have the advantage of being abundant, easily scored and compared, and undergo evolutionary change according to extensively studied and modelled processes. In genomics, it is often the case that what is of interest is the phylogeny of the sequences themselves, rather than the phylogeny of the organisms. Nonetheless, it is important to recognise that even if phylogenetic trees could always be inferred with perfect accuracy (which is not the case), the phylogenetic trees for genes sampled from a given set of taxa may vary from one gene to the next. This may be due to horizontal transfer, or, for closely related taxa, to the variable nature of gene genealogies in populations. In addition to inferring the order and timing of divergence events among sequences, molecular phylogenetic trees can provide a framework for asking a wide variety of evolutionary and functional questions about the sequences and the organisms that host them.

One can distinguish different kinds of homologues using molecular phylogenetics (Fitch, 1970). *Orthologues* are a subset of homologues where sequence divergence has occurred after a speciation event (i.e. the most recent common ancestor of the two sequences corresponds to the most recent common ancestor of the two species from which the sequences were obtained). With the few exceptions mentioned above, one can make the generalisation that the true phylogeny of a set of orthologues is the same as the true phylogeny of the taxa in which they are found. *Paralogues*, on the other hand, are homologues that have diverged after a gene duplication event and may co-exist in the same genome. While duplicated genes may be paralogues of each other, they can still both be orthologues (sometimes called semi-orthologues, or co-orthologues; Sharman, 1999; Taylor *et al*., 2001) of the corresponding gene in those lineages that diverged prior to the gene duplication event. It is generally more problematic to assume that paralogous genes (i.e. different lineages of a gene family) have conserved functions than it is for orthologous genes. Knowing orthology and paralogy within a gene family can

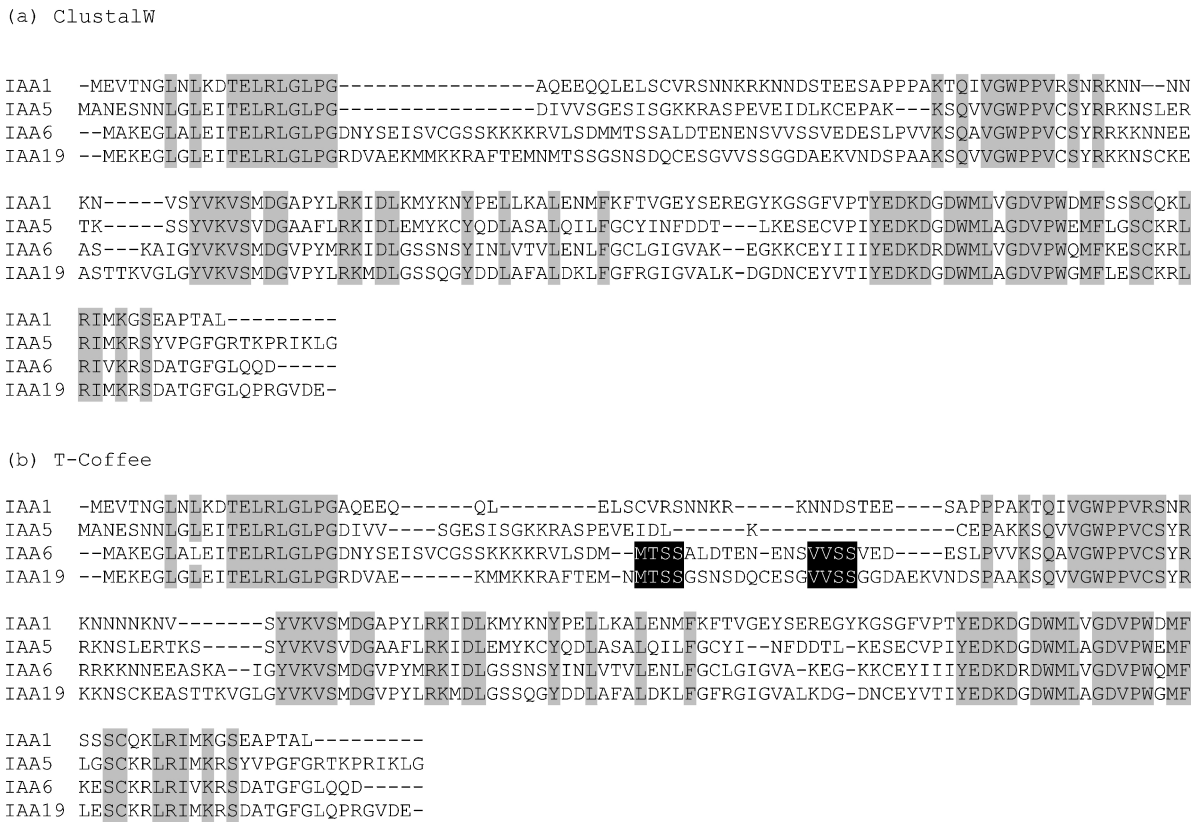


Fig. 4.2 Alignments generated using two different software programs (with default settings) for a set of indole acetic acid responsive proteins from *Arabidopsis thaliana* (IAA1: gi12644289; IAA5: gi1168608; IAA6: gi12484195; IAA19: gi17365900). (a) ClustalW, (b) T-Coffee. Apparent conserved residues are shaded grey. The most obvious difference between these two alignments is that T-Coffee appears to have much more reliable alignments in and around gaps (e.g. the residues shaded in black)

thus aid predictions of functional conservation. While alignment can be used to find homologues, only phylogenetics can distinguish orthologues from paralogues.

