【平成１４年度　未踏ユース事業】

４次元グラフによるゲノムの可視化

Genome Visualization on 4-th Dimensional Space

西尾 泰和 [1)]　　比戸 将平 [2)]
Hirokazu NISHIO　Shohei HIDO

1) 奈良先端科学技術大学院大学情報学研究科情報生命科学専攻比較ゲノム学分野
2) 京都大学工学部情報学科数理工学コース

ABSTRACT. Many genome sequences were specified by research in recent years. However, since the amount of information is huge, it is hard to grasp a whole image of them. Then, we suggested two visualization technology using the ability of genome arrangement to match with the curve on 4-dimensional space. Simultaneously, the application using the technique was developed. Moreover, visualization of relation nature applicable to the larger range was challenged, and the prototype was developed.

１．背景

前世紀末に始まった国際的なヒトゲノム計画によって、予想以上の速さでヒト遺伝子情報が明らかになった。今日では、その過程で開発された高速なＤＮＡシークエンスの決定技術によって、様々な生物のゲノム塩基配列が、凄まじいスピードで解明され、膨大な量のゲノム情報が公的機関のＷＥＢページにおいて公開されている。

しかし、それら大量に蓄積された塩基配列データから、生物学的に有意義な情報を探し出すことを研究するゲノム解析の分野は、いまだ黎明期にある。様々なアイデアに基づく新しい手法の開発に、世界中の研究者がしのぎを削っている段階と言えるだろう。つまり、２１世紀の科学を牽引していくと思われるバイオインフォマティクス研究分野において、欧米に比べ出遅れたと言われて久しい日本の研究者にも、まだまだ大発見のチャンスが残されていると考えられる。

２．目的

そこで本プロジェクトでは、独自の発想に基づき、高速かつ高性能で、また子供でも直感的に操作が可能な、汎用性の高いゲノム解析ツールを開発する。具体的には、4 種類の塩基を完全に等価に扱うための３Ｄグラフ表示機能を実装することによって、ユーザーのマウス操作によって角度やスケールを自由に変更して観察することを可能にする。またパブリックなゲノム・データベースとの連携機能を実装するなど、すぐにでも研究者に使用していただけるツールを目指して開発に取り組み、ゲノム・データマイニング分野に新たな視点を導入したいと考えている。

３．提案手法

（１）配列を曲線に対応付ける

$\{S_i\}$をゲノムの配列とする。S_iは'A','T','G','C' のいずれかの値をとる。$A_i;T_i;G_i;C_i$を

$$A_i := \sum_{1}^{i} d(S_i, 'A') \tag{1}$$

と定義する。d はクロネッカーのデルタであり、A_iは 1 塩基目から i 塩基目までにあった A の数になる。

ここで、4 次元空間上の点列$\{P_i\}$を

$$\vec{P}_i := (A_i, T_i, G_i, C_i) \tag{2}$$

で定義する。この点列をつないだものが以下の可視化技術の基礎になる 4 次元空間上の曲線である。この曲線を便宜的に「ゲノム曲線」と呼ぶことにする。

（２）3 次元表示

ゲノム曲線を可視化することを考える。ゲノム曲線は 4 次元空間上にあるので、そのままの形では表示することができない。そこで、3 次元空間上に投影する。

4 元空間上の構造を 3 次元空間上に投影するということは、1 次元分の情報を捨てるということである。ここで、「塩基 N 個あたりの各塩基の個数」を可視化することを考えると、一番必要のない情報は「塩基 N 個あたりの A,T,G,C の個数の和」である。この値は配列データに欠損がない限り常に N になるからである。そこで、次のような変換行列T を用いて座標を変換し、W_iの値を捨

