Managing All Those Bytes: The Human Genome Project

A. Jamie Cuticchia, Michael A. Chipperfield, Christopher J. Porter, William Kearns, Peter L. Pearson

As we prepare for the next 5 years of the Human Genome Project, it is crucial to assess the contribution of informatics to the ultimate success of the project. For the enormous amount of information in the human genome to be useful (see figure), we must be able to access and manipulate it. Databases will provide this facility.

The three databases of primary importance to the Human Genome Project each store a different kind of information—DNA sequences (GenBank), chromosome mapping information (Genome Data Base), and protein sequence and structure (Protein Information Resource). Currently, these databases are independently administered and are separate physical entities, each with its own system for data collection, storage, and presentation. However, the community would be better served by the convenience of "one-stop shopping," provided by a seamless integration of these primary databases into a single "virtual database." This integration will necessitate adoption of a standard protocol such as SQL (Structured Query Language) for the interrogation and retrieval of related information simultaneously from several distributed databases. At present, data is accessed from databases across networks with simple document-based protocols such as GOPHER and WAIS (Wide Area Information Server). Although such protocols allow easy access to a wide variety of information (and undoubtedly will be used extensively), they do not provide the connectivity needed for a virtual database.

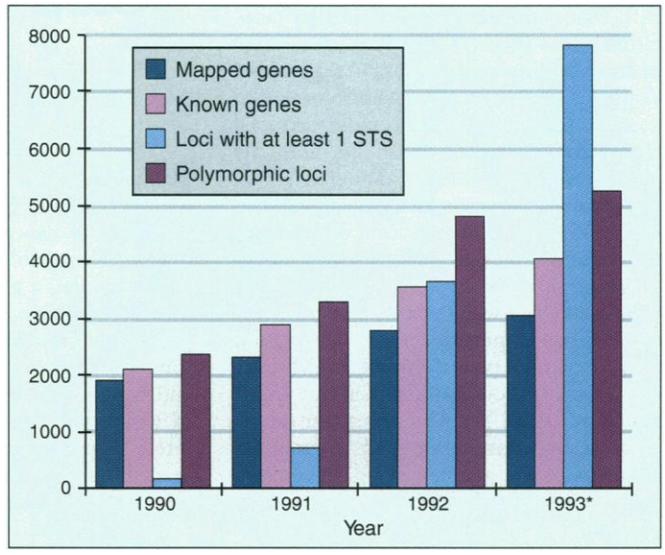
To set up a virtual database, linkages between the primary databases must be established and maintained. First, information synthesis is required to produce the link. The DNA sequence and protein databases can be easily linked through the translation of the genetic code and subsequent homology comparison. A link between the sequence and mapping database can be made by using the common sequence identity between the sequence record and sequences stored in the mapping database such as STSs (sequence tagged sites) or identity between shared attributes such as locus name, locus symbol, or probe symbol. This method requires stable, informative, and nonoverlapping nomenclature.

These simple database linkages can be performed by computer algorithms with little or no human intervention. However, in many cases there is an inconsistency between the shared attributes in the sequence and mapping records due to different nomenclature for mapping and sequencing. These cases already complicate the linkages between relatively well-characterized objects such as coding sequences. As large-scale anonymous sequences are produced from the genome program, it will become even more difficult to establish these links reliably unless researchers adhere to protocols for identifying mapped and sequenced objects. Thus, a newly cloned and characterized gene should be named by established naming conventions (*1*) and the name submitted to the Human Gene Mapping Nomenclature Committee for approval.

The second and somewhat less difficult component in the establishment of these linkages involves the mechanism by which the links are represented. In order for the links to be established and maintained, it is imperative that each of the linked databases has a stable and unique identifier (commonly referred to as an accession number) associated with their records. The links are denoted by the use of accession numbers as pointers between records in different databases. Additionally, the databases must work out a protocol for the exchange of the pairs of accession numbers involved in each linkage. In principle, this exchange can be performed automatically.

Although such linkages could, in theory, be expanded to include all of the data repositories collecting data from the Human Genome Project, this would not be advisable. A retrieval from this array of networked databases might require exchange of information among several of the data repositories, which could make timely on-line retrievals impractical. The unreliability of network connections worldwide would also adversely affect the ability to retrieve information. The practicality of data exchange between many repositories must also be considered. The exchange between any two repositories requires the establishment of a protocol or arrangement for that exchange. In the case of only 100 repositories this would require, if unregulated, each repository to establish an individual arrangement with the 99 other repositories, resulting in a total of 4950 protocols! Preferably, the responsibility of data collection should rest with a single primary database for each class of data so that individual data repositories need only establish a single protocol with the primary database.

Several genome centers and chromosome communities in the United States have recently established FTP (file transfer protocol) servers to provide the genome community with ready access to recently produced data. Unfortunately, there has been no coordination of the formats for the data placed on the FTP servers. Therefore a researcher wishing to access data, such as STS sequences including polymerase chain reaction (PCR) primer information, from multiple sources is confronted with the challenge of having to reformat each set of data to a single standard format for their own uses. However, if the data repositories were to standardize their formats, then the researcher need only utilize a single protocol for data retrieval from any of the FTP servers. Furthermore, if the data repositories were to transmit their data to the pri-



Growth of information from the Human Genome Project.

The authors are in the Department of Medicine at the Johns Hopkins University School of Medicine, 2024 East Monument Street, Baltimore, MD 21205-2100 and are affiliated with the Genome Data Base.

mary database, in addition to making it available on their own FTP servers, it would facilitate one-stop shopping and eliminate the need to query each data repository individually.

