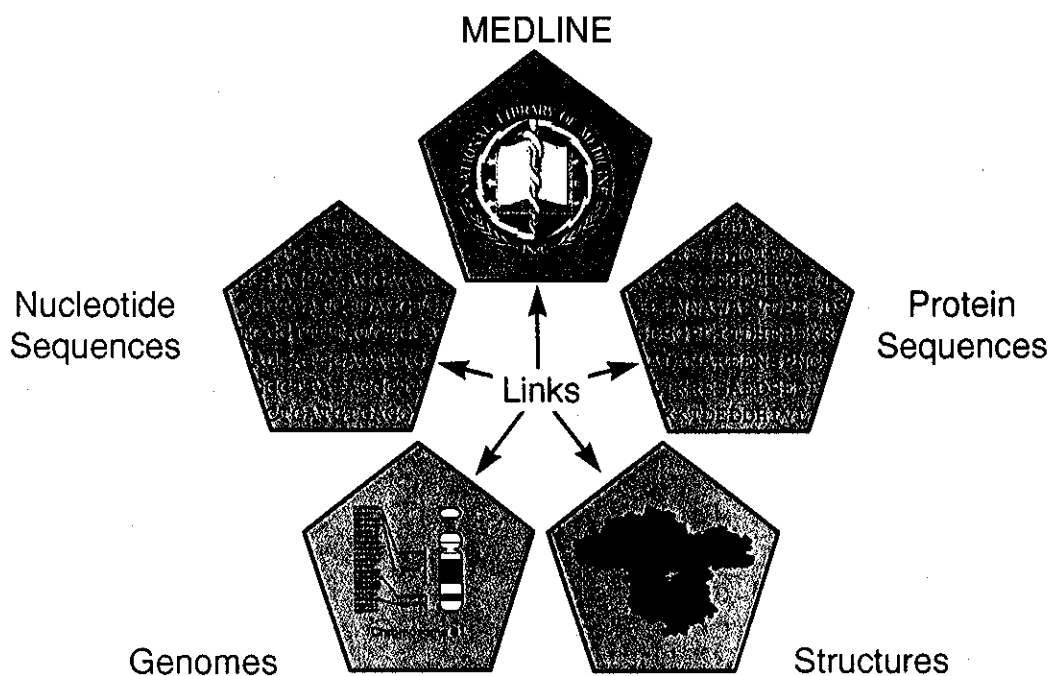


## INTEGRATED INFORMATION RETRIEVAL

The ability to traverse different information spaces is well demonstrated with the use of the World Wide Web and hyperlinks, but to have the capacity to track different databases in a stable, rigorous way requires a different type of infrastructure. Entrez is one example of integrated information retrieval within the field of molecular biology. Using Entrez, investigators can search nucleotide, protein, structure, and genome databases as well as the genetics subset of the MEDLINE bibliographic database, all by issuing a single query (Figure 5). It is quite logical to have a program that allows such travel between databases. All of these data are interrelated in the sense that protein sequences are derived from nucleotide sequences, structures are derived from isolated proteins of known sequence, articles are written on protein purification and gene mapping, and so on. Without becoming overly technical, what actually makes travel possible in Entrez is that all data elements from each of the constituent databases are converted into a single format called abstract syntax notation (ASN.1), in which all of the same elements (e.g., a bibliographic reference) are structured in the same way. Connections, or links, are then made between the different databases to allow users to traverse this information space.

Another type of integrated information retrieval system is ACeDB, initially developed for UNIX systems but now available for Macintosh systems and for



**Figure 5** Relationships between the component databases available through the Entrez system for integrated information retrieval.

Web browsers, which provide a graphical user interface. The power of ACeDB is based on its simplicity and adaptability to many database models, making it the tool of choice for many popular organism or chromosome-specific databases. Unfortunately, this strength also proves to be a weakness, since each database using ACeDB will be structured differently, making it more difficult to integrate information from various sources. Nevertheless, the popularity and power of ACeDB warrant further discussion.

## The Entrez System

When a "significant" match is identified by a database search for similarities, what does one do next? The information in the BLAST output (or the output of any other database search program) is necessary but seldom sufficient to evaluate the full significance and implications of a sequence match. The full database record(s), as well as references cited therein and other relevant literature, must then be studied. Users might also want to retrieve some of the matching sequences and perform additional database searches to confirm the results.

As discussed above, Entrez integrates these tasks so that users can search a single database, find all of the relevant information for that query within that database, and then move on to related information in all of the remaining databases without having to start another search. The ease with which users can jump from one database to the next allows a tremendous amount of information to be found in a fraction of the time it would have taken to search each constituent database separately. Entrez can also be used to link information to documents outside the databases; for example, in the electronic version of this chapter, each reference is linked to its corresponding MEDLINE entry.

To make the interconnections *within* databases, Entrez uses a procedure called neighboring to retrieve sets of related entries. Neighboring allows the user to ask the question: What publications are similar to a given publication? or What sequences are similar to a given sequence? Within the sequence databases, neighbors are determined by comparing each sequence with all others by using BLAST. For the MEDLINE subset, each entry is compared with all others through the use of weighted key terms, looking for the occurrence of words or phrases in the titles, keywords, and abstracts of other publications (Wilbur and Coffee 1994). All neighbors are precomputed, thereby substantially improving the speed with which related entries can be located and returned to the user.

Connections *between* databases are made through specific connections called hard links. For example, a publication on *BRCA1* found in MEDLINE may contain the nucleotide sequence for the *BRCA1* gene. If so, a hard link is established between the MEDLINE entry and the related entry in the nucleotide database. All hard links are reciprocal (i.e., the user can move between databases in any direction). As discussed above, hard links are established *anywhere* there is a logical connection between entries in different databases. Two examples of how both neighboring and hard links can be used to traverse a "biological discovery space" are discussed below.

*Example 5*

For the search result obtained with the BRCA1 exon in Example 1, the BLAST output (Figure 1b) shows a significant similarity to a sequence called rpt-1. Entrez can now be used to determine the function of this protein. The accession number (P15533) is entered into the Term: query box; the buttons labeled Accept and Retrieve 1 Document are then "pressed" in sequence. This leads the user to the actual database record for this protein (Figure 6a). The database record indicates that rpt-1 is a nuclear protein that contains a C3HC4-class zinc finger domain and regulates gene expression.

Entrez can now be used to determine whether this protein is homologous to other proteins in the databases. To do this, the user backtracks to the Document window and clicks the computer mouse on the small box to the left of the aa icon. This instructs Entrez to find all sequences related to the one just selected. In this case, 74 homologs, or sequence neighbors, were found, some of which are shown in Figure 6b. The protein named estrogen-responsive finger protein appears to be of interest in the context of the original search on rpt-1. To obtain the MEDLINE record for this estrogen-responsive finger protein, the user selects MEDLINE from the pop-up menu at the bottom of the Document window and clicks the computer mouse on the box to the left of the aa icon for this entry. Pressing the Lookup 1 at the bottom of the Document window produces a new window with a single MEDLINE entry listed (Figure 6c). The associated abstract for this entry indicates that the estrogen-responsive finger protein may mediate estrogen effects at the transcriptional level in certain cells isolated from mammary glands. This example illustrates how Entrez can be used to establish putative relationships between different proteins and biological function.

In addition to precomputed sequence homologies, Entrez performs a similar operation on MEDLINE records creating literature neighbors (i.e., articles that are related to each other based on their frequency of use of significant terms). From the retrieved MEDLINE record in Figure 6c, users can continue to branch out in the database and find other relevant publications. By taking advantage of these precomputed neighbors, users can readily assemble an entire bibliography on a new subject through just a few mouse clicks and keystrokes. For other examples of Entrez neighboring, see Cockerill (1994) and Harper (1994).

