Enhancement of the Integrated Database "HyperGenome" for Genome Maps and Sequence Information

Takahiko SUZUKI[1] Susumu NAKASHIMA[2] Toshihisa TAKAGI[1]
suzuki@ims.u-tokyo.ac.jp susumu@grt.kyushu-u.ac.jp takagi@ims.u-tokyo.ac.jp

Satoru KUHARA[2] Minoru KANEHISA[1 3]
kuhara@grt.kyushu-u.ac.jp kanehisa@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1, Shiroganedai, Minato-ku, Tokyo 108, Japan

[2] Graduate School of Genetic Resources and Technology, Faculty of Agriculture, Kyushu University, 6-10-1, Hakozaki, Fukuoka 812, Japan

[3] Institute for Chemical Research, Kyoto University Uji, Kyoto 611 Japan

Abstract

An integrated database system "HyperGenome" for genome maps and DNA sequences was developed. The system can handle two different types of data, each of which has an unique complex structure. Graphical user interface (GUI) enables ready retrieval of information obtained from genome mapping data and data on DNA sequences. Data on mapping are derived from the Genome Data Base (GDB) and sequence data are from GenBank.

The following information was added to the system. 1. Mendelian Inheritance in Man(MIM) entries can be linked to a locus in our system. 2. Amino asid sequences from Protein Identification Resources(PIR) can be displayed, in conjunction with the nucleotide sequence.

[1]東京大学医科学研究所ヒトゲノム解析センター, 〒108 東京都港区白金台 4-6-1
[2]九州大学大学院農学研究科遺伝子資源工学専攻, 〒812 福岡市東区箱崎 6-10-1
[3]京都大学化学研究所, 〒611 京都府宇治市五ヶ庄

1 Introduction

The Human Genome Project is rapidly accumulating various kinds of data on the genome. Genome is a term used to refer to all of the genes carried by a single gamete, that is, by a single representative of each of all chromosome pairs. The objective of the Human Genome Project is to acquire knowledge on the structure, function and evolution of the human genome. Databases were developed in order to organize data obtained during experiment. The following is a list of well known databases:

- Databases for nucleotide sequences of DNA (GenBank[1], EMBL[2], DDBJ).
- Databases for amino acid sequences of proteins(PIR, Swissprot [3, 4]).
- A database for crystallographic structures of proteins (PDB[5]).
- A database for physical and genome maps of genes (GDB[7]).
- A database for bibliographic information (MEDLINE).

These databases had been developed and maintained independently. It can be difficult for a sometime user to merge related information stored in different databases, the user needs two different databases (GDB and GenBank) when studying functions around a locus in GDB and wants to retrieve a DNA sequence data referenced in GDB.

Several systems have been developed to merge related information present in different databases. Entrez[9] integrates biological information in MEDLINE with DNA and sequence information of DNA and Protein. ODS [8] is designed to integrate sequence and bibliographic data, as well as data on functions of proteins, etc.

We developed a database system called "HyperGenome" which integrates data on genome mapping and DNA sequences.

The HyperGenome system is an user friendly integrated interface for a database of genome maps and DNA sequences. We assume that the average users of the HyperGenome will not enter their own data into the database, rather they retrieve information which is currently stored separately in genome mapping databases and/or in DNA sequence databases. In the HyperGenome System, the user can retrieve information related to a locus by directly pointing to the locus on a chromosome displayed on a window.

Two enhancements in the HyperGenome are outlined here. 1. Mendelian Inheritance in Man(MIM) is a catalog of human genes and genetic disorders. MIM number is included for all GDB loci with a corresponding MIM entry. The MIM database is incorporated into the HyperGenome. 2. Amino acid sequences from PIR are registered in HyperGenome, on the basis of GenBank accession number included for PIR.

2 System Overview

Fig. 1 shows organization of the HyperGenome system. Data are organized around loci data from the genome mapping database (GDB). Graphic images of chromosomes (chromosome windows) are generated from the locus data. A user of the HyperGenome selects a locus in the chromosome window and retrieves information related to the locus.

