

The Merck Gene Index browser: an extensible data integration system for gene finding, gene characterization and EST data mining

Barbara A. Eckman², Jeffrey S. Aaronson²,
Joseph A. Borkowski, Wendy J. Bailey, Keith O. Elliston³,
Alan R. Williamson¹ and Richard A. Blevins

Department of Bioinformatics, Merck Research Laboratories, West Point, PA and Rahway, NJ, USA and ¹Department of Immunology, Merck Research Laboratories, Rahway, NJ, USA

Received on April 28, 1997; accepted on June 3, 1997

Abstract

Motivation: To make effective use of the vast amounts of expressed sequence tag (EST) sequence data generated by the Merck-sponsored EST project and other similar efforts, sequences must be organized into gene classes, and scientists must be able to 'mine' the gene class data in the context of related genomic data.

Results: This paper presents the Merck Gene Index browser, an easily extensible, World Wide Web-based system for mining the Merck Gene Index (MGI) and related genomic data. The MGI is a non-redundant set of clones and sequences, each representing a distinct gene, constructed from all high-quality 3' EST sequences generated by the Merck-sponsored EST project. The MGI browser integrates data from a variety of sources and storage formats, both local and remote, using an eclectic integration strategy, including a federation of relational databases, a local data warehouse and simple hypertext links. Data currently integrated include: LENS cDNA clone and EST data, dbEST protein and non-EST nucleic acid similarity data, WashU sequence chromatograms, Entrez sequence and Medline entries, and UniGene gene clusters. Flatfile sequence data are accessed using the Bioapps server, an internally developed client-server system that supports generic sequence analysis applications. Browser data are retrieved and formatted by means of the Bioinformatics Data Integration Toolkit (B-DIT), a new suite of Perl routines.

Availability: Software is available on request from the authors.

Contact: barbara_eckman@sbphrd.com

Introduction

The explosion of genomics data

Biological research is generating data at an explosive rate. Nucleotide sequence databases alone are growing at a rate of >210 million base pairs (bp)/year, and it has been estimated that if the present rate of growth continues, by the end of the millennium the sequence databases will have grown to 4 billion bp (Benton, 1996). In 1994, a public effort to generate both 5' and 3' expressed sequence tag (EST) sequences from the majority of human genes was initiated (Williamson *et al.*, 1995). This collaboration, consisting of Merck & Co., Inc., the Integrated Molecular Analysis of Genomes and their Expression (IMAGE) Consortium (Lennon *et al.*, 1996), the Genome Sequencing Center at the Washington University of St Louis (WashU) School of Medicine, the National Center for Biotechnology Information (NCBI), and the Computational Biology and Informatics Laboratory (CBIL) at the University of Pennsylvania, has concentrated on the characterization of normalized libraries, and has produced >430 000 EST sequences from >250 000 IMAGE cDNA clones derived from 45 distinct libraries. These sequences have been placed in the public EST database dbEST (Boguski *et al.*, 1993) as well as GenBank (Benson *et al.*, 1994), and comprise >75% of human EST sequences in dbEST.

The Merck-sponsored EST project

The ultimate goal of the Merck-sponsored EST project is to produce a gene index to the human genome [the Merck Gene Index (MGI)], a non-redundant set of clones and sequences, each representing a distinct gene. Such an index can be a critical resource in studies of gene finding, gene characterization and gene expression. The project was designed specifically to facilitate the creation of the index. All sequenced cDNAs are oligo-dt primed and directionally cloned. Consequently, 3' ESTs are anchored at the poly(A) tail and repre-

²Present address: Department of Bioinformatics, SmithKline Beecham Pharmaceuticals, King of Prussia, PA, USA

³Present address: Gene Logic, Inc., Columbia, MD, USA

sent comparable 3' untranslated region (UTR) sequences. The 3' UTR is the most diverse region of the transcript (Ko *et al.*, 1994) and thus may serve as a unique identifier of the genes tagged in the project. When a 3' EST is assigned to an index class, all ESTs on its clone, both 5' and 3', are also assigned to the class. In this way, the cDNA clones are clustered into gene index classes by sequence comparison of their 3' ESTs. Where a gene's splice forms have alternative 3' ends, there will be a cluster corresponding to each unique 3' end. In such cases, a cluster will represent an individual transcript rather than a gene. The MGI data set is coordinated with the associated cDNA clones, as individual clones, sets of clones, and high-density gridded filters available through the current suppliers of the complete clone set.

The Merck Gene Index

As of January 1997, Version 1 of the MGI has identified ~45 000 distinct genes represented in the data produced by the Merck-sponsored EST project. A similar effort considering additional public domain sequence data estimates roughly 50 000 distinct genes through analysis of 3' sequence (Boguski and Schuler, 1995). Another effort, more limited in scope and focused on differential tissue expression, has identified 12 000 distinct genes using a similar 3' sequence clustering approach (Matsubara, 1995). Other efforts have adopted different EST clustering methodologies: the 'shotgun sequence assembly' approach (Adams *et al.*, 1995) and clustering using both 3' and 5' sequence (Houlgatte *et al.*, 1995). Estimates of coverage of known genes by ESTs are in the 60–70% range (Adams *et al.*, 1995; Aaronson *et al.*, 1996; Hillier *et al.*, 1996; Schuler *et al.*, 1996), suggesting that the EST dataset represents a large fraction of the genes in the human genome.

The MGI browser

The best dataset is of limited utility in biological research if it cannot be accessed and 'mined' by research scientists from their desktops with an easy-to-use interface. Further, a dataset that is isolated from related datasets in the genomics community is of limited value. The MGI dataset must be integrated with sequence annotation and other genomic data distributed at many sites, both within Merck and outside. Further, since biological research often proceeds at high speed, scientists need access to the most up-to-date data available, and it must be relatively easy to integrate new types of annotation from new data sources as they arise. Our engineering challenge was to design a robust, easily extensible, user-friendly system to facilitate mining of the MGI dataset, integrating data from a variety of sources in a variety of storage formats, both inside and outside Merck. Our solution was to build a World Wide Web-based browser with a point-and-click interface for portability and ease of use. We utilized a modular system architecture to ensure ease of ex-

tensibility, in terms of both adding data from new sources and also upgrading our query capabilities as the browser evolved.