*Example 6*

In addition to moving between sequence and MEDLINE entries, Entrez can also be used to find related three-dimensional structural information. As an example, consider the case of disease genes that have been isolated through positional cloning (Collins 1995), whereby a previously identified sequence is known to be linked to a disease locus; this sequence is then tested for mutations that segregate with the disease phenotype. Such was the case for the Cu/Zn superoxide dismutase gene (*SOD1*) and the autosomal dominant form of amyotrophic lateral sclerosis (Table 7) (Rosen et al. 1993). A variant strategy is one where there is a strong indication of the biochemical basis for a given phenotype, where the sequence of a relevant enzyme is known but the location of its gene is not. In this case, identification of the gene occurs through a two-step process in which linkage must first be demonstrated and then the search for mutations

Query

Database: Protein    Field: Accession    Mode: Selection    Accept

Term: P15533

P15533    DOWN REGULATORY PROTEIN OF INTERLEUKIN 2 RECEPTOR.    01-APR-1993

P15534

P15535

P15538

P15539

P15540

P15541

↑ Term S

[P15533

Ret

Neigh

Document

P15533

Report    Graphic    GenPept    FASTA

133482    353 aa    01-APR-1993

LOCUS    DOWN REGULATORY PROTEIN OF INTERLEUKIN 2 RECEPTOR.

DEFINITION    133482

ACCESSION    SWISS-PROT: locus RPT1\_MOUSE, accession P15533

DESOURCE    class: standard.

KEYWORDS    created: Apr 1, 1990.

SOURCE    sequence updated: Apr 1, 1990.

ORGANISM    annotation updated: Apr 1, 1993.

REFERENCE    xrefs: genbank accession J03776, pir locus A30891.

AUTHORS    TRANS (non-sequence databases): HSSP P28990, PROSITE PS00518

   DURFEE T., SHENG F.-Y.H., SINGH P., JOHNSON K.A., GURRAGIA S.M.,

   BLATTNER F. and CANTOR H.

   PROC. NATL. ACAD. SCI. U.S.A. 85, 2733-2737 (1988)

JOURNAL    ZINC-FINGER, NUCLEAR PROTEIN.

MEDLINE    MOUSE.

REMARK    UNclassified.

COMMENT    PATAICA R., SCHWARTZ J., SINGH R.P., KONG Q.-T., MURPHY E.,

   ANDERSON Y., SHENG F.-Y.H., SINGH P., JOHNSON K.A., GURRAGIA S.M.,

   DURFEE T., BLATTNER F. and CANTOR H.

   PROC. NATL. ACAD. SCI. U.S.A. 85, 2733-2737 (1988)

   [FUNCTION] TRANS-ACTING FACTOR THAT REGULATES GENE EXPRESSION OF

Figure 6a GenPept format of a sequence record as it appears in Network Entrez. The equivalent view for the World Wide Web version would be a series of pages instead of the cascade of windows shown here. See following pages for Figure 6b and 6c.

Query

Database:  Field:  Mode:

Term:

**Document**

<input checked="" type="checkbox"/>	A49656	estrogen-responsive finger protein, efp (RING finger, coiled-coil domains) - human gi 542812 pir  A49656
<input type="checkbox"/>	SCYBR114W cds1	ORF YBR114w gi 536453
<input type="checkbox"/>	HUMHIP116A cds1	ATPase gi 531196
<input type="checkbox"/>	YSCRAD16B cds1	RAD16 gi 487900
<input type="checkbox"/>	SCRACII cds24	UY damage repair protein gi 476069
<input type="checkbox"/>	Q03605	HYPOTHETICAL 18.7 KD PROTEIN T02C1.1 IN CHROMOSOME III. gi 466015 sp Q03605 YNN1_CAEEL
<input type="checkbox"/>	Q03601	HYPOTHETICAL 104.4 KD PROTEIN F5468.4 IN CHROMOSOME III. gi 465914 sp Q03601 YMB4_CAEEL

↑ Term S [P15533] Ret

Target Database:

Select:

Figure 6b Use of Entrez to find sequence neighbors (homologous proteins) for nucleoprotein rpt-1.

Query

Database: Protein    Field: Accession    Mode: Selection    Accept

Term: P15533

P15533  
P15534  
P15535  
P15538  
P15539  
P15540  
P15541

↑ Terms S

[P15533

Ret

Neigh

Document

Inoue, 1993

Genomic binding-site cloning reveals an estrogen-responsive gene that encodes a RING finger protein.  
Proc Natl Acad Sci USA 90, 11117-21 (1993)

Inoue, 1993

MEDLINE    MEDLARS

Proc Natl Acad Sci USA 90: 11117-21 (1993) [94068555]

**Genomic binding-site cloning reveals an estrogen-responsive gene that encodes a RING finger protein.**

S. Inoue, A. Onimo, T. Hosoi, S. Kondo, H. Toyoshima, T. Kondo, A. Ikegami, Y. Ouchi, H. Onimo & M. Muramatsu  
Department of Biochemistry, Saitama Medical School, Japan.

Estrogen receptor (ER)-binding fragments were isolated from human genomic DNA by using a recombinant ER protein. Using one of these fragments as a probe, we have identified an estrogen-responsive gene that encodes a putative zinc finger protein. It has a RING finger motif present in a family of apparent DNA-binding proteins and is designated estrogen-responsive finger protein (epf). epf cDNA contains a consensus estrogen-responsive element at the 3'-untranslated region that can act as a downstream estrogen-dependent enhancer. Moreover, epf is regulated by estrogen as demonstrated at both the mRNA and the protein level in ER-positive cells derived from mammary gland. These data suggest that epf may represent an estrogen-responsive transcription factor that

Figure 6c MEDLINE record for the estrogen-responsive finger protein epf. This record represents the bibliographic reference corresponding to the sequence neighbor found in Figure 6b (A49656).

is performed. One of the genes for hereditary nonpolyposis colon cancer, *hMLH1* (Table 7), was cloned in this way (Bronner et al. 1994; Papadopoulos et al. 1994). Experiments had suggested that the pathophysiology of this cancer might be due to a defect in DNA mismatch repair, and genes encoding mismatch repair enzymes in bacteria (*mutL*) and yeast (*MLH1*) were already known. Isolation of the human homolog, mapping, and mutation detection ensued.

It should now be apparent that Entrez is useful in assessing candidate genes and functions via sequence data and MEDLINE literature. Starting with a biochemical function or an EC number, users can quickly find all relevant sequences and published articles. From the original Query window, a query on human *sod1* against the nucleotide databases returns six entries (Figure 7a). To determine whether any structural information is available, the user would click the computer mouse on the box next to the record of interest and then change the target database to Structure. In addition to providing atomic coordinates and textual descriptions of structures (Figure 7b), Entrez can interface with both the RasMol (Figure 7c) (Sayle 1994) and Kinemage (Figure 7d) (Richardson and Richardson 1992) programs to view and manipulate the structures graphically. The easy availability of three-dimensional structural information is of infinite value in the design of experiments intended to elucidate structure-function relationships. (Rosen et al. 1993). By the time this volume is published, the Entrez client will have its own built-in graphical viewer, named CN3D.

## Specialized Data Sets

In addition to Entrez, there are specialized data sets (often organism-specific) that also represent integrated information systems. These are often available on the World Wide Web, although they are also sometimes distributed on CD-ROM or are available by anonymous ftp. Two examples of these specialized databases are presented here (*C. elegans* and *D. melanogaster*). Additional databases, such as the SWISS-PROT annotated sequence database, are presented in Table 4.

