隠れマルコフモデルを利用した遺伝子発現情報と位置情報の統合

加納 真[†] 石川俊平[††] 油谷浩幸[††]

bioinformatics の分野では多種多様な生物情報が存在し，これらの情報の統合解析が重要な課題となっている．その中でも遺伝子発現情報と位置情報の統合は，癌発生のメカニズムを担う染色体異常領域（欠損/増幅）を推定するうえで重要である．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．提案モデルは，(1) 状態遷移確率はノード（遺伝子）間の距離に依存（遺伝子は染色体上で非等間隔），(2) 観察される出力値は連続値（発現量），という 2 つの特性を有し，遺伝子間の距離を考慮に入れて，領域としての発現量の増加・減少を評価することができる．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り，従来手法を上回る精度（61%）と再現率（74%）で染色体欠損・増幅領域を推定できることを示した．

Integration of Gene Expression Data and Locus Information Using Hidden Markov Model

Makoto Kano,[†] Shunpei Ishikawa[††] and Hiroyuki Aburatani[††]

In the field of bioinformatics, integration of various types of biological information is essential. Particularly, integration of expression profiles and locus information should be effective in detecting chromosomal structural abnormalities such as genomic gains and losses. We describe a new method based on Hidden Markov Model (HMM) for detecting chromosomal abnormalities. This method has two features: 1) probabilities of state transitions depend on distances between genes, 2) observable outputs are continuous quantities. 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

bioinformatics の分野で近年注目を集めている生体情報として遺伝子発現情報がある．**図 1** に示すように，遺伝子は中間生成物である 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，すなわち 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，遺伝子間の相互関係を解明するうえで重要な役割を果たすと考えられている．遺伝子発現量解析により，症例の分類[1),2)]や，既存の分類では鑑別困難な癌の予後の経過と遺伝子発現の相関[3)]が明らかにされてきた．従来の遺伝子発現情報の解析の多くは，クラスタリング解析[4),5)]に代表されるように，遺伝子発現データ単独での解析であったが，今後は文献や位置情報，配列情報，蛋白質情報など他の種類の生体情報と統合することが重要な課題となってきている．その中でも，遺伝子発現情報と染色体上の遺伝子の位置情報を統合することは，癌発生のメカニズムを担う染色体異常領域（欠損/増幅）を推定するうえで重要である．

染色体異常には，染色体増幅と染色体欠損の 2 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 2**）．癌抑制遺伝子が含まれる領域が

† 日本アイ・ビー・エム株式会社東京基礎研究所
Tokyo Research Laboratory, IBM Japan Ltd.
†† 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
Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo

図 1 遺伝子の発現の流れ
Fig. 1 Expression from genes to proteins.

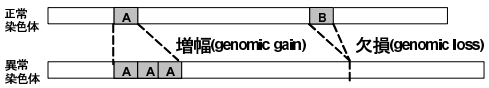

図 2 染色体増幅と欠損
Fig. 2 Genomic gains and losses.

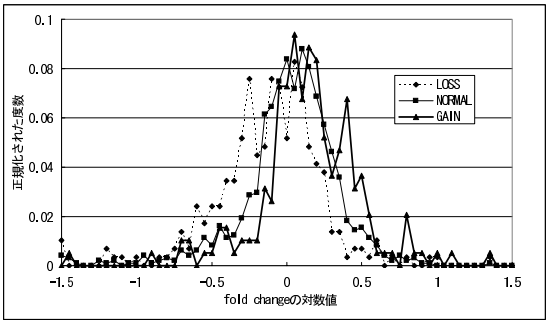

図 3 染色体の状態ごとの遺伝子発現量の増減の度数分布
Fig. 3 Histograms of changes in gene expression by chromosomal states.