てることにする。

$$T=\frac{1}{2}\begin{pmatrix}1 & 1 & -1 & -1\\ 1 & -1 & 1 & -1\\ 1 & -1 & -1 & 1\\ 1 & 1 & 1 & 1\end{pmatrix} \tag{3}$$

$$\left(X_{i,}Y_{i,}Z_{i,}W_{i}\right)=T\,{}^{t}\vec{P}_{i} \tag{4}$$

この行列Tは、4 次元空間上の回転行列になっている。つまり、ここで行われていることは、3 次元空間から 2 次元空間への写像に例えれば「立方体の体対角線が投影面に垂直になるように回転し、並行投影によって垂直方向の成分を捨てる」という射影である。

この投影は、3 本の座標軸が互いに 120 度の角度を持つことから「等角投影(isometric-projection)」と呼ばれる。つまり、先ほどの行列Tによる変換は等角投影の 4 次元空間への拡張ということになる。そのため、この投影によって 4 の座標軸は、正四面体の重心から各頂点への直線に写像され、互いに等しい 109 度 28 分の角度で交わる。

（3）既存の手法との比較

GC-Content,GC-Skew という既存の手法を解説し、ゲノム曲線が優れている点と劣っている点を述べる。

まず、便宜的にG'_{i},C'_{i}を

$$G'_{i}:=G_{i+w}-G_{i} \tag{5}$$

と定義する。ここで w はウィンドウ幅であり、1000 や 10000 などの数字が選ばれることが多い。このときG'_{i}はゲノム配列の i + 1 塩基目から i + w 塩基目までの間に含まれる G の個数になる。

すると GC-Content,GC-Skew は

$$\text{GC-Content(i)}:=\frac{G'_{i}-C'_{i}}{w} \tag{6}$$

$$\text{GC-Skew(i)}:=\frac{G'_{i}-C'_{i}}{G'_{i}+C'_{i}} \tag{7}$$

と定義される。

GC-Content はゲノムが複製される起点(orogin) や終点(tarminus) で値が極値をとる傾向があり[1]、GC-Skew は値の符号が変わる傾向がある[2]。

ここで GC-Content は$P_{i+w}-P_{i}$の(0, 0, 1, 1) 方向成分を$\sqrt{2}$で割ったものになる。また、図 2 は 4 次元空間を GC 平面に並行投影したものである。P_{i}を原点 O とし、P_{i+w}を投影した点をP'_{i}とする。P'_{i}を通り、G 軸および C 軸と 45 度で交差する直線 L を考え、原点から L に下ろした垂線の足を N とすると

$$\left|\vec{ON}\right|=\frac{G+C}{\sqrt{2}} \tag{8}$$

$$\left|\vec{NP}\right|=\frac{G-C}{\sqrt{2}} \tag{9}$$

となり、GC-Skew は$\tan q$になる。

つまり、GC-Content や GC-Skew は、ゲノム曲線をある恣意的な直線もしくは平面に投影したものであり、これらのグラフだけを見て議論をすることには、存在する特徴を見落としてしまう危険性がある。

そのため最近では GC-Content で特徴がわかりにくい場合に GT-Content の累積和である Keto-Excess などで特徴を抽出しようとする試みも存在する[3]。塩基は 4 種類存在するので、これらの 4 塩基を等価に扱わない方法には単純に${}_{4}C_{2}$もしくは${}_{4}P_{2}$通りの方法があることになる。

一方、ゲノム曲線の等角投影による可視化は、等角投影の性質上全ての軸が対等な関係にある。もちろん等角投影によって 3 次元に落とす過程で恣意的な視線ベクトルを採用しているが、4 塩基を等価に扱えることのメリットは大きい。また、現在任意の視線方向で観察することが出来るプログラムを開発中であり、完成の暁には視線の恣意性も克服できる。このことは、たとえば環状の DNA を持つ生物のゲノム配列を、環状になるような視線で観察することが出来る、ということであり、生物学的に重要な意味を持つ。

これらのプログラムの問題点は、3 次元空間上の曲線をインタラクティブに回転させて観察することを前提に開発されているため印刷に適さないという点と、一定塩基あたりの曲線の長さが一定しないため特徴的な形状がゲノム上のどの位置の配列であるかが見ただけではわからない点である。