Data integration

We chose an eclectic data integration strategy, taking advantage of three different methods. First, for flatfile protein and nucleotide sequence data, we used the data warehouse or 'instantiated' approach (Davidson *et al.*, 1995). Local copies of the data are maintained and updated nightly, to ensure fast access at runtime, and to allow us to add value to the datasets by eliminating redundancy and grouping sequences into useful subsets. Second, with respect to data already stored in public-access relational databases (IMAGE clone data and dbEST sequence similarity data), we chose a federated architecture (Heimbigner and McLeod, 1985; Alonso *et al.*, 1987; Sheth and Larson, 1990). Unlike flatfile sequence data, data from these remote databases are not mirrored internally at Merck; rather, they are accessed 'on the fly' whenever the index is queried, thereby ensuring that only the most current view of the data is provided to research scientists. This approach was chosen because it is cost effective, scaleable and flexible: if remote datasets are not mirrored locally, the cost of maintaining each remote database link is significantly reduced, many database links may be simultaneously maintained, and replacing one data source with another is simply a matter of plugging in a different query module. Third, browsing access to data from additional related sources is provided using WWW hypertext links.

System and methods

Hardware and software

All local computation is performed on a 4 × 200 MHz processor Silicon Graphics Challenge L. This machine functions as a database, WWW, file, and compute server for both the Bioinformatics Department research effort and >400 users. The following software is used: BLAST (Altschul *et al.*, 1990), FastA (Pearson, 1991), the Apache v1.1.3 HTTP Server (Apache, 1995–1997), Sybase SQL server version 11.02 (Sybase, 1996) and Perl (Wall *et al.*, 1996) with the Syberl (Pepler, 1996) and CGI (Stein, 1997) modules. Java (Flanagan, 1996) was not used in this version of the browser, due to security problems; we plan to convert the browser to Java as soon as Merck network security has cleared it for use. The plughole system of Trusted Information Systems' (TIS) Internet Firewall Toolkit (TIS, 1996, 1997) was used to ensure network security during accesses of remote relational databases through the Merck firewall.

Data sources

Local flatfile sequence data. Flatfile sequence data comprise Nbase (Blevins *et al.*, 1995), a local data warehouse consisting

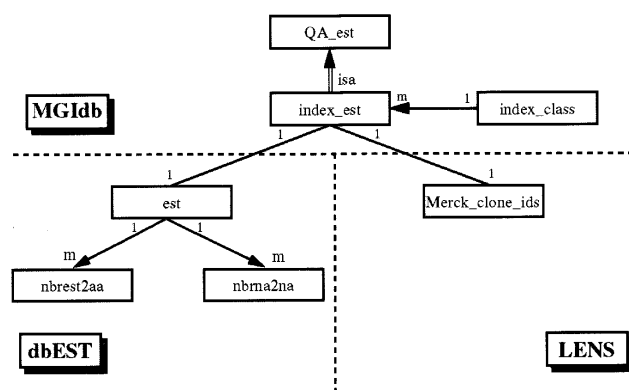


Fig. 1. The schema of the relational database federation accessed by the MGI browser. Boxes correspond to tables/views in the underlying databases. Relationships between tables are represented by arcs labeled with the cardinality of the relationship: a 1-m arc between the *index_class* and *index_est* tables indicates that for each row in *index_class* there are many associated rows in *index_est*. An *isa* arc indicates a subclass relationship: ESTs that have been indexed (*index_est* table) form a subclass of the class of ESTs subjected to the quality analysis (*QA_est* table).

of comprehensive non-redundant protein and nucleotide sequence databases built nightly from the latest release of PIR (Release 51.00) (George *et al.*, 1986, 1994), Swiss-Protein (Release 34) (Bairoch and Boeckmann, 1994), NRL-3D (Release 20.00) (Pattabiraman *et al.*, 1990), GenPept (Release 100.0) and GenBank (Release 100.0) (Benson *et al.*, 1994). Nbase is accessed using the Bioapps server, an internally developed client-server system that supports generic computational biology applications (e.g. fetching sequence entries, BLAST and FastA sequence comparisons, multiple alignments) in addition to non-biological applications such as newsreaders. The Bioapps server runs on multiple platforms and is callable from C (Kernighan and Ritchie, 1988) and Perl.

Local relational database. The data relating directly to index classes and the index EST quality analysis are stored in a local Sybase relational database, MGIdb. This database is very simple, because our design relies on the two remote relational databases for most of our ancillary data. Its schema is shown in the MGIdb quadrant of the federated database schema in Figure 1. The *QA_est* table contains a row for each of the Merck-WashU ESTs, with the EST's accession number and an integer code reflecting the results of the index quality analysis. The *index_est* table contains a row for each index-quality EST, and specifies attributes of the EST, including the starting and ending position of index-quality sequence, the length of index-quality sequence, and the FastA bestscore (perfect match score) of the sequence. The *index_class* table contains a row for each index class. Currently, the attributes of a class are its assigned integer identifier and the GenBank accession number of its representative EST sequence.

Remote relational databases. (i) LENS. The Linking ESTs and their associated Name Space database is a Sybase data warehouse maintained by CBIL at the University of Pennsylvania and updated nightly. LENS provides a mapping between EST/cDNA clone identifiers from dbEST, GDB, GenBank, WashU and IMAGE databases, as well as additional information on cDNA clones. LENS data items accessed by the MGI browser include: a clone's IMAGE id, array address, insert size and GDB id; the name and GDB id of the library from which the clone was derived; and the GenBank accession number, WashU id, and orientation (3'/5') of its EST sequence reads. While building the LENS data warehouse, CBIL adds value by monitoring the integrity of the data emanating from the IMAGE project and identifying errors and inconsistencies.

(ii) dbEST. The National Center for Biotechnology Information (NCBI) provides a Sybase database that stores the results of pairwise BLAST comparisons between all EST sequences and all known protein and nucleic acid sequences, updated on a nightly basis. dbEST data items accessed by the MGI browser include: an EST's GenBank accession number, the GenBank definition line of the matching nucleic acid/protein sequence, and the *P* value of the sequence comparison, indicating the significance of the match.

Distributed relational access. The core of the MGI browser's data store is logically a single relational database, whose schema is given in Figure 1. The dbEST and LENS quadrants reflect only the subset of their respective databases which is accessed by the MGI browser. The *Merck_clone_ids* table is an unnormalized view in the LENS database with one row for each EST. The dbEST schema has been slightly simplified; two linking tables have been omitted for the sake of clarity. The *nbrest2aa* table contains the results of BLAST comparisons between ESTs and amino acid sequence; the *nbrna2na* table contains the results of pairwise comparisons of nucleic acid sequences, including ESTs. The local MGI database was designed specifically to minimize redundancy with the other relational databases: all clone information is retrieved from LENS and all sequence similarity data are retrieved from dbEST. The ESTs' GenBank accession numbers function as links between the three databases.

Currently, the three federated databases are accessed using separate Syperl query modules. Distributed joins are performed using a 'semi-join' strategy (Ozsu and Valduriez, 1991) in which relevant tuples are retrieved from one database and the relevant identifiers passed into the query module for the next database. This results in excellent performance: in our current hardware configuration, it takes only a few seconds for all data on an index class from the three locations to be retrieved, integrated, formatted and displayed in response to an interactive user request. On a dedicated server, we would expect even better performance.