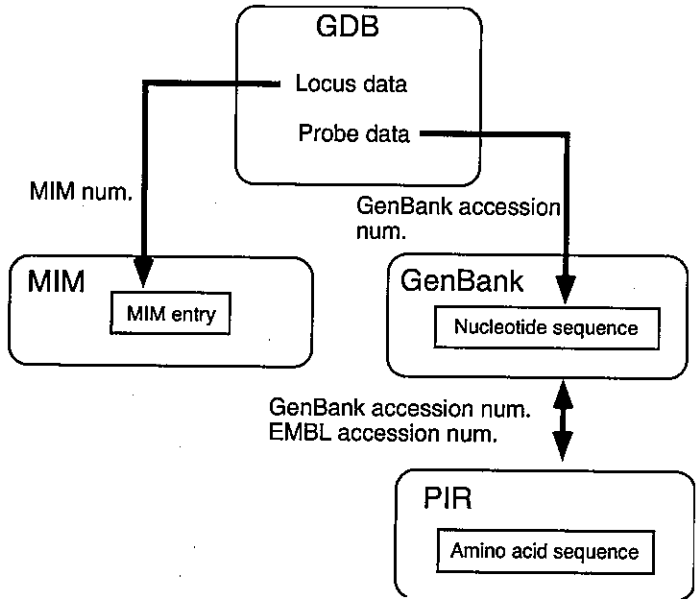

Figure 1: Organization of the HyperGenome system

DNA sequence data are imported from GenBank. Sequence data are retrieved using accession numbers (a unique identifier assigned to each GenBank entry) from cross reference fields in GDB. The cross reference between nucleotide sequence and amino acids sequence is generated from a GenBank accession number included for PIR cross-reference. MIM information is linked by MIM number(a unique identifier assigned to each entry) included for GDB loci.

The HyperGenome is currently developed on SUN Sparc series workstations running SUN OS 4.1.X and OpenWindows Ver.2. There are two versions of the HyperGenome which use different data management methods for data on genome mapping. One version uses an internal database management routine while the other has a Sybase relational database interface and manages the genome mapping data in relational tables.

2.1 Function and Graphical Interface of the HyperGenome

Investigators can use the HyperGenome with no difficulty because its functions are tightly coupled with its graphical user interface. The HyperGenome has the following functions:

Selection from the Chromosome Window

The chromosome window (**Fig. 2**) is the main window of the HyperGenome. The user can select a chromosome with the chromosome selection dialog. A locus can be selected using the mouse. Other functions are assigned to menu buttons on the window.

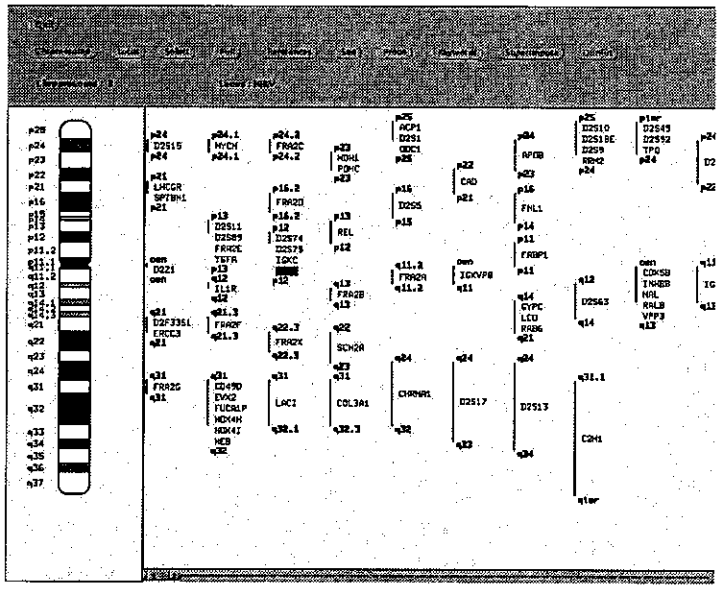

Figure 2: Chromosome Window

The HyperGenome supports retrieval of a locus in genome maps even when only the locus identifier (Locus symbol in GDB) is known. Locus symbol and name are listed in the locus listing box . After selecting a locus in the list, the chromosome window and other windows will display the related data on maps.

