

Presentation: School of Library and Information Studies, University of Michigan, 1 November 1994.

**TOWARDS A FEDERATED INFORMATION
INFRASTRUCTURE FOR BIOLOGY:**

**ELECTRONIC DATA PUBLISHING ON THE GLOBAL
INFORMATION HIGHWAY**

Robert J. Robbins

Johns Hopkins University
rrobbins@gdb.org

US Department of Energy
robbins@er.doe.gov

TABLE OF CONTENTS

<u>TOWARDS A FEDERATED INFORMATION INFRASTRUCTURE FOR BIOLOGY:</u>	1
CHALLENGES OF BIOLOGICAL DATA MANAGEMENT	2
Data Acquisition	2
Data Integration	2
ELECTRONIC DATA PUBLISHING	3
DATABASES AS LITERATURE	4
ACHIEVING INTEROPERABILITY	8
EVOLUTION OF COMPLEX SYSTEMS	10
HISTORY OF BIOLOGICAL INFORMATION SYSTEMS	12
RECENT ADVANCES	14
FIIB AND THE GLOBAL INFORMATION INFRASTRUCTURE	16
DBPS: FEDERATED OBJECT SERVER MODEL	18
FOSM Data Structures	18
THE POWER OF MIDDLEWARE—THE EXAMPLE OF GENQUEST	25
<u>BIBLIOGRAPHY</u>	31

TOWARDS A FEDERATED INFORMATION INFRASTRUCTURE FOR BIOLOGY:

ELECTRONIC DATA PUBLISHING ON THE GLOBAL INFORMATION HIGHWAY

Robert J. Robbins
Johns Hopkins University
rrobbins@gdb.org

Biology is entering a new era in which findings are generated that cannot be published in traditional literature. Databases and other electronic resources serve as literature and distribute this information to the community. The ultimate success of many biological undertakings, such as the Human Genome Project or the National Biological Survey, will require electronic data publishing. Many biological databases began as secondary literature—reviews in which certain kinds of data were collected from the primary literature. Now, these databases are becoming a new kind of primary literature with findings being submitted directly to the database and never being published in print form.

Although some electronic data publishing is here now, already changing the way biology is done, many challenges still exist: structured data cannot be adequately presented via hypertext; semantic connectivity among databases is weak or non-existent; query interoperability across multiple resources is rare; resource discovery and resource filtering are only poorly addressed; the involvement of information-management professionals is lacking; few electronic publications support the traditional role of the editor.

To be sure, the problems are not biology's alone—all science needs good information-infrastructure services. But biological problems have a special claim: In chemistry and physics all things of interest in a particular class (hydrogen atoms, electrons, quarks, etc.) are held to be genuinely, not metaphorically interchangeable. All living things, on the other hand, are genuinely unique, and their individual properties are determined in significant part by the unique, frequently contingent historical events that happened to each of their unique ancestors. The number of living things that now exist, that have existed, or that ever will exist is sufficiently small in relation to their information content, that a law of large numbers cannot be applied to them so that they might be described in all interesting ways as essentially, if not actually interchangeable items. Understanding biology will depend in some very real way on the ability to manage information in a way that preserves the individuality of the subjects. Nearly all widely used statistical tests were first invented to solve biological problems. We may hope that methods for biological information management will have equal applicability.

Current research offers guidance for developing systems for the future. We need to build a federated information infrastructure dedicated to the principles of componentry, of open and anonymous interoperability, of technical and social scalability, and of value-additivity. Industry will spend billions of dollars over the next decade providing the physical components of an information infrastructure. If the non-commercial sector—the education, library, and research communities—successfully articulate their special needs so that they may be addressed, we will see a transformation in the way information is managed and science is done.

CHALLENGES OF BIOLOGICAL DATA MANAGEMENT

Data Acquisition

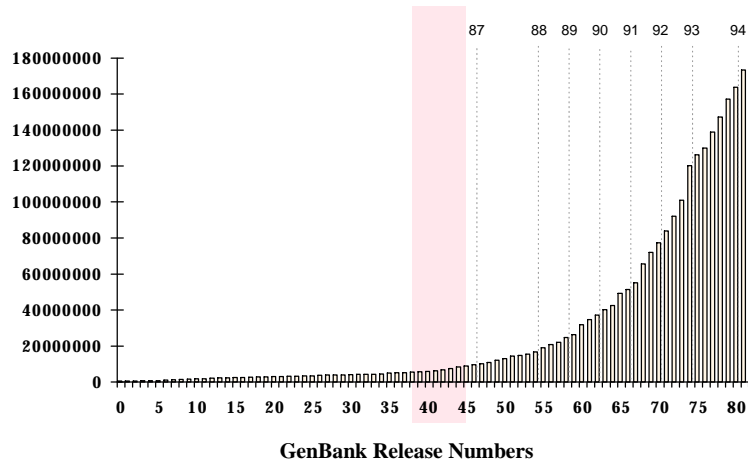


Figure 1. Growth in the world's collection of nucleotide sequence data, shown as the number of bases contained in every release of GenBank from 1 through 82. The numbers at the tops of the dotted lines show years (which do not necessarily coincide with a particular number of releases). The shaded bar in the middle represents the period in the mid 1980s when the data volume was, for a time, more than the databases could handle. (Data supplied by Michael Cinkosky and Dennis Benson.)

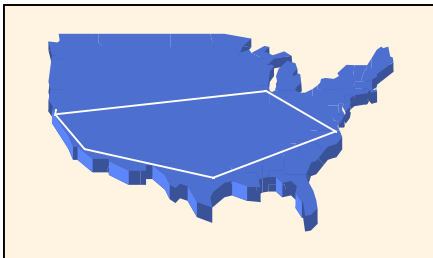


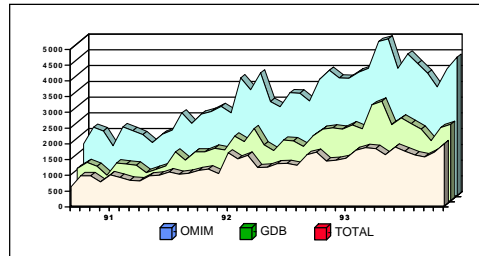
Figure 2. The DNA sequence in one human sperm cell, typed on a continuous ribbon in 10-pitch type, would reach from San Francisco to Chicago to Washington to Houston to Los Angeles and back to San Francisco, with 60 miles of ribbon to spare. The amount of human sequence in the databases now is equal to less than that extra 60-mile piece.

Data Integration

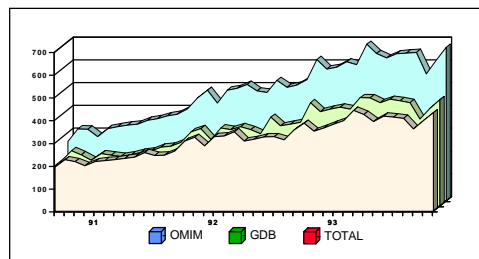
Molecular biology has survived the data-acquisition crisis of the 1980s. Now it faces a data-integration crisis. Data are flowing into databases at an exponential rate, yet the ability to link data objects across data objects is inadequate. As data are entered without proper cross-referencing, the utility of the data declines and the work deferred mounts.

ELECTRONIC DATA PUBLISHING

Sessions



Unique Users



**Sessions
(third-party software)**

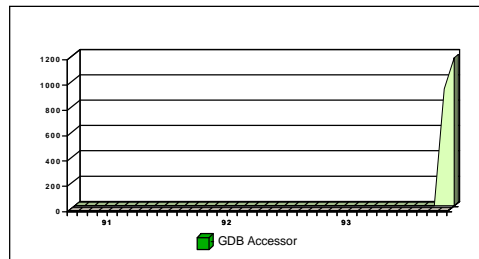


Figure 3. GDB/OMIM monthly usage statistics for sessions and unique users. A “session” is defined as one login session of any duration and a “unique user” is defined as at least one login during the month for an individual user.

Publishing on Demand

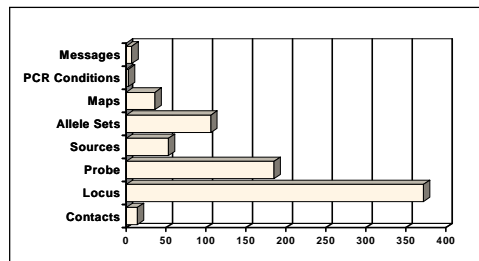
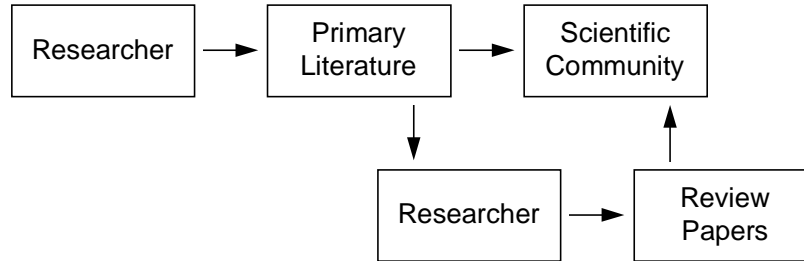


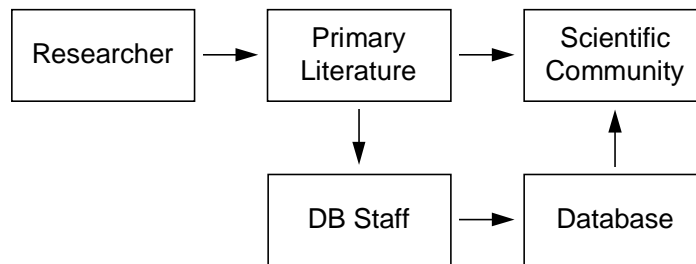
Figure 4. Over 750 custom reports were generated during the month of October, 1993.

DATABASES AS LITERATURE

Traditional Literature:



Database Development:



Early Electronic Data Publishing:

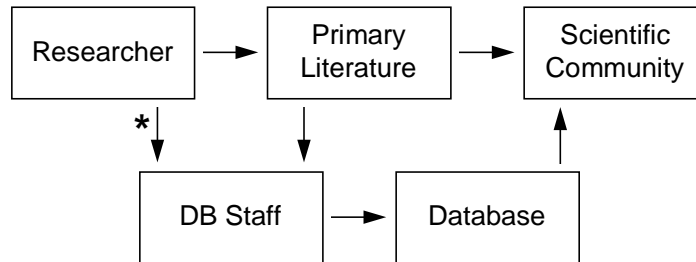
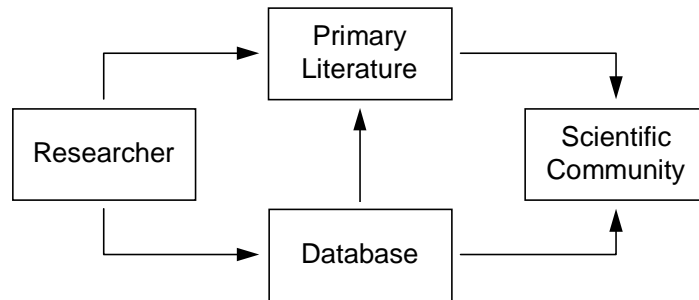


Figure 5. Many biological databases began like traditional review papers. An individual researcher combed the primary literature, collecting relevant data. When the general utility of the collection was recognized, funding was obtained, and a staff hired to accelerate the process of literature scanning. With an exponential growth in some of the data, this approach could not scale, so electronic data publishing was developed, with the creators of the data now responsible for submitting their findings directly to the database, as well as to the traditional literature.

Mature Electronic Data Publishing:



Federated Information Infrastructure:

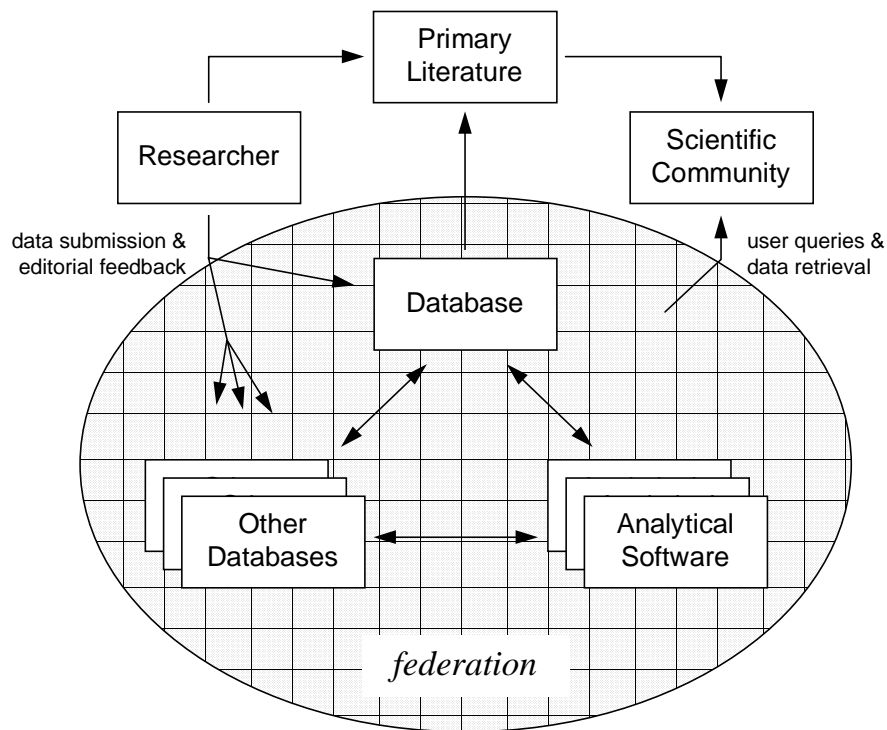


Figure 6. As electronic data publishing matures, the database staff seem to disappear and researchers feel like they are communicating directly with each other. Now, the call is for developing a federated information infrastructure (FII) so that researchers may interact with a single, virtual information resource.

