Perspective

The Human Genome Initiative—Do Databases Reflect Current Progress?

P. L. PEARSON, B. MAIDAK, M. CHIPPERFIELD, R. ROBBINS

THE TWO MAJOR COMPONENTS OF THE HUMAN GENOME Initiative are concerned with defining the map location of map objects such as genes, anonymous DNA sequences, and polymorphic markers, and defining their associated DNA sequence composition. Estimates given in the accompanying wall chart on the number of objects currently mapped to the human genome and the number of genes and proportion of the genome that has been sequenced are based on the content of two databases, namely, the Genome Data Base (GDB) at Johns Hopkins University (*1*) for the map information and GenBank at Los Alamos (*2*) for the DNA sequence information. It is pertinent to ask whether the contents of these databases reflect the true proportion of the human genome that has been mapped and sequenced, and whether the information provided constitutes an "adequate" level of completeness.

DNA sequence information has been assembled mainly in GenBank and an associated database, the European Molecular Biology Laboratory (EMBL) Data Library (*3*), for the last decade, and the tradition has been established of submitting sequence information to the databases either at the time of or prior to publication. More recently, submissions have been made to the databases without any intent of publication, that is, in the form of personal communications. Indeed, many journals no longer accept DNA sequence information for publication and are only prepared to publish comments or annotations on sequences directly submitted, and made available through an appropriate database. The major means of data sharing and communication will probably be through electronic data publishing (databases) in future and not through the printed word (*4*). The peer review process will generally occur after data entry; the data will be checked for accuracy and consistency with other database entries once it has been added to the database. Thus, as with traditional publication, "publication" status is achieved only after data has passed the validation and review procedures used by the database. A further analogy with traditional publication is that the responsibility of submitting information to the database in a timely fashion lies with the data generators and not the database managers. The community is already moving toward this situation in the case of sequence information and is likely to adopt a similar attitude for mapping information in the near future; in both cases this trend is driven by the mass of information that can no longer be published in a traditional form.

Because GenBank now accepts direct submissions in the form of electronic files, its content is probably more representative of the total amount of human DNA sequenced than information based solely on extractions from printed articles. In addition, direct electronic submissions should be less prone to transcriptional errors than those entered via the printed word. This in no sense implies that published sequences should not be considered. There are, however, certain provisos in using the information from GenBank to establish the proportion of the total genome and the number of human genes that have been sequenced. For example, each length of sequence is submitted to GenBank as an individual entry and may duplicate other sequence information already present. The figures present in the wall chart for the amount of sequence information on individual chromosomes have not been corrected for such possible duplications and in that sense are probably overestimates. Another problem encountered in attributing sequences to individual chromosomes or chromosome regions is the failure of sequence contributors to use a standardized gene nomenclature (*5*) in describing or annotating the DNA sequences or to include map information where that is known. Only when the identity of a gene has been established by the use of an officially designated gene symbol and name, is it possible to optimally link the sequence to its map location in a mapping database. Notwithstanding these drawbacks, the current sequence information within the public databases is a reasonable representation of the total amount of human genome sequenced to date, albeit for expressed sequences.

However, as the Human Genome Initiative gathers momentum and the proportion of nonexpressed sequences that have been mapped and sequenced increases, the problems of keeping track of sequence identity and map position are going to increase enormously. It will be extremely important to create efficient links between the sequence information in GenBank and EMBL and their map location as represented within GDB. Currently such links do not exist. A consequence of this is that it is not possible to retrieve data from the sequence database by using map location as one of the search criteria. The DNA sequence totals presented on the wall chart for each chromosome have had to be assembled by hand. Possible differences in the selection criteria have led to the anomalous situation that the sequence totals given for some chromosomes this year are less than those described by Stephens *et al.* (*6*) last year, despite the fact that the total amount of human sequence in GenBank has doubled approximately in the last year, and the total amount of mapped sequence this year (approximately 6.4 Mb) has increased substantially. This implies that these estimates are not really representative of the absolute amount of sequence data on each chromosome. However, the relative distribution of sequence data between chromosomes is probably realistic.

