5

STRUCTURE DATABASES

Christopher W. V. Hogue

Samuel Lunenfeld Research Institute
Mount Sinai Hospital
Toronto, Ontario, Canada

INTRODUCTION TO STRUCTURES

This chapter introduces biomolecular structures from a bioinformatics perspective, with special emphasis on the sequences that are contained in three-dimensional structures. The major goal of this chapter is to inform the reader about the contents of structure database records and how they are treated, and sometimes mistreated, by popular software programs. This chapter does not cover the computational processes used by structural scientists to obtain three-dimensional structures, nor does it discuss the finer points of comparative protein architecture. Several excellent monographs regarding protein architecture and protein structure determination methods are already widely available and often found in campus bookstores (e.g., Branden and Tooze, 1999).

The imagery of protein and nucleic acid structures has become a common feature of biochemistry textbooks and research articles. This imagery can be beautiful and intriguing enough to blind us to the experimental details an image represents—the underlying biophysical methods and the effort of hard-working X-ray crystallographers and *nuclear magnetic resonance* (NMR) spectroscopists. The data stored in structure database records represents a practical summary of the experimental data. It is, however, not the data gathered directly by instruments, nor is it a simple mathematical transformation of that data. Each structure database record carries assumptions and biases that change as the state of the art in structure determination advances. Nevertheless, each biomolecular structure is a hard-won piece of crucial information and provides potentially critical information regarding the function of any given protein sequence.

Since the first edition of this book was released, the software for viewing and manipulating three-dimensional structures has improved dramatically. Another major change has come in the form of an organizational transition, with the Protein Data Bank (PDB) moving from the Brookhaven National Laboratories to the Research Collaboratory for Structural Biology. The result has been a complete change in the organization of the PDB web site. The impact of these changes for biologists will be discussed herein.

The Notion of Three-Dimensional Molecular Structure Data

Let us begin with a mental exercise in recording the three-dimensional data of a biopolymer. Consider how we might record, on paper, all the details and dimensions of a three-dimensional ball-and-stick model of a protein like myoglobin. One way to begin is with the sequence, which can be obtained by tracing out the backbone of the three-dimensional model. Beginning from the NH_2-terminus, we identify each amino acid side chain by comparing the atomic structure of each residue with the chemical structure of the 20 common amino acids, possibly guided by an illustration of amino acid structures from a textbook.

Once the sequence has been written down, we proceed with making a two-dimensional sketch of the biopolymer with all its atoms, element symbols, and bonds, possibly taking up several pieces of paper. The same must be done for the heme ligand, which is an important functional part of the myoglobin molecule. After drawing its chemical structure on paper, we might record the three-dimensional data by measuring the distance of each atom in the model starting from some origin point, along some orthogonal axis system. This would provide the *x*-, *y*-, and *z*-axis distances to each atomic ''ball'' in the ball-and-stick structure.

The next step is to come up with a bookkeeping scheme to keep all the (*x*, *y*, *z*) coordinate information connected to the identity of each atom. The easiest approach may be to write the (*x*, *y*, *z*) value as a coordinate triple on the same pieces of paper used for the two-dimensional sketch of the biopolymer, right next to each atom. This associates the (*x*, *y*, *z*) value with the atom it is attached to.

This mental exercise helps to conceptualize what a three-dimensional structure database record ought to contain. There are two things that have been recorded here: the chemical structure and the locations of the individual atoms in space. This is an adequate ''human-readable'' record of the structure, but one probably would not expect a computer to digest it easily. The computer needs clear encoding of the associations of atoms, bonds, coordinates, residues, and molecules, so that one may construct software that can read the data in an unambiguous manner. Here is where the real exercise in structural bioinformatics begins.

Coordinates, Sequences, and Chemical Graphs

The most obvious data in a typical three-dimensional structure record, regardless of the file format in use, is the *coordinate data*, the locations in space of the atoms of a molecule. These data are represented by (*x*, *y*, *z*) triples, distances along each axis to some arbitrary origin in space. The coordinate data for each atom is attached to a list of labeling information in the structure record: which element, residue, and molecule each point in space belongs to. For the standard biopolymers (DNA, RNA, and proteins), this labeling information can be derived starting with the raw sequence.

Implicit in each sequence is considerable chemical data. We can infer the complete chemical connectivity of the biopolymer molecule directly from a sequence, including all its atoms and bonds, and we could make a sketch, just like the one described earlier, from sequence information alone. We refer to this ''sketch'' of the molecule as the *chemical graph* component of a three-dimensional structure. Every time a sequence is presented in this book or elsewhere, remember that it can encode a fairly complete description of the chemistry of that molecule.

When we sketch all the underlying atoms and bonds representing a sequence, we may defer to a textbook showing the chemical structures of each residue, lest we forget a methyl group or two. Likewise, computers could build up a sketch like a representation of the chemical graph of a structure in memory using a *residue dictionary*, which contains a table of the atom types and bond information for each of the common amino acid and nucleic acid building blocks. What sequence is unable to encode is information about posttranslational modifications. For example, in the structure databases, a phosphorylated tyrosine residue is indicated as ''X'' in the one letter code—essentially an unknown! Any residue that has had an alteration to its standard chemical graph will, unfortunately, be indicated as X in the one-letter encoding of sequence.

Atoms, Bonds, and Completeness

Molecular graphics visualization software performs an elaborate ''connect-the-dots'' process to make the wonderful pictures of protein structure we see in textbooks of biomolecular structure, like the structure for insulin (3INS; Isaccs and Agarwa, 1978) shown in Figure 5.1. The connections used are, of course, the chemical bonds between all the atoms. In current use, three-dimensional molecular structure database records employ two different ''minimalist'' approaches regarding the storage of bond data.

The original approach to recording atoms and bonds is something we shall call the *chemistry rules* approach. The rules are the observable physical rules of chemistry, such as, ''the average length of a stable C—C bond is about 1.5 angstroms.'' Applying these rules to derive the bonds means that any two coordinate locations in space that are 1.5 Å apart and are tagged as carbon atoms always form a single bond. With the chemistry rules approach, we can simply disregard the bonds. A perfect and complete structure can be recorded without any bond information, provided it does not break any of the rules of chemistry in atomic locations. Obviously, this is not always the case, and specific examples of this will be presented later in this chapter.

The chemistry rules approach ended up being the basis for the original three-dimensional biomolecular structure file format, the PDB format from the Protein Data Bank at Brookhaven (Bernstein et al., 1977). These records, in general, lack complete bond information for biopolymers. The working assumption is that no residue dictionary is required for interpretation of data encoded by this approach, just a table of bond lengths and bond types for every conceivable pair of bonded atoms is required.

Every software package that reads in PDB data files must reconstruct the bonds based on these rules. However, the rules we are describing have never been explicitly codified for programmers. This means that interpreting the bonding in PDB files is left for the *programmer* to decide, and, as a result, software can be inconsistent in

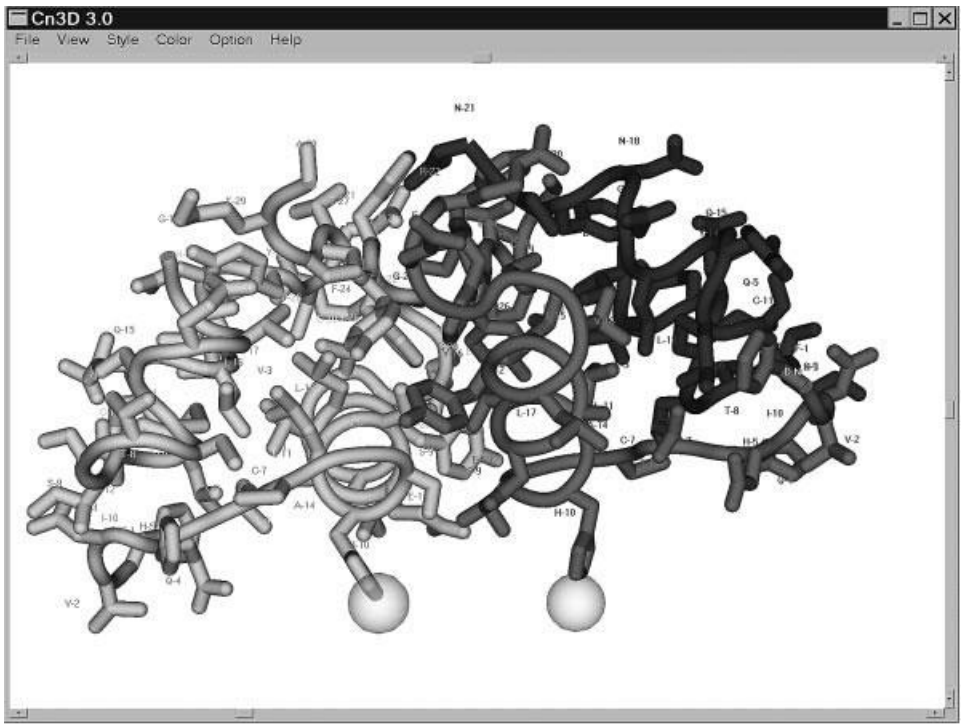

Figure 5.1. The insulin structure 3INS illustrated using Cn3D with OpenGL. Four chains are depicted in the crystallographic unit. This structure illustrates two of many bioinformatics bridges that must be spanned between sequence and structure databases, the lack of encoding of the active biological unit, and the lack of encoding of the relationship of the observed structure to the parent gene. (See color plate.)

the way it draws bonds, especially when different algorithms and distance tolerances are used. The PDB file approach is minimalist in terms of the data stored in a record, and deciphering it often requires much more sophisticated logic than would be needed if the bonding information and chemical graph were explicitly specified in the record. Rarely is this logic properly implemented, and it may in fact be impossible to deal with all the exceptions in the PDB file format. Each exception to the bonding rules needs to be captured by complicated logic statements programmed on a case-by-case basis.

The second approach to describing a molecule is what we call the *explicit bonding approach*, the method that is used in the database records of the Molecular Modeling Database (MMDB), which is, in turn, derived from the data in PDB. In the MMDB system, the data file contains all of its own explicit bonding information. MMDB uses a standard residue dictionary, a record of all the atoms and bonds in the polymer forms of amino acid and nucleic acid residues, plus end-terminal variants. Such data dictionaries are common in the specialized software used by scientists to solve X-ray or NMR structures. The software that reads in MMDB data can use the bonding information supplied in the dictionary to connect atoms together, without trying to enforce (or force) the rules of chemistry. As a result, the three-dimensional coordinate data are consistently interpreted by visualization software, regardless of

type. This approach also lends itself to inherently simpler software, because exceptions to bonding rules are recorded within the database file itself and read in without the need for another layer of exception-handling codes.

Scientists that are unfamiliar with structure data often expect all structures in the public databases to be of ''textbook'' quality. They are often surprised when parts of a structure are missing. The availability of a three-dimensional database record for a particular molecule does not ever imply its completeness. Structural completeness is strictly defined as follows: *At least one coordinate value for each and every atom in the chemical graph is present*.

Structural completeness is quite rare in structure database records. Most X-ray structures lack coordinates for hydrogen atoms because the locations of hydrogens in space are not resolved by the experimental methods currently available. However, some modeling software can be used to predict the locations of these hydrogen atoms and reconstruct a structure record with the now-modeled hydrogens added. It is easy to identify the products of molecular modeling in structure databases. These often have overly complete coordinate data, usually with all possible hydrogen atoms present that could not have been found using an experimental method.

PDB: PROTEIN DATA BANK AT THE RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS (RCSB)

Overview

The use of computers in biology has its origins in biophysical methods, such as X-ray crystallography. Thus, it is not surprising that the first ''bioinformatics'' database was built to store complex three-dimensional data. The Protein Data Bank, originally developed and housed at the Brookhaven National Laboratories, is now managed and maintained by the Research Collaboratory for Structural Bioinformatics (RCSB). RCSB is a collaborative effort involving scientists at the San Diego Supercomputing Center, Rutgers University, and the National Institute of Standards and Technology. The collection contains all publicly available three-dimensional structures of proteins, nucleic acids, carbohydrates, and a variety of other complexes experimentally determined by X-ray crystallographers and NMR spectroscopists. This section focuses briefly on the database and bioinformatics services offered through RCSB.

RCSB Database Services

The World Wide Web site of the Protein Data Bank at the RCSB offers a number of services for submitting and retrieving three-dimensional structure data. The home page of the RCSB site provides links to services for depositing three-dimensional structures, information on how to obtain the status of structures undergoing processing for submission, ways to download the PDB database, and links to other relevant sites and software.

PDB Query and Reporting

Starting at the RCSB home page, one can retrieve three-dimensional structures using two different query engines. The SearchLite system is the one most often used,

providing text searching across the database. The SearchFields interface provides the additional ability to search specific fields within the database. Both of these systems report structure matches to the query in the form of Structure Summary pages, an example of which is shown in Figure 5.2. The RCSB Structure Summary page links are to other Web pages that themselves provide a large number of links, and it may be confusing to a newcomer to not only sift through all this information but to decide which information sources are the most relevant ones for biological discovery.

Submitting Structures. For those who wish to submit three-dimensional structure information to PDB, the RCSB offers its ADIT service over the Web. This

Figure 5.2. Structure query from RCSB with the structure 1BNR (Bycroft et al., 1991). The Structure Explorer can link the user to a variety of other pages with information about this structure including sequence, visualization tools, structure similarity (neighbors), and structure quality information, which are listed on subsequent Web pages.

service provides a data format check and can create automatic validation reports that provide diagnostics as to the quality of the structure, including bond distances and angles, torsion angles, nucleic acid comparisons, and crystal packing. Nucleic acid structures are accepted for deposition at NDB, the Nucleic Acids Database.

It has been the apparent working policy of PDB to reject three-dimensional structures that result from computational three-dimensional modeling procedures rather than from an actual physical experiment; submitting data to the PDB from a nonexperimental computational modeling exercise is strongly discouraged.

PDB-ID Codes. The structure record accessioning scheme of the Protein Data Bank is a unique four-character alphanumeric code, called a PDB-ID or PDB code. This scheme uses the digits 0 to 9 and the uppercase letters A to Z. This allows for over 1.3 million possible combinations and entries. Many older records have mnemonic names that make the structures easier to remember, such as 3INS, the record for insulin shown earlier. A different method is now being used to assign PDB-IDs, with the use of mnemonics apparently being abandoned.

Database Searching, PDB File Retrieval, mmCIF File Retrieval, and Links. PDB's search engine, the Structure Explorer, can be used to retrieve PDB records, as shown in Figure 5.2. The Structure Explorer is also the primary database of links to third-party annotation of PDB structure data. There are a number of links maintained in the Structure Explorer to Internet-based three-dimensional structure services on other Web sites. Figure 5.2 shows the Structure Summary for the protein barnase (1BNR; Bycroft et al., 1991). The Structure Explorer also provides links to special project databases maintained by researchers interested in related topics, such as structural evolution (FSSP; Holm and Sander, 1993), structure-structure similarity (DALI; Holm and Sander, 1996), and protein motions (Gerstein et al., 1994). Links to visualization tool-ready versions of the structure are provided, as well as authored two-dimensional images that can be very helpful to see how to orient a three-dimensional structure for best viewing of certain features such as binding sites.

Sequences from Structure Records

PDB file-encoded sequences are notoriously troublesome for programmers to work with. Because completeness of a structure is not always guaranteed, PDB records contain two copies of the sequence information: an *explicit sequence* and an *implicit sequence*. Both are required to reconstruct the chemical graph of a biopolymer.

Explicit sequences in a PDB file are provided in lines starting with the keyword SEQRES. Unlike other sequence databases, PDB records use the three-letter amino acid code, and nonstandard amino acids are found in many PDB record sequence entries with arbitrarily chosen three-letter names. Unfortunately, PDB records seem to lack sensible, consistent rules. In the past, some double-helical nucleic acid sequence entries in PDB were specified in a 3′-to-5′ order in an entry above the complementary strand, given in 5′-to-3′ order. Although the sequences may be obvious to a user as a representation of a double helix, the 3′-to-5′ explicit sequences are nonsense to a computer. Fortunately, the NDB project has fixed many of these types of problems, but the PDB data format is still open to ambiguity disasters from the standpoint of computer readability. As an aside, the most troubling glitch is the inability to encode element type separately from the atom name. Examples of where

this becomes problematic include cases where atoms in structures having FAD or NAD cofactors are notorious for being interpreted as the wrong elements, such as neptunium (NP to Np), actinium (AC to Ac), and other nonsense elements.

Because three-dimensional structures can have multiple biopolymer chains, to specify a discrete sequence, the user must provide the *PDB chain identifier*. SEQRES entries in PDB files have a chain identifier, a single uppercase letter or blank space, identifying each individual biopolymer chain in an entry. For the structure 3INS shown in Figure 5.1, there are two insulin molecules in the record. The 3INS record contains sequences labeled A, B, C, and D. Knowledge of the biochemistry of insulin is required to understand that protein chains A and B are in fact derived from the same gene and that a posttranslational modification cuts the proinsulin sequence into the A and B chains observed in the PDB record. This information is not recorded in a three-dimensional structure record, nor in the sequence record for that matter. A place for such critical biological information is now being made within the BIND database (Bader and Hogue, 2000). The one-letter chain-naming scheme has difficulties with the enumeration of large oligomeric three-dimensional structures, such as viral capsids, as one quickly runs out of single-letter chain identifiers.

The *implicit* sequences in PDB records are contained in the embedded stereochemistry of the (x, y, z) data and names of each ATOM record in the PDB file. The implicit sequences are useful in resolving explicit sequence ambiguities such as the backward encoding of nucleic acid sequences or in verifying nonstandard amino acids. In practice, many PDB file viewers (such as RasMol) reconstruct the chemical graph of a protein in a PDB record using only the *implicit* sequence, ignoring the *explicit* SEQRES information. If this software then is asked to print the sequence of certain incomplete molecules, it will produce a nonphysiological and biologically irrelevant sequence. The implicit sequence, therefore, is not sufficient to reconstruct the complete chemical graph.

Consider an example in which the sequence `ELVISISALIVES` is represented in the SEQRES entry of a hypothetical PDB file, but the coordinate information is missing all (x, y, z) locations for the subsequence `ISA`. Software that reads the implicit sequence will often report the PDB sequence incorrectly from the chemical graph as `ELVISLIVES`. A test structure to determine whether software looks only at the implicit sequence is 3TS1 (Brick et al., 1989) as shown in the Java three-dimensional structure viewer WebMol in Figure 5.3. Here, both the implicit and explicit sequences in the PDB file to the last residue with coordinates are correctly displayed.

Validating PDB Sequences

To properly validate a sequence from a PDB record, one must first derive the *implicit* sequence in the ATOM records. This is a nontrivial processing step. If the structure has gaps because of lack of completeness, there may only be a set of *implicit sequence fragments* for a given chain. Each of these fragments must be aligned to the *explicit* sequence of the same chain provided within the SEQRES entry. This treatment will produce the complete chemical graph, including the parts of the biological sequence that may be missing coordinate data. This kind of validation is done on creation of records for the MMDB and mmCIF databases.

The best source of validated protein and nucleic acid sequences in single-letter code derived from PDB structure records is NCBI's MMDB service, which is part

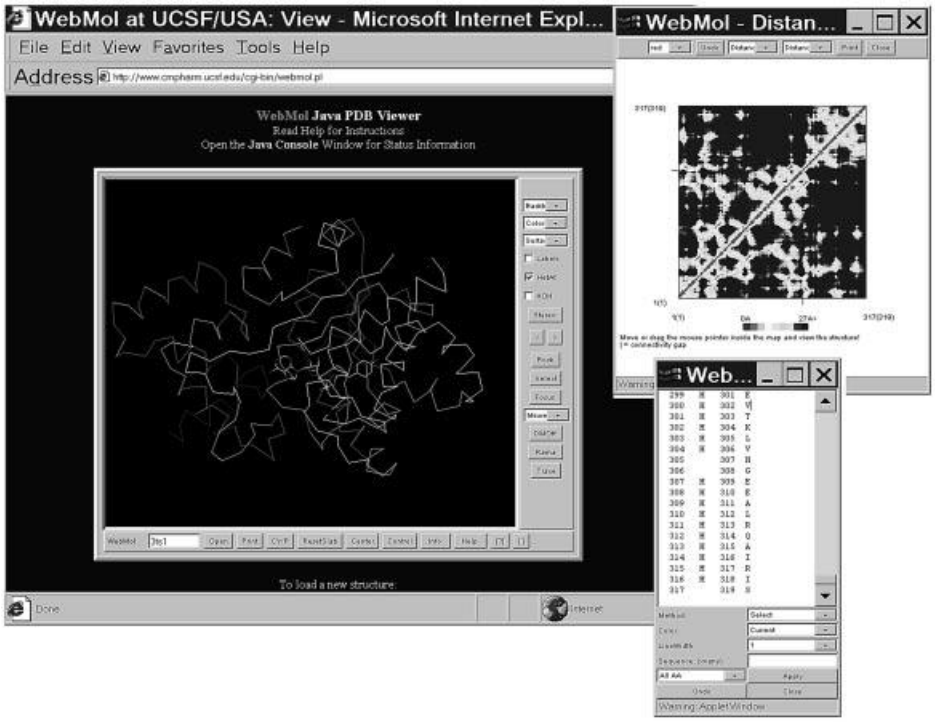

Figure 5.3. Testing a three-dimensional viewer for sequence numbering artifacts with the structure 3TS1 (Brick et al., 1989). WebMol, a Java applet, correctly indicates both the explicit and implicit sequences of the structure. Note the off-by-two difference in the numbering in the two columns of numbers in the inset window on the lower right. The actual sequence embedded in the PDB file is 419 residues long, but the COOH-terminal portion of the protein is lacking coordinates; it also has two missing residues. (See color plate.)

of the Entrez system. The sequence records from our insulin example have database accessions constructed systematically and can be retrieved from the protein sequence division of Entrez using the accessions `pdb|3INS|A`, `pdb|3INS|B`, `pdb|3INS|C`, and `pdb|3INS|D`. PDB files also have references in `db_xref` records to sequences in the SWISS-PROT protein database. Note that the SWISS-PROT sequences will not necessarily correspond to the structure, since the validation process described here is not carried out when these links are made! Also, note that many PDB files currently have ambiguously indicated taxonomy, reflecting the presence in some of three-dimensional structures of complexes of molecules that come from different species. The PDBeast project at NCBI has incorporated the correct taxonomic information for each biopolymer found within a given structure.

MMDB: MOLECULAR MODELING DATABASE AT NCBI

NCBI's Molecular Modeling Database (MMDB; Hogue et al., 1996) is an integral part of NCBI's Entrez information retrieval system (cf. Chapter 7). It is a compilation of all the Brookhaven Protein Data Bank (Bernstein et al., 1977) three-dimensional

structures of biomolecules from crystallographic and NMR studies. MMDB records are in ASN.1 format (Rose, 1990) rather than in PDB format. Despite this, PDB-formatted files can also be obtained from MMDB. By representing the data in ASN.1 format, MMDB records have value-added information compared with the original PDB entries. Additional information includes explicit chemical graph information resulting from an extensive suite of validation procedures, the addition of uniformly derived secondary structure definitions, structure domain information, citation matching to MEDLINE, and the molecule-based assignment of taxonomy to each biologically derived protein or nucleic acid chain.

Free Text Query of Structure Records

The MMDB database can be searched from the NCBI home page using Entrez. (MMDB is also referred to as the NCBI Structure division.) Search fields in MMDB include PDB and MMDB ID codes, free text from the original PDB REMARK records, author name, and other bibliographic fields. For more specific, fielded queries, the RCSB site is recommended.

MMDB Structure Summary

MMDB's Web interface provides a Structure Summary page for each MMDB structure record, as shown in Figure 5.4. MMDB Structure Summary pages provide the FASTA-formatted sequences for each chain in the structure, links to MEDLINE references, links to the 3DBAtlas record and the Brookhaven PDB site, links to protein or nucleic acid sequence neighbors for each chain in the structure, and links to VAST structure-structure comparisons for each domain on each chain in the structure.

BLAST Against PDB Sequences: New Sequence Similarities

When a researcher wishes to find a structure related to a new sequence, NCBI's BLAST (Altschul et al., 1990) can be used because the BLAST databases contain a copy of all the validated sequences from MMDB. The BLAST Web interface can be used to perform the query by pasting a sequence in FASTA format into the sequence entry box and then selecting the ''pdb'' sequence database. This will yield a search against all the validated sequences in the current public structure database. More information on performing BLAST runs can be found in Chapter 8.