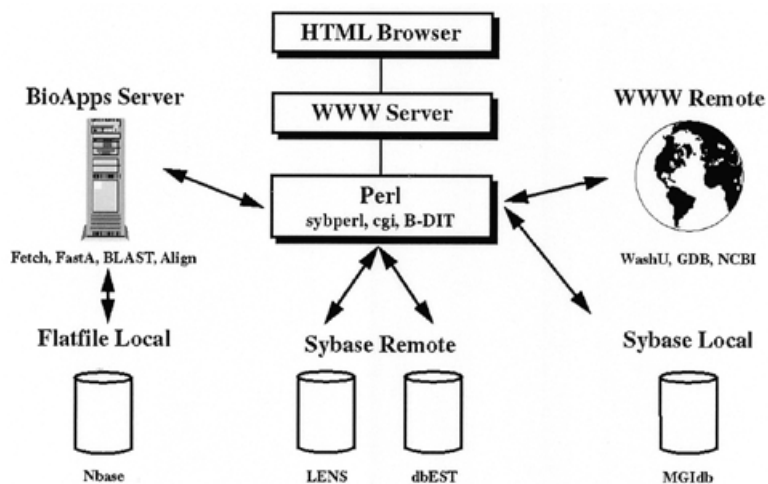


Fig. 2. In the current MGI browser system architecture, Perl scripts are used to integrate data from local and remote relational databases and the local Bioapps server (flatfile sequence and results of sequence analysis), and to insert hypertext links to other remote WWW servers. B-DIT modules format the results for viewing in the browser.

Data sources linked through hypertext. (i) WashU. Washington University of St Louis makes the original ABI sequence chromatogram trace files available for all ESTs sequenced in the Merck–WashU project. A hypertext link to the trace for each EST sequence is provided in the MGI browser.

(ii) GDB. The Genome Data Base at the Johns Hopkins University School of Medicine in Baltimore provides information on the biological reagents (clones and libraries) used in the EST project, including map locations. Hypertext links from the browser allow several entry points into the GDB WWW server, enabling direct browsing of the GDB database.

(iii) UniGene. The NCBI provides the UniGene web browser, designed to facilitate high-throughput physical mapping of ESTs. ESTs are grouped into classes, each representing a single gene. Several laboratories then make sequence-tagged-sites (STSs) from the UniGene classes and map them. The resulting map locations are reported for UniGene classes in their browser.

(iv) Entrez. The NCBI's Entrez Web tool is used to access Medline entries and entries from the underlying Entrez sequence databases through hypertext links from Nbase sequence entries.

System architecture

The MGI browser is built on the system architecture shown in Figure 2. Relational and non-relational data sources are integrated using Sybperl CGI scripts. Data from either source are formatted by the Bioinformatics Data Integration Toolkit (B-DIT), a library of generic Sybperl routines. Data are retrieved by a relational query or a call to the Bioapps server and written to STDOUT. The resulting data stream is piped through a series of B-DIT formatting modules, each of

which is designed to receive input on STDIN and send output to STDOUT. Examples of B-DIT modules are: *heat*, which converts identifiers embedded in the data stream into hypertext hotlinks that link to or perform a query in the relevant database; *tabularize*, which reformats a relational table as an HTML table with standard features such as caption, description, and column headers and widths; *nest*, which takes as input a relational table which is the result of a one-to-many join and outputs a relational table with a single row for each key value; and *merckify*, which adds the Merck logo to the beginning of the data stream and a standard Bioinformatics Department footer to the end. B-DIT modules also facilitate remote RDBMS log-ins and error handling.

Algorithm

The MGI Version 1 is constructed iteratively from all index-quality 3' ESTs generated by the Merck–WashU EST project. The index quality screening, fully described in (Aaronson *et al.*, 1996) and similar to the quality screening utilized in the BodyMap project (Okubo *et al.*, 1992), is designed to ensure the integrity of the clustering results, which may be compromised by poor-quality or low-complexity sequence. Poor-quality portions on the trailing end of sequence reads and any untrimmed poly(A) tails [appearing as leading poly(T) runs within 3' ESTs] are trimmed. Trimmed sequences must be at least 100 bp in length in order to pass screening, resulting in the removal of ~11% of ESTs from consideration. In addition, ESTs containing repetitive sequences (~10%) are removed during the screening process.

Incremental runs of the indexing algorithm are performed nightly on any new EST data that appear in GenBank

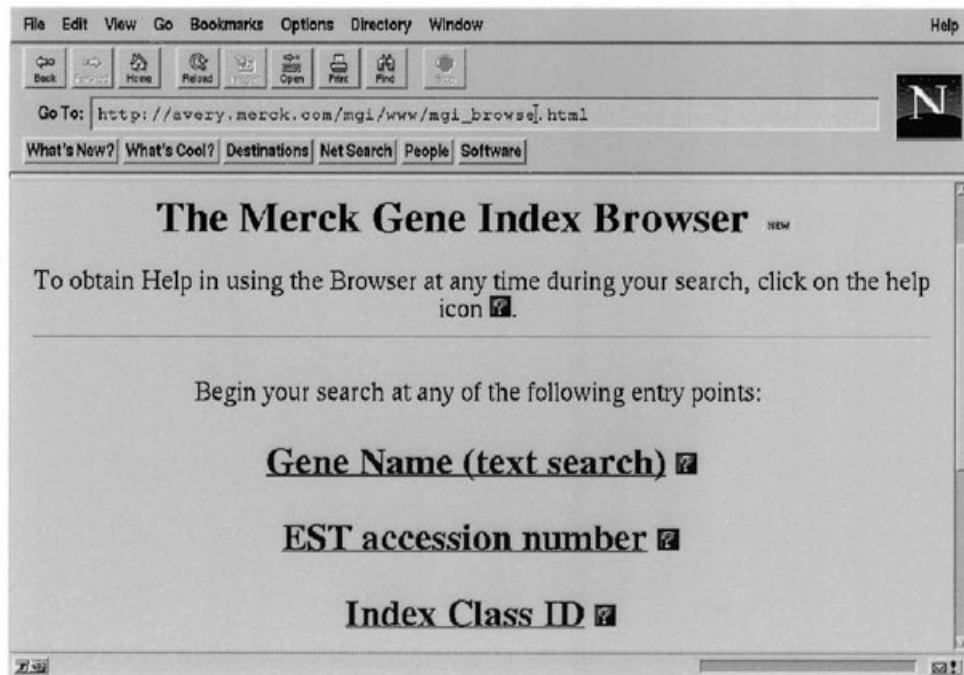


Fig. 3. The top-level browser menu, illustrating the three modes of access currently provided to the MGI data.

