染色体異常領域検出のための遺伝子位置情報と発現情報の統合

Integration of Gene Locus Information and Expression Data for Detection of Chromosomal Structural Abnormalities

加納 真♠　石川 俊平♣　油谷 浩幸♣

Makoto KANO　Shumpei ISHIKAWA
Hiroyuki ABURATANI

遺伝子位置情報と発現情報の統合は，癌発生のメカニズムを担う染色体異常領域（欠損／増幅）を推定する上で重要である．本稿では，染色体の状態（欠損／正常／増幅）を隠れ状態，発現量を出力値とした，隠れマルコフモデルを用いた情報統合手法を提案する．また，肺癌細胞株から得られた遺伝子発現データに本手法を適用することにより，その有効性を示した．

Integration of locus information and expression profiles should be effective in detecting chromosomal structural abnormalities such as genomic gains and losses. In this article, a new integration method based on Hidden Markov Model (HMM) is described. We applied this novel method to gene expression data extracted from lung cancer cell lines and confirmed its effectiveness compared to conventional methodologies.

1. はじめに

近年，bioinformaticsの分野では多種多様な生物情報が急速に蓄積・公開されつつあり，これらの情報の統合解析が重要な課題となっている．しかしながら，新しい生物学的知見を生み出すためには，SQLで記述されるような単純な条件を用いたデータベース間の統合では不十分で，より高度な情報技術の適用が必要となる場合が多い．

様々な生物情報の統合の中に，染色体異常領域の検出を目的とした，遺伝子位置情報と発現情報の統合がある．遺伝子位置情報とは，文字通り各遺伝子の染色体上の位置を指す．ヒトゲノムプロジェクト[1]により，多くの遺伝子の染色体上の位置が解明され，その情報がNCBI[2]などの公共Webサイトで公開されている．一方，遺伝子発現情報とは，本稿においては，遺伝子がタンパク質として発現する際の中間生成物であるmRNAの数を指す．　遺伝子はmRNAに転写(コピー)され，さらにそのmRNAがタンパク質に翻訳されることで発現する．マイクロアレイ技術の発達によって，転写段階における遺伝子発現量，即ちmRNAのコピー数を網羅的に同時測定することが可能となった．近似的にはmRNAのコピー数はタンパク質の数であり，細胞内における各遺伝子の持つ機能の活性化度合いの尺度と見なすことができる．現状においては，この遺伝子発現量が，全遺伝子に関する網羅測定が可能な唯一の生物量であり，生体システムのメカニズム解明に重要な役割を果たすと考えられている．遺伝子発現データは，世界中の研究施設のWebサイト[3-5]で公開されている．

本稿の解析の目的である染色体異常には，染色体増幅と染色体欠損の二種類が存在する．染色体増幅とは，染色体上の数メガ塩基対（base-pairs 以下bpと記す）程度の部分領域が複数コピーされて縦列に挿入される現象である．一方，染色体欠損とはその逆で，染色体からある部分領域が完全に抜け落ちてしまう現象のことである（図1）．癌抑制遺伝子が含まれる領域が染色体欠損し癌抑制の機能が失われたり，癌遺伝子の含まれる領域が染色体増幅し癌発生・進行の作用が増幅されることがあり，欠損・増幅といった染色体異常は発癌や癌の悪性度決定に大きく寄与すると考えられている．逆に言えば，染色体異常領域に含まれる遺伝子は，癌のメカニズムにおいて重要な役割を担っている可能性があり，染色体異常領域の同定は極めて重要である．

図 1　染色体増幅と欠損

Fig.1　Genomic Gains and Losses.