Entrez Neighboring: Known Sequence Similarities

If one is starting with a sequence that is already in Entrez, BLAST has, in essence, already been performed. Structures that are similar in sequence to a given protein sequence can be found by means of Entrez's neighboring facilities. Details on how to perform such searches are presented in Chapter 7.

Figure 5.4. Structure query from NCBI with the structure 1BNR (Bycroft et al., 1991). The Structure Summary links the user to RCSB through the PDB ID link, as well as to validated sequence files for each biopolymer, sequence, and three-dimensional structure neighbors through the VAST system. This system is more efficient than the RCSB system (Fig. 5.2) for retrieval because visualization, sequence, and structure neighbor links are made directly on the structure summary page and do not require fetching more Web pages.

STRUCTURE FILE FORMATS

PDB

The PDB file format is column oriented, like that of the punched cards used by early FORTRAN programmers. The exact file format specification is available through the PDB Web site. Most software developed by structural scientists is written in FORTRAN, whereas the rest of the bioinformatics world has adopted other languages, such as those based on C. PDB files are often a paradox: they look rather easy to parse, but they have a few nasty surprises, as already alluded to in this chapter. To the uninitiated, the most obvious problem is that the information about biopolymer bonds is missing, obliging one to program in the rules of chemistry, clues to the identity of each atom given by the naming conventions of PDB, and robust exception handling. PDB parsing software often needs lists of synonyms and tables of exceptions to correctly interpret the information. However this chapter is not intended to be a manual of how to construct a PDB parser.

Two newer chemical-based formats have emerged: mmCIF (MacroMolecular Chemical Interchange Format) and MMDB (Molecular Modeling Database Format). Both of these file formats are attempts to modernize PDB information. Both start by using data description languages, which are consistently machine parsable. The data description languages use ''tag value'' pairs, which are like variable names and values used in a programming language. In both cases, the format specification is composed in a machine-readable form, and there is software that uses this format specification document to validate incoming streams of data. Both file formats are populated from PDB file data using the strategy of alignment-based reconstruction of the implicit ATOM and HETATM chemical graphs with the explicit SEQRES chemical graphs, together with extensive validation, which is recorded in the file. As a result, both of these file formats are superior for integrating with biomolecular sequence databases over PDB format data files, and their use in future software is encouraged.

mmCIF

The mmCIF (Bourne et al., 1995) file format was originally intended to be a biopolymer extension of the CIF (Chemical Interchange Format; Hall et al., 1991) familiar to small-molecule crystallographers and is based on a subset of the STAR syntax (Hall et al., 1991). CIF software for parsing and validating format specifications is not forward-compatible with mmCIF, since these have different implementations for the STAR syntax. The underlying data organization in an mmCIF record is a set of relational tables. The mmCIF project refers to their format specification as the mmCIF dictionary, kept on the Web at the Nucleic Acids Database site. The mmCIF dictionary is a large document containing specifications for holding the information stored in PDB files as well as many other data items derivable from the primary coordinate data, such as bond angles. The mmCIF data specification gives this data a consistent interface, which has been used to implement the NDB Protein Finder, a Web-based query format in a relational database style, and is also used as the basis for the new RCSB software systems.

Validating an incoming stream of data against the large mmCIF dictionary entails significant computational time; hence, mmCIF is probably destined to be an archival

and advanced query format. Software libraries for reading mmCIF tables into relational tables into memory in FORTRAN and C are available.

MMDB

The MMDB file format is specified by means of the ASN.1 data description language (Rose, 1990), which is used in a variety of other settings, surprisingly enough including applications in telecommunications and automotive manufacturing. Because the US National Library of Medicine also uses ASN.1 data specifications for sequence and bibliographic information, the MMDB format borrows certain elements from other data specifications, such as the parts used in describing bibliographic references cited in the data record. ASN.1 files can appear as human-readable text files or as a variety of binary and packed binary files that can be decoded by any hardware platform. The MMDB standard residue dictionary is a lookup table of information about the chemical graphs of standard biopolymer residue types. The MMDB format specification is kept inside the NCBI toolkit distribution, but a browser is available over the Web for a quick look. The MMDB ASN.1 specification is much more compact and has fewer data items than the mmCIF dictionary, avoiding derivable data altogether.

In contrast to the relational table design of mmCIF, the MMDB data records are structured as hierarchical records. In terms of performance, ASN.1-formatted MMDB files provide for much faster input and output than do mmCIF or PDB records. Their nested hierarchy requires fewer validation steps at load time than the relational scheme in mmCIF or in the PDB file format; hence, ASN.1 files are ideal for three-dimensional structure database browsing.

A complete application programming interface is available for MMDB as part of the NCBI toolkit, containing a wide variety of C code libraries and applications. Both an ASN.1 input/output programming interface layer and a molecular computing layer (MMDB-API) are present in the NCBI toolkit. The NCBI toolkit supports x86 and alpha-based Windows' platforms, Macintosh 68K and PowerPC CPUs, and a wide variety of UNIX platforms. The three-dimensional structure database viewer (Cn3D) is an MMDB-API-based application with source code included in the NCBI toolkit.

VISUALIZING STRUCTURAL INFORMATION

Multiple Representation Styles

We often use multiple styles of graphical representation to see different aspects of molecular structure. Typical images of a protein structure are shown in Figure 5.5 (see also color plate). Here, the enzyme barnase 1BN1 (Buckle et al., 1993) appears both in wire-frame and space-filling model formats, as produced by RasMol (Sayle and Milner-White, 1995).

Because the protein structure record 1BN1 has three barnase molecules in the crystallographic unit, the PDB file has been hand-edited using a text editor to delete the superfluous chains. Editing data files is an accepted and widespread practice in three-dimensional molecular structure software, forcing the three-dimensional structure viewer to show what the user wants. In this case, the crystallographic data

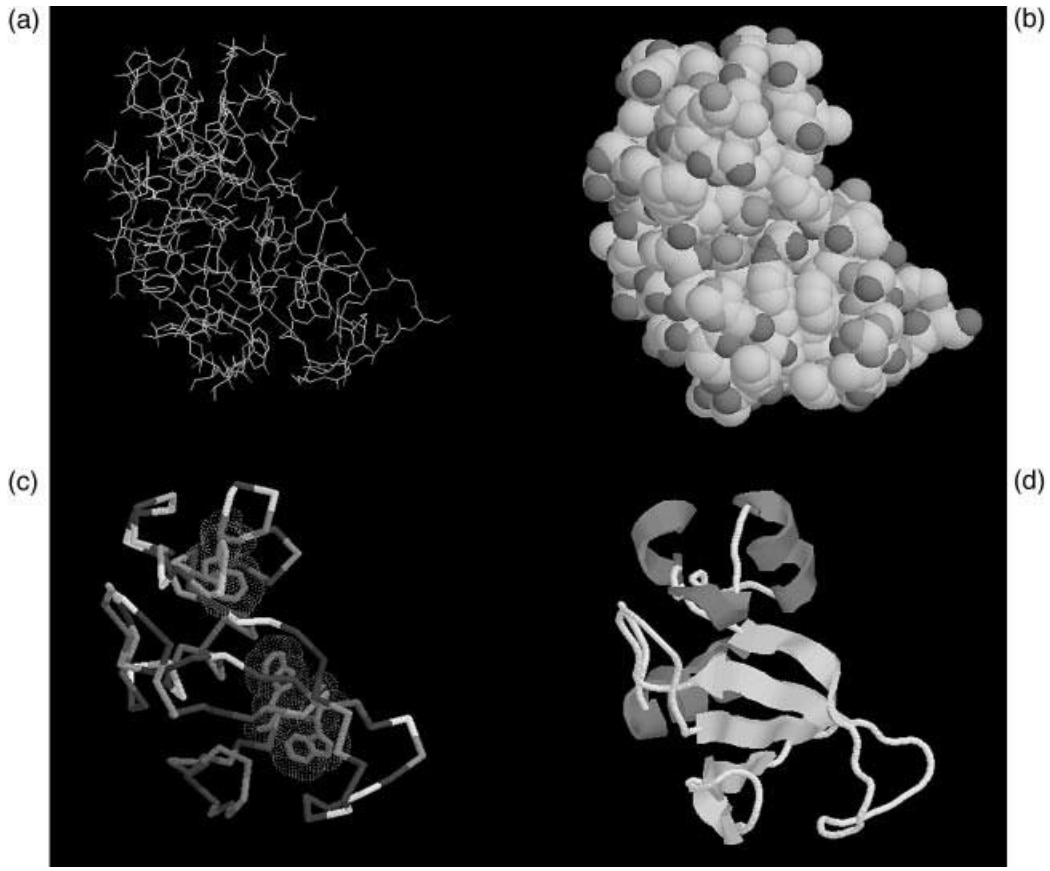

Figure 5.5. A constellation of viewing alternatives using RasMol with a portion of the barnase structure 1BN1 (Buckle et al., 1993). 1BN1 has three barnase molecules in the asymmetric unit. For this figure, the author edited the PDB file to remove two extra barnase molecules to make the images. Like most crystal structures, 1BN1 has no hydrogen locations. (a) Barnase in CPK coloring (element-based coloring) in a wire-frame representation. (b) Barnase in a space-filling representation. (c) Barnase in an α-carbon backbone representation, colored by residue type. The command line was used to select all the tryptophan residues, render them with "sticks," color them purple, and show a dot surface representation. (d) Barnase in a cartoon format showing secondary structure, α-helices in red; β-strands in yellow. Note that in all cases the default atom or residue coloring schemes used are at the discretion of the author of the software. (See color plate.)

recorded in the three-dimensional structure does not represent the functional *biological* unit. In our example, the molecule barnase is a monomer; however, we have three molecules in the crystallographic unit. In our other example, 3TS1 (Brick et al., 1989) (Fig. 5.3), the molecule is a dimer, but only one of the symmetric subunits is recorded in the PDB file.

The wire-frame image in Figure 5.5a clearly shows the chemistry of the barnase structure, and we can easily trace of the sequence of barnase on the image of its biopolymer in an interactive computer display. The space-filling model in Figure

5.5b gives a good indication of the size and surface of the biopolymer, yet it is difficult to follow the details of chemistry and bonding in this representation. The composite illustration in Figure 5.5c shows an α-carbon backbone in a typical pseudo-structure representation. The lines drawn are not actual chemical bonds, but they guide us along the path made by the α-carbons of the protein backbone. These are also called ''virtual bonds.'' The purple tryptophan side chains have been selected and drawn together with a dot surface. This composite illustration highlights the volume taken up by the three tryptophan side chains in three hydrophobic core regions of barnase, while effectively hiding most of the structure's details.

The ribbon model in Figure 5.5d shows the organization of the structural path of the secondary structure elements of the protein chain (α-helix and β-sheet regions). This representation is very often used, with the arrowheads indicating the N-to-C-terminal direction of the secondary structure elements, and is most effective for identifying secondary structures within complex topologies.

The variety of information conveyed by the different views in Figure 5.5 illustrates the need to visualize three-dimensional biopolymer structure data in unique ways that are not common to other three-dimensional graphics applications. This requirement often precludes the effective use of software from the ''macroscopic world,'' such as computer-aided design (CAD) or virtual reality modeling language (VRML) packages.

Picture the Data: Populations, Degeneracy, and Dynamics

Both X-ray and NMR techniques infer three-dimensional structure from a *synchronized* population of molecules—synchronized in space as an ordered crystal lattice or synchronized in behavior as nuclear spin states are organized by an external magnetic field. In both cases, information is gathered from the population as a whole. The coordinate (x, y, z) locations of atoms in a structure are derived using numerical methods. These fit the expected chemical graph of the sample into the three-dimensional data derived from the experimental data. The expected chemical graph can include a mixture of biopolymer sequence-derived information as well as the chemical graph of any other known small molecules present in the sample, such as substrates, prosthetic groups, and ions.

One somewhat unexpected result of the use of molecular populations is the assignment of degenerate coordinates in a database record, i.e., more than one coordinate location for a single atom in the chemical graph. This is recorded when the population of molecules has observable conformational heterogeneity.

NMR Models and Ensembles

Figure 5.6 (see also color plate) presents four three-dimensional structures (images on the left were determined using X-ray crystallography and the right using NMR). The NMR structures on the left appear ''fuzzy.'' In fact, there are several different, complete structures piled one on top of another in these images. Each structure is referred to as a *model*, and the set of models is an *ensemble*. Each model in the ensemble is a chirally correct, plausible structure that fits the underlying NMR data as well as any other model in the ensemble.

The images from the ensemble of an NMR structure (Fig. 5.6, b and d) show the dynamic variation of a molecule in solution. This reflects the conditions of the

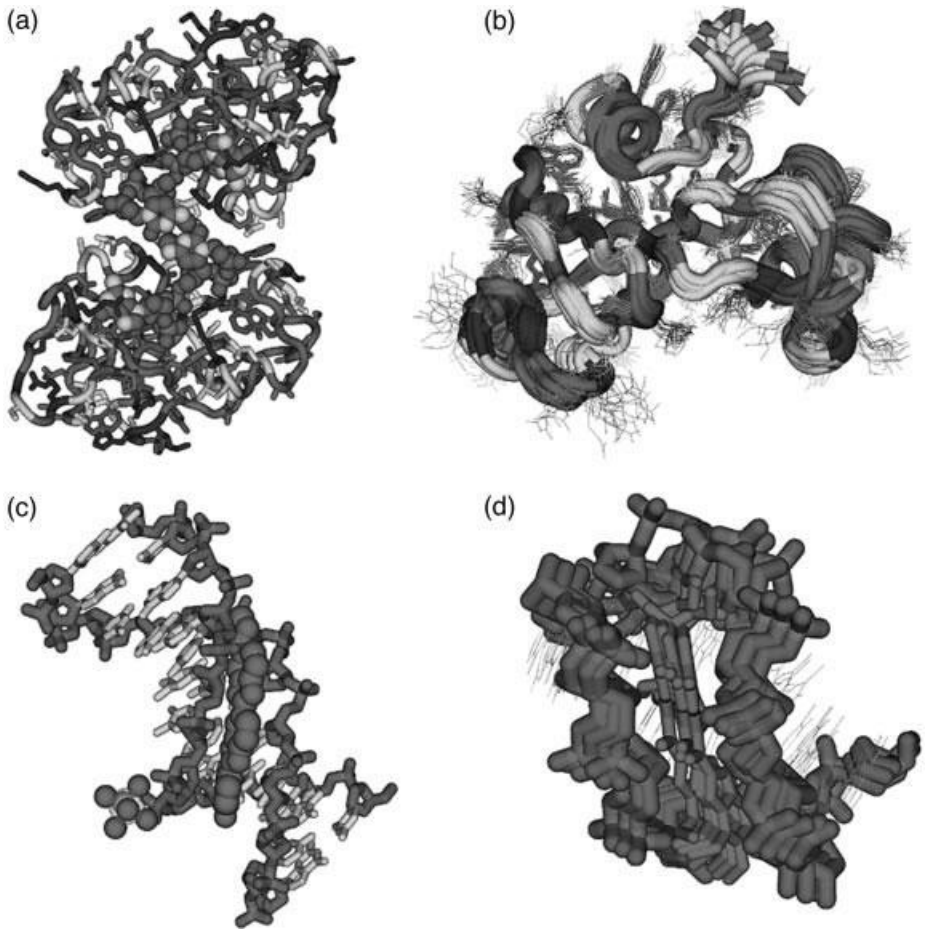

Figure 5.6. A comparison of three-dimensional structure data obtained by crystallography (left) and NMR methods (right), as seen in Cn3D. (a) The crystal structure 1BRN (Buckle and Fersht, 1994) has two barnase molecules in the asymmetric unit, although these are not dimers in solution. The image is rendered with an α-carbon backbone trace colored by secondary structure (green helices and yellow sheets), and the amino acid residues are shown with a wire-frame rendering, colored by residue type. (b) The NMR structure 1BNR (Bycroft et al., 1991) showing barnase in solution. Here, there are 20 different models in the ensemble of structures. The coloring and rendering are exactly as the crystal structure to its left. (c) The crystal structure 109D (Quintana et al., 1991) showing a complex between a minor-groove binding bis-benzimidazole drug and a DNA fragment. Note the phosphate ion in the lower left corner. (d) The NMR structure 107D showing four models of a complex between a different minor-groove binding compound (Duocarmycin A) and a different DNA fragment. It appears that the three-dimensional superposition of these ensembles is incorrectly shifted along the axis of the DNA, an error in PDB's processing of this particular file. (See color plate.)

experiment: molecules free in solution with freedom to pursue dynamic conformational changes. In contrast, the X-ray structures (Fig. 5.6, a and c) provide a very strong mental image of a static molecule. This also reflects the conditions of the experiment, an ordered crystal constrained in its freedom to explore its conformational dynamics. These mental images direct our interpretation of structure. If we measure distance between two atoms using an X-ray structure, we may get a single value. However, we can get a range of values for the same distance in each model looking at an ensemble of an NMR structure. Clearly, our interpretation of this distance can be dependent on the source of the three-dimensional structure data! It is prudent to steer clear of any software that ignores or fails to show the population degeneracy present in structure database records, since the absence of such information can further skew *biological* interpretations. Measuring the distance between two atoms in an NMR structure using software that hides the other members of the ensemble will give only one value and not the true range of distance observed by the experimentalist.

Correlated Disorder

Typically, X-ray structures have one and only one model. Some subsets of atoms, however, may have degenerate coordinates, which we will refer to as *correlated disorder* (Fig. 5.7a; see also color plate). Many X-ray structure database records show correlated disorder. Both correlated disorder and ensembles are often ignored by three-dimensional molecular graphics software. Some programs show only the first model in an ensemble, or the first location of each atom in a correlated disorder set, ignoring the rest of the degenerate coordinate values. Worse still, sometimes, erroneous bonds are drawn between the degenerate locations, making a mess of the structure, as seen in Figure 5.7b.

Local Dynamics

A single technique can be used to constrain the conformation of some atoms differently from others in the same structure. For example, an internal atom or a backbone atom that is locked in by a multitude of interactions may appear largely invariant in NMR or X-ray data, whereas an atom on the surface of the molecule may have much more conformational freedom (consider the size of the smears of different residues in Fig. 5.6b). Interior protein side chains typically show much less flexibility in ensembles, so it might be concluded that the interiors of proteins lack conformational dynamics altogether. However, a more sensitive, biophysical method, time-resolved fluorescence spectroscopy of single tryptophan residues, has a unique ability to detect heterogeneity (but not the actual coordinates) of the tryptophan side-chain conformation. Years of study using this method has shown that, time and time again, populations of interior tryptophans in pure proteins are more often in heterogeneous conformations than not (Beechem and Brand, 1985). This method was shown to be able to detect rotamers of tryptophan within single crystals of erabutoxin, where X-ray crystallography could not (Dahms and Szabo, 1995). When interpreting three-dimensional structure data, remember that heterogeneity does persist in the data, and that the NMR and X-ray methods can be blind to all but the most populated conformations in the sample.

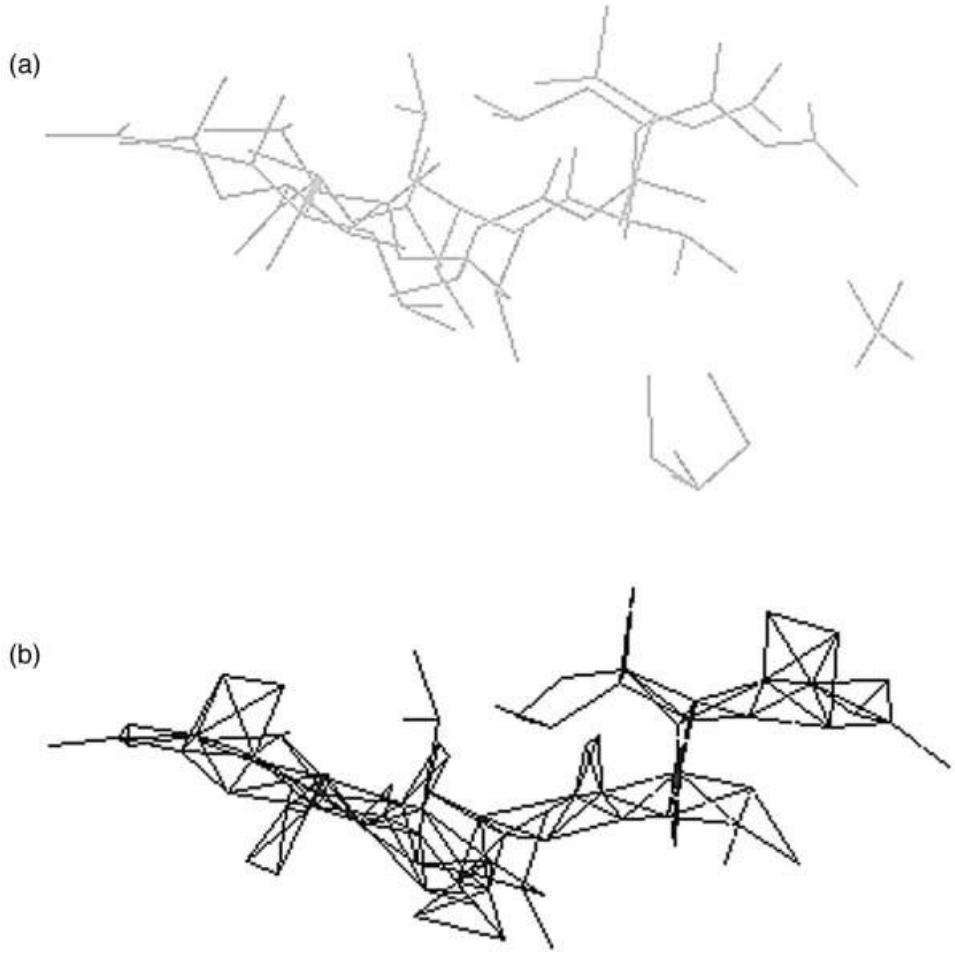

Figure 5.7. An example of crystallographic correlated disorder encoded in PDB files. This is chain C of the HIV protease structure 5HVP (Fitzgerald et al., 1990). This chain is in asymmetric binding site and can orient itself in two different directions. Therefore, it has a single chemical graph, but each atom can be in one of two different locations. (a) The correct bonding is shown with an MMDB-generated Kinemage file; magenta and red are the correlated disorder ensembles as originally recorded by the depositor, bonding calculated using standard-residue dictionary matching. (b) Bonding of the same chain in RasMol, wherein the disorder ensemble information is ignored, and all coordinates are displayed and all possible bonds are bonded together. (See color plate.)

DATABASE STRUCTURE VIEWERS

In the past several years, the software used to examine and display structure information has been greatly improved in terms of the quality of visualization and, more importantly, in terms of being able to relate sequence information to structure information.

Visualization Tools

Although the RCSB Web site provides a Java-based three-dimensional applet for visualizing PDB data, the applet does not currently support the display of nonprotein structures. For this and other reasons, the use of RasMol v2.7 is instead recommended for viewing structural data downloaded from RCSB; more information on RasMol appears in the following section. If a Java-based viewer is preferred, WebMol is recommended, and an example of WebMol output is shown in Figure 5.3. With the advent of many homemade visualization programs that can easily be downloaded from the Internet, the reader is strongly cautioned to *only* use mature, well-established visualization tools that have been thoroughly tested and have undergone critical peer review.

RasMol and RasMol-Based Viewers

As mentioned above, several viewers for examining PDB files are available (Sanchez-Ferrer et al., 1995). The most popular one is RasMol (Sayle and Milner-White, 1995). RasMol represents a breakthrough in software-driven three-dimensional graphics, and its source code is a recommended study material for anyone interested in high-performance three-dimensional graphics. RasMol treats PDB data with extreme caution and often recomputes information, making up for inconsistencies in the underlying database. It does not try to validate the chemical graph of sequences or structures encoded in PDB files. RasMol does not perform internally either dictionary-based standard residue validations or alignment of explicit and implicit sequences. RasMol 2.7.1 contains significant improvements that allow one to display information in correlated disorder ensembles and select different NMR models. It also is capable of reading mmCIF-formatted three-dimensional structure files and is thus the viewer of choice for such data. Other data elements encoded in PDB files, such as disulfide bonds, are recomputed based on rules of chemistry, rather than validated.

RasMol contains many excellent output formats and can be used with the Molscript program (Kraulis, 1991) to make wonderful PostScript™ ribbon diagrams for publication. To make optimal use of RasMol, however, one must master its command-line language, a familiar feature of many legacy three-dimensional structure programs.