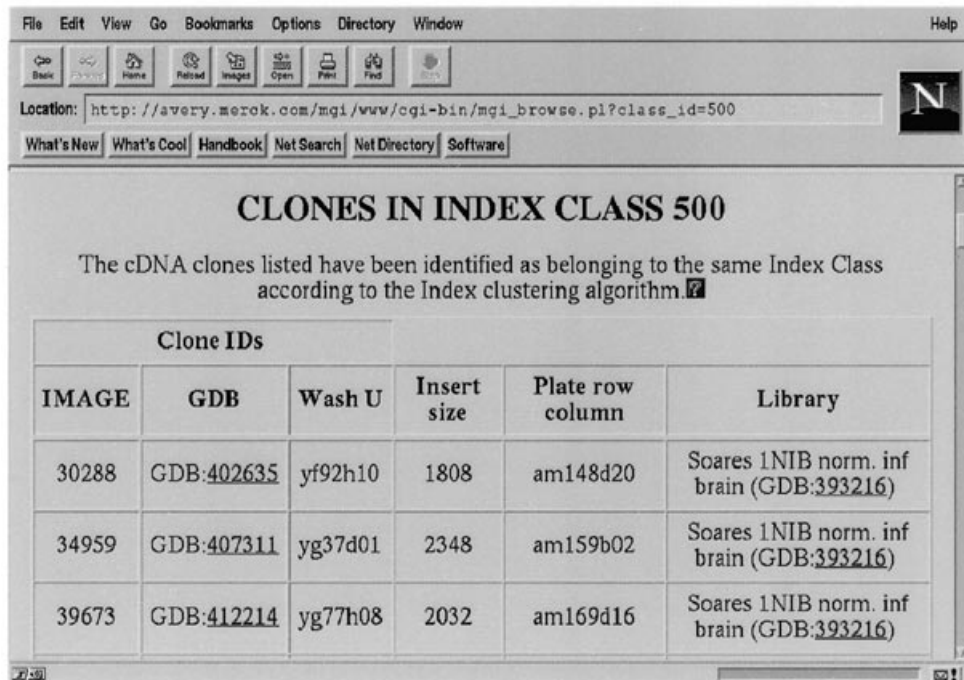


Fig. 4. A sample clone data module from the MGI class report.

updates, ensuring that the index remains up to date. Each new index-quality 3' EST is compared in turn against the index. If the EST is equivalent to an index entry, then the underlying

cDNA clones of the new EST and the index entry EST are assumed to be derived from the same gene. The EST is then placed into the index class represented by that entry. If the

ESTS IN INDEX CLASS 500

The EST's listed encompass all successful sequence runs from each end of the cDNA clones belonging to this Index Class.

Clone ID	3' End of Clone		5' End of Clone	
	GB Accession	Trace File	GB Accession	Trace File
30288	GB:R42549	WU:yf92h10.s1	GB:R14779	WU:yf92h10.r1
34959	GB:R44423	WU:yg37d01.s1	GB:R19635	WU:yg37d01.r1
39673	GB:R51648	WU:yg77h08.s1	GB:R51725	WU:yg77h08.r1
44121	GB:H05303	WU:yl80e03.s1	GB:H05353	WU:yl80e03.r1
52716	GB:H29245	WU:ym59f11.s1	GB:H29244	WU:ym59f11.r1

Retrieve sequences in format for saving to a file:

Fig. 5. A sample EST data module from the MGI class report.

EST is not equivalent to any existing index entry, we assume that the gene represented by the underlying cDNA clone is not represented by any cDNA clone already in the index. A new index class is then created with the EST as its representative sequence. Single-stranded comparisons are performed using the FastA sequence similarity program. Sequences are judged to be equivalent if their FastA opt score is at least 40% of their perfect match score, after normalizing for sequence length. This is a conservative method, designed to minimize the rate of false positives and to cope with the error-prone nature of EST sequence data. Upon completion of an incremental run, results are loaded into the local MGI relational database by means of the Sybase bcp utility. The indexing algorithm has been well optimized, and currently classifies 1000 new ESTs per hour.

Implementation

There are at present three main ways to search the integrated databases that the browser accesses: by EST accession number, text search of EST definition lines and index class id (Figure 3). These modes of access will be discussed in greater detail later in the paper. We plan to add new modes of access as the browser project develops.

The index class report

An index class report is composed of self-contained modules, each reporting on a different aspect of the class. Currently, the modules are: Clone Data, EST Data, Protein Similarity Data and Non-EST Nucleic Acid Similarity Data. Each module has a help icon, which when clicked displays the relevant section of the on-line browser manual. In the browser, the modules form a single page of output, but they will be displayed here in separate figures for the sake of clarity.

Clone data module (Figure 4). All data in the clone and EST data modules are retrieved from the LENS database. There is one row in the table for each cDNA clone in the index class. Hotlinks to GDB are provided by means of the clone and library GDB ids. The library indicates the tissue type used as the source of the clone. The address of the clone in the IMAGE array is also provided, to enable access to the actual clone for further laboratory experimentation.

EST data module (Figure 5). The EST data module is also organized by clone. The Clone ID is a cross-reference to the IMAGE ID in the clone data module. ESTs from the two ends of the clone (3' and 5', respectively) are grouped together in each row. Hotlinks are provided to the local sequence databases (using the Bioapps server) and the sequence chromatograms at

Location: http://avery.merck.com/mgi/www/cgi-bin/mgi_browse.pl?class_id=500

What's New | What's Cool | Handbook | Net Search | Net Directory | Software

PROTEIN SIMILARITIES TO ESTS IN INDEX CLASS 500

The EST's in this Index Class were compared using BLAST against a comprehensive set of protein sequences currently in the public domain. The (up to) top 50 hits between proteins and EST's in this Index Class are displayed.

P-value	EST	Protein Description
1.56e-70	GB:H29244	gil1139548 (D64009) seizure-related gene product 6 type 2 precursor [Mus musculus]
1.56e-70	GB:H29244	gil1585810 prfl2202175A SEZ-6 gene:ISOTYPE-2 [Mus musculus]
4.28e-53	GB:R51725	gil1095324 prfl2108345A seizure-related protein SEZ-6 [Mus musculus]
4.28e-53	GB:R51725	gil693910 (D29763) seizure-related gene product 6 precursor [Mus musculus]

Fig. 6. A sample protein similarity data module from the MGI class report.

WashU. A single mouse click will enable the user to save all the EST sequences in a file for further analysis, e.g. assembling with Sequencher (GeneCodes, 1995). A choice of three common sequence file formats is provided.

