

4 生命科学分野・ゲノムデータの 可視化技術

西村 邦裕

東京大学大学院情報理工学系研究科

はじめに

情報爆発、情報大航海といった、多量の情報をどのように取り扱っていくのかという研究が数多くなされるようになってきた。主な取り扱いの対象としては、Web上の情報、データベース、ロボットの動き、テレビなどの動画などの多量情報である。このような分野で研究されている成果が役に立つと考えられる1つのアプリケーション分野が生命科学分野である。生命科学分野においても、多量の情報が蓄積されており、それを処理して、生命科学の新しい知見の発見につなげようという研究が進められている。具体的なデータとその解析目的を持ち、現実的な情報爆発に直面している分野である。このような多量の情報の理解、全体像の理解、解析の助けになるのが可視化技術であると考えられる。そこで、本稿では、生命科学技術分野における可視化、特にゲノム情報の可視化を対象として、紹介を行う。

多量のゲノム情報

生命科学技術分野、特にゲノムサイエンス分野において、whole genome shotgun 法などによる塩基配列の解読や、半導体技術を用い DNA チップを利用した microarray 法による発現量情報の取得など、生命科学情報の高速取得技術が発展してきている。この結果、2001年にヒトゲノムの解析が完了し、アメリカのNCBI（国立バイオテクノロジー情報センター）やUCSC（カリフォルニア大学サンタクルーズ校）のデータベースに格納されてパブリックドメインとして公開されている例や、さまざまなデータベースを統合データベースとして整理しようという事例(DBCLS)に代表されるように、さまざまな生命科学情報が公共データベースに格納されるようになり、データベースのエントリ数も急激に増えている。さらには、塩基配列読み取り処理の並列化などにより、一度にテラバイトオーダーで実験データを算出する装置な

ども生まれ、情報科学による支援、特に、多量情報の解析およびその支援技術が急務になっていると言える。

情報科学と生命科学の融合的な分野として、近年、ゲノム情報科学、情報生物学、bioinformatics と呼ばれる分野が生まれた。そのため、生命科学分野において、生物学的な実験をすることも必要であることは変わらないものの、より多くの生命情報を利用した解析の重要性が高まってきている。手に入るゲノム情報の全体像をつかみ、俯瞰しながら注目すべき点を見つけることや、データ全体の傾向や特徴を捉え、新規機能や新規メカニズムを発見あるいは類推していく作業が必要となってきている。この解析作業の支援として可視化技術が重要な役割を担いつつある。本特集の他稿で触れられている2009年3月に開催された第2回BioHackathon¹⁾においても、可視化についてはワーキンググループが設けられ、活発な議論が行われた。

生命科学分野における可視化対象

生命科学分野において、取り扱う情報が広範囲にわたることから分かるように、可視化対象も広範囲にわたる。2010年3月に開催されるEMBO Workshop on Visualizing Biological Data (VizBi)²⁾では、可視化の対象を6分野に分け、セッションとして、生物科学分野の可視化について議論を行う予定になっている。

- (1) 配列アセンブリとゲノム：塩基配列の並べ方、染色体から遺伝子、塩基配列の情報までを閲覧するためのゲノムブラウザ、複数の種のゲノムを比較し共通性・差を研究する比較ゲノムなど
- (2) 配列アライメントと系統発生：塩基配列同士の類似性、遺伝的進化、進化系統樹など
- (3) システム生物学：転写やmRNA（トランスクリプトーム）、タンパク質（プロテオーム）、代謝（メタボローム）、タンパク質や遺伝子の相互作用ネットワーク、生体内の反応経路を描くパスウェイなど

- (4) 構造：小分子同士の相互作用、分子間の反応、タンパク質の構造、分子動力学など
- (5) 光学顕微鏡、時系列情報：光学顕微鏡などで観測された細胞、腫瘍、組織、1分子の動きなど
- (6) MRI（核磁気共鳴装置）：MRIで観測された細胞、腫瘍、組織など（3次元再構成なども含む）

以上のように、生命科学分野における可視化対象は、可視化対象が物理的形を持つ科学的可視化（Scientific Visualization）から、可視化対象が物理的形を持たない情報の可視化（Information Visualization）にまでわたる。本稿では主に、筆者が取り組んできた（1）のゲノム情報に焦点を絞り、情報の可視化について述べることにする。