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たりすることがあり，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．逆にいえば，癌細胞の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きわめて重要である．

染色体欠損・増幅を，染色体を直接観察することによって調べる手法として，Comparative genomic hybridization（CGH）[6]と CGH アレイ[7]がある．しかし，前者は蛍光顕微鏡によって染色体の状態を観察する方法であるため，解像度が低いという問題があり，10 メガ塩基対（以下 10M (bp) と記す）より細かい染色体の状態の変化を検出することはできない．一方後者は，染色体上の特定の 1 点におけるゲノムの状態を検出する方法であるが，実験コストが高く，また技術的にも全染色体領域を網羅することは難しい．現状では，3G (bp) のヒトゲノムに対して 200〜300 個程度の測定が限界であり，約 30,000 の全遺伝子を網羅して測定可能な遺伝子発現量と比べ，測定領域のカバー率が 2 桁も少ない．したがって，染色体を直接観察する方法は不十分であり，現状において，全遺伝子に関する網羅測定が可能な唯一の生物量である遺伝子発現量と，NCBI [8]などの公共 Web サイトで公開されている遺伝子の位置情報との統合によって，染色体異常領域を推定する意義は大きい．

一般に，欠損領域（Loss）の遺伝子の発現量は減少し[9]，増幅領域（Gain）の遺伝子の発現量は増加する傾向がある[10]．**図 3** 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（すべて正常領域）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，本稿では簡単のため，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．なお，各染色体につき 2 本ずつ染色体を有するため（両親から由来），一方の染色体の特定領域で欠損が生じたとしても，通常はもう一方の染色体の対応領域が正常であり，発現量は 0 にはならない．

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考えられる．Genome-wide transcriptome map [11]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たりする場合がある．このため，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のは難しい．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，周囲の遺伝子の発現量と見比べたうえで領域として評価する必要がある．Expression Imbalance Map [12]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，予測精度の点に問題があった．

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（61%）と再現率（74%）で染色体異常領域を予測できることを確認した．

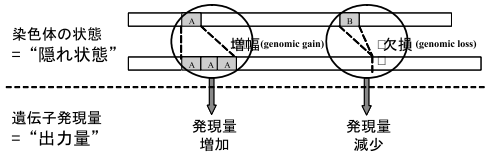

図 4 欠損・増幅領域推定のための隠れマルコフモデル
Fig. 4 Hidden Markov Model for detecting genomic gains and losses.

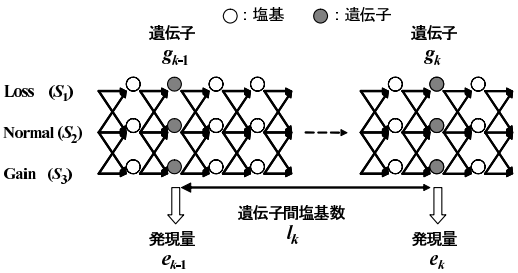

図 5 塩基をノードとした状態遷移
Fig. 5 State transitions when nodes are bases.

2. 隠れマルコフモデルによる異常領域の推定

2.1 概要と特徴

2.1.1 隠れマルコフモデル

隠れマルコフモデルとは，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．出力系列として観測されるシンボルは，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とき，最適状態遷移系列を動的計画法で効率良く求める方法が知られている（Viterbi アルゴリズム）[13]．

2.1.2 隠れマルコフモデルへのあてはめ

染色体の状態と遺伝子発現量の関係は，この隠れマルコフモデルによってきれいにモデル化することが可能である．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，隠れマルコフモデルの問題と見なすことができる（**図 4**）．隠れマルコフモデルは，音声認識や遺伝子発見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．

2.1.3 他の機械学習モデルと比較した際の利点

遺伝子発現量から染色体の状態を推定するうえで基本をなす前提は，(1) 遺伝子発現量と染色体の状態の間には相関関係がある，(2) 距離が近い遺伝子ほど同じ状態である確率が高い，という 2 点である．このため，隠れマルコフモデル以外の機械学習モデルを用いる場合，近傍の遺伝子の発現量から染色体の状態を予測することになる．しかしながら，遺伝子は非等間隔で染色体上に存在するために，予測に用いる近傍遺伝子の個数を固定することができないし，そもそも「近傍」を明確に定義することは難しい．また，予測に用いる近傍遺伝子の個数の増加にともない，計算量が膨大なものとなるという問題がある．一方，隠れマルコフモデルの場合は，染色体上の全遺伝子の発現量を考慮したうえで，最適な状態遷移系列を高速に求めることが可能である．したがって，染色体の欠損・増幅領域を推定するうえでは，隠れマルコフモデルが最も適したモデルであると考えられる．

2.1.4 提案モデルの特徴

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，隠れマルコフモデルを考える．本稿で扱うモデルの，従来の隠れマルコフモデルと比較した際の特長は以下の 2 点である．

(1)　状態遷移確率はノード（遺伝子）間距離に依存．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．すなわち，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とりうる．

