Bioinformatics Resources From the National Center for Biotechnology Information: An Integrated Foundation for Discovery

Barbara A. Rapp  
National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894. E-mail: RappB@mail.nih.gov

David L. Wheeler  
National Center for Biotechnology Information, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894. E-mail: wheeler@ncbi.nlm.nih.gov

The National Center for Biotechnology Information (NCBI) provides access to more than 30 publicly available molecular biology resources, offering an effective discovery space through high levels of data integration among large-scale data repositories. The foundation for many services is GenBank®, a public repository of DNA sequences from more than 133,000 different organisms. GenBank is accessible through the Entrez retrieval system, which integrates data from the major DNA and protein sequence databases, along with resources for taxonomy, genome maps, sequence variation, gene expression, gene function and phenotypes, protein structure and domain information, and the biomedical literature via PubMed®. Computational tools allow scientists to analyze vast quantities of diverse data. The BLAST® sequence similarity programs are instrumental in identifying genes and genetic features. Other tools support mapping disease loci to the genome, identifying new genes, comparing genomes, and relating sequence data to model protein structures. A basic research program in computational molecular biology enhances the database and software tool development initiatives. Future plans include further data integration, enhanced genome annotation and protein classification, additional data types, and links to a wider range of resources.

Introduction

Established in 1988 as a national resource for organizing and delivering molecular biology information, the National Center for Biotechnology Information (NCBI) provides an information infrastructure for molecular biology research. The NCBI is a division of the National Library of Medicine (NLM®) within the National Institutes of Health (NIH) and the Department of Health and Human Services (DHHS). Since its inception, NCBI’s approach to building and providing access to bioinformatics resources has centered on data integration. The goal has been to achieve high levels of data integration among large-scale and varied data repositories to create an effective discovery space.

In its first 15 years, NCBI has grown from providing one database, the GenBank® repository for DNA and protein sequence information (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2003), and one analysis tool, the BLAST® program that compares one sequence against all others in a database to identify similar ones (Altschul, Gish, Miller, Myers, & Lipman, 1990; Altschul et al., 1997), to offering more than 30 publicly available database resources and search tools. Some are archival resources that are essentially repositories of data submitted to NCBI by the scientific community. Others are more highly curated resources produced by analysis and synthesis of data in the large archives. Many are closely tied to GenBank, as derivative databases representing specialized subsets, as sequence analysis tools, or as services that provide important links between nucleotide sequence data and other types of information important to molecular biology research.

In addition to molecular sequences, NCBI services cover a broad range of data types, including genome maps, phenotype and functional information for genes and proteins, common sequence variations in populations, quantitative information on levels of gene expression, a taxonomy resource for classifying organisms on the basis of sequence data, nomenclature resources and related finding tools, and the biomedical literature. The Entrez (Schuler, Epstein, Ohkawa, & Kans, 1996; Wheeler et al., 2003) retrieval system lays the foundation for text-based access to the diverse databases, offering a rich set of links among records within databases and across varied resources.
Computational access to the vast quantities of research data is provided by a suite of analysis tools developed by NCBI scientists. The BLAST programs for determining sequence similarity are used throughout NCBI services and are instrumental in many areas of computational biology, including identification of genes and genetic features. Other computational tools allow researchers to map disease loci to the genome, identify new genes and variants, compare whole genomes, and relate sequence data to model three-dimensional (3D) structures.

Bioinformatics is an integral component of public and private research throughout the world, and access to online resources serves as a virtual extension of the experimental laboratory. Nearly 100,000 BLAST searches are done each day on NCBI servers alone. The article describing the original BLAST algorithm, published in 1990 (Altschul et al., 1990), was one of the most highly cited papers of the decade by mid-1999 with nearly 9,000 citations—a powerful indicator of its central role in molecular biology research methodology (Russo & Bunk, 1999). Subsequent papers describing enhancements and customizations are also leaders of the pack relative to other more recently published papers.

In this article, we will provide an overview of the various types of biological data managed by NCBI, describe some of the diverse database resources it offers, describe key access and analysis tools, and highlight the role of bioinformatics research in building the computational resources.

All resources discussed herein are available from the NCBI Web site at http://www.ncbi.nlm.nih.gov. Additional information about the resources can be found in the NCBI Handbook, located on NCBI’s Web site in the Books database.

Diverse Biological Data—A Challenge for Integrated Services

The challenge in managing the wealth of molecular biology data available to biomedical researchers is in providing tools that facilitate the ability of scientists to make new connections between disparate data and broaden their understanding of biological relationships.

Sequence data is the basis for many of NCBI’s data analysis and retrieval services. Although data in this form are certainly human-readable as strings of letters, they are not readily human-understandable until compared with other sequence data. To obtain the complete genome sequence of an organism, individually sequenced segments of DNA are assembled in the correct linear order. Figuring out how to do that is a huge computational challenge that is aided by the use of genome maps. Genome maps contribute to the genome sequence assembly process at NCBI and elsewhere by pinpointing sequence landmarks and identifying biological features along a genome. They are important in directing the hunt for genes and other features of the genome that affect development of the organism. Insight into the function of a gene can be obtained by studying the three-dimensional structure of the protein it encodes. A gene’s function and influence on phenotype can be affected by variations in the sequence or in levels of expression within the cell.

The databases at NCBI are designed to facilitate biological investigation in addition to basic information retrieval. In that regard, they can be viewed as a hierarchical system that reflects the natural relationships existing between biological entities. For instance, nucleotide sequences are related to amino acid sequences via the biological process of translation. This relationship is reflected in the NCBI databases by a link between a gene and its implied protein product. Nucleotide sequences are found within cells at various levels of the cell’s information-processing system. Within NCBI databases, an indexed molecule-type field with values such as “DNA” for genomic DNA sequences, or “mRNA” for transcript sequences, mirrors two biological levels of information processing and allows researchers to focus on either of the two datasets in isolation, or to explore links between the sets.

In the case of protein sequences and 3D structures, biological relationships are also reflected in the database design. These relationships are reflected at NCBI by the linear sequence of amino acid letters in the protein sequence databases and in the 3D coordinate sets of protein structures in the structure data resources. Tools for detecting similarities in amino acid sequences or in amino acid proximity in 3D protein structures allow biologists to probe the databases seeking patterns at either level. Tools to integrate sequence and structural patterns allow biologists to investigate one of the most subtle and fundamental of biological relationships—the relationship between a protein sequence and the 3D structure that is the basis of its function.