一般に，欠損領域(Loss)の遺伝子の発現量は減少し[6]，増幅領域(Gain)の遺伝子の発現量は増加する傾向がある[7]．図2は，肺癌細胞株において染色体欠損領域(Loss)，正常領域(Normal)，増幅領域(Gain)で測定された発現量と，正常検体のゲノム（全て正常領域）から測定された発現量の比（*fold change*)の対数値の度数分布を表している．以下，本文では簡単のため，この*fold change*の対数値を，単に「遺伝子発現量」と呼ぶこととする．

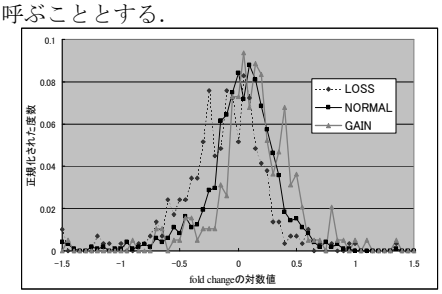

図 2　染色体の状態ごとの遺伝子発現量の増減の度数分布

Fig.2　Histograms of Changes in Gene Expression by Chromosomal States

染色体の状態と遺伝子発現量には相関関係があるため，染色体上の遺伝子発現の増減から，染色体欠損・増幅領域を推定することができると期待されている．Genome-wide transcriptome map[8]は，発現量を単純に染色体上にグラフとしてマップすることによって，発現量の増減と位置情報の関係を表現した．しかし，生体内には複数の制御パスウェイが存在し，多様なフィードバック制御を受けるため，染色体増幅領域の遺伝子の発現量が逆に減少していたり，染色体欠損領域の遺伝子の発現量が逆に増加している場合がある．こ

♠正会員　日本アイ・ビー・エム（株）東京基礎研究所
mkano@jp.ibm.com
♣東京大学先端科学技術研究センター システム生物医学ラボラトリー 癌システム生物学部門
shumpei@genome.rcast.u-tokyo.ac.jp
haburata-tky@umin.ac.jp

のため，単純なマッピングではノイズに埋もれて有益な情報を得ることは難しい．異常領域を推定するためには，単に個別の遺伝子に関する発現の増減に着目するだけでは不十分で，遺伝子の発現の増減を領域として評価する必要がある．Expression Imbalance Map[9]は，領域内で発現が一定基準以上増加（／減少）した遺伝子の数を超幾何分布としてモデル化し，発現量の増減を領域として評価した．しかし，遺伝子間の距離を考慮したモデルではなかったため，予測精度が十分ではなかった．

本稿では，染色体の状態（Loss／Normal／Gain）を隠れ状態，発現量を出力値とした，隠れマルコフモデルを用いた解析手法を提案する．提案モデルは，ノード（遺伝子）間の距離に依存する状態遷移確率を導入し，遺伝子間距離を考慮した染色体異常領域の推定を行う．本手法を肺癌細胞株から得られた遺伝子発現データに適用することによって，従来手法を上回る精度と再現率で染色体異常領域を予測できることを確認した．

2. 隠れマルコフモデルによる異常領域推定

2.1 モデル化の概要

隠れマルコフモデル(Hidden Markov Model: 以下HMMと記す)とは，隠された状態遷移系列と，観測される出力系列からなるモデルである．出力値は，その時点の状態に応じて異なる確率分布で出力されると仮定し，観測された出力系列から隠れた状態遷移系列を予測する．状態遷移確率，状態ごとのシンボル出力確率が与えられた時，最適状態遷移系列を動的計画法で効率よく求める方法が知られている（Viterbiアルゴリズム）[10].

染色体の状態と遺伝子発現量の関係は，このHMMによってモデル化が可能である．図3に示したように，各遺伝子から観測される遺伝子発現量は，その遺伝子が存在する周辺のゲノムの状態（Loss／Normal／Gain）ごとに異なる分布を示す．このため，網羅的な観察が可能な出力量（遺伝子発現量）から，隠れ状態（染色体の状態）を推測する，HMMの問題とみなすことができる．遺伝子間距離が非等間隔であり，また染色体上の遺伝子数が多いため，Support Vector MachineやNeural Networksといったその他のPattern Classification技術[11]は適用が難しい．その点，HMMの場合は，染色体上の全遺伝子の発現量を考慮した上で，最適な状態遷移系列を高速に求めることが可能である．HMMは，音声認識[12]や遺伝子発見[10]などの様々な分野で幅広く利用されているが，遺伝子発現量と染色体の状態の関係への適用は本研究が初めてである． 本稿では，染色体の隠れ状態 S_i（i = 1,2,3）を以下のように定義し，