Several new programs are becoming available and are free for academic users. Based on RasMol's software-driven three-dimensional-rendering algorithms and sparse PDB parser, these programs include Chime™, a Netscape™ plug-in. Another program, WebMol, is a Java-based three-dimensional structure viewer apparently based on RasMol-style rendering, as seen in Figure 5.3.

MMDB Viewer: Cn3D

Cn3D (for ''see in 3-D'') is a three-dimensional structure viewer used for viewing MMDB data records. Because the chemical graph ambiguities in data in PDB entries have been removed to make MMDB data records and because all the bonding information is explicit, Cn3D has the luxury of being able to display three-dimensional database structures consistently, without the parsing, validation, and exception-handling overhead required of programs that read PDB files. Cn3D's default image of

a structure is more intelligently displayed because it works without fear of misrepresenting the data. However, Cn3D is dependent on the complete chemical graph information in the ASN.1 records of MMDB, and, as such, it does not read in PDB files.

Cn3D 3.0 has a much richer feature set than its predecessors, and it now allows selection of subsets of molecular structure and independent settings of rendering and coloring aspects of that feature. It has state-saving capabilities, making it possible to color and render a structure, and then save the information right into the ASN.1 structure record, a departure from the hand-editing of PDB files or writing scripts. This information can be shared with other Cn3D users on different platforms.

The images shown in Figures 5.1 and 5.6 are from Cn3D 3.0, now based on OpenGL three-dimensional graphics. This provides graphics for publication-quality images that are much better than previous versions, but the original Viewer3D version of Cn3D 3.0 is available for computers that are not capable of displaying OpenGL or that are too slow.

Also unique to Cn3D is a capacity to animate three-dimensional structures. Cn3D's animation controls resemble tape recorder controls and are used for displaying quickly the members of a multiple structure ensemble one after the other, making an animated three-dimensional movie. The GO button makes the images animated, and the user can rotate or zoom the structure while it is playing the animation. This is particularly useful for looking at NMR ensembles or a series of time steps of structures undergoing motions or protein folding. The animation feature also allows Cn3D to provide superior multiple structure alignment when used together with the VAST structure-structure comparison system, described later in this chapter.

Other 3D Viewers: Mage, CAD, and VRML

A variety of file formats have been used to present three-dimensional biomolecular structure data lacking in chemistry-specific data representations. These are viewed in generic three-dimensional data viewers such as those used for ''macroscopic'' data, like engineering software or virtual-reality browsers. File formats such as VRML contain three-dimensional graphical display information but little or no information about the underlying chemical graph of a molecule. Furthermore, it is difficult to encode the variety of rendering styles in such a file; one needs a separate VRML file for a space-filling model of a molecule, a wire-frame model, a ball-and-stick model, and so on, because each explicit list of graphics objects (cylinders, lines, spheres) must be contained in the file.

Biomolecular three-dimensional structure database records are currently not compatible with ''macroscopic'' software tools such as those based on CAD software. Computer-aided design software represents a mature, robust technology, generally superior to the available molecular structure software. However, CAD software and file formats in general are ill-suited to examine the molecular world, owing to the lack of certain ''specialty'' views and analytical functions built in for the examination of details of protein structures.

Making Presentation Graphics

To get the best possible publication-quality picture out of any molecular graphics software, first consider whether a bitmap or a vector-based graphic image is needed. Bitmaps are made by programs like RasMol and Cn3D—they reproduce exactly

what you see on the screen, and are usually the source of trouble in terms of pixellation (''the jaggies''), as shown in Figure 5.7, a bitmap of 380–400 pixels. High-quality print resolution is usually at 300–600 dots per inch, but monitors have far less information in pixels per inch (normally 72 dpi), so a big image on a screen is quite tiny when printed at the same resolution on a printer. Expanding the image to fit a page causes exaggeration of pixel steps on diagonal lines.

The best advice for bitmaps is to use as big a monitor/desktop as possible, maximizing the number of pixels included in the image. This may mean borrowing a colleagues' 21-in monitor or using a graphics card that offers a ''virtual desktop'' that is larger than the monitor being used in pixel count. In any case, always fill the entire screen with the viewer window before saving a bitmap image for publication.

ADVANCED STRUCTURE MODELING

Tools that go beyond simple visualization are now emerging and are freely available. Biologists often want to display structures with information about charge distribution, surface accessibility, and molecular shape; they also want to be able to perform simple mutagenesis experiments and more complex structure modeling. SwissPDB Viewer, shown in Figure 5.8, also known as Deep View, is provided free of charge to academics and can address a good number of these needs. It is a multi platform (Mac, Win, and Linux) OpenGL-based tool that has the ability to generate molecular surfaces, align multiple proteins, use scoring functions, as well as do simple, fast modeling, including site-directed mutagenesis and more complex modeling such as loop rebuilding. An excellent tutorial for SwissPDB Viewer developed by Gale Rhodes is one of the best starting points for making the best use of this tool. It has the capability to dump formatted files for the free ray-tracing software POV-Ray, and it can be used to make stunning images of molecular structures, easily suitable for a journal cover.

STRUCTURE SIMILARITY SEARCHING

Although a sequence-sequence similarity program provides an alignment of two sequences, a structure-structure similarity program provides a three-dimensional structural superposition. This superposition results from a set of three-dimensional rotation-translation matrix operations that superimpose similar parts of the structure. A conventional sequence alignment can be derived from three-dimensional superposition by finding the α-carbons in the protein backbone that are superimposed in space. Structure similarity search services are based on the premise that some similarity metric can be computed between two structures and used to assess their similarity, much in the same way a BLAST alignment is scored. A structure similarity search service can take a three-dimensional protein structure, either already in PDB or a new one, and compare that structure, making three-dimensional superpositions with other structures in the database and reporting the best match without knowing anything about the sequence. If a match is made between two structures that are not related by any measurable sequence similarity, it is indeed a surprising discovery. For this type of data to be useful, the similarity metric *must* be meaningful. A large fraction of structures, for example, have β-sheets. Although a similar substructure may include a single β-hairpin turn with two strands, one can find an incredibly

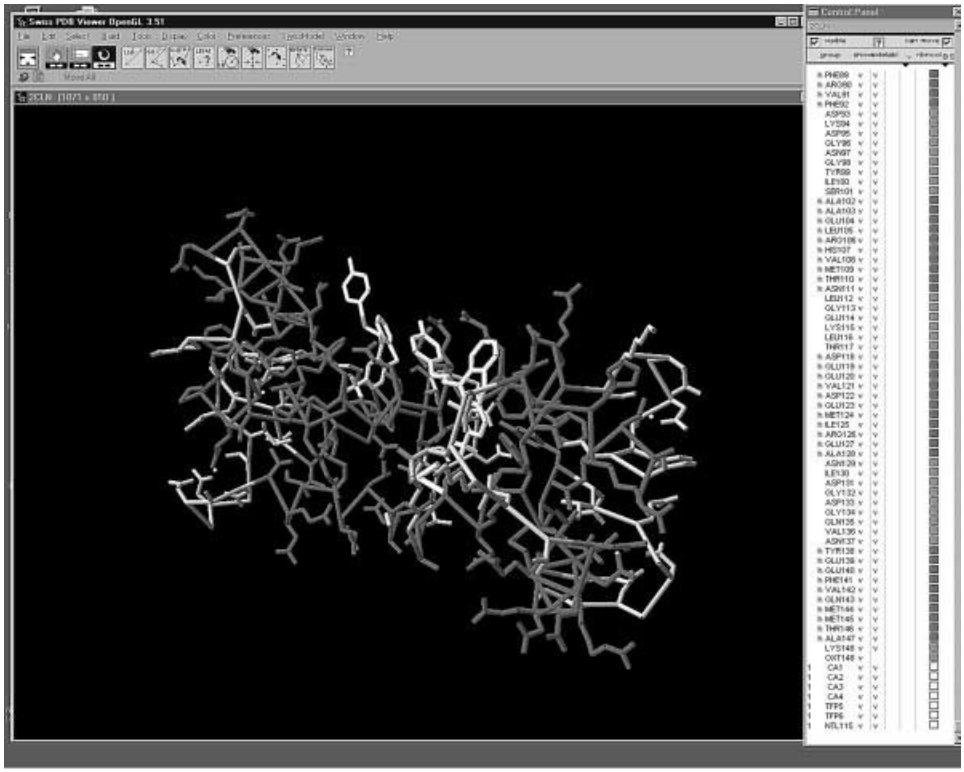

Figure 5.8. SwissPDB Viewer 3.51 with OpenGL, showing the calmodulin structure 2CLN. The binding of the inhibitor TFP is shown in yellow. The side panel allows great control over the rendering of the structure image, and menus provide a wealth of options and tools for structure superposition and modeling including mutagenesis and loop modeling, making it a complete structure modeling and analysis package. (See color plate.)

large number of such similarities in the PDB database, so these similarities are simply not surprising or informative. A number of structure similarity searching systems are now available on the Internet, and almost all of them can be found following links from the RCSB Structure Summary page. The process of similarity searching presents some interesting high-performance computational challenges, and this is addressed in different ways, ranging from human curation, as the SCOP system provides, to fully automated systems, such as DALI, SCOP, or the CE system provided by RCSB.

The Vector Alignment Search Tool (VAST; Gibrat et al., 1996) provides a similarity measure of three-dimensional structure. It uses vectors derived from secondary structure elements, with no sequence information being used in the search. VAST is capable of finding structural similarities when no sequence similarity is detected. VAST, like BLAST, is run on all entries in the database in an $N \times N$ manner, and the results are stored for fast retrieval using the Entrez interface. More than 20,000 domain substructures within the current three-dimensional structure database have been compared with one another using the VAST algorithm, the structure-structure (Fig. 5.9) superpositions recorded, and alignments of sequence derived from the

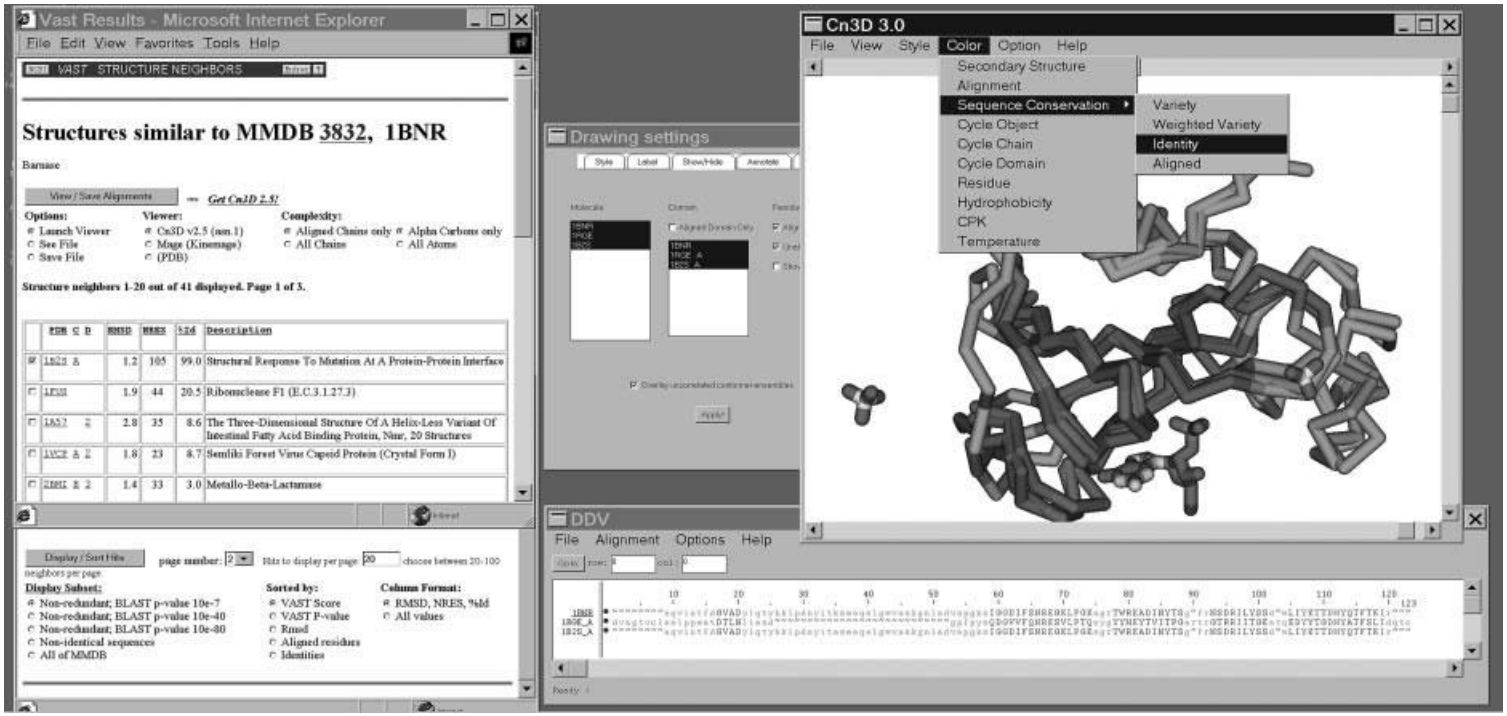

Figure 5.9. VAST structure neighbors of barnase. On the left is the query window obtained by clicking on the Structure Neighbors link from Figure 5.4. Structures to superposition are selected with the check boxes on the left, and Cn3D is launched from the top of the Web page. At the bottom left, controls that change the query are shown from the bottom of the VAST results page. The results shown here are selected as examples from a nonredundant set based on a BLAST probability of 10^{-7}, for the most concise display of hits that are not closely related to one another by sequence. The list may be sorted by a number of parameters, including RMSD from the query structure, number of identical residues, and the raw VAST score. More values can be displayed in the list as well. Cn3D is shown on the right, launched from the Web page with the structures 1RGE and 1B2S. Menu options show how Cn3D can highlight residues in the superposition (top right) and in the alignment (bottom right). The Cn3D drawing settings are shown in the top middle, where one can toggle structures on or off in the superposition window. (See color plate.)

superposition. The VAST algorithm focuses on similarities that are surprising in the *statistical* sense. One does not waste time examining many similarities of small substructures that occur by chance in protein structure comparison. For example, very many small segments of β-sheets have obvious, but not surprising, similarities. The similarities detected by VAST are often examples of remote homology, undetectable by sequence comparison. As such, they may provide a broader view of the structure, function, and evolution of a protein family.

The VAST system stands out amongst these comparative tools because (a) it has a clearly defined similarity metric leading to surprising relationships, (b) it has an adjustable interface that shows nonredundant hits for a quick first look at the most interesting relationships, without seeing the same relationships lots of times, (c) it provides a domain-based structure comparison rather than a whole protein comparison, and (d) it has the capability to integrate with Cn3D as a visualization tool for inspecting surprising structure relationships in detail. The interface between a VAST hit list search and the Cn3D structure superposition interface can be seen in Figure 5.9. In addition to a listing of similar structures, VAST-derived structure neighbors contain detailed residue-by-residue alignments and three-dimensional transformation matrices for structural superposition. In practice, refined alignments from VAST appear conservative, choosing a highly similar ''core'' substructure compared with DALI (Holm and Sander, 1996) superpositions. With the VAST superposition, one easily identifies regions in which protein evolution has modified the structure, whereas DALI superpositions may be more useful for comparisons involved in making structural models. Both VAST and DALI superpositions are excellent tools for investigating relationships in protein structure, especially when used together with the SCOP (Murzin et al., 1995) database of protein families.

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 5

BIND	*http://bioinfo.mshri.on.ca*
Imagemagick	*http://http://www.wizards.dupont.com/cristy/ImageMagick.html*
mmCif Project	*http://ndbserver.rutgers.edu/NDB/mmcif/index.html*
National Center for Biotechnology Information (NCBI)	*http://www.ncbi.nlm.nih.gov/*
NCBI Toolkit	*http://www.ncbi.nlm.nih.gov/Toolbox*
Nucleic Acids Database (NDB)	*http://ndbserver.rutgers.edu/*
POV-RayAY	*http://www.povray.org/*
Protein Data Bank at RCSB	*http://www.rcsb.org/*
RasMol	*http://www.bernstein-plus-sons.com/*
SwissPDB Viewer/Deep View	*http://http://www.expasy.ch/spdbv/mainpage.html*
WebMol	*http://www.cmpharm.ucsf.edu/~walther/webmol/*

PROBLEM SET

1. Calmodulin is a calcium-dependent protein that modulates other protein activity via protein interactions. The overall structure of calmodulin is variable, and is modulated by calcium. An NMR structure of calmodulin is found in PDB record 2BBN, complexed with a peptide. How many models are in this structure? Find other calmodulin structures from the PDB site, and inspect them using RasMol. How many ''gross,'' unique conformations can this protein be found in? Where, in terms of secondary structure, is the site of the largest structural change?

2. The black beetle virus coat protein (pdb|2BBV) forms a very interesting trianglular shape. Using VAST, examine the list of neighbors to the A chain. Examine some of the pairwise alignments in Cn3D. What is the extent of the similarity? What does this list of neighbors and the structure similarity shared by these proteins suggest about the origin and evolution of eukaryotic viruses?

3. Compare substrate binding in Rossman fold structures, between the tyrosinyl-5′-adenylate of tyrosyl-tRNA synthetase and the NADH of malate dehydrogenase. Describe the similarities and differences between the two substrates. Do you think these are homologous structures or are they related by convergent evolution?

4. Ribosomal protein synthesis or enzyme-based peptide synthesis—which came first?

Repeat the analysis you did for question 3, examining the difference between substrates bound to tyrosyl-tRNA synthetase and D-Ala:D-Ala ligase (pdb|1IOV). Note the substrate is bound to domains 2 and 3 of 1IOV, but domain 1 is aligned with 3TS1. What does the superposition suggest about the activity of domain 1 of 1IOV? According to VAST, what is similar to domain 2 of 1IOV? How do you think D-Ala:D-Ala ligase arose in evolution? Speculate on whether enzyme-catalyzed protein synthesis such as that seen in 1IOV arose before or after ribosomal protein synthesis.

REFERENCES

Ahmed, F. R., Przybylska, M., Rose, D. R., Birnbaum, G. I., Pippy, M. E., and MacManus, J. P. (1990). Structure of oncomodulin refined at 1.85 angstroms resolution. An example of extensive molecular aggregation via Ca^2. *J. Mol. Biol*. 216, 127–140.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Bader, G. D., and Hogue, C. W. V. (2000). BIND, a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* 16, 465–477.

Beechem, J. M., and Brand, L. (1985). Time-resolved fluorescence of proteins. *Annu. Rev. Biochem.* 54, 43–71.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank. *J. Mol. Biol.* 112, 535–542.

Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J., and Fitzgerald, P. M. D. (1995). The macromolecular crystallographic information file (mmCIF). *Methods Enzymol*. 277.

Branden, C., and Tooze, J. (1999). *Introduction to Protein Structure* (New York: Garland).

Brick, P., Bhat, T. N., and Blow, D. M. (1989). Structure of tyrosyl-tRNA synthetase refined at 2.3 angstroms resolution. Interaction of the enzyme with the tyrosyl adenylate intermediate. *J. Mol. Biol.* 208, 83–98.

Buckle, A. M., and Fersht, A. R. (1994). Subsite binding in an RNase: Structure of a barnase-tetranucleotide complex at 1.76-angstroms resolution. *Biochemistry* 33, 1644–1653.

Buckle, A. M., Henrick, K., and Fersht, A. R. (1993). Crystal structural analysis of mutations in the hydrophobic cores of barnase. *J. Mol. Biol.* 23, 847–860.

Bycroft, M., Ludvigsen, S., Fersht, A. R., and Poulsen, F. M. (1991). Determination of the three-dimensional solution structure of barnase using nuclear magnetic resonance spectroscopy. *Biochemistry* 30, 8697–8701.

Dahms, T., and Szabo, A. G. (1995). Conformational heterogeneity of tryptophan in a protein crystal. *J. Am. Chem. Soc.* 117, 2321–2326.

Fitzgerald, P. M., McKeever, B. M., Van Middlesworth, J. F., Springer, J. P., Heimbach, J. C., Leu, C. T., Herber, W. K., Dixon, R. A., and Darke, P. L. (1990). Crystallographic analysis of a complex between human immunodeficiency virus type 1 protease and acetyl-pepstatin at 2.0-angstroms resolution. *J. Biol. Chem.* 265, 14209–14219.

Gerstein, M., Lesk, A., and Chothia, C. (1994). Structural mechanisms for domain movements in proteins. *Biochemistry* 33, 6739–6749.

Gibrat, J. F., Madej, T., and Bryant, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol*. 6, 377–385.

Hall, S. R. (1991). The STAR file: A new format for electronic data transfer and archiving. *J. Chem. Inf. Comput. Sci*. 31, 326–333.

Hall, S. R., Allen, A. H., and Brown, I. D. (1991). The crystallographic information file (CIF): A new standard archive file for crystallography. *Acta Crystallogr. Sect. A* 47, 655–685.

Hogue, C. W. V., Ohkawa, H., and Bryant, S. H. (1996). A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database. *Trends Biochem. Sci.* 21, 226–229.

Hogue, C. W. V (1997) Cn3D: a new generation of three-dimensional molecular structure viewer. *Trends Biochem. Sci.* 22, 314–316.

Holm, L., and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233, 123–138.

Holm, L., and Sander, C. (1996). The FSSP database: Fold classification based on structure-structure alignment of proteins. *Nucl. Acids Res.* 24, 206–210.

Issacs, N. W., and Agarwal, R. C. (1978). Experience with fast Fourier least squares in the refinement of the crystal structure of rhombohedral 2 zinc insulin at 1.5 angstroms resolution. *Acta Crystallogr. Sect. A* 34, 782.

Kraulis, P. J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 24, 946–950.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 24, 536–540.

Quintana, J. R., Lipanov, A. A., and Dickerson, R. E. (1991). Low-temperature crystallographic analyses of the binding of Hoechst 33258 to the double-helical DNA dodecamer C-G-C-G-A-A-T-T-C-G-C-G. *Biochemistry* 30, 10294–10306.

Richardson, D. C., and Richardson, J. S. (1992). The Kinemage: A tool for scientific communication. *Protein Sci.* 1, 3–9.

Richardson, D. C., and Richardson, J. S. (1994). KinemagesSimple macromolecular graphics for interactive teaching and publication. *Trends Biochem. Sci.* 19, 135–138.

Rose, M. T. (1990). *The Open Book, A Practical Perspective on OSI* (Englewood Cliffs, NJ: Prentice-Hall), p. 227–322.

Sanchez-Ferrer, A., Nunez-Delicado, E., and Bru, R. (1995). Software for viewing biomolecules in three dimensions on the Internet. *Trends Biochem. Sci.* 20, 286–288.

Sayle, R. A., and Milner-White, E. J. (1995). RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci.* 20, 374–376.

Schuler, G. D., Epstein, J. A., Ohkawa, H., and Kans, J. A. (1996). Entrez: Molecular biology database and retrieval system. *Methods Enzymol.* 266, 141–162.

Walther D. (1997) WebMol—a Java based PDB viewer. *Trends Biochem. Sci.* 22, 274–275.

6

GENOMIC MAPPING AND MAPPING DATABASES

Peter S. White

Department of Pediatrics
University of Pennsylvania
Philadelphia, Pennsylvania

Tara C. Matise

Department of Genetics
Rutgers University
New Brunswick, New Jersey

A few years ago, only a handful of ready-made maps of the human genome existed, and these were low-resolution maps of small areas. Biomedical researchers wishing to localize and clone a disease gene were forced, by and large, to map their region of interest, a time-consuming and painstaking process. This situation has changed dramatically in recent years, and there are now high-quality genome-wide maps of several different types containing tens of thousands of DNA markers. With the pending availability of a finished human sequence, most efforts to construct genomic maps will come to a halt; however, integrated maps, genome catalogues, and comprehensive databases linking positional and functional genomic data will become even more valuable. Genome projects in other organisms are at various stages, ranging from having only a handful of available maps to having a complete sequence. By taking advantage of the available maps and DNA sequence, a researcher can, in many cases, focus in on a candidate region by searching public mapping databases in a matter of hours rather than by performing laboratory experiments over a course of months.

Subsequently, the researcher's burden has now shifted from mapping the genome to navigating a vast *terra incognita* of Web sites, FTP servers, and databases. There are large databases such as the National Center for Biotechnology Information (NCBI) Entrez Genomes Division, Genome Database (GDB), and Mouse Genome Database (MGD), smaller databases serving the primary maps published by genome centers, sites sponsored by individual chromosome committees, and sites used by smaller laboratories to publish highly detailed maps of specific regions. Each type of resource contains information that is valuable in its own right, even when it overlaps with the information found at others. Finding one's way around this information space is not easy. A recent search for the word ''genome'' using the AltaVista Web search engine turned up 400,000 potentially relevant documents.