Sequence Data Lays the Foundation

GenBank

At the core of NCBI’s services is the DNA sequence database called GenBank. GenBank contains not only human sequence data, such as that generated by the human genome project, but also DNA sequences from more than 133,000 other species. This allows for the very important cross-species comparative analyses that have always been important for biology, and remain at the core of molecular biology research. In addition to source DNA sequences, GenBank also contains the protein sequence translations that are specified by regions of the DNA that code for proteins. GenBank currently contains more than 26 million DNA sequences, representing more than 33 billion base pairs.

As a comprehensive public resource, GenBank depends on the participation of the scientific community and the continued support of journal editors in requiring that authors submit their data to a public repository as a condition of publication. Sequence data, with supporting bibliographic and biological annotation, is submitted directly to GenBank by individual scientists, genome centers conducting large-scale sequencing projects, and the U.S. Office of Patents and Trademarks. In addition, worldwide coverage is facilitated through collaborative data collection policies and nightly
data exchange with its international sequence database partners, the EMBL Data Library (Stoesser et al., 2003) and the DNA Data Bank of Japan (Tateno et al., 2002).

October 2002 marked the 20th anniversary of the creation of GenBank, which had grown from 680,338 base pairs in 1982 to 22 billion in 2002. The database continues to grow at an exponential rate, doubling approximately every 15 months. Access methods have changed over time as well. In the 1980s, access was primarily through local installations of the database with commercially supplied search and analysis software. In 1984 there were 120 magnetic tape subscribers and an average of 5 online users per day. Magnetic tape was replaced by CD-ROM in the 1990s, accompanied by a steady increase in online access as well. The surge in Internet access that accompanied widespread use of the World Wide Web, along with rapid database growth, led to discontinuation of CD-ROM distribution in the mid-1990s. Today, while there are still many local installations of GenBank in universities and private companies, the Internet is the predominant method of access, supporting more than 30,000 online users per day.

The content of a GenBank record is actually just text, so the overall structure of a GenBank record is very much like that of a bibliographic record from an abstract database. There are text fields for data elements such as the accession number, a descriptive "title" for the record (termed a Definition Line), the taxonomic classification of the organism represented, the names and affiliations of submitters or "authors," and journal citations for sequences that have been published. In place of the bibliographic record's abstract is the DNA sequence, which is a string of letters. In place of the index terms that highlight key concepts in a journal article, the GenBank record contains structured annotations that point to regions of biological significance within the sequence data. The biological annotations are constructed following a detailed set of guidelines developed jointly by NCBI and its international collaborating databases. When a DNA coding region is specified as part of the biological annotation, the corresponding protein sequence translation is also included in the annotation section of the record. The nonsequence components of a GenBank record are searchable as text fields by the Entrez retrieval system. The sequence data is accessed by the BLAST suite of sequence similarity search programs.

As with other scientific arenas, molecular biology is characterized by multiple laboratories conducting simultaneous research on the same problem. As an archival repository, GenBank accepts data submissions from all contributing scientists, without regard to controlling for data redundancy. Duplicate submissions of essentially the same sequence data can be useful for purposes of verification and quality control, and scientists often contribute unique information through the biological annotation that accompanies their submissions of sequence data. However, the data redundancy and scattering of pieces of biological annotation across many different records can also confound efforts to analyze and understand the data and to apply it for further research purposes. As a result, derivative sequence databases such as UniGene, UniSTS, and RefSeq have been developed to remove redundancy and consolidate information. UniGene and UniSTS offer nonredundant views of GenBank subsets and are described in the section covering Entrez. RefSeq, on the other hand, generates new sequence records as a result of the data curation, and is described below as a second resource for DNA sequence data.

**RefSeq—Standard Reference Sequences to Support Genome Annotation**

The Reference Sequence (RefSeq) initiative aims to develop a nonredundant source of reference sequences that can serve as sequence standards for purposes of computation and genome annotation. They provide a stable reference for gene characterization, mutation analysis, expression studies, and polymorphism discovery.

Relying heavily on computational analysis as well as expert review and synthesis of the published literature, RefSeq is a database of curated reference sequences for mRNA, genomic DNA, computationally derived transcripts, and protein sequences for human and more than 2,000 other organisms. New sequence records comprising the composite information from multiple GenBank records and other database sources are created for RefSeq and assigned their own set of accession numbers.

The most reliable of NCBI’s human gene models are produced from RefSeq transcript sequences aligned to the human genomic sequence and used as the basis of gene annotation for the human genome. These transcript-based gene assignments can then be supplemented by assignments based on the predictions of gene finding programs such as GenomeScan. GenomeScan predictions can be reinforced with Expressed Sequence Tag (EST) alignments and similarities between predicted gene products and proteins already in the databases. All the above processes make use of variants of the BLAST sequence similarity search programs and are based on finding similarities between DNA and protein sequences. As other complex genomes become available, RefSeq will continue to fulfill the requirement for a nonredundant database of reliable species-specific transcript sequences upon which to base gene models.

The RefSeq approach has also been applied to viral genomes and sequence variations.

In the case of viral sequences, the data redundancy situation is further complicated by the large number of strains, isolates and mutants, making it particularly important to compare available sequences and choose one full-length genomic sequence for each virus as the “reference sequence.” In the case of sequence variations, variations mapping to the same genomic location are assigned to a single RefSNP cluster.

The sequence data in GenBank and RefSeq is central to many aspects of molecular biology research and computational analysis. Additional database services provided by NCBI serve to relate the sequence data in these resources to the other types of data that NCBI supports.
Organizing and Accessing Diverse Resources

Integrated access to diverse data resources is the goal that drives development of NCBI’s services. This is accomplished through the organization of databases within the Entrez retrieval system, through links to related resources, and through computational tools that support discovery of biological relationships.

The ASN.1 data standard was adopted by NCBI as an efficient format for the encapsulation of biological data and its direct and computed relationships to other data. The standard was chosen with the knowledge that it must accommodate data types that emerge as new experimental techniques are devised. To date, the standard has proven robust and has allowed NCBI to expand its range of databases to cover everything from sequences and sequence alignments gene expression data and protein structures.