There are numerous methods for phylogenetic tree estimation. The most commonly used are *neighbour-joining*, *maximum parsimony* and *maximum likelihood*. We will describe these, as well as some more recent methods, although our treatment will necessarily be brief. The usual starting point for molecular phylogenetic inference is a multiple sequence alignment.

Some methods (e.g. neighbour-joining) use only the pairwise distance estimates obtained from the MSA, while others (maximum parsimony, maximum likelihood) explicitly model character state (residue) changes along the branches of the tree.

Neighbour-Joining

Saitou and Nei's (1987) neighbour-joining (NJ) algorithm is a heuristic method for obtaining a tree from a matrix of distance values according to the criterion of *minimum evolution* (in which the goal is the tree with the minimum sum of branch lengths). NJ is extremely rapid and can handle very large datasets. In addition, it is guaranteed to produce the minimum evolution tree if sequences are evolving in a clock-like fashion. Though NJ is *not* guaranteed to find the optimal tree when that condition does not hold, it is not as sensitive to branch length variation as other rapid tree-building methods.

NJ has been implemented by many programs, including the ClustalW and T-Coffee MSA

software packages. While the MSA guide tree (discussed above) is constructed from a provisional distance matrix based on pairwise alignments, and thus is not to be relied upon, the user may also specify that a (more reliable) tree be drawn using the distances obtained from the MSA itself.

QA2

Maximum Parsimony

The philosophy behind maximum parsimony (MP) is that the true evolutionary history of a set of genes can be traced through the path of fewest sequence residue (or some other heritable trait) changes. An algorithm referred to as *branch-and-bound* is guaranteed to find the optimal tree according to the MP criterion (Hendy and Penny, 1982), but it cannot be applied to datasets with larger than 20–30 sequences at present. Thus, one typically searches for the MP tree (or trees) using various quick-and-dirty strategies that evaluate some subset of tree space. A given search may or may not succeed in finding the maximum parsimony tree or trees. When one obtains many equally or similarly parsimonious trees, as is often the case, it is necessary to compute a *consensus tree* that collapses ambiguous branches. For computational efficiency, MP is generally applied in an unweighted fashion (i.e. all substitutions are assumed to occur with equal frequency). However, this is a problematic assumption for both DNA and protein sequences and can give very misleading results. MP is also a questionable criterion for choosing among alternative topologies because it does not adequately account for multiple hits. Because of this, in situations where there are two or more unrelated long branches in the true phylogeny, MP will provide ever-increasing support for a false relationship between those taxa as the amount of data increases (the so-called long branch attract phenomenon; Felsenstein, 1978a). MP is also fairly slow on large datasets, even when using heuristic search algorithms, because the number of possible trees grows to astronomical proportions very quickly as the number of sequences increases (Felsenstein, 1978b).

Maximum Likelihood

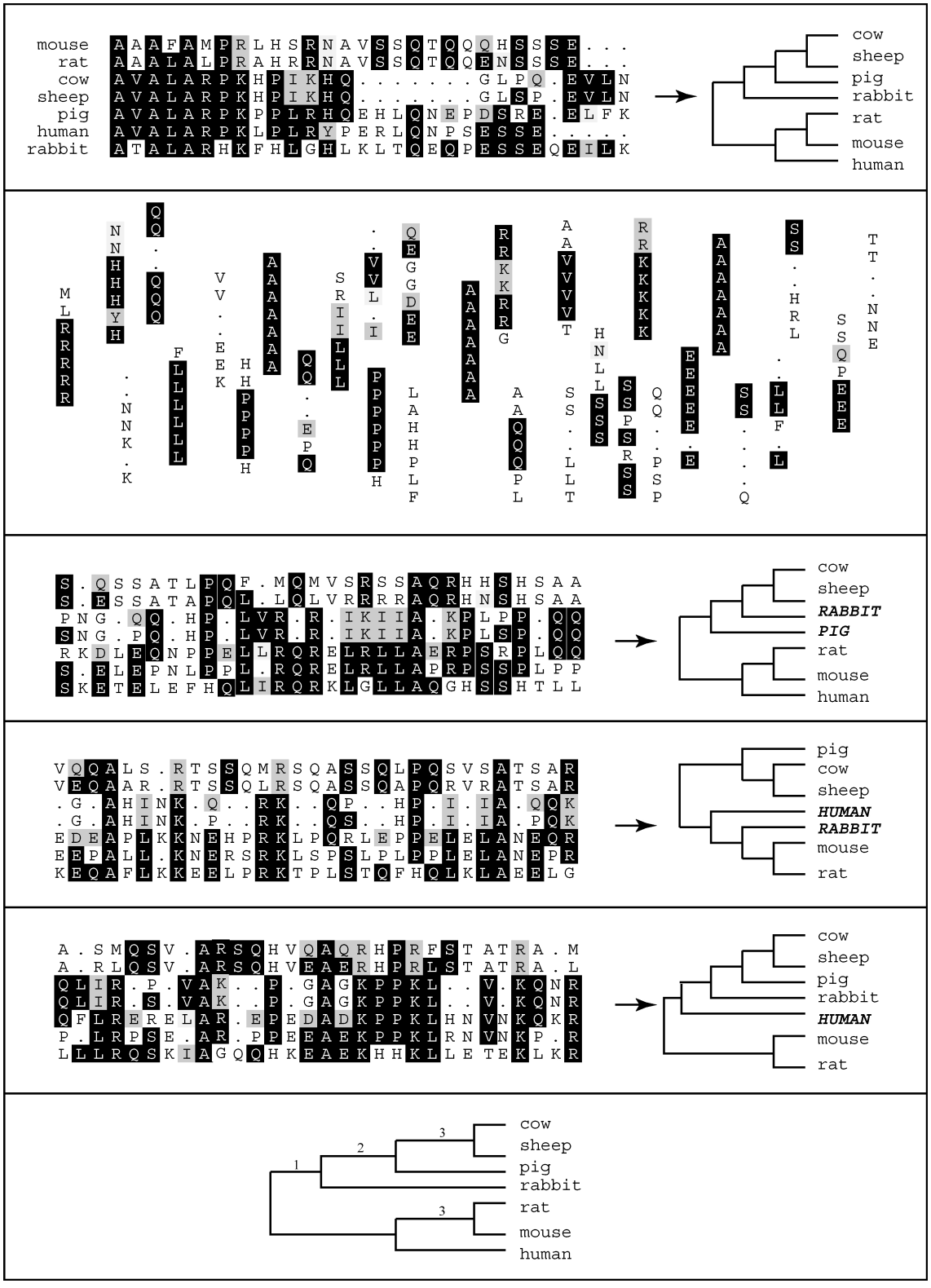
Maximum likelihood (ML) is an explicitly statistical criterion for tree construction (Felsenstein, 1981). The likelihood of a phylogenetic tree is the probability of observing the sequence data under a specific model of substitution and for a given tree topology, assuming both to be correct. The principle of ML is to find the topology with the highest likelihood. In order to do this, the algorithm must investigate the probability of all possible residues (amino acids, or nucleotides) for each position in the alignment at each of the internal nodes of the tree. Some residues will be more likely than others to give rise to the residues at the tips of the tree (i.e. the residues in the extant sequences) and therefore will contribute more to the likelihood of that topology. Since ML is only a criterion to choose among trees and does not provide a quick way to search for the best tree, it can be very computationally intensive, requiring examination of many different trees—as with MP. Recently, fast ML search algorithms have become available (e.g. Strimmer and von Haeseler, 1996).