Primary databases should establish working relations with each of the data repository centers to ensure that newly produced data is transmitted to them quickly. In some cases, this will require working with the informatics group at the centers to output the data in a standard format. However, genome data is often generated in laboratories that do not have the luxury of dedicated individuals to assist in the data collection and cataloging. It becomes largely the responsibility of the primary database to provide the tools necessary to allow such groups to submit their data efficiently [for example, Authorin, a program developed to transmit data to GenBank (*2*)].

Genome-wide research is now in vogue in a few major centers, and this is resulting in the generation of large amounts of low-resolution mapping data. The challenge to the database is to integrate this information with that garnered by all the smaller laboratories who produce high-resolution maps of their favorite chromosomal region. For example, all the detailed map information derived from a particular yeast artificial chromosome (YAC) must be correctly placed on the low-resolution map of YACs. The primary mapping database must both store and allow integration of all information relating to that one YAC, irrespective of the original source of the data. Even if a virtual database was a practical proposition, the essential integration of information would not take place without a single authority taking responsibility for carrying it out. It is logical that the primary mapping database, in consultation with appropriate members of the genome mapping community, provide such integration.

Although it is in the interest of the community that all relevant mapping information be made available in a central fashion, it cannot be the function of the primary database to act as a "data police" for ensuring timely entry. Foremost responsibility must rest on the research community itself to promptly submit data so that the primary database increases in utility for the entire genome community. The requirements of scientific journals to have the data submitted to a primary database as a condition for publication of an article is a significant motivating factor for timely data submission. Furthermore, this relieves the journals from having to publish large amounts of sequence information that is largely unintelligible on a printed page but can be studied and analyzed extensively once in electronic form. A similar requirement for mapping information must also be established so that detailed information on, for example, probe descriptions, polymorphisms, and chromosome breakpoints are consigned to the primary mapping database and can simply be referred to in journal articles by their accession numbers. This mechanism would also allow retrieval of all information relating to one publication from the database in a single action, such as retrieving all the PCR primers and their associated polymorphic information from a particular linkage study.

Although it is the major directive of each primary database to collect data, maintain its integrity, and make data available in as many different formats as possible, the limited resources do not permit an unlimited number of access methods. Many tools for accessing and analyzing the data contained with the DNA sequence and protein databases have already been developed by both the research community and commercial enterprises. As increasing amounts of mapping information are accumulated in the primary mapping database, tools for accessing such data are already being developed by several groups throughout the world using sophisticated graphics for map displays. Although it will still be the responsibility of the primary mapping database to provide some tools to the public for accessing the data, it is also their task to provide information to the research community to assist in the development of new access methods.

References

1. M. J. Cinkosky, J. W. Fickett, P. Gilna, C. Burks, *Science* **252**, 1273 (1991).
2. T. B. Shows *et al.*, *Cytogenet. Cell Genet.* **46**, 11 (1987).

Presymptomatic Diagnosis: A First Step Toward Genetic Health Care

C. Thomas Caskey

Rapid progress is being made by the international human genome initiative in the discovery of genes responsible for human disease. The establishment of a genetic map of the human, a goal of the initiative, will provide medicine with the most rapid expansion of new knowledge in recent history. The application of this knowledge will begin a new era of molecular medicine, in which the risk of disease can be accurately assessed by DNA-based diagnostic procedures. Furthermore, disease pathogenesis and progression can be logically and efficiently studied, once the genes associated with a particular disease are known. The ability to detect individuals at risk for a disease prior to any pathologic evidence of the disease theoretically offers to medicine a new strategy—anticipation of disease and preemptive therapy. This vision will not be realized by the mapping and sequencing efforts of the Human Genome Project alone, but also requires study of gene function by disease specialists.

DNA-based diagnostics, coupled with the discoveries of new genes, can benefit health care in the short term by reducing the incidence of severe or presently untreatable diseases. Screening programs for Tay-Sachs and β-thalassemia initiated in the 1970s have reduced the incidence of these diseases by 20-fold (*1*). This remarkable success can be attributed to recognition of the requisite features of accurate diagnosis, education (public and professional), and freedom for individuals to make reproductive decisions in avoiding the disease in their families. The discovery in recent years of common disease genes, such as those causing cystic fibrosis (CF) (*2*) (1 in 3700) and fragile X syndrome (*3*) (1 in 1200), and their ancestral mutations or premutations, makes possible the expansion of genetic screening. The Ethical, Legal, and Social Implications (ELSI) component of the Human Genome Project within the National Institutes of Health is studying the feasibility and utility of population-based screening for CF in the United States. The first reports of this study will be available in the fall of 1993. The experiences from Tay-Sachs, β-thalassemia, sickling hemoglobinopathies, and CF screening should point the way for effective and acceptable broader usage of genetic screening in other common heritable diseases, such as Gaucher disease, α_1-antitrypsin deficiency, myotonic dystrophy, and spinal muscular atrophy (*4*). Couples at risk for disease in their offspring have had increased options since the 1970s for reproductive planning, including prenatal diagnosis, preimplantation diagnosis of embryos, artificial insemi-

The author is an investigator in the Howard Hughes Medical Institute and Henry and Emma Meyer Professor in the Institute for Molecular Genetics, One Baylor Plaza, Baylor College of Medicine, Houston, TX 77030.