S_1 : Loss, S_2 : Normal S_3 : Gain

状態ごとに異なる確率分布から生成される遺伝子発現量が観測されるようなHMMを考える．本稿で扱うモデルの，従来のHMMと比較した際の特徴は以下の二点である．

(1) 出力値（遺伝子発現量）は連続量

各ノード（遺伝子）からの出力値は連続量である．本稿では，状態 S_i における遺伝子発現量の出力の確率密度関数を平均 μ_i 分散 σ_i の正規分布としてモデル化する．即ち，遺伝子 g_k が状態 S_i である場合に，出力される遺伝子発現量が e_k～ $e_k + \Delta e$ の範囲である確率 $b_i(e_k)$は，

$$b_i(e_k) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(e_k-\mu_i)^2}{2\sigma_i^2}\right) \times \Delta e$$

で与えられるものとする．

(2)状態遷移確率は，ノード（遺伝子）間の距離に依存

遺伝子は非等間隔で染色体上に存在するために，遺伝子 g_{k-1} と次の遺伝子 g_k 間の状態遷移確率は，遺伝子間の距離 l_k の関数となる．即ち，距離が近い遺伝子ほど同じ状態となる確率が高く，距離が離れるほど前の遺伝子の状態に依存せずに状態を取り得る．遺伝子間距離に依存する遷移確率を算出するために，一時的に全ての塩基をノードとしたHMMを考える(図 3)．簡単のため，遺伝子の塩基長を考慮せず染色体上のある１塩基で遺伝子を表すものとする．

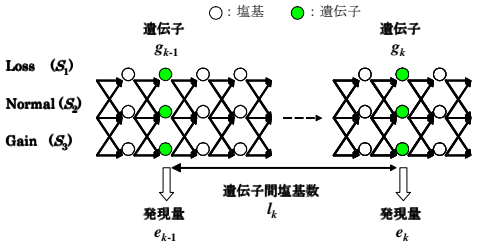

図３ 塩基をノードとした状態遷移

Fig.3 State Transitions when Nodes Are Bases

ここで，ある塩基の状態を S_i とした時，次の塩基が状態 S_j になる遷移確率 q_{ij} とその遷移確率行列 Q を次のように定義する．

$$Q = \{q_{ij}\} = \begin{pmatrix} 1-\alpha & \alpha & 0 \\ \beta_1 & 1-\beta_1-\beta_2 & \beta_2 \\ 0 & \gamma & 1-\gamma \end{pmatrix}$$

（ただし，$0 < \alpha, \beta_1, \beta_2, \gamma < 1$ ）

この Q を用いて，状態 S_i のある塩基の l 塩基後が状態 S_j である遷移確率 $q_{ij}(l)$ とその遷移確率行列 $Q(l)$ を Q^l として算出する．

$$Q(l) = \{q_{ij}(l)\} = Q^l$$

なお，Loss 領域と Gain 領域が直接隣接することはないものとして，q_{13}= q_{31}=0 とした．間に１塩基以上の正常領域をはさめば，Loss 領域の後に Gain 領域が存在することを許容しているため，実用上この仮定は問題ないと考えられる．この仮定を置くことにより，Q は正則な遷移確率行列となり，$l \to \infty$の時，$Q(l)$が以下の行列 A に収束することが保証される．

$$Q(\infty) = \lim_{l\to\infty} Q(l) = \lim_{l\to\infty} Q^l = \begin{pmatrix} \vec{w}_1 \\ \vec{w}_1 \\ \vec{w}_1 \end{pmatrix} = A$$

ここで $\vec{w}_1$ は，$\vec{w}_1 Q = \vec{w}_1$ を満たす Q の固有値１に対応する固有ベクトルで，本稿では収束状態確率と呼ぶこととする．

$$\vec{w}_1 = (\hat{w}_1, \hat{w}_2, \hat{w}_3) = \frac{1}{\alpha\beta_2 + \beta_1\gamma + \gamma\alpha}(\beta_1\gamma, \gamma\alpha, \alpha\beta_2)$$