The Entrez System—A Foundation for Integrated Access

Entrez is a database retrieval system that provides integrated access to the varied molecular biology databases developed by NCBI. This includes DNA and protein sequence data, assembled genomes, genome maps, sequence variations, gene expression data, protein structures and domains, a sequence-based taxonomy, and biomedical literature. It is important that all these domains be interrelated because one ultimately wants to relate the primary sequence data to its place in the genome, to the structure and function of the proteins that are encoded, and to its relationship with other organisms.

What makes Entrez more powerful than many services is that most of its records are linked to other records. It has extensive crossreferencing within each database and across databases within the system. It also provides external links to related records in outside databases, and offers methods for outside data providers to link to Entrez records from their Web sites. Although the Web now facilitates extensive database linking, it was a particularly significant innovation for bioinformatics in the early 1990s when the Entrez system was developed as a standalone application delivered on CD-ROM. Previous molecular biology database services had lacked even the most basic of links between nucleotide and protein sequences.

Entrez was first developed with just three database components—nucleotide sequences, protein sequences, and the biomedical literature via molecular biology subsets of the National Library of Medicine’s MEDLINE® database. See Figure 1 for an illustration of the basic model, to which several additional resources have been added, all the while maintaining the information science principles of linking related records within and across databases in the system.

For the bibliographic data represented by PubMed® in Figure 1, the within-database relationships are defined in terms of text similarity between records. For the nucleotide and protein sequence databases, the within-database relationships are defined as sequences that share a certain degree of similarity. Structure relationships are defined by vector alignment of protein structures. Across databases, for example, nucleotide sequences are linked with corresponding protein sequence translations, and published sequences are linked to corresponding bibliographic records. Links to the published literature were important contributors to the early success of Entrez, and strong ties to information in journals as well as books continue to be a priority. Substantial informatics research and development were required to develop the precalculated sequence relationships and to create the bibliographic links between databases with differing citation formats.

Following the launch of Entrez in 1991, new database components were added year by year as additional data became available, enhancing the richness of the integrated discovery space. The databases included in the Entrez network in 2003 are listed in Table 1 and described below.

### TABLE 1. The Entrez databases.

<table>
<thead>
<tr>
<th>DNA and Protein Sequences</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide: DNA sequence database (primarily GenBank)</td>
<td>Taxonomy: Organisms with sequence data in GenBank</td>
</tr>
<tr>
<td>Protein: Protein sequence database</td>
<td>Sequence Variation</td>
</tr>
<tr>
<td>PopSet: Sequence alignments from population and phylogenetic studies</td>
<td>SNP: Single nucleotide polymorphisms</td>
</tr>
<tr>
<td>Genome: Complete genome assemblies</td>
<td>Gene Expression</td>
</tr>
<tr>
<td>UniGene: Gene-oriented clusters of transcript sequences</td>
<td>GEO: Gene expression hybridization array repository</td>
</tr>
<tr>
<td>UniSTS: Markers and mapping data</td>
<td>GEO Datasets: Curated gene expression and microarray datasets</td>
</tr>
<tr>
<td>Structure: Three-dimensional macromolecular structures</td>
<td>Phenotype and Function</td>
</tr>
<tr>
<td>3D Domains: Protein domains from Entrez Structure Domain</td>
<td>OMIM: Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>Domains: Conserved protein domains (CDD database)</td>
<td>Molecular Structure Databases</td>
</tr>
<tr>
<td>Bibliographic Databases</td>
<td>Structure: Three-dimensional macromolecular structures</td>
</tr>
<tr>
<td>PubMed: The biomedical journal literature</td>
<td>3D Domains: Protein domains from Entrez Structure</td>
</tr>
<tr>
<td>MeSH: Medical subject headings</td>
<td>Domains: Conserved protein domains (CDD database)</td>
</tr>
<tr>
<td>PubMed Central: Full-text journal articles</td>
<td>Bibliographic Databases</td>
</tr>
<tr>
<td>Books: Online textbooks and manuals</td>
<td></td>
</tr>
<tr>
<td>Journals: Journal titles represented in Entrez</td>
<td></td>
</tr>
</tbody>
</table>
The Entrez Databases

Nucleotide—DNA sequences from GenBank and RefSeq. The Nucleotide database of Entrez is comprised primarily of nucleotide sequence records from GenBank and RefSeq. Relationships within the Nucleotide database are defined in terms of sequence similarity and are calculated using NCBI’s BLAST family of sequence comparison algorithms. If a user locates a sequence of interest, other nucleotide sequences that meet NCBI’s predefined threshold of similarity can be displayed. Going across databases, nucleotide sequences are crossreferenced and linked to their corresponding protein sequence translations. Sequence records that include bibliographic references are linked to the corresponding abstracts in PubMed. Links to additional resources are created for sequences that are included or referenced in other Entrez components.

Protein—Protein sequences from multiple sources. The Protein database is comprised of protein sequence translations from GenBank and RefSeq, plus several protein sequence databases produced outside NCBI. These include SwissProt (Boeckmann et al., 2003), Protein Information Resource (PIR; Wu et al., 2003), Protein Research Foundation (PRF; Protein Research Foundation, 2004), and the Protein Data Bank (PDB) (Westbrook, Feng, Chen, Huanwang, & Berman, 2003).

As with nucleotides, the within-database relationships for protein sequences are defined in terms of sequence similarity and are calculated using NCBI’s BLAST family of sequence comparison algorithms. If a user locates a sequence of interest, other database sequences that meet NCBI’s predefined threshold of similarity can be displayed. Going across databases, protein sequences that are translated from nucleotide sequences in GenBank are linked to their source record in the Nucleotides database. Those that include bibliographic references are linked to the corresponding abstracts in PubMed. As with Nucleotides, links to additional resources are created for protein sequences that are included or referenced in other Entrez components, particularly the protein structure.

PopSets—Sequence alignments. The PopSet database contains aligned sequences submitted as a set resulting from a population, phylogenetic, or mutation study. These alignments can be used to analyze population variation or the evolutionary relatedness of organisms from different species. PopSet is derived from data submitted to GenBank and contains both nucleotide and protein sequence data. Links to source sequence records in Nucleotides are provided.

Genome—Assembled genome sequences. The Genome database houses assembled genomic data contributed by the scientific community for over 900 species whose sequencing and mapping is complete or in progress. This includes microbial genomes as well as higher eukaryotic genomes such as Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, and Arabidopsis thaliana. All three main domains of life—bacteria, archaea, and eukaryota—are represented, as well as many viruses and organelles.