Keyword Search

The HyperGenome has a keyword search facility, as collected from gene symbols, gene and locus names, titles of references, probe names, and GenBank accession numbers. The list of hit loci for a given keyword is shown in the search result listing box. The investigator uses a mouse to retrieve a locus.

Retrieval of Data

The HyperGenome system comes equipped with four windows for displaying data related to the selected locus, these are reference window, sequence window (**Fig. 3**), probe window (**Fig. 4**)and MIM window(**Fig. 5**) .

The reference window contains bibliographic data imported from GDB. The sequence window shows data on DNA sequences derived from the GenBank. There may be more than one entry in the sequence window because the relationship between map loci and DNA sequences exceeds one. The sequence windows(GenBank and PIR) have three optional menu buttons(Fig. 3). The next button in the GenBank window shows the next GenBank entry linked to the loci,

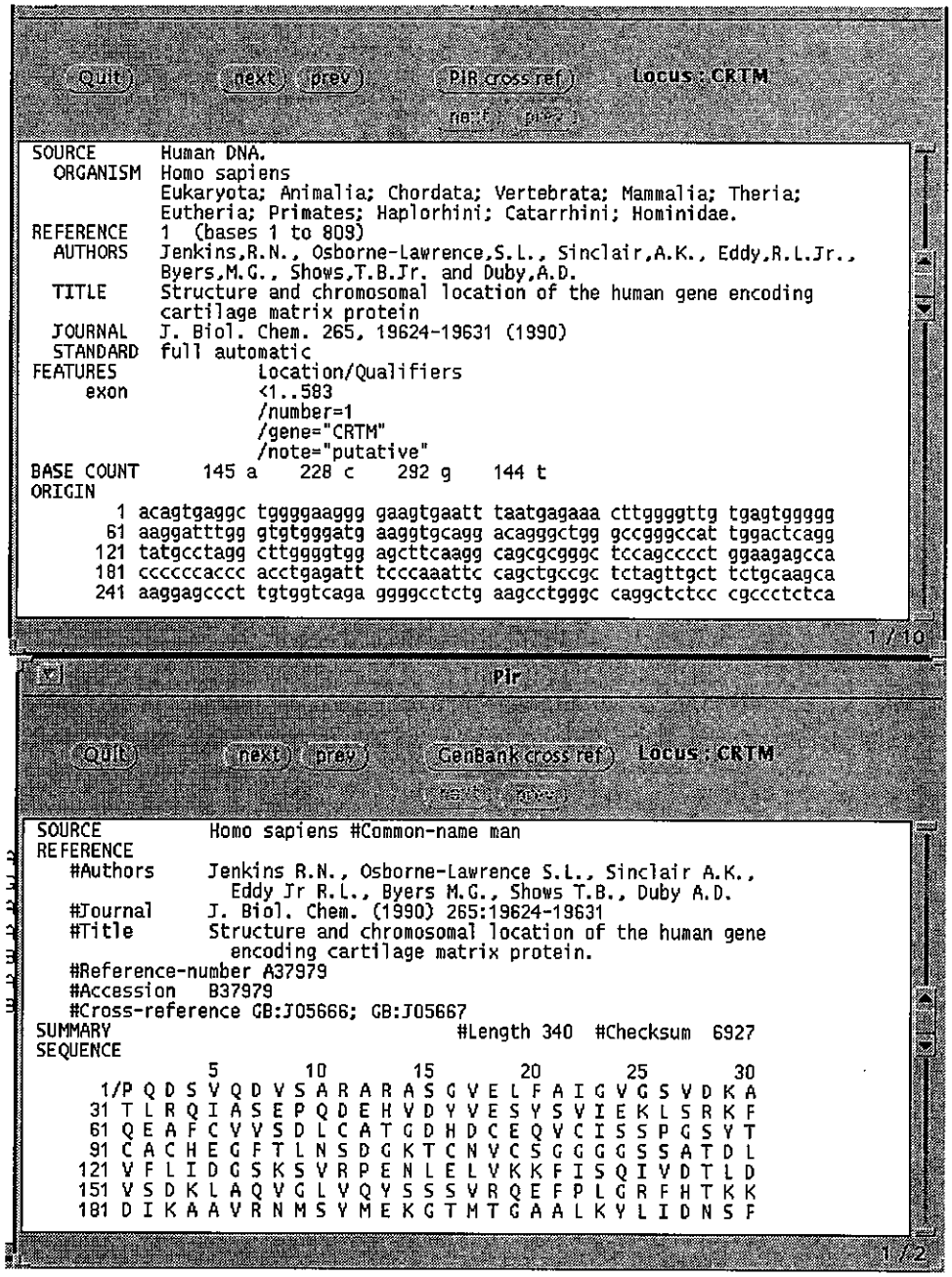

Figure 3: Data in the Sequence Window

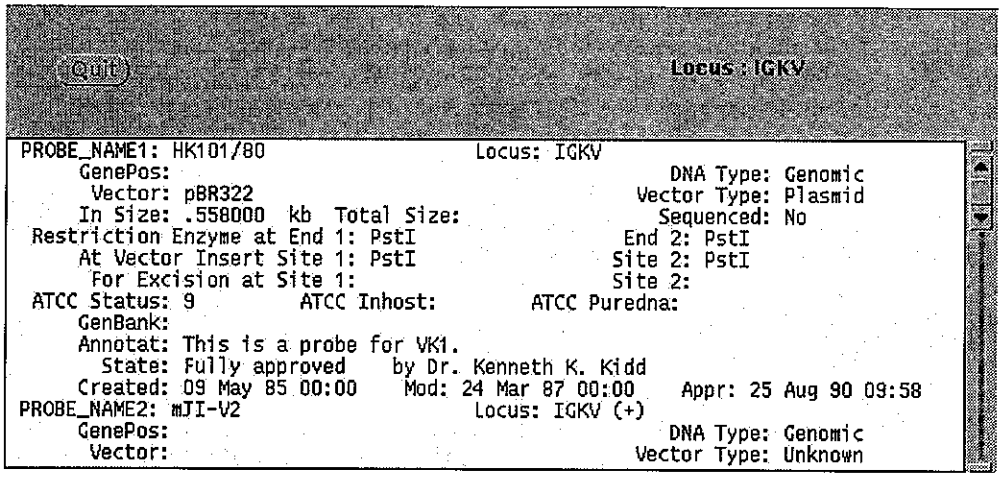

Figure 4: Data in the Probe Window

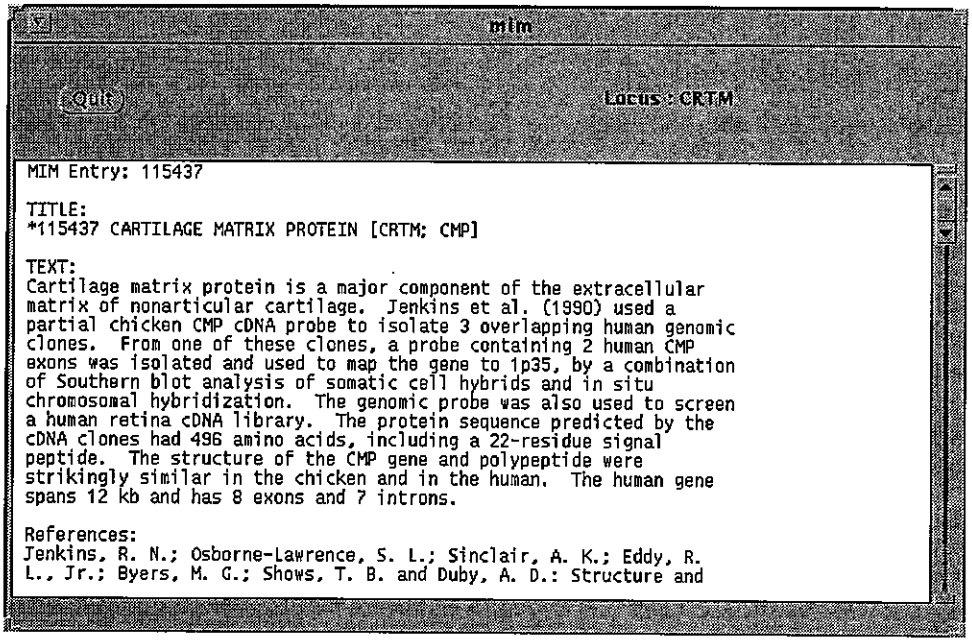

Figure 5: Data in the MIM Window

if many sequence entries are assigned to one loci. The PIR cross ref.button in the GenBank window shows the coded amino acid sequence in the protein window.

