ゲノムコピー数異常検出のための可視化手法

Visualization Methods for Detection of Genomic Imbalance

西村 邦裕T♥ 石川 俊平♦ 広田 光一■
油谷 浩幸♠ 廣瀬 通孝♦

Kunihiro NISHIMURA Shumpei ISHIKAWA
Koichi HIROTA Hiroyuki ABURATANI
Michitaka HIROSE

ゲノムサイエンス分野において，DNA チップを利用した実験結果はほとんど公共データベースに格納され，公開され始めている．これらのデータを利用して，医学・生物学的知見を発見する解析作業が重要となっている．疾患に関連しているため重要と考えられるゲノムコピー数情報が，DNA チップ技術の進展により，対立遺伝子レベルで網羅的に取得でき，コピー数を予測することも可能となった．解析には情報の整理・抽出，関係性の把握が必要となる．本論文では、ゲノムサイエンスの研究者の要請および思考方法を考慮に入れ，可視化によりゲノム情報解析支援をする手法を提案する．染色体イメージを利用したコピー数異常領域を検出する可視化手法を提案し，データの抽出・強調方法を提案する．

In the field of genome science, experimental data taken by DNA chips are accumulated rapidly in the public database. Genome researchers want to analyze these data and to get medical and biological knowledge. Using DNA microarray technology, we can estimate copy number of each allele for all genome. These data are important because they relate to some diseases. Genome researchers need to analyze the data and to extract relationships between them. In this paper, we developed visualization methods of these data to support the analysis. The visualization methods provide us to compare allelic copy numbers over many samples and to detect chromosomal abnormal region in 2 ways. First one is to visualize only the abnormal regions; second one is to emphasize the abnormal regions using blinking, color, and shape.

1. はじめに

ゲノムサイエンス分野において，ヒトゲノム塩基配列の解読が完了した．DNAチップを利用するマイクロアレイ技術に代表されるように，解読結果を利用した，ゲノム全体に渡って網羅的に情報を調べる技術が確立され，ほとんどのデータが論文公開とともにNCBI[1]やUCSC[2]といった公共データベースに蓄積されるようになった．DNAチップを利用したゲノム全体にわたるゲノムコピー数の取得技術(Genotyping microarray技術)が生まれ[3][4]，NCBIのGene Expression Omnibus[5]といったデータベースに実験データが格納され始めてきている．

癌などの遺伝子の変異による疾患になるとゲノムのコピー数，つまり染色体の本数が増加したり減少したりするという異常が起きる．そのため，ゲノムのコピー数と疾患などの関係を明らかにすることで，疾患のメカニズムが解明できるのではないか，と考えられている．DNAチップを利用することで，様々な疾患サンプルについて大量のゲノムコピー数情報を得ることが可能になったため，この新しいデータに対する解析手法が要請されている．

しかし，ゲノムのコピー数を網羅的に推定する技術が登場してきたものの，データベースに格納されている生データを目で見ていくことは不可能である．ゲノムサイエンスの知見を総合して情報を解読していくためには，ゲノムサイエンス研究者の思考に沿ってデータを絵として見えるように（可視化）し，インタラクティブに解析していく必要がある．

筆者らはこれまでに，遺伝子発現量情報の解析支援システムを，没入型多面ディスプレイを用いた可視化によるアプローチで構築してきており，データを可視化することが有効であることを示してきた[6][7]．本研究では，ゲノムのコピー数という新しい情報について解析を支援するシステムの構築を目指す．そこで本研究の目的は，ゲノムのコピー数（染色体の欠失・増幅）を処理し，わかりやすい形で提示すること，つまり可視化することで，ゲノムのコピー数異常領域を検出する支援手法の構築，である．