This chapter is intended as a ''map of the maps,'' a way to guide readers through the maze of publicly available genomic mapping resources. The different types of markers and methods used for genomic mapping will be reviewed and the inherent complexities in the construction and utilization of genome maps will be discussed. Several large community databases and method-specific mapping projects will be presented in detail. Finally, practical examples of how these tools and resources can be used to aid in specific types of mapping studies such as localizing a new gene or refining a region of interest will be provided. A complete description of the mapping resources available for all species would require an entire book. Therefore, this chapter focuses primarily on humans, with some references to resources for other organisms.

INTERPLAY OF MAPPING AND SEQUENCING

The recent advent of whole-genome sequencing projects for humans and select model organisms is dramatically impacting the use and utility of genomic map-based information and methodologies. Genomic maps and DNA sequence are often treated as separate entities, but large, uninterrupted DNA sequence tracts can be thought of and used as an ultra-high-resolution mapping technique. Traditional genomic maps that rely on genomic markers and either clone-based or statistical approaches for ordering are precursory to finished and completely annotated DNA sequences of whole chromosomes or genomes. However, such completed genome sequences are predicted to be publicly available only in 2002 for humans, 2005 for the mouse, and even later for other mammalian species, although complete sequences are now available for some individual human chromosomes and selected lower eukaryotes (see Chapter 15). Until these completed sequences are available, mapping and sequencing approaches to genomic analysis serve as complementary approaches for chromosome analysis.

Before determination of an entire chromosome's sequence, the types of sequences available can be roughly grouped into marker/gene-based tags [e.g., expressed sequence tags (ESTs) and sequence-tagged sites (STSs)], single gene sequences, prefinished DNA clone sequences, and completed, continuous genomic sequence tracts. The first two categories provide rich sources of the genomic markers used for mapping, but only the last two categories can reliably order genomic elements. The human genome draft sequence is an example of a prefinished sequence, in which >90% of the entire sequence is available, but most continuous sequence tracts are relatively short (usually <100 kb and often <10 kb), thus providing high

local resolution but little long-range ordering information. Genomic maps can help provide a context for this sequence information. Thus, two or more sequences containing unique genomic markers can be oriented if these markers are ordered on a map. In this way, existing maps serve as a scaffold for orienting, directing, and troubleshooting sequencing projects. Similarly, users can first define a chromosomal region of interest using a traditional map approach and then can identify relevant DNA sequences to analyze by finding long sequences containing markers mapping within the defined region. NCBI tools such as BLAST and electronic PCR (e-PCR) are valuable for finding marker/sequence identities, and several of the resources discussed below provide marker/sequence integration.

As large sequence tracts emerge from the human and model organism projects, sequence-based ordering of genomic landmarks will eventually supplant map-based ordering methods. The evolution from a mapped chromosome to the determination of the chromosome's complete sequence is marked by increasing incorporation of partial genomic sequence tracts into the underlying map. Once complete, finished sequences can be used to confirm map-determined marker orders. Given the error rates inherent in both map and sequence-assembly methodology, it is good practice to use both map and sequence information simultaneously for independent verification of regional order.

GENOMIC MAP ELEMENTS

DNA Markers

A DNA *marker* is simply a uniquely identifiable segment of DNA. There are several different types of markers, usually ranging in size from one to 300–400 nucleotide bases in size. Markers can be thought of as landmarks, and a set of markers whose relative positions (or order) within a genome are known comprises a *map*. Markers can be categorized in several ways. Some markers are polymorphic, and others are not (monomorphic). Detection of markers may be either PCR based or hybridization based. Some markers lie in a sequence of DNA that is expressed; some do not, or their expression status may be unknown.

PCR-based markers are commonly referred to as sequence-tagged sites (STSs). An STS is defined as a segment of genomic DNA that can be uniquely PCR amplified by its primer sequences. STSs are commonly used in the construction of physical maps. STS markers may be developed from any genomic sequence of interest, such as from characterized and sequenced genes, or from expressed sequence tags (ESTs, Chapter 12). Alternatively, STSs may be randomly identified from total genomic DNA. The EST database (dbEST) at NCBI stores information on most STS markers.

Polymorphic Markers

Polymorphic markers are those that show sequence variation among individuals. Polymorphic markers are used to construct genetic linkage maps. The number of alleles observed in a population for a given polymorphism, which can vary from two to >30, determines the degree of polymorphism. For many studies, highly polymorphic markers (>5 alleles) are most useful.

Polymorphisms may arise from several types of sequence variations. One of the earlier types of polymorphic markers used for genomic mapping is a restriction fragment length polymorphism (RFLP). An RFLP arises from changes in the sequence of a restriction enzyme recognition site, which alters the digestion patterns observed during hybridization-based analysis. Another type of hybridization-based marker arises from a variable number of tandem repeat units (VNTR). A VNTR locus usually has several alleles, each containing a different number of copies of a common motif of at least 16 nucleotides tandemly oriented along a chromosome.

A third type of polymorphism is due to tandem repeats of short sequences that can be detected by PCR-based analysis. These are known variously as microsatellites, short tandem repeats (STRs), STR polymorphisms (STRPs), or short sequence length polymorphisms (SSLPs). These repeat sequences usually consist of two, three, or four nucleotides and are plentiful in most organisms. All PCR-converted STR markers (those for which a pair of oligonucleotides flanking the polymorphic site suitable for PCR amplification of the locus has been designed) are considered to be STSs. The advent of PCR-based analysis quickly made microsatellites the markers of choice for mapping.

Another polymorphic type of PCR-based marker is a single nucleotide polymorphism (SNP), which results from a base variation at a single nucleotide position. Most SNPs have only two alleles (biallelic). Because of their low heterozygosity, maps of SNPs require a much higher marker density than maps of microsatellites. SNPs occur frequently in most genomes, with one SNP occurring on average approximately once in every 100–300 bases in humans. SNPs lend themselves to highly automated fluidic or DNA chip-based analyses and have quickly become the focus of several large-scale development and mapping projects in humans and other organisms. Further details about all of these types of markers can be found elsewhere (Chakravarti and Lynn, 1999; Dietrich et al., 1999).

DNA Clones

The possibility of physically mapping eukaryotic genomes was largely realized with the advent of cloning vehicles that could efficiently and reproducibly propagate large DNA fragments. The first generation of large-insert cloning was made possible with yeast artificial chromosome (YAC) libraries (Burke et al., 1987). Because YACs can contain fragments up to 2 Mb, they are suitable for quickly making low-resolution maps of large chromosomal regions, and the first whole-genome physical maps of several eukaryotes were constructed with YACs. However, although YAC libraries work well for ordering STSs and for joining small physical maps, the high rate of chimerism and instability of these clones makes them unsuitable for DNA sequencing.

The second and current generation of large-insert clones consists of bacterial artificial chromosomes (BACs) and P1-artificial chromosomes, both of which act as episomes in bacterial cells rather than as eukaryotic artificial chromosomes. Bacterial propagation has several advantages, including higher DNA yields, ease-of-use for sequencing, and high integrity of the insert during propagation. As such, despite the relatively limited insert sizes (usually 100–300 kb), BACs and PACs have largely replaced YACs as the clones of choice for large-genome mapping and sequencing projects (Iaonnou et al., 1994; Shizuya et al., 1992). DNA fingerprinting has been

applied to BACs and PACs to determine insert overlaps and to construct clone contigs. In this technique, clones are digested with a restriction enzyme, and the resulting fragment patterns are compared between clones to identify those sharing subsets of identically sized fragments. In addition, the ends of BAC and PAC inserts can be directly sequenced; clones whose insert-end sequences have been determined are referred to as sequence-tagged clones (STCs). Both DNA fingerprinting and STC generation now play instrumental roles in physical mapping strategies, as will be discussed below.

TYPES OF MAPS

Cytogenetic Maps

Cytogenetic maps are those in which the markers are localized to chromosomes in a manner that can be directly imaged. Traditional cytogenetic mapping hybridizes a radioactively or fluorescently labeled DNA probe to a chromosome preparation, usually in parallel with a chromosomal stain such as Giemsa, which produces a banded karyotype of each chromosome (Pinkel et al., 1986). This allows assignment of the probe to a specific chromosomal band or region. Assignment of cytogenetic positions in this manner is dependent on some subjective criteria (variability in technology, methodology, interpretation, reproducibility, and definition of band boundaries). Thus, inferred cytogenetic positions are often fairly large and occasionally overinterpreted, and some independent verification of cytogenetic position determinations is warranted for crucial genes, markers, or regions. Probes used for cytogenetic mapping are usually large-insert clones containing a gene or polymorphic marker of interest. Despite the subjective aspects of cytogenetic methodology, karyotype analysis is an important and relatively simple clinical genetic tool; thus, cytogenetic positioning remains an important parameter for defining genes, disease loci, and chromosomal rearrangements.

Newer cytogenetic techniques such as interphase fluorescence in situ hybridization (FISH) (Lawrence et al., 1990) and fiber FISH (Parra and Windle, 1993) instead examine chromosomal preparations in which the DNA is either naturally or mechanically extended. Studies of such extended chromatin have demonstrated a directly proportional relationship between the distances measured on the image and the actual physical distance for short stretches, so that a physical distance between two closely linked probes can be determined with some precision (van den Engh et al., 1992). However, these techniques have a limited ordering range ($\leq$1–2 Mb) and are not well-suited for high-throughput mapping.

Genetic Linkage Maps

Genetic linkage (GL) maps (also called meiotic maps) rely on the naturally occurring process of recombination for determination of the relative order of, and map distances between, polymorphic markers. Crossover and recombination events take place during meiosis and allow rearrangement of genetic material between homologous chromosomes. The likelihood of recombination between markers is evaluated using genotypes observed in multigenerational families. Markers between which only a few

recombination occur are said to be linked, and such markers are usually located close to each other on the same chromosome. Markers between which many recombinations take place are unlinked and usually lie far apart, either at opposite ends of the same chromosome or on different chromosomes.

Because the recombination events cannot be easily quantified, a statistical method of maximum likelihood is usually applied in which the likelihood of two markers being linked is compared with the likelihood of being unlinked. This likelihood ratio is called a ''lod'' score (for ''log of the odds''), and a lod score greater than 3 (corresponding to odds of 1,000:1 or greater) is usually taken as evidence that markers are linked. The lod score is computed at a range of recombination fraction values between markers (from 0 to 0.5), and the recombination fraction at which the lod score is maximized provides an estimate of the distance between markers. A map function (usually either Haldane or Kosambi) is then used to convert the recombination fraction into an additive unit of distance measured in centiMorgans (cM), with 1 cM representing a 1% probability that a recombination has occurred between two markers on a single chromosome. Because recombination events are not randomly distributed, map distances on linkage maps are not directly proportional to physical distances.

The majority of linkage maps are constructed using multipoint linkage analysis, although multiple pairwise linkage analysis and minimization of recombination are also valid approaches. Commonly used and publicly available computer programs for building linkage maps include LINKAGE (Lathrop et al., 1984), CRI-MAP (Lander and Green, 1987), MultiMap (Matise et al., 1994), MAPMAKER (Lander et al., 1987), and MAP (Collins et al., 1996). The MAP-O-MAT Web server is available for estimation of map distances and for evaluation of statistical support for order (Matise and Gitlin, 1999).

Because linkage mapping is a based on statistical methods, linkage maps are not guaranteed to show the correct order of markers. Therefore, it is important to be critical of the various available maps and to be aware of the statistical criteria that were used in map construction. Typically, only a subset of markers (framework or index markers) is mapped with high statistical support. The remainder are either placed into well-supported intervals or bins or placed into unique map positions but with low statistical support for order (see additional discussion below).

To facilitate global coordination of human linkage mapping, DNAs from a set of reference pedigrees collected for map construction were prepared and distributed by the Centre d'Etude du Polymorphism Humain (CEPH; Dausset et al., 1990). Nearly all human linkage maps are based on genotypes from the CEPH reference pedigrees, and genotypes for markers scored in the CEPH pedigrees are deposited in a public database maintained at CEPH. Most recent maps are composed almost entirely of highly polymorphic STR markers. These linkage maps have already exceeded the maximum map resolution possible given the subset of CEPH pedigrees that are commonly used for map construction, and no further large-scale efforts to place STR markers on human linkage maps are planned. Thousands of SNPs are currently being identified and characterized, and a subset are being placed on linkage maps (Wang et al., 1998).

Linkage mapping is also an important tool in experimental animals, with many maps already produced at high resolution and others still under development (see *Mapping Projects and Associated Resources*, below).

Radiation Hybrid Maps

Radiation hybrid (RH) mapping is very similar to linkage mapping. Both methods rely on the identification of chromosome breakage and reassortment. The primary difference is the mechanism of chromosome breakage. In the construction of radiation hybrids, breaks are induced by the application of lethal doses of radiation to a donor cell line, which is then rescued by fusion with a recipient cell line (typically mouse or hamster) and grown in a selective medium such that only fused cells survive. An RH panel is a library of fusion cells, each of which has a separate collection of donor fragments. The complete donor genome is represented multiple times across most RH panels. Each fusion cell, or radiation hybrid, is then scored by PCR to determine the presence or absence of each marker of interest. Markers that physically lie near each other will show similar patterns of retention or loss across a panel of RH cells and behave as if they are linked, whereas markers that physically lie far apart will show completely dissimilar patterns and behave as if they are unlinked. Because the breaks are largely randomly distributed, the break frequencies are roughly directly proportional to physical distances. The resulting data set is a series of positive and negative PCR scores for each marker across the hybrid panel.

These data can be used to statistically infer the position of chromosomal breaks, and, from that point on, the procedures for map construction are similar to those used in linkage mapping. A map function is used to convert estimates of breakage frequency to additive units of distance measured in centirays (cR), with 1 cR representing a 1% probability that a chromosomal break has occurred between two markers in a single hybrid. The resolution of a radiation hybrid map depends on the size of the chromosomal fragments contained in the hybrids, which in turn is proportional to the amount of irradiation to which the human cell line was exposed.

Most RH maps are built using multipoint linkage analysis, although multiple-pairwise linkage analysis and minimization of recombination are also valid approaches. Three genome-wide RH panels exist for humans and are commercially available, and RH panels are available for many other species as well. Widely used computer programs for RH mapping are RHMAP (Boehnke et al., 1991), RHMAPPER (Slonim et al., 1997), and MultiMap (Matise et al., 1994), and on-line servers that allow researchers to place their RH mapped markers on existing RH maps are available. The Radiation Hybrid Database (RHdb) is the central repository for RH data on panels available in all species. The Radiation Hybrid Information Web site also contains multi-species information about available RH panels, maps, ongoing projects, and available computer programs.

Transcript Maps

Of particular interest to researchers chasing disease genes are maps of transcribed sequences. Although the transcript sequences are mapped using one of the methods described in this section, and thus do not require a separate mapping technology, they are often set apart as a separate type of map. These maps consist of expressed sequences and sequences derived from known genes that have been converted into STSs and usually placed on conventional physical maps. Recent projects for creating large numbers of ESTs (Adams et al., 1991; Houlgatte et al., 1995; Hillier et al., 1996) have made tens of thousands of unique expressed sequences available to the

mapping laboratories. Transcribed sequence maps can significantly speed the search for candidate genes once a disease locus has been identified. The largest human transcript map to date is the GeneMap '99, described below.

Physical Maps

Physical maps include maps that either are capable of directly measuring distances between genomic elements or that use cloned DNA fragments to directly order elements. Many techniques have been created to develop physical maps. The most widely adopted methodology, due largely to its relative simplicity, is STS content mapping (Green and Olson, 1990). This technique can resolve regions much larger than 1 Mb and has the advantage of using convenient PCR-based positional markers.

In STS content maps, STS markers are assayed by PCR against a library of large-insert clones. If two or more STSs are found to be contained in the same clone, chances are high that those markers are located close together. (The fact that they are not close 100% of the time is a reflection of various artifacts in the mapping procedure, such as the presence of chimeric clones.) The STS content mapping technique builds a series of contigs (i.e., overlapping clusters of clones joined together by shared STSs). The resolution and coverage of such a map are determined by a number of factors, including the density of STSs, the size of the clones, and the depth of the clone library. Maps that use cloning vectors with smaller insert sizes have a higher theoretical resolution but require more STSs to achieve coverage of the same area of the genome. Although it is generally possible to deduce the relative order of markers on STS content maps, the distances between adjacent markers cannot be measured with accuracy without further experimentation, such as by restriction mapping. However, STS content maps have the advantage of being associated with a clone resource that can be used for further studies, including subcloning, DNA sequencing, or transfection.

Several other techniques in addition to STS content and radiation hybrid mapping have also been used to produce physical maps. Clone maps rely on techniques other than STS content to determine the adjacency of clones. For example, the CEPH YAC map (see below) used a combination of fingerprinting, inter-Alu product hybridization, and STS content to create a map of overlapping YAC clones. Fingerprinting is commonly used by sequencing centers to assemble and/or verify BAC and PAC contigs before clones are chosen for sequencing, to select new clones for sequencing that can extend existing contigs, and to help order genomic sequence tracts generated in whole-genome sequencing projects (Chumakov et al., 1995). Sequencing of large-insert clone ends (STC generation), when applied to a whole-genome clone library of adequate coverage, is very effective for whole-genome mapping when used in combination with fingerprinting of the same library. Deletion and somatic cell hybrid maps relying on large genomic reorganizations (induced deliberately or naturally occurring) to place markers into bins defined by chromosomal breakpoints have been generated for some human chromosomes (Jensen et al., 1997; Lewis et al., 1995; Roberts et al., 1996; Vollrath et al., 1992). Optical mapping visualizes and measures the length of single DNA molecules extended and digested with restriction enzymes by high-resolution microscopy. This technique, although still in its infancy, has been successfully used to assemble whole chromosome maps of bacteria and lower eukaryotes and is now being applied to complex genomes (Aston et al., 1999; Jing et al., 1999; Schwartz et al., 1993).

Comparative Maps

Comparative mapping is the process of identifying conserved chromosome segments across different species. Because of the relatively small number of chromosomal breaks that have occurred during mammalian radiation, the order of genes usually is preserved over large chromosomal segments between related species. Orthologous genes (copies of the same genes from different species) can be identified through DNA sequence homology, and sets of orthologous genes sharing an identical linear order within a chromosomal region in two or more species are used to identify conserved segments and ancient chromosomal breakpoints.

Knowledge about which chromosomal segments are shared and how they have become rearranged over time greatly increases our understanding of the evolution of different plant and animal lineages. One of the most valuable applications of comparative maps is to use an established gene map of one species to predict positions of orthologous genes in another species. Many animal models exist for diseases observed in humans. In some cases, it is easier to identify the responsible genes in an animal model than in humans, and the availability of a good comparative map can simplify the process of identifying the responsible genes in humans. In other cases, more might be known about the gene(s) responsible in humans, and the same comparative map could be used to help identify the gene(s) responsible in the model species. There are several successful examples of comparative candidate gene mapping (O'Brien et al., 1999).

As mapping and sequencing efforts progress in many species, it is becoming possible to identify smaller homologous chromosome segments, and detailed comparative maps are being developed between many different species. Fairly dense gene-based comparative maps now exist between the human, mouse, and rat genomes and also between several agriculturally important mammalian species. Sequence- and protein-based comparative maps are also under development for several lower organisms for which complete sequence is available (Chapter 15). A comparative map is typically presented either graphically or in tabular format, with one species designated as the index species and one or more others as comparison species. Homologous regions are presented graphically with nonconsecutive segments from the comparison species shown aligned with their corresponding segments along the map of the index species.

Integrated Maps

Map integration provides interconnectivity between mapping data generated from two or more different experimental techniques. However, achieving accurate and useful integration is a difficult task. Most of the genomic maps and associated Web sites discussed in this section provide some measure of integration, ranging from the approximate cytogenetic coordinates provided in the Généthon GL map to the inter-associated GL, RH, and physical data provided by the Whitehead Institute (WICGR) Web site. Several integration projects have created truly integrated maps by placing genomic elements mapped by differing techniques relative to a single map scale. The most advanced sources of genomic information provide some level of genomic cataloguing, where considerable effort is made to collect, organize, and map all available positional information for a given genome.

COMPLEXITIES AND PITFALLS OF MAPPING

It is important to realize that the genomic mapping information currently available is a collection of a large number of individual data sets, each of which has unique characteristics. The experimental techniques, methods of data collection, annotation, presentation, and quality of the data differ considerably among these data sets. Although most mapping projects include procedures to detect and eliminate and/or correct errors, there are invariably some errors that occur, which often result in the incorrect ordering or labeling of individual markers. Although the error rate is usually very low (5% or less), a marker misplacement can obviously have a great impact on a study. A few mapping Web sites are beginning to flag and correct (or at least warn) users of potential errors, but most errors cannot be easily detected. Successful strategies for minimizing the effects of data error include (1) simultaneously assessing as many different maps as possible to maximize redundancy (note that ideally ''different'' maps use independently-derived data sets or different techniques); (2) increased emphasis on utilizing integrated maps and genomic catalogues that provide access to all available genomic information for the region of interest (while closely monitoring the map resolution and marker placement confidence of the integrated map); and (3) if possible, experimentally verifying the most critical marker positions or placements.

In addition to data errors, several other, more subtle complexities are notable. Foremost is the issue of nomenclature, or the naming of genomic markers and elements. Many markers have multiple names, and keeping track of all the names is a major bioinformatics challenge. For example, the polymorphic marker D1S243 has several assigned names: AFM214yg7, which is actually the name of the DNA clone from which this polymorphism was identified; SHGC-428 and stSG729, two examples of genome centers renaming a marker to fit their own nomenclature schemes; and both GDB:201358 and GDB:133491, which are database identifier numbers used to track the polymorphism and STS associated with this marker, respectively, in the Genome Database (GDB). Genomic mapping groups working with a particular marker often assign an additional name to simplify their own data management, but, too often, these alternate identifiers are subsequently used as a primary name. Furthermore, many genomic maps display only one or a few names, making comparisons of maps problematic. Mapping groups and Web sites are beginning to address these inherent problems, but the difficulty of precisely defining ''markers,'' ''genes,'' and ''genomic elements'' adds to the confusion. It is important to distinguish between groups of names defining different elements. A gene can have several names, and it can also be associated with one or more EST clusters, polymorphisms, and STSs. Genes spanning a large genomic stretch can even be represented by several markers that individually map to different positions. Web sites providing genomic cataloguing, such as LocusLink, UniGene, GDB, GeneCards, and eGenome, list most names associated with a given genomic element. Nevertheless, collecting, cross-referencing, and frequently updating one's own sets of names for markers of interest is also a good practice (see Chapter 4 for data management using Sequin), as even the genomic cataloguing sites do not always provide complete nomenclature collections.

Each mapping technique yields its own resolution limits. Cytogenetic banding potentially orders markers separated by $\geq 1-2$ Mb, and genetic linkage (GL) and RH analyses yields long-range resolutions of $\geq 0.5-1$ Mb, although localized ordering can achieve higher resolutions. The confidence level with which markers are

ordered on statistically based maps is often overlooked, but this is crucial for assessing map quality. For genomes with abundant mapping data such as human or mouse, the number of markers used for mapping often far exceeds the ability of the technique to order all markers with high confidence (often, confidence levels of 1,000:1 or lod 3 are used as a cutoff, which usually means that a marker is $\geq$1,000:1 times more likely to be in the given position than in any other). Mappers have taken two approaches to address this issue. The first is to order all markers in the best possible linear order, regardless of the confidence for map position of each marker [examples include GeneMap '99 (GM99) and the Genetic Location Database; Collins et al., 1996; Deloukas et al., 1998]. Alternatively, the high confidence linear order of a subset of markers is determined, and the remaining markers are then placed in high confidence ''intervals,'' or regional positions (such as Généthon, SHGC, and eGenome; Dib et al., 1996; Stewart et al., 1997; White et al., 1999). The advantage of the first approach is that resolution is maximized, but it is important to pay attention to the odds for placement of individual markers, as alternative local orders are often almost equally likely. Thus, beyond the effective resolving power of a mapping technique, increased resolution often yields decreased accuracy, and researchers are cautioned to strike a healthy balance between the two.