The probe window shows information on the probe, PCR and ASO linked to the locus symbol. The MIM window indicates the MIM entry related to the loci.

Superimposition of Maps

When one chromosome can be superimposed on another one, differences between related chromosomes of different species can be readily discerned.

3 Discussion

The HyperGenome contains over 14,000 gene symbols and DNA fragments for humans and mice, from GDB. DNA sequences associated with the fragment are from the GenBank. The User of the system can readily retrieve information related to a locus, using the graphic interface. Several database systems which manage genome mapping data have been developed. The SIGMA (System for Integrated Genome Map Assembly)[10] can manipulate and retrieve map data using GUI. Acedb (A C. elegans Database)[6] also manipulates map data through a similar GUI. Sequence data can also be retrieved by using acedb. The objective of these programs is to build data on mapping and to manage mapping projects, the HyperGenome integrates existing maps and sequence data. Features of systems lacking in the HyperGenome are:

- Map building potential.
- Zooming display of data on maps.

HyperGenome is not a system for Map building. Zooming is necessary in retrieval as the maps become more detailed and dense.

Both GDB and GenBank are updated frequently. The database of the HyperGenome can be updated by importing data from GDB and GenBank, which is relatively easy because the DNA sequence database in the HyperGenome is similar to the text file format distributed by GenBank.

GDB distributes its genome mapping database as dump files from relational tables. One might think that the import of GDB data into the HyperGenome could be done automatically, using scripts that convert dump files to the internal database format, however, there are at least two weak points in the conversion script method:

- The GDB database schema can change when the GDB system is updated, this happened recently when the GDB version 5.0 was released.
- Incremental update is not possible because relational tables are not independent. Even subtle changes in the data of relational tables can produce a completely different output database.

Acknowledgements

We thank Ms M.Ohara for valuable advice on the presentation. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Informatics', from the Ministry of Education, Science and Culture of Japan.

References

[1] Burks, C., Cinkosky, M. J., Fischer, W. M., Gilna, P., Hayden, J. E.-D., Keen, G. M., Kelly, M., Kristofferson, D. and Lawrence, J., "GenBank," Nucleic Acids Research, 20, pp. 2065–2069(1992).

[2] Higgins, D. G., Fuchs, R., Stoehr, P. J. and Cameron, G. N., "The EMBL Data Library," Nucleic Acids Research, 20, pp.2071–2074(1992).

[3] Barker, W. C., George, D. G., Mewes, H.-W. and Tsugita, A., "The PIR-International Protein Sequence Database," Nucleic Acids Research, 20, pp.2023–2026(1992).

[4] Bairoch, A. and Boeckmann, B., "The SWISS-PROT protein sequence data bank," Nucleic Acids Research, 20, pp.2019–2022(1992).

[5] Bernstein, F. C. et al., "The Protein Data Bank: A Computer-based Archival File for Macromolecule Structures," Journal of Molecular Biology, 112, pp.535–542(1977).

[6] Durbin, R. and Thierry-Mieg, J., "acedb A C.elegans Database I. Users' guide," (1993).

[7] Pearson, P. L., Matheson, N. W., Flescher, D. C. and Robbins, R. J., "The GDB™ Human Genome Data Base Anno 1992," Nucleic Acids Research, 20, pp.2201-2206(1992).

[8] Sakamoto, N., Takagi, T. and Sakaki, Y., "Development of the Overlapping Oligonucleotide Database and its application to Signal Sequence Search of the Human Genome," in Computer Applications in Biosciences (CABIOS), 9, pp.427-434(1993).

[9] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, "Entrez User's Guide," (1993).

[10] Cinkosky, M. and Fickett, J., Human Genome Information Resource, Theoretical Biology and Biophysic Group, Center for Human Genome Studies, Los Alamos National Laboratory, "S.I.G.M.A System for Integrated Genome Map Assembly," (1992).