ゲノム情報解析における要求事項

ここで、対象となるゲノム情報の解析作業の流れを考えてみる。生命科学の研究者は、従来まで注目する遺伝子やタンパク質といった対象に対して研究を進めていくことが主であったのに対して、ゲノム情報の網羅的取得が可能となったために、ゲノム情報を全体として対象とするようになった。もちろん、具体的な遺伝子や見たい染色体領域、機能などがある場合は検索により情報抽出を行うが、何かの変異や特徴、新規的な部分を発見したい場合などは、全体像の中から興味のある部位・注目部位を見いだす、という作業を行うようになった。つまり、ぼやけた対象を試行錯誤しながら発見する作業であると言える。この作業を分解して考えてみると、以下の4点を繰り返す試行錯誤を行っている。

- 全体の傾向を捉える
- データの特徴検出
- 注目する部分を詳細に見る
- 他の種類・他のデータと比較をする

この試行錯誤においては、パラメータを変更し注目する部分・特徴を見いだすとともに、全体の中での特徴から詳細を眺める。そして、異なる情報（軸やデータ）と比較することにより共通性・差異を見だし、多面的に見る。さまざまなパラメータ変化の可能性を試すことにより、データを解釈していく。このような作業を繰り返し行う。

ゲノム情報解析は、試行錯誤をする過程で、データの特徴を捉え、ある事象の証拠となり得るデータを抽出する。そして、妥当性のある証拠を積み重ね、そこから仮説を立てる。その仮説に対して、生物学的実験や、他のゲ

ノム情報などを利用して裏付け、検証する、という過程を行っているといえる。

ゲノム情報解析支援の要求事項としては、解析をする際のインタラクションとして、上記の試行錯誤が可能なように、

- 多面的に見ることができるインタラクティブ性
- パラメータの変更機能
- 全体像と詳細像の提供
- 情報の抽出機能
- 比較可能な仕組み

が必要である。多種の情報に対して試行錯誤を行うと、情報の参照回数も多くなるため、情報の参照に対する支援も必要である（図-1）。

筆者としては、解析における判断は最終的には人間が行う必要があるが、その判断を可能とするのが可視化技術、という立場である。つまり、全自動ではなく、半自動で情報を抽出し提示することを行っている。可視化については単に絵にするだけではなく、上記の要求事項を満たす方法が必要であると考える。

遺伝子のコピー数情報の可視化

ここでゲノム情報の可視化について、全体像を分かりやすく見るための可視化手法などを紹介する。ゲノムサイエンスの研究者は、ヒトゲノムなどを解析している場合、ゲノムの染色体位置情報が頭の中に入っているため、ある意味、地図として染色体情報が役に立つ。そのため、主に、染色体情報を用いた可視化手法なども多い。

本稿では一例として、遺伝子のコピー数の可視化を取り上げる。

ヒトの場合、遺伝子は1つの細胞に2個（2コピー）あ

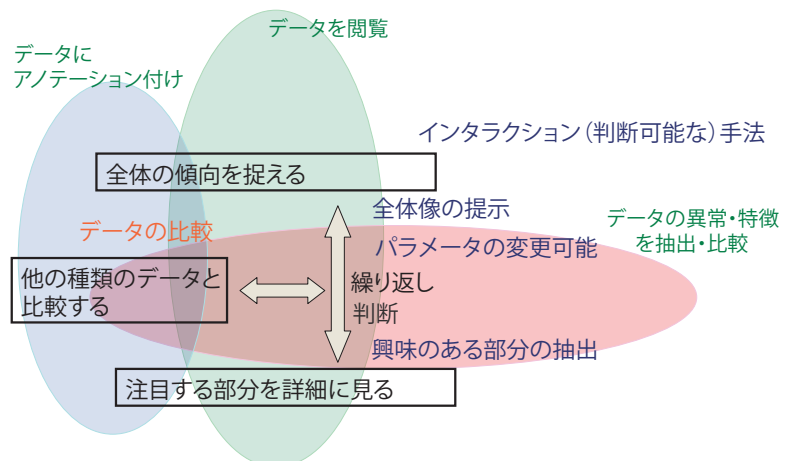


図-1 ゲノム情報解析への要求事項

り、1つは父親由来、もう1つは母親由来であると言われている。父親由来、母親由来の遺伝子是对立遺伝子と呼ばれる。近年、人によって1つの細胞に遺伝子が1個(1コピー)しかないことや、3個(3コピー)以上あることなど、個人差がある、ということが分かってきた。この遺伝子の「数」の個人差はコピー数多型(CNV: Copy Number Variation)と呼ばれる。このCNVの違いが、個人ごとの病気になるリスクや薬に対する副作用などを含めた反応の違いにつながると考えられている。図-2にCNVの概念図を示す³⁾。