### C. ELEGANS

The *C. elegans* community was one of the first groups to integrate and make available its genetic and physical map information through a graphical user interface. This was possible because of the pioneering work of J. Thierry-Mieg and R. Durbin and the great cooperation that exists among the members of the worm community. The software package built to disseminate this information was ACeDB, a *C. elegans* database. ACeDB is widely used by worm biologists and has been "cloned" as a database management tool for many nonworm genome projects. Hence, reference to the ACeDB tool does not automatically refer to *C. elegans* data. The ACeDB model has been used, for example, for yeast (SacchDB), *Arabidopsis* (AtDB), and rice.

Disease	Gene Symbol	Accession Number	OMIM Number	Low-Complexity Segments		Possible Non-Globular Regions	
				Number	Percent	Number	Percent
Aarskog-Scott syndrome	FGD1	U11690	305400	10	16.8%	1	38.2%
Achondroplasia	FGFLR	M64347	100800	4	8.3%	2	23.3%
Adenomatous polyposis coli	APC	M74088	175100	27	15.4%	10	62.9%
Adrenoleukodystrophy, X-linked	ALD	Z21876	300100	3	6.7%	3	40.7%
Alzheimer disease (Chromosome 14)	AD3	L42110	104311	4	14.1%	3	42.6%
Alzheimer disease (Chromosome 1)	AD4	L44577	600759	3	13.4%	1	27.5%
Amyotrophic lateral sclerosis	SOD1	K00065	105400	0	0.0%	0	0.0%
Aniridia	PAX6	M77844	106210	5	14.2%	2	59.7%
Ataxia telangiectasia	ATM	U26455	208900	1	0.9%	3	13.9%
Bloom syndrome	BLM	U39817	210900	9	11.2%	3	25.3%
Breast cancer type 1	BRCA1	U14680	113705	9	6.3%	5	32.6%
Breast cancer type 2	BRCA2	U43746	600185	7	2.9%	4	8.5%
Chedial-Higashi Syndrome	CHS		214500				
Choroideremia	CHM	X78121	303100	1	2.6%	2	27.6%
Chondrodysplasia punctata	ARSE	X83573	302950	4	11.2%	2	19.7%
Chronic granulomatous disease	NCF1	M55067	233700	1	3.3%	2	40.3%
Congenital adrenal hyperplasia	CYP21	M26856	201910	2	8.6%	1	19.4%
Cystic fibrosis	CFTR	M28668	219700	3	3.4%	2	12.2%
Diastrophic dysplasia	DTD	U14528	222600	5	11.0%	2	19.5%
Duchenne muscular dystrophy	DMD	M18533	310200	13	6.4%	13	44.1%
Emery-Dreifuss muscular dystrophy	STA	X82434	310300	3	20.9%	1	37.8%
Epidermolytic palmoplantar keratoderma	KRT9	X75015	144200	3	48.4%	3	63.3%
Fragile-X syndrome	FMR1	S65791	309550	2	5.4%	1	22.8%
Glycerol kinase deficiency	GK	L13943	307030	0	0.0%	0	0.0%
Gonadal dysgenesis	SRY	L08063	480000	0	0.0%	0	0.0%
Hereditary multiple exostoses	EXT1	S79639	133700	2	3.5%	0	0.0%

**Table 7** Inherited Disease Genes Identified by Positional Cloning. In the columns labeled low-complexity segments and possible nonglobular regions, the first number represents the number of segments or regions, and the second number indicates what 89% of the protein (amino acids) are found in those regions. Low-complexity segments are compositionally biased regions, which include different types of near-homopolymeric residue clusters, short-period repeats, and aperiodic mosaics of only a few residue types (Wootton 1994a,b). Gene, sequence, and complexity information is being gathered for those disease genes where no information is currently listed. (Continued on following page.)



Hereditary non-polyposis colon cancer	MLH1	U07343	120436	2	3.2%	1	10.2%
Hereditary non-polyposis colon cancer	MSH2	U03911	120435	2	2.5%	1	9.0%
Huntington disease	HD	L12392	143100	13	7.4%	6	22.6%
Hyperplexia	GLYRA2	X52009	149400	4	12.5%	0	0.0%
Hypophosphatemic rickets	XLH		307800				
Kallman syndrome	KAL	M97252	308700	3	7.5%	1	8.1%
Long QT syndrome	LQT1	U40990	192500	3	10.1%	0	0.0%
Lowe oculocerebrorenal syndrome	OCRL	M88162	309000	1	1.2%	1	12.2%
Machado-Joseph disease	MJD1	S75313	109150	1	7.2%	1	36.7%
McLeod syndrome	XK	Z32684	314850	2	7.9%	2	36.7%
Menkes disease	MNK	X69208	309400	1	2.4%	1	20.1%
Miller-Dieker lissencephaly	PAF	L13385	247200	0	0.0%	0	0.0%
Multiple endocrine neoplasia type 2a	RET	M57464	171400	2	2.9%	1	7.6%
Myotonic dystrophy	DM	L19268	160900	4	13.5%	3	61.4%
Myotubular Myopathy 1	MTM1		310400				
Neurofibromatosis type 1	NF1	M89914	162200	3	1.3%	1	2.5%
Neurofibromatosis type 2	NF2	L11353	101000	3	8.1%	2	44.0%
Norrie disease	NDP	X65882	310600	0	0.0%	0	0.0%
Obesity	OBS	U18915	164160	0	0.0%	1	40.7%
Ocular Albinism	OA1	Z48804	300500	3	15.3%	1	24.1%
Polycystic kidney disease	PKD1	L33243	173900	14	16.0%	4	69.4%
Retinoblastoma	RB1	M15400	180200	1	3.2%	2	7.9%
Spinal muscular atrophy	SMA	U18423	253300	4	37.4%	1	32.3%
Spinocerebellar ataxia 1	SCA1	X79204	164400	6	13.2%	4	72.7%
Thomsen disease	CLC1	Z25884	160800	7	9.8%	3	37.1%
Treacher Collins syndrome	TCS		154500				
Tuberous sclerosis	TSC	X75621	191090	9	8.2%	4	31.3%
von Hippel-Lindau syndrome	VHL	L15409	193300	2	20.4%	0	0.0%
Waardenburg syndrome	PAX3	U02309	268200	0	0.0%	1	51.8%
Werner syndrome	WRN		277700				
Wilms tumor	WT1	X51630	194070	4	12.9%	1	40.3%
Wilson disease	WND	L25591	277900	0	0.0%	0	0.0%
Wiskott-Aldrich syndrome	WASP	U12707	301000	6	8.7%	2	60.4%

Table 7 (Continued.)

Query

Database:  Field:  Mode:

Nucleotide  Gene Symbol

Term: sod1

	Document
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Human Cu/Zn superoxide dismutase gene. gi 181207 gb M13267 HUMCUZNDI
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Human superoxide dismutase (SOD-1) mRNA, complete cds. gi 338275 gb K00065 HUMSOD1
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Human Cu/Zn superoxide dismutase (SOD1) gene, exon 4, with a mutation causing amyotrophic lateral sclerosis. gi 927069 gb L44141 HUMSOD104M
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Human superoxide dismutase (SOD-1) gene, exon 5 with a mutation causing familial amyotrophic lateral sclerosis. gi 928824 gb L44746 HUMSOD1MUB
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Homo sapiens superoxide dismutase (SOD-1) gene, exon 2 with a mutation causing familial amyotrophic lateral sclerosis. gi 939772 gb L46374 HUMSOD1MUA
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Homo sapiens superoxide dismutase (SOD-1) gene, exon 5 with a mutation causing familial

↑ Term S

[ human  
[ sod1

Ret

Lookup 0

Target Database:

Figure 7a Use of multiple terms in Entrez to narrow the results returned from a query. Six nucleotide entries are returned here by using the terms human and sod1 together. See following pages for Figures 7b, 7c, and 7d.