Each mapping technique also yields very different measures of distance. Cytogenetic approaches, with the exception of high-resolution fiber FISH, provide only rough distance estimates, GL and STS content mapping provide marker orientation but only relative distances, and RH mapping yields distances roughly proportional to true physical distance. For GL analysis, unit measurements are in centMiorgans, with 1 cM equivalent to a 1% chance of recombination between two linked markers. The conversion factor of 1 cM $\simeq$ 1 Mb is often cited for the human genome but is overstated, as this is just the *average* ratio genome-wide, and many chromosomal regions have recombination hotspots and coldspots in which the cM-to-Mb ratio varies as much as 10-fold. In general, cytogenetic maps provide subband marker regionalization but limited localized ordering, GL and STS content maps provide excellent ordering and limited-to-moderate distance information, and RH maps provide the best combination of localized ordering and distance estimates.

Finally, there are various levels at which genomic information can be presented. *Single-resource maps* such as the Généthon GL maps use a single experimental technique and analyze a homogeneous set of markers. Strictly *comparative maps* make comparisons between two or more different single-dimension maps either within or between species but without combining data sets for integration. GDB's Mapview program can display multiple maps in this fashion (Letovsky et al., 1998). *Integrated maps* recalculate or completely integrate multiple data sets to display the map position of all genomic elements relative to a single scale; GDB's Comprehensive Maps are an example of such integration (Letovsky et al., 1998). Lastly, *genome cataloguing* is a relatively new way to display genomic information, in which many data sets and/or Web sites are integrated to provide a comprehensive listing and/or display of all identified genomic elements for a given chromosome or genome. Completely sequenced genomes such as *C. elegans* and *S. cerevisiae* have advanced cataloguing efforts (see Chapter 15), but catalogues for complex genome organisms are in the early stages. Examples include the interconnected NCBI databases, MGD, and eGenome (Blake et al., 2000; Wheeler et al., 2000). Catalogues provide a ''one-stop shopping'' solution to collecting and analyzing genomic data and are recommended as a maximum-impact means to begin a regional analysis. However, the

individual data sets provide the highest quality positional information and are ultimately the most useful for region definition and refinement.

DATA REPOSITORIES

There are several valuable and well-developed data repositories that have greatly facilitated the dissemination of genome mapping resources for humans and other species. This section covers three of the most comprehensive resources for mapping in humans: the Genome Database (GDB), the National Center for Biotechnology Information (NCBI), and the Mouse Genome Database (MGD). More focused resources are mentioned in the *Mapping Projects and Associated Resources* section of this chapter.

GDB

The Genome Database (GDB) is the official central repository for genomic mapping data created by the Human Genome Project (Pearson, 1991). GDB's central node is located at the Hospital for Sick Children (Toronto, Ontario, Canada). Members of the scientific community as well as GDB staff curate data submitted to the GDB. Currently, GDB comprises descriptions of three types of objects from humans: Genomic Segments (genes, clones, amplimers, breakpoints, cytogenetic markers, fragile sites, ESTs, syndromic regions, contigs, and repeats), Maps (including cytogenetic, GL, RH, STS-content, and integrated), and Variations (primarily relating to polymorphisms). In addition, contributing investigator contact information and citations are also provided. The GDB holds a vast quantity of data submitted by hundreds of investigators. Therefore, like other large public databases, the data quality is variable. A more detailed description of the GDB can be found in Talbot and Cuticchia (1994).

GDB provides a full-featured query interface to its database with extensive online help. Several focused query interfaces and predefined reports, such as the Maps within a Region search and Lists of Genes by Chromosome report, present a more intuitive entry into GDB. In particular, GDB's Mapview program provides a graphical interface to the genetic and physical maps available at GDB.

A Simple Search is available on the home page of the GDB Web site. This query is used when searching for information on a specific genomic segment, such as a gene or STS (amplimer, in GDB terminology) and can be implemented by entering the segment name or GDB accession number. Depending on the type of segment queried and the available data, many different types of segment-specific information may be returned, such as alternate names (aliases), primer sequences, positions in various maps, related segments, polymorphism details, contributor contact information, citations, and relevant external links.

At the bottom of the GDB home page is a link to Other Search Options. From the Other Search Options page there are links to three customized search forms (Markers and Genes within a Region, Maps within a Region, and Genes by Name or Symbol), sequence-based searches, specific search forms for subclasses of GDB elements, and precompiled lists of data (Genetic Diseases by Chromosome, Lists of Genes by Chromosome, and Lists of Genes by Symbol Name).

A particularly useful query is the Maps within a Region search. This search allows retrieval of all maps stored in GDB that span a defined chromosomal region.

In a two-step process, the set of maps to be retrieved is first determined, and, from these, the specific set to be displayed is then selected.

Select the Maps within a Region link to display the search form. To view an entire chromosome, simply select it from the pop-up menu. However, entire chromosomes may take considerable time to download and display; therefore, it is usually best to choose a subchromosomal region. To view a chromosomal region, type the names of two cytogenetic bands or flanking genetic markers into the text fields labeled From and To. An example query is shown in Figure 6.1. If the flanking markers used in the query are stored in GDB as more than one type of object, the next form will request selection of the specific type of element for each marker. For the example shown in Figure 6.1, it is appropriate to select Amplimer.

The resulting form lists all maps stored in GDB that overlap the selected region. Given the flanking markers specified above, there are a total of 21 maps. The user selects which maps to display by marking the respective checkboxes. Note that GDB's Comprehensive Map is automatically selected. If a graphical display is requested, the size of the region and the number of maps to be displayed can significantly affect the time to fetch and display them. The resulting display will appear in a separate window showing the selected maps in side-by-side fashion.

While the Mapview display is loading, a new page is shown in the browser window. If your system is not configured to handle Java properly, a helpful message will be displayed in the browser window. (*Important*: Do not close the browser window behind Mapview. Because of an idiosyncrasy of Java's security specification, the applet cannot interact properly with GDB unless the browser window remains open.) To safely exit the Mapview display, select Exit from Mapview's File menu.

Mapview has many useful options, which are well described in the online help. Some maps have more than one *tier*, each displaying different types of markers, such as markers positioned with varying confidence thresholds on a linkage or radiation hybrid map. It is possible to zoom in and out, highlight markers across maps, color code different tiers, display markers using different aliases, change the relative position of the displayed maps, and search for specific markers. To retrieve additional information on a marker from any of the maps, double-click on its name to perform a *Simple Search* (as described above). A separate browser window will then display the GDB entry for the selected marker.

Two recently added GDB tools are GDB BLAST and e-PCR. These are available from the Other Search Options page and enable users to employ GDB's many data resources in their analysis of the emerging human genome sequence. GDB BLAST returns GDB objects associated with BLAST hits against the public human sequence. GDB's e-PCR finds which of its many amplimers are contained within queried DNA sequences and is thereby a quick means to determine or refine gene or marker localization. In addition, the GDB has many useful genome resource Web links on its Resources page.

NCBI

The NCBI has developed many useful resources and tools, several of which are described throughout this book. Of particular relevance to genome mapping is the Genomes Division of Entrez. Entrez provides integrated access to several different types of data for over 600 organisms, including nucleotide sequences, protein structures and sequences, PubMed/MEDLINE, and genomic mapping information. The

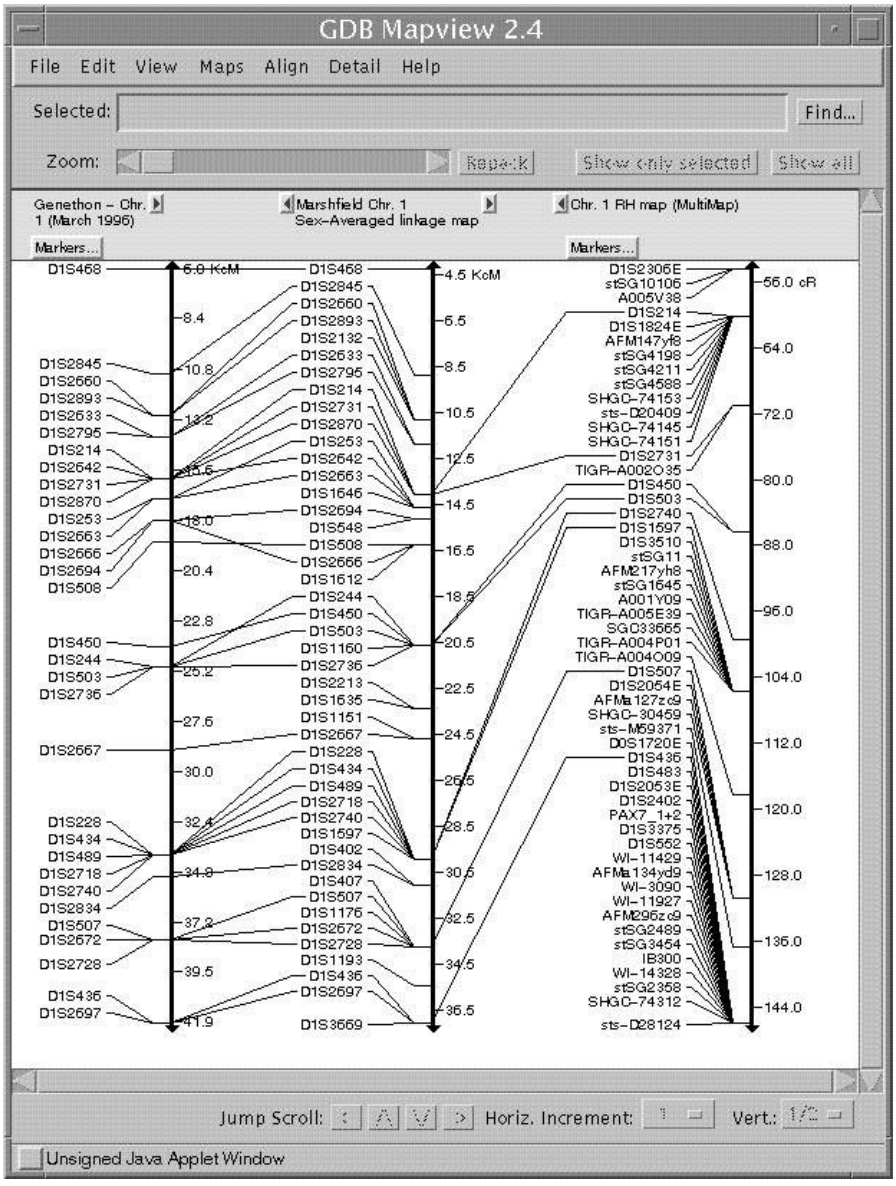

Figure 6.1. Results of a *Maps within a Region* GDB query for the region *D1S468–D1S214,* with no limits applied to the types of maps to be retrieved. Twenty-one maps were available for display. Only the Genethon and Marshfield linkage maps, as well as the Chromosome 1 RH map were selected for graphical display. Markers that are shared across maps are connected by lines.

NCBI Human Genome Map Viewer is a new tool that presents a graphical view of the available human genome sequence data as well as cytogenetic, genetic, physical, and radiation hybrid maps. Because the Map Viewer provides displays of the human genome sequence for the finished contigs, the BAC tiling path of finished and draft sequence, and the location of genes, STSs, and SNPs on finished and draft sequences,

it is an especially useful tool for integrating maps and sequence. The only other organisms for which the Map Viewer is currently available is *M. musculus* and *D. melanogaster*.

The NCBI Map Viewer can simultaneously display up to seven maps that are selected from a set of 19, including cytogenetic, linkage, RH, physical, and sequence-based maps. Some of the maps have been previously published, and others are being computed at NCBI. An extensive set of help pages is available. There are many different paths to the Map Vieweron the NCBI Web site, as described in the help pages. The Viewer supports genome-wide or chromosome-specific searches.

A good starting point is the *Homo sapiens* Genome View page. This is reached from the NCBI home page by connecting to Human Genome Resources (listed on the right side), followed by the link to the Map Viewer (listed on the left side). From the Genome View page, a genome-wide search may be initiated using the search box at the top left, or a chromosome-specific search may be performed by entering a chromosome number(s) in the top right search box or by clicking on a chromosome idiogram. The searchable terms include gene symbol or name and marker name or alias. The search results include a list of hits for the search term on the available maps. Clicking on any of the resulting items will bring up a graphical view of the region surrounding the item on the specific map that was selected. For example, a genome-wide search for the term `CMT*` returns 33 hits, representing the loci for forms of Charcot-Marie-Tooth neuropathy on eight different chromosomes. Selecting the Genes_seq link for the PMP22 gene (the gene symbol for CMT1A, on chromosome 17) returns the view of the sequence map for the region surrounding this gene. The Display Settings window can then be used to select simultaneous display of additional maps (Fig. 6.2).

The second search box at the top right may be used to limit a genome-wide search to a single chromosome or range of chromosomes. Alternatively, to browse an entire chromosome, click on the link below each idiogram. Doing so will return a graphical representation of the chromosome using the default display settings. Currently, the default display settings select the STS map (shows placement of STSs using electronic PCR), the GenBank map (shows the BAC tiling path used for sequencing), and the contig map (shows the contig map assembled at NCBI from finished high-throughput genomic sequence) as additional maps to be displayed. To select a smaller region of interest from the view of the whole chromosome, either define the range (using base pairs, cytogenetic bands, gene symbols or marker names) in the main Map Viewer window or in the display settings or click on a region of interest from the thumbnail view graphic in the sidebar or the map view itself. As with the GDB map views, until all sequence is complete, alignment of multiple maps and inference of position from one map to another must be judged cautiously and should not be overinterpreted (see Complexities and Pitfalls of Mapping section above).

There are many other tools and databases at NCBI that are useful for gene mapping projects, including e-PCR, BLAST (Chapter 8), the GeneMap ’99 (see Mapping Projects and Associated Resources), and the LocusLink, OMIM (Chapter 7), dbSTS, dbSNP, dbEST (Chapter 12), and UniGene (Chapter 12) databases. e-PCR and BLAST can be used to search DNA sequences for the presence of markers and to confirm and refine map localizations. In addition to EST alignment information and DNA sequence, UniGene reports include cytogenetic and RH map locations. The GeneMap ’99 is a good starting point for finding approximate map

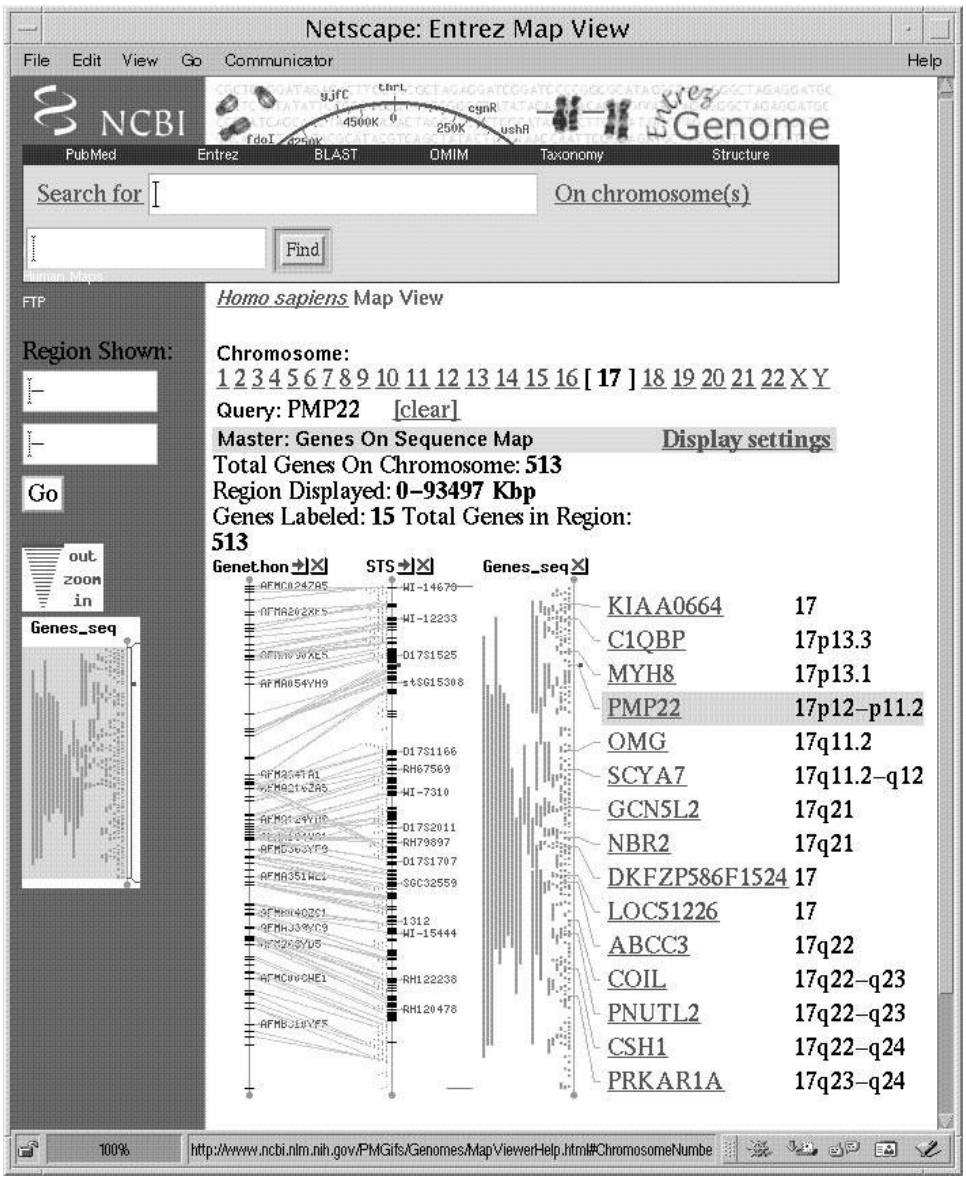

Figure 6.2. NCBI's Map View of the region surrounding the PMP22 gene. The Généthon, STS, and Genes_seq maps are displayed with lines connecting markers in common.

positions for EST markers, although additional fine-mapping should be performed to confirm order in critical regions. LocusLink, OMIM, and UniGene are good starting points for genome catalog information about genes and gene-based markers. LocusLink (Pruitt et al., 2000) presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related Web sites. The dbSTS and dbEST databases themselves play a lesser role in human and mouse gene mapping endeavors as their relevant information has already been captured by other more detailed resources

(such as LocusLink, GeneMap '99, UniGene, MGD, and eGenome) but are currently the primary source of genomic information for other organisms. The dbSNP database stores population-specific information on variation in humans, primarily for single nucleotide repeats but also for other types of polymorphisms. In addition, the NCBI's Genomic Biology page provides genomic resource home pages for many other organisms, including mouse, rat, *Drosophila*, and zebrafish.

MGI/MGD

The Mouse Genome Initiative Database (MGI) is the primary public mouse genomic catalogue resource. Located at The Jackson Laboratory, the MGI currently encompasses three cross-linked topic-specific databases: the Mouse Genome Database (MGD), the mouse Gene Expression Database (GXD), and the Mouse Genome Sequence project (MGS). The MGD has evolved from a mapping and genetics resource to include sequence and genome information and details on the functions and roles of genes and alleles (Blake et al., 2000). MGD includes information on mouse genetic markers and nomenclature, molecular segments (probes, primers, YACs and MIT primers), phenotypes, comparative mapping data, graphical displays of linkage, cytogenetic, and physical maps; experimental mapping data, and strain distribution patterns for recombinant inbred strains (RIs) and cross haplotypes. As of November 2000, there were over 29,500 genetic markers and 11,600 genes in MGD, with 85% and 70% of these placed onto the mouse genetic map, respectively. Over 4,800 genes have been matched with their human ortholog and over 1,800 matched with their rat ortholog.

Genes are easily searched through the Quick Gene Search box on the MGD home page. Markers and other map elements may also be accessed through several other search forms. The resulting pages contain summary information such as element type, official symbol, name, chromosome, map positions, MGI accession ID, references, and history. Additional element-specific information may also be displayed, including links to outside resources (Fig. 6.3). A thumbnail linkage map of the region is shown to the right, which can be clicked on for an expanded view.

The MGD contains many different types of maps and mapping data, including linkage data from 13 different experimental cross panels and the WICGR mouse physical maps, and cytogenetic band positions are available for some markers. The MGD also computes a linkage map that integrates markers mapped on the various panels. A very useful feature is the ability to build customized maps of specific regions using subsets of available data, incorporating private data, and showing homology information where available (see *Comparative Resources* section below). The MGD is storing radiation hybrid scores for mouse markers, but to date, no RH maps have been deposited at MGD.

MAPPING PROJECTS AND ASSOCIATED RESOURCES

In addition to the large-scale mapping data repositories outlined in the previous section, many invaluable and more focused resources also exist. Some of these are either not appropriate for storage at one of the larger-scale repositories or have never been deposited in them. These are often linked to specific mapping projects that primarily use only one or a few different types of markers or mapping approaches.

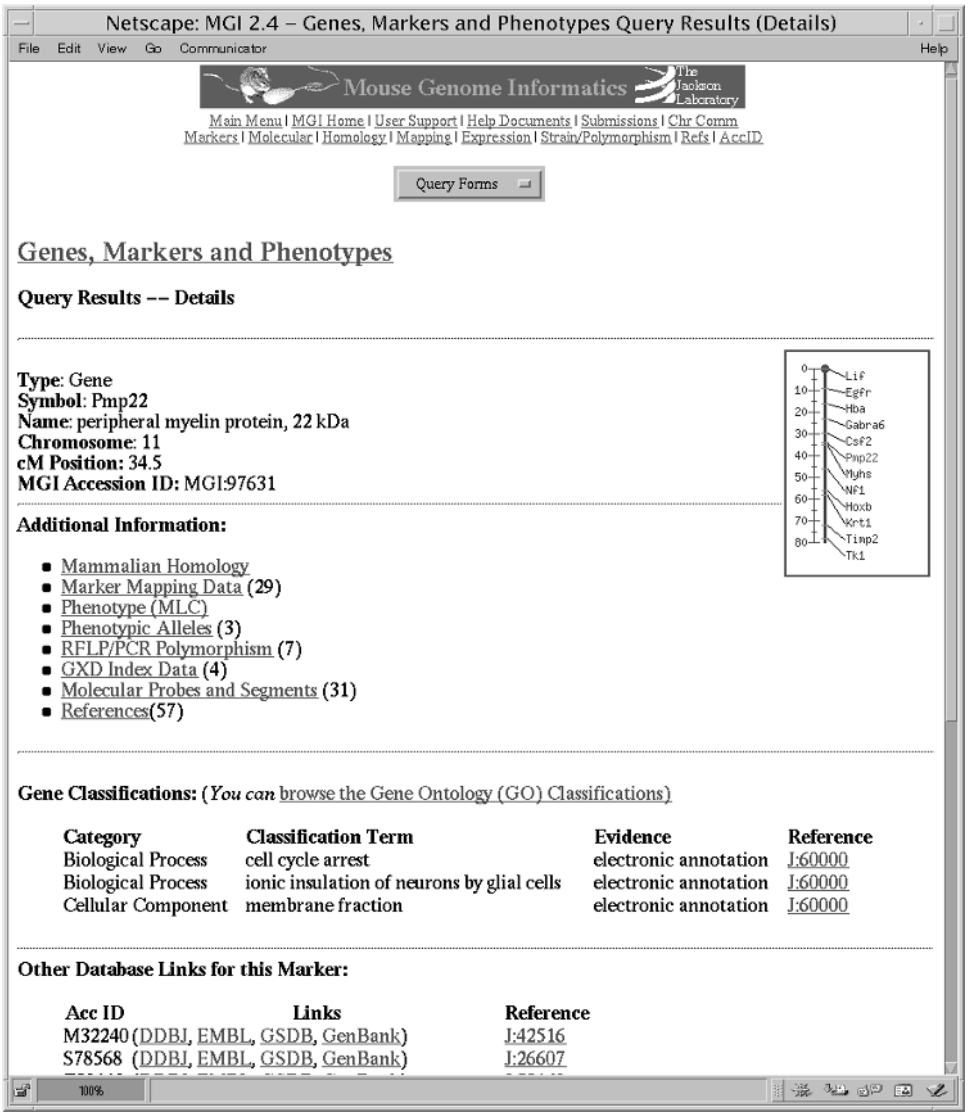

Figure 6.3. Results of an MGD Quick Gene Search for *pmp22*.

For most studies requiring the use of genome maps, it remains necessary to obtain maps or raw data from one or more of these additional resources. By visiting the resource-specific sites outlined in this section, it is usually possible to view maps in the form preferred by the originating laboratory, download the raw data, and review the laboratory protocols used for map construction.