2. ゲノムのコピー数

2.1 ゲノムのコピー数とは

ヒトの場合，染色体には父親由来の部分と母親由来の部分があり2本で対をなし，合計46本（22対×2本+性染色体2本）ある．その対を成す父親・母親由来の遺伝子を対立遺伝子と呼ぶ．癌などの遺伝子の変異に起因する疾患になると，染色体の合計本数が通常46本のところが，50本にも100本にもなることが多い．またダウン症では約9割が，21番染色体が3本になることに起因することが知られている．さらに染色体の一部あるいは全体が増幅・欠失する場合もある．片方の対立遺伝子が欠失していた場合，もう片方が傷つく，または，不活化した場合に，機能が補完されずに病気が発症してしまう．以上のような背景から，ゲノムコピー数と疾患の関係を見ることで，疾患メカニズムの解明につながると考えられている．

対立遺伝子のコピー数が0本のところはLOH（loss of heterozygosity，ヘテロ接合性消失）と呼ばれ，正常組織ではLOHになっておらず，癌などの腫瘍組織においてLOHが起きている部分に癌抑制遺伝子の存在が示唆されている．また，対立遺伝子の1対ともに0本になっているときはHomozygous deletion(ホモ接合性欠失)と呼ばれ，ゲノム構造の変異が疾患に結びついていると考えられている．

2.2 ゲノムコピー数解析への要請と問題点

DNAチップから取得されたゲノムコピー数の解析にあたり，

♥学生会員　東京大学大学院 工学系研究科 博士課程
kuni@cyber.rcast.u-tokyo.ac.jp
♦東京大学先端科学技術研究センター
shumpei@genome.rcast.u-tokyo.ac.jp,
hirose@cyber.rcast.u-tokyo.ac.jp
■東京大学大学院 新領域創成科学研究科
hirota@cyber.rcast.u-tokyo.ac.jp
♠東京大学 国際・産学共同研究センター
haburata-tky@umin.ac.jp

ゲノムサイエンス研究者からの要請は以下の3点であった．

1. 実験ノイズへの対策
2. 対立遺伝子レベルでのコピー数の推定
3. 染色体方向(染色体上での遺伝子の位置情報)を利用して整理してコピー数を見られること

コピー数は0本，1本，2本，･･･と自然数を取るために離散化して本数を推定することを求められた．1，2に対しては，Gaussian Mixtureといった確率モデルを利用した手法[8]やHMMを利用した手法[9]などが提案されている．しかし，3の全体像として染色体方向に整理して把握する手法は提案されていない状態であった．

そして解析に対する問題点として，以下の3点があった．

1. インタラクティブな情報の抽出ができない
2. 複数サンプルに渡る比較手法がない
3. 対立遺伝子1対内の関係性が同時に見えない

情報の抽出については，領域が小さい部分を評価したい場合，LOHや増幅，Homozygous deletionなどそれぞれの事象に注目して見たい場合，染色体全体像と詳細像の両方を切り替えて見たい場合，などがあった．また，対立遺伝子1対のコピー数を両方見たい，あるいは，片方を見ているときにもう片方のコピー数も知りたい，という要求があり，それに答える手法が存在していなかった．

さらに課題は，ゲノム研究者にわかりやすい可視化手法や解析手法を提供することであった．そこで，本論文ではゲノム研究者の思考方法を考慮に入れた可視化手法を提案する．上記の問題点を解決する，解析に適し，小さい領域を見出せ，かつ，複数サンプルが比較できる手法である．

2.3 ゲノム研究者の思考方法

共同研究の中で，ゲノムサイエンス研究者の思考方法を観察した．研究者は，有名な遺伝子の染色体上における位置，ある疾患におけるゲノムコピー数の増幅・欠失の報告事例，などの知識を共通して持っていた．つまり，染色体という空間地図と，疾患とコピー数の増幅･欠失の情報を持っており，可視化された結果を見ただけで，その結果の妥当性を判断することができることがわかった．