これまで、疾患リスクなどの個人差として遺伝子の「配列」が異なる一塩基多型(SNP)が知られているが、CNVは遺伝子の「数」の違いであることが特徴であり、遺伝子を丸ごと含むような大きな塩基配列の領域の数が個人間で異なるものである。たとえば、アフリカのある民族は、ヒト免疫不全ウイルス(HIV)に感染しても発症しないことが知られている。遺伝子のコピー数を調べると、他の民族では4個から6個(4コピーから6コピー)特定の遺伝子を持っていることに対して、この民族では10個(10コピー)以上持っている。このことから、この特定の遺伝子のコピー数がHIVの発症に関係しているのでは、ということなどが分かってくる⁴⁾。この事例のように遺伝子のコピー数と疾患の関係を見ることが、疾患のメカニズムの解明につながると考えられている。

近年、DNAチップを利用して、そのまま遺伝子のコピー数を計測する手法が生まれ、統計的解析をすることでコピー数の推定が可能になった。遺伝子のコピー数情報の可視化の目的は、遺伝子のコピー数異常領域の検出支援である。可視化により、全体像の把握が可能となり、遺伝子のコピー数の異常領域が検出できるようになる。

遺伝子のコピー数を分かりやすく表現するとともに、対立遺伝子の区別をし、染色体の地図を利用した可視化例が図-3である⁵⁾。遺伝子のコピー数を本数で表現し、ヒトゲノム全体で、遺伝子のコピー数変化がひと目で分かるようになっている。

さらに複数人のデータ同士を比較するために、遺伝子のコピー数情報を並べた可視化事例がある。コピー数の異常領域の検出が目的であるため、0コピー領域のみを抽出して可視化することにより、異常領域だけが表示される。図-4においては、0コピーの部分のみを抽出し、22サンプルを横に並べている。た

えば染色体3番の短腕においては、複数のサンプルが共通して0コピーとなっている領域であることが見てとれる。このデータは肺癌のデータであり、小細胞性肺癌(SCLC)と大細胞性神経・内分泌系肺癌(LCNEC)の部分のデータに分けることができる。背景を色で変えることで、癌の種類によりコピー数が0本の領域(欠損領域)の比較をすることが可能となっている。

筆者と共同研究をしている複数の研究者が以上のような可視化結果を用い、実際の解析を行っている。特に、生物学的な意味論まで踏み込み、全体像の把握が可能な可視化結果は、解析支援として有効である、最終的には論文の図録として利用されるなどに至っている。