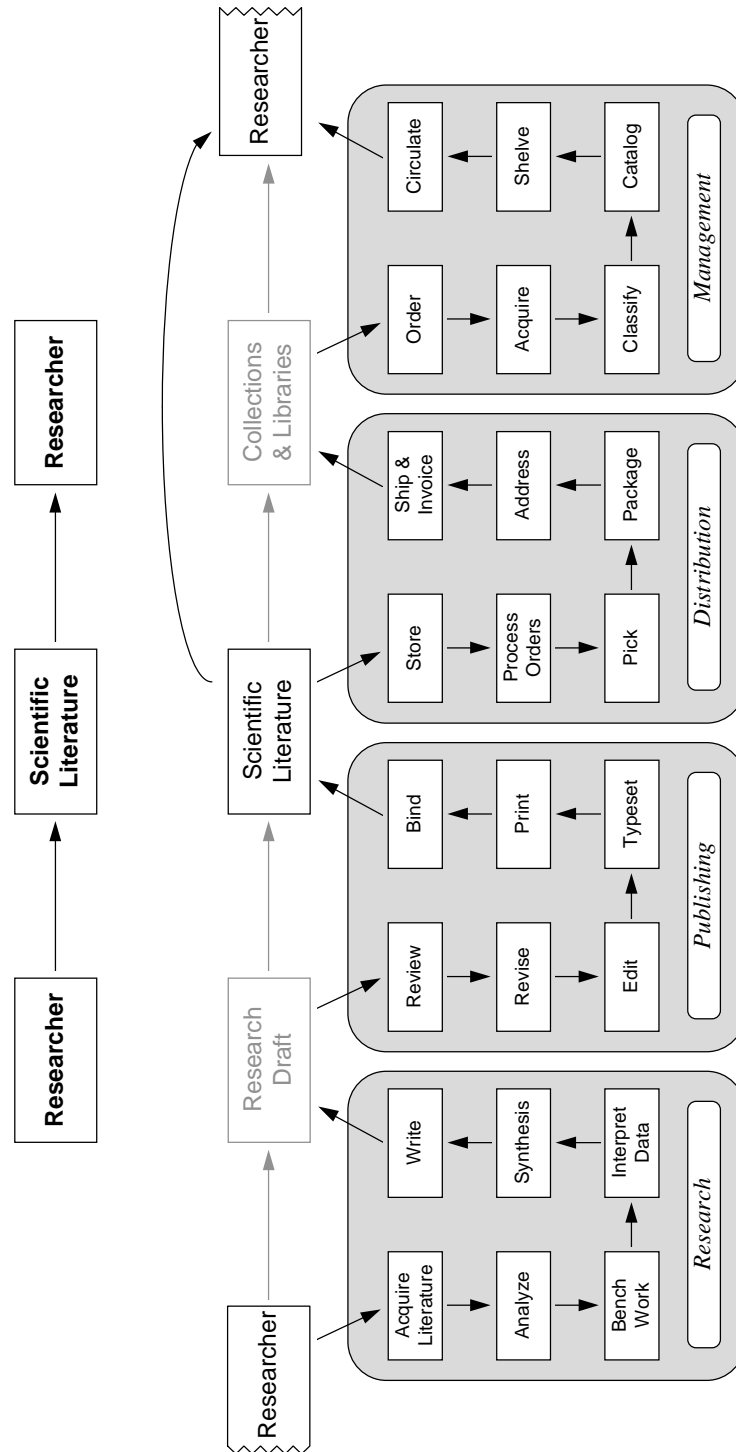


Figure 7. Scientists imagine that they communicate with each other fairly directly via the scientific literature (upper figure). In fact, their communication is based on a substantial infrastructure, without which the communication would be difficult or even impossible. For print publication, this infrastructure functions so well that its presence is barely noticed.

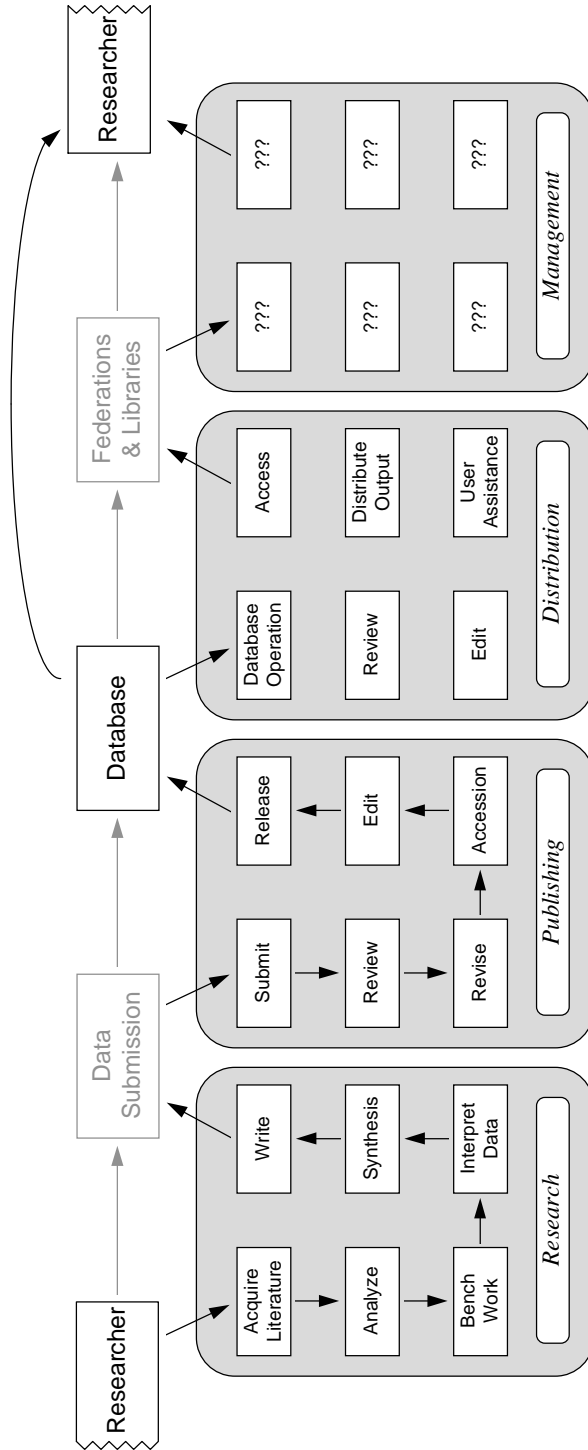


Figure 8. As with print literature, mature electronic data publishing feels like a direct communication among researchers via transactions in the database. Again, a substantial infrastructure is required to make the system work. Although many are calling for the development of an interoperating federated information infrastructure, implementing such a system among heterogeneous databases is still a research challenge in computer science. Defining the tasks for managing collections of such systems, and developing the means for accomplishing those tasks are among the challenges facing the developers of digital libraries. If the system is to scale gracefully, some of the tools (such as name-resolution for individual data objects as well as for information resources) for managing these collections must be deeply embedded in networking protocols.

ACHIEVING INTEROPERABILITY

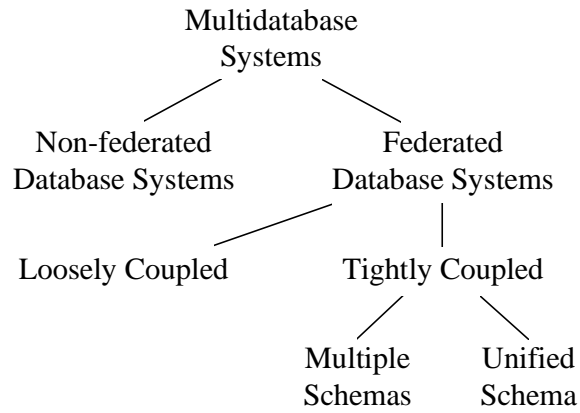


Figure 9. A taxonomy of multidatabase systems, according to Sheth and Larson (1990).

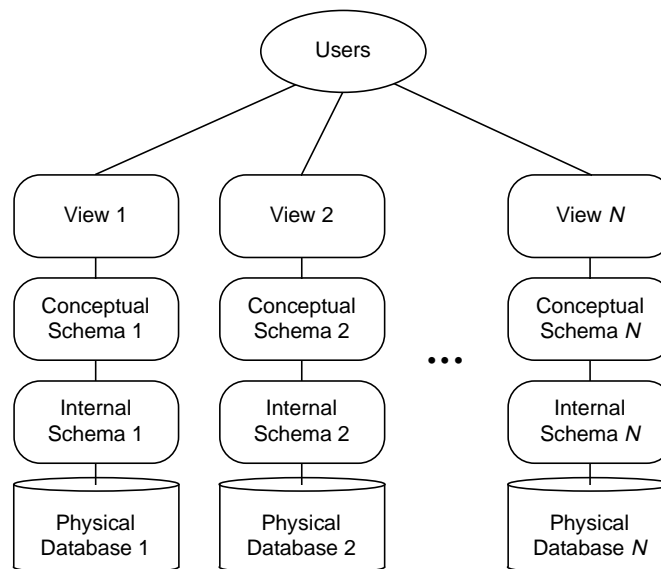


Figure 10. At present, most biological databases present their own local views and schemas to the users, who are then responsible for integrating the information.

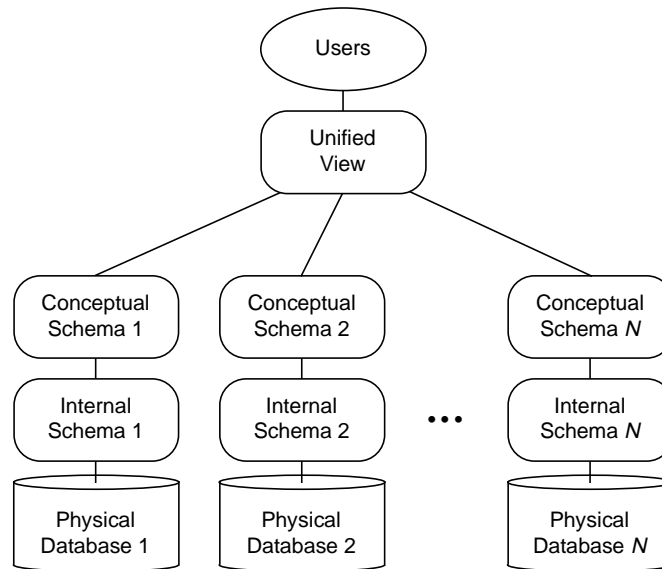


Figure 11. A common, unified view into multiple databases is needed. Here “view” is used in the common sense, not in the formal sense of a database view.

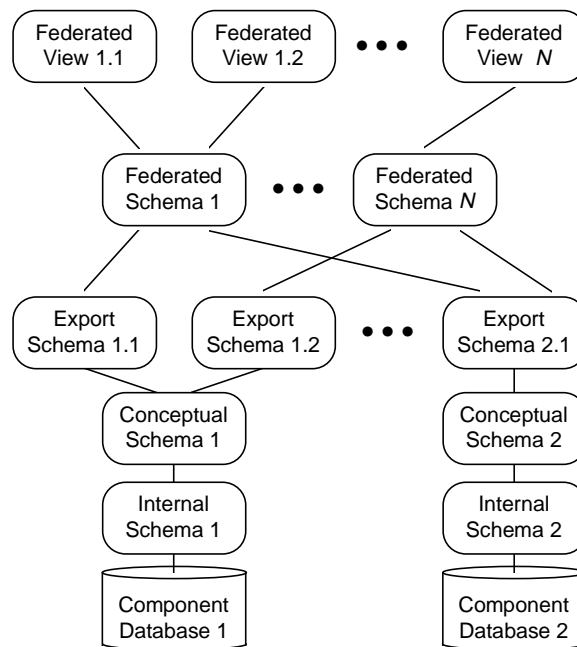


Figure 12. Many researchers believe that an essential first step toward database interoperability is the preparation of one or more export schemas by each participating database. The export schemas are then integrated into one or more federated schemas, which serve as the basis for one or more federated views into the underlying integrated information resource. (Figure adapted from Sheth and Larson, 1990).