（4）方向によらない尺度

そこで、ゲノム曲線の曲がり具合を 2 次元上に、1 軸を配列上の位置として可視化することを考える。そこで、方向によらない尺度$\{\Theta_{i}\}$を以下の式で定義する。

$$\Theta_{i}=\frac{(\vec{P}_{i}-\vec{P}_{i-w})\cdot(\vec{P}_{i}-\vec{P}_{i+w})}{\left|\vec{P}_{i}-\vec{P}_{i-w}\right|\left|\vec{P}_{i}-\vec{P}_{i+w}\right|} \tag{10}$$

これは方向余弦である。ゲノム曲線が直線状に延びているときは-1 に近づきゲノム曲線が急激に曲がるほど増加する。これは i 番目の塩基の、前 w 個と後 w 個の塩基組成がどの程度変化したかを示す尺度である。

これを実装したプログラムでは$\{\Theta_{i}\}$が$\Theta(i,w)$とみなせることを利用し、値を明度で表すことによって w の値による変化も表すことが出来るようにした。これを以下では Cos$q$ Map と呼ぶ。

４．開発成果概要

（１）100%PureJAVA ３Ｄエンジン作成

Java3D などの既存のライブラリに頼らず、３Ｄ処理を全て自前で行う描画エンジンを開発した。

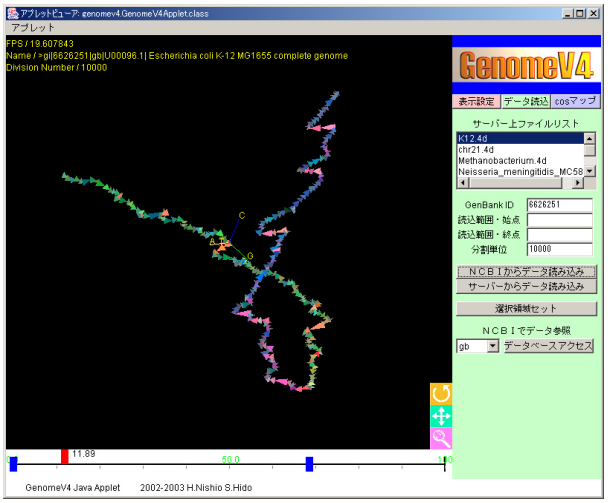
３Ｄ処理に必要な基本モジュールを開発した後に、エンジン自体のゲノム可視化への最適化・高速化を何段階にも渡って行った。外部ライブラリを使用した場合には、このようなことは不可能である。これにより、数十万ポリゴンにも及ぶヒトゲノムの表示にも対応できる、極めて高速なレンダリングエンジンとなった。本来３次元上の"線"でしかないベクトルデータを、どの角度から見ても太さを持つように、ポリゴンで表示することが可能である。その結果、ゲノム全体を表示させたときにも、形や色がはっきりとわかるように表示することが可能になった。

（２）GenomeV4 Applet 開発

３．手法で述べた独自の可視化アルゴリズムを搭載した Java アプレット、GenomeV4 Applet を開発した。

複数のシークエンスを同時に表示してその構造を見比べることができる。同じく４次元ベクトルを対応させた色をポリゴンにのせて表示させ、特定の色だけ強調できる色調調節機能も搭載した。また、範囲指定を行うことで、ゲノムシークエンスの一部分だけを詳細に表示させることもできる。