インタラクティブなゲノムビューワ

ゲノム情報解析支援の要求事項としてのインタラクティブ性などについて上述したが、その実現例としてのゲノムビューワを紹介する。ゲノム情報のパブリックドメ

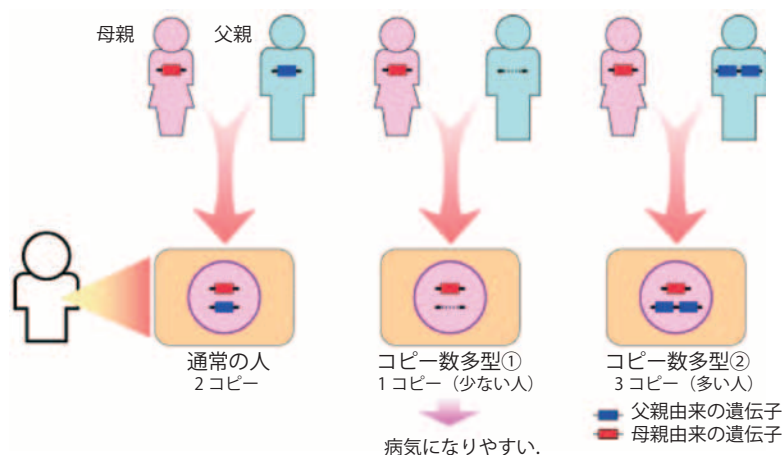


図-2 CNV (コピー数多型) の概念図³⁾

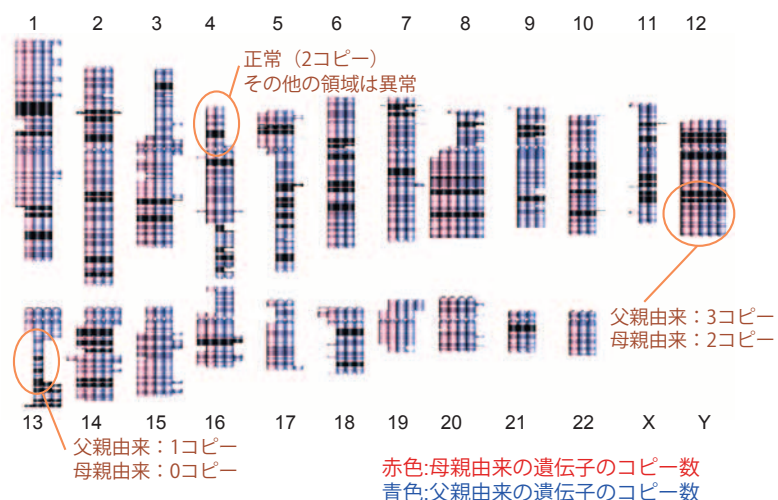


図-3 肺癌1サンプルについて対立遺伝子のコピー数を可視化

インにあるデータベースにおいては、Web を通したインタフェースにより閲覧機能や検索機能を提供している事例が多い。ただ、Web ブラウザを利用した場合には現状ではまだ限られたインタラクションしかできないことが多い。本章においては、スタンドアローンのソフトウェアとしてゲノム情報をシームレスに閲覧することが可能なゲノムビューの試みについて紹介する。

ゲノム情報の可視化をする際に、染色体情報を地図として用いること、その上に実験データなどを情報として配置し、それらを染色体全体の全体像から遺伝子レベルまでシームレスにズーム可能に可視化した事例が図-5である。図-5はヒトを対称とし、遺伝子発現量という数値データやメチル化などの修飾情報を色により可視化した例である。染色体1本を選択し、マウスを動かすことにより、連続的に拡大・縮小・移動をすることができ、研究者が閲覧したい部分、あるいは、全体的に見て、色の変化の大きい部分に注目し、詳細像を拡大して見ることや3次元的に斜めに見ることが可能となっている。

類似したコンセプトを持ち、ゲノムを比較するために

3次元に配置し、グラフなども一緒に可視化した事例として Sockeye がある⁶⁾。3次元に並べることにより表示空間を大きくし、比較しやすくする試みである(図-6)。また、ヒトのミトコンドリアゲノム⁷⁾の全体像と塩基配列を同時にインタラクティブに可視化しているビューとして mitoWheel⁸⁾がある(図-7)。mitoWheel は Flash で実装されており、Web 上で提供している事例である。

さらに図-5のゲノムビューを発展させ、より綺麗なグラフィックスを備え、データアクセスやさまざまな機能を備えたゲノムビューが genoDive⁹⁾である。図-8が原核生物用のゲノムビュー genoDiveProである。染色体情報からシームレスにアミノ酸配列・塩基配列情報まで拡大可能であり、全体像を3次元的に俯瞰することもできる。genoDive は上述の第2回 BioHackathon で筆者らが紹介し、参加者からの注目を集めた。genoDive はまだ開発途中のゲノムビューであるが、可視化以外の部分にも特徴を持っている。それは、本特集の他稿でも触れられている分散型データ共有の方法として XML で記述された仕組みである