The complete genome sequence is assembled from thousands of smaller sequence segments, all of which also are included in GenBank along with the associated biological annotations. A separate Genome database was created within Entrez to make the assembled genome available as a whole, to be searched and displayed as a contiguous data set complete with biological annotation of genetic and other biological features.

Data can be accessed using the Entrez text search syntax, but also by specifying a map position or a range of base pairs to get to a particular region of the genome.

Tools for visualizing genomes to get a global view of this very large and complex dataset have been developed at NCBI, and the need for such tools will grow as the database expands. A tool called the Map Viewer allows researchers to start with a graphical view of sets of parallel genetic, physical, and sequence maps for a genome and drill down quickly from a bird’s-eye view to the raw data for a region of interest. The Map Viewer for the human genome provides more than 35 maps and makes it a simple matter for researchers to download or analyze the DNA sequence of a gene or the region surrounding it. Another tool, called TaxPlot, depicts similarities in the protein complements of two genomes when compared to a third reference genome, allowing researchers to easily discern areas of specialization or differentiation between two organisms.

UniGene—Gene-based clusters of transcript sequences from GenBank. Designed to focus specifically on sequences of gene transcripts, UniGene was the first of several curated sequence resources developed by NCBI. Retrieving useful information from the GenBank database is often complicated by the high level of data redundancy, particularly due to submissions of large batches of ESTs, which are short cDNA sequences that represent segments of expressed genes and are therefore useful as gene markers. UniGene is a system for removing much of the redundancy found in GenBank, and operates by automatically partitioning mRNA and EST sequences into a non-redundant set of gene-oriented clusters.

This clustering has been performed for several organisms, including human, mouse, rat, zebrafish, cow, and frog in the animal kingdom, and barley, rice, thale cress, maize, and wheat from the plant kingdom. Each cluster is annotated with computed similarities of transcript translations to protein sequences in the databases and with mapping information from cluster members. Researchers seeking to link a novel EST to a gene in the genome can do so by first finding the EST in UniGene using BLAST, then checking for a known or predicted gene using the Map Viewer or UniGene cluster annotations.
The UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences—representing an enormous bioinformatics analysis in and of itself. Entrez links are provided to component sequences in Nucleotides, to genome maps in which UniGene clusters are cited in gene annotations, and to LocusLink, which, in turn, provides a gateway to a wide array of gene-related resources.

UniSTS—Sequence-based genome markers. Within the range of nucleotide sequence data in GenBank, Sequence Tagged Sites (STs) have importance and utility for finding genes and mapping genomes. Sequence Tagged Sites are short genomic sequences that are unique to a particular place in the genome and therefore serve as markers, or landmarks, for a particular genome location. While also a part of GenBank, these sequences have been broken out into a separate database and curated to remove the high level of redundancy that is particularly characteristic of this type of sequence data. The resulting database of unique STS sequences is called UniSTS, and is used in place of GenBank for BLAST and other computational analyses on STS data.

Links from UniSTS go to Nucleotides and any genome maps in which they are used as markers.

Taxonomy—Sequence-based taxonomy for classification of organisms. Sequence data is classified and can be queried using NCBI’s Taxonomy database, which was developed in the mid-1990s to create a consistent taxonomy for the growing number of species represented in GenBank. Standing at more than 133,000 species, new organisms are added at a rate of approximately 1400 per month.

The Taxonomy database contains the names of all organisms that are represented in the NCBI databases by at least one nucleotide or protein sequence. It does not follow a single taxonomic treatise, but rather attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of sequence submitters and outside taxonomy experts.

The Taxonomy database can be searched by common or scientific name. Links are provided to the source nucleotide sequence records. Links into Taxonomy come from any database or computational output report that includes an organism name.

SNP—Sequence variations. A key aspect of research in genetics is associating sequence variations with heritable phenotypes. The most common variations are single nucleotide polymorphisms (SNPs), which occur frequently throughout the genome and serve as excellent biological markers, which are segments of DNA with an identifiable physical location along a chromosome. Due to improvements in sequencing technology and the promise of SNPs for facilitating large-scale association genetics studies, there is great interest in SNP discovery and detection.

In collaboration with the National Human Genome Research Institute at the NIH, NCBI has established the dbSNP database to serve as a central repository for SNPs and other types of polymorphisms. The resource contains more than 9 million records submitted from academic and commercial sources, and continues to grow at a rapid pace. A RefSNP database addresses the issue of managing high rates of data redundancy, generating a curated set of unique clusters of SNPs mapping to a single location.

While not included as sequences in GenBank, SNPs are important for mapping landmarks along the genome sequence. DbSNP records are integrated with several NCBI information services, including GenBank, Genomes, the LocusLink nomenclature database, UniSTS, and PubMed. Users may query dbSNP directly through the Entrez retrieval system, or start a search in any part of the NCBI discovery space to construct a set of dbSNP records that satisfy their search conditions. Records are also integrated with external information resources for more detailed information that is beyond the scope of dbSNP content.

GEO and GEO datasets—Gene expression and microarray data. The Gene Expression Omnibus (GEO) is an archival gene expression and hybridization array data repository, as well as an online resource for browsing and retrieval of gene expression data from any organism or artificial source. Gene expression technology is passing through a phase of rapid technological evolution so that many different methods are used. As a result, GEO currently contains the results of over 7000 gene expression experiments using over 240 different protocols. It can be searched by elements such as author, protocol, organism, tissue, gene name, GenBank accession, and free text.

GEO Datasets contains curated sets of molecular abundance profiles originating from GEO, and provides an experiment-based view of GEO data. The Entrez system facilitates powerful searching on the GEO resources, linking the search results to internal and external resources where possible. For instance, using the nucleotide links from a GEO record for a gene expression analysis covering thousands of genes, a scientist can obtain any GenBank records that exist for the genes covered in the analysis.