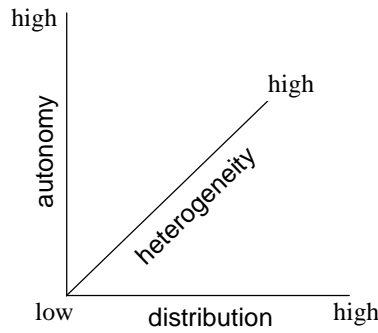


Figure 13. Some difficulty dimensions for integrating multiple database systems. More dimensions could be added. Difficulty increases with distance from the origin on any dimension. Interactions among high difficulties on multiple dimensions could lead to difficulty increases at a greater than linear rate.

UNFEASIBLE	COMMON GLOBAL SCHEMA
	MEDIUM TO LONGER TERM SOLUTION
	SHORT TERM SOLUTION
UNACCEPTABLE	DO NOTHING IN ASSURING INTEROPERABILITY

Figure 14. Options for integrating networked databases (adapted from Chorafas and Steinmann, 1993).

EVOLUTION OF COMPLEX SYSTEMS

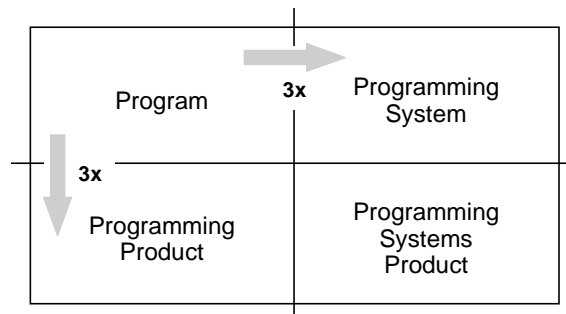


Figure 15. The evolution of a programming systems product (adapted from Figure 1.1 of Brooks, 1982, page 5). Each transition across a complexity boundary multiplies the difficulty significantly.

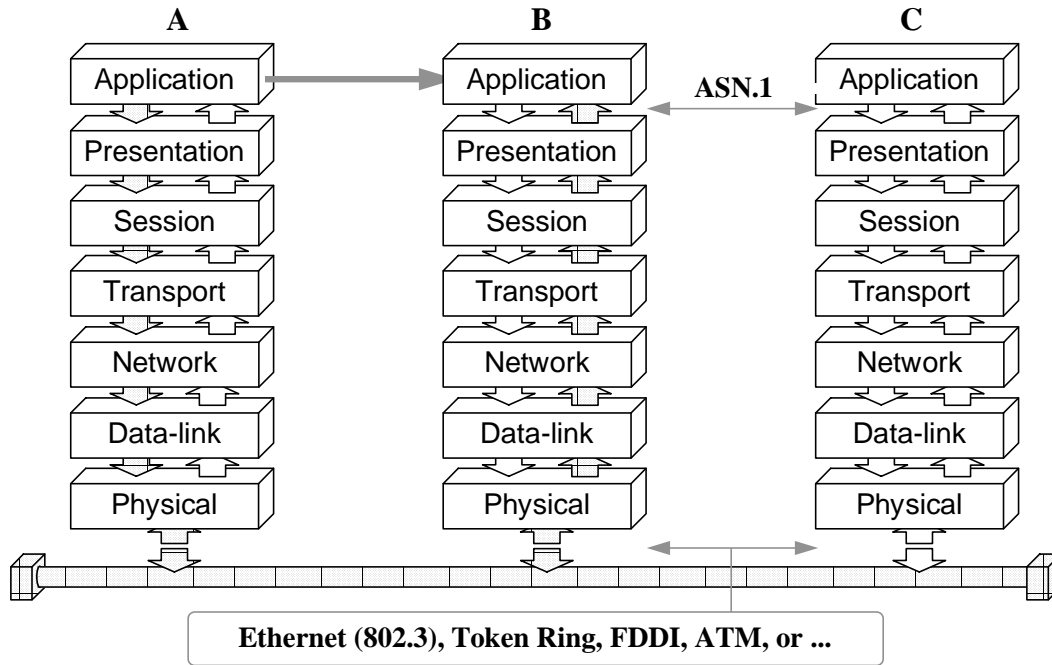


Figure 16. The seven layers of the ISO–OSI reference model. Virtual connectivity between applications on different computers is accomplished via direct communication between adjacent layers within each system, according to well-defined protocols. For example, abstract syntax notation (ASN.1) defines the communication protocols between the application and presentation layers.

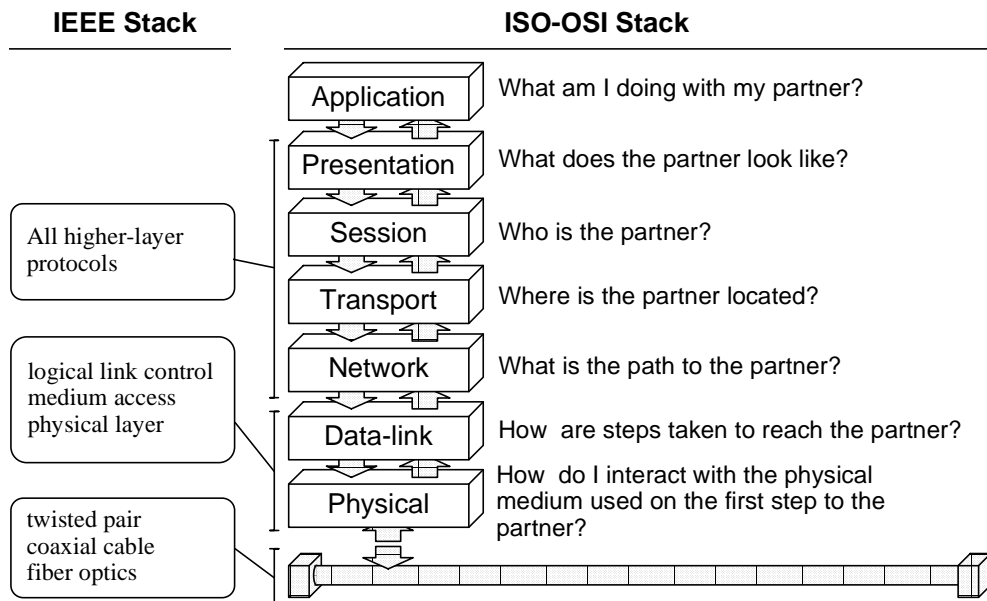


Figure 17. Functional responsibilities of different networking layers in the ISO-OSI model.

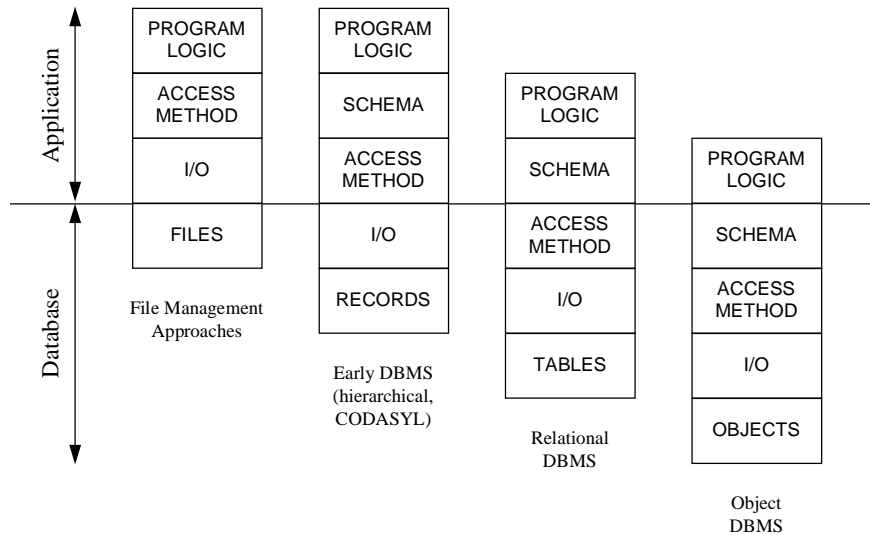


Figure 18. The evolution in database management methods (adapted from Figure 5.1 of Chorafas and Steinmann, 1993, page 75). Although some take exception with specific aspects of this figure, few would dispute that two trends are evident in the evolution of data-management systems: (1) an increasingly complex and abstract data structure at the bottom layer (files to records to tables to objects), and (2) a movement of more and more components away from the specific domain of the application and into the generic tools of the database management system.

HISTORY OF BIOLOGICAL INFORMATION SYSTEMS

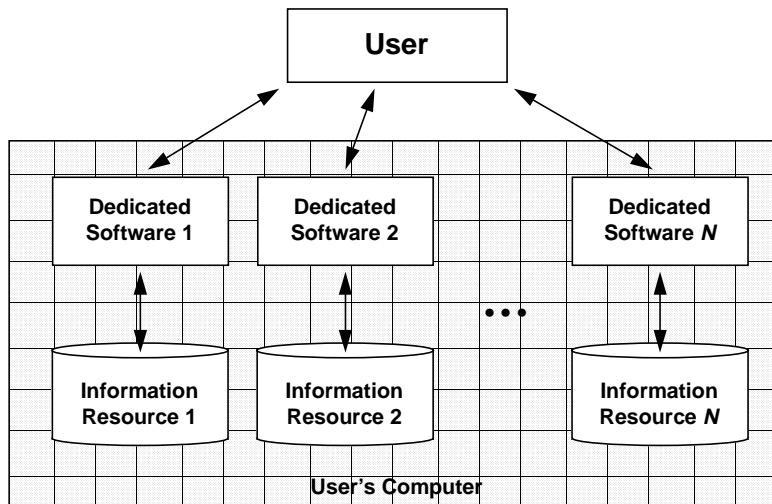


Figure 19. Early biological information systems operated as complete packages on single computers.

Figure 20. Client-server systems moved the information resource to the server, leaving only dedicated client software on the user's computer. The dedicated clients usually communicate with the dedicated servers via custom, proprietary protocols. And, the dedicated clients have no way of communicating with each other, leaving integration of the information from the different resources up to the user

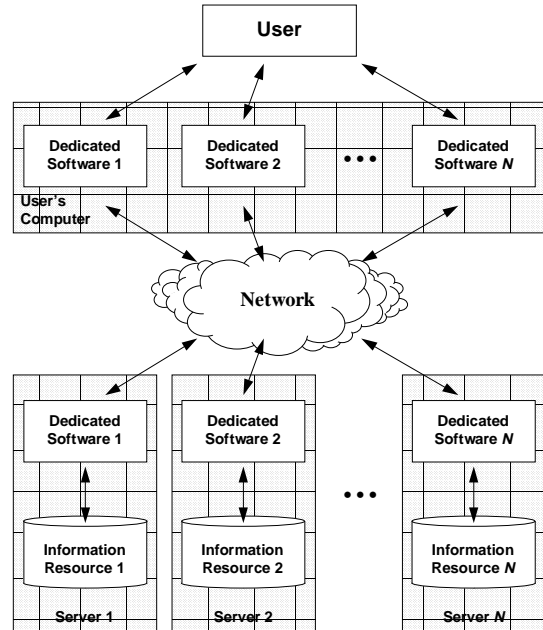
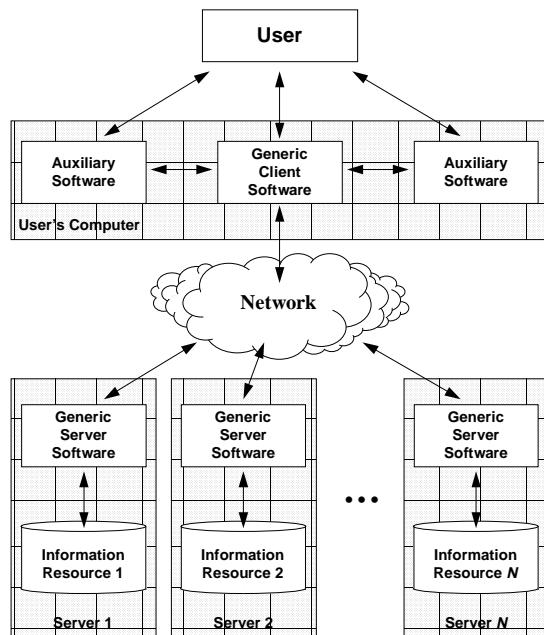


Figure 21. Generic client-server systems are data-driven systems that allow access to multiple servers through a single client. Because multiple servers are accessed through the same client, technical integration among the different resources is automatically achieved. Full semantic integration depends upon the presence of appropriate cross-references in the data provided by the various servers.



RECENT ADVANCES

Figure 22. Growth in the use of gopher and World Wide Web (WWW) protocols for accessing data on the internet since January 1993. Use of these protocols is increasing much more rapidly than is use of the net itself.