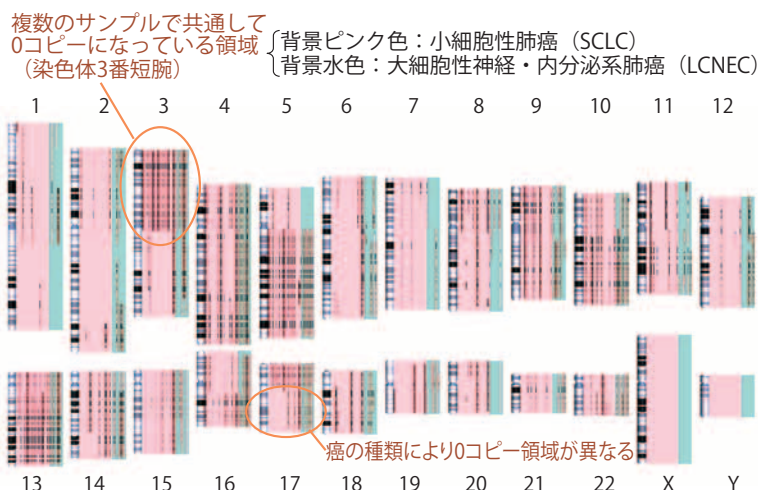


図-4 肺癌22サンプルについて対立遺伝子が0コピー部分の可視化。癌の種類により、コピー数が0コピー部分の分布が異なることが分かる

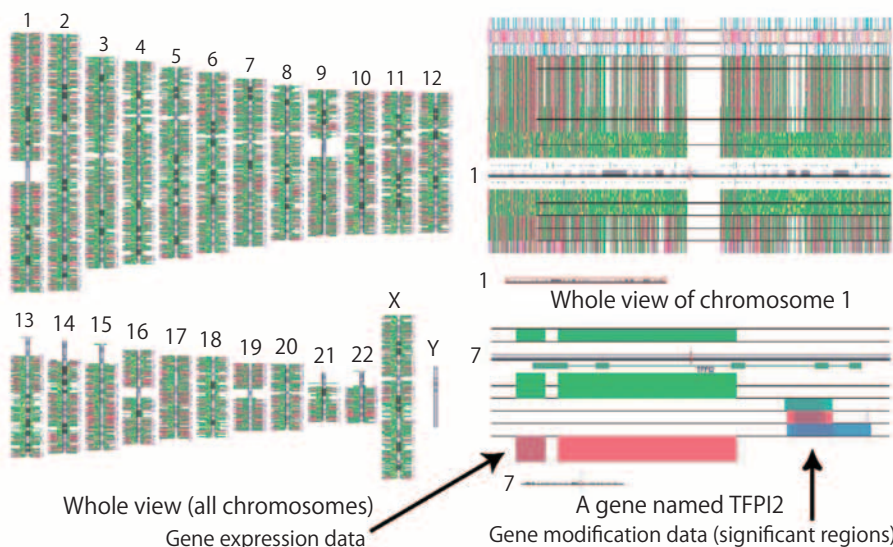


図-5 ヒトの遺伝子発現情報などをシームレスにズームできるゲノムビュー

Distributed Annotation System (DAS) のクライアントとして genoDive は機能し、ユーザの選択により可視化しようとする領域に応じて、必要な情報をデータベースから非同期的にダウンロードして可視化する仕組みを持っている点である。また、キャッシュ機能も併せ持ち、ネットワークの負担を最小限にすることや、ローカル PC にある規格のある XML ファイルフォーマットに合わせたデータを取り込むことが可能となっている点である。そのため、さまざまな公開されているデータベースのクライアントとして利用することが可能であり、複数のユースケースが考えられるビューワである。

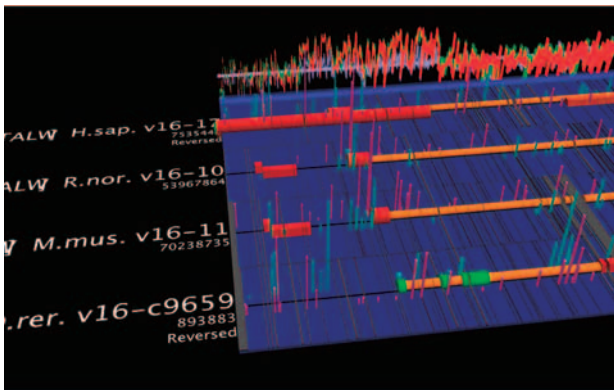


図-6 ゲノムを比較するための3次元状に種を配置した Sockeye⁶⁾

ゲノム情報の可視化とこれから

ゲノム情報の可視化の要求事項や可視化例について述べてきたが、最後に、これからのゲノム情報の可視化について、第2回 BioHackathon における議論なども参考にしつつ述べる。