OMIM—Phenotypes and function. The Online Mendelian Inheritance in Man (OMIM) database is a catalog of human genes and genetic disorders authored and edited at the Johns Hopkins University under the direction of Victor A. McKusick (McKusick, 1998). Primarily a text-based resource, it contains information on disease genes and phenotypes, including gene names, inheritance patterns, map locations, gene polymorphisms, and extensive state-of-the-art summaries based on published literature. This is a highly curated resource that depends on expert review and synthesis of the literature as well as extensive use of bioinformatics tools to make important data connections. Links are provided to related databases including GenBank, RefSeq, the protein sequence of the helicase and the UniGene database of gene transcripts.
**Structure—Three-dimensional macromolecular structures.** The Structure database contains three-dimensional structure information from the Molecular Modeling Database (MMDB) (Chen et al., 2003). The Molecular Modeling Database is a molecular structure database developed at NCBI using source data obtained from PDB. It contains more than 20,000 experimentally determined 3D macromolecular structures, including proteins and as well as some polynucleotides. NCBI processes the data derived from PDB by adding explicit chemical bond definitions and secondary structure assignments to the 3D coordinates. In addition, the taxonomic information in the source PDB record is reviewed, corrected if needed, and integrated with the NCBI taxonomy.

Molecular structures provide a wealth of information on biological function, on mechanisms linked to the function, and on the evolutionary history of and relationships between macromolecules. Wherever possible, the structure information is linked to data in other NCBI resources, including sequences, taxonomic classifications, and bibliographic citations.

The Structure database is considerably smaller than Entrez's protein or nucleotide databases, but a large fraction of all known protein sequences have homologs in this set of structures, and one may often learn more about a particular protein by examining 3D structures of its homologs. The goals in adding structure data to Entrez were to make this information easily accessible to biologists, and to facilitate comparative analysis involving 3D structure.

**3D Domains: Protein domains from Entrez Structure.** Proteins often contain several modules or domains, each with a distinct evolutionary origin and function. 3D Domains contains protein domains that have been identified by analyzing 3D structures of its homologs. The goals in adding structure data to Entrez were to make this information easily accessible to biologists, and to facilitate comparative analysis involving 3D structure.

**Domains: Conserved protein domains.** Although many functional modules of proteins have been identified at the level of 3D structure, many more are known only from their sequences. These domains can still be recognized through the alignment of many proteins sharing a similar stretch of sequence. Such domains are used during molecular evolution as building blocks that may be recombined in various arrangements to make proteins with different functions.

The Conserved Domains Database (CDD) database, referred to as Domains within Entrez, houses thousands of protein domain signatures known from sequence alignments, plus domains known from 3D structures. It serves as a directory of sequence and structure alignments, representing more than 11,000 conserved functional domains within proteins. It also serves as a classification resource that groups proteins based on the presence of predefined domains.

Data sources for the alignments in CDD include PFAM (Letunic et al., 2002) and SMART (Bateman et al., 2002), two databases outside of the NCBI that specialize in conserved domain definitions; NCBI’s Clusters of Orthologous Groups (COG) database (Tatusov et al., 2001); the NCBI Library of Ancient Domains (LOAD); and an internal database of curated alignments assembled by CDD staff.

Domains are linked to protein sequence records, similar domains and co-occurring domains. The co-occurrence link allows a researcher to see a protein containing a particular domain within the context of other associated domains. Because domains and functions are strongly correlated, this sort of domain analysis can provide valuable clues to the significance of a protein in the cell.

**PubMed and MeSH®—Bibliographic data and controlled vocabulary.** PubMed includes primarily the more than 12 million records in NLM’s MEDLINE database of bibliographic citations and abstracts from more than 4,600 biomedical journals. It also contains some additional content from selected life sciences journals not included in MEDLINE.

Relationships within the PubMed database are defined in terms of text similarity and are calculated using a weighted term co-occurrence algorithm developed by scientists at NCBI (NCBI Documentation, 2004). The text fields used in the algorithm include title, abstract, and MeSH terms from NLM’s controlled vocabulary. If a user locates a journal title or abstract of interest, other bibliographic records that meet NCBI’s predefined threshold of similarity can be displayed. Citations are displayed in rank order, from most to least relevant. Going across databases, bibliographic records are crossreferenced and linked to the corresponding nucleotide or protein sequence records in which they were cited. Links based on text similarity have also been computed for traveling from PubMed to the Books database.

MeSH (Medical Subject Headings) is NLM’s controlled vocabulary thesaurus used for indexing articles for MEDLINE. It can be searched by MeSH term, MeSH entry term, subheading, publication type, or text words within the MeSH scope notes that define use of the terms. Search results are displayed in relevance-ranked order if there is no exact match.

**PubMed Central.** PubMed Central™ (PMC) is a digital archive of peer-reviewed journals in the life sciences. More than 130 journals now deposit the full text of their articles in PMC. Participation in PMC requires a commitment to free access to full text, allowing for some delay after publication before the free access is available. For some journals, access to the full text is provided directly in PMC; others are accessed by a link to the journal’s own Web site, where full text is generally available free within 6 months to a year of publication.
publication. All articles that are free in PMC are identified as such in PubMed search results.

Books. In collaboration with book publishers, NCBI is adapting a growing number of scientific textbooks for the Web and linking them to PubMed records. When linking to Books from PubMed, the words and phrases in the abstract are highlighted with hypertext links. These phrases correspond to terms that are also found in the 24 online books available through Entrez. Links lead to a list of book sections in which the term is found. The Books database can also be searched directly using the Entrez search feature.

Journals. The Journals database contains the title, publisher information, and other descriptive and journal authority information for all journals represented within the Entrez database system.

LocusLink—A Central Finding Tool

The LocusLink (Pruitt & Maglott, 2001) database of official gene names and other gene identifiers was developed at NCBI, in conjunction with several international collaborators, as a central repository and finding tool for curated sequences and descriptive information about genetic loci. It is the resource that addresses a user’s request to “Show me everything you have about gene x.” As a nomenclature resource, it has a name authority and vocabulary control function. As a central linking point, it has an information directory function. For a given gene locus, LocusLink presents such information as official nomenclature recognized by appropriate nomenclature committees, alternative names and symbols that have been used, phenotypes, genome map location, EC numbers, associated UniGene clusters, the reference sequence for the gene if available, a list of DNA and protein sequence records associated with the locus, sequence homology, and protein domains from NCBI’s Conserved Domain Database that are detected in gene translation products. Links from the LocusLink report lead to the Gene Ontology developed by Promteome, Inc. (Beverly, MA).

LinkOut—External Links Widen the Horizon

External links are also an important aspect of NCBI’s data services and are designed to facilitate bioinformatics and biomedical research. LinkOut is the component of the Entrez system that provides for external links to a varied set of resources, including, for example, full-text journal databases, organism-specific genome databases, disease-specific resources, suppliers of experimental reagents, directories of scientists, and the online and print collections of participating university libraries.