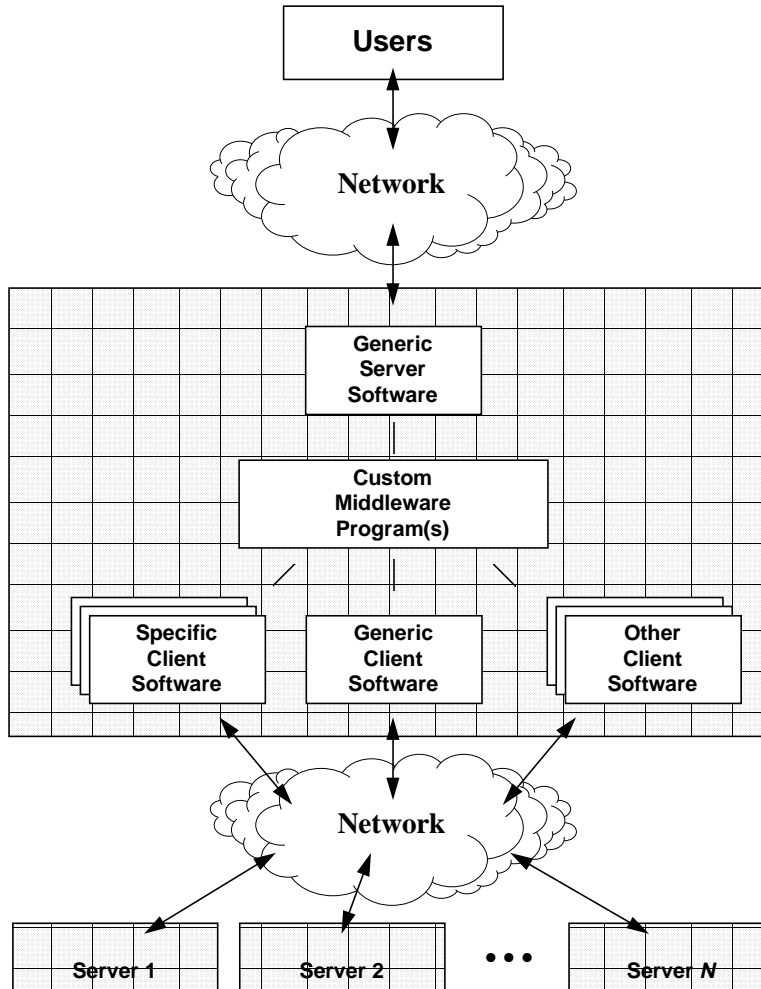
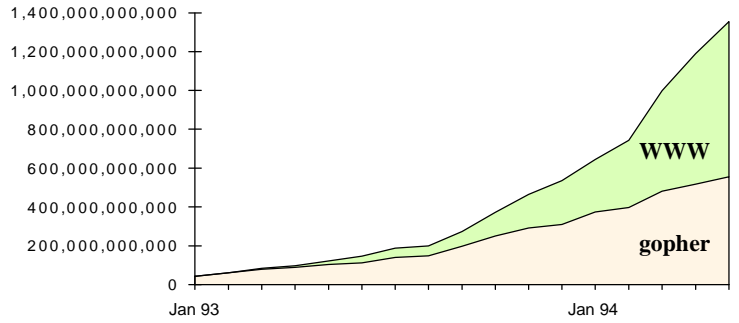


Figure 23. Powerful middleware can be built from generic client-server components, by sandwiching custom programs between server and client modules.

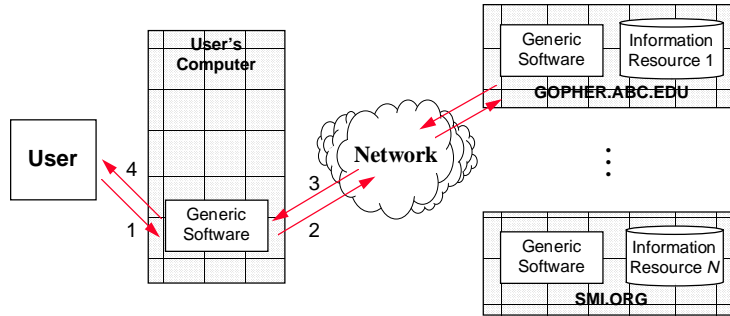


Figure 24. At startup, the client sends a default request for the top level menu to the server named gopher.abc.edu. The server responds with the contents of the menu, resulting in the menu being displayed on the user's system. The steps of the process are: (1) the user starts the generic software, (2) the client requests a menu from a server, (3) the server provides the menu (as structured text), and (4) the client displays the menu to the user.

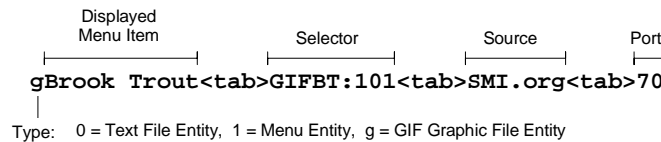


Figure 25. Gopher menu files consist of a set of five-part records. The first part is a single character that indicates the type of the file that will be returned when the menu choice is invoked. (There are many other types available, besides those indicated in the figure.) The second part is a character string to be displayed by the gopher generic menu display system. The third part is the selector, a structured string that identifies the file on the host system to be returned. Fourth is the name of the source—the computer that will be asked for the file. Fifth is the port identifier that is to be included with the request. This record would result in a menu choice being offered to the user that would presumably return a picture of a brook trout in gif file format.

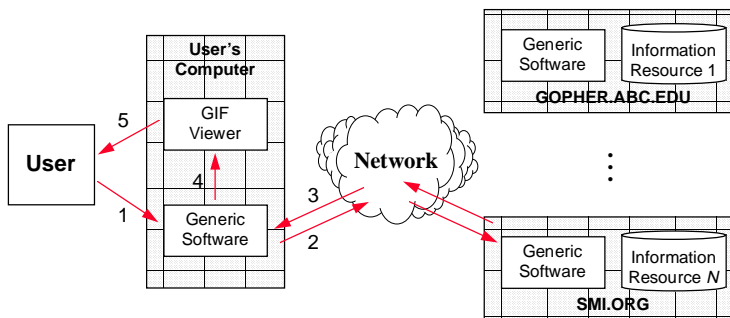


Figure 26. A request for a gif file is sent to host SMI.ORG. Because the menu record had identified the file type as gif, the generic client is automatically prepared to spawn an appropriate gif-viewing program when the file is returned from the host computer. The steps of the process are: (1) the user makes a menu choice on the client, (2) the client requests a file item from the server named in the menu record, (3) the server supplies the file (as binary), (4) the client launches an appropriate viewer for the file, and (5) the viewer displays the file contents to the user.

FIIB AND THE GLOBAL INFORMATION INFRASTRUCTURE

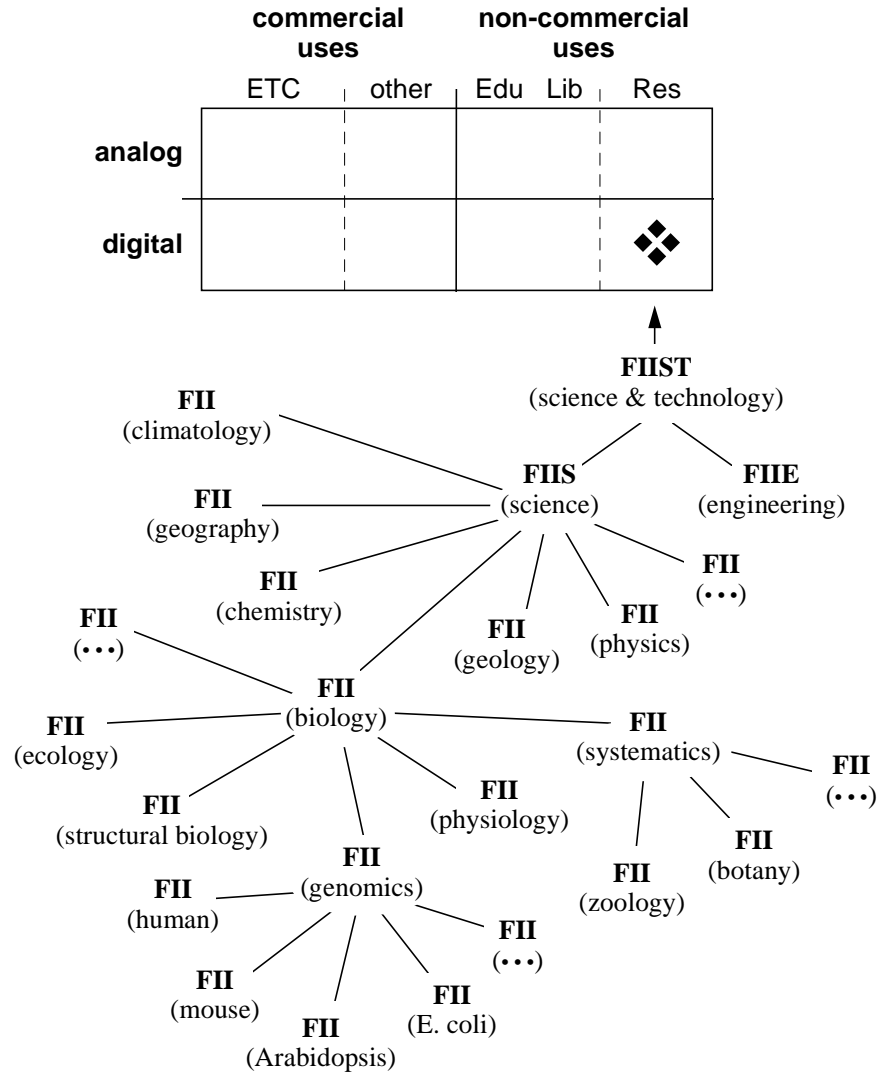


Figure 27. Many different fields in biology and other sciences are becoming increasingly dependent upon access to a coherent information infrastructure of electronically published text and data. Interoperability among different electronic resources is required, at least at the level of a loose federation. None of these subgroups is completely independent of any other, and this is true at all levels in the hierarchy. Understanding the genome will ultimately require integrating genome findings with protein structure (structural biology) and metabolic information (physiology). Comparative genomics involves systematics and other areas of comparative biology. This non-independence can involve merely the need to cross reference objects in other databases or the mutual need to access shared resources or a parallel need for similar resources (e.g., bibliographic information, geographic reference data, molecular structure, etc.)

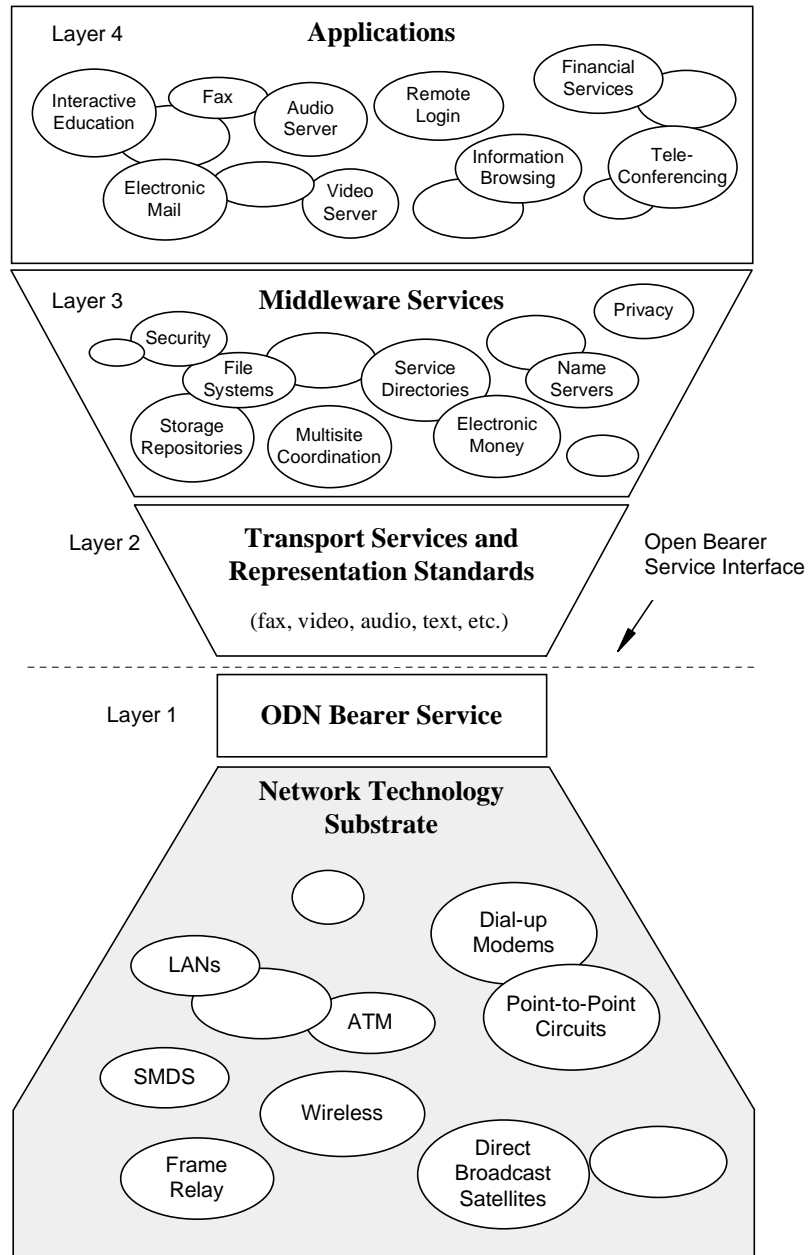


Figure 28. A four-layer model for the Open Data Network. (adapted from NREnaissance Committee, 1994, *Realizing the Information Future*. Washington, D.C.: National Academy Press.)