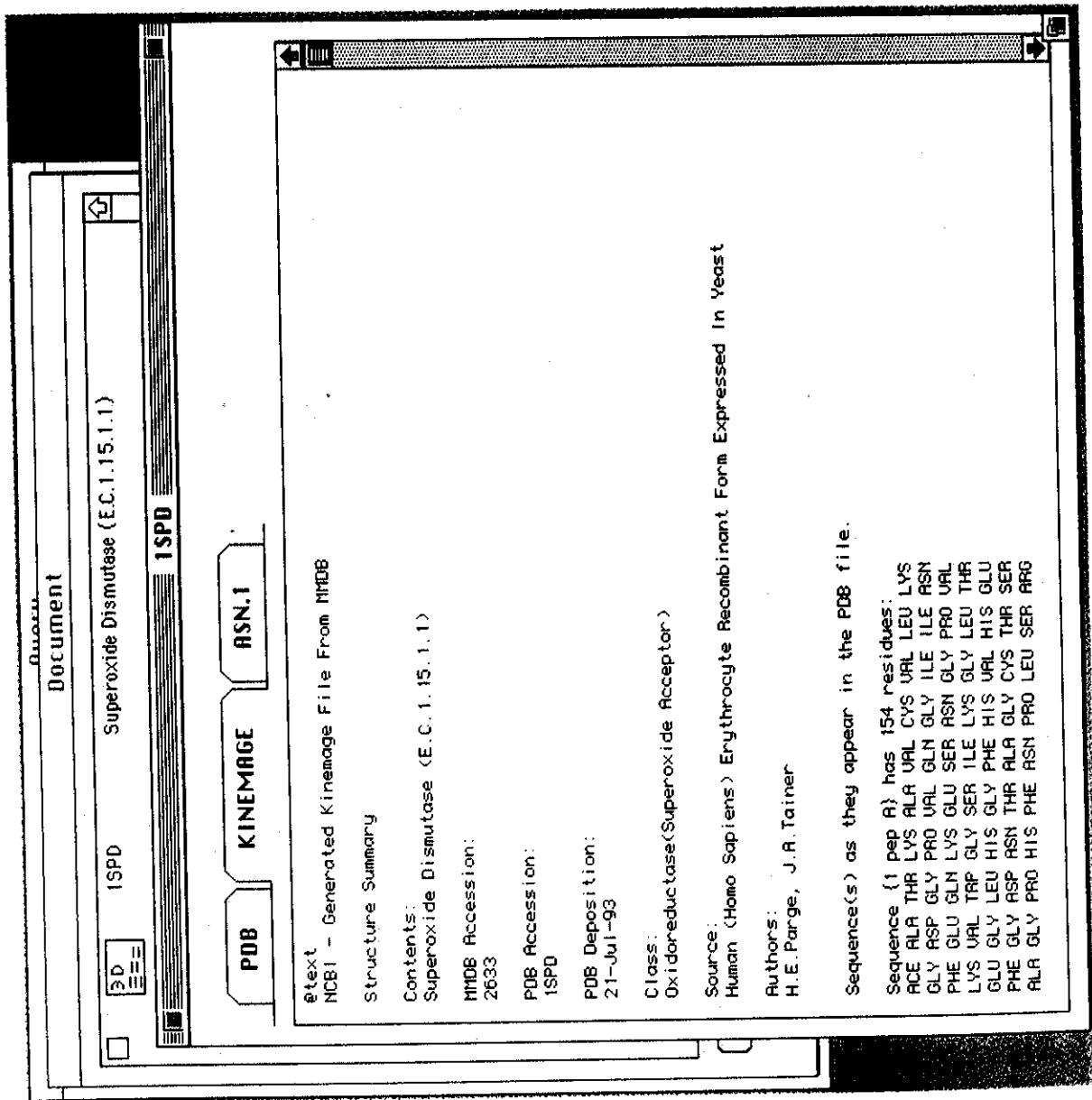


Figure 7b MMDB representation of a structure record, providing both coordinate information and a textual description of the structure.

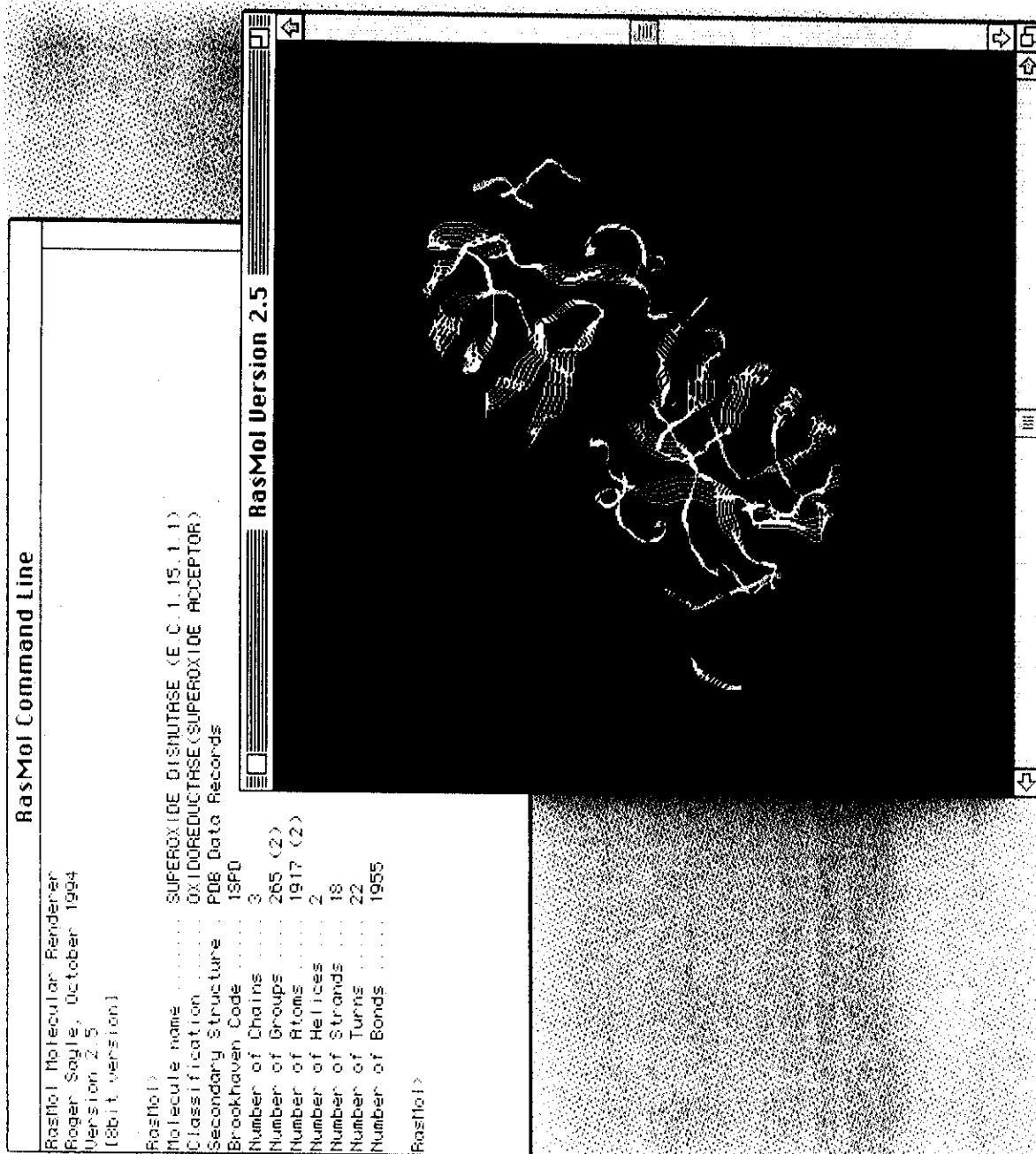


Figure 7c Use of RasMol to view three-dimensional structure information available through Entrez. This structure is superoxide dismutase.