To participate, external data providers register with NCBI, provide an Entrez search strategy that is used to tag those records that are to be linked to their resource, and give specifications for creating the appropriate URL for linking to records in their system. The search strategy can consist of an accession number, an author name, a gene name, a Boolean combination of text words, or any valid Entrez search statement. Through this mechanism, publishers, electronic journal aggregators, libraries, sequencing centers, biological databases, and other Web data providers can display links to their sites on records within Entrez. A user profile component called Cubby allows users to customize the set of links to be routinely displayed, but the full set is also always available upon request.

Linking Into Entrez

Links into the Entrez from external providers is another important aspect of the service. NCBI has always stressed open access to its resources and provides several methods for linking to Entrez records or generating URLs that will retrieve data automatically from within an external application. Information on how to link to Entrez is available through sidebar links on Entrez search pages.

Computational Access and Analysis

Integrated access based on text searching across varied resources is important, but not unique to bioinformatics. The need to compute on the data to carry out any type of molecular biology research is what drives the continued fast pace of development of bioinformatics resources. NCBI has developed several powerful analysis tools to allow users to do computational research and to visualize relationships that are important in guiding their research.

BLAST Sequence Similarity Searching—A BLAST for Every Purpose

Because nucleotide and protein sequences are usually understood by comparison with one another, tools to detect similarity or relationships between sequences of the same or of different types are invaluable to scientists. The most widely used tools of this type are the BLAST family of sequence-similarity search programs, developed at NCBI. BLAST programs are able to detect similarities between groups of nucleotide sequences, protein sequences, or between protein and nucleotide sequences as they are related by a genetic code. Using BLAST, a researcher can enter the Entrez system with a novel sequence, find a similar sequence in the database, and rapidly discover related data using Entrez’s pre-computed links. BLAST searches can be performed on NCBI’s Web site via a simple form into which a query sequence can be uploaded or pasted.

The BLAST program is also vital in the creation of many of the pre-computed analyses that interlink the Entrez databases. BLAST is used to link each nucleotide or protein sequence in Entrez with a number of similar sequences that are called Entrez sequence neighbors. Sequence neighbors allow a researcher to quickly assemble a set of related sequences for download or inspection. Often, a possible function for a sequence that is annotated as “unknown” is
suggested by the existence of a sequence neighbor to which a function has been assigned. Protein neighbors detected by BLAST are presented graphically with sets of color-coded alignments, using a service called BLASTLink, or BLink. The proteins included in the alignments shown can be grouped or filtered according to the organism of origin, the database of origin, known or suspected functional class, or relationship to a 3D structure. BLink allows a researcher to use pre-computed comparisons and the integration of multiple databases to quickly determine if an unknown protein is similar to a protein with a suspected function, 3D structure, or genomic context. BLink links are displayed for protein records in Entrez as well as within LocusLink reports.

Similarity searches through the 3 billion letters of large genomes, such as the human or mouse genomes, represent a recently acquired bioinformatics challenge. Researchers need to quickly map a sequence or large batches of sequences, to positions on their genome of origin or the genomes of related species. NCBI has developed a program called MegaBLAST that streamlines this mapping by seeking nearly exact sequence matches, rather than subtle similarities, and by combining a batch of sequences into a single large query to reduce the overhead. While up to 10 times faster than conventional BLAST, it is not effective for cross-species searches in which greater sequence variation is expected. For cross-species searches, a variant of MegaBLAST called discontiguous MegaBLAST has been developed that is still very fast, but also sensitive to similarities between nucleotide sequences coding for the same protein product.

Variants of BLAST tailored for sensitive protein similarity-searches are described under CD-Search below.

**e-PCR**

Electronic PCR (e-PCR) finds STSs within nucleotide sequences by comparing the query against UniSTS, nonredundant database of more than 134,000 human and 80,000 non-human STSs. Using an interactive web version of the program, researchers can quickly map a new DNA sequence to its location in a genome on the basis of STSs detected. NCBI uses e-PCR internally to map sequences to particular regions of chromosomes for assembly of large genomes from smaller pieces, as is done with the human genome.

**ProtEST**

ProtEST provides a pre-computed link between the huge EST database of transcript sequences coding for proteins and the protein sequences in the databases. The computations required to support this link are not feasible for the typical user, but access to the pre-computed results can allow researchers to quickly determine whether transcripts for a particular portion of a protein, or arising from a particular organism, exist in the databases without the need to run the BLAST search. Analogous to BLink, ProtEST shows pre-computed BLAST alignments between protein sequences from model organisms such as human, mouse, yeast and \textit{E. coli}, and the protein translations of UniGene nucleotide sequences. ProtEST reports are linked to UniGene reports.

**VAST**

The Vector-Alignment Search Tool (VAST) is used to align protein structures to support the inference of related protein function from similarities in protein shape. NCBI uses the VAST alignments as the source of the Entrez “structure neighbors.” Using structure neighbors, together with sequence neighbors, a researcher can begin with a protein sequence, use BLink to find a similar sequence linked to a 3D structure, then find other related structures and wind up with a biologically relevant linkage between a sequence and a structure to which it bears little or no sequence-similarity; all without performing a single calculation.

Since there are currently over 20,000 structures in the Molecular Modeling Database at NCBI, pairwise alignments between structures must be performed as rapidly as possible. To align two structures, VAST takes advantage of the fact that the unique 3D structure of a protein is determined, in large part, by the arrangement in space of two types of “secondary structural” elements—helices and strands—both of which are linear structures and can be represented as vectors. The approximation of a protein’s 3D structure with a set of vectors allows a rapid initial “vector” alignment of the two structures to be performed, bringing the structures into an approximate alignment. If a successful vector alignment is made, atoms of the peptide backbone, the alpha carbons, are then aligned at a higher level of precision using a computationally intensive process. As new structures enter the MMDB they are aligned with the structures already in the database.