DBPS: FEDERATED OBJECT SERVER MODEL

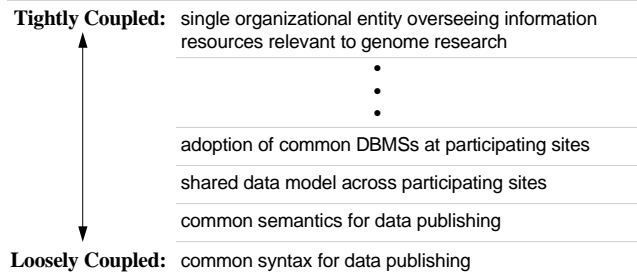


Figure 29. The distinction between tightly coupled and loosely coupled systems, seen as designating the ends of a continuum of relationships among database publishing systems. The tightest level of coupling yields a completely integrated, single management structure. The loosest level of coupling involves a collection of wholly independent organizations that share in common only a willingness to publish their data in a common syntax.

FOSM Data Structures

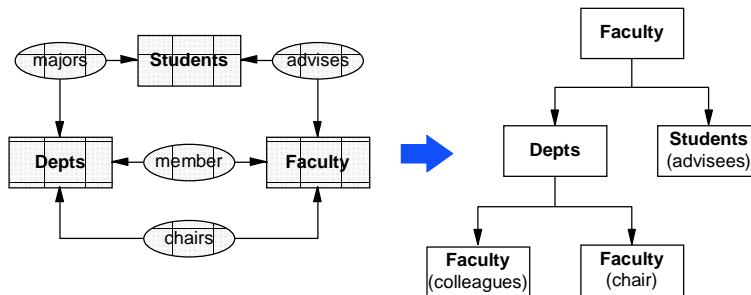


Figure 30. Tree data objects can be easily extracted from EER schemas. Here a “faculty” object is extracted from a portion of a university database schema.

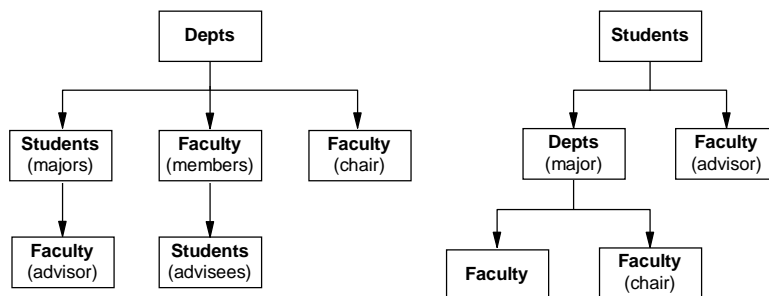


Figure 31. Other trees may be extracted from the same overall data structure.

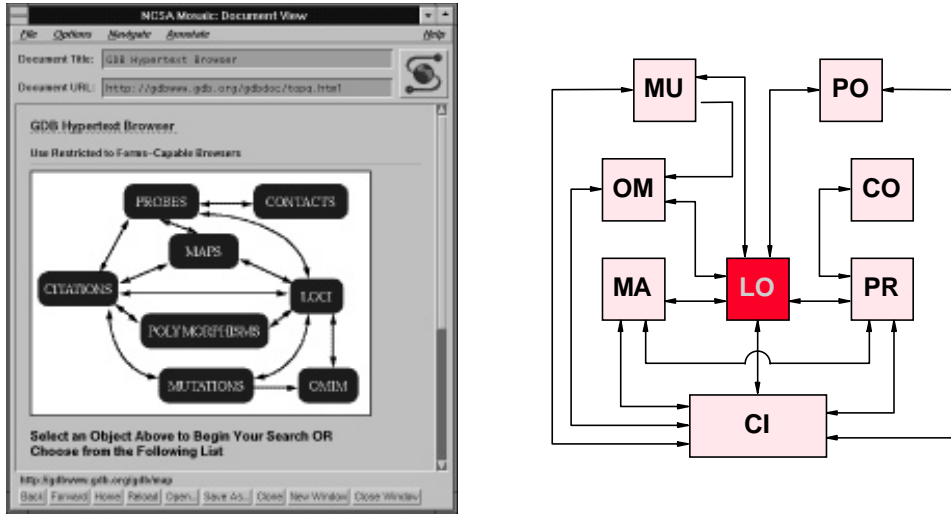


Figure 32. Tree data structures arise naturally in real biological databases. At left is a screen from the browser interface into the Genome Data Base at Johns Hopkins. At right is that same figure converted into a more formal graph.

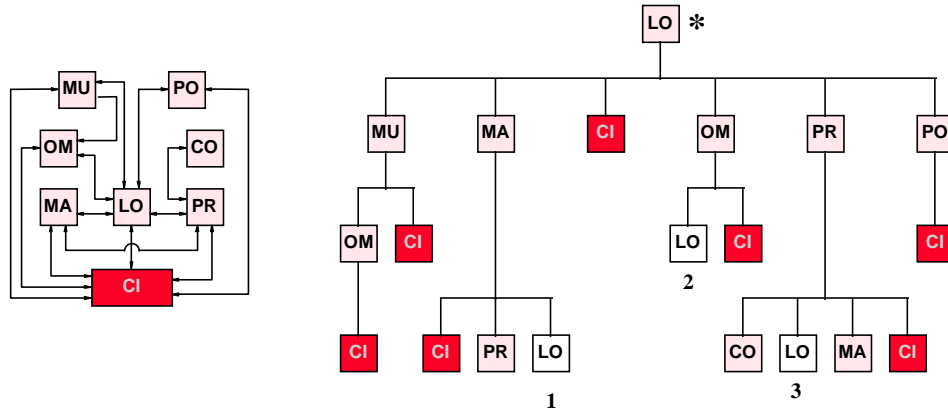


Figure 33. A “locus” object is extracted from a portion of the Genome Data Base schema. (LO = locus, MU = mutation, MA = map, C = citation, OM = OMIM, PR = probe, PO = polymorphism, CO = contact.). Notice that the citation node is repeated several times, each time with a different meaning. Even the root node can be repeated with different (and useful) semantics at each location.

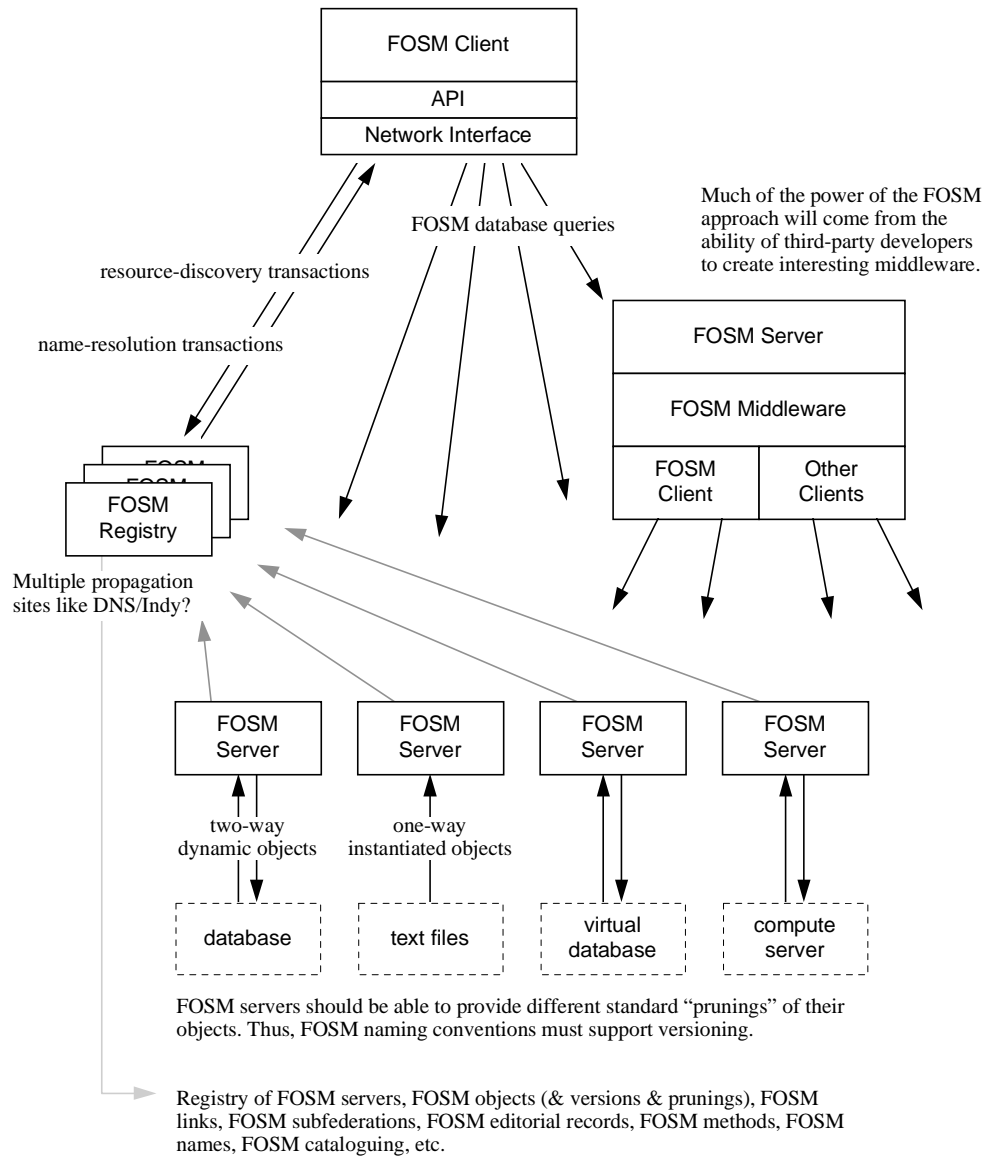


Figure 34. The basic client-server FOSM architecture also includes a public resource registry and explicit support for nth-party developers.

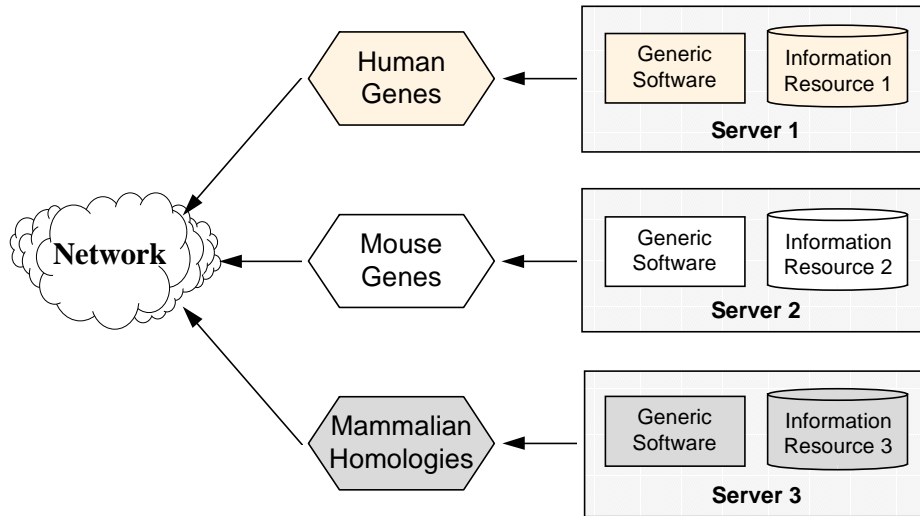


Figure 35. In a FOSM environment, individual data resources would publish their holdings to the network in a standard format, according to standard protocols.