The situation with regards to interpreting the proportion of the human genome mapped is far less clear than for sequence information for various reasons. First and foremost is that the tradition has not yet been established of submitting map data directly to public databases, and much of the information currently resides solely within private databases or, worse still, solely within laboratory notebooks. Secondly, the level of detail on currently available map information is sparse. For example, GDB contains map information on approximately 2300 coding sequences, the majority of which are mapped only to an entire chromosome or chromosome band. However, few would agree that the localization of genes to individual cytogenetic bands is of sufficient resolution to be of use in detailed studies on genome structure. The National Advisory Council For Human Genome Research has provided a suitable definition (*7*) of an adequate level of mapping information as part of its recommendation in generating a map of sequenced tagged sites (STSs). STSs are regions of DNA for which pairs of polymerase chain reaction (PCR) primers have been defined. The Council's plan recommends generating markers at approximately 100-kb intervals as one of the goals to be achieved within the first 5 years.

On the basis of the above definition of "adequate" map resolution, we may conclude that very few of the existing map locations of genes and anonymous DNA sequences included in the current GDB entries are adequately defined. On the other hand, some information at this level of detail is present in the databases of individual centers

P. L. Pearson is Director of the Genome Data Base, B. Maidak is a Research Associate, M. Chipperfield is DNA Data Coordinator, and R. Robbins is Director of the Welch Lab for Applied Research in Academic Information, Johns Hopkins University School of Medicine, Baltimore, MD 21205.

that have concentrated on the physical or genetic mapping of particular chromosomes. For example, the Human Genome Centers of the Department of Energy at Los Alamos and Lawrence Livermore have been working on the physical mapping of chromosomes 16 and 19, respectively, and both groups have now mapped more than 60% of their chromosomes at the contig level (*8*). This type of map information has not been available to GDB until recently, and consequently the information is not included in the wall chart estimates. One of the major challenges over the coming year is going to be developing the means of directly entering diverse types of map information into GDB from other databases. Recently, GDB has collaborated with the Lawrence Livermore Center to develop means of transferring information from their chromosome 19 repository into GDB by direct database-to-database communication.

Currently GDB provides a forum in which consensus map information can be stored and retrieved. All maps are stored within GDB according to a standard format. For example, linkage and contig maps are both defined by map objects (such as genes or anonymous DNA segments) arranged according to order and distance. The only essential difference between one type of map and another is the unit of distance used; in the case of linkage maps, centimorgans are used, whereas contig maps are based on kilobases. A map grammar has been developed permitting entry of map information as simple alpha-numeric strings that indicate whether map objects are ordered, grouped, overlapping, or contained within other map objects, as well as the associated distance parameters. This system permits map information to be stored and represented identically over all chromosomes irrespective of the origin and method of mapping used.

Over the last year GDB has experienced a modest increase (5%) in the number of genes and anonymous DNA sequences mapped. For the reasons advanced above we believe this to be an underestimate. The development of appropriate means of direct electronic entry from centers or chromosome-specific databases will result in a rapid rise in the content of the database until a steady state has been achieved between the rate of data generation and entry into GDB. Because of the restructuring currently taking place within the Human Genome Initiative in terms of establishing genome centers and other groups organized around specific chromosomes, we can expect that a steady state will not be achieved within the next 3 years. Until a steady state is reached, any estimates on the state of mapping of the human genome will be prone to high levels of uncertainty.

The last year has also witnessed a distinct change in the types of genetic markers present in the database, with a clear shift toward inclusion of nucleotide repeat polymorphisms with polymerase chain reaction (PCR)–derived probes or target DNA. At present more than 200 loci include known simple repeat polymorphisms, most of which are dinucleotide repeats. Repeat polymorphisms are particularly useful for genetic mapping, because the large number of alleles (usually >5) increases their information content. GDB currently contains approximately 1000 primer pairs for generating PCR-based probes and, once again, this is an area where the database content does not reflect the amount of information available within the genome community. Conservative estimates based on discussions with representatives from genome centers and other groups suggest that information on several thousand PCR-based probes or STSs exists within the community. It is essential that this information be made widely available as soon as possible through submission to a centralized database.