Protein similarity data module (Figure 6). Protein similarity data for each EST in the index class are retrieved from NCBI's nightly updated dataset of pairwise BLAST comparisons between ESTs and their comprehensive protein database. Up to 50 hits are displayed in order of significance (*P* value). Protein similarity data enable the scientist to characterize the index class by its similarity to known, well-characterized proteins. An index class can be related to known genes, or a putative placement of a class within a gene family can be suggested, based on the protein hits displayed here. Hypertext links are provided to the Nbase EST and protein sequence database entries.

Non-EST nucleic acid similarity data module (Figure 7). Nucleic acid similarity data are retrieved from NCBI's nightly updated dataset of pairwise BLAST comparisons between sequences in their comprehensive nucleotide database. Up to 20 hits are displayed in order of significance (*P* value); only non-EST hits are displayed. Nucleic acid similarity data may enable the investigator to recognize whether a class corresponds to a known or novel gene. Hypertext links are provided to the Nbase sequence database entries.

Modes of access to the MGI data

EST search. The EST search is probably the most common mode of access. It assumes that an EST or set of ESTs has already been identified by preliminary gene-finding efforts, e.g. a BLAST or FastA search of Nbase using the Bioapps server. Future releases of the MGI browser will incorporate sequence similarity searches as a direct mode of access available from the top-level search menu.

The results of the search are presented as two reports: classes represented by the ESTs of interest and ESTs that have no index class assignment. The search may be unable to identify classes for ESTs from some cDNA clones, due to the absence of a suitable 3' EST to serve as an identifier for a gene. The MGI browser is designed to avoid 'dead ends', by providing two methods to investigate further ESTs that have not been assigned to an index class (Figure 8). The first method invokes the Bioapps server to perform a BLAST search of the EST against the full set of Merck-WashU ESTs. Examining the BLAST output may identify ESTs with index classes that are effectively identical to the unclassified EST. The BLAST output enables easy browsing of alignments, facilitating the evaluation of the biological significance of the results. The second method of investigating unclassified ESTs executes a search of the UniGene web browser. Since UniGene's sequence quality screening method and clustering

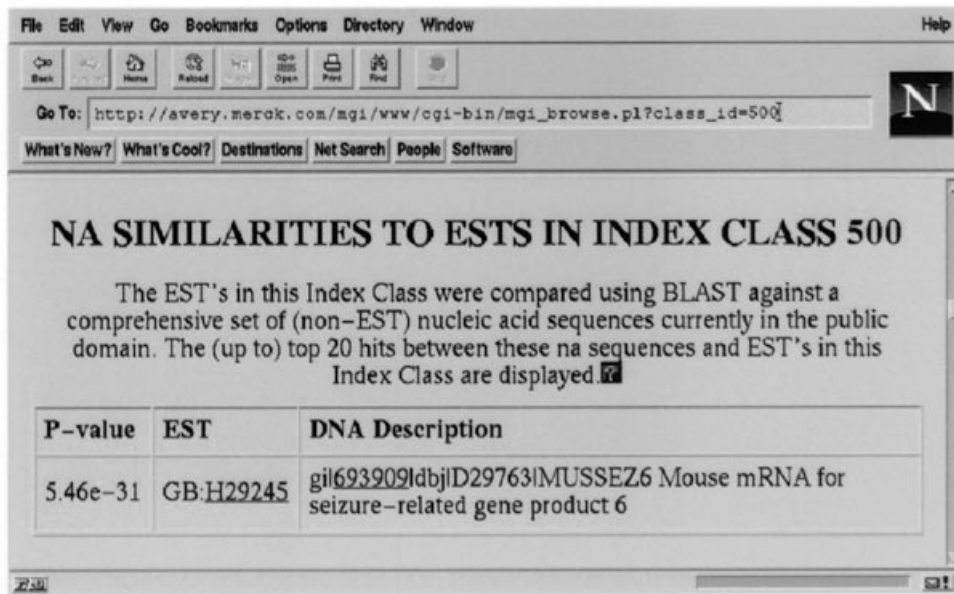


Fig. 7. A sample non-EST nucleic acid similarity data module from the MGI class report.

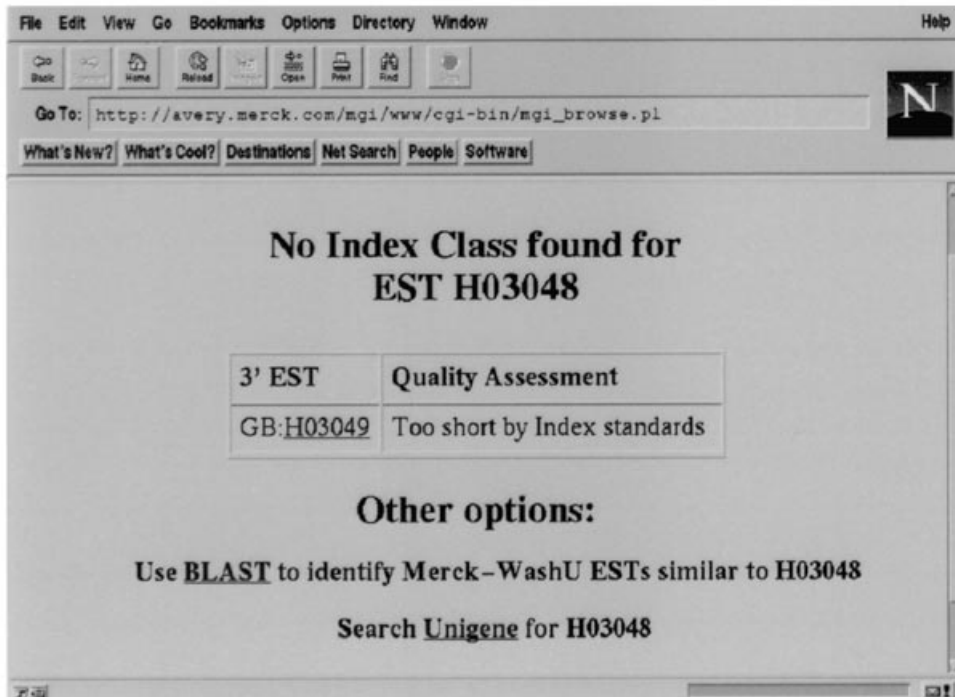


Fig. 8. If no index class is found for an input EST, the browser explains why and offers alternative avenues for investigation. In this case, the 3' sequence read from the same cDNA clone as the input (5) EST H03048 contained <100 bp of high-quality sequence.

algorithm differ from those employed in the MGI, UniGene may provide a classification for these ESTs and their underlying cDNA clones.