Cytogenetic Resources

Cytogenetic-based methodologies are instrumental in defining inherited and acquired chromosome abnormalities, and (especially gene-based) chromosomal mapping data is often expressed in cytogenetic terms. However, because cytogenetic markers are

not sequence based and the technique is less straightforward and usually more subjective than GL, RH, or physical mapping, there is only a modicum of integration between chromosomal band assignments and map coordinates derived from other techniques in humans and very little or none in other species. Thus, it is often difficult to determine the precise cytogenetic location of a gene or region. Useful human resources can be divided into displays of primary cytogenetic mapping data, efficient methods of integrating cytogenetic and other mapping data, and resources pertaining to specific chromosomal aberrations.

The central repository for human cytogenetic information is GDB, which offers several ways to query for marker and map information using cytogenetic coordinates (see above). GDB is a useful resource for cross-referencing cytogenetic positions with genes or regions of interest. NCBI's LocusLink and UniGene catalogues, as well as their other integrated mapping resources, are also valuable repositories of cytogenetic positions. LocusLink and NCBI's Online Mendelian Inheritance in Man (OMIM) list cytogenetic positions for all characterized genes and genetic abnormalities, respectively (McKusick, 1998; Pruitt et al., 2000). The National Cancer Institute (NCI)-sponsored project to identify GL-tagged BAC clones at 1 Mb density throughout the genome is nearing completion. This important resource, which is commercially available both as clone sets and as individual clones, provides the first complete integration of cytogenetic band information with other genome maps. At this site, BACs can be searched for individually by clone name, band position, or contained STS name, and chromosome sets are also listed. Each clone contains one or more microsatellite markers and has GL and/or RH mapping coordinates along with a FISH-determined cytogenetic band assignment. This information can be used to quickly determine the cytogenetic position of a gene or localized region and to map a cytogenetic observation such as a tumor-specific chromosomal rearrangement using the referenced GL and physical mapping reagents.

Three earlier genome-wide efforts to cytogenetically map large numbers of probes are complementary to the NCI site. The Lawrence Berkeley National Laboratory-University of California, San Francisco, Resource for Molecular Cytogenetics has mapped large-insert clones containing polymorphic and expressed markers using FISH to specific bands and also with fractional length (flpter) coordinates, in which the position of a marker is measured as a percentage of the length of the chromosome's karyotype. Similarly, the Genetics Institute at the University of Bari, Italy, and the Max Planck Institute for Molecular Genetics have independently localized large numbers of clones, mostly YACs containing GL-mapped microsatellite markers, onto chromosome bands by FISH. All three resources have also integrated the mapped probes relative to existing GL and/or RH maps.

Many data repositories and groups creating integrated genome maps list cytogenetic localizations for mapped genomic elements. These include GDB, NCBI, the Unified Database (UDB), the Genetic Location Database (LDB), and eGenome, all of which infer approximate band assignments to many or all markers in their databases. These assignments rely on determination of the approximate boundaries of each band using subsets of their marker sets for which accurate cytogenetic mapping data are available.

The NCI's Cancer Chromosome Aberration Project (CCAP; Wheeler et al., 2000), Infobiogen (Wheeler et al., 2000), the Southeastern Regional Genetics Group (SERGG), and the Coriell Cell Repositories all have Web sites that display cytogenetic maps or descriptions of characterized chromosomal rearrangements. These sites

are useful resources for determining whether a specific genomic region is frequently disrupted in a particular disease or malignancy and for finding chromosomal cell lines and reagents for regional mapping. However, most of these rearrangements have only been mapped at the cytogenetic level.

Nonhuman resources are primarily limited to displays or simple integrations of chromosome idiograms. ArkDB is an advanced resource for displaying chromosomes of many amniotes; MGD incorporates mouse chromosome band assignments into queries of its database; and the Animal Genome Database has clickable chromosome idiograms for several mammalian genomes (Wada and Yasue, 1996). A recent work linking the mouse genetic and cytogenetic maps consists of 157 BAC clones distributed genome-wide (Korenberg et al., 1999) and an associated Web site is available for this resource at the Cedars-Sinai Medical Center.

Genetic Linkage Map Resources

Even with the ''sequence era'' approaching rapidly, linkage maps remain one of the most valuable and widely used genome mapping resources. Linkage maps are the starting point for many disease-gene mapping projects and have served as the backbone of many physical mapping efforts. Nearly all human linkage maps are based on genotypes from the standard CEPH reference pedigrees. There are three recent sets of genome-wide GL maps currently in use, all of which provide high-resolution, largely accurate, and convenient mapping information. These maps contain primarily the conveniently genotyped PCR-based microsatellite markers, use genotypes for only 8–15 of the 65 available CEPH pedigrees, and contain few, if any, gene-based or cytogenetically mapped markers. Many chromosome-specific linkage maps have also been constructed, many of which use a larger set of CEPH pedigrees and include hybridization- and gene-based markers. Over 11,000 markers have been genotyped in the CEPH pedigrees, and these genotypes have been deposited into the CEPH genotype database and are publicly available.

The first of the three genome-wide maps was produced by the Cooperative Human Linkage Center (CHLC; Murray et al., 1994). Last updated in 1997, the CHLC has identified, genotyped, and/or mapped over 3,300 microsatellite repeat markers. The CHLC Web site currently holds many linkage maps, including maps comprised solely of CHLC-derived markers and maps combining CHLC markers with those from other sources, including most markers in CEPHdb. CHLC markers can be recognized by unique identifiers that contain the nucleotide code for the tri- or tetranucleotide repeat units. For example, CHLC.GATA49A06 (D1S1608) contains a repeat unit of GATA, whereas CHLC.ATA28C07 (D1S1630) contains an ATA repeat. There are over 10,000 markers on the various linkage maps at CHLC, and most CHLC markers were genotyped in 15 CEPH pedigrees. The highest resolution CHLC maps have an average map distance of 1–2 cM between markers. Some of the maps contain markers in well-supported unique positions along with other markers placed into intervals.

Another set of genome-wide linkage maps was produced in 1996 by the group at Généthon (Dib et al., 1996). This group has identified and genotyped over 7,800 dinucleotide repeat markers and has produced maps containing only Généthon markers. These markers also have unique identifiers; each marker name has the symbols ''AFM'' at the beginning of the name. The Généthon map contains 5,264 genotyped in 8–20 CEPH pedigrees. These markers have been placed into 2,032 well-supported

map positions, with an average map resolution of 2.2 cM. Because of homogeneity of their marker and linkage data and the RH and YAC-based mapping efforts at Généthon that incorporate many of their polymorphic markers, the Généthon map has become the most widely utilized human linkage map.

The third and most recent set of human maps was produced at the Center for Medical Genetics at the Marshfield Medical Research Foundation (Broman et al., 1998). This group has identified over 300 dinucleotide repeats and has constructed high-density maps using over 8,000 markers. Like the CHLC maps, the Marshfield maps include their own markers as well as others, such as markers from CHLC and Généthon. These maps have an average resolution of 2.3 cM per map interval. Markers developed at the Marshfield Foundation have an MFD identifier at the beginning of their names. The authors caution on their Web site that because only eight of the CEPH families were used for the map construction, the orders of some of the markers are not well determined. The Marshfield Web site provides a useful utility for displaying custom maps that contain user-specified subsets of markers.

Two additional linkage maps have been developed exclusively for use in performing efficient large-scale and/or genome-wide genotyping. The ABI PRISM linkage mapping sets are composed of dinucleotide repeat markers derived from the Généthon linkage map. The ABI marker sets are available at three different map resolutions (20, 10, and 5 cM), containing 811, 400, and 218 markers, respectively. The Center for Inherited Disease Research (CIDR), a joint program sponsored by The Johns Hopkins University and the National Institutes of Health, provides a genotyping service that uses 392 highly polymorphic tri- and tetranucleotide repeat markers spaced at an average resolution of 9 cM. The CIDR map is derived from the Weber v.9 marker set, with improved reverse primers and some additional markers added to fill gaps.

Although each of these maps is extremely valuable, it can be very difficult to determine marker order and intermarker distance between markers that are not all represented on the same linkage map. The MAP-O-MAT Web site at Rutgers University is a marker-based linkage map server that provides several map-specific queries. The server uses genotypes for over 12,000 markers (obtained from the CEPH database and from the Marshfield Foundation) and the CRI-MAP computer program to estimate map distances, perform two-point analyses, and assess statistical support for order for user-specified maps (Matise and Gitlin, 1999). Thus, rather than attempting to integrate markers from multiple maps by rough interpolation, likelihood analyses can be easily performed on any subset of markers from the CEPH database.

High-resolution linkage maps have also been constructed for many other species. These maps are often the most well-developed resource for animal species' whose genome projects are in early stages. The mouse and rat both have multiple genome-wide linkage maps (see MGD and the Rat Genome Database); other species with well-developed linkage maps include zebrafish, cat, dog, cow, pig, horse, sheep, goat, and chicken (O'Brien et al., 1999).

Radiation Hybrid Map Resources

Radiation hybrid maps provide an intermediate level of resolution between linkage and physical maps. Therefore, they are helpful for sequence alignment and will aid in completion of the human genome sequencing project. Three human whole-genome panels have been prepared with different levels of X-irradiation and are available for

purchase from Research Genetics. Three high-resolution genome-wide maps have been constructed using these panels, each primarily utilizing EST markers. Mapping servers for each of the three human RH panels are available on-line to allow users to place their own markers on these maps. RH score data are deposited to, and publicly available from, The Radiation Hybrid Database (RHdb). Although this section covers RH mapping in humans, many RH mapping efforts are also underway in other species. More information regarding RH resources in all species are available at The Radiation Hybrid Mapping Information Web site.

In general, lower-resolution panels are most useful for more widely spaced markers over longer chromosomal regions, whereas higher-resolution panels are best for localizing very densely spaced markers over small regions. The lowest-resolution human RH panel is the Genebridge4 (GB4) panel (Gyapay et al., 1996). This panel contains 93 hybrids that were exposed to 3000 rads of irradiation. The maximum map resolution attainable by GB4 is 800–1,200 kb. An intermediate level panel was produced at the Stanford Human Genome Center (Stewart et al., 1997). The Stanford Generation 3 (G3) panel contains 83 hybrids exposed to 10,000 rads of irradiation. This panel can localize markers as close as 300–600 kb apart. The highest resolution panel (''The Next Generation,'' or TNG) was also developed at Stanford (Beasley et al., 1997). The TNG panel has 90 hybrids exposed to 50,000 rads of irradiation and can localize markers as close as 50–100 kb.

The Whitehead Institute/MIT Center for Genome Research constructed a map with approximately 6,000 markers using the GB4 panel (Hudson et al., 1995). Framework markers on this map were localized with odds ≥300:1, yielding a resolution of approximately 2.3 Mb between framework markers. Additional markers are localized to broader map intervals. A mapping server is provided for placing markers (scored in the GB4 panel) relative to the MIT maps.

The Stanford group has constructed a genome-wide map using the G3 RH panel (Stewart et al., 1997). This map contains 10,478 markers with an average resolution of 500 kb. Markers localized with odds = 1,000:1 are used to define ''high-confidence bins,'' and additional markers are placed into these bins with lower odds. A mapping server is provided for placing markers scored in the G3 panel onto the SHGC G3 maps.

A fourth RH map has been constructed using both the G3 and GB4 panels. This combined map, the Transcript Map of the Human Genome (GeneMap '99; Fig. 6.4), was produced by the RH Consortium, an international collaboration between several groups (Deloukas et al., 1998). This map contains over 30,000 ESTs localized against a common framework of approximately 1,100 polymorphic Généthon markers. The markers were localized to the framework using the GB4 RH panel, the G3 panel, or both. The map includes the majority of human genes with known function. Most markers on the map represent transcribed sequences with unknown function. The order of the framework markers is well supported, but most ESTs are mapped relative to the framework with odds <1,000:1. The majority of markers on the GeneMap have a lod score <2.0, and many are <1.0. Such markers are localized with relatively low support for local order, and their map positions should be confirmed by other means if critical. A mapping server for placing markers on GeneMap '99 is available at the Sanger Centre.

The Radiation Hybrid Database (RHdb) is the central repository for all RH data. It is maintained at the European Bioinformatics Institute (EBI) in Cambridge, UK (Rodriguez-Tome and Lijnzaad, 2000). RHdb is a sophisticated Web- and FTP-based

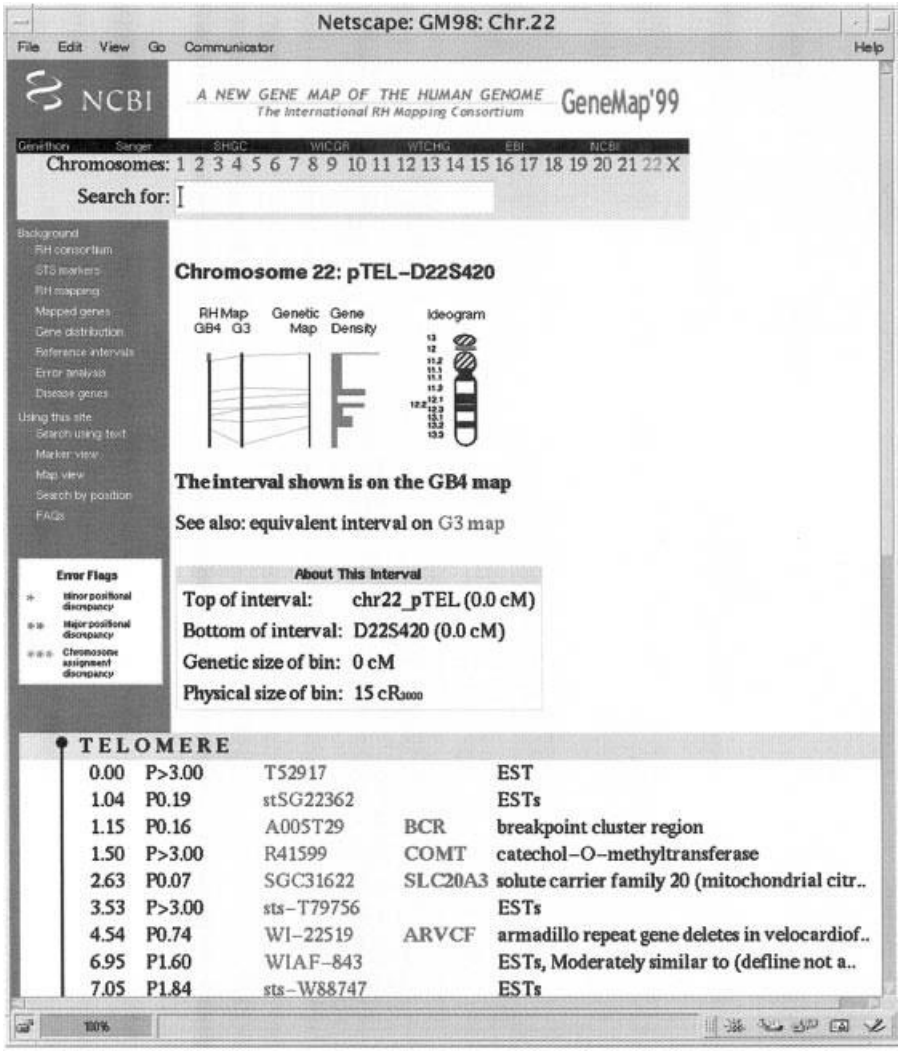

Figure 6.4. GeneMap '99. Example segment of The Human Gene Map, showing the first map interval on human chromosome 22q. Although the figure indicates that the map begins at a telomere, on this acrocentric chromosome, it actually begins near the centromere. The lower section of the figure contains 6 columns describing the elements mapped to this interval: column 1 gives cM linkage map positions for the polymorphic markers (none shown here); column 2 shows the computed cR position on either the GB4 or G3 portion of the GeneMap; column 3 contains either an `F` (for framework markers), or `P` followed by a number. This value represents the difference in statistical likelihood (lod score) for the given map position versus the next most likely position. A lod score of 3 is equivalent to odds of 1000:1 in favor of the reported marker position, 2 is equivalent to odds of 100:1, and a lod score of 1 represents odds of 10:1. Columns 4, 5, and 6 provide marker and gene names (if known).

searchable relational database that stores RH score data and RH maps. Data submission and retrieval are completely open to the public. Data are available in multiple formats or as flatfiles. Release 18.0 (September 2000) contained over 126,000 RH entries for 100,000 different STSs scored on 15 RH panels in 5 different species, as well as 91 RH maps.

STS Content Maps and Resources

Many physical mapping techniques have been used to order genomic segments for regional mammalian genome mapping projects. However, only RH and STS content/large-insert clone mapping methods have yielded the high throughput and automation necessary for whole-genome analysis to date, although advances in sequencing technology and capacity have recently made sequence-based mapping feasible. Two landmark achievements by the CEPH/Généthon and WICGR groups have mapped the entire human genome in YACs. The most comprehensive human physical mapping project is the collection of overlapping BAC and PAC clones being identified for the human DNA sequencing project, along with the now complete draft sequence of the human genome. This information is being generated by many different labs, and informatics tools to utilize the data are rapidly evolving.

The WICGR physical map is STS content based and contains more than 10,000 markers for which YAC clones have been identified, thus providing an average resolution of approximately 200 kb (Hudson et al., 1995). This map has been integrated with the Généthon GL and the WICGR RH maps. Together, the integration provides STS coverage of 150 kb, and approximately half the markers are expressed sequences also placed on GM99. The map was generated primarily by screening the CEPH MegaYAC library with primers specific for each marker and then by assembling the results by STS content analysis into sets of YAC contigs. Contigs are separately divided into ''single-linked'' and ''double-linked,'' depending on the minimum number of YACs (one or two) required to simultaneously link markers within a contig. Predictably, the double-linked contigs are shorter and much more reliable than the single-linked ones, largely because of the high chimeric rate of the MegaYAC library. Thus, some skill is required for proper interpretation of the YAC-based data.

The WICGR Human Physical Mapping Project Home Page provides links to downloadable (but large) GIFs of the maps, a number of ways to search the maps, and access to the raw data. Maps can be searched by entering or selecting a marker name, keyword, YAC, or YAC contig. Text-based displays of markers list marker-specific information, YACs containing the marker, and details of the associated contig. Contig displays summarize the markers contained within them, along with their coordinates on the GL and RH maps, which is a very useful feature for assessing contig integrity. Details of which YACs contain which markers and the nature and source of each STS/YAC hit are also shown. Clickable STS content maps are also provided from the homepage, and users have the option of viewing the content map alone or integrated with the GL and RH maps. Although there are numerous conflicts between the GL, RH, and STS content maps that often require clarification with other techniques, this resource is very informative once its complexities and limitations are understood, especially where BAC/PAC/sequence coverage is not complete and in linking together BAC/PAC contigs.

The CEPH/Généthon YAC project is a similar resource to the WICGR project, also centered around screening of the CEPH MegaYAC library with a large set of STSs (Chumakov et al., 1995). Much of the CEPH YAC screening results have been incorporated into the WICGR data (those YAC/STS hits marked as C). However, the CEPH data includes YAC fingerprinting, hybridization of YACs to inter-Alu PCR products, and FISH localizations as complementary methods to confirm contig holdings. As with WICGR, these data suffer from the high YAC chimerism rate; long-range contig builds should be interpreted with caution, and the data are best used only as a supplement to other genomic data. The CEPH YAC Web site includes a rudimentary text search engine for STSs and YACs that is integrated with the Généthon GL map, and the entire data set can be downloaded and viewed using the associated QUICKMAP application (Sun OS only; Chumakov et al., 1995).

Much of the human draft sequence was determined from BAC libraries that have been whole-scale DNA fingerprinted and end sequenced. To date, over 346,000 clones have been fingerprinted by Washington University Genome Sequencing Center (WUGSC), and the clone coverage is sufficient to assemble large contigs spanning almost the entire human euchromatin. The fingerprinting data can be searched by clone name at the WUGSC Web site and provides a list of clones overlapping the input clone, along with a probability score for the likelihood of each overlap. Alternatively, users can download the clone database and analyze the raw data using the Unix platform software tools IMAGE (for fingerprint data) and FPC (for contig assembly), which are available from the Sanger Centre.

In parallel with the BAC fingerprinting, a joint project by The Institute for Genome Research (TIGR) and the University of Washington High-Throughput Sequencing Center (UWHTSC) has determined the insert-end sequences (STCs) of the WUGSC-fingerprinted clones (743,000 sequences). These data can be searched by entering a DNA sequence at the UWHTSC site or by entering a clone name at the TIGR site. Together with the fingerprinting data, this is a convenient way to build and analyze maps in silico. The fingerprinting and STC data have been widely used for draft sequence ordering by the human sequencing centers, and the BAC/PAC contigs displayed by the NCBI Map Viewer are largely assembled from these data.

Many human single-chromosome or regional physical maps are also available. Because other complex genome mapping projects are less well developed, the WICGR mouse YAC mapping project is the only whole-genome nonhuman physical map available. This map is arranged almost identically to its human counterpart and consists of 10,000 STSs screened against a mouse YAC library (Nusbaum et al., 1999). However, whole-genome mouse fingerprinting and STC generation projects similar to their human counterparts are currently in production by TIGR/UWHTSC and the British Columbia Genome Sequence Centre (BCGSC), respectively.

DNA Sequence

As mentioned above, the existing human and forthcoming mouse draft genomic sequences are excellent sources for confirming mapping information, positioning and orienting localized markers, and bottom-up mapping of interesting genomic regions. NCBI tools like BLAST (Chapter 8) can be very powerful in finding marker/sequence links. NCBI's LocusLink lists all homologous sequences, including genomic sequences, for each known human gene (genomic sequences are type ''g'' on the LocusLink Web site; Maglott et al., 2000). e-PCR results showing all sequences

containing a specific marker are available at the GM99, dbSTS, GDB, and eGenome Web sites, where each sequence and the exact base pair position of the marker in the sequence are listed. Large sequence contigs can also be viewed schematically by NCBI's Entrez contig viewer and the Oakridge National Laboratory's Genome Channel web tool (Wheeler et al., 2000).

As the mammalian sequencing projects progress, a ''sequence first'' approach to mapping becomes more feasible. As an example, a researcher can go to the NCBI's human genome sequencing page and click on the idiogram of the chromosome of interest or on the chromosome number at the top of the page. Clicking on the idiogram shows an expanded idiogram graphically depicting all sequence contigs relative to the chromosome. Clicking on the chromosome number instead displays a list of all sequence contigs listed in order by cytogenetic and RH-extrapolated positions. These contigs can then be further viewed for clone, sequence, and marker content, and links to the relevant GenBank and dbSTS records are provided.

Integrated Maps and Genomic Cataloguing

GDB's Comprehensive Maps provide an estimated position of all genes, markers, and clones in GDB on a megabase scale. This estimate is generated by sequential pairwise comparison of shared marker positions between all publicly available genome-wide maps. This results in a consensus linear order of markers. At the GDB Web site, the Web page for each genomic element lists one or more maps on which the element has been placed, with the estimated Mb position of the marker on each map:

Element	Chromosome	Map	Coordinate	Units	EST MB	+/−
D1S228	1	GeneMap '99	782.0000	cR	32.2	0.0

This example shows that marker D1S228 has been placed 782 cR from the 1p telomere on GM99, and this calculates to 32.2 Mb from the telomere with the GDB mapping algorithm. Well-mapped markers such as the Généthon microsatellites generally have more reliable calculated positions than those that are mapped only once and/or by low-resolution techniques such as standard karyotype-based FISH. For chromosomes with complete DNA sequence available, the Mb estimates are very precise.

LDB and UDB are two additional sites that infer physical positions of a large, heterogeneous set of markers from existing maps using algorithms analogous to GDB's. Both Web sites have query pages where a map region can be selected by Mb coordinates, cytogenetic band, or specific marker names. The query results show a text-based list of all markers in the region ordered by their most likely positions, along with an estimated physical distance in Mb from the p telomere. LDB also displays the type of mapping technique(s) used to determine the comprehensive position, the position of the marker in each underlying single-dimension map, and appropriate references. An added feature of the UDB site is its provision of marker-specific links to other genomic databases. At present, there are no graphical depictions for either map.

Physical map positions derived from the computationally based algorithms used by GDB, LDB, and UDB are reliant on the accuracy and integrity of the underlying maps used to determine the positions. Therefore, these estimates serve better as initial localization guides and as supportive ordering information rather than as a primary ordering mechanism. For instance, a researcher defining a disease locus to a chromosome band or between two flanking markers can utilize these databases to quickly collect virtually all mapped elements in the defined region, and the inferred physical positions serve as an approximate order of the markers. This information would then be supplanted by more precise ordering information present in single-dimension maps and/or from the researcher's own experimental data.