2.2 最適状態遷移系列の推定

本稿のHMMは，遺伝子を表す塩基のみに着目した状態遷移を扱う（図4）．k番目の遺伝子(g_k)から出力される発現量をe_k ，隠れた状態をX_kとし，k-1番目の遺伝子(g_{k-1})からの距離（塩基数）をl_kとする．ここで，今着目している染色体上の遺伝子の数をNとする．また，出力遺伝子発現量系列 E_m^n（ただし，$m \leqq n$）を以下のように定義する．

$$E_m^n = \{e_m, e_{m+1}, ..., e_l, ..., e_{n-1}, e_n\}$$

ただし e_lは遺伝子 g_lから観測された発現量

また，遺伝子g_mから遺伝子g_nまでの状態の遷移の系列X_m^nを以下のように定義する．

$$X_m^n = \{X_m, X_{m+1}, ..., X_l, ..., X_{n-1}, X_n\}$$

ただし X_l は，遺伝子 g_lの隠れ状態

ここで，出力系列E_1^Nに対する最適な状態遷移系列X_1^N，即ち，$P(X_1^N, E_1^N)$を最大化する最適状態遷移系列を動的計画法（Viterbi アルゴリズム）によって求める[10].

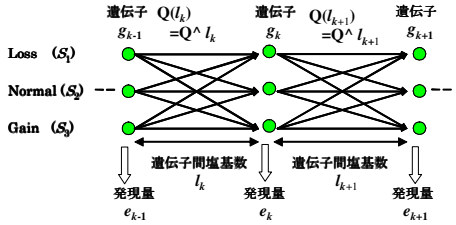

図 4　遺伝子をノードとした状態遷移

Fig.4　State Transitions when Nodes Are Genes

2.3　パラメタの決定

本稿では HMM のパラメタを以下のように決定した.

(1) 遺伝子発現量の出力の確率密度関数の平均と分散

ゲノムの状態がわかっている遺伝子に関して，Loss, Gain，Normal ごとに発現量を測定しパラメタを決定する.

(2) 状態遷移確率行列

ユーザからの二種類の入力を基に，パラメタを決定する.

(2-1) Loss, Gain 領域の長さの期待値

(Loss の長さの期待値)

$$= \sum_{k=1}^{\infty} k\,\alpha(1-\alpha)^k = \alpha \lim_{n\to\infty} \sum_{k=1}^{n} k\,(1-\alpha)^k = \alpha \lim_{n\to\infty} \left\{ \frac{1-\alpha}{\alpha^2} - \frac{(1-\alpha)^{n+1}}{\alpha^2} - \frac{n(1-\alpha)^{n+1}}{\alpha} \right\}$$

$$= \frac{1-\alpha}{\alpha} \quad (\because 0<1-\alpha<1) \qquad \approx \frac{1}{\alpha} \quad (\because \alpha \ll 1)$$

同様にして，(Gain の長さの期待値) $\approx \frac{1}{\gamma}$

(2-2) Loss, Normal, Gain の領域の占有比の期待値

収束状態確率より，占有比の期待値は以下のようになる.

Loss : Normal : Gain = $\beta_1\gamma : \gamma\alpha : \alpha\beta_2$

一般に，bioinformatics の分野で用いられている解析ツールは，様々なパラメタの設定を必要とする．しかし，それらのパラメタはプログラム作成者側の視点に基づいていて，ユーザである生物学者にとって解釈しづらい場合が多い．そこで本稿では，パラメタと具体的な生物量を関連付けることによって，直観的にパラメタを解釈できるように配慮した.

3. 評価実験

3.1　実験データ

Affimetrix 社の発現マイクロアレイ(GeneChip U133AB) [13]を用いて測定された，肺癌細胞株 6 検体と正常細胞 1 検体に関する遺伝子発現量データ（未発表データ）に対して本手法を適用し，染色体欠損・増幅領域を予測した．また，本手法と Expression Imbalance Map（以下 EIM）[9]の比較実験を行った.

[遺伝子発現情報と位置情報の対応付け]

GeneChipU133AB 上の 44592 個のプローブのうち, 最終的に 12485 個をゲノム上にマップした.

(1) 以下の閾値を満たすプローブを取り出す
　(正常検体での発現量)が 40 以上　もしくは,
　(肺癌細胞株検体での平均発現量)が 40 以上

(2) (1)で取り出したプローブのターゲット配列をクエリとして，BLAST(build31) [2]で染色体上の位置を検索.

