情 報 学 基 礎 23－3
（1991. 9. 10）

Xウインドウの上で動くヒト遺伝子マップデータベースの構築

蓑島 伸生[1] 土肥 浩[2] 石塚 満[2] 清水 信義[1]

[1]慶應義塾大学医学部分子生物学教室
[2]東京大学生産技術研究所

長大なヒトゲノムの解析研究には、様々なタイプのマップが必要である。現在マップデータは、ワークステーション上のデータベースGDBとして集積されている。その作業及びそのための世界共通のソフトウェアは大変重要であるが、そういった集積システムとは別の専用データ利用システムを作成することにも大きな意義がある。我々は、汎用のXウインドウの上で動く、優れたユーザインタフェースと高速なデータ検索機能を持つ遺伝子マップライブラリシステム(JHGM/X)を開発している。このシステムでは、データも日本語で記述しており、研究者をはじめとする医学・生物学のあらゆる人々に使い易いものである。今回、JHGM/Xのプロトタイプについて報告する。

Construction of Human Gene Mapping Library System on X-Window

Shinsei Minoshima,[1] Hiroshi Dohi,[2] Mitsuru Ishizuka,[2] Nobuyoshi Shimizu[1]

[1] Department of Molecular Biology, Keio University School of Medicine
[2]Institute of Industrial Science, University of Tokyo

Construction of various types of map is important for facilitating the human genome project. Currently, mapping data is compiled in the GDB (Genome DataBase) on a workstation. Development of softwares to utilize the mapping data is also important. We have been constructing a human gene mapping library system, JHGM/X. The JHGM/X operates on X-Window, is equipped with an excellent searching capability, and offers a good user interface. It handles mapping data in Japanese. A prototype of JHGM/X runs on a SPARC Station(SUN4.)

1 はじめに

ヒトのゲノムは22対の常染色体とX,Y性染色体の計24種類46本の染色体から構成され、それら染色体DNAの上に5万〜10万の遺伝子が並んでいる。染色体DNAの長さは最長の第1染色体が2億4900万塩基対(249Mb)、最短の第21染色体が4800万塩基対(48Mb)であり、ハプロイドゲノムのDNAの全長は実に30億塩基対(3000Mb)にもおよぶ[1]。しかし、現在までに染色体にマップされた遺伝子の総数は3000に満たず、塩基配列が決定されたcDNAやゲノムDNAの総塩基対数も全ゲノムの1%に達していない。数年来全世界で論議され、最近欧米や日本で実際に開始されている「ヒトゲノム解析プロジェクト」は、そのような長大なゲノムを根こそぎ解析して多くの未知の遺伝子や染色体の機能ドメインを明らかにし、医学やヒトの遺伝学、分子生物学に役立てようという壮大なものである[2]。

ヒトゲノムの解析研究には様々な方法が用いられるので、既に蓄積しているデータにも今後出て来るデータにも色々なタイプのものがある。上述のように染色体DNAは長大なので、ゲノム研究を進めるにはそれらの様々なタイプのデータをマップの形に表わし、異なるタイプのマップを適宜利用していく必要がある。データ量も膨大になってきている。そのようなデータやマップの蓄積・検索・利用には、計算機を用いたデータベース及び専用のデータベース利用ソフトウェアが必須である。我々は、様々なタイプのデータを扱える総合的な専用データベースソフトウェアの開発を目指している。本報告ではその一環として作成している日本語ヒト遺伝子マップデータベースについて述べる。

2 ゲノム解析で用いられるマップの種類とデータベースの現状

ゲノム解析で用いられる主なマップを挙げる。

1. 遺伝子マップ
   染色体を特殊な方法で染色すると染色体の種類に特異的な縞模様状の染色を行なうことができる。これを染色体バンドと呼ぶ。染色体バンドに番号をつけて、位置を表す番地として用いている。この番地を用いて遺伝子の存在位置や、染色体異常の場所を表すことができる。このような地図を遺伝子マップという。

2. リンケージマップ
   同じ染色体上で離れている遺伝子ほど減数分裂の時に組換えを起こし易い。大きな家系の構成員の遺伝形質を調べて組換え頻度を計算することにより遺伝子間の相対距離及び遺伝子の並びの順序を決めることができる。このようにしてつくられるマップをリンケージマップと呼ぶ。

3. フィジカルマップ
   染色体DNAの上に存在する制限酵素認識部位等の特殊な配列を位置の基準として、DNA断片の整列順序を決定し、その上で遺伝子等の物理的位置を決定してつくられるマップがフィジカルマップである。

これらのマップについては、毎年ヒト遺伝子マッピング国際ワークショップ(HGMW)でデータが集積・吟味されてデータベース化されGDB (Genome DataBase)に収められる。GDB[3]はヒトゲノムに関する唯一の世界共通のデータベースであり、コンピュータネットワークを通して全世界からアクセスできる。GDBは関係データベースソフトウェアSybaseの上で作られたキャラクターベースのデータベースで、主な内容は遺伝子マップである。最近のversion upによりリンケージマップやフィジカルマップの入力も行えるようになり、本年8月にロンドンで行われたワークショップHGM11でそれらの機能も実際に使われ始めた。しかし、現在のところGDBは現場でデータを整理したり、視覚的にマップを利用していくためのものでなく、マッピングデータを集積するための世界共通のフォーマットを提供し、実際にデータを集積していることに大きな意義がある。