NCBI - Generated Kinemage File From MMDB

Structure Summary

Contents:  
Superoxide Dismutase (E.C.1.15.1.1)

MMDB Accession: 2633

PDB Accession: 1SPD

PDB Deposition Date: 21-Jul-9

Class: Oxidoreductase

Source: Human (Homo sapiens)

Authors: H.E.Parge

**KINEMAGE Color Graphics**

1 pep A

backbone

residues

H-links

2 pep B

backbone

Virtual

Real

Residues

pickcenter

zclip

100 - ZOOM +

200 - ZSLAB +

0 - ZTRAN +A

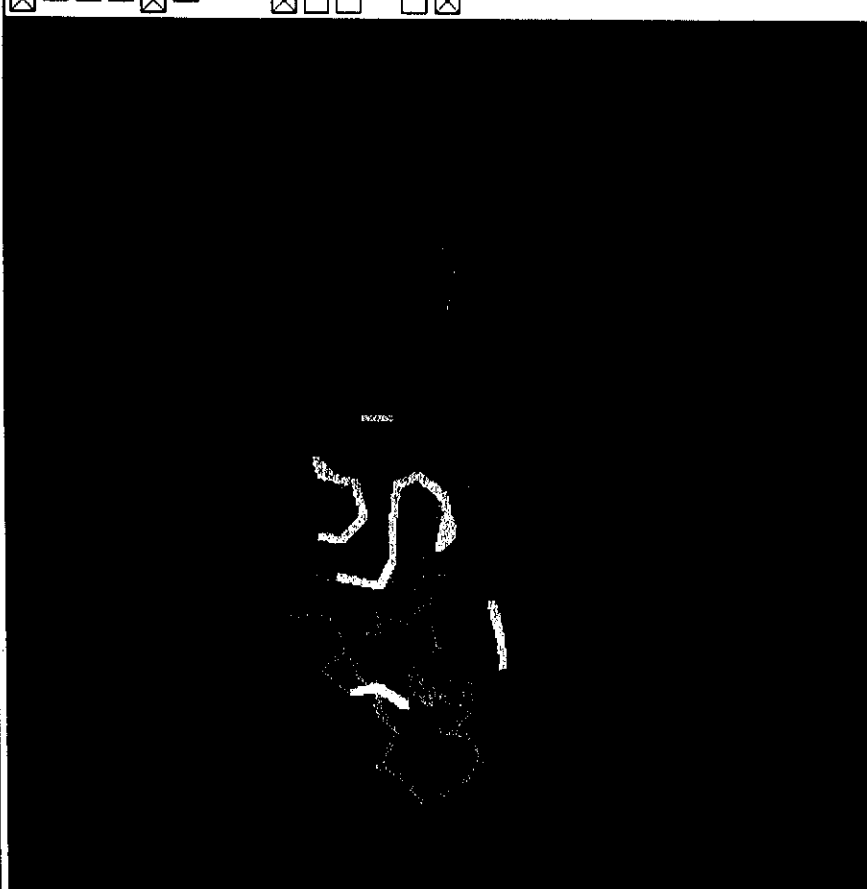


Figure 7d Use of Kinemage to view three-dimensional structure information available through Entrez. This structure is superoxide dismutase.

In its native form, ACeDB runs as a UNIX application, presenting users with a graphical interface through which they can move between genetic maps, physical maps, DNA sequences, clone grids, bibliographic information, Southern blot GIF files, genetic cross data, and so on. A version of ACeDB is now available for the Macintosh, and a similar text-based form (tace) is available for use on Gopher servers. A copy of frequently asked questions for ACeDB can be obtained from the Sanger Center at:

<http://www.sanger.ac.uk/HTML/acedbfaq.html>

#### **D. MELANOGASTER**

The *Drosophila* community began sharing its information in a slightly different way. D. Gilbert was instrumental in making much of the Red Book available on Gopher servers before the advent of the World Wide Web. Flybase, as it is now called, is a comprehensive database for information on the genetics and biology of *Drosophila* and is available via ftp, Gopher, and the World Wide Web. Access information can be obtained at:

<http://morgan.harvard.edu/About-flybase.html>

Like ACeDB, Flybase represents a substantial body of integrated information on and about *D. melanogaster*. A comparison of Flybase to ACeDB shows that more genetic information is available on Flybase, whereas more molecular sequence data are available on ACeDB. Flybase is also represented within GenBank as a cross-referenced database, i.e., related GenBank records will contain a cross-reference qualifier (`/db_xref`) as well as a hyperlink to Flybase.

# MULTIPLE ALIGNMENT, SEQUENCE MOTIFS, AND STRUCTURE INFERENCE

## Multiple Alignment

As in the case of amyotrophic lateral sclerosis and the *SOD1* gene product, there may not be a crystal structure upon which to evaluate the possible effects of mutations. More often, homologs in other organisms are available that can be used for comparative sequence analysis. Multiple alignments are performed to study similarities and differences in a group of related sequences. The purpose here is to assess whether mutational changes are expected to lead to loss or diminution of function. If a change in a sequence is one that results in a non-conservative amino acid substitution, particularly in a residue that multiple alignment has shown to be conserved in a group of evolutionarily distant sequences, then that change is more likely to represent a deleterious mutation than a silent polymorphism.

Multiple alignment is too large and active a research area in computational biology for justice to be done to it here. Instead, some useful programs that are freely available in the public domain are discussed. CLUSTAL W is among the most powerful multiple sequence alignment packages available, and it performs progressive multiple sequence alignments based on the method of Feng and Doolittle (1987). Each pair of sequences is aligned and the distance between each pair is calculated; from this distance matrix, a guide tree is calculated, and all of the sequences are progressively aligned based on this tree. A major feature of the program is its sensitivity to the effect of gaps on the alignment; gap penalties are varied to encourage the insertion of gaps in probable loop regions instead of in the middle of structured regions. Users can specify gap penalties, choose between a number of scoring matrices, or supply their own scoring matrix for both the pairwise alignments and the multiple alignments. The output can be obtained in a number of file formats, allowing users to proceed directly to phylogenetic packages such as PHYLIP (Felsenstein 1993) or publication-quality alignment formatters such as ALSCRIPT (Barton 1993). CLUSTAL W for UNIX and VMS systems is freely available by anonymous ftp at:

`ftp.ebi.ac.uk`

or by E-mail at:

`netserv@ebi.ac.uk`

CLUSTAL W can also be accessed as an external module through SeqApp, a Macintosh sequence editor and analysis program (Gilbert 1992) available by anonymous ftp at:

`ftp.bio.indiana.edu`

Another useful program is MACAW (Schuler et al. 1991), for which both Macintosh and Microsoft Windows versions are available. MACAW uses a

graphical interface, provides a choice of several alignment algorithms, and is available by anonymous ftp at:

ncbi.nlm.nih.gov (directory/pub/macaw)

MACAW has been used, for example, to analyze the product of the diastrophic dysplasia gene (Table 7) (de la Hastbacka et al. 1994) that encodes a novel sulfate transporter related to several previously described sequences. MACAW has also been used to align the product of the human *MLHI* gene for hereditary nonpolyposis colon cancer with its various yeast and bacterial homologs (Papadopoulos et al. 1994). As concerted genomic and cDNA sequencing progresses, it will become more and more likely that a positionally cloned gene will already have homologs in the database at the time of its isolation. Multiple sequence alignment will therefore become an increasingly important method of data analysis.