Text search. The text search is intended as a quick way to start a search, not a rigorous or comprehensive gene-finding method. Through this option, the user may search the

The screenshot shows a web browser window with the address bar containing 'http://avery.merck.com/mgi/vvw/cgi-bin/mgi_browse.pl'. Below the address bar are navigation buttons for Back, Home, Reload, Images, Open, Print, and Find. A search bar is present with a 'N' logo. Below the search bar are tabs for 'What's New', 'What's Cool', 'Handbook', 'Net Search', 'Net Directory', and 'Software'. The main content area displays a table with two columns: 'EST' and 'Description'. The table contains five rows of search results, each with a checkbox in the 'EST' column and a corresponding description in the 'Description' column.

EST	Description
<input type="checkbox"/> GB:AA001109	zh82f04.r1 Soares fetal liver spleen 1NFLS S1 Homo sapiens cDNA clone 427807 5' similar to PIR:JC4014 JC4014 steroid hormone-nuclear receptor NER - human ;
<input type="checkbox"/> GB:AA035697	ze25e02.r1 Soares fetal heart NbHH19W Homo sapiens cDNA clone 360026 5' similar to GB:X51416_cds1 STEROID HORMONE RECEPTOR ERR1 (HUMAN);
<input type="checkbox"/> GB:AA045123	zk63a05.r1 Soares pregnant uterus NbHPU Homo sapiens cDNA clone 487472 5' similar to PIR:JC4014 JC4014 steroid hormone-nuclear receptor NER - human ;
<input type="checkbox"/> GB:AA179970	zp14f07.r1 Stratagene fetal retina 937202 Homo sapiens cDNA clone 609445 5' similar to TR:G1117915 G1117915 STEROID RECEPTOR COACTIVATOR. ;
<input type="checkbox"/> GB:AA180462	zp14f07.s1 Stratagene fetal retina 937202 Homo sapiens cDNA clone 609445 3' similar to TR:G1117915 G1117915 STEROID RECEPTOR COACTIVATOR. ;

Fig. 9. Partial results of a text search using the input 'steroid*receptor'.

sequence similarity annotation on the GenBank definition line of Merck–WashU EST entries for text of interest. The annotation is provided by the sequencers at WashU at the time of submission of the EST, and reflects the state of the public databases at that time. The wildcard '*' matches anything or nothing: 'steroid*receptor' matches 'steroid hormone receptor err1' and 'steroid receptor tr2 (human)'. A summary of all EST entries whose definition lines include the specified text is displayed (Figure 9). The user is given the opportunity to review the set of ESTs identified, including the ability to investigate, with a single mouse click, the characterized protein or gene that was found to be similar to the EST. Sequence entries are fetched by the Bioapps server. Entries include many value-added features, including hyperlinks to related databases and the ability to save the entire sequence or individual sequence features in popular sequence file formats, to facilitate further study (Figure 10). ESTs found by the text search that are not of interest may be eliminated by clicking on the associated checkbox. After all matching ESTs have been reviewed, a single mouse click retrieves their associated index classes.

Index class ID. If the user has already identified an index class of interest, its class report may be accessed directly by entering the unique integer ID associated with the class.

Discussion

The usefulness of the MGI browser

The MGI browser has greatly enhanced the utility of the Merck-sponsored EST data for scientists, by offering an integrated, gene-centered view of the data and enabling them to mine the data from their desktop in the context of a variety of related genomics data. For drug discovery, one of the key uses of a gene index lies in the identification of novel molecular targets for therapeutic intervention. Some approaches to target identification at Merck which are currently making use of the MGI browser are: identifying ESTs that are potentially novel members of gene families of interest; choosing a broadly representative set of cDNA clones for use in differential expression studies; and identifying candidate disease genes.

Plans for further development

Indexing algorithm. ESTs containing repeat sequences, currently removed during the screening process, will undergo a repeat-masking step, and will then be subject to the index analysis. Plans are also being developed to incorporate 3' ESTs from other public EST data sets, e.g. Auffray *et al.* (1995), Berry *et al.* (1995), Okubo *et al.* (1992) and Wilcox *et al.* (1991). New methodology will be developed to validate

```

File Edit View Go Bookmarks Options Directory Window Help
Location: http://avery.merck.com/cgi-bin/ds_entry?db=pir&entry=defaults&seq=JC4014
What's New What's Cool Handbook Net Search Net Directory Software

ENTRY          JC4014      #type complete
TITLE          steroid hormone-nuclear receptor NER - human
ORGANISM       #forma_name Homo sapiens #common name man
DATE          13-Jul-1995 #sequence_revision 14-Jul-1995 #text_change
              14-Jul-1995
ACCESSIONS     JC4014
REFERENCE      JC4014
              #authors Shinar, D.N.; Endo, H.; Rutledge, S.J.; Vogel, R.; Rodan,
              G.A.; Schmidt, A.
              #journal Gene (1994) 147:273-276
              #title NER, a new member of the gene family encoding the human
              steroid hormone nuclear receptor.
              #accession JC4014
              #molecule_type mRNA
              #residues 1-461 ##label SHI
              #cross-references GB:U07132
              #experimental_source osteosarcoma cells SAOG-2/310
KEYWORDS       steroid hormone receptor
FEATURE        87-154      #domain DNA-binding #status predicted #label BIN
SUMMARY        #length 451 #molecular-weight 51102 #checksum 3C87
SEQUENCE
              5          10         15         20         25         30
1  M S S P T T S S L D T F L P G N G P P F D P P G A P S S S P T D V
31  K R R G F R P W P G G F D P D V P P G T D R A S S A C S T D W
61  V I P D F E E E P E R K R K K E G C K G P F R R S V V R G C A
91  C D K A S G F H Y N V L S C D F M R K K C Q Q C R L R K C K
121 R R Y A C G G T C C M D A F M R K K I R K Q Q Q Q L S S Q S
151 E A G M R E Q C V L S E E Q I R K K K I R K Q Q Q Q L S S Q S
181 Q S Q S F V G P Q G S S S A S G P C A S P G G S E A G S Q
211 G S G E G S G V Q L T A A Q E L M I C Q L Y A A Q L Q C N K
241 R S P S C Q P K V T P W P L G A D P C S R D A R Q G R P A H
271 F T E L A I I S V Q E I V D P A K O V P G F L O L G R E D D O
301 - A L L K A S C I E I M L L E T A R R V N H E T E C I T F L
331 K D P T Y S K D D D P H R A G L Q V E P I N P I P E F S R R A M
361 R R L G L C D A E Y A L L I A I N I P S A D R P N V Q E P P G
391 R V E A L Q Q P Y V E A L L S Y T R I X R P Q D Q L R P P R
421 M L N K I L V S L R T L S V H S E Q V P A L R L Q D E X K L F
451 P L L S E I W D V H E
///

```

Fig. 10. A PIR protein sequence entry. Using hypertext links, the user may: visit the corresponding nucleic acid sequence entry in GenBank; extract the sequence of the DNA binding domain; and extract the full protein sequence in FastA format.

existing index classes with 3' ESTs that fail the screening process, and to assign putative index classes for non-directionally cloned cDNAs and directionally cloned cDNAs that have not been successfully sequenced on the 3' end. A rigorous comparison of the index classes against other clusterings, such as the UniGene dataset, is also planned.