(3) *fold change* の対数値を計算

一方の発現量の絶対値が極端に小さい場合の誤差の影響を緩和するために，定数 C を分母と分子に加えて *fold change* を算出する．今回実験に用いた発現マイクロアレイでは，発現量の絶対値が 20 以下のものは SN 比が小さく，信頼性が低いと考えられている．このため，本稿では C=20 として解析を行った.

$$\text{fold change の対数値} = \log\left\{\frac{(\text{癌細胞株での発現量}) + C}{(\text{正常細胞での発現量}) + C}\right\}$$

[正解セット及び評価尺度]

ゲノム上の特定の限られた地点に関しては，欠損・増幅の状態を CGH アレイ[14]によって調べることができる．そこで，Vysis 社の Genosensor 300[15]を用いて，肺癌細胞株 6 検体に関するゲノム欠損・増幅を調べ，その結果を正解セットとして精度評価を行った．なお図 3 は，Loss／Normal／Gain と判定された地点の，近傍 1Mbp(前後 500kbp)に存在する GeneChipU133AB 上のプローブの *fold change* の対数値(3.1(3))の分布である．CGH アレイでゲノムの状態が確認できた地点のうち，近傍 1Mbp 以内に，5 個以上 3.1(1)の閾値を満たす GeneChipU133AB のプローブが存在する 419 箇所について評価を行った. 評価尺度として, precision, recall, f-value を以下のように定義した.

$$precision = \frac{\text{Loss/Gainの正解数}}{\text{Loss/Gainと予測された箇所の数}}$$

$$recall = \frac{\text{Loss/Gainの正解数}}{\text{CGHでLoss/Gainと判定された箇所の数}}$$

$$f\text{-}value = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

3.2　精度・再現率の評価

EIM において，f-value が最大になるようにパラメタ調整を行った結果を表 1 に示す．この結果と比較した際の本稿における　提案手法　の有効性を評価する．図 5 は，Loss:Normal:Gain の占有比を 3:4:3 に固定した状態で, Loss 及び Gain の長さを 1M(bp)から 20M(bp)まで変化させた際の予測結果を示す. 長さを長く設定するほど recall が下がり，precision が上がる傾向が観察されたが, f-value は常に EIM における f-value の最大値を上回った. 特に, Loss 及び Gain の長さが 5M(bp)の時には，recall=0.74, precision=0.61, f-value=0.67 となり, EIM における f-value の最大値を 17%も上回った. また, 図 6 は, Loss 及び Gain の長さを 5M(bp)に固定した状態で, Loss:Normal:Gain の占有比を x:(1-2x):x として変化させた際の予測結果を示す. 異常領域の比率を大きくするほど recall が上がり，precision が下がる傾向が観察されたが, この場合も f-value は常に EIM における f-value の最大値を上回った.