3. ゲノムコピー数の可視化

3.1 基本方針

上述した問題点や要請にこたえる可視化手法の基本的方針として，以下の 3 点を提案する．

1. 染色体の位置情報を利用した情報の統合
2. 染色体のバンドイメージを利用
3. 必要に応じた情報の抽出と強調

バンドイメージは，染色体をギムザ染色したものから得られるバンドパターンのことであり，一般的に利用される．濃いバンドは塩基配列 AT が豊富，薄い部分は GC が豊富な部分であり，塩基配列の傾向を表しているものである．

研究者が染色体の空間地図を知識として持っていることから，空間地図を示す染色体バンドイメージを利用して可視化することにする．また，染色体方向で情報を統合することにより，バンドイメージとも矛盾なくデータを可視化することができる．

また，必要な情報にアクセスできるようにするための，抽出と強調を備えたインタラクション手法を提供する．

上記の方針を元にデータ処理・可視化手法を開発した．

3.2 ゲノムコピー数検出処理

ゲノムコピー数は，ひとつのサンプルにおいて，正常細胞と疾患細胞のDNAチップ出力結果の比をとることにより計測する．正常細胞とのコピー数比を見ることで，コピー数を同定する．しかし，データには実験ノイズが入っているために，データ全体を補正する必要がある．本論文では，本グループが開発したGIM(Genome Imbalance Map)[10]という5つの実験パラメータに対して，最小二乗法などで補正するアルゴリズムを用いて，データ補正を行った．本手法を用いることで，従来手法とは異なり，対立遺伝子レベルでのゲノムのコピー数が計測できる．

ここで便宜的に1対の対立遺伝子をそれぞれ上側対立遺伝子と下側対立遺伝子と呼ぶことにする．対立遺伝子のコピー数が1対の中で異なる場合，コピー数が多いほうを上側，少ないほうを下側対立遺伝子と定義する．例えば，対立遺伝子のコピー数が2本と0本からなる場合，2本のほうが上側対立遺伝子，0本のほうが下側対立遺伝子である．ここで，対立遺伝子 i の持つデータの値を上側･下側対立遺伝子の順に（G_{hi}, G_{li}）とする．

DNAチップからの実験結果を，染色体方向を横軸にとり，グラフを作ると図1のようになる．対立遺伝子のコピー数が最低0本からなるとすると，0本，1本，2本，･･･をグラフから読み取ることが出来る．

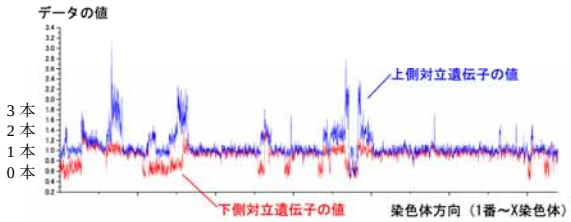

図 1　ゲノムのコピー数データ
Fig.1　Raw Data of Genomic Copy Number

3.3 コピー数推定とゲノムコピー数の可視化

図 1 で示したデータからゲノムのコピー数の推定を行った．推定は，まず（G_{hi}, G_{li}）に対して，染色体方向に移動平均（$\grave{G}_{hi}$, $\grave{G}_{li}$）をとり，データの値に対する（$\grave{G}_{hi}$, $\grave{G}_{li}$）のヒストグラムをとった．ヒストグラムの山の分布から，図 1 における最小の値のグループ（山）が染色体の本数 0 本，次が本数 1 本，として，コピー数なので自然数を取ることを仮定し，本数推定を行った．その結果が図 2 である．

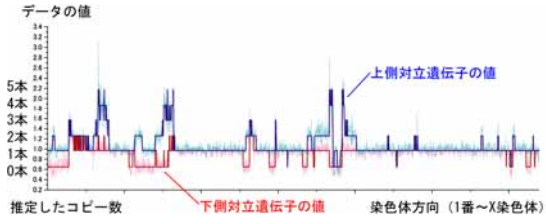

図 2　ゲノムコピー数のデータと本数推定の結果
Fig.2　Estimation Result of Genomic Copy Number

この結果を，染色体情報に合わせ，バンドイメージを利用して可視化する．可視化手法としてわかりやすいように，以下の 3 点を利用した手法を提案する．

1. 染色体のコピー数を染色体の本数で可視化すること

2. 対立遺伝子ごとに色分けして可視化すること
3. 染色体の全体像と 1 本 1 本に見せる詳細像を切り替えて見せること

図 2 のデータをもとに可視化した結果が図 3（全体像）と図 4（各染色体ごとの像；1 番染色体）である．バンドイメージは，UCSC のサイト[2]からデータに基づいて作成している．

