ESSAY #04

臨床ゲノム研究とプログラミング

埼玉医科大学
萩原 弘一

ゲノム研究の進歩により、さまざまな疾患遺伝子、疾患関連遺伝子多型が徐々にではあるが明らかになってきている。しかしながら、私は呼吸器内科学が専門だが、私の関連する学会でゲノム研究の発表を見ることはあまりない。私はゲノム研究に少々手を染めているが、なぜこれほど面白そうな分野に参入する人が少ないのか、そして研究発表が少ないのかと疑問に思っていた。数学を使うからだろうか、と当初は考えていたが、遺伝学やゲノム研究で使用する数学は高校レベルの数学で十分理解できるものが多く、それほど支障にはなりそうもない。もしかしたらコンピュータプログラムが必要となることがあるからか、と最近は考えるようになったが、実際の原因は謎である。

ゲノムワイド SNP データを得るために患者末梢血 DNA を専門機関で外注検査してもらうと、図 1 のようなデータが返ってくる。とても単純なデータだが、延々と 100 万行も続くその量が、一筋縄では行かない理由である。最近は Excel も 100 万行を読み込めるので、1 名分のデータは Excel でもぎりぎり処理できる量である。しかし数百名のデータを処理しようとすると、専用のプログラムが必要になる。そして定型的な解析以外の工夫を加える場合、自分でプログラムを書くことになる。医学研究者にはちょっと畑違いに思えるこの点が、研究者が増えない理由なのだろうか。

私はプログラミングも遺伝学も独学で、ほぼすべての知識は医学部卒業後に本から得た。プログラミングについて述べると、プログラミングを職業とする人にお会いしたこともなく、100%独学である。分子生物学は知っていたが、遺伝学はつい数年前、初めて教科書を手に取った。私の経験はやや特殊であるが、ゲノム研究を一人で始めたいと思う人の参考になるかもしれないと思い、とくにプログラミングに関して、簡単に記載してみたいと思う。

世にプログラミング言語は数多あるが、私のお気に入りは次の 4 つである。C、Ruby、Objective C、PostScript。C++ や Java など、主流とされる言語は他にあるが、例えば C++ は 10 冊くらい入門書を読んだだろうか、何冊本を読んでも使えるようにならなかった。あまりにも複雑すぎるのはどうもダメなようだ。Ruby や Objective C は、本を読んだその日のうちに使い始めることができた。専門のプログラマを目指している訳ではないので、このあたりはわがままだし、自分が気楽に使えるものに手が伸びる。また、前記の 4 つで十分こと足りるようだ。

C はコンピュータのすべての能力を引き出せる、基本中の基本言語として知られている。巨大なデータを高速に扱うために必須であり、最も良く使用する。巨大データに対して、高速化は非常に大事だ。私がこの前行なった計算は、Intel Xeon 搭載のコンピュータを含めた複数台のコンピュータで半年かかった。コンピュータが 2 倍速ければ 3 カ月、10 倍速ければ半月で済むはずだが、ちょっと気を抜けば一年を超えることになる。この時間差は大きい。

Ruby は、日本人のまつもとゆきひろ氏が開発した言語。目の前に草原が広がるように、それまでの言語がややこしく行っていた作業を見晴らし良く美しく行ってくれる、日本的な美意識を感じる言語である。

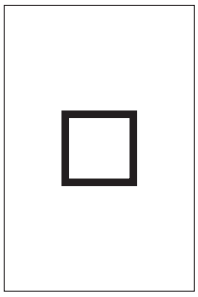
```
#CHP File=¥¥.PSF¥.Home¥Desktop¥A394_01-12¥A394_01-12_Raw_files¥20080503_211950¥A394-01.birdseed-v2.chp
#Exec GUID=0000050010-1209817196-0000018315-0000000354-0000014482
SNPID    Call     Confidence
SNP_A-2131660     AB       0.004934122
SNP_A-1967418     AB       0.004656699
SNP_A-1969580     BB       0.002190549
SNP_A-4263484     AB       0.008556345
SNP_A-1978185     AA       0.003239007
SNP_A-4264431     AB       0.0007773066
SNP_A-1980898     BB       0.003351128
SNP_A-1983139     AA       0.00306015
SNP_A-4265735     BB       0.001669541

以下約 100 万行続く
```

図 1　Affymetrix 社 SNP Array 6.0で取得した全ゲノム SNP データ出力の一部。SNP-ID は検索した SNP の名前。Call は SNP の遺伝型、Confidence はデータの信頼性を示す尺度。Annotation file といわれる、SNP-ID と染色体上の位置との対応表とともに使用する。

```
%%!PS-Adobe-3.1 EPSF-3.0
2.83 2.83 scale
1 setlinewidth
0 0 moveto
0 10 lineto
10 10 lineto
10 0 lineto
closepath
stroke
```

(a)



(b)

図 2
（a）四角を描く PostScript プログラム。最初の行は PostScript プログラムであることを示す指示文。PostScript はパラメータが前に来るという特徴がある。2.83 2.83 scale は x 軸、y 軸方向に2.83倍する（PostScript の 1 単位が1/72インチのため、それを mm 単位に変更する）。線の幅を 1 mm とし、0 0（x 座標、y 座標）に移動、0 10まで線を引き、10 10まで線を引き、10 0まで線を引き、線を閉じる。そしてそれを描画する。10mm 各の四角ができる。
（b）Illustrator で開いたところ