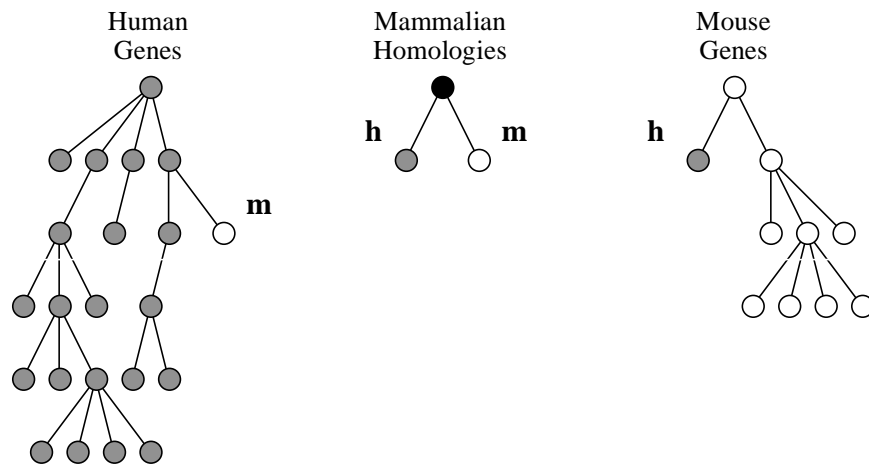
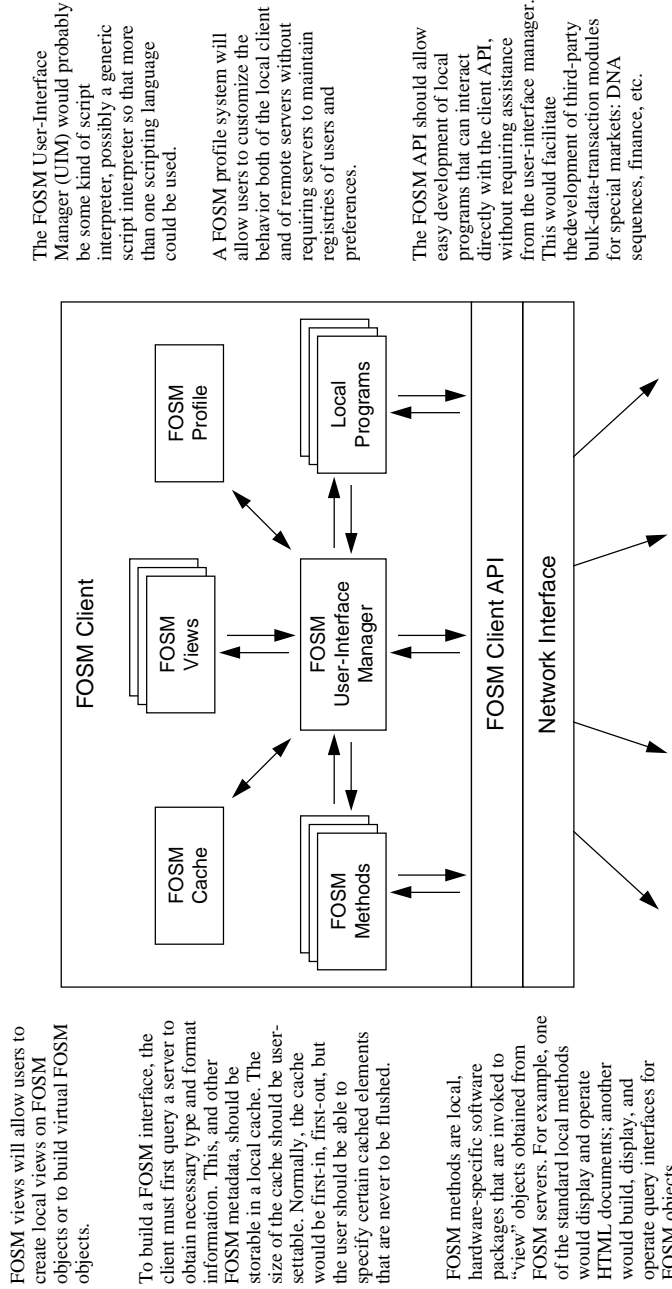


Figure 36. Possible tree structures for data objects published by FOSM servers. Nodes marked with “m” and “h” represent sets of tokens that would correspond to the root nodes for mouse–gene and human–gene objects respectively. The inclusion of these external references as leaf nodes indicates that the designer of the local database believes that these external objects are related to the database’s primary objects in some role (which is defined in the local database). The decision to include such references, and the populating of them with values, would be the responsibility of the local FOSM server.



FOSM views will allow users to create local views on FOSM objects or to build virtual FOSM objects.

To build a FOSM interface, the client must first query a server to obtain necessary type and format information. This, and other FOSM metadata, should be storable in a local cache. The size of the cache should be user-settable. Normally, the cache would be first-in, first-out, but the user should be able to specify certain cached elements that are never to be flushed.

FOSM methods are local, hardware-specific software packages that are invoked to "view" objects obtained from FOSM servers. For example, one of the standard local methods would display and operate HTML documents; another would build, display, and operate query interfaces for FOSM objects.

Figure 37. The FOSM client provides much of its functionality through its component-based design. All aspects of the FOSM system are intended to facilitate the value-adding activities of third-part developers.

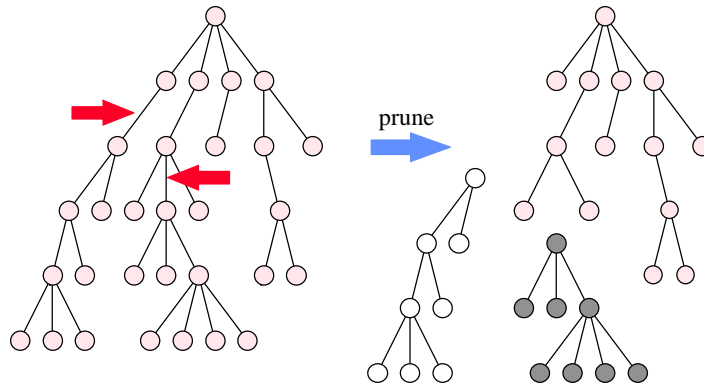


Figure 38. The *prune* operator is similar to the relational *project* operation.

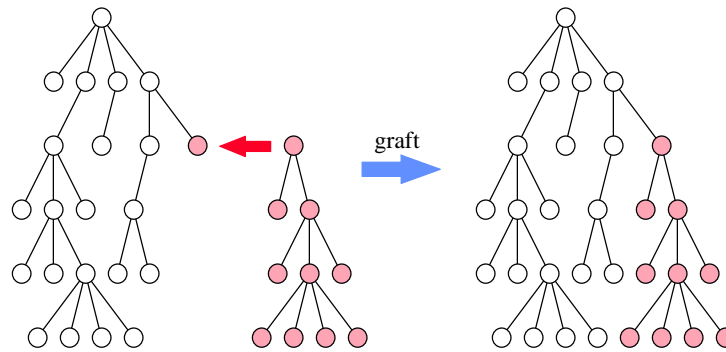


Figure 39. The *graft* operator is similar to the relational *join* operation.

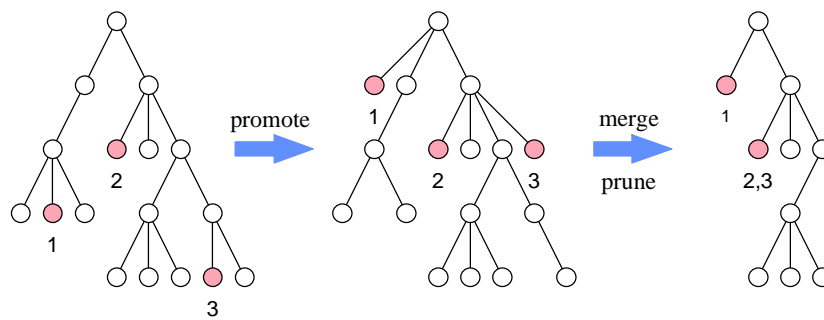


Figure 40. The *promote* operator would allow users to move nodes upwards in the data tree. If this were to result in multiple “union compatible” nodes being attached at the same point in the tree, these nodes could be *merged*.