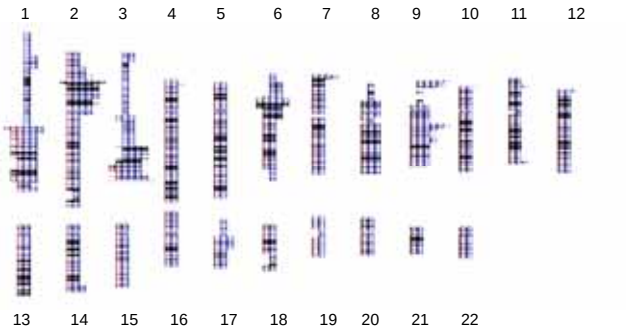

図 3　染色体バンドイメージを利用したゲノムコピー数の可視化（染色体全体像）

Fig.3　Visualization of Genomic Copy Number Data using Chromosomal Band Image

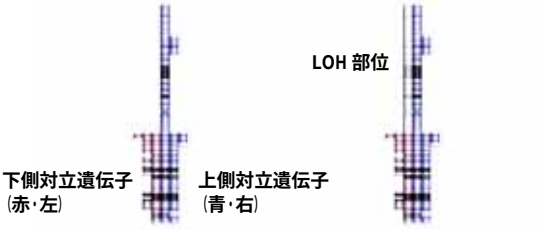

図 4　染色体 1 番のゲノムコピー数の可視化と LOH 部位に影をつけた場合（右）

Fig.4 Visualization of Genomic Copy Number Data on Chromosome 1 and Shadow Visualization of LOH (left)

3.4 コピー数異常領域検出のための可視化

コピー数の異常領域を容易に検出しやすくする可視化手法として以下の 2 つの方法が考えられる．

1. 異常領域のみを抽出して可視化
2. 異常領域を強調して可視化

異常領域としては，染色体 0 本である LOH 部位，染色体の増幅部位，Homoxygous deletion 部位などがあげられ，抽出する場合は，インタラクティブに抽出部位を変更できるようにする．また，強調する手法として，点滅，色の変更，動き（振動），形の違いを利用することにする．

例えば図 4 左の場合 LOH 部分は表示していない．その LOH 部位に関してのみ影をつけたのが図 4 右である．LOH 部位の検出をしやすくするための仕組みとして，影をつけた LOH 部位を点滅させる，つまり図 4 左と図 4 右を交互に表示させることを利用する．

3.5 複数サンプルを比較する可視化手法

複数サンプルにわたって，対立遺伝子ごとにコピー数を比較することができる可視化手法として，対立遺伝子を交互に色で区別しながら並べる手法を考案した．X 軸方向にサンプル，Y 軸方向に染色体，Z 軸方向にコピー数を取り，3 次元上に可視化した結果が図 5 である．

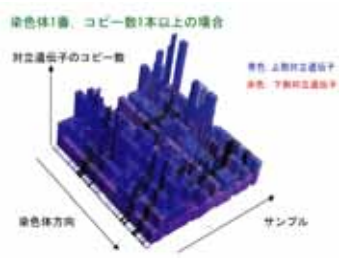

図 5　複数サンプルにおける対立遺伝子ごとのゲノムコピー数の可視化結果（染色体 1 番）

Fig.5　Visualization of Genomic Copy Number Data across Multiple Samples (Chromosome 1)