Class reports. Since the field of genomics is generating data at an explosive rate, there are many additional types of data that might be fruitfully integrated into the MGI browser, both Merck proprietary and publicly available data. With this in mind, the MGI browser was specifically designed to facilitate the addition of new data modules. Some examples are: mapping data from the NCBI human gene map (Schuler *et al.*, 1996); amino acid sequence motifs from, for example, PROSITE searches (Bairoch and Bucher, 1994); the presence of polyadenylation signals in the ESTs; and the results of differential gene expression experiments. Sequence assemblies could be presented which combine the ESTs in a class with other published sequence, producing a consensus sequence. Relationships among index classes could be indicated, as a means of exploring such biological phenomena as alternative splicing, gene families and overlapping genes.

Smith-Waterman alignments could be generated on demand between an EST and a similar protein or nucleotide sequence (Smith and Waterman, 1981; Hardy and Waterman, 1996).

True heterogeneous, distributed query capability. The next step in the development of interfaces to the MGI is to provide true query capability over heterogeneous, distributed data sources. For this, we plan to use Sybase's OmniCONNECT middleware, since it has a proven track record and is commercially available.

Sybase provides modules which enable access to multiple relational DBMSs (Sybase, Oracle, Informix, etc.) as if they were a single Sybase database. Further, using the libraries provided, one can write access modules for other data sources, e.g. the Bioapps server. The planned system architecture after the addition of this middleware is shown in Figure 11. Queries would be posed to the OmniSQL server, which would then pass individual queries to the underlying database servers, whether individual relational databases or the Bioapps server. Results of on-the-fly sequence analyses, e.g. BLAST searches or Phrap assemblies (Green, 1994-1996), could then be viewed as data sources in their own right, parallel to MGIdb or LENS. The use of this middleware

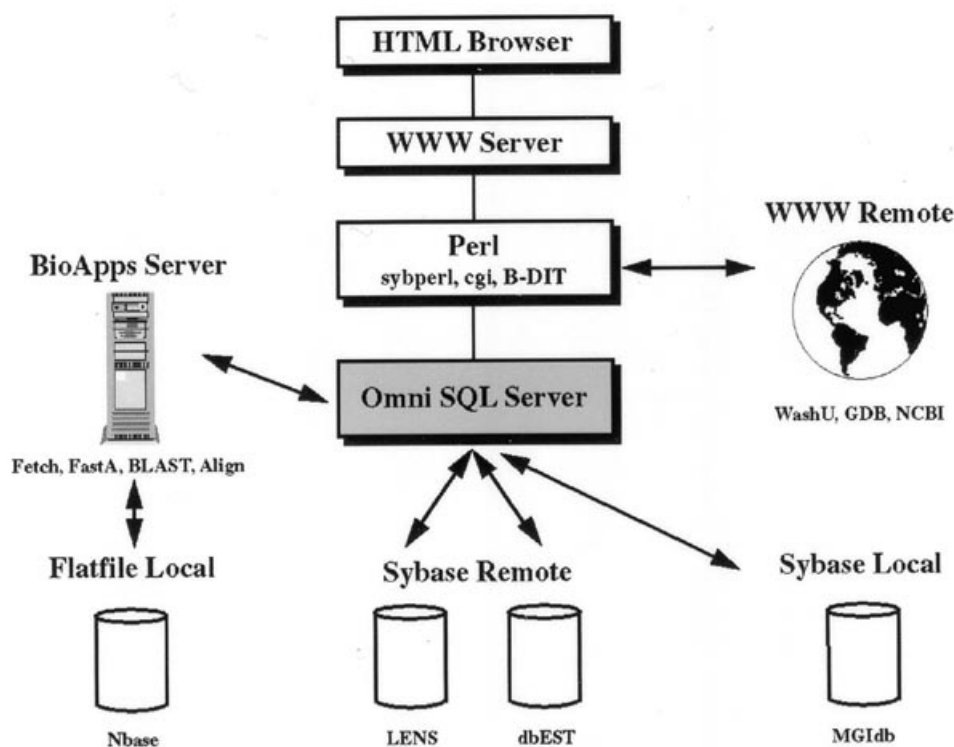


Fig. 11. The Sybase OmniSQL server middleware is the core of the system architecture in our vision of the future of the MGI browser.

would enable relational algebra operations (join, selection, projection, union and difference) to be performed across relational and non-relational sources. In addition to using a browsing strategy to winnow through the data to identify the desired classes, scientists would be able to pose a single highly targeted query such as: 'Retrieve all index classes which are differentially expressed in breast, are similar to known GPCRs or contain a TM7 motif, and are mapped to chr 8q12'.

Richer data models. We expect that the relational data model (Codd, 1970) and the 'industrial strength' Sybase middleware will enable us to provide most of the query functionality that we need in a relatively short time. However, for modeling of highly hierarchical data, e.g. three-dimensional chemical structures or hierarchies of sequence features, we plan to investigate middleware supporting a richer data model. Some examples of such middleware projects currently under development are the University of Pennsylvania's Collection Programming Language (CPL) (Buneman *et al.*, 1995), Lawrence Berkeley Laboratory's Object-Protocol Model (OPM) multidatabase query tools (Chen *et al.*, 1995), and IBM Almaden's object-oriented Garlic project (Carey *et al.*, 1995).

Acknowledgements

We would like to thank the following people: all those at Merck who made the Merck-sponsored EST project possible:

Kamil Ali-Jackson, Mary Bartkus, Werten Bellamy, C. Thomas Caskey, Oliver Johnson, Jeffrey Sturchio, Eileen Undercoffler, Cindy Zarsky; our external collaborators: Mark Boguski (NCBI), Ken Fasman (GDB), Greg Lennon (IMAGE), Chris Overton (UPenn), Bento Soares (Columbia), Robert Waterston (WashU); our alpha- and beta-version testers at Merck, including: Angela Amend, Dave Gerhold, Fred Hess, Jim Liu and Michael Phillips. We also thank Mark Gibson of UPenn for curating the LENS database, Mark Boguski of NCBI for providing relational database access to dbEST and Carolyn Tolstoshev for curating relational dbEST. Finally, we thank Anthony Starks for providing a way of performing Sybase client-server access through the firewall that conforms with Merck network security standards.