現在、日本からもWIDE等を経由してGDBにアクセスできるが、アクセスのスピードが遅いことと、画面上の機能ボタンの上のカーソルをキーボードから動かして操作する方式の使いにくさがある。ただし、このことはリモートアクセスであるが故であり、GDBがデータ集積の機能を重点にしてつくられていることとは別のことである。

ゲノム解析研究をGDBに集積されているデータなしで行なうことはほとんど不可能である。しかし、GDBのデータをもっと実験室の現場で使いやす

くすることには、別の大きな意義があると考えられる。我々は、データベースを手元に置き、集積されたデータをもっと容易に利用できるシステムの開発を目指した。

3 日本語ヒト遺伝子マップライブラリの設計

我々が以前から作成していたデータベース JHGM/PC (Japanese Human Gene Mapping library on PC) [4] は国内で広く普及している NEC 社製のパーソナルコンピュータで動く。データベース管理システム (DBMS) として MS-DOS 上の Let's アイリスを用いている。遺伝子の形質等は日本語で記述しており、既に 6,000 件以上のマップデータと 10,000 件以上の文献データを独自に入力した。データはディスク上にあるので小まわりがきき使いやすいが、キャラクターベースであり検索速度はあまり速くない。この JHGM/PC で蓄積されたデータを利用することを前提に、我々は次のような特徴を持つ新しい遺伝子マップライブラリーシステム JHGM/X の開発を目指した。

- 汎用の X ウインドウの上で動く。
- データベース全体をオンメモリーに持って迅速に動作する。
- EUC コードによる日本語データを扱える。
- マウスをポインティングデバイスとして用いる優れたユーザインタフェースを持つ。
- 染色体バンドを模式化してグラフィクスで表示し、マップ表示を大幅に視覚化する。

以下に JHGM/X の特徴を項目別に解説する。

3.1 システム構成

JHGM/X、JHGM/PC 及び GDB のシステム構成を図 1 に示す。JHGM/PC も GDB も汎用を目的とした商用 DBMS の上に遺伝子マップ及び文献のデータを入力してつくられている。

専用のシステムを開発した方が汎用のシステムよりも目的に適合するものを実現できるであろうことは間違いない。ヒトゲノム解析研究のためのデータベース利用システムに必要な機能は限られているので専用システムの開発も現実的なことである。そこで、JHGM/X のデータ処理部分には独自のプログラムを開発し、計算機の性能を十分生かせるデータ構造を採用した。

JHGM/X の画面は、多くのワークステーションで採用されている X ウインドウ上に表示される。これにより、高速のネットワーク環境に接続されたマシンで X ウインドウが動くものならばリモートでも JHGM/X を用いることができる。画面制御部分は、MIT で開発された Athena ウイジェットセットを用いて記述した。JHGM/X を起動した画面を図 2 に示す。以下の解説には適宜図 2 を参照いただきたい。

3.2 データ構造

JHGM/X は日本語データを含む独立した二つのデータファイル（遺伝子マップデータと文献データ）を扱う。

遺伝子マップデータは、遺伝子記号、遺伝形質、遺伝子の存在領域などのフィールドを持ち、各遺伝子を記述している。文献データは、遺伝子記号、著者名、タイトル、ジャーナル名などのフィールドを持ち、各遺伝子のマップ決定に寄与した論文を記述している。

この二つのデータは、遺伝子記号をキーとし関連付けられ、システム起動の際に一つのデータ構造にまとめられて仮想記憶上に展開されるので、非常に高速なデータ検索が可能となる。

3.3 日本語機能

日本の研究者、臨床医、看護婦（士）、学生にとって遺伝子マップデータベースが日本語で使えることの意義は大きい。JHGM/X では、遺伝形質・備考については JHGM/PC と同じ日本語データを用いている。また、画面表示も可能な限り日本語を用いてつくられている。

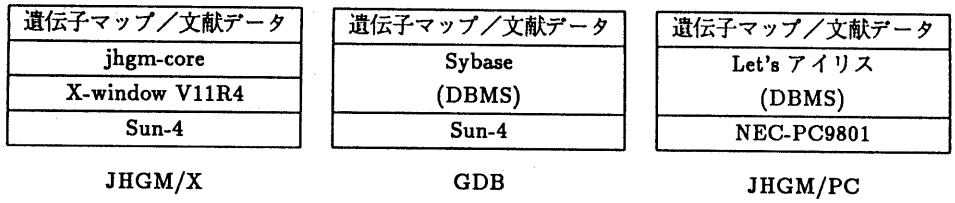

図 1: JHGM/X,GDB および JHGM/PC のシステム構成

3.4 容易な操作

3.4.1 マウスによる選択・実行

JHGM は、生物学者・医師等、計算機の専門家ではないユーザが使うことを想定しているので、そのような人々が容易に操作できることが強く望まれる。JHGM/X の操作は、基本的には検索キーワード等最低限の情報をキーボードから与えて、希望する出力形式の機能ボタンをマウスで選択するだけである。