図 5 ではコピー数 1 本以上を表示しているが，異常領域を抽出するために，表示するコピー数をユーザがインタラクティブに変化させることを可能にした．具体的には，異常領域である LOH 部位のみ表示，その他の増幅部位に関しては閾値の本数を決め，その本数以上の部分の表示を可能とした．図 6 が全染色体のコピー数 0 本の LOH 部位のみを表示した場合，図 7 は染色体 1 番についてコピー数が 0 本のみ，1 本以上，2 本以上を表示した場合である．コピー数が 0 本と 1 本以上は相補的であることがわかる．これらの方法によって，増幅部位のみ表示すること，LOH 部位のみ表示することが可能になり，把握が容易になる．

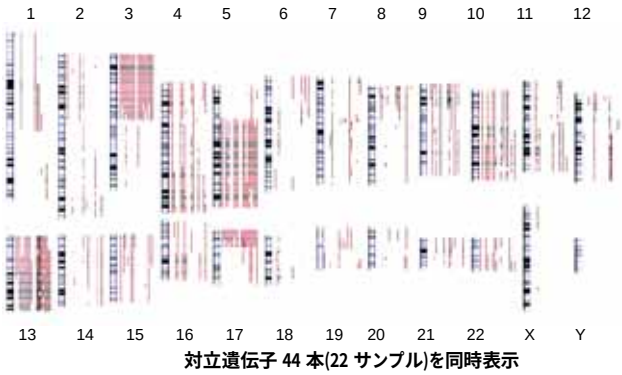

図 6　対立遺伝子のコピー数 0 本（LOH）部位の場合

Fig.6　Visualization of LOH Region

図 7　コピー数 0 本，1 本以上，2 本以上の対立遺伝子

Fig.7　Visualization of Allelic Copy Number is 0, more than 1, and more than 2 (Chromoseme 1)

また，対立遺伝子レベルで表示しているために，染色体が

完全に0本になってしまうhomozygous deletion部位を可視化することができる．この部位は小さい領域であることが多いため，検出しやすいように振動をさせることを行った．振動は手前に飛び出してくるようにし，表示させている幅を2倍と大きくして，強調させる可視化を行った．これにより小さい領域の検出も可能となる．例えば染色体9番における可視化結果が図8である．染色体9番短腕（染色体の上側半分）には癌抑制遺伝子に近い性格を持つ p16 という遺伝子があり，癌などの疾患において欠失が起きることが有名であり，その部分が表示されていることが確認できた．

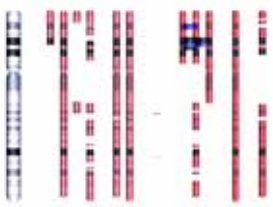

図8 染色体9番のHomozygous Deletion部位の強調（振動）

Fig.8 Emphasis (Vibration) of Homozygous Deletion at Chromosome 9

図9 LOH部位における上側対立遺伝子の本数を色と高さで可視化

Fig.9 Visualization of Allelic Copy Number in the LOH region using Color and Height

さらに，対立遺伝子1対内の関係，つまり，下側対立遺伝子が 0 本（LOH）の時，もう片方の対立遺伝子のコピー数を同時に見せる手法を開発した．これは対立遺伝子の一方が欠失しているときに，もう片方がそのままの本数で合計1本か，あるいはもう片方の対立遺伝子が2本になって補完して，合計の染色体の本数は見かけ上通常と同じ2本か，を知るためである．そのため，下側対立遺伝子のLOH部分のみをベースとして，上側対立遺伝子のコピー数を色や高さで表示できるようにした．染色体9番について，色を利用して上側対立遺伝子のコピー数が2本以上のところを色付けと高さで表現した結果が図9であり，補完の有無などを把握することが可能である．