References

- Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S. and Elliston, K.O. (1996) Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Res.*, **6**, 829–845.
- Adams, M.D. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377**, 3–174.
- Alonso, R., Garcia-Molina, H. and Salem, K. (1987) Concurrency control and recovery for global procedures in federated database systems. *IEEE Data Eng. Bull.*, **10**.

- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Apache (1995–1997) Apache HTTP Server Version 1.1.3. <http://www.apache.org>
- Auffray,C. *et al.* (1995) IMAGE: molecular integration of the analysis of the human genome and its expression. *C. R. Acad. Sci. III, Sci. Vie*, **318**, 263–272.
- Bairoch,A. and Boeckmann,B. (1994) The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Res.*, **22**, 3578–3580.
- Bairoch,A. and Bucher,P. (1994) PROSITE: recent developments. *Nucleic Acids Res.*, **22**, 3583–3589.
- Benson,D., Boguski,M., Lipman,D. and Ostell,J. (1994) GenBank. *Nucleic Acids Res.*, **22**, 3441–3444.
- Benton,D. (1996) Bioinformatics—principles and potential of a new multidisciplinary tool. *Trends Biotechnol.*, **14**, 261–272.
- Berry,R. *et al.* (1995) Gene-based sequence-tagged-sites (STSs) as the basis for a human gene map. *Nature Genet.*, **10**, 415–423.
- Blevins,R., Aaronson,J., Myerson,J., Hamm,G. and Elliston,K. (1995) PROFILER: a tool for automatic searching of internally maintained databases. *Comput. Applic. Biosci.*, **11**, 667–673.
- Boguski,M.S. and Schuler,G.D. (1995) ESTablishing a human transcript map. *Nature Genet.*, **10**, 369–371.
- Boguski,M.S., Lowe,T.M.J. and Tolstoshev,C.M. (1993) dbEST—database for ‘expressed sequence tags’. *Nature Genet.*, **4**, 332–333.
- Buneman,P., Davidson,S., Hart,K., Overton,C. and Wong,L. (1995) A data transformation system for biological data sources. In *VLDB 1995*. Zurich, Switzerland.
- Carey,M.J. *et al.* (1995) Towards heterogeneous multimedia information systems: the garlic approach. In *5th International Workshop on Research Issues in Data Engineering (RIDE): Distributed Object Management*.
- Chen,I., Kosky,A., Markowitz,V. and Szeto,E. (1995) OPM*QS: The object-protocol model multidatabase query system. Technical Report LBNL-38181. Lawrence Berkeley National Laboratory, Berkeley, CA.
- Codd,E.F. (1970) A relational model of data for large shared data banks. *Commun. ACM*, **13**, 377–387.
- Davidson,S.B., Overton,C. and Buneman,P. (1995) Challenges in integrating biological data. *J. Comput. Biol.*, **2**, 557–572.
- Flanagan,D. (1996) *Java in a Nutshell*, 1st edn. O’Reilly & Associates, Inc.
- GeneCodes (1995) Sequencher. GeneCodes Corp., Ann Arbor, MI.
- George,D.G., Barker,W.C. and Hunt,L.T. (1986) The protein identification resource (PIR). *Nucleic Acids Res.*, **14**, 11–15.
- George,D.G., Barker,W.C., Mewes,H.-W., F., P. and Tsugita,A. (1994) The PIR-International protein sequence database. *Nucleic Acids Res.*, **22**, 3569–3573.
- Green,P. (1994–1996) Phrap. <http://bozeman.mbt.washington.edu/phrap.docs/phrap.html>
- Hardy,P. and Waterman,M.S. (1996) The Sequence Alignment Software Library at U.S.C.
- Heimbigner,D. and McLeod,D. (1985) A federated architecture for information management. *ACM Trans. Office Inf. Systems*, **3**, 253–278.
- Hillier,L. *et al.* (1996) Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.*, **6**, 807–828.
- Houlgatte,R., Mariage-Samson,R., Duprat,S., Tessier,A., Bentolila,S., Lamy, B. and Auffray,C. (1995) The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res.*, **5**, 272–304.
- Kernighan,B.W. and Ritchie,D.M. (1988) *The C Programming Language*, 2nd edn. Prentice Hall, Englewood Cliffs, NJ.
- Ko,M.S., Wang,X., Horton,J.H., Hagen,M.D., Takahashi,N., Maezaki,Y. and Nadeau,J.H. (1994) Genetic mapping of 40 cDNA clones on the mouse genome by PCR. *Mamm. Genome*, **5**, 349–355.
- Lennon,G.G., Auffray,C., Polymeropoulos,M. and Soares,M.B. (1996) The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics*, **33**, 151–152.
- Matsubara,K. (1995) Untitled slide presentation. In *Ventures in Genetics: Advances and Applications in Research and Technology* (Tokyo, Japan: <http://cookie.imcb.osaka-u.ac.jp/bodymap/slide/summary.html>).
- Okubo,K., Hori,N., Matoba,R., Niiyama,T., Fukushima,A., Kojima,Y. and Matsubara,K. (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.*, **2**, 173–179.
- Ozsu,M.T. and Valduriez,P. (1991) *Principles of Distributed Database Systems*. Prentice Hall, Englewood Cliffs, NJ.
- Pattabiraman,N., Nambodiri,K., Lowrey,A. and Gaber,B.P. (1990) NRL—3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment. *Protein Sequences Data Analysis*, **3**, 387–405.
- Pearson,W.R. (1991) Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.
- Peppler,M. (1996) Syperl 2.05. http://reality.sgi.com/pablo/Sybase_FAQ/Q9.4.html
- Schuler,G. *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540–546.
- Sheth,A.P. and Larson,J.A. (1990) Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Comput. Surveys*, **22**, 183–236.
- Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
- Stein,L.D. (1997) CGI.pm v 2.30. <http://www-genome.wi.mit.edu/ftp/pub/software/WWW>
- Sybase (1996) *Sybase SQL Server Reference Manual (2 vols)*. Server Publication Group, Sybase, Inc.
- Trusted Information Systems (1996, 1997) Trusted Information Systems Internet Firewall Toolkit. <http://www.tis.com>
- Wall,L., Christiansen,T. and Schwartz,R. (1996) *Programming Perl*, 2nd edn. O’Reilly & Associates, Inc., Sebastopol, CA.
- Wilcox,A.S., Khan,A.S., Hopkins,J.A. and Sikela,J.M. (1991) Use of 3’ untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Res.*, **19**, 1837–1843.
- Williamson,A.R., Elliston,K.O. and Sturchio,J.L. (1995) The Merck Gene Index, a public resource for genomics research. *J. NIH Res.*, **7**, 61–63.