**CD-Search**

Protein structures are characterized by conserved 3-D structural motifs that are usually associated with particular functions. To allow the sensitive detection of conserved domains in proteins, the domains are first characterized by the multiple sequence alignments of proteins known to contain each domain. As mentioned earlier in the section describing the CDD database, NCBI imports many of these alignments from two databases specializing in conserved domain definitions, PFam (Letunic, et al., 2002) and SMART (Bateman, et al., 2002). Additional domain alignments are obtained from NCBI-developed resources, including the COG database, the LOAD collection, and from an internal database curated by CDD staff. The alignments are then encapsulated in compact form within a Position Specific Score Matrix (PSSM) created by an NCBI program called Position Specific Iterated BLAST (PSI-BLAST). PSSM’s for thousands of conserved domains are combined into a library of conserved domain signatures for rapid searching at NCBI using a program called Reverse PSI-BLAST (RPS-BLAST). NCBI runs RPS-BLAST against all the proteins in Entrez to
produce links to the CDD database containing 11,000 conserved protein domains.

Interactive forms of both PSI-BLAST and RPS-BLAST, also known as CD-Search, are available on the NCBI website. Interactive PSI-BLAST allows researchers to create their own PSSM’s on the fly and use them to perform sensitive searches of the protein databases. Because the presence of a conserved domain in a protein provides a strong clue as to its function, the pre-computed links to the CDD and the interactive CD-Search program are valuable aids to protein annotation.

CDART

Many proteins contain multiple domains, or functional modules, that can be compared to words in a sentence. Knowledge of a single word may provide a clue to the meaning of the sentence but still leave a great deal of ambiguity. The knowledge of several words in a sentence, however, reduces the level of ambiguity considerably. In the same manner, knowledge of the combination of domains present in a protein and the ability to find other proteins in the database sharing all or portions of this “domain architecture” is very useful to researchers seeking to understand a protein’s function.

The Conserved Domain Architecture Retrieval Tool (CDART) allows researchers to view the domain architecture of a query protein and compare it with architectures found in other proteins in the databases via the pre-computed CDD searches described above.

Data Visualization—Map Viewer and Cn3D

Entrez provides a variety of computational links between disparate data types. However, it is often advantageous to merge objects from two or more Entrez databases into a single view. NCBI implements this sort of approach in several ways.

In the case of complete genomes where data for genetic, physical, and sequence-based mapping are all available, the Map Viewer can be used to display all the data as a set of parallel tracks, synchronizing unequal scales on the basis of shared landmarks, or genomic markers, such as STSs. The graphical displays created by the Map Viewer make it possible for researchers to visually correlate NCBI’s pre-computed tracks such as the alignment of transcripts to the genome, the locations of predicted transcript models, and SAGE tag mappings, with tracks showing the location of known physical and genetic markers, and disease loci.

Another case in which a simultaneous graphical display of two distinct data types is advantageous is in the case of a protein structure and its sequence. NCBI produces and distributes multi-platform versions of a program called Cn3D that displays MMDB protein and nucleic acid structures along with their sequences in such a way that sequence elements can be directly correlated with structural elements. In addition, sequence alignments can be constructed using BLAST or PSI-BLAST that combine the information of many similar sequences with a structural display to reveal how the patterns of sequence-conservation are reflected in the three-dimensional structure. Cn3D can also be used to test the feasibility of using an existing structure as a template for the modeling of another protein. Promising templates for proteins already in the databases have been identified by NCBI and are easily selected for viewing using Cn3D from BLink reports.

Building the Databases—Bioinformatics Research Behind the Resources

NCBI data collections are not only used as bioinformatics tools and information resources. They are also the result of much computational research and development, and much bioinformatics analysis. This is true for all of NCBI’s services, whether it is a seemingly straightforward archival data repository, an enhanced version of BLAST, a system for correlating genome maps, a method of relating molecular sequence and structure information to infer protein function, or an approach to comparative genomics.

As part of the National Institutes of Health intramural research program, NCBI supports a strong basic research program in computational molecular biology. Scientists studying a wide range of topics develop computational tools to conduct their own research and also work collaboratively with development teams to contribute to NCBI services. The strong synergy between the basic research and development teams greatly enhances the effectiveness of NCBI’s database and software tool development initiatives. Many of the public resources—UniGene, MapViewer, MMDB, and Cn3D to name just a few—have grown out of basic research programs in computational molecular biology that often required development of new bioinformatics tools and data resources before the research questions could be addressed.

Some examples of the research behind the resources will illustrate the role of NCBI’s computational research program in creating the tools that are, in turn, offered to the public to support bioinformatics research. They will also highlight some of the bioinformatics challenges that go into building and maintaining data resources of this size, magnitude, and variation.

Building GenBank—Data Links, Quality Assurance, Accession Numbers

The GenBank database itself has offered a number of bioinformatics challenges, beginning with data links. When NCBI assumed responsibility for producing GenBank in 1992, it added protein sequence translations directly into the GenBank record because of the obvious importance of the relationship. Creating bibliographic links from the
Sequence identifiers and accession numbers. The fact that sequence database records are used directly in computational analysis affects something as seemingly straightforward as a record’s unique identification number. As with most databases, each GenBank DNA sequence record is assigned an accession number, which is a stable and unique identifier for the GenBank entry as a whole, and does not change, even when the record is updated to reflect a change in sequence, annotation, or bibliographic information. However, it is important to be able to track changes to the sequence data. Sequence records are works in progress, results of an ongoing experimental process, and not static documents; they may be revised as knowledge improves. Therefore, NCBI assigns a second type of unique identifier, termed a “gi” number, to the sequence. When a change in a sequence occurs, a new gi number is assigned to the sequence component of the record.

Quality assurance. Quality assurance procedures utilize the BLAST algorithms to screen for cloning vector contamination and other unwanted or erroneous sequence data, but also to identify similar sequences in the database to aid in the annotation process.

Data submission and annotation software. A highly trained scientific staff reviews the data submissions and uses a suite of analysis tools and record maintenance software to create GenBank records according to a standard set of specifications developed in collaboration with its international database partners. Most GenBank data submissions are prepared using the BankIt input system on the Web or the stand-alone Sequin input system. While a relatively straightforward Web-based form, BankIt does rely on a number of informatics-based quality-checking routines as it processes the incoming data and informs the user of possible errors or omissions. Sequin, on the other hand, is a sophisticated bioinformatics tool that is designed to facilitate submission of large batches of sequences, including those from phylogenetic, population and mutation studies, and environmental samples. Used by GenBank annotation staff as well as the scientific community, it can serve as an interactive ASN.1 editor, check data for conformance with standard format requirements, incorporate alignment data, and automate some aspects of the data annotation process.