(2)　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．すなわち，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 $e_k \sim e_k + \Delta e$ 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．

2.2 遺伝子間の距離に依存した状態遷移確率行列

遺伝子間距離に依存する遷移確率を算出するために，一時的にすべての塩基をノードとしたマルコフモデルを考える（**図 5**）．一般に，染色体欠損・増幅は染色体上の任意の場所に生じうるため，1 つの遺伝子領域に異なる複数の染色体状態が存在する可能性もある．

しかし，1 つの遺伝子領域に異なる状態が存在する場合，遺伝子の発現の機構がうまく働かず，発現量として発現マイクロアレイで測定することはできない．すなわち，1 つの遺伝子領域に 1 つの染色体状態しか存在しない場合にのみ，遺伝子発現量から染色体の状態を推定することが可能である．ただし，現実には，遺伝子の長さに比べて遺伝子間の塩基長の方がはるかに長く，また，1 つの染色体欠損・増幅領域の方が 1 つの遺伝子の領域より長いと考えられるため，1 つの遺伝子領域に異なる状態が存在するケースはほとんど無視できると考えられる．

本稿では，1 つの遺伝子の領域が，1 つの染色体状態を持つようにするために，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．ある塩基の状態を S_i としたとき，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，$0 < \alpha, \beta_1, \beta_2, \gamma < 1$

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 は正則なチェーンの遷移確率行列となり，$l \to \infty$ のとき，Q^l が収束することが保証される．ここで，Q の特性方程式の解は以下のようになる．

$$\begin{aligned} |Q-\lambda| &= (1-\lambda) \\ &\quad \times\{(1-\lambda)^2+(\alpha+\beta_1+\beta_2+\gamma)(1-\lambda) \\ &\quad +(\alpha\beta_2+\beta_1\gamma+\gamma\alpha)\} \\ &= 0 \end{aligned}$$

これを解き，固有値を求める．

$\lambda = \lambda_1(=1),\ \lambda_2,\ \lambda 3$
ただし，$0 < \lambda_2 < \lambda_3 < 1$

各固有値に対応して，以下の式を満たす固有ベクトル（行ベクトル）を求める．

$$\overrightarrow{w_i} Q = \lambda_i \overrightarrow{w_i}$$

求めた固有値と固有ベクトルを用いて，行列 W と Λ を以下のように定義する．

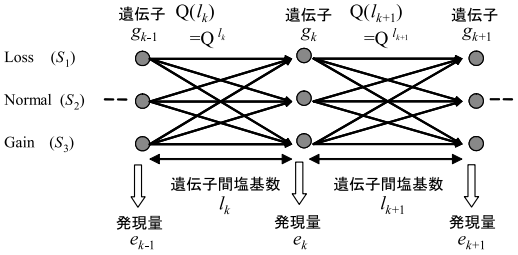

図 6 遺伝子をノードとした状態遷移
Fig. 6 State transitions when nodes are genes.

$$W = \begin{pmatrix} \overrightarrow{w_1} \\ \overrightarrow{w_2} \\ \overrightarrow{w_3} \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

この行列 W と Λ を用いて，Q を次のように書き直すことができる．

$$Q = W^{-1}\Lambda W$$

したがって，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 $Q(l)$ は，以下のようにして求めることができる．

$$Q(l) = \{q_{ij}(l)\} = Q^l = W^{-1}\Lambda^l W$$
ただし，$\Lambda^l = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda_2^l & 0 \\ 0 & 0 & \lambda_3^l \end{pmatrix}$

また，Q が正則なチェーンの遷移確率行列であるため，$l \to \infty$ のとき，$Q(l)$ は以下の行列 A に収束する．

$$Q(\infty) = \lim_{l\to\infty} Q^l = \begin{pmatrix} \overrightarrow{w_1} \\ \overrightarrow{w_1} \\ \overrightarrow{w_1} \end{pmatrix} = A$$

ここで $\overrightarrow{w_1}$ は，$\overrightarrow{w_1}Q = \overrightarrow{w_1}$ 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．

$$\begin{aligned} \overrightarrow{w_1} &= (\hat{w}_1, \hat{w}_2, \hat{w}_3) \\ &= \frac{1}{\alpha\beta_2+\beta_1\gamma+\gamma\alpha}(\beta_1\gamma, \gamma\alpha, \alpha\beta_2) \end{aligned}$$