情報が急速に蓄積しているゲノム情報に対しては、研

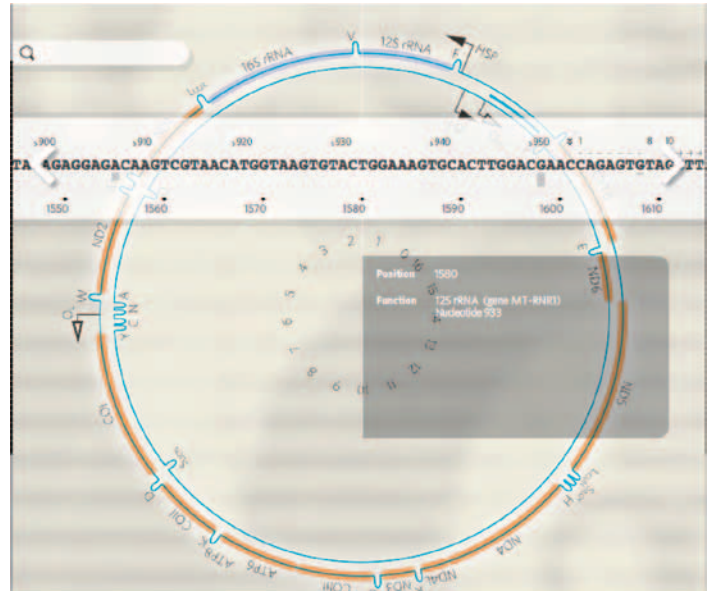


図-7 ヒトミトコンドリアゲノムのビューワである mitoWheel⁸⁾

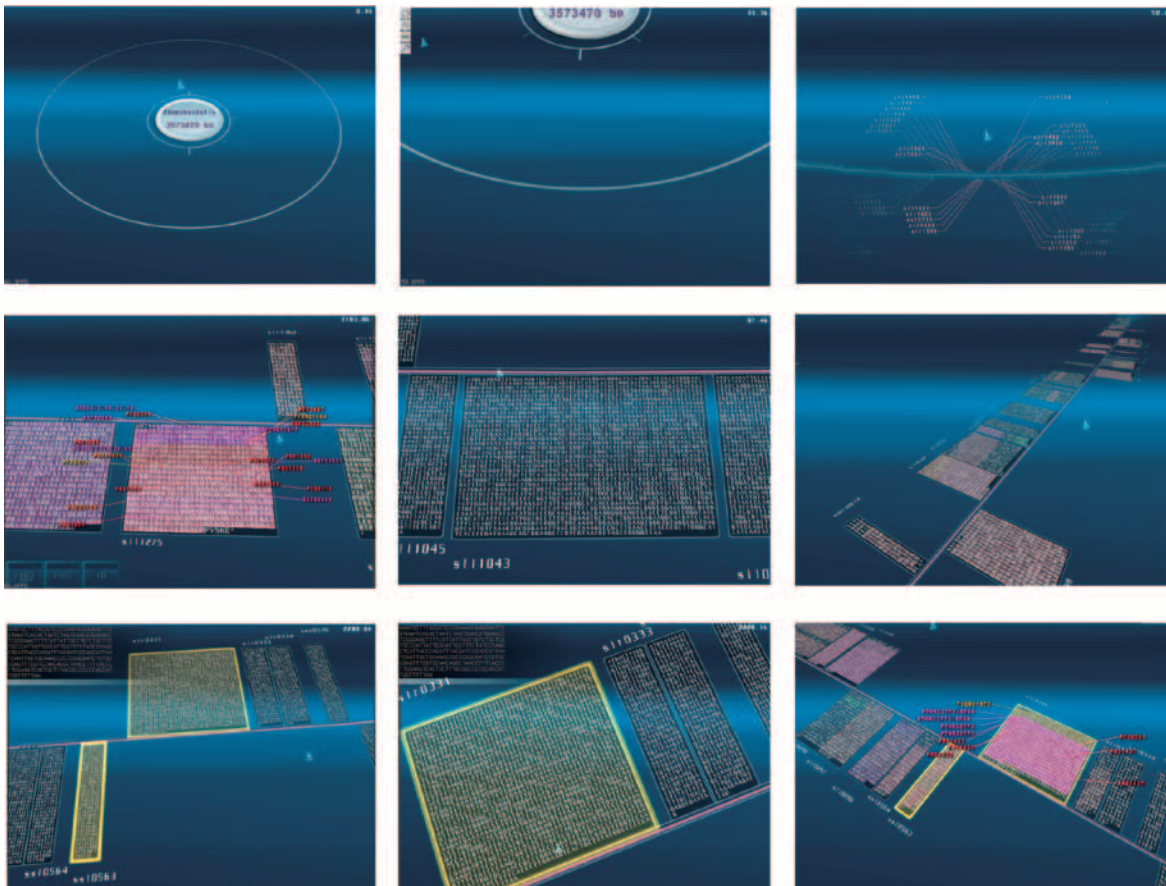


図-8 原核のゲノム情報をデータベースから自動的に取得し可視化するゲノムビューワ genoDivePro⁹⁾

究者(ユーザ)の処理が追いつかなくなってきており、より高度な可視化技術が今後とも要請されている。可視化の方向としては、(1) より大規模、かつ、よりゲノム情報の意味論を考慮した可視化、(2) より分かりやすく利用しやすいインタフェースを備えた可視化、があると考ええる。

(1) より大規模、かつ、よりゲノム情報の意味論を考慮した可視化

今後、次世代シーケンサやさまざまな網羅的実験装置の普及により、多量の情報、多人数のデータの算出が予測され、それらを解析、理解するための可視化が必要である。1,000 ドルゲノム計画などとも言われつつあり、診断の際に、ゲノムを読んでからオーダーメイド医療を行う、という青写真も描かれつつあるが、そのためには、

それらを支える情報基盤が必須である。その1つとして、医者・研究者が理解するための可視化、また、医者から患者に見せるための可視化、などが必要である。

本稿ではゲノム情報の中でも分かりやすい染色体情報に沿った情報を取り上げたが、遺伝子同士やタンパク質同士のネットワークなどの情報も可視化の対象である。第2回 BioHackathon の可視化ワーキンググループでも議論されたネットワークを対象とした Cytoscape¹⁰⁾ (図-9) のようなさまざまな統合環境における可視化も、解析インタフェースと密接にかかわってきているため重要であろう。