UniGene—Finding Unique Genes

In the early 1990s, soon after NCBI had assumed responsibility for production and distribution of GenBank, the technique for producing ESTs, came into prominent use. Because of the utility of ESTs as short tags for whole genes, most human gene discoveries today rely heavily on EST approaches. However, this was not always the case.

In 1992, following computational research that demonstrated the utility of ESTs for identifying genes in the sequence databases, NCBI began accepting sequence data of this type, formally adding it to GenBank a year later in 1993. Much of GenBank’s subsequent growth was due to the high volume and redundancy of EST sequences, which soon presented problems in data presentation and analysis. Recognizing the need to organize the EST data to enhance its utility for gene discovery, NCBI scientists developed the UniGene system to identify and cluster the unique genes represented in GenBank. The result of extensive computational research, the database organized matching sequences into clusters, each representing a unique human gene.

From the beginning, UniGene served as an important springboard for gene hunting. The first transcript map of expressed human genes, GeneMap 98, was produced by an international mapping consortium that relied on UniGene as a central resource for identifying novel, nonredundant genome mapping candidates (Deloukas et al., 1998). GeneMap 98 integrated STS mapping data, sequence data, and UniGene clustering data and provided a mapping framework upon which to mount the human sequence data that was coming out of the Human Genome Project.

Today, the millions of EST sequences in GenBank continue to be a rich source of data on gene expression. Finding an EST arising from a gene is one of the easiest ways to recognize the existence of a gene. If a complete genomic sequence is available for the organism in question, and ESTs of interest can be mapped by sequence-similarity to the genomic sequence, the location the gene on the genome can be determined. With gene-based clusters for over a dozen plants and animals, UniGene greatly simplifies this kind of analysis.

Visualization of Genome Mapping Data

In 1995, NCBI added the Genomes database to Entrez, complete with a method of simultaneously displaying genome maps constructed by different experimental techniques and based on different units of measurement. This achievement represented years of work towards the development of a framework of shared markers that allow the proper alignment of the maps for simultaneous display. The resulting public service made it possible to bring together work of experimental research groups in the mapping and sequencing communities who had operated independently and had little communication across sectors.

The specific maps displayed by the MapViewer vary according to the subject organism, but include cytogenetic maps, such as chromosomal ideograms showing banding patterns, sequence-based maps, such as those showing genes, and physical maps, such the radiation-hybrid maps used in placing genomic sequence within the human genome.
assembly. Maps showing predicted gene models, EST, and mRNA alignments are also available.

**MMDB and Cn3D—Building Protein Modeling Resources**

The 3D structure services at NCBI, centered around the foundation of the Molecular Modeling Database, grew out of years of research and development work needed to conduct basic research on protein function, the relationship between protein sequences and structures, and protein structure modeling and prediction.

The PDB database of protein crystal structures provides the primary data for the MMDB but lacks the explicit chemical bond information required for computational analysis. NCBI scientists developed methods of representing the needed information, relating it to sequence data, and building a robust imaging system that allowed for data editing as well as simultaneous viewing and manipulation of sequence and structure data. Later research led to a method for locating structural motifs or domains with a protein and developing a system for a functional classification of proteins.

The processing of the 3D atomic coordinates from PDB for inclusion in the MMDB involves several enhancements to the original record. Uniform secondary structural assignments are made to each record. These are necessary for computational analyses such as the 3D structural comparisons performed by the Vector Alignment Search Tool (VAST) to produce the Entrez structural neighbors. Explicit chemical bond-graphs are added to ensure that chemical bonds do not have to be inferred by display software, such as NCBI's Cn3D molecular viewer, on the basis of inter-atomic distances. Structural folding domains are detected on the basis of geometric measures such as the “compactness” and indicated in the record. These domains form the basis of the Entrez 3D-Domains database. The taxonomic assignments on PDB records are also checked and corrected if required to link the structure properly to NCBI's Taxonomy database.

To allow researchers to make the best use of the enhanced MMDB structural records, NCBI developed a 3D viewer called Cn3D that was capable of displaying all the information computed for the structure in the context of a linked display of the protein sequence. Various coloring and rendering schemes allow for the display of the secondary structural assignments as well as the domain assignments made during the processing of the MMDB record. The linkage between the display of the structure and the display of the protein sequence gives a user a simultaneous view of both a 3D domain and the linear sequence that comprises it.

Because the number of protein sequences far exceeds the number of known protein structures, additional tools were developed at NCBI to link patterns seen at the sequence level to the 3D structures responsible for protein function. One such tool is PSI-BLAST, itself a product of basic research at NCBI, which can detect the subtle sequence similarities which often indicate a common 3D structure. An outgrowth of PSI-BLAST of great utility to users is the CD-Search in which a collection of profiles for known conserved protein domains can be scanned with great rapidity for a match to a protein sequence query.

With the machinery in place for the enhancement and standardization of the primary data from the PDB, the MMDB has grown rapidly and become tightly integrated with other NCBI resources such as the sequence and taxonomy databases. The public services into which this research program was translated include MMDB, Cn3D, VAST, CDART, and CD-Search.

**Future Directions**

Standing items in NCBI's development plans include further data integration and incorporation of new data types as they become available from the scientific community. Current expansion efforts include linking the sequence variation and expression data to 3D structures and human gene maps. To meet the challenges of analyzing the data generated by the Human Genome Project and other whole genome initiatives, continued emphasis will be placed on initiatives related to information organization and characterization, such as genome annotation, protein classification, and development of curated gene-based databases. Projects to expand the literature-based services are also a high priority, particularly to include a wider range of textbook and increase links to the full-text journal literature.

Prior to the mid 1990s, GenBank consisted primarily of nucleotide sequence data on the scale of individual genes or small genomic regions. Over the past 15 years, NCBI has consistently produced resources to meet the changing scientific needs for managing individual sequences and genes within large-scale data repositories. Many of these are based on pre-computed analyses of large data sets used to build end-user services that facilitate use of the data for scientific discovery. Examples include UniGene, RefSeq, and BLink. Recent research initiatives have focused on developing the databases and tools needed for data management on a genome scale. Today, more than 130 complete microbial genomes and the genomes of more than 10 higher eukaryotes, including human, have been deposited in GenBank. One can envision a future for genome-scale services in which analogous pre-computed resources and relationships for whole genomes are as rich as those that have been created for data in GenBank.

**References**