Bayesian Methods

Closely related to ML methods are so-called Bayesian methods (Huelsenbeck *et al*., 2001). These are also computationally intensive, but because of algorithmic differences, can be used to infer trees for very large datasets (Karol *et al*., 2001). Bayesian methods also provide a natural way to measure statistical support for particular relationships within the tree. For other methods, support can only be approximated by bootstrapping (see below). The results from both ML and Bayesian methods are dependent on the quality of the substitution model, which specifies the probability of change between all character states and may also allow for such features as rate variation among sites (see below). Software is available for comparing the fit of different models of substitution (Posada and Crandall, 1998).

Rooting Trees

It is necessary to *root* a phylogenetic tree before inference can be made about the relative order of branching events, but identifying the root (i.e. the branch ancestral to all sequences) can sometimes be a challenge.


mouse
rat
cow
sheep
pig
human
rabbit
cow
sheep
pig
rabbit
rat
mouse
human
cow
sheep
RABBIT
PIG
rat
mouse
human
pig
cow
sheep
HUMAN
RABBIT
mouse
rat
cow
sheep
pig
rabbit
HUMAN
mouse
rat
1
2
3
3
cow
sheep
pig
rabbit
rat
mouse
human

Quick and dirty methods place the root at the midpoint of the tree or along the longest internal branch. In cases where rates of substitution are non clock-like, these are a poor choice. A preferred method is to include an *outgroup* sequence (or sequences) in the analysis. This should be a related sequence, or set of them, known to have diverged before the *ingroup* sequences under study. The branch connecting the outgroup(s) to the rest of the tree can then be taken as the root. The challenge is to identify a sequence that is sufficiently divergent to be confident that it really is an outgroup yet not so divergent that it is difficult to determine its point of attachment to the tree or that it cannot be aligned with the other sequences.

Bootstrapping

Quite often there is not enough information in an alignment to resolve some or all of the relationships of the sequences under study with confidence. This may happen if the sequences are all very similar or if they are sufficiently divergent that many sites are saturated (have undergone multiple substitutions along each branch). In other words, not all positions in an alignment are informative of the evolutionary relationship of the sequences. In cases where there are very few informative sites, there will be low statistical support for the tree.

Bootstrapping (Felsenstein, 1985) is a resampling technique used to calculate the degree of statistical support for each branch (Figure 4.3) and is the *de facto* standard method for assessing confidence in trees obtained via NJ, MP and ML. It works by resampling (with replacement) positions from the original alignment to construct a new alignment of the same length as the original. A tree is then inferred from the resampled alignment. This is typically done for 500–1000 replicates. The bootstrap value for a given branch is the proportion of replicates in which the same set of taxa descended from that branch. If only very few sites of the original alignment support a branch, then it will be rarely seen among the replicates and will therefore have a low bootstrap value. As a general rule, a bootstrap value of 80% or higher is considered strong support for a branch.

It is common practice, when reporting the results of a phylogenetic analysis, to merge the nodes on either end of those branches that have low bootstrap support (Berry and Gascuel, 1996). The resultant tree will then contain some nodes that are *multifurcating* (having three or more descendent branches), rather than *bifurcating* (having exactly two descendent branches).

Special Considerations

Alignment Gaps

Gaps (or *indels*, short for insertions/deletions) can be difficult to incorporate into phylogenetic analysis. If one were willing to assume that a contiguous gap represents a single mutational event, it would be appropriate to score it as a single binary character (present or absent). However, models that incorporate both substitutions and gaps are yet to be widely accepted. In addition, multiple sequence alignments tend to be unreliable in the neighbourhood of gaps. Thus, it is recommended to exclude positions with gaps when inferring a phylogenetic tree from an MSA. Most phylogenetic software packages provide it as an option. The common exception to this is in unweighted MP, where gaps are treated as a fifth base or an extra amino acid residue.