また、可視化の方法として、2次元に頼らず3次元を利用するというのもアイデアとしては考えられる。3次元を利用した際にはオクルージョンの問題(ある物の背後が隠れて見えなくなる問題)が出てくるが、動きで対応することや、近年、徐々に発売されてきている3次元ディスプレイの普及により解決できると考えられる。

3次元可視化の一例としては、細胞のシミュレータであり、その上に反応シミュレーションの結果を載せている E-Cell 3D¹¹⁾ (図-10) などがあげられる。細胞が3次元情報を持っており、科学的可視化の一部とも言えるが、多面的に情報を見せるために3次元を利用する側面もあり、情報の可視化でもある。また、第2回 BioHackathon において、筆者が Wellcome Trust Sanger Institutes の Jan Aerts 氏と試みたゲノム情報コピー数の可視化の3次元化¹²⁾ (図-11) においても、プロトタイプではあるが、見たいデータの比較が容易になりそうな感触をつかめるなど、可視化の表現力として追求できると考えている。

(2) より分かりやすく利用しやすいインタフェースを備えた可視化

ゲノム情報の研究者は、必ずしもコンピュータに強い研究者ばかりとは限らない。そのため、より直観的なインタフェースや操作方法を提供する必要があると考える。また、ゲノム情報解析は知的作業でもあり、そのような知的作業を支援できる仕組みが必要であろう。筆者は、ゲノム情報解析において、試行錯誤をする際、その「試行」を支援するための可視化が必要であることに加え、試行錯誤をした「プロセス」自体の可視化が必要ではないか、と考えている。上述した genoDive の設計においては、そのような「プロセスの可視化」の一例になるように、注目した部分を「ブックマーク」して保存する機能や、自分が興味のある部分に対してリストを作ってスライドショー的に情報