## Sequence Motifs

The cornerstone of sequence analysis is database searching for pairwise similarity between sequences. As described above, multiple alignment adds another useful dimension to sequence analysis. Sequence motifs are derived from multiple alignments and can be used to examine individual sequences or an entire database for subtle patterns. With motifs, it is sometimes possible to detect distant relationships that may not be demonstrable based on comparisons of primary sequences alone. The derivation and use of sequence motifs are very active research areas with a sizable literature. Only a few examples from this field are discussed here. For access to some of the latest developments, see Tatusov et al. (1994 and references cited therein).

Currently, the largest collection of sequence motifs in the world is PROSITE (Bairoch and Bucher 1994). PROSITE can be accessed via either the ExPASy server on the World Wide Web or anonymous ftp site. A free software package named MacPattern (Fuchs 1991) is available for searching PROSITE motifs; MacPattern is available through anonymous ftp from EBI and other sites (Table 2). Many commercial sequence analysis packages also provide search programs that use PROSITE data.

One of the most useful resources for searching protein motifs is the BLOCKS E-mail server (Table 1) developed by S. Henikoff (Henikoff and Henikoff 1991; Henikoff 1993). BLOCKS searches a protein or nucleotide sequence against a database of protein motifs or "blocks." Blocks are defined as short, ungapped multiple alignments that represent highly conserved protein patterns. The blocks themselves are derived from entries in PROSITE as well as other sources. Either a protein or nucleotide query can be submitted to the BLOCKS server; if a nucleotide sequence is submitted, the sequence is translated in all six reading frames and motifs are sought in these conceptual translations. Once the search is completed, the server will return a ranked list of significant matches, along with an alignment of the query sequence to the matched BLOCKS entries.

PROSITE and BLOCKS represent collected families of protein motifs. Thus, searching these databases entails submitting a single sequence to determine whether or not that sequence is similar to the members of an established family.



Programs working in the opposite direction compare a collection of sequences with individual entries in the protein databases. An example of such a program is the Motif Search Tool, or MoST (Tatusov et al. 1994). On the basis of an aligned set of input sequences, a weight matrix is calculated by using one of four methods (selected by the user); a weight matrix is simply a representation, position by position in an alignment, of how likely a particular amino acid will appear. The calculated weight matrix is then used to search the databases. To increase sensitivity, newly found sequences are added to the original data set, the weight matrix is recalculated, and the search is performed again. This procedure continues until no new sequences are found.

An increasingly important use of motifs in the future will be to "preprocess" query sequences to determine the presence of known domains and then mask these regions (see pp. 543–559) before a full-scale BLAST search. This should simultaneously increase the speed of the search and improve the ability to detect subtle matches that would otherwise be swamped by abundant strong matches to other sequence regions, such as kinase catalytic domains (Altschul et al. 1994).

## Structure Prediction and Protein Modeling

In the course of studies aimed at rational drug design or determining the biochemical function of a protein, knowledge of the structure of the protein is of critical importance. X-ray crystallography and NMR spectroscopy are two main experimental methods by which the tertiary structure of a protein can be determined. Both of these methods, however, are technically demanding, time-consuming, and not amenable to automation. These methods will therefore not be able to keep pace with the discovery of new sequences. For this reason, predictive methods are necessary to address the need for structural insights in the absence of direct experimental data. As with all such methods, it must be kept in mind that no matter how good the method, the results are still predictions, and different methods will give different predictions. With this precaution in mind, and in conjunction with supporting biochemical data, these methods can provide valuable insights into protein structure.

As the structures of more and more proteins are being solved, it is becoming increasingly apparent that there is a relatively small set of three-dimensional motifs into which proteins are observed to fold (Chothia 1992). This observation, coupled with the concept that protein structure is conserved to a greater extent than sequence (Chothia and Lesk 1986), has led to the development of sophisticated methods by which the three-dimensional structure of a protein of interest can be predicted. One of the most promising methods is known as homology model building, or threading (for review, see Fetrow and Bryant 1993). In this method, a query sequence of unknown structure is threaded through the coordinates of a homologous protein of known structure (Bryant and Lawrence 1993). All possible placements of the query sequence subject to a number of set physical constraints are attempted; for example, core regions ( $\alpha$ -helices or  $\beta$ -sheets) are kept at a fixed length, and loop regions are allowed to have variable lengths within limits. By evaluating pairwise and hydrophobic in-

teractions between nonlocal residues, this method is able to identify the most energetically favorable and conformationally stable placements of the query sequence with respect to the known structure.

The threading technique was recently applied to a DNA-binding motif found within high-mobility-group proteins HMG-1 and HMG-2, a motif called the HMG-1 box. In this case, a number of non-HMG DNA-binding proteins were shown to form the HMG-1 box motif despite the absence of statistically significant sequence similarity. Another recent example involves the *ob* gene product which, when mutated, is associated with obesity and type II diabetes in mice (Madej et al. 1995). Here, the *ob* gene product was found to be similar to a family of helical cytokines that includes interleukin-2 and growth hormone. Although threading programs are not yet widely available, their development will provide the molecular biologist with a very powerful tool with which to deduce structural similarities that are not necessarily obvious through traditional sequence alignment techniques.

For a newly discovered gene product, it is possible to determine if there is a homologous protein in the structure database that might make tertiary structure modeling feasible via threading. If there is no such structure available, methods exist for predicting possible secondary structure of a new sequence. Since the original method of Chou and Fasman (1974) was introduced, a number of different algorithms have been developed to predict the secondary structure of proteins. The end product of all of these methods is the same: a prediction of how likely protein sequences are to assume an  $\alpha$ -helical,  $\beta$ -sheet, or random-coil conformation. Most of these methods rely on patterns that can be deduced from sets of proteins for which three-dimensional structure information is already available.

Several E-mail servers are currently available for such secondary structure prediction, each based on a slightly different method (see Table 1). The nnpredict algorithm (Kneller et al. 1990) uses a FASTA-formatted sequence as its input and allows the user to select the tertiary structure class of the protein; the server then returns the most likely structure for each individual residue in the query. PredictProtein (Rost and Sander 1993) first takes the query sequence and performs a database search, from which a multiple sequence alignment is derived. The information in this alignment (a sequence profile) is then used to improve the accuracy of the prediction. A detailed report of the probability of each individual amino acid assuming an  $\alpha$ -helical,  $\beta$ -sheet, or random-coil conformation is then returned. The accuracy rate for the best-case prediction using nnpredict is reported to be 79%; for PredictProtein, the average accuracy rate is reported to be 71.6% for all residues and 92% for the most reliably scored residues in the query. One important caveat to all of these methods is that they are based on extrapolations from the existing structure database, which is biased toward globular sequences. As work by Wootton (1994a,b) and others has shown, however, a large fraction of sequences or sequence domains have non-globular structures.

A relevant example of where motif analysis and tertiary structure modeling have advanced our understanding of a positionally cloned human disease gene is the case of Norrie disease, an X-linked disorder characterized by progressive blindness, deafness, and mental retardation. When the Norrie disease gene was

first cloned, no significant homologies were found during database searches and no motifs from PROSITE were detected (Berger et al. 1992; Chen et al. 1992). However, on the basis of a particular arrangement of conserved cysteine residues and database searches with consensus patterns, Meitinger et al. (1993) constructed a three-dimensional model showing that the Norrie disease protein is most likely to have a structure similar to that of transforming growth factor- $\beta$ .