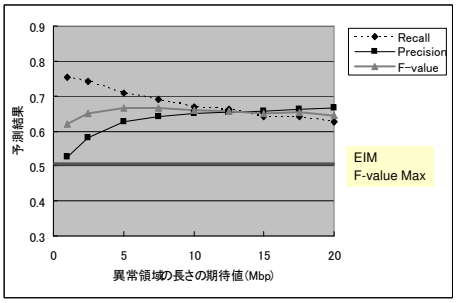

図 5 異常領域の長さの期待値と予測結果

Fig.5　Length of Chromosomal Abnormalities vs. Prediction Results

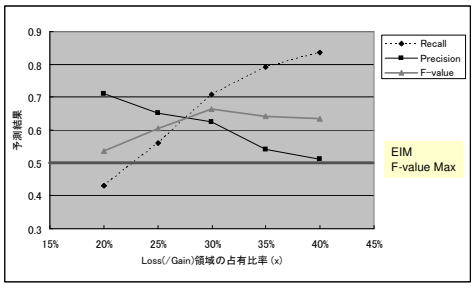

図 6 異常領域の占有比率の期待値と予測結果

Fig.6 Ratio of Chromosomal Abnormalities vs. Prediction Results

	Recall	Precision	F-value
EIM F-value Max	0.65	0.41	0.5

表 1 EIM による予測結果

Table 1 Results of Predictions by EIM

3.3 可視化

図 7 は，Loss:Normal:Gain の占有比を 3:4:3，Loss 及び Gain の長さを 5M(bp)に設定した際の染色体異常領域の予測結果の可視化である．使用した細胞株では男女が混在しており，X 染色体と Y 染色体上の異常領域を発現量から推定することは難しいため解析対象から除いた．Loss と判定された領域を染色体の左側に，Gain と判定された領域を右側に，6 つ癌細胞株検体ごとに並べて表示した．グレイスケールは発現量が観察された条件下における異常領域の事後確率を表しており，輝度が高い領域ほど異常領域である可能性が高い．図 7 右側は 1 番染色体の拡大図である．A の領域は，全ての検体で Loss と予測されているが，B の領域は 4 検体のみ Gain と予測された．このように，染色体異常領域の分布から検体ごとの特徴を読み取ることができる．

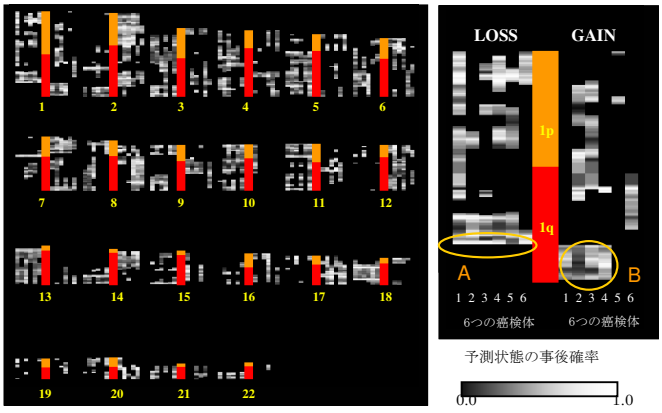

図7 染色体異常領域の予測結果の可視化

Fig.7 Visualization of Prediction of Chromosomal Abnormalities

4. まとめ

本稿では，染色体の状態（欠損／正常／増幅）を隠れ状態，発現量を出力値とする隠れマルコフモデルを用いた，遺伝子発現情報と位置情報の統合手法を提案した．ノード（遺伝子）間の距離に依存する状態遷移確率を導入することで，遺伝子間距離を考慮した染色体異常領域の推定を行い，従来手法を上回る精度と再現率で染色体異常領域 を予測できることを示した．

[文献]

[1] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
[2] http://www.ncbi.nlm.nih.gov/BLAST/
[3] http://research.dfci.harvard.edu/meyersonlab/lungca/
[4] http://microarray.cnmcresearch.org/
[5] http://www.lsbm.org/db/index.html
[6] Mukasa, A. et al.: Distinction in gene expression profiles of oligodendrogliomas with and without allelic loss of 1p. Oncogene, 21: 3961, 2002.
[7] Virtaneva K. et al.: Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. Proc Natl Acad Sci USA 98:1124-1129, 2001.
[8] Fujii T. et al.: A preliminary transcriptome map of non-small cell lung cancer. Cancer Res 62: 3340-3346, 2002.
[9] Kano M. et al.: Expression imbalance map: a new visualization method for detection of mRNA expression imbalance regions. Physiol Genomics. 2003 Mar 18;13(1):31-46. Epub 2003.
[10] Durbin R. et al.: Biological sequence analysis. Cambridge University Press. 1998
[11] Duda R O. et al.: Pattern Classification (Second Edition). John Wiley & Sons Inc. 2000
[12] Rabiner L. & Juang, B H.: Fundamentals of Speech Recognition. Prentice Hall. 1995
[13] http://www.affymetrix.com
[14] Veltman JA. et al.: High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. Am J Hum Genet 70:1269-1276, 2002.
[15] http://www.vysis.com/

加納 真 Makoto KANO
日本アイ・ビー・エム（株）東京基礎研究所副主任研究員．2000 東京大学工学系研究科修士課程修了．遺伝子発現データの解析手法・可視化技術の研究に従事．日本データベース学会会員．

石川 俊平 Shumpei ISHIKAWA
東京大学大学院医学系研究科人体病理学専攻博士課程在学中．2000 東京大学医学部医学科卒業．

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．1988 医学博士取得．