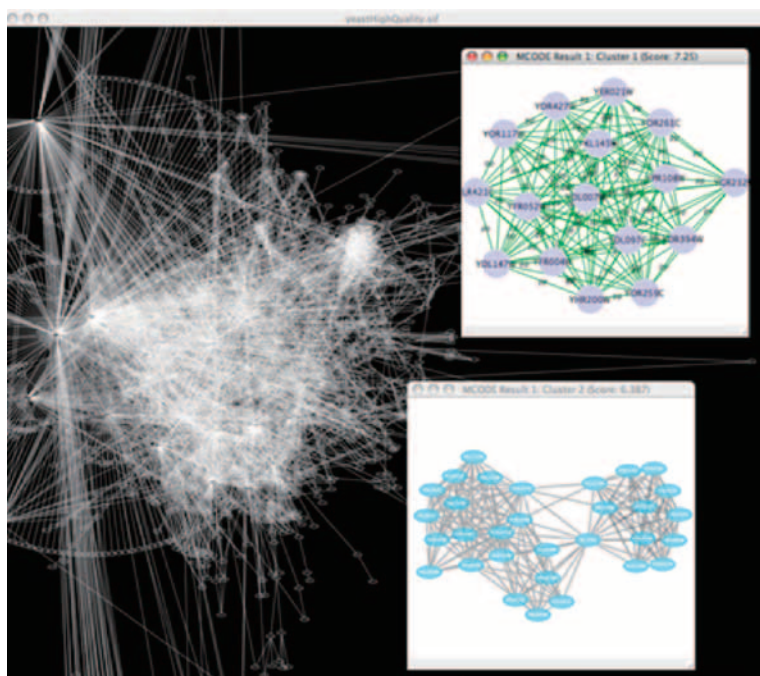


図-9 ネットワークを対象とした可視化もサポートする解析環境 Cytoscape¹⁰⁾

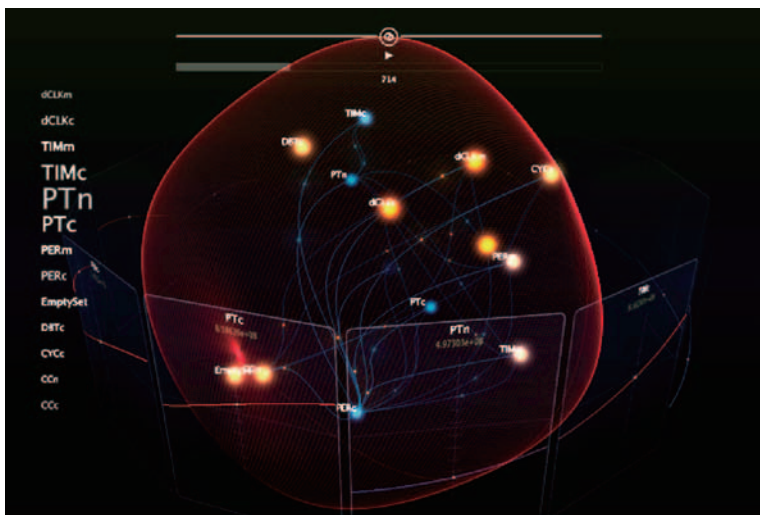


図-10 細胞のシミュレータ E-Cell 3D¹¹⁾

を眺めることができる「ツアー」機能などを実装している。このような機能が知的作業支援となるかなど、今後、評価などを行いたいと考えている。

また、インタフェースとしては、第2回 BioHackathonなどで議論したが、音楽のフィジカルコントローラなどを利用して、パラメータを変更し、ゲノム情報を閲覧・解析する試みなどが考えられる。図-11のpARPの3次元プロトタイプにおいて、フィジカルコントローラによる回転操作等の操作体系は分かりやすかったため、今後、可視化と合わせたインタフェースの検討も必要だろう。

最後に、よりリッチなインタフェース・環境として、バーチャルリアリティ環境の利用も考えられる。複数人で議論する際など、大画面を利用し、解析作業を行うことが重要であると考えられ、図-12のような壁面型ディスプレイに情報を提示し、解析支援できる環境も有効であると考えられる。

以上、ゲノム情報の可視化について論じてきたが、まだまだ解決すべき課題や挑戦しがいのある部分が多い分野である。情報学・情報工学の専門の研究者が生命科学の課題により一層挑戦することで、今後、さらなる発展が期待できると考える。

参考文献

- 1) <http://hackathon2.dbcls.jp/>
- 2) http://www.embl-heidelberg.de/training/courses_conferences/conferences/index.php?p_conferenceld=128
- 3) <http://www.lsbm.org/news/2006/1123.html>
- 4) Gonzalez, E., et al. : The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility, Science, Vol.307, pp.1434-1440 (2005).
- 5) 西村, 石川, 広田, 油谷, 廣瀬: ゲノムコピー数異常検出のための可視化手法, 日本データベース学会 Letters (DBSJ Letters), Vol.4, No.3, pp.1-4 (2005).
- 6) Montgomery, S., et al. : Sockeye : A 3D Environment for Comparative Genomics, Genome Research, Vol.5, pp.956-962 (2004).
- 7) MITOMAP : A Human Mitochondrial Genome Database (2007), <http://www.mitomap.org/>
- 8) <http://mitowheel.org/>
- 9) <http://genodive.org/>
- 10) <http://www.cytoscape.org/>
- 11) <http://ecell3d.iab.keio.ac.jp/>
- 12) <http://github.com/jandot/parp/tree/master>

(平成21年7月24日受付)

西村 邦裕 (正会員)

kuni@cyber.t.u-tokyo.ac.jp

東京大学大学院情報理工学系研究科助教。2001年東京大学工学部卒業、2006年同大学院工学系研究科博士課程修了、博士(工学)。バーチャルリアリティ技術を用いた多量情報およびゲノム情報の可視化研究に従事。

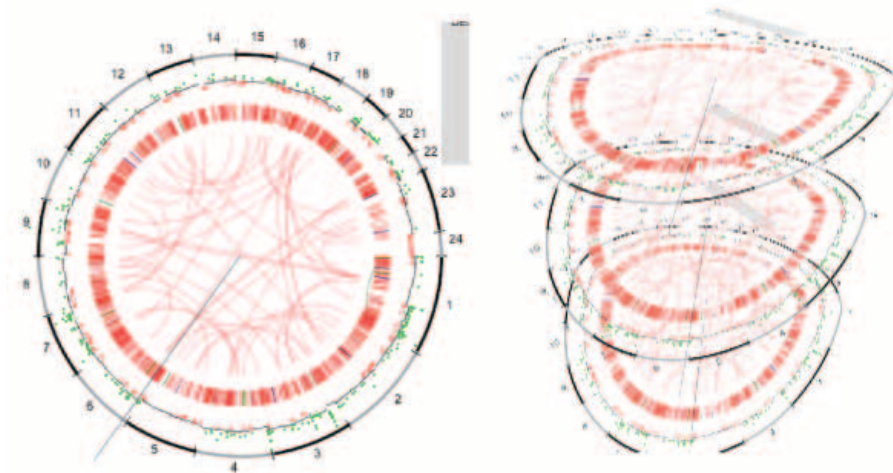


図-11 ヒトゲノム情報のコピー数変化同士の関係を可視化したpARP¹²⁾と3次元pARPの試み



図-12 壁面型ディスプレイを利用したゲノム情報の可視化