## SUBMITTING DATA TO PUBLIC DATABASES

An important responsibility for investigators is to make their sequence data available to the scientific community by submitting sequence data to a public sequence database. Historically, most funding agencies have encouraged submissions, and journals have required a database accession number as proof of submission as a condition of publication. (An accession number is a unique identifier for a particular sequence and allows retrieval of the sequence and associated annotation.) Accession numbers should be referenced in reports of new sequences or descriptions of experiments and analyses based on existing sequences. It is best to refer to accession numbers in all publications (in a footnote if possible) in the following fashion: These sequence data have been submitted to the DDBJ/EMBL/GenBank databases under accession number AA123456. This is an essential component of research reproducibility in the electronic age. Accession numbers have been in the one letter plus five digits format (e.g., U12345) since their inception, but growth of the databases has recently caused the databases to switch to a different format where two letters are used with six digits (e.g., AA123456).

GenBank, EMBL, and DDBJ are the three partners in the International Collaboration of Nucleotide Sequence Databases and users may submit their sequences to whichever of these databases is most convenient. The general E-mail addresses and World Wide Web URL used to submit data to these databases are listed below:

Database	E-mail submissions	WWW submissions
DDBJ	<a href="mailto:ddbjsub@ddbj.nig.ac.jp">ddbjsub@ddbj.nig.ac.jp</a>	<a href="http://sakura.ddbj.nig.ac.jp/">http://sakura.ddbj.nig.ac.jp/</a>
EMBL	<a href="mailto:datasubs@ebi.ac.uk">datasubs@ebi.ac.uk</a>	<a href="http://www.ebi.ac.uk/subs/emblsubs.html">http://www.ebi.ac.uk/subs/emblsubs.html</a>
GenBank	<a href="mailto:gb-sub@ncbi.nlm.nih.gov">gb-sub@ncbi.nlm.nih.gov</a>	<a href="http://www3.ncbi.nlm.nih.gov/BankIt/">http://www3.ncbi.nlm.nih.gov/BankIt/</a>

Since submitted data are exchanged between all three partners (as well as other sites) on a daily basis, users should submit their sequences to *only one* of these locations. For example, once a sequence is sent to GenBank and is made public, the submission is automatically shared with EMBL and DDBJ during daily updates. Therefore, a newly submitted sequence will be available within a few days at the other two sites as well. Overall, this arrangement allows the best management of the data with optimal accessibility at sites located throughout the world. Protein sequences contained in these databases are also distributed to a number of protein databases (SWISS-PROT, PIR International, and GenPept) in a similar fashion. The protein databases use the nucleotide databases as their primary source of new amino acid sequences, so there is no need to submit to them separately. The protein databases will get the sequences from the DNA sequence records.

## Preparing a New Submission to GenBank

Many different tools are available for submitting data to the three data repositories described above. All three sites have their own experts that can provide advice on submission, annotation, presentation, and even data analysis issues. Details are provided here for the interaction between NCBI and investigators submitting sequences to GenBank.

At NCBI, more than 80% of sequences (not including ESTs and entries from genome centers) are submitted using the World Wide Web-based tool BankIt. BankIt is a simple, document-based tool that enables easy step-by-step entry of sequence and associated biological information. This information is, in turn, converted into a format that allows the rapid biological and computational checks performed on all sequence data submitted to GenBank.

A diminishing number of submissions are still coming from files prepared by using a tool called Authorin. Available for both Macintosh (albeit only the 16-bit variety) and PC systems (without Windows), Authorin was among the first generation of submission tools, providing a standard interface through which sequences and information related to those sequences could be submitted. The ease of using the World Wide Web, however, has steered users to BankIt.

NCBI, in collaboration with the other nucleotide sequence databases, has recently released a beta version of Sequin, which is a specialized network or stand-alone (i.e., not requiring a network connection) application for the submission of sequence data and associated annotation. It is intended as a replacement for Authorin. Sequin differs from BankIt in having more sophisticated capabilities, with specialized editors allowing more accurate annotations, as well as built-in validating capabilities. All objects that can be viewed in Sequin, through either a flatfile view or the graphical interface, can be clicked on to bring up the appropriate editor. A simple submission can be generated from stored FASTA files in a matter of minutes. Sequin software is available for the Macintosh (all varieties), PC (Windows 16/32 bit), and UNIX systems.

As mentioned above, sequences submitted to GenBank go through a series of validation checks conducted by a professional database staff. These quality-assurance procedures are designed to detect vector and other sequence contaminants (e.g., mitochondrial sequences, which should be absent from nucleus-encoded genes) and mistranslations of coding sequences and to ensure proper taxonomic classification. As a final check, all records submitted through the direct submission process at GenBank are checked by at least one Ph.D.-level molecular biologist. Accession numbers are issued to the submitters in less than 24 hours, and provisional GenBank records (e.g., see Figure 8) are made available for approval before to public release. Authors may request that their data remain confidential until publication.

The most important component of a submission is, of course, the new sequence itself. Investigators using GenBank data want to be assured that submitted data are free of sequencing errors, cloning artifacts, and so on; routine checks can be done in this regard. With the advent of single-pass strategies for both cDNA and genomic data, however, a great deal of lower-quality but nevertheless very useful data have been included in special divisions of GenBank. In these cases, it is important for the submitter to provide an estimate of sequencing accuracy. The database staffs have been very insightful in creating divisions that permit the sequestration of subsets of the sequences, so that they

can be manipulated, interpreted, and used in specific ways that allow for maximum benefit. This has been the case for ESTs and STSs (unique, sequence-defined landmarks in the genome that represent a specific PCR) and is now the case for the GSS division, which represents low-pass genomic sequences coming from a variety of genome projects. Although it was not long ago when the usefulness of ESTs was questioned because of the lower accuracy inherent in single-pass sequence readings, the full advantage and benefit of ESTs are no longer challenged.

What basic information does one need to know about a sequence before submitting it to GenBank? Many features can be attached to a sequence, as detailed in the GenBank Feature Table Document, available on the World Wide Web at:

<http://www.ncbi.nlm.nih.gov/collab/FT/index.html>

or by anonymous ftp at:

[ncbi.nlm.nih.gov \(directory/genbank/docs\)](http://ncbi.nlm.nih.gov/directory/genbank/docs)

The list of features and their qualifiers are sometimes quite overwhelming, but the list is designed (and updated) to cover all possible needs clearly and unambiguously. The most convenient way to become familiar with the GenBank annotation system is to examine some existing GenBank records by looking them up by using either Entrez or the QUERY E-mail server (see Table 1). As an example, the GenBank record for the hereditary breast cancer gene, *BRCA1*, is shown in Figure 8. A larger and more complex example is the record for the breakpoint cluster region gene on human chromosome 22. Finally, although the GenBank entries for positionally cloned genes shown in Table 7 are not all sterling examples of well-done records, they do give a good sense of the variety of types of information that GenBank annotations can accommodate.

One of the most important characteristics of a DNA sequence is its coding potential (the CDS feature in GenBank jargon). All mRNA (cDNA) sequences and exons should have their coding regions precisely specified, and the conceptual translations should be supplied. An equally important and very useful piece of information is the gene name and product name for the feature of interest. These items are stored in the GenBank record, which is the best way to ensure that the encoded protein is represented in protein sequence databases, without having to perform separate submissions. Proper annotation of the CDS feature also ensures that the records will be properly linked in Entrez.