Multiple Hits

When the time since divergence of two sequences is short and/or the rate of evolution is slow, then

Fig. 4.3 The bootstrap procedure to evaluate the statistical support for the branches in a phylogenetic tree topology. (a) The original alignment and the tree generated from that alignment. (b) For the purposes of bootstrap analysis, all positions in the alignment are considered to be independent of one another and constitute the sample space for bootstrap sampling. (c) Bootstrap alignments are generated by randomly sampling, with replacement, k positions from the original alignment of length k. In each bootstrap alignment, a given site may be sampled more than once while others may not be sampled at all. A phylogenetic tree is constructed from each individual bootstrap alignment. (d) and (e) Different bootstrap samples and corresponding trees. (f) The original tree with bootstrap values on each of the branches. The bootstrap value is the proportion of bootstrap trees in which an identical branch was found

the chance of more than one substitution having occurred at the same site is negligible. But, when divergence times are longer or rates are faster, this chance increases and must be taken into consideration when calculating evolutionary distance. The observable number of substitutions is almost always considerably less than the actual number of substitutions that have taken place. There are several different methods of correcting for multiple hits in DNA sequences (Jukes and Cantor, 1969; Uzzell and Corbin, 1970; Kimura, 1980) and protein sequences (Dayhoff, 1978; Henikoff and Henikoff, 1992). One must explicitly perform this correction when calculating pairwise distances for a method such as NJ. However, the substitution matrices used by ML and Bayesian methods inherently correct for multiple substitutions.

The simplest correction methods are based on models in which the probability of substitution is equal over sites. The gamma correction for multiple hits (Uzzell and Corbin, 1970; Yang, 1996; Gu and Zhang, 1997) is more realistic in that it allows for rate variation among sites (as may happen when sites differ in their degrees of selective constraint). However, the gamma correction is not implemented in every tree inference package and, because the shape parameter for the gamma distribution must be estimated from the data, it may not be appropriate for small datasets. An alternative approach is to model some fraction of the sites as invariant while allowing the others to substitute with equal probabilities.

Silent and Replacement Substitutions

Due to the degeneracy of the genetic code, some fraction of DNA base substitutions within protein coding regions are *silent* (or *synonymous*), in that they do not lead to an alteration in the protein sequence, while some lead to an amino acid *replacement* (and are therefore *nonsynonymous*). For a protein sequence under no selective constraint, the ratio of silent to replacement substitutions, normalised by the number of silent and replacement sites, has an expectation of one. A change in this ratio along a protein coding sequence, or among lineages within a protein family, can be informative about the forces of functional constraint and selection acting on the sequence. Sophisticated methods for sequence analysis have been developed based upon this simple principle. Many of these can be implemented in the PAML and HyPHY software packages (Table 4.1). An example of this type of analysis for NBS-LRR genes is discussed below.

Phylogenetics Software

PHYLIP

The PHYLogeny Inference Package (Table 4.1) provides numerous software modules for a wide variety of phylogenetic methods. PHYLIP includes: NEIGHBOR, for NJ trees, PROTPARS and DNAPARS for MP trees (from protein or DNA alignments, respectively) and PROTML and DNAML for ML trees. PHYLIP can also be accessed through a user friendly web interface hosted by the Institut Pasteur. PHYLIP is widely used and comes with extensive documentation. In addition, the author maintains a comprehensive list of other phylogenetics software at the Phylip website.

Drawing Trees

The most commonly used way to represent trees in computer files is the Newick format, in which nested pairs of parentheses indicate which taxa are descended from each branch in the tree. Numbers within the parentheses indicate branch lengths. This format has the advantage of allowing the tree image to be represented only by text so that it is (i) independent of any image format and (ii) takes up negligible hard-disk space. However, Newick format is not easily read by humans. Several freely available software tools allow one to convert any Newick formatted tree into a print-quality image, including Phylodendron and TreeView (Table 4.1).

Other Analysis Tools

Gene Expression Analysis

Unlike the other areas we have discussed so far, software for highly parallel gene expression analysis is predominantly commercial in nature. Still, there are a few webservers that provide standard

analysis tools (e.g. ep.ebi.ac.uk/EP/EPCLUST/ and bioinfo.cnio.es/dnarray/). In addition, several academic laboratories and research institutes offer their software packages for download (a useful list is maintained at ep.ebi.ac.uk/Links.html). Many of the challenges associated with microarray expression data concern image processing. The positions and boundaries of the spots must be determined, the background signal must be subtracted, the intensity values must be normalised to account for differences in loading and dye strength and a host of other factors. Once these substantial tasks are accomplished, the typical downstream analyses that are done with microarray expression data (or other highly parallel expression techniques) fall into several classes: detecting which genes are differentially regulated between treatments, identifying clusters of genes with similar transcriptional profiles and classification of different samples on the basis of transcriptional profile. There are many recent reviews of these issues (e.g. Quackenbush, 2001; Sherlock, 2001).

Protein Structural Analysis

The structure of a protein is crucial to its function. As a general rule, substitutions which alter the structure of a protein will tend to alter its function in some way while substitutions that leave the structure unchanged are usually neutral (Chothia and Lesk, 1986). While *ab initio* secondary structure prediction (i.e. prediction of alpha helices, beta sheets and turns) can be done with a great deal of accuracy, tertiary structure prediction is one of the major open problems in computational biology. Prediction of the tertiary structures of a protein is greatly facilitated if there is a homologous sequence for which the 3D crystal structure has been experimentally determined (reviewed by Baker and Sali, 2001). The website of the Sali lab at Rockefeller University (guitar.rockefeller.edu/bioinformatics_resources.shtml) has links to many different comparative protein structural analysis tools, including meta-servers that provide centralised web interfaces for multiple algorithms.