The eGenome project uses a slightly different approach for creating integrated maps of the human genome (White et al., 1999). All data from RHdb are used to generate an RH framework map of each chromosome by a process that maximizes the number of markers ordered with high confidence (1,000:1 odds). This extended, high-resolution RH framework is then used as the central map scale from which the high-confidence intervals for additional RH and GL markers are positioned. As with GDB, the absolute base pair positions of all markers are calculated for chromosomes that have been fully sequenced. eGenome also integrates UniGene EST clusters, large-insert clones, and DNA sequences associated with mapped markers, and it also infers cytogenetic positions for all markers. The eGenome search page allows querying by marker name or GenBank accession ID or by defining a region with cytogenetic band or flanking marker coordinates. The marker displays include the RH and GL (if applicable) positions, large-insert clones containing the marker, cytogenetic position, and representative DNA sequences and UniGene clusters. Advantages of eGenome include the ability to view regions graphically using GDB's Mapview, exhaustive cataloguing of marker names, and an extensive collection of marker-specific hypertext links to related database sites. eGenome's maps are more conservative than GDB, LDB, and UDB as they show only the high-confidence locations of markers (often quite large intervals). Researchers determining a regional order *de novo* would be best advised to use a combination of these integrated resources for initial data collection and ordering.

Because of the large number of primary data sources available for human genome mapping, ensuring that the data collected for a specific region of interest are both current and all-inclusive is a significant task. Genomic catalogues help in this regard, both to provide a single initial source containing most of the publicly available genomic information for a region and to make the task of monitoring new information easier. Human genomic catalogues include the NCBI, GDB, and eGenome Web sites. NCBI's wide array of genomic data sets and analysis tools are extremely well integrated, allowing a researcher to easily transition between marker, sequence, gene, and functional information. GDB's concentration on mapped genomic elements makes it the most extensive source of positional information, and its inclusion of most genomic maps provides a useful mechanism to collect information about a defined region. eGenome also has powerful ''query-by-position'' tools to allow rapid collection of regional information. No existing database is capable of effectively organizing and disseminating all available human genomic information. However, the eGenome, GDB, and NCBI Web sites faithfully serve as genomic Web portals by providing hyperlinks to the majority of data available for a given genomic locus.

WICGR's mouse mapping project and the University of Wisconsin's Rat Genome Database (RGD; Steen et al., 1999) have aligned the GL and RH maps for the

respective species in a comparative manner. MGD's function as a central repository for mouse genomic information makes it useful as a mouse genomic catalogue, and, increasingly, RGD can be utilized as a rat catalogue. Unfortunately, other complex species' genome projects have not yet progressed to the point of offering true integrated maps or catalogues.

Comparative Resources

Comparative maps provide extremely valuable tools for studying the evolution and relatedness of genes between species and finding disease genes through position-based orthology. There are several multispecies comparative mapping resources available that include various combinations of most animal species for which linkage maps are available. In addition, there are also many sequence-based comparative analysis resources (Chapter 15). Each resource has different coverage and features. Presently, it is necessary to search multiple resources, as no single site contains all of the currently available homology information. Only the most notable resources will be described here.

A good starting point for homology information is NCBI's LocusLink database. The LocusLink reports include links to HomoloGene, a resource of curated and computed cross-species gene homologies (Zhang et al., 2000). Currently, HomoloGene contains human, mouse, rat, and zebrafish homology data. For example, a LocusLink search of all organisms for the gene PMP22 (peripheral myelin protein) returns three entries, one each for human, mouse, and rat. At the top of the human PMP22 page is a link to HOMOL (HomoloGene). HomoloGene lists six homologous elements, including the rat and mouse Pmp22 genes, as well as additional mouse UniGene cluster and a weakly similar zebrafish UniGene cluster. The availability of both curated and computed homology makes this a unique resource. However, the lack of integrated corresponding homology maps is a disadvantage.

The MGD does provide homology maps that simplify the task of studying conserved chromosome segments. Homologies are taken from the reported literature for mouse, human, rat, and 17 other species. Homology information can be obtained in one of three manners: searching for genes with homology information, building a comparative linkage map, or viewing an Oxford Grid. The simple search returns detailed information about homologous genes in other species, including map positions and codes for how the homology was identified, links to the relevant references, and links for viewing comparative maps of the surrounding regions in any two species. For example, a homology search for the Pmp22 gene returns a table listing homologous genes in cattle, dog, human, mouse, and rat. Figure 6.5 shows the mouse-human comparative map for the region surrounding Pmp22 in the mouse. A comparative map can also be obtained by using the linkage map-building tool to specify a region of the mouse as the index map and to select a second, comparison, species. The resulting display is similar to that shown in Figure 6.5. An Oxford Grid can also be used to view a genome-wide matrix in which the number of gene homologies between each pair of chromosomes between two species is shown. This view is currently available for seven species. Further details on the gene homologies can be obtained via the links for each chromosome pair shown on the grid. The map-viewing feature of MGD is quite useful; however, the positions of homologous nonmouse genes are only cytogenetic, so confirmation of relative marker order within

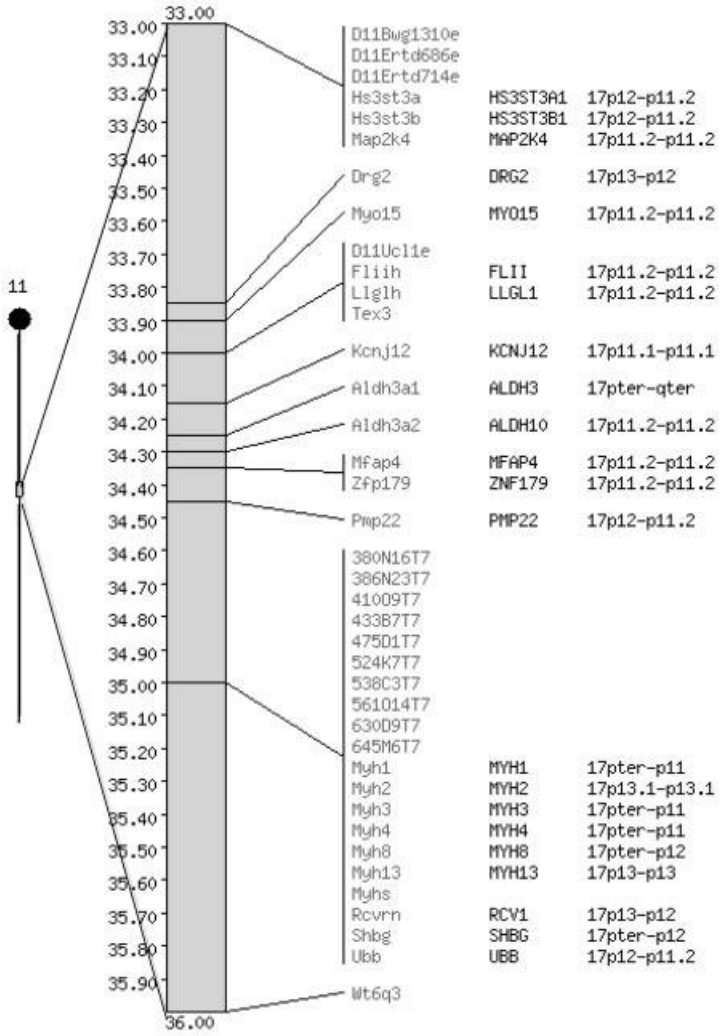

Figure 6.5. MGD mouse-human comparative map of the region surrounding the mouse Pmp22 gene. Pmp22 is on mouse chromosome 11 at the position 34.5 cM on the mouse linkage map. As shown by the human genes displayed on the right, a segment of human chromosome 17 is homologous to this mouse region.

small regions is not possible. It is also possible to view MGD homology information using GDB (Gatewood and Cottingham, 2000).

In silico mapping is proving to be a very valuable tool for comparative mapping. The Comparative Mapping by Annotation and Sequence Similarity (COMPASS) approach (Ma et al., 1998) has been used by researchers studying the cattle genome to construct cattle-human comparative maps with 638 identified human orthologs (Band et al., 2000). Automated comparison of cattle and human DNA sequences, along with the available human mapping information, facilitated localization predictions for tens of thousands of unmapped cattle ESTs. The COMPASS approach has been shown to have 95% accuracy. The Bovine Genome Database displays the gene-

based comparative maps, which also integrate mouse homologies. A similar approach is being used at the Bioinformatics Research Center at the Medical College of Wisconsin. Here, human, rat, and mouse radiation hybrid maps are coupled with theoretical gene assemblies based on EST and cDNA data (such as the UniGene set at NCBI) for all three species and provide the fundamental resources allowing for the creation of iteratively built comparative maps (Tonellato et al., 1999). Homologies with uniformly mapped ESTs form the anchor points for the comparative maps. This work has, so far, identified 8,036 rat-human, 13,720 rat-mouse, and 9,745 mouse-human UniGene homologies, most mapped on one or all of the organisms. The creation of these comparative maps is an iterative exercise that is repeated as the radiation hybrid maps, ESTs, and UniGene rebuilds are developed. In addition, the algorithm predicts the placement of unmapped assemblies relative to the anchor information, providing a powerful environment for ''virtual mapping'' before radiation hybrid or other wet-lab methods are used to confirm the predictions.

Another project utilizing electronic mapping has developed a high-resolution human/mouse comparative map for human chromosome 7. Recent efforts have greatly increased the number of identified gene homologies and have facilitated the construction of sequence-ready BAC-based physical maps of the corresponding mouse regions (Thomas et al., 2000).

An additional notable resource details homology relationships between human, mouse, and rat. Derived from a high-resolution RH maps, homologies for over 500 genes have been identified and are available in tabular format at a user-friendly Web site (Watanabe et al., 1999).

Single-Chromosome and Regional Map Resources

Although whole-genome mapping resources are convenient for initial collection and characterization of a region of interest, data generated for only a single chromosome or a subchromosomal region are often important for fine mapping. In many cases, these regional maps contain more detailed, better integrated, and higher resolution data than the whole-genome maps can provide. There are numerous such data sets, databases, and maps available for each human chromosome, although little regional information is yet available on-line for other complex genomes. Most published human chromosome maps are listed and can be viewed at GDB's Web site (see above).

Another excellent resource is the set of human chromosome-specific Web sites that have been created by various groups. Recently, the Human Genome Organization (HUGO) has developed individual human chromosome Web pages, each of which is maintained by the corresponding HUGO chromosome committees. Each page has links to chromosome-specific information from a variety of mapping sources, most of them being chromosome-specific subsets of data derived from whole-genome resources (such as the chromosome 7 GL map from Généthon). At the top of most HUGO chromosome pages are links to other chromosome pages designed by groups mapping the specific chromosome. These sites vary widely in their utility and content; some of the most useful are briefly mentioned below. The sites offer a range of resources, including chromosome- and/or region-specific GL, RH, cytogenetic, and physical maps; DNA sequence data and sequencing progress, single chromosome databases and catalogues of markers, clones, and genomic elements; and links to

related data and resources at other sites, single chromosome workshop reports, and chromosome E-mail lists and discussion forums.

The major genome centers often include detailed mapping and sequence annotation for particular chromosomes at their sites. The Sanger Centre and the WUGSC have two of the most advanced collections of chromosome-specific genomic data, informatics tools, and resources. Sanger has collected and generated most available mapping data and reagents for human chromosomes 1, 6, 9, 10, 13, 20, 22, and X. These data are stored and displayed using ACeDB, which can be utilized through a Web interface (WEBACE) at the Sanger Web site or, alternatively, downloaded onto a local machine (Unix OS). ACeDB is an object-oriented database that provides a convenient, relational organizational scheme for storing and managing genomic data, as well as for viewing the information in both text-based and graphical formats. ACeDB is the database of choice for most researchers tackling large genomic mapping projects. WUGSC has recently implemented single-chromosome ACeDB sequence and mapping databases for most human chromosomes, each of which has a Web interface.

The Human Chromosome 1 Web site is an example of a community-based approach to genomic research. This site includes a repository for chromosome data generated by several labs, an extensive list of hyperlinks to chromosome 1 data, an E-mail list and discussion forum, a listing of chromosome 1 researchers and their interests, and several workshop reports. The University of Texas at San Antonio's chromosome 3 site contains a database of large-insert clones and markers along with GL, RH, cytogenetic, and comparative maps. The University of California-Irvine has an on-line chromosome 5 ACeDB database, whereas the Joint Genome Institute (JGI) maintains chromosome 5 large-insert clone maps and some external Web links at their site. The University of Toronto chromosome 7 Web site includes a searchable comprehensive chromosome 7 database containing markers, clones, and cytogenetic information; this site also has a long list of chromosome links. Also, the National Human Genome Research Institute's chromosome 7 Web site contains a YAC/STS map, a list of ESTs, and integration with chromosome 7 sequence files. The University College London maintains a good comprehensive resource of chromosome 9 genomic links, an E-mail group, workshop reports, and a searchable chromosome 9 database. Genome Therapeutics Corporation has developed an inclusive Web site for chromosome 10. This site has both GL/physical and integrated sequence-based maps, links to related data, and workshop reports.

Imperial College maintains a searchable chromosome 11 database at their chromosome 11 Web site, whereas the chromosome 16 Web site at JGI contains restriction-mapped BAC and cosmid contigs and determined sequence, along with a list of chromosome 16 hyperlinks. A similar JGI resource for chromosome 19 includes a completely integrated physical map with sequence links and a list of external resources. The University of Colorado, the RIKEN Genomic Sciences Center, and the Max Planck Institute for Molecular Genetics (MPIMG) have an interconnected set of resources that together integrate genomic clones, markers, and sequence for the completely sequenced chromosome 21. The Sanger Centre and the LDB have comprehensive resources for the viewing and analysis of chromosome 22. It is expected that additional resources for all completely sequenced chromosomes will be available soon. The resources for the X chromosome are most impressive. The MPIMG has established a complete genomic catalogue of this chromosome that features integration of genomic mapping and sequence data derived from many sources and ex-

perimental techniques. These data can be viewed graphically with the powerful online Java application derBrowser. Finally, the sequenced and well-characterized mitochondrial genome is well displayed at Emory University, where a highly advanced catalogue encompassing both genomic and functional information has been established.

PRACTICAL USES OF MAPPING RESOURCES

Potential applications of genomic data are numerous and, to a certain extent, depend on the creativity and imagination of the researcher. However, most researchers utilize genomic information in one of three ways: to find out what genomic elements—usually transcribed elements—are contained within a genomic region, to determine the order of defined elements within a region, or to determine the chromosomal position of a particular element. Each of these goals can be accomplished by various means, and the probability of efficient success is often enhanced by familiarity with many of the resources discussed in this chapter. It is prudent to follow a logical course when using genomic data. During the initial data acquisition step, in which genomic data are either generated experimentally or retrieved from publicly available data sources, simultaneous evaluation of multiple data sets will ensure both higher resolution and greater confidence while increasing the likelihood that the genomic elements of interest are represented. Second, the interrelationships and limitations of the data sets must be sufficiently understood, as it is easy to overinterpret or underrepresent the data. Finally, it is important to verify critical assignments independently, especially when using mapping data that are not ordered with high confidence. Below, we give some brief suggestions on how to approach specific map-related tasks, but many modifications or alternative approaches are also viable. The section is organized in a manner similar to a positional cloning project, starting with definition of the region's boundaries, determining the content and order of elements in the region, and defining a precise map position of the targeted element.

Defining a Genomic Region

A genomic region of interest is best defined by two flanking markers that are commonly used for mapping purposes, such as polymorphic Généthon markers in humans or MIT microsatellites in mice. Starting with a cytogenetically defined region is more difficult due to the subjective nature of defining chromosomal band boundaries. Conversion of cytogenetic boundaries to representative markers can be approximated by viewing the inferred cytogenetic positions of markers in comprehensive maps such as GDB's universal map, UDB, LDB, or eGenome. Because these cytogenetic positions are inferred and approximate, a conservative approach is recommended when using cytogenetic positions for region definition. The choice of flanking markers will impact how precisely a region's size and exact boundary locations can be defined. Commonly used markers are often present on multiple, independently derived maps, so their ''position'' on the chromosome provides greater confidence for anchoring a regional endpoint. In contrast, the exact location of less commonly used markers is often locally ambiguous. These markers can sometimes be physically tethered to other markers if a large sequence tract that contains multiple markers can be found.

This can be performed by BLASTing marker sequences against GenBank or by scanning e-PCR results in UniGene or eGenome for a particular marker.

Determining and Ordering the Contents of a Defined Region

Once a region has been defined, there are a number of resources available for determining what lies within the region. A good way to start is to identify a map that contains both flanking markers, either from a chromosome-wide or genome-wide map from the sources listed above, from a genomic catalogue, or from a local map that has been generated by a laboratory interested in this particular region. For humans, GDB is the most inclusive map repository, although many regional maps have not been deposited in GDB but can be found with a literature search of the corresponding cytogenetic band or a gene known to map to the region. Many localized maps are physically based and are more accurate than their computationally derived, whole-chromosome counterparts. For other species, the number of maps to choose from is usually limited, so it is useful to first define flanking markers known to be contained in the available maps.

The map or maps containing the flanking markers can then be used to create a consensus integrated map of the region. This is often an inexact and tedious process. To begin, it is useful to identify from the available maps an index map that contains many markers, high map resolution, and good reliability. Integration of markers from additional maps relative to the index map proceeds by comparing the positions of markers placed on each map. For example, if an index map contains markers in the order A-B-C-D and a second map has markers in the order B-E-D, then marker E can be localized to the interval between markers B and D on the index map. Importantly, however, the relative position of marker E with respect to marker C usually cannot be accurately determined by this method. Repeated iterations of this process should allow localization of all markers from multiple maps relative to the index map. This process is of course significantly reinforced by experimental verification, such as with STS content mapping of large-insert clones identified for the region-specific markers or, ideally, by sequence-determined order.

Each marker represents some type of genomic element: a gene, an EST, a polymorphism, a large-insert clone end, or a random genomic stretch. In humans, identifying what a marker represents is relatively straightforward. Simply search for the marker name in GDB or eGenome, and, in most cases, the resulting Web display will provide a summary of what the marker represents, usually along with hyperlinks to relevant functional information. For mice, MGD provides a similar function to GDB. For other organisms, the best source is usually either dbSTS or, if present, Web sites or publications associated with the underlying maps. GenBank and dbSTS are alternatives for finding markers, but, because these repositories are passive (requiring researchers to submit their markers rather than actively collecting markers), many marker sets are not represented. If a marker is known to be expressed, UniGene, LocusLink, and dbEST are excellent sources of additional information. Many genes and some polymorphisms have been independently discovered and developed as markers multiple times, and creating a nonredundant set from a collection of markers is often challenging. GDB, eGenome, MGD, and (for genes) UniGene are good sources to use for finding whether two markers are considered equivalent but even more reliable is a DNA sequence or sequence contig containing both

marker's primers. BLAST and the related BLAST2 are efficient for quickly determining sequence relatedness (Chapter 8).

Obviously, the most reliable tool for marker ordering is a DNA sequence or sequence contig. For expressed human markers, searching with the marker name in UniGene or Entrez Genomes returns a page stating where (or if) the marker has been mapped in GeneMap '99 and other maps, a list of mRNA, genomic, and EST sequences, and with Entrez Genomes, a Mapviewer-based graphical depiction of the maps, sequence-ready contigs, and available sequence of the region. Similarly, GDB and eGenome show which DNA sequences contain each displayed marker. For other markers, the sequence from which the marker is derived, or alternatively one of the primer sequences, may be used to perform a BLAST search that can identify completely or nearly homologous sequences. The nonredundant, EST, GSS, and HTGS divisions of GenBank are all potentially relevant sources of matching sequence, depending on the aim of the project. Only long sequences are likely to have worthwhile marker-ordering capabilities. Finished genomic sequence tracts have at least some degree of annotation, and scanning the GenBank record for the large sequence will often yield an annotated list of what markers lie within the sequence and where they are. Keep in mind that such annotations vary considerably in their thoroughness and most are fixed in time; that is, they only recognize markers that were known at the time of the annotation. BLAST, BLAST2, or other sequence-alignment programs are helpful in identification or confirmation of what might lie in a large sequence. Also, the NCBI e-PCR Web interface can be used to identify all markers in dbSTS contained within a given sequence, and this program can be installed locally to query customized marker sets with DNA sequences (Schuler, 1997).

For genomes for which DNA sequencing is complete or is substantially underway, it may be possible to construct local clone or sequence contigs. Among higher organisms, this is currently possible only for the human and mouse genomes. Although individual clone sequences can be found in GenBank, larger sequence contigs —sequence tracts comprising more than one BAC or PAC—are more accessible using the Entrez Genomes Web site (see above). Here, by entering a marker or DNA accession number into the contigs search box, researchers can identify sequence contigs containing that marker or element. This site also provides a graphical view of all other markers contained in that sequence, the base pair position of the markers in the sequence, and, with the Mapviewer utility, graphical representations of clone contigs. This process can also be performed using BLAST or e-PCR, although it is somewhat more laborious.

Once a sequence has been identified for markers in a given region, YAC clone, DNA fingerprinting, and STC data can be used to bridge gaps. For humans and mice, the WICGR YAC data provide a mechanism for identifying YAC clones linking adjacent markers. However, caution should be exercised to rely mainly on double-linked contigs and/or to experimentally confirm YAC/marker links. Also for human genome regions, the UWHTSC and TIGR Web sites for identifying STCs from DNA sequence or BAC clones are very useful. For example, researchers with a sequence tract can go to the UWHTSC TSC search page, enter their sequence, and find STCs contained in the sequence. Any listed STC represents the end of a BAC clone whose insert contains a portion of the input sequence (Venter et al., 1996). The TIGR search tool is complementary to the UWHTSC search, as the TIGR site requires input of a large-insert clone name, which yields STC sequences. STCs represent large-insert clones that potentially extend a contig or link two adjacent, nonoverlapping contigs.

Similarly, ~375,000 human BAC clones have been fingerprinted for rapid identification of overlapping clones (Marra et al., 1997). The fingerprinting data are available for searching at the Washington University Human Genome Sequencing Center (WUGSC). Combined use of Entrez, BLAST, the STC resources, and the BAC fingerprinting data can often provide quick and reliable contig assembly by in silico sequence and clone walking.

Defining a Map Position From a Clone or DNA Sequence

Expressing the chromosomal position of a gene or genomic element in physical, RH, GL, or cytogenetic terms is not always straightforward. The first approach is to determine whether the element of interest has already been localized. The great majority of human transcripts are precisely mapped, and many genes have been well localized in other organisms as well. For species with advanced DNA sequencing projects, it is helpful to identify a large DNA sequence tract containing the genomic element of interest and then determine what markers it contains by looking at the sequence annotation record in GenBank or by e-PCR. Identified human and mouse genes are catalogued in GDB and LocusLink or MGD, respectively, and searching UniGene with a marker name, mRNA or EST sequence accession number, or gene name will often provide a localization if one is known. Here again, nomenclature difficulties impede such searches, making it necessary to search each database with one or more alternate names in some cases. Another alternative is to determine if the genomic element is contained in a genomic sequence by a simple BLAST search. Most large genomic sequences have been cytogenetically localized, and this information is contained in the sequence annotation record (usually in the title).

If gene-specific or closely linked markers have been used previously for mapping, a position can usually be described in terms specific to the mapping method that was employed. For example, if an unknown gene is found to map very close to a Généthon marker, then the gene position can be reported relative to the Généthon GL centiMorgan coordinates. Most human markers and many maps have been placed in GDB, so this is a good first step in determining whether a marker has been mapped. Simply search for the relevant marker and see where it has been placed on one or several maps listed under ''cytogenetic localizations'' and ''other localizations.'' Inferred cytogenetic positions of human genes and markers are usually listed in GDB, UniGene, and eGenome if the elements have been previously mapped. If not, band or band range assignments can usually be approximated by finding the cytogenetic positions of flanking or closely linked markers and genes. Many sequenced large-insert clones have been assigned by FISH to a cytogenetic position; this information can usually be found in the sequence annotation or at the clone originator's Web site. The process of determining whether a transcript or genomic element from another organism has been mapped varies somewhat due to the lack of extensive genomic catalogs, making it usually necessary to cross-reference a marker with the GL and/or RH maps available for the species.