## Updates and Corrections to GenBank

In addition to new data, investigators are strongly encouraged to submit updates and corrections to their GenBank records. A submission remains the investigator's publication, and although the databases do have an editorial role in maintaining a certain standard, only biologists themselves can help maintain the quality of the biological information in the databases. For instance, if a gene name is now known or if there is a better understanding of the enzyme for which a sequence was previously submitted (e.g., it now has an EC number), it is highly desirable to inform the databases of this new information so that records can be properly updated.

```

LOCUS       HSU14680      5711 bp    mRNA          PRI          05-AUG-1995
DEFINITION Human breast and ovarian cancer susceptibility (BRCA1) mRNA,
            complete cds.
ACCESSION  U14680
NID        g555931
KEYWORDS
SOURCE     human.
  ORGANISM Homo sapiens
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE  1 (bases 1 to 5711)
AUTHORS    Miki Y., Swensen J., Shattuck-Eidens D., Futreal P.A., Harshman K.,
            Tavtigian S., Liu Q., Cochran C., Bennett L.M., Ding W., Bell R.,
            Rosenthal J., Hussey C., Tran T., McClure M., Frye C., Hattier T.,
            Phelps R., Haugen-Strano A., Katcher H., Yakumo K., Gholami Z.,
            Shaffer D., Stone S., Bayer S., Wray C., Bogden R., Dayananth P.,
            Ward J., Tonin P., Narod S., Bristow P.K., Norris F.H., Helvering
            L., Morrison P., Rosteck P., Lai M., Barrett J.C., Lewis C.,
            Neuhausen S., Cannon-Albright L., Goldgar D., Wiseman R., Kamb A.
            and Skolnick M.H.
TITLE      A strong candidate for the breast and ovarian cancer susceptibility
            gene BRCA1
JOURNAL    Science 266 (5182), 66-71 (1994)
MEDLINE    95025896
REFERENCE  2 (bases 1 to 5711)
AUTHORS    Skolnick, M.H.
TITLE      Direct Submission
JOURNAL    Submitted (14-SEP-1994) Mark H. Skolnick, Myriad Genetics Inc. and
            the University of Utah, 421 Wakara Way, Suite 201, Salt Lake City,
            UT 84108, USA
FEATURES   Location/Qualifiers
  source   1..5711
            /organism="Homo sapiens"
            /note="For sequence of alternatively spliced exon 4, see
            GenBank Accession Number U15595"
            /chromosome="17"
            /map="17q21; spans D17S855"
  5'UTR    1..119
  exon     1..100
            /number=1
  exon     101..199
            /number=2
  CDS      120..5711
            /gene="BRCA1"
            /note="influences susceptibility to breast and ovarian
            cancer"
            /codon_start=1
            /db_xref="PID:g555932"
            /translation="MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFC
            KFCMLKLLNQKKGPSQCPLCKNDITKRSLQESTRFSQLV ELLKIIICAFQLDTGLEYA
            <translation truncated for brevity>
            QLCGASVVKELSSFTLGTGVHPVIVVQPDAWTEDNGFHAIGQMCEAPVVTREWVLSV
            ALYQCQELDTYLIPQIPHSY"
  exon     200..253
            /gene="BRCA1"
            /number=3
  exon     254..331
            /gene="BRCA1"
            /number=5

  <exon list in feature table truncated for brevity>

BASE COUNT  1956 a   1099 c   1274 g   1382 t
ORIGIN
   1 agctcgetga gacttcctgg accccgcacc aggctgtggg gtttctcaga taactggggc
  61 cctgcgetca ggaggccttc accctctgct ctgggtaaag ttcattggaa cagaagaaaa
 121 tggatttacc tgctcttcgc gttgaagaag tacaaaatgt cattaatgct atgcagaaaa
 181 tcttagagtg tccatctgt ctggagtga tcaaggaacc tgctccaca aagtgtgacc

  <sequence truncated for brevity>

```

Figure 8 GenBank record for the hereditary breast cancer gene, *BRCA1*.

As the genomes of various organisms are completed, errors, omissions, and misplacements will become apparent to some users, and notification to the databases will be critical to ensure the quality of the databases. It will become the responsibility of all to participate in the gigantic task of maintaining order in this sea of data. The staffs of the databases themselves must be very diligent in their task to keep up with the flow of information, but users also need to direct the focus of the databases, since they are the users of this information. If the users do not voice their opinions, the database staffs have no way to speculate on what may be needed (although they will probably hazard a guess).

Users of the databases who notice a problem, error, or omission in a given entry are therefore encouraged to bring the discrepancy to the attention of the GenBank staff. For example, it is not uncommon for submitters to forget to notify GenBank that a confidential sequence has been published and can be made public; this is almost certainly the case when the entry for a published GenBank accession number cannot be retrieved. GenBank will release the entry when the complete journal citation, including the full title of the article, is sent via E-mail to [update@ncbi.nlm.nih.gov](mailto:update@ncbi.nlm.nih.gov) or when the first page of the article and the page containing the cited accession number is sent via fax to NCBI at 301-480-9241. Updates can also be submitted through the World Wide Web by using BankIt's Update option.

The updated information should be sent to only *one* of the databases, because the information will be shared with the other collaborating databases as described earlier. The addresses for submitting updates to each of the collaborating databases are listed below.

Database	E-mail update	WWW update
DDBJ	<a href="mailto:ddbjudt@ddbj.nig.ac.jp">ddbjudt@ddbj.nig.ac.jp</a>	Use E-mail for DDBJ updates
EMBL	<a href="mailto:update@ebi.ac.uk">update@ebi.ac.uk</a>	<a href="http://www.ebi.ac.uk/ebi_docs/update.html">http://www.ebi.ac.uk/ebi_docs/update.html</a>
GenBank	<a href="mailto:update@ncbi.nlm.nih.gov">update@ncbi.nlm.nih.gov</a>	<a href="http://www3.ncbi.nlm.nih.gov/BankIt/">http://www3.ncbi.nlm.nih.gov/BankIt/</a>

### Special Arrangements for Large Projects

During the past several years, large projects with the goal of sequencing entire chromosomes or genomes have become more common, and single-pass survey sequencing is generating tens of thousands of new sequences every month. The database submission needs for these projects are not conveniently met by the standard methods discussed so far. Large, well-organized laboratories or consortia that are carrying out this work usually have their own sophisticated informatics capabilities and laboratory information management systems. In this milieu, the professional database staffs become close collaborators in getting the information into the public repositories and making it available in a form that allows the entire scientific community to benefit.



Several years ago, NCBI devised special streamlined submission procedures for rapidly accumulating EST and STS data and has also worked with major sequencing groups to provide convenient interfaces between local laboratory information management systems (such as ACeDB) and GenBank. EBI and DDBJ have also made special arrangements with high-throughput sequencing laboratories. If data handling requirements exceed the capacity or capabilities of existing submission tools, the GenBank staff can discuss alternatives that will ensure accurate, efficient, and timely submission, annotation, and distribution of such sequence data.

Since data throughput may soon approach hundreds of megabases a year, a major challenge will be to provide up-to-date annotation for this sequence data. A new class of data from high-throughput genome sequencing centers will soon be present in the database in the HTG division. These data will be primarily large records (>100 kb) and will have automatic annotations for repeats, structural RNAs, and homology, as well as some experimentally carefully annotated features. The data will initially be updated frequently (once or twice a month) and will be retrievable through the normal channels (e.g., Entrez), which will have all the typical neighbor information, but which will now be anchored to the various genetic maps present in the genomes division in Entrez.

The public databases will necessarily have an increasingly important role in this endeavor. Keeping homology information and links to relevant literature current, as in the Entrez system, will be a useful and essential approach to this task.