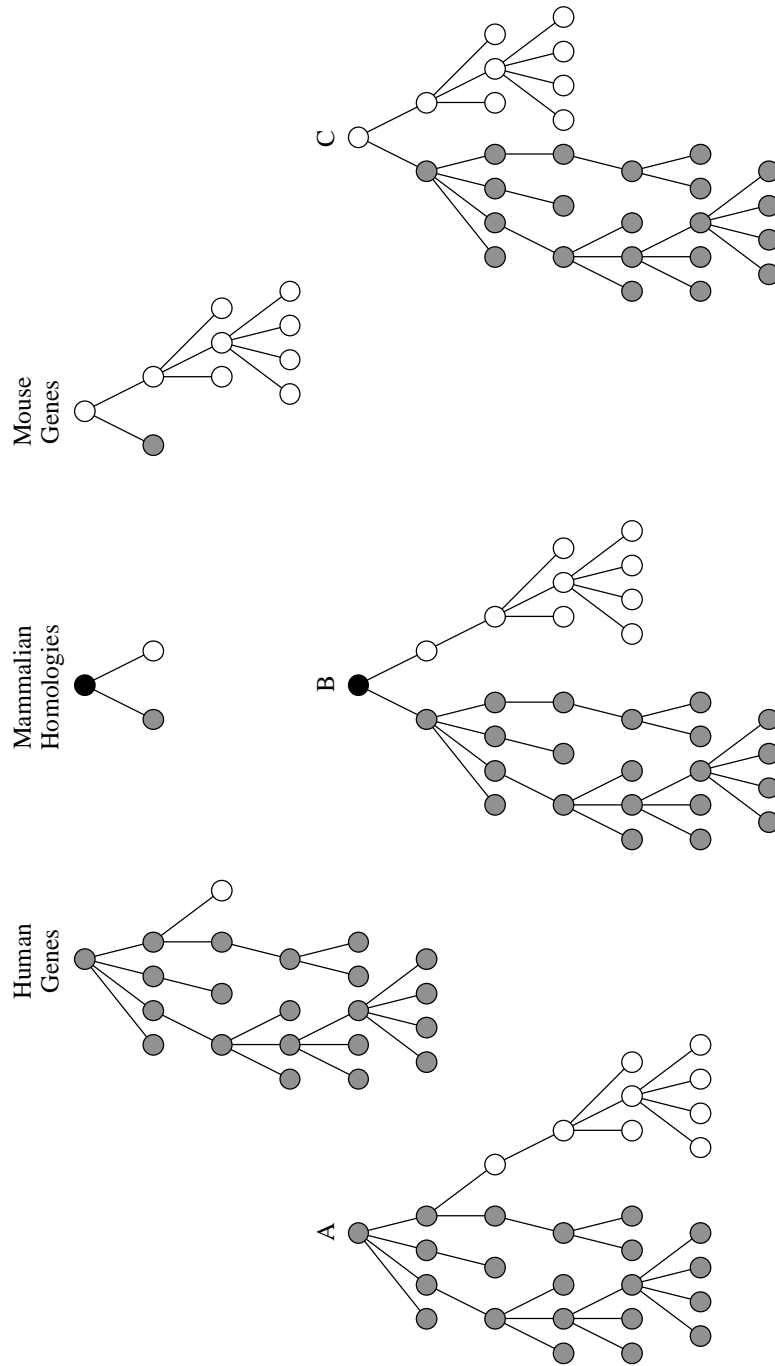
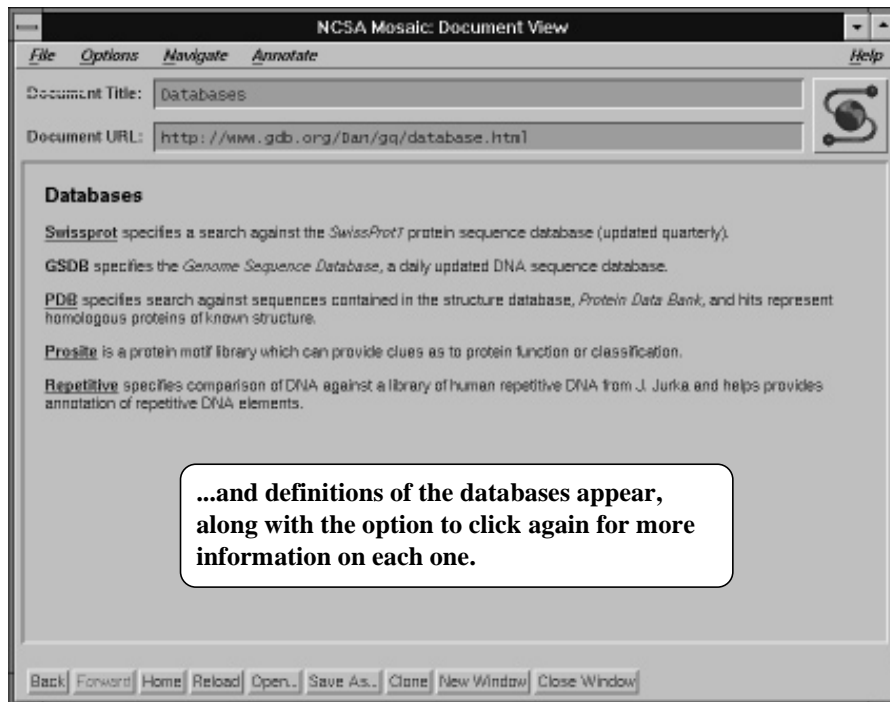
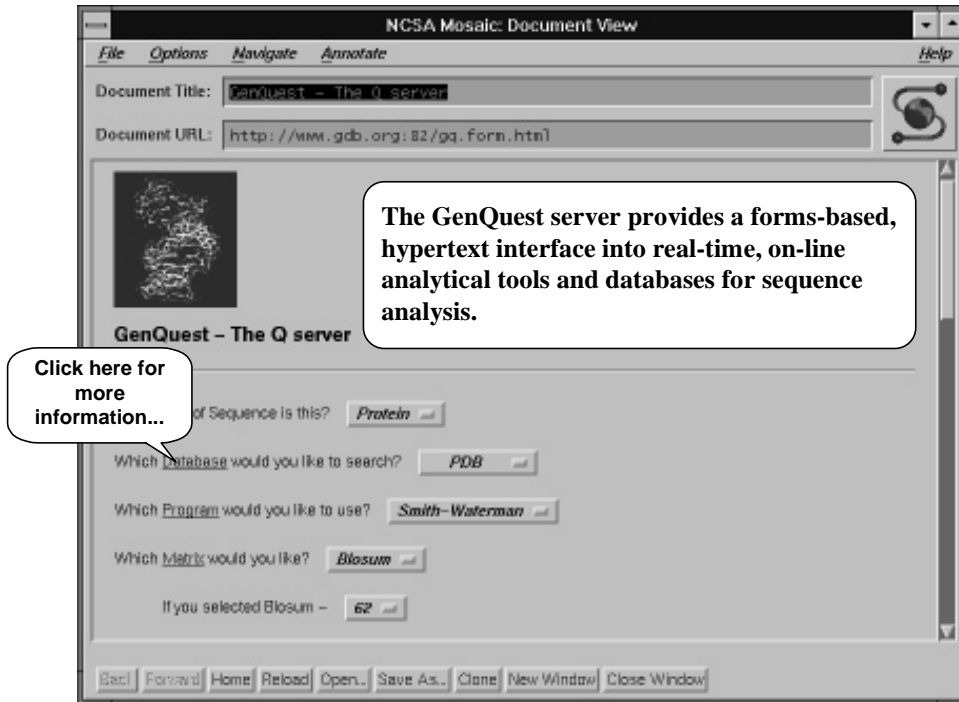
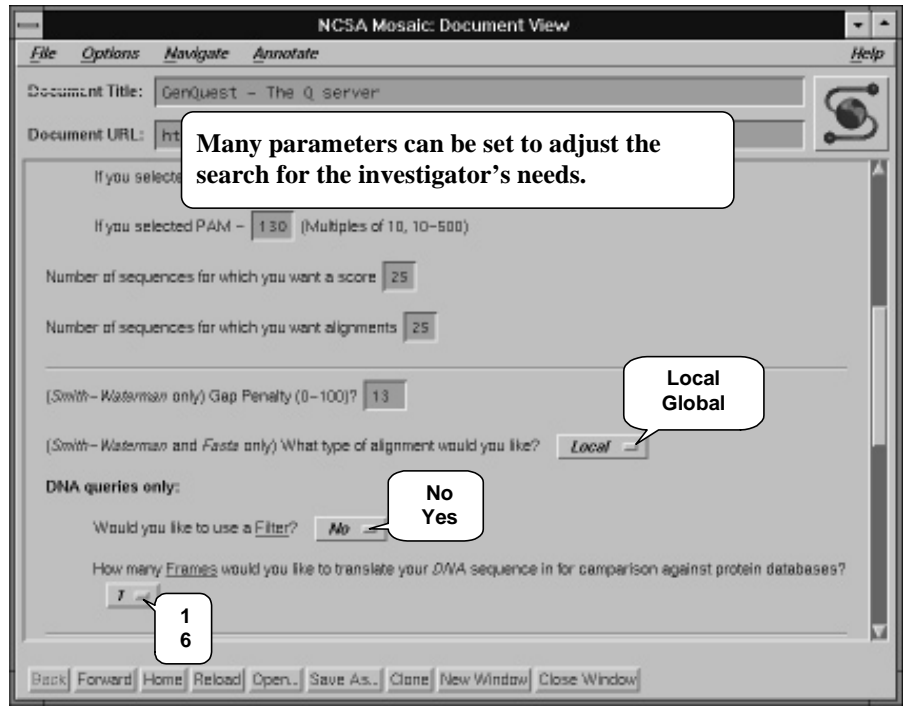
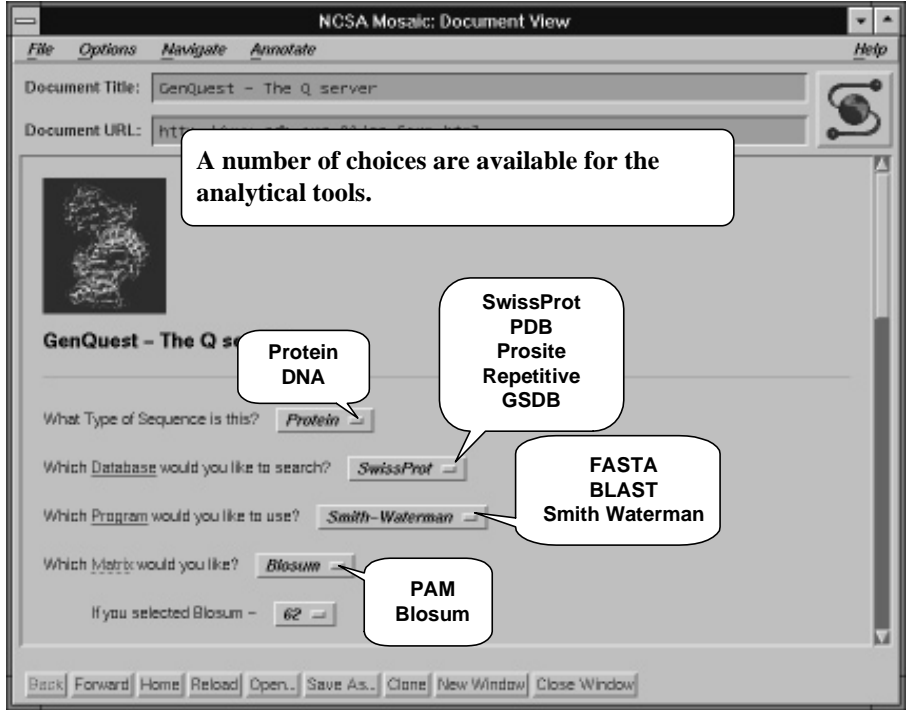
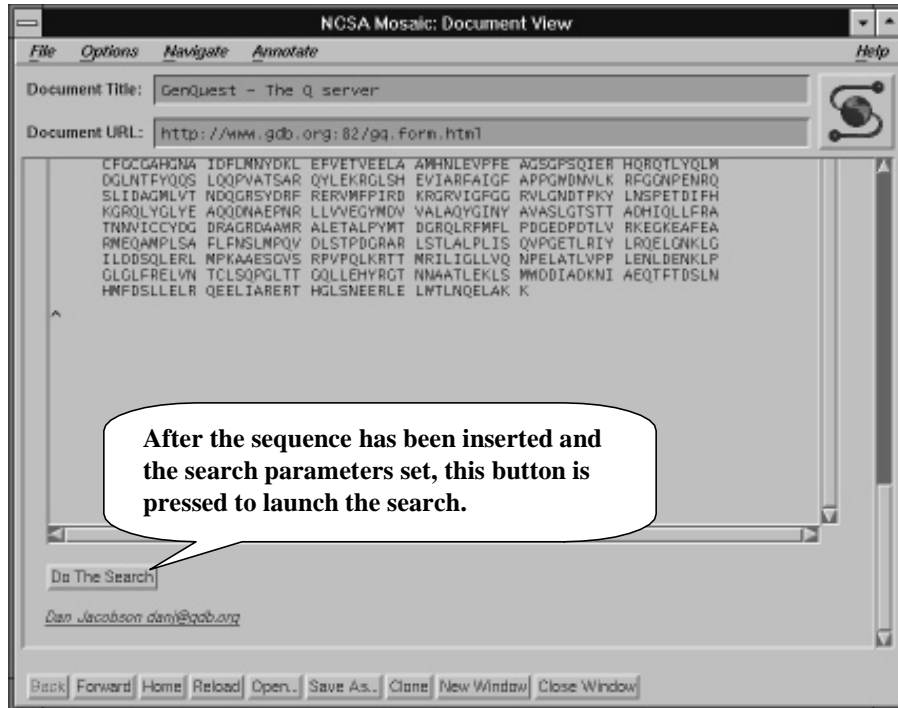
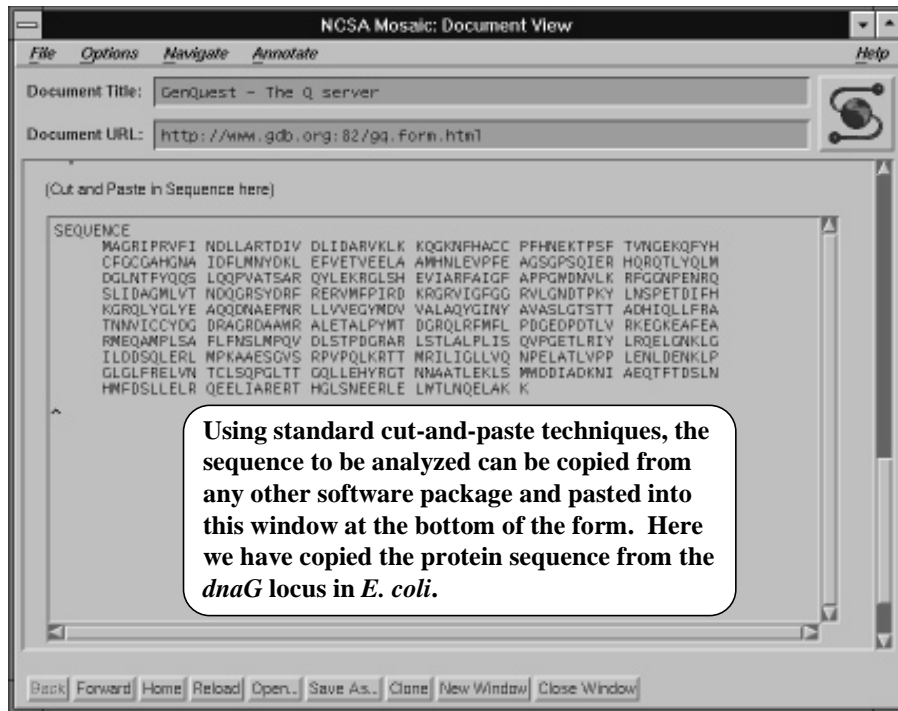


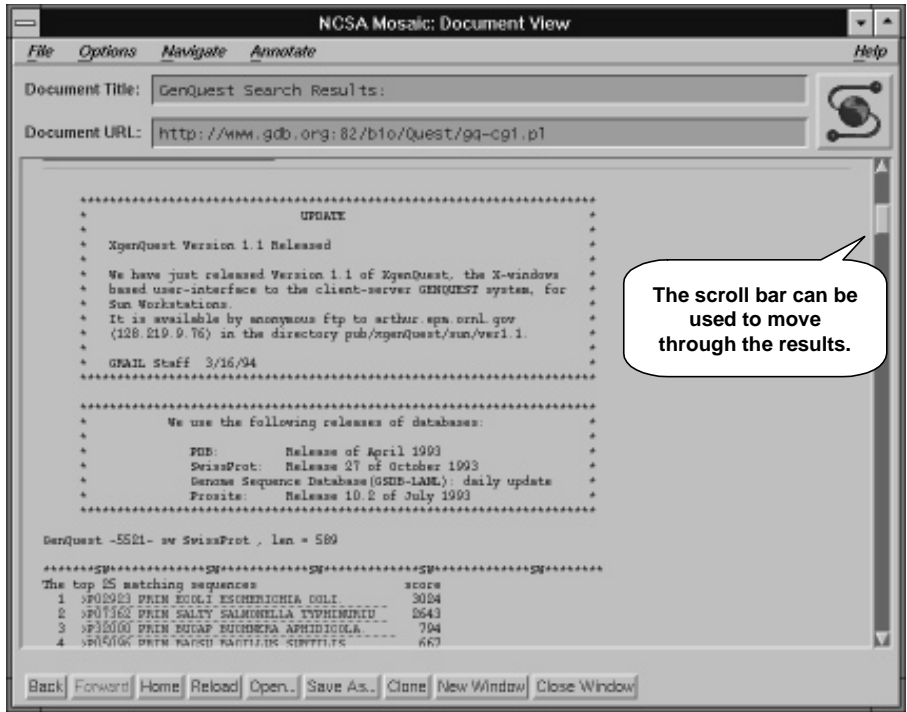
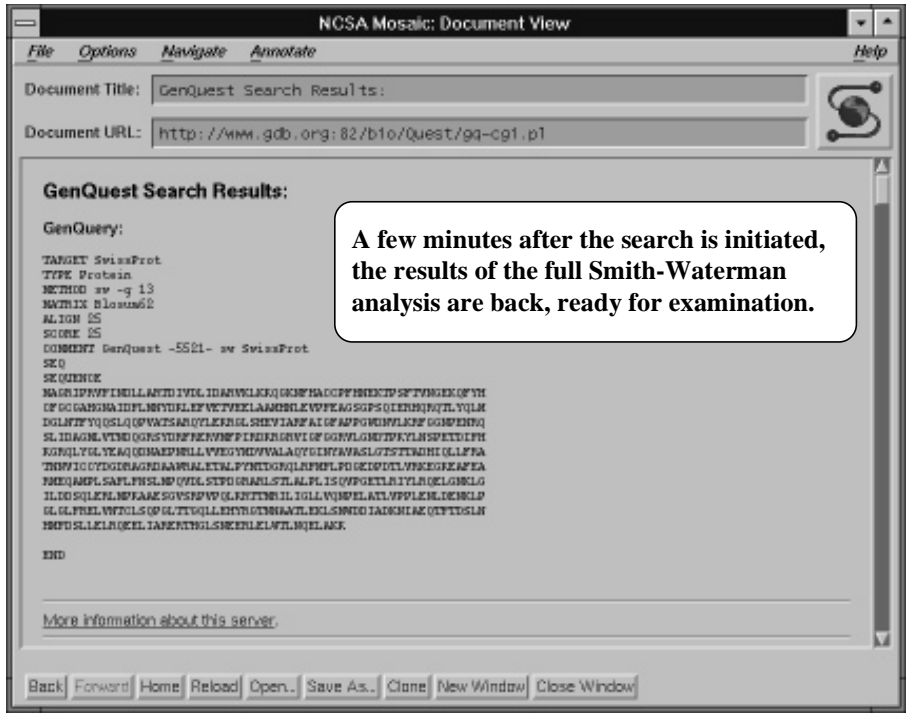
Figure 41. The graft operation can be used in the FOSM client to create new, virtual data structures composed of objects from different FOSM servers. A = human genes with mouse genes as attributes; B = mammalian homologies with both human and mouse genes as attributes; and C = mouse genes with human genes as attributes. A FOSM client should support object composition to an arbitrary depth.