GenomeV4 Applet はＷＥＢブラウザ上で動作する Java アプレットなので、ユーザーが専用モジュールをダウンロードする必要は一切無い。



<Figure 1: GenomeV4 Applet 実行画面>

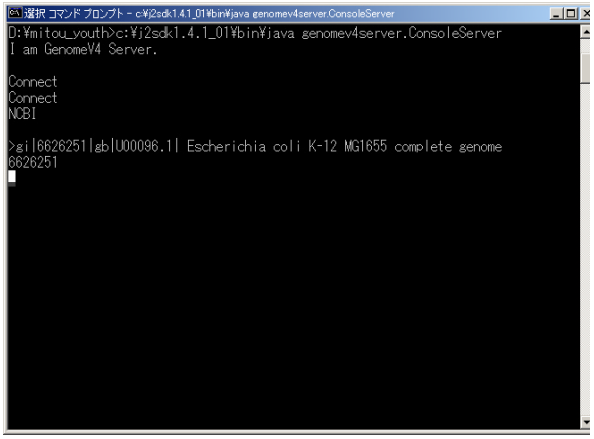
（３）サーバープログラム GenomeV4 Server の作成

アプレットでは、セキュリティ上の制約から、任意のネットワークサーバーにアクセスすることができない。そこで、アプレットと同じサーバー上で動作するデータ仲介役とも言うべきサーバープログラムを作成した。

GenomeV4 Server は、Java のマルチスレッド技術を用い、実際にクライアントである GenomeV4 Applet からデータのダウンロード要求を受理した後は、メインスレッドとは別のスレッドを生成して、以降の通信処理を全て委任するため、同時に複数のクライアントから接続要求があっても十分耐えうる仕様となっている。

GenomeV4 Server が扱うのは、アメリカのゲノム・パブリックデータベース NCBI(http://www.ncbi.nlm.nih.gov/)で公開されている、GenBank 形式と呼ばれるフォーマットのゲノム配列データである。GenomeV4 Applet からユーザーが可視化したい GenBank データのＩＤを受け取ると、GenomeV4 Server 上で等角写像に必要なデータの抽出・処理を行って、描画に必要な情報だけを GenomeV4 Applet に返す。これにより、数メガバイトから数十メガバイトにも及ぶ配列データを、閲覧のためだけにユーザーが自分のＰＣまでダウンロードしてくる必要がなくなった。

将来的に、外部に高速回線でアクセスできる環境や、ゲノムのシークエンスデータ自体をローカルネットワーク内に持っている研究所のサーバー上で、GenomeV4 Server と GenomeV4 Applet を運用することを考えている。現時点でも、サーバー上に研究者が公開したいデータを、ユーザーに提供可能な形に加工して登録しておけるようになっている。そうすればアプレットを実行するユーザーに、さらに迅速に可視化情報を提供できるようになる。



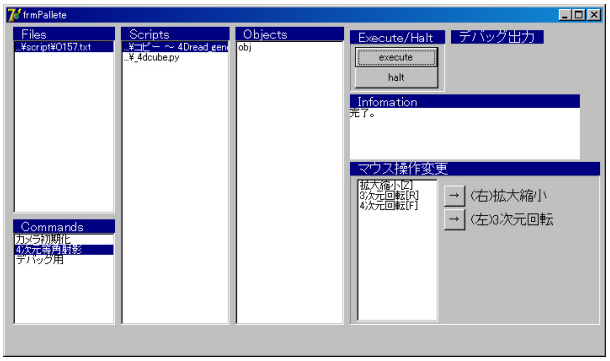
<Figure 2:Genome V4 Server 実行画面>

（４）プロ向け汎用可視化ツール GenomeV4 Pro 開発

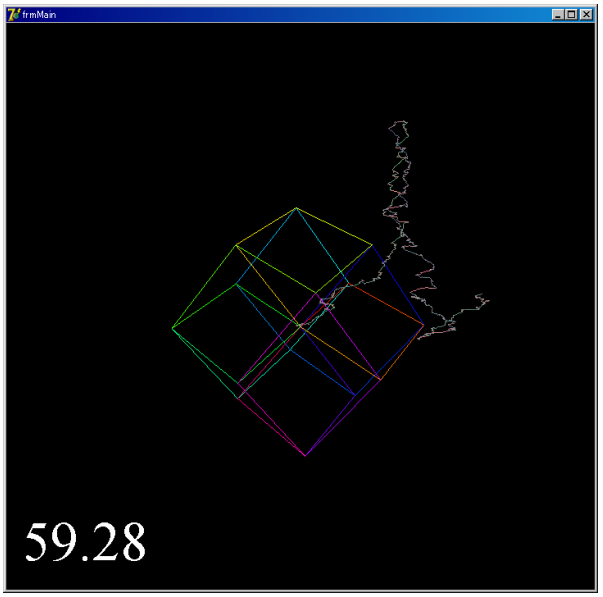
GenomeV4 Applet は一般のユーザーでも WEB 上で気軽にアクセスできることを重視して開発したツールだが、同時にさらに高機能で汎用性の高い研究支援ツールも作成した。開発言語には Delphi、3D レンダリング部分には OpenGL を採用しているので Linux 上への移植も可能である。