In the interests of space, we must pass over several major areas of sequence analysis, such as sequence assembly and gene prediction, but a list of more comprehensive bioinformatics references is provided at the end of this chapter.

SPECIAL CONSIDERATIONS FOR PLANT BIOINFORMATICS

The plants for which the most genomic data are available are, not surprisingly, the major crops (rice, maize, soya bean and tomato) plus a handful of model organisms such as *Arabidopsis thaliana* and *Medicago truncatula*. Although many of these fall into a limited set of plant families (particularly, the grasses, legumes, mustards and nightshades), there are also a number of orphan species of economic interest, such as cotton, banana and the citrus crops. This taxonomic diversity of study organisms presents both challenges and opportunities to plant genomics and bioinformatics. On the one hand, genomic information is woefully incomplete for many important systems. On the other hand, it is sometimes possible to use data from related plant species to address this deficiency. This use of comparative genomic methodology is more easy in some domains than in others, and generally works best when investigating those fundamental cellular processes that are conserved among plants. A classic case is flower development, for which models formulated in *Arabidopsis* and *Antirrhinum* provide insight into the developmental genetics of angiosperms in general (Shepard and Purugganan, 2002). With respect to details of biology specific to each organism (the production of lint in cotton, or xylem in poplars), it is more difficult to take advantage of information obtained from a model such as *Arabidopsis*.

Another consideration is that many plant genomes are complex in the sense that they have much repetitive DNA and a high degree of duplication even in genic regions (due to polyploidy and other processes). This complexity poses a challenge to many kinds of genomic analyses, among them: (i) the assembly of genomic sequences (because of spurious matches between repeat units); (ii) the computational prediction of protein-coding genes, intronic splice sites and functionally important non-coding features and (iii) the analysis of mutant phenotypes (due to increased functional redundancy). While it is only a partial solution, comparative sequence analysis can help to identify those sequences that

are conserved across taxa and thus likely to be functionally important (e.g. Koch *et al*., 2001).

BIOINFORMATICS-BASED DISCOVERIES IN PLANT GENOMES

Some of the most important contributions bioinformatics has made to plant biology have been as part of larger projects in which experimental and computational studies were closely intertwined. Nevertheless, it is possible to point to studies in which bioinformatics played a critical role in connecting ideas or testing hypotheses.

Comparative Mapping

Comparative mapping centres on the identification of homologous chromosomal regions based upon the conservation of linkage relationships among homologous markers. In the late 1990s, the availability of extensive sequence information from *Arabidopsis* motivated a number of researchers working on genetic mapping in crops to start using *Arabidopsis* as a point of comparison. Such comparative maps would allow researchers to exploit gene content information from *Arabidopsis* to provide candidate genes for traits mapped to homologous regions in other species. Unlike the earlier generation of comparative maps, which generally involved closely related species within a family (e.g. Tanksley *et al*., 1992), non-sequence based molecular markers could no longer be used for such distant comparisons. Thus, a bioinformatics approach must be used. Sequences from the crop species are characterised and mapped using traditional linkage analysis. The orthologous sequence is then mapped *in silico* to the *Arabidopsis* genome using available data. For example, this strategy has been employed in the development of a tomato-*Arabidopsis* comparative map, in which approximately 1000 markers have been developed to identify regions of chromosomal homology (described at the website of the Solanaceae Genome Network, Table 4.1). An unexpected outcome of such comparative mapping studies has been the discovery of multiple, ancient, large-scale genome duplications in the lineage that gave rise to *Arabidopsis* and many other eudicot species (Grant *et al*., 2000; Ku *et al*., 2000; Vision *et al*., 2000; Mayer *et al*., 2001).

Evolution of Disease Resistance Genes

Many disease resistance proteins in plants have a very characteristic domain architecture consisting of a nucleotide binding site (NBS) region and a leucine-rich region (LRR). Other domains, chiefly related to intercellular signalling, typify different subfamilies of NBS-LRR proteins. The evolution of new recognition specificities in these proteins is an area of considerable practical interest. Evolutionary theory predicts that genes evolving under divergent selection (also known as positive selection), where there is continual pressure for new functionality such as recognition specificity, will show an accelerated rate of replacement relative to silent DNA substitution (as discussed above). Statistical methods have been developed to estimate these rates on a per-site basis so that the number of silent and replacement substitutions can be directly compared. Using these measures, one can test whether the ratio is significantly greater than one, which would be indicative of positive selection (Hughes, 2000). Application of these methods to genes in the NBS-LRR class has revealed that hypervariable solvent-exposed residues in the leucine-rich-repeat region have elevated ratios of replacement to silent substitutions (Michelmore and Meyers, 1998). This suggests that they have evolved to detect variation in pathogen-derived ligands.

In addition, NBS-LRR proteins are frequently clustered within the genome in tandem or near-tandem arrays. Because of this, it was thought for some time that new resistance specificities evolved via unequal crossing-over and gene conversion between tandemly arrayed paralogues. However, phylogenetic studies of the genes in several of the clusters (such as the *Pto* and *Cf4/9* clusters of tomato) indicate that this is not generally the case, because the alleles are more closely related within than between the tandemly arranged loci. Thus, computational techniques from the field of molecular evolution have provided important insights into the dynamics of resistance gene evolution. (Michelmore and Meyers, 1998).