3.4.2 キーワード・ウインドウ

キーボードからのキーワード入力もさらに減らすために、マウスを用いたコピー&ペースト機能を用いることができる。キーワードのプールとしてこのキーワード・ウインドウを用意してある。キーワード・ウインドウはそれぞれのユーザが自分が使う頻度の高いキーワードを自由に登録しておくホワイトボードである。その内容はファイルとして保存／再利用できる。キーワード・ウィンドウは機能ボタンを選択すれば、JHGM/X システムを起動したユーザのものが自動的に開く。

3.5 検索機能

JHGM/X の検索機能は、次の二種に大別される。

- 遺伝子マップ検索機能
  マップデータを構成するフィールドから遺伝子及びその文献を検索する。
- 文献検索機能
  文献データを構成するフィールドから文献を検索する。

以下に、それぞれの検索機能の特徴について述べる。

3.5.1 遺伝子マップ検索機能

遺伝子記号、染色体番号、存在領域、MIM 番号（V. McKusick 博士の著書 Mendelian Inheritance in Man [5] での遺伝子の整理番号）、遺伝形質の各フィールドから検索できる。複数項目にわたってキーワードを指定した場合、検索は AND をとって行なわれる。出力件数・行数に制限はない。結果はウインドウ下部に表示されるが、ウインドウに入らない部分はスクロールバーをマウスで操作して順次見ることができる。

検索の結果の出力を遺伝子にするか文献にするかはボタンで選択する。いずれの場合も、結果の 1 件を 1 行に簡略表示するモードとすべてのデータを表示する詳細表示モードのどちらかを選択できる。

検索結果の表示の順序は、領域順（短腕の端からセントロメアを通って長腕の端へ向かう順序）または遺伝子記号のアルファベット順のどちらかを選択できる。後者はさらに、染色体番号毎に遺伝子をソーティングした上でのアルファベット順と、染色体番号に無関係にアルファベット順とを選べる。

マップデータには、いわゆる遺伝子と anonymous（機能不明の）DNA 断片の両方が含まれるので、検索対象をそれらのどちらかだけに限定することもできる。

3.5.2 文献検索機能

著者名（単名の場合その人、2 名連名の場合 A & B、3 名以上の場合 A et al のフォーマット）、全著者リスト、文献タイトル、ジャーナル名・巻・ページ、発行年の各フィールドから検索できる。主な特

表 1: JHGM の文字列探索ルーチン

	α	β	γ
完全一致／部分一致	完全	部分	部分
ワイルドカードの使用	可	不可	不可
大文字／小文字の区別	有	無	有
日本語の使用	不可	不可	可
複数キーワード検索	不可	可	可
検索の柔軟性	大	中	小
検索速度	小	中	大

徴は遺伝子マップ検索機能と同じようにあてはまる。検索結果の表示順序については、著者名のアルファベット順または発行年順のどちらかを選択できる。

3.6 高速検索

高速の検索を実現するために 3 種類の文字列探索ルーチン（表 1 参照）を使い分けている。遺伝子記号の検索にはワイルドカードが使用でき、アルファベット大文字／小文字の区別のある探索ルーチン α が使われる。完全一致検索であるがワイルドカードが使える。ワイルドカードとして使える文字は次の 2 種である。ワイルドカードを複数組み合わせても目的どおり動く。

* ...0 文字以上の任意の文字と一致する。

? ...任意の一文字と一致する。

遺伝形質の検索には、日本語も字列が検索できる探索ルーチン γ が使われる。アルファベット大文字／小文字は区別される。自ずから部分一致検索であるので、指定したキーワードを含むものが検索される。キーワードを "&&" で接続して並べることにより、複数のキーワードの AND 検索を行なうことができる。

上記以外の検索には、大文字／小文字の区別をしない β 探索ルーチンが使われる。これは著者名や論文タイトルなどからの検索に有効である。これも自ずから部分一致検索であり、"&&" も使用できる。

参考文献

[1] A.E.M.Southern, *Application of DNA analysis to mapping the human genome*, Cytogenet. Cell Genet. 32, pp.52-57, 1982

[2] 清水 信義,「ヒトゲノム解析の現状と戦略」< 代謝> 第 27 巻臨時増刊号, 先端の医生物学とバイオサイエンス, pp.369-376, 中山書店

[3] *HGM 10.5*, Cytogenet. Cell Genet. 55, pp1-3, 1990

[4] S.Minoshima et al. *The human gene map database for a personal computer*, Jpn. J. Hum. Genet. 36, p77, 1991

[5] C.V.A.McKusick, *Mandelian Inheritance in Man.*, 9th Edition, The Johns Hopkins University Press, 1990

図２：JHGM/Xでの検索例
染色体番号（第７染色体）で遺伝子のみを検索し、そのマップを領域順に一覧表示したようす。キーワード画面も表示されている。