また、コアの 4 次元ベクトルを表示する部分を拡張し、任意の 4 次元的な方向から投影できるようになった。これを用いることで 4 次元的な構造、たとえば複素関数などを可視化することも可能である。

表示する物体は Python スクリプトで動的に追加削除できるので、分野を問わず可視化ツールとして利用できる。



<Figure 3: GenomeV4 Pro サブウィンドウ>



<Figure 4: GenomeV4 Pro メインウィンドウ>

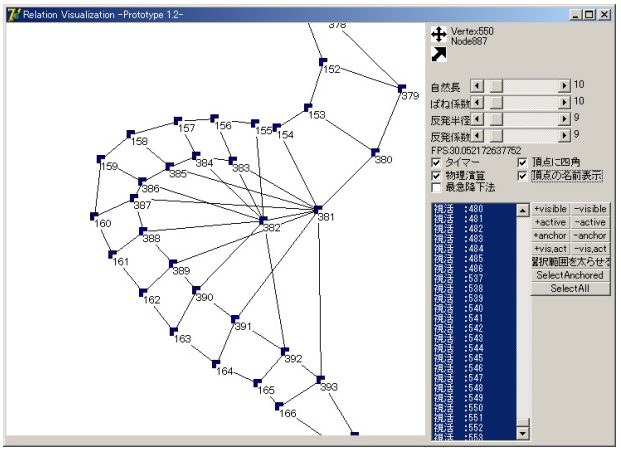
（５）Cos θ Map

4 次元空間上のゲノム曲線の折れ曲がり具合を、白黒の濃淡で表現することで塩基組成の変化を効果的に可視化するツールを作成した。この可視化手法によって、外来性の毒性遺伝子が存在する位置などに特徴的な模様が確認された。

<Figure 5: Cos θ Map 実行画面>

（６）関係の可視化ツール プロトタイプ開発

関係を可視化するツールのプロトタイプを作成した。このソフトによって、従来の方法では可視化が難しかった情報を可視化できるようになる。



<Figure 6:関連可視化ツールプロトタイプ>

５．今後の展望

プロジェクト期間中に、途中成果発表という形で、ＧＩＷ（国際バイオインフォマティクス学会）においてポスター発表を、情報処理学会のプログラミングシンポジウムにおいては登壇発表をさせていただき、本プロジェクトのきっかけとなったアイデアの理論的な説明と、開発中のゲノム解析ツールについてデモンストレーションを行った。そこで、分子生物学の研究者の方々には、我々の発想とアルゴリズムの独創性や、得られる結果の新規性について、高い評価を頂くことができた。また情報科学の専門家の方々は、３Ｄ技術と Java 言語が、今まさに旬であるゲノム解析の分野への応用として十分実用化されていることについて、驚きの声をあげておられた。

今後、本プロジェクトの成果は、国立遺伝学研究所の池村淑道教授、並びに奈良先端科学技術大学院大学の金谷重彦助教授との、共同研究プロジェクトへと繋がっていく予定である。最先端の研究者の方々に我々のツール類を実際に使用していただき、そこで得られる貴重な御意見を取り入れることによって、さらにソフトウェアとして磨きをかけていたい。

６．参考文献

[1]http://linkage.rockefeller.edu/wli/dnacorr/grigoriev98.pdf
[2] Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol.13, 660-665
[3]http://www.sciencemag.org/cgi/content/full/279/5358/1827a