Organellar–Nuclear Gene Transfer

Multiple genomes co-exist within all eukaryotic cells. In plants, there are three genomes: nuclear,

mitochondrial and plastid. The mitochondrion and plastid are believed to be descended from an alpha proteo-bacterial and a cyanobacterial endosymbiont, respectively (Margulis, 1970; Lang *et al*., 1999; McFadden, 2001). The genomes of these erstwhile free-living organisms have not remained static. Mitochondrial and plastid genomes now contain only a subset of those genes thought to have been present in their free-living ancestors; most of the genes that remain are involved directly in organelle function. An unexpected discovery revealed by recent phylogenetic studies is that parallel gene transfers to the nucleus among different plant lineages have occurred more often than have single unique transfer events–particularly in certain genes (Martin *et al*., 1998; Adams *et al*., 2000; Millen *et al*., 2001). In some cases the protein products of the transferred genes continue to function within the organelle. In these, the change from organellar to nuclear encoding has entailed the acquisition of organelle targeting signal peptides (chloroplast or mitochondrial)—obtained from another gene in the genome and prepended to the ancestral protein sequence. The resulting proteins have hybrid phylogenetic signals reflecting the different origins of the two parts of the protein.

FUTURE PROSPECTS

Computers play an increasingly important role in biology. In fact, it is likely that future generations of biologists will perform many of their experiments *in silico*. This perhaps has a parallel in the early years of molecular genetics, when classical studies of phenotypic inheritance came to be enriched by the powerful tools of molecular biology. The tools of bioinformatics and computational biology are to us now what molecular biology was then. As researchers continue to accumulate large amounts of biological information, we anticipate that a lot of biological insight will be gained from the innovative application of computational methods.

ACKNOWLEDGMENTS

The authors wish to thank Andrew Lloyd and Sam Cartinhour for helpful comments and advice.

REFERENCES

Adams KL, Daley DO, Qiu YL, Whelan J, Palmer JD (2000). Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. *Nature* **16**: 354–7.

Altschul SF, Gish W (1996). Local alignment statistics. *Meth Enzym* **266**: 460–80.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–10.

Altschul SF, Koonin EV (1998). Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem Sci* **23**: 444–7.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–402.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* **29**: 37–40.

Attwood TK, Blythe MJ, Flower DR, Gaulton A, Mabey JE, Maudling N, McGregor L, Mitchell AL, Moulton G, Paine K, Scordis P (2002). PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res* **30**: 239–41.

Bairoch A, Apwelier R (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**: 45–8.

Baker D, Sali A (2001). Protein structure prediction and structural genomics. *Science* **294**: 93–6.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002). The Pfam protein families database. *Nucleic Acids Res* **30**: 276–80.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2002). Genbank. *Nucleic Acids Res* **30**: 17–20.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank. *Nucleic Acids Res* **28**: 235–42.

Berry V, Gascuel O (1996). On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. *Mol Biol and Evol* **13**: 999–1011.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001). Minimum information about a microarray experiment (MIAME): toward standards for microarray data. *Nature Genet* **29**: 365–71.

Capela D, Barloy-Hubler F, Gouzy J, Bothe G, Ampe F, Batut J, Boistard P, Becker A, Boutry M, Cadieu E, Dreano S, Gloux S, Godrie T, Goffeau A, Kahn D, Kiss E, Lelaure V, Masuy D, Pohl T, Portetelle D, Puehler A, Purnelle B, Ramsperger U, Renard C, Thebault P, Vandenbol M, Weidner S, Galibert F (2001). Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti*. *Proc Natl Acad Sci USA* **98**: 9877–982.

Chothia C, Lesk AM (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**: 823–6.

Corpet F, Servant F, Gouzy J, Kahn D (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* **28**: 267–9.

Dayhoff MO, Schwartz RM, Orcutt BC (1978). A model of evolutionary change in proteins. pp 345–52 In Dayhoff MO (ed) *Atlas of Protein Sequence and Structure, Vol 5, Suppl 3*. National Biomedical Research Foundation, Washington DC.

Durbin R, Thierry-Mieg J (1991) A *C. elegans* database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.

Eddy S (1998). Profile hidden Markov models. *Bioinformatics* **14**: 755–63.

Edgar R, Domrachev M, Lash AE (2002): Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**: 207–10.

Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A (2002). The PROSITE database, its status in 2002. *Nucleic Acids Res* **30**: 235–8.

Felsenstein J (1978a). Cases in which parsimony or compatability methods will be positively misleading. *Systematic Zoology* **27**: 401–10.

Felsenstein J (1978b). The number of evolutionary trees. *Systematic Zoology* **27**: 27–33.

Felsenstein J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. **17**: 368–76.

Felsenstein J (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–91.

Feng DF, Doolittle RF (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* **25**: 351–60.

Fitch WM (1970). Distinguishing homologous from analogous proteins. *Syst Zool* **19**: 99–113.

Fitch WM (2000). Homology a personal view on some of the problems. *Trends Genet* **16**: 227–31.

The Gene Ontology Consortium (2001). Gene ontology: tool for the unification of biology. *Nature Genet* **25**: 25–9.

Gibrat JF, Madej T, Bryant SH (1996). Surprising similarities in structure comparison. *Curr Opin Struct Biol* **6**: 377–85.

Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, Slater S (2001). Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294**: 2323–8.

Gotoh O (1996). Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol* **264**: 823–38.