In a review of the status of the human gene map made one year ago, Stephens *et al.* (*6*) based their estimates of the completeness of the human gene map on assumptions that gene density is equal across all chromosomes or chromosome regions and that the DNA content of each band is proportional to band lengths depicted in the ISCN banding nomenclature report (*9*). At best these can only be very rough approximations because coding sequence content is not a simple function of chromosome or band size and because the ISCN band sizes, as depicted, are not based on quantitative measurements. McKusick has recently pointed out the apparent discrepancy between the number of genes expected on a chromosome on the basis of size and equal gene density and the number now emerging from gene mapping observations (*10*). One of the clearest examples is chromosome 19, which seems to have a much higher density of genes than its closest size neighbors. We must be a little circumspect in overinterpreting this result, since chromosome 19 has also received a great deal of attention from gene mappers. However, when this result is compared to those for other chromosomes, including chromosomes 22, 18, 17, 13, and 11, the overall pattern of gene density is compatible with GC-rich chromosomes or chromosome regions having an increased density of coding sequences and AT-rich chromosomes or regions being relatively gene poor. Chromosomes 13 and 18 show the least gene density and are two of the only three chromosomes resulting in trisomic states compatible with life in man.

Mapping and sequencing the entire human genome in a timely fashion requires organization of all available resources to the common goal. Federal funding agencies have established individual genome centers that will focus on one or more chromosomes. Further, chromosome-specific workshops are being organized to permit individual centers, researchers, or groups to pool their results with other colleagues working on the same chromosome. These activities imply the following: (i) each chromosome community should have its own database; (ii) the databases should permit inclusion of data from many different groups and give different map interpretations of the same chromosome region; and (iii) similar formats for data storage and representation should be used across the databases to simplify data exchange and interpretation.

However, no matter how sophisticated modern database management systems may be, they cannot realistically fulfill their responsibilities until all parties concerned are prepared to submit their data to centralized databases. To do this they need to be provided with adequate tools and incentives. Provision of the tools is the task of the database organizations. Provision of incentives is partly a question of adequate peer recognition for direct submission, partly a willingness to openly share information with the community at large, and partly the need for funding organizations to insist on data sharing as a requisite for their continued support.

REFERENCES AND NOTES

1. P. L. Pearson, *Nucleic Acids Res.* **19** (suppl.), 2237 (1991).
2. C. Burks *et al.*, *Comput. Appl. Biosci.* **1**, 225 (1985); C. Burks *et al.*, *Methods Enzymol.* **183**, 3 (1990).
3. P. Kahn and G. Cameron, *Methods Enzymol.* **183**, 23 (1990).
4. M. J. Cinkosky, J. W. Fickett, P. Gilna, C. Burks, *Science* **252**, 1273 (1991).
5. T. B. Shows *et al.*, *Cytogenet. Cell. Genet.* **46**, 11 (1987); H. F. Willard, *ibid.* **40**, 360 (1985); P. L. Pearson *et al.*, *ibid.* **46**, 390 (1987); K. K. Kidd *et al.*, *ibid.* **51**, 622 (1989).
6. J. C. Stephens *et al.*, *Science* **250**, 237 (1990).
7. U.S. Department of Health and Human Services and U.S. Department of Energy, *The U.S. Genome Project: The First Five Years FY 1991–1995* (National Technical Information Service, Springfield, VA, 1990), pp. 11–14.
8. E. Branscomb *et al.*, *Genomics* **8**, 3511 (1990); A. V. Carrano *et al.*, *Genome* **31**, 1059 (1989); R. L. Stallings *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 6218 (1990); R. L. Stallings, personal communication.
9. D. G. Harnden and H. P. Klinger, Eds., *An International System for Human Chromosome Nomenclature (1985): Report of the Standing Committee on Human Cytogenetics Nomenclature* (Karger, New York, 1985).
10. V. A. McKusick, *FASEB J.* **5**, 12 (1991).