If no previous localization exists for a genomic element, some experimental work must be undertaken. For human and mouse markers, an efficient and precise way to map a sequence-based element is to develop and map an STS derived from the element by RH analysis. A set of primers should be designed that uniquely amplify a product in the species of interest, but not in the RH panel background genome. An STS is usually unique if at least one primer is intronic. Primers designed from

an mRNA sequence will not amplify the correct-sized product in genomic DNA if they span an intron, but a good approximation is to use primers from the 3′ untranslated region, as these stretches only rarely contain introns and usually comprise sequences divergent enough from orthologous or paralogous sequences. However, beware of pseudogenes and repetitive sequences, and genomic sequence stretches are superior for primer design. Suitable primers can then be used to type an appropriate RH panel; currently, human (G3, GB4, or TNG), mouse, rat, baboon, zebrafish, dog, horse, cow and pig panels are available commercially. After the relevant panel is typed, the resulting data can be submitted to a panel-specific on-line RH server (see above) for the human, mouse, rat, and zebrafish panels. For other species, isolation and FISH of a large-insert clone or GL mapping with an identified or known flanking polymorphism may be necessary.

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 6

DATA REPOSITORIES	
The Genome DataBase (GDB)	*http://www.gdb.org/*
National Center for Biotechnology Information (NCBI)	
Home Page	*http://www.ncbi.nlm.nih.gov*
Entrez Genomes Division	*http://www.ncbi.nlm.nih.gov/Entrez/Genome/main_genomes.html*
LocusLink	*http://www.ncbi.nlm.nih.gov/LocusLink/*
GeneMap'99	*http://www.ncbi.nlm.nih.gov/genemap99/*
OMIM	*http://www.ncbi.nlm.nih.gov/Omim/*
HomoloGene	*http://www.ncbi.nlm.nih.gov/HomoloGene/*
BLAST	*http://www.ncbi.nlm.nih.gov/BLAST/*
ePCR	*http://www.ncbi.nlm.nih.gov/STS/*
Entrez sequence viewer	*http://www.ncbi.nlm.nih.gov/genome/seq/*
GenBank	*http://www.ncbi.nlm.nih.gov/Genbank*
Genomic Biology page	*http://www.ncbi.nlm.nih.gov/Genomes*
dbSTS	*http://www.ncbi.nlm.nih.gov/dbSTS*
Mouse Genome Informatics (MGD/MGI)	*http://www.informatics.jax.org/*
RESOURCES AND PROJECTS	
Cytogenetic	
BAC	*http://bacpac.med.buffalo.edu/human/overview.html*
LBNL/UCSF RMC	*http://ioerror.ucsf.edu:8080/~dfdavy/rmc/OUTSIDE.html*
U. of Bari	*http://bioserver.uniba.it/fish/rocchi*
Cytogenetic/YAC data	*http://www.mpimg-berlin-dahlem.mpg.de/~cytogen/probedat.htm*
NCI	*http://www.ncbi.nlm.nih.gov/CGAP/*
CCAP	*http://www.ncbi.nlm.nih.gov/CCAP/mitelsum.cgi*
Infobiogen	*http://www.infobiogen.fr/services/chromcancer/*
SERGG	*http://www.ir.miami.edu/genetics/sergg/chromosome.html*

Coriell	*http://locus.umdnj.edu/nigms/ideograms/1.html*
ARKdb	*http://www.ri.bbsrc.ac.uk/bioinformatics/ark_overview.html*
Animal Genome Database	*http://ws4.niai.affrc.go.jp/jgbase.html*
Cedars-Sinai	*http://www.csmc.edu/genetics/korenberg/korenberg.html#A*
Genetic Linkage	
CEPH Genotype Database	*http://www.cephb.fr/cephdb/*
CHLC	*http://lpg.nci.nih.gov/CHLC/*
Généthon	*http://www.genethon.fr/genethon_en.html*
Marshfield	*http://www.marshmed.org/genetics/*
MAP-O-MAT	*http://compgen.rutgers.edu/mapomat*
Rat Genome Database	*http://www.lgr.mcw.edu/projects/rgd.html*
Radiation Hybrid	
RHdb	*http://www.ebi.ac.uk/RHdb/*
RH Information Page	*http://compgen.rutgers.edu/rhmap/*
Research Genetics	*http://www.resgen.com*
WICGR RH Maps	*http://www-genome.wi.mit.edu/cgi-bin/contig/phys_map*
WICGR GB4 RH Map Server	*http://www-genome.wi.mit.edu/cgi-bin/contig/rhmapper.pl*
SHGC RH Maps	*http://shgc-www.stanford.edu/Mapping/rh/*
SHGC G3 Map Server	*http://shgc-www.stanford.edu/RH/*
Sanger Centre GB4/GM Map server	*http://www.sanger.ac.uk/Software/RHserver/*
STS content	
WICGR human physical map	*http://carbon.wi.mit.edu:8000/cgi-bin/contig/phys_map*
CEPH/Généthon YAC map	*http://www.cephb.fr/bio/ceph-genethon-map.html*
WUGSC home	*http://genome.wustl.edu/gsc/index.shtml*
TIGR STCs	*http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_intro.html*
UWHTSC STCs	*http://www.htsc.washington.edu/human/info/index.cfm*
WUGSC BAC fingerprints	*http://genome.wustl.edu/gsc/human/human_database.shtml*
UBGSC mouse BAC fingerprints	*http://www.bcgsc.bc.ca/projects/mouse_mapping/*
Trask	*http://fishfarm.biotech.washington.edu/BACResource/Random/index.html*
WICGR mouse physical/genetic map	*http://carbon.wi.mit.edu:8000/cgi-bin/mouse/index*
DNA Sequence	
see NCBI links	
ORNL Genome Channel	*http://compbio.ornl.gov/tools/channel/*
Integrated and Catalogs	
UDB	*http://bioinformatics.weizmann.ac.il/udb/*
LDB	*http://cedar.genetics.soton.ac.uk/public_html/ldb.html*

LDB Sequence-based maps	*http://cedar.genetics.soton.ac.uk/public_html/LDB2000.html*
eGenome	*http://genome.chop.edu*
Comparative	
Mouse Homology	*http://www.informatics.jax.org/menus/homology_menu.shtml*
Otsuka/Oxford rat-mouse-human	*http://ratmap.ims.u-tokyo.ac.jp/*
Human Chromosome 7–mouse map	*http://genome.nhgri.nih.gov/chr7/comparative/*
Bovine Genome Database	*http://bos.cvm.tamu.edu/bovgbase.html*
MCW Rat-Mouse-Human	*http://rgd.mcw.edu*
Single-chromosome/regional	
1 Rutgers	*http://linkage.rockefeller.edu/chr1/*
3 UTSA	*http://apollo.uthscsa.edu/*
5 UCI	*http://chrom5.hsis.uci.edu*
5 JGI	*http://jgi.doe.gov/Data/JGI_mapping.html*
7 HSC	*http://www.genet.sickkids.on.ca/chromosome7/*
7 NHGRI	*http://www.nhgri.nih.gov/DIR/GTB/CHR7*
9 UCL	*http://www.gene.ucl.ac.uk/chr9/*
10 GTC	*http://www.cric.com/sequence_center/chromosome10/*
11 Imperial College	*http://chr11.bc.ic.ac.uk/*
16 JGI	*http://jgi.doe.gov/Data/JGI_mapping.html*
19 JGI	*http://jgi.doe.gov/Data/JGI_mapping.html*
21 Colorado	*http://www-eri.uchsc.edu/chromosome21/frames.html*
21 RIKEN	*http://hgp.gsc.riken.go.jp/chr21/index.html*
21 MPIMG	*http://chr21.rz-berlin.mpg.de/*
X	*http://www.mpimg-berlin-dahlem.mpg.de/~xteam/*
Mito Emory	*http://infinity.gen.emory.edu/mitomap.html*
HUGO Chromosome resources	*http://www.gdb.org/hugo/*
Sanger Centre	*http://www.sanger.ac.uk/HGP/*
ACEDB	*http://www.acedb.org/*

PROBLEM SET

You have performed a large-scale genome-wide search for the gene for the inherited disorder Bioinformatosis. Initial analyses have identified one region with significant results, flanked by the markers D21S260–D21S262. There are many genes mapping within this region, one of which is particularly interesting, superoxide dismutase 1 (SOD1).

1. What is the cytogenetic location of this gene (and hence, at least part of the region of interest)?

2. How large is this region in cM?

3. What polymorphic markers can be identified in this region (that you might use to try to narrow the region)? Choose six of these polymorphic markers. Based on the chosen markers, can a map based on these markers be identified or constructed?

4. What STS markers have been developed for SOD1? What are their map positions on the Human Transcript Map (GeneMap '99)? Are these positions statistically well-supported? Have any SNP markers been identified within SOD1?

5. What other genes are in this region?

6. Has the region including the SOD1 gene been sequenced? What contigs and/or clones contain SOD1?

7. Have orthologous regions been identified in any other species?

8. Have mutations in SOD1 been associated with any diseases other than Bioinformatosis?

REFERENCES

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656.

Aston, C., Mishra, B., and Schwartz, D. C. (1999). Optical mapping and its potential for large-scale sequencing projects. *Trends Biotechnol.* 17, 297–302.

Band, M. R., Larson, J. H., Rebeiz, M., Green, C. A., Heyen, D. W., Donovan, J., Windish, R., Steining, C., Mahyuddin, P., Womack, J. E., and Lewin, H. A. (2000). An ordered comparative map of the cattle and human genomes. *Genome Res.* 10, 1359–1368.

Beasley, E., Stewart, E., McKusick, K., Aggarwal, A., Brady-Hebert, S., Fang, N., Lewis, S., Lopez, F., Norton, J., Pabla, H., Perkins, S., Piercy, M., Qin, F., Reif, T., Sun, W., Vo, N., Myers, R., and Cox, D. (1997). The TNG4 radiation hybrids improve the resolution of the G3 panel. *Am. J. Hum. Genet.* 61(Suppl.), A231.

Blake, J. A., Eppig, J. T., Richardson, J. E., and Davisson, M. T. (2000). The mouse genome database (MGD): expanding genetic and genomic resources for the laboratory mouse. *Nucleic Acids Res.* 28, 108–111.

Boehnke, M., Lange, K., and Cox, D. R. (1991). Statistical methods for multipoint radiation hybrid mapping. *Am. J. Hum. Genet.* 49, 1174–1188.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., and Weber, J. L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* 63, 861–869.

Burke, D. T., Carle, G. F., and Olson, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 236, 806–812.

Chakravarti, A., and Lynn, A. (1999). Meiotic mapping in humans. In *Genome Analysis: A Laboratory Manual, Vol. 4, Mapping Genomes*, B. Birren, E. Green, P. Hieter, S. Klapholz, R. Myers, H. Riethman, and J. Roskams, eds. (Cold Spring Harbor: Cold Spring Harbor Laboratory Press).

Chumakov, I. M., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billault, A., Guillou, S., Soularue, P., Guasconi, G., Poullier, E., Gros, I., Belova, M., Sambucy, J.-L., Susini, L., Gervy, P., Glibert, F., Beaufils, S., Bul, H., Massart, C., De Tand, M.-F., Dukasz, F., Lecoulant, S., Ougen, P., Perrot, V., Saumier, M., Soravito, C., Bahouayila, R., Cohen-

Akenine, A., Barillot, E., Bertrand, S., Codani, J.-J., Caterina, D., Georges, I., Lacroix, B., Lucotte, G., Sahbatou, M., Schmit, C., Sangouard, M., Tubacher, E., Dib, C., Faure, S., Fizames, C., Cyapay, G., Millasseau, P., NGuyen, S., Muselet, D., Vignal, A., Morissette, J., Menninger, J., Lieman, J., Desai, T., Banks, A., Bray-Ward, P., Ward, D., Hudson, T., Gerety, S., Foote, S., Stein, L., Page, D. C., Lander, E. S., Weissenbach, J., Le Paslier, D., and Cohen, D. (1995). A YAC contig map of the human genome. *Nature* 377, 175–297.

Collins, A., Frezal, J., Teague, J., and Morton, N. E. (1996). A metric map of humans: 23,500 loci in 850 bands. *Proc. Natl. Acad. Sci. USA* 93, 14771–14775.

Collins, A., Teague, J., Keats, B., and Morton, N. (1996). Linkage map integration. *Genomics* 35, 157–162.

Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.-M., and White, R. (1990). Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative genetic mapping of the human genome. *Genomics* 6, 575–577.

Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tomé, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S., Benolila, S., Bihoreau, M.-T., Birren, B. B., Browne, J., Butler, A., Castle, A. B., Chiannikulchai, N., Clee, C., Day, P. J. R., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Fox, S., Gelling, S., Green, L., Harison, P., Hocking, R., Holloway, E., Hunt, S., Keil, S., Lijnzaad, P., Louis-Dit-Sully, C., Ma, J., Mendis, A., Miller, J., Morissette, J., Muselet, D., Nusbaum, H. C., Peck, A., Rozen, S., Simon, D., Slonim, D. K., Staples, R., Stein, L. D., Stewart, E. A., Suchard, M. A., Thangarajah, T., Vega-Czarny, N., Webber, C., Wu, X., Auffray, C., Nomura, N., Sikela, J. M., Polymeropoulos, M. H., James, M. R., Lander, E. S., Hudson, T. J., Myers, R. M., Cox, D. R., Weissenbach, J., Boguski, M. S., and Bentley, D. R. (1998). A physical map of 30,000 human genes. *Science* 282, 744–746.

Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J., and Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380, 152–154.

Dietrich, W., Weber, J., Nickerson, D., and Kwok, P.-Y. (1999). Identification and Analysis of DNA Polymorphisms. In *Genome Analysis: A Laboratory Manual, Vol. 4, Mapping Genomes*, B. Birren, E. Green, P. Hieter, S. Klapholz, R. Myers, H. Riethman and J. Roskams, eds. (Cold Spring Harbor: Cold Spring Harbor Laboratory Press).

Gatewood, B., and Cottingham, R. (2000). Mouse-human comparative map resources on the web. *Briefings in Bioinformatics* 1, 60–75.

Green, E. D., and Olson, M. V. (1990). Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: a model for human genome mapping. *Science* 250, 94–98.

Gyapay, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillet, D., Muselet, D., Prud'homme, J., Dib, C., Auffray, C., Morissette, J., Weissenbach, J., and Goodfellow, P. N. (1996). A radiation hybrid map of the human genome. *Hum. Mol. Genet.* 5, 339–358.

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Marra, M., and et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6, 807–828.

Houlgatte, R., Mariage-Samson, R., Duprat, S., Tessier, A., Bentolila, S., Lamy, B., and Auffray, C. (1995). The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res.* 5, 272–304.

Hudson, T. J., Stein, L. D., Gerety, S. S., Ma, J., Castle, A. B., Silva, J., Slonim, D. K., Baptista, R., Kruglyak, L., Xu, S.-H., Hu, X., Colbert, A. M. E., Rosenberg, C., Reeve-Daly, M. P., Rozen, S., Hui, L., Wu, X., Vestergaard, C., Wilson, K. M., Bae, J. S., Maitra, S., Ganiatsas, S., Evans, C. A., DeAngelis, M. M., Kngalls, K. A., Nahf, R. W., Horton

Jr., L. T., Anderson, M. O., Collymore, A. J., Ye, W., Kouyoumijan, V., Zemsteva, I. S., Tam, J., Devine, R., Courtney, D. F., Renaud, M. T., Nguyen, H., O'Connor, T. J., Fizames, C., Fauré, S., Gyapay, G., Dib, C., Morissette, J., Orlin, J. B., Birren, B. W., Goodman, N., Weissenbach, J., Hawkins, T. L., Foote, S., Page, D. C., and Lander, E. S. (1995). An STS-based map of the human genome. *Science* 270, 1945–1954.

Iaonnou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A., and de Jong, P. J. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat. Genet.* 6, 84–89.

Jensen, S. J., Sulman, E. P., Maris, J. M., Matise, T. C., Vojta, P. J., Barrett, J. C., Brodeur, G. M., and White, P. S. (1997). An integrated transcript map of human chromosome 1p35–36. *Genomics* 42, 126–136.

Jing, J., Lai, Z., Aston, C., Lin, J., Carucci, D. J., Gardner, M. J., Mishra, B., Anantharaman, T. S., Tettelin, H., Cummings, L. M., Hoffman, S. L., Venter, J. C., and Schwartz, D. C. (1999). Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Res.* 9, 175–181.

Korenberg, J. R., Chen, X.-N., Devon, K. L., Noya, D., Oster-Granite, M. L., and Birren, B. W. (1999). Mouse Molecular Cytogenetic Resource: 157 BACs link the chromosomal and genetic maps. *Genome Res.* 9, 514–523.

Lander, E. S., and Green, P. (1987). Construction of multi-locus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* 84, 2363–2367.

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E., and Newburg, L. (1987). MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1, 174–181.

Lathrop, G. M., Lalouel, J. M., Julier, C., and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. USA* 81, 3443–3446.

Lawrence, J. B., Singer, R. H., and NcNeil, J. A. (1990). Interphase and metaphase resolution of different distances within the human dystrophin gene. *Science* 249, 928–932.

Letovsky, S. I., Cottingham, R. W., Porter, C. J., and Li, P. W. D. (1998). GDB: the Human Genome Database. *Nucleic Acids Res.* 26, 94–99.

Lewis, T. B., Nelson, L., Ward, K., and Leach, R. J. (1995). A radiation hybrid map of 40 loci for the distal long arm of human chromosome 8. *Genome Res.* 5, 334–341.

Ma, R. Z., van Eijk, M. J., Beever, J. E., Guerin, G., Mummery, C. L., and Lewin, H. A. (1998). Comparative analysis of 82 expressed sequence tags from a cattle ovary cDNA library. *Mamm Genome* 9, 545–549.

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. (2000). NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* 28, 126–128.

Marra, M. A., Kucaba, T. A., Dietrich, N. L., Green, E. D., Brownstein, B., Wilson, R. K., McDonald, K. M., Hillier, L. W., McPherson, J. D., and Waterston, R. H. (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res.* 7, 1072–1084.

Matise, T., and Gitlin, J. (1999). MAP-O-MAT: marker-based linkage mapping on the World Wide Web. *Am. J. Hum. Genet.* 65, A2464.

Matise, T. C., Perlin, M., and Chakravarti, A. (1994). Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nat. Genet.* 6, 384–390.

McKusick, V. A. (1998). *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders*, 12th Edition (Baltimore: Johns Hopkins University Press).

Murray, J. C., Buetow, K. H., Weber, J. L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V. C., Sunden, S., Duyk, G. M., Weissenbach, J., Gyapay, G., Dib, C., Morrissette, J., Lathrop, G. M., Vignal, A., White, R., Matsunami, N., Gerken, S., Melis, R., Albertsen, H., Plaetke, R., Odelberg, O., Ward, D., Dausset, J., Cohen, D., and

Cann, H. (1994). A comprehensive human linkage map with centimorgan density. *Science* 265, 2049–2054.

Nusbaum, C., Slonim, D., Harris, K., Birren, B., Steen, R., Stein, L., Miller, J., Dietrich, W., Nahf, R., Wang, V., Merport, O., Castle, A., Husain, Z., Farino, G., Gray, D., Anderson, M., Devine, R., Horton, L., Ye, W., Kouyoumjian, V., Zemsteva, I., Wu, Y., Collymore, A., Courtney, D., Tam, J., Cadman, M., Haynes, A., Heuston, C., Marsland, T., Southwell, A., Trickett, P., Strivens, M., Ross, M., Makalowski, W., Wu, Y., Boguski, M., Carter, N., Denny, P., Brown, S., Hudson, T., and Lander, E. (1999). A YAC-based physical map of the mouse genome. *Nat. Genet.* 22, 388–393.

O'Brien, S. J., Menotti-Raymond, M., Murphy, W. J., Nash, W. G., Wienberg, J., Stanyon, R., Copeland, N. G., Jenkins, N. A., Womack, J. E., and Marshall Graves, J. A. (1999). The promise of comparative genomics in mammals. *Science* 286, 458–462.

Parra, I., and Windle, B. (1993). High resolution visual mapping of stretched DNA by fluorescent hybridization. *Nat. Genet.* 5, 17–21.

Pearson, P. L. (1991). The genome data base (GDB)—a human gene mapping repository. *Nucleic Acids Res.* 19 Suppl, 2237–9.

Pinkel, D., Straume, T., and Gray, J. W. (1986). Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc. Natl. Acad. Sci. USA* 83, 2934–2938.

Pruitt, K., Katz, K., Sicotte, H., and Maglott, D. (2000). Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* 16, 44–47.

Roberts, T., Auffray, C., and Cowell, J. K. (1996). Regional localization of 192 genic markers on human chromosome 1. *Genomics* 36, 337–340.

Rodriguez-Tome, P., and Lijnzaad, P. (2000). RHdb: the radiation hybrid database. *Nucleic Acids Res.* 28, 146–147.

Schuler, G. D. (1997). Sequence mapping by electronic PCR. *Genome Res.* 7, 541–550.

Schwartz, D. C., Li, X., Hernandez, L. I., Ramnarain, S. P., Huff, E. J., and Wang, Y. K. (1993). Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science* 262, 110–114.

Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA* 89, 8794–8797.

Slonim, D., Kruglyak, L., Stein, L., and Lander, E. (1997). Building human genome maps with radiation hybrids. *J. Comput. Biol.* 4, 487–504.

Steen, R., Kwitek-Black, A., Glenn, C., Gullings-Handley, J., Etten, W., Atkinson, S., Appel, D., Twigger, S., Muir, M., Mull, T., Granados, M., Kissebah, M., Russo, K., Crane, R., Popp, M., Peden, M., Matise, T., Brown, D., Lu, J., Kingsmore, S., Tonellato, P., Rozen, S., Slonim, D., Young, P., Knoblauch, M., Provoost, A., Ganten, D., Colman, S., Rothberg, J., Lander, E., and Jacob, H. (1999). A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Res.* 9, AP1–AP8.

Stewart, E. A., McKusick, K. B., Aggarwal, A., Bajorek, E., Brady, S., Chu, A., Fang, N., Hadley, D., Harris, M., Hussain, S., Lee, R., Maratukulam, A., O'Connor, K., Perkins, S., Piercy, M., Qin, F., Reif, T., Sanders, C., She, X., Sun, W., Tabar, P., Voyticky, S., Cowles, S., Fan, J., Mader, C., Quackenbush, J., Myers, R. M., and Cox, D. R. (1997). An STS-based radiation hybrid map of the human genome. *Genome Res.* 7, 422–433.

Talbot, C. A., and Cuticchia, A. J. (1994). Human Mapping Databases. In *Current Protocols in Human Genetics*, N. Dracopoli, J. Haines, B. Korf, D. Moir, C. Morton, C. Seidman, J. Seidman and D. Smith, eds. (New York: J. Wiley), p. 1.13.1–1.13.21.

Thomas, J. W., Summers, T. J., Lee-Lin, S. Q., Maduro, V. V., Idol, J. R., Mastrian, S. D., Ryan, J. F., Jamison, D. C., and Green, E. D. (2000). Comparative genome mapping in

the sequence-based era: early experience with human chromosome 7. *Genome Res.* 10, 624–633.

Tonellato, P. J., Zho, H., Chen, D., Wang, Z., Stoll, M., Kwitek-Black, A., and Jacob, H. *Comparative Mapping of the Human and Rat Genomes with Radiation Hybrid Maps*, RECOMB '99, Lyon, France, April 1999.

van den Engh, G., Sachs, R., and Trask, B. J. (1992). Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science* 257, 1410–1412.

Venter, J. C., Smith, H. O., and Hood, L. (1996). A new strategy for genome sequencing. *Nature* 381, 364–366.

Vollrath, D., Foote, S., Hilton, A., Brown, L. G., Beer-Romero, P., Bogan, J. S., and Page, D. C. (1992). The human Y chromosome: a 43-interval map based on naturally occurring deletions. *Science* 258, 52–59.

Wada, Y., and Yasue, H. (1996). Development of an animal genome database and its search system. Comput. Appl. Biosci. 12, 231–235.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lander, E. S., and et al. (1998). Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082.

Watanabe, T. K., Bihoreau, M. T., McCarthy, L. C., Kiguwa, S. L., Hishigaki, H., Tsuji, A., Browne, J., Yamasaki, Y., Mizoguchi-Miyakita, A., Oga, K., Ono, T., Okuno, S., Kanemoto, N., Takahashi, E., Tomita, K., Hayashi, H., Adachi, M., Webber, C., Davis, M., Kiel, S., Knights, C., Smith, A., Critcher, R., Miller, J., James, M. R., and et al. (1999). A radiation hybrid map of the rat genome containing 5,255 markers. *Nat. Genet.* 22, 27–36.

Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., and Rapp, B. A. (2000). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 28, 10–14.

White, P. S., Sulman, E. P., Porter, C. J., and Matise, T. C. (1999). A comprehensive view of human chromosome 1. *Genome Res.* 9, 978–988.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 7, 203–214.