Grant D, Cregan P, Shoemaker RC (2000). Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *Proc Natl Acad Sci USA* **97**: 4168–73.

Gu X, Zhang J (1997). A simple method for estimating the parameter of substitution rate variation among sites. *Molecular Biology and Evolution* **14**: 1106–13.

Hendy MD, Penny D (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* **60**: 133–42.

Henikoff S, Henikoff JG (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**: 10915–9.

Henikoff S, Henikoff JG, Pietrokovski S (1999). Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **5**: 471–9.

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999). Plant *cis*-acting regulatory DNA elements (PLACE) database: *Nucleic Acids Res* **27**: 297–300.

Hokamp K, Wolfe K (1999). What's new in the library? What's new in Genbank? Let PubCrawler tell you. *Trends Genet* **15**: 471–2.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**: 2310–4.

Hughes AL (2000). *Adaptive Evolution of Genes and Genomes*. Oxford University Press.

Ilipoulis I, Enright AJ, Ouzounis CA (2001). Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. *Pac Symp Biocomput* **2001**: 384–95.

Karol KG, McCourt RM, Cimino MT, Delwiche CF (2001). The closest living relatives of land plants. *Science* **294**: 2351–3.

Kimura M (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J Mol Evolut* **16**: 111–20.M.A.

Koch MA, Weisshaar B, Kroymann J, Haubold B, Mitchell-Olds T (2001). Comparative genomics and regulatory evolution: conservation and function of the *Chs* and *Apetala3* promoters. *Mol Biol Evol* **18**: 1882–91.

Kolpakov FA, Ananko EA, Kolesov GB, Kolchanov NA (1998). GeneNet: a database for gene networks and its automated visualization. *Bioinformatics* **14**: 529–37.

Koski LB, Golding BG (2001). The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* **52**: 540–2.

Ku HM, Vision T, Liu J, Tanksley SD (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc Natl Acad Sci USA* **97**: 9121–6.

Lang BF, Gray MW, Burger G (1999). Mitochondrial genome evolution and the origin of eukaryotes. *Annu Rev Genet* **33**: 351–97.

Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S (2002). PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res* **30**: 325–7.

Letondal C (2001). A Web interface generator for molecular biology programs in Unix, *Bioinformatics* **17**: 73–82.

Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res* **30**: 242–4.

Li WH (1993). So, what about the molecular clock hypothesis? *Curr Opin Gen Dev* **3**: 896–901.

Lonsdale D, Crowe M, Arnold B, Arnold BC (2001). Mendel-GFDb and Mendel-ESTS: databases of plant gene families

and ESTs annotated with gene family numbers and gene family names. *Nucleic Acids Res* **29**: 120–2.

Margulis L (1970). *Origin of Eukaryotic Cells*. Yale University Press, New Haven, CT, USA.

Martin W, Stoebe B, Goremykin V, Hapsmann S, Hasegawa M, Kowallik KV (1998). Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**: 162–5.

Mayer K, Murphy G, Tarchini R, Wambutt R, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N, Lemcke K, Haase D, Hall CR, van Dodeweerd AM, Tingey SV, Mewes HW, Bevan MW, Bancroft I (2001). Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana*. *Genome Res* **11**: 1167–74.

McFadden GI (2001). Chloroplast origin and integration. *Plant Physiol* **125**: 50–3.

Michelmore RW, Meyers BC (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res* **11**: 1113–30.

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH (2001). Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* **13**: 645–58.

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structuRes *J Mol Biol* **247**: 536–40.

Needleman SB, Wunsch CD (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–53.

Notredame C, Higgins D, Heringa J (2000). T-Coffee: A novel method for multiple sequence alignments. *J Mol Biol* **302**: 205–17.

The Plant Ontology Consortium The Plant Ontology Consortium and Plant Ontologies. *Comp Func Genom* (in press).

Pearl FMG, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM, Orengo CA (2000). Assigning genomic sequences to CATH. *Nucleic Acids Res* **28**: 277–82.

Pearson WR (2000). Flexible sequence similarity searching with the Fasta3 program package. *Methods Mol Biol* **132**: 185–219.

Posada D, Crandall KA (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–8.

Price CA, Reardon EM. Related Articles (2001). Mendel, a database of nomenclature for sequenced plant genes. *Nucleic Acids Res* **29**: 118–9.

Quackenbush J (2001). Computational analysis of microarray data. *Nature Rev Genet* **2**: 418–27.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res* **29**: 159–64.

Reeck GR, de Haen C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, Zuckerkandl E (1987). "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* **50**: 667.

Rice P, Longden I, Bleasby A (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–7.

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406–25.

Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, Arlat M, Billault A, Brottier P, Camus JC, Cattolico L, Chandler M, Choisne N, Claudel-Renard C, Cunnac S, Demange N, Gaspin C, Lavie M, Moisan A, Robert A, Saurin W, Schiex T, Siguier P, Thebault P, Whalen M, Wincker P, Levy M, Weissenbach J, Boucher CA (2002). Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415**: 497–502.

Schuler GD, Epstein JA, Ohkawa H, Kans JA (1996). Entrez: molecular biology database and retrieval system. *Methods Enzymol* **266**: 141–62.

Sharman AC (1999). Some new terms for duplicated genes. *Semin Cell Dev Biol* **10**: 561–3.

Shepard KA, Purugganan MD (2002). The genetics of plant morphological evolution. *Curr Opin Plant Biol* **5**: 49–55.

Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM (2001). The Stanford Microarray Database. *Nucleic Acids Res* **29**: 152–5.

Sherlock G (2001). Analysis of large-scale gene expression data. *Brief Bioinform* **2**: 350–62.

Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, Alves LM, Araya JE, Baia GS, Baptista CS, Barros MH, Bonaccorsi ED, Bordin S, Bove JM, Briones MR, Bueno MR, Camargo AA, Camargo LE, Carraro DM, Carrer H, Colauto NB, Colombo C, Costa FF, Costa MC, Costa-Neto CM, Coutinho LL, Cristofani M, Dias-Neto E, Docena C, El-Dorry H, Facincani AP, Ferreira AJ, Ferreira VC, Ferro JA, Fraga JS, Franca SC, Franco MC, Frohme M, Furlan LR, Garnier M, Goldman GH, Goldman MH, Gomes SL, Gruber A, Ho PL, Hoheisel JD, Junqueira ML, Kemper EL, Kitajima JP, Krieger JE, Kuramae EE, Laigret F, Lambais MR, Leite LC, Lemos EG, Lemos MV, Lopes SA, Lopes CR, Machado JA, Machado MA, Madeira AM, Madeira HM, Marino CL, Marques MV, Martins EA, Martins EM, Matsukuma AY, Menck CF, Miracca EC, Miyaki CY, Monteriro-Vitorello CB, Moon DH, Nagai MA, Nascimento AL, Netto LE, Nhani A Jr, Nobrega FG, Nunes LR, Oliveira MA, de Oliveira MC, de Oliveira RC, Palmieri DA, Paris A, Peixoto BR, Pereira GA, Pereira HA Jr, Pesquero JB, Quaggio RB, Roberto PG, Rodrigues V, de M Rosa AJ, de Rosa VE Jr, de Sa RG, Santelli RV, Sawasaki HE, da Silva AC, da Silva AM, da Silva FR, da Silva WA Jr, da Silveira JF, Silvestri ML, Siqueira WJ, de Souza AA, de Souza AP, Terenzi MF, Truffi D, Tsai SM, Tsuhako MH, Vallada H, Van Sluys MA, Verjovski-Almeida S, Vettore AL, Zago MA, Zatz M, Meidanis J, Setubal JC (2000). The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature* **406**: 151–7.

Smith TF, Waterman MS (1981). Identification of common molecular subsequences. *J Mol Biol* **147**: 195–7.

Strimmer K, von Haeseler A (1996). Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol Biol Evol* **13**: 964–9.

Tanksley SD, Ganal MW, Prince JP, de Vicente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, Grandillo S, Martin GB, Messeguer R, Miller JC, Miller L,

Paterson AH, Pineda O, Ršder M, Wing RA, Wu W, Young ND (1992). High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132**: 1141–60.

Taylor S, Van de Peer Y, Braasch I, Meyer A (2001). Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philosl Trans R Soc Lond BBiol Sci* **356**: 1661–79.

Taylor WR (1988). A flexible method to align large numbers of biological sequences. *J Mol Evol* **28**: 161–9.

Thompson JD, Higgins DG, Gibson TJ (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**: 4673–80.

Wang Y, Anderson JB, Chen J, Geer LY, He S, Hurwitz DI, Liebert CA, Madej T, Marchler GH, Marchler-Bauer A, Panchenko AR, Shoemaker BA, Song JS, Thiessen PA, Yamashita RA, Bryant SH (2002). MMDB: Entrez's 3D-structure database. *Nucleic Acids Res* **30**: 249–52.

Uzzell T, Corbin KW (1971). Fitting discrete probability distributions to evolutionary events. *Science* **172**: 1089–96.

Vision TJ, Brown DG, Tanksley SD (2000). The origins of genomic duplications in *Arabidopsis*. *Science* **290**: 2114–7.

Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S, McCouch S, Stein L (2002). Gramene: a resource for comparative grass genomics. *Nucleic Acids Res* **30**: 103–5.

Watson JD, Crick FHC (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**: 737–8.

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüβ M, Reuter I, Schacherer F (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**: 316–9.

Wootton J, Federhen S (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266**: 554–71.

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S, Eisenberg D (2002). DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**: 303–5.

Yang Z (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* **11**: 367–72.

Zhu H, Choi S, Johnston AK, Wing RA, Dean RA (1997). A large-insert (130 kbp) bacterial artificial chromosome library of the rice blast fungus *Magnaporthe grisea*: genome analysis, contig assembly, and gene cloning. *Fungal Genet Biol* **21**: 337–47.

Zuckerkandl E, Pauling L (1965). Evolutionary divergence and convergence in proteins. In Bryson V, Vogel HJ (eds). *Evolving Genes and Proteins*. Academic Press, New York, pp 97–165.

BIBLIOGRAPHY

We list here some reading material for the biologist who desires to learn more about genomic databases and bioinformatic tools:

Baxevanis A, Ouellette F (2001). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 2nd edn. John Wiley and Sons: Cambridge, UK.

Durbin R, Eddy S, Krogh A, Mitchison G (1998). *Biological Sequence Analysis*. Cambridge University Press: Cambridge, UK.

Gibas C, Jambeck P (2001). *Developing Bioinformatics Computer Skills*. O'Reilly & Associates: Sebastopol, CA, USA.

Hall BG (2000). *Phylogenetic Trees Made Easy: A How-To Manual for Molecular Biologists*. Sinauer Associates: Cambridge, MA, USA.

Liu BH (1998). *Statistical Genomics: Linkage, Mapping and QTL Analysis*. CRC Press: Boca Raton FL, USA.

Mount DW (2001). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press: Plainview, NY, USA.

QUERIES TO THE AUTHOR

QA1. Please check 'bioperl**b**' ok?
QA2. Please specify the correct section.