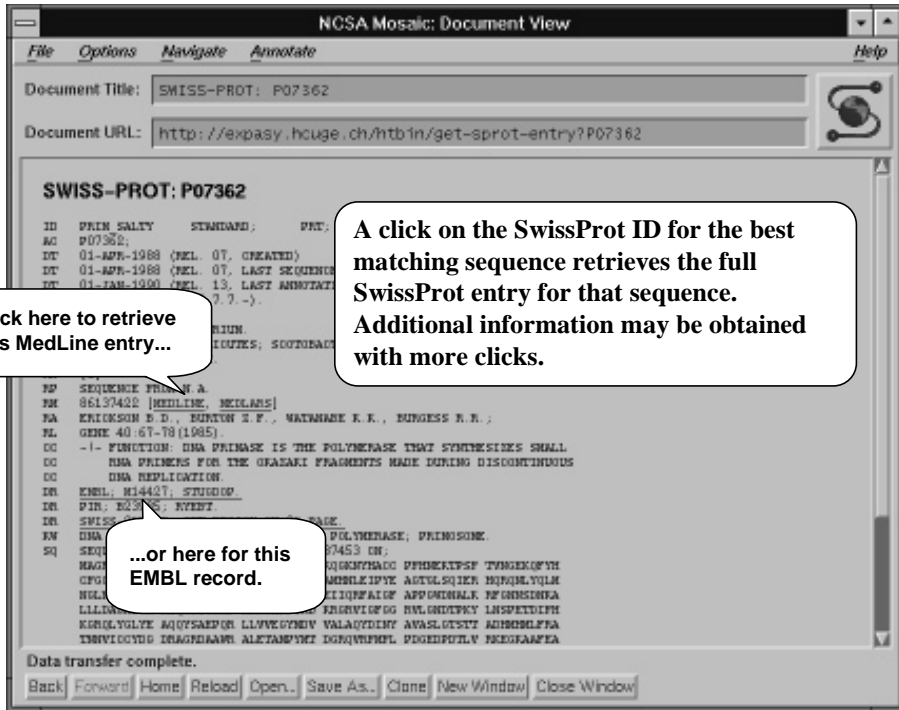
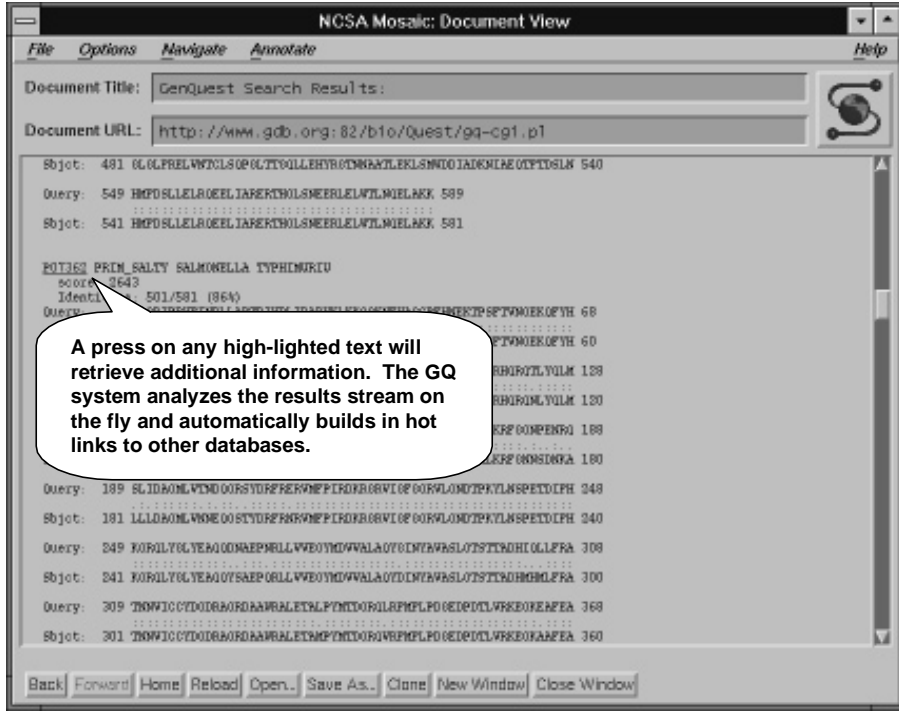
THE POWER OF MIDDLEWARE—THE EXAMPLE OF GENQUEST











NCSA Mosaic: Document View

File Options Navigate Annotate Help

Document Title: MEDLINE report for UID 86137422

Document URL: http://atlas.nih.nih.gov:5700/htbin/eF/entreznr?86137422

Gene 40: 67-78 (1985) [86137422]

Nucleotide sequence of the rpsU-dnaG-rpoD operon from *Salmonella typhimurium* and a comparison of this sequence with the homologous operon of *Escherichia coli*.

B. D. Erickson, Z. F. Burton, K. K. Watanabe & R. R. Burgess

In *Escherichia coli* the genes encoding ribosomal protein S21 (rpsU), DNA primase (dnaG), and the 70-kDa sigma subunit of RNA polymerase (rpoD) are contained in a single operon. These gene products are involved in the initiation of translation, DNA replication, and transcription, respectively. We have examined the homologous region in the closely related bacterium *Salmonella typhimurium* and have found that the same three genes are similarly organized. We have sequenced the DNA for this operon in *S. typhimurium* and have compared the (nt) nucleotide and amino acid (aa) sequences with *E. coli*. In the coding regions, the sequence conservation varies from extremely high for rpsU to moderate for dnaG with respect to both nt and aa sequence. In the noncoding regions, sequences thought to be important for the regulation of transcription are conserved, while other sequences are not conserved. aa differences in DNA primase and sigma are not randomly distributed and suggest regions that may be important for protein structure or function.

MeSH Terms:

- [Amino Acid Sequence](#)
- [Base Sequence](#)
- [DNA Replication](#)
- [DNA Restriction Enzymes](#)
- [Escherichia coli](#)
- [Genes, Bacterial](#)
- [Genes, Structural](#)
- [rpoD](#)

Back Forward Home Reload Open... Save As... Clone New Window Close Window

Here is the MedLine entry. Clicks on the MeSH terms will return their definitions...

NCSA Mosaic: Document View

File Options Navigate Annotate Help

Document Title: M14427 from EMBL

Document URL: http://expasy2.houge.ch/htbin/get-emb1-entry?M14427

EMBL: M14427

ID STU609 standard; DNA; 4631 BP.

XX

AC M14427;

XX

DT 16-JUL-1988 (Vol. 16, Created)

DT 04-SEP-1993 (Vol. 37, Last update)

XX

DE *S. typhimurium* rpsU-dnaG-rpoD operon

XX

FN .

XX

OS *Salmonella typhimurium*

OC Prokaryota; Bacteria; Gramicicutes; Scotobacteria;

OC Facultatively anaerobic rods; Enterobacteriaceae; *Salmonella*.

XX

XX [1]

FP 1-4631

FA Erickson B.D., Burton Z.F., Watanabe K.K., Burgess R.R.;

NT "Nucleotide sequence of the rpsU-dnaG-rpoD operon from *Salmonella typhimurium* and a comparison of this sequence with the homologous operon of *Escherichia coli*";

NT Gene 40:67-78 (1985).

XX

IN SWISS-PROT; P02379; NS21 ECOLI.

IN SWISS-PROT; P07336; NP70 SALT1.

IN SWISS-PROT; P07361; PR18 SALT1.

IN IPI; 001349; Release Y.D.

XX

XX

XX Key Location/Qualifiers

Back Forward Home Reload Open... Save As... Clone New Window Close Window

Here is the EMBL record. Clicks on any high-lighted text will retrieve additional information...

BIBLIOGRAPHY

- Anklesaria, F., McCahill, M., Lindner, P., Johnson, D., and Torrey, D. 1993. F.Y.I. on the Internet Gopher Protocol
- Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F., and Secret, A. 1994. The World-Wide Web. *Communications of the ACM*, 37(8):76–82.
- Brooks, Frederick P., Jr. 1982 *The Mythical Man-month*. Reading, Massachusetts: Addison-Wesley Publishing Company.
- Chorafas, D. N., and Steinmann, H. 1993. *Solutions for Networked Databases: How to Move from Heterogeneous Structures to Federated Concepts*. New York: Academic Press, Inc.
- Cinkosky, M. J., Fickett, J. W., Gilna, P., and Burks, C. 1991. Electronic data publishing and GenBank. *Science*, 252:1273–1277.
- Comer, D. E. 1988. *Internetworking with TCP/IP: Principles, Protocols, and Architecture*. Englewood Cliffs, NJ: Prentice-Hall. (A second edition of this book was published in 1991. A three volume treatment of the same topic has appeared in 1994, now with D. Stevens as co-author.)
- Committee on the Formation of the National Biological Survey. 1993. *A Biological Survey for the Nation*. Washington, D.C.: National Academy Press.
- Donoghue, M. J. 1992.. Homology. In: Keller, E. F., and Lloyd, E. A. (eds). *Keywords in Evolutionary Biology*. Cambridge, Massachusetts, Harvard University Press. pp. 170–179
- GeSTeC Directors. 1994. Report: NCHGR GeSTeC Director's meeting on genome informatics. Available electronically from Johns Hopkins WWW server, <http://www.gdb.org/Dan/nchgr/report.html>.
- Hurson, A. R., Bright, M. W., and Pakzad, S. H. (Eds.) 1994. *Multidatabase Systems: An Advanced Solution for Global Information Sharing*. Los Alamitos, California: IEEE Computer Society Press
- Lewin, R. 1986. DNA databases are swamped. *Science*, 232:1599.
- Murray, J. C., Buetow, K. H., Weber, J. L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V. C., Sunden, S., Duyk, G. M., Weissenbach, J., Gyapay, G., Dib, C., Morrisette, J., Lathrop, G. M., Vignal, A., White, R., Matsunami, N., Gerken, S., Melis, R., Albertsen, H., Plaetke, R., Odelberg, S., Ward, D., Dausset, J., Cohen, D., and Cann, H. 1994. A comprehensive human linkage map with centimorgan density. *Science*, 265:2049–2054.
- NRENnaissance Committee. 1994. *Realizing the Information Future: The Internet and Beyond*. Washington, DC: National Academy Press.
- Robbins, R. J. 1992. Database and computational challenges in the human genome project. *IEEE Engineering in Medicine and Biology Magazine*, 11:25–34.
- Robbins, R. J. 1993. Genome informatics: Requirements and challenges. In: Lim, H. A., Fickett, J. W., Cantor, C. R., and Robbins, R. J. (eds). *Bioinformatics, Supercomputing and Complex Genome Analysis*. Singapore: World Scientific Publishing Company.
- Robbins, R. J. 1994a. Biological databases: A new scientific literature. *Publishing Research Quarterly*, 10:1–27.
- Robbins, R. J. 1994b. Representing genomic maps in a relational database. In: Suhai, S. (ed). *Computational Methods in Genome Research*. New York: Plenum Publishing Company. (in press)
- Robbins, R. J. (Ed.) 1994c. Genome informatics I: Community databases. *Journal of Computational Biology*, in press.
- Robbins, R. J. 1994d. Genome Informatics: Toward a Federated Information Infrastructure (keynote address). The Third International Conference on Bioinformatics and Genome Research; Tallahassee, Florida; 1–4 June 1994.
- Robbins, R. J. 1995. *Data Publishing and the Global Information Infrastructure*, manuscript in preparation. Prepublication copies available for review from the author.
- Schatz, B. R., and Hardin, J. B. 1994. NCSA Mosaic and the World Wide Web: Global hypermedia protocols for the internet. *Science*, 265:895–901.

- Sheth, A. P., and Larson, J. A. 1990. Federated databases systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22:183–236.
- United States Department of Energy. 1990. *Understanding Our Genetic Inheritance. The U. S. Human Genome Project: The First Five Years*.
- United States National Academy of Sciences, National Research Council, Commission on Life Sciences, Board on Basic Biology, Committee on Mapping and Sequencing the Human Genome. 1988. *Mapping and Sequencing the Human Genome*. Washington, DC: National Academy Press.
- Vetter, R. J., Spell, C., and Ward, C. 1994. Mosaic and the World-Wide Web. *IEEE Computer*, 27(10):49–57.