4. まとめと今後の課題

本研究では，公共データベースに蓄積されている情報の解析手法として，ゲノムコピー数の異常領域の可視化手法を提案した．ゲノムサイエンスの研究者の知識である染色体地図を利用し，異常領域を検出しやすくするために点滅や振動といった動きを利用した可視化手法を提案した．本可視化手法は実用性が高いため，現在，ゲノムサイエンスの研究者に利用されている．今後，生データや他の情報を同時に見たい，などという要求があるために，生物学的情報を含む公共データベースを利用した既知情報の重畳提示などを考えている．

[謝辞]

本研究の一部は，文部科学省科学研究費補助金(特別研究員奨励費 15-11570)の援助を受けた．ここに深く感謝の意を表する．

[文献]

[1] NCBI (National Center for Biotechnology Information, National Library of Medicine, National Institute of Health, U.S.A), http://www.ncbi.nlm.nih.gov/
[2] UCSC(University of California, Santa Cruz) Genome Bioinformatics Site, http://genome.ucsc.edu/
[3] Ishkanian AS, et.al, "A tiling resolution DNA microarray with complete coverage of the human genome", Nature Genetics, Vol.36(3), pp.299-303, 2004.
[4] Zhao X, et.al, "An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays", Cancer Research, Vol.64(9), pp.3060-3071, 2004.
[5] Gene Expression Omnibus(GEO), NCBI, NIH, U.S.A., http://www.ncbi.nlm.nih.gov/geo/
[6] Nishimura K, et.al, "Virtual Environment Design Guidelines for Gene Expression Analysis: The Utility of a Lab Bench Metaphor and a Road Metaphor", IEEE Virtual Reality Conference 2004, pp.247-248, 2004.
[7] Nishimura K, et.al, "Virtual Environment Design for Gene Selection Using Gene Expression Data", 10th International Conference on Human - Computer Interaction (HCI International 2003), Vol.1, pp.1213-1217, 2003.
[8] J. Wang, et.al, "Analysing Microarray-based CGH Experiments", BMC Bioinformatics, Vol.5, p.74, 2004.
[9] M. Lin, et.al, "dChipSNP: Significance Curve and Clustering of SNP-Array-Based Loss-of-Heterozygosity Data", Bioinformatics, Vol. 20, No. 8, pp. 1233–40, 2004.
[10] Ishikawa S, Komura D, Tsuji S, Nishimura K, et.al, "Allelic dosage analysis with genotyping microarrays", Biochem Biophys Res Commun., Vol.333(4), pp.1309-1314, 2005.

西村 邦裕 Kunihiro NISHIMURA
東京大学大学院工学系研究科 博士課程在学中．日本学術振興会特別研究員．2003 東京大学大学院情報理工学系研究科修士課程修了，VR 技術のゲノム科学への応用や情報の可視化の研究に従事．日本データベース学会学生会員．

石川 俊平 Shumpei ISHIKAWA
東京大学先端科学技術研究センター 特任助手．2000 東京大学医学部卒業，2004 同大学大学院医学系研究科 博士課程修了，博士（医学）．病理学および DNA チップを利用したゲノム情報解析研究に従事．

広田 光一 Koichi HIROTA
東京大学大学院新領域創成科学研究科 助教授. 1994 東京大学大学院産業機械工学専攻 博士課程修了，博士（工学）．2000 東京大学先端科学技術研究センター助教授．主にヒューマンインタフェースの研究に従事．

油谷 浩幸 Hiroyuki ABURATANI
東京大学国際・産学共同研究センター 教授．1980 東京大学医学部卒業．博士（医学）．1988 東京大学医学部第三内科助手, マサチューセッツ工科大学癌研究センター研究員．1999 東京大学先端科学技術研究センター助教授． 主にゲノムサイエンスの研究に従事。

廣瀬 通孝 Michitaka HIROSE
東京大学先端科学技術研究センター 教授．1982 東京大学大学院産業機械工学専攻 博士課程修了，工学博士．1982 東京大学工学部専任講師，1983 同大学助教授，1999 同大学大学院工学系研究科教授．主にシステム工学，ヒューマンインタフェース，バーチャルリアリティの研究に従事．