Ruby の文字列処理は強力で、それを期待して使うことが多い。1 名分のヒトゲノムは 30 億塩基対。コンピュータに入れると 3GB。ほとんどの人のコンピュータの RAM は、これ以下のはずだ。ヒトゲノムが解読され、何となく簡単に扱える代物になったような印象があるが、とんでもない誤解といえる。30 億塩基対のゲノムは、一塩基を 1mm の文字で書き並べて行くと 3000km になる。日本列島を往復する長さである。この長さの文字列の中から特定の文字を見つける、そんな文字列処理がゲノム研究の基本となる。Ruby では、最初が G、次が A か C、10 個おいて CA が 10 回以上、そんな見つけ方も一行のプログラムで実現可能だ。C は文字列処理がそれほど上手ではないので、速度がそれほど問題にならないなら Ruby を使ってみるといいと思う。

Objective C はユーザーインターフェース作成用だ。人に使ってもらえるようなプログラムができたら、ボタンとかメニューとかのあるプログラムにしておくと良い。Objective C は Mac や iPhone の開発に使われている言語で、しゃれたユーザーインターフェースがすぐに作れる。

PostScript はページ記述言語と呼ばれるもので、解析データを図として作成するのに使用している。最初数行の定型的な行を入れておくと、生の数字データをそのまま絵にすることができる。プログラミングの出力を、どのようにプレゼンテーションに有効な美しい絵にするか、というのはいつも悩むところだが、PostScript はそのための言語と言って良い。具体例を示そう。

図 2 は、簡単な四角を書く PostScript プログラム。ファイル名の最後を .ps として保存し、Adobe Illustrator で開くと、きちんと図形になっていて編集も保存もできる。Mac の場合、描画が PostScript に基づいているので、ダブルクリック

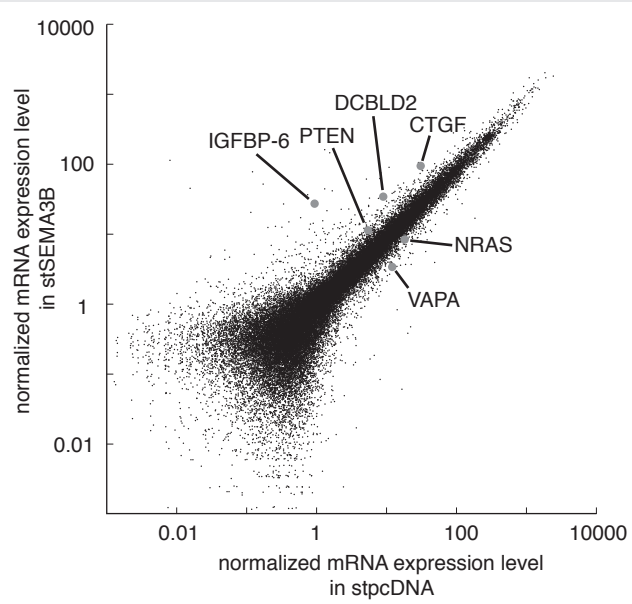


図3　発現アレイデータ。大量の数字データを一度に絵にしてしまう

するだけでプレビューというプログラムが絵にしてくれる。PostScriptの文法は1時間程度で身につけられるくらい簡単で、なぜみな使わないのか、といつも思う。例えば、図3は発現アレイの数万個の点の座標を示すExcelの数字データをプログラムの形にし、Illustratorで開き、Illustratorで軸や文字を付けたものだが、おそらく所要時間は10分程度。十分に論文に使用できる品質と思う。

私はMacを使っているので、4つのプログラム言語はMacOSXにはすべて組み込まれてくる。実質的に無料である。Windowsではobjective Cは使えないが、CやRubyはインターネットからダウンロードできる。ただ、UNIX環境を作るため、Cygwinなどの追加ソフトが必要となることがある。

最後に遺伝学を独習しようとする人のために、2＋1冊書籍を記載しておく。世に良書は多いが、最初の2冊はその中でも最上位に位置するものと思う。部分部分を読むだけでもその分だけ知識が得られる、という本なので、移動中や就眠前に眺めているだけでも勉強になる気がする。後の一冊はゲノム研究を巡るさまざまなエピソードをちりばめた自伝で、ヒトゲノム解読レースの実に興味深い舞台裏が記載されている。いろいろな意味で勉強になる本である。

《参考文献》

【プログラミング】
・プログラミング言語C ANSI規格準拠 第2版　石田 晴久訳, 共立出版, 1989.
・プログラミング言語Ruby　David Flanagan, まつもとゆきひろ. オライリージャパン, 2009.
・Mac OS X Cocoaプログラミング　アーロン・ヒレガス. ピアソンエデュケーション, 2002.
・PostScriptリファレンスマニュアル 第3版 Adobe Systems　ASCII電子出版シリーズ, 2001.
【遺伝学】
・ヒトの分子遺伝学 第3版, 村松 正實(監修), メディカル・サイエンス・インターナショナル.
・遺伝学概説, J.F. クロー, 培風館, 1991.
・Craig Venter. A life decoded: My Genome: My life. Penguin Books Ltd, 2008. 邦訳は「ヒトゲノムを解読した男クレイグ・ベンター自伝」東京化学同人, 2008.