2.3 状態遷移系列の推定

本稿で扱う隠れマルコフモデルは，遺伝子を表す塩基のみに着目した状態遷移である（**図 6**）．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，$m \leq n$）を以下のように

定義する．

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

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．

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

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，すなわち，$P(X_1^N, E_1^N)$ を最大化するような状態遷移系列 X_1^N を Viterbi アルゴリズムによって求める[13]．なお，モデルが，出力系列 E_1^k を生成して，k 番目の遺伝子で状態 S_i に到達する状態遷移系列に関して，最大の確率値を $\delta_k(i)$ とする．

$$\delta_k(i) = \max_{X_1^{k-1}} P(X_1^{k-1}, X_k = S_i, E_1^k)$$

また，簡便のため状態 S_i を単に状態 i と記すこととし，この表記を利用して，最大の確率値 $\delta_k(j)$ を与える直前の状態 i を $\varphi_k(j)$ とする．

[Viterbi アルゴリズム]

(1) 各状態 $j(j = 1, 2, 3)$ に対して $\delta_0(j)$ と $\varphi_0(j)$ を初期化．

$$\delta_0(j) = \hat{w}_j \quad \varphi_0(j) = 0$$

(2) 各遺伝子 $g_k(k = 1, 2, \ldots, N)$，各状態 $j(j = 1, 2, 3)$ に対して，再帰的に計算する．

$$\delta_k(j) = \max_i \lfloor \delta_{k-1}(i) q_{ij}(l_k) \rfloor b_j(e_k)$$
$$\varphi_k(j) = \arg\max_i \lfloor \delta_{k-1}(i) q_{ij}(l_k) \rfloor$$

(3) 再起計算の終了．

$$\hat{P} = \max_i \delta_N(i)$$
$$\hat{X}_N = \operatorname{argmax}_i \delta_N(i)$$

(4) バックトラックによる最適状態遷移系列の復元．$k = N-1, \ldots, 1$ に対して以下を実行．

$$\hat{X}_k = \varphi_{k+1}(\hat{X}_{k+1})$$

2.4 事後確率の算出

本手法では，欠損・増幅領域の推定結果を可視化する際に，出力系列 E_1^N を観測したうえでの，各遺伝子 g_k の隠れ状態 X_k が S_i である事後確率 $P(X_k = S_i | E_1^N)$ を利用する．そこで，出力系列 E_1^k を生成して k 番目の遺伝子で状態 S_i に到達する前向き確率 $F_k(i)$ と，k 番目の遺伝子で状態 S_i に到達後，出力系列 E_{k+1}^N を生成する後向き確率 $B_k(i)$ を以下に示すアルゴリズムで求め，前向き確率と後ろ向き確率を用いて事後確率を算出する[13]．

$$F_k(i) = P(X_k = S_i | E_1^k)$$
$$B_k(i) = P(E_{k+1}^N | X_k = S_i)$$

[前向きアルゴリズム]

(1) 各状態 $j(j = 1, 2, 3)$ に対して $F_0(j)$ を初期化．

$$F_0(j) = \hat{w}_j$$

(2) 各遺伝子 $g_k(k = 1, 2, \ldots, N)$，各状態 $j(j = 1, 2, 3)$ に対して，再帰的に計算．

$$F_k(j) = \left[\sum_{i=1}^{3} F_{k-1}(i) q_{ij}(l_k) \right] b_j(e_k)$$

(3) 最終確率の計算．

$$P(E_1^N) = \sum_{i=1}^{3} F_N(i)$$

[後向きアルゴリズム]

(1) 各状態 $i(i = 1, 2, 3)$ に対して $B_N(i)$ を初期化．

$$B_N(i) = 1$$

(2) 各遺伝子 $g_k(k = N, N-1, \ldots, 1)$，各状態 $i(i = 1, 2, 3)$ に対して，再帰的に計算．

$$B_{k-1}(i) = \sum_{j=1}^{3} q_{ij}(l_k) b_j(e_k) B_k(j)$$

(3) 最終確率の計算．

$$P(E_1^N) = \sum_{j=1}^{3} \hat{w}_i B_0(j)$$

事後確率は，前向き確率と後向き確率から以下のように算出することができる．

$$P(X_k = S_i | E_1^N) = \frac{F_k(i) B_k(i)}{P(E_1^N)}$$

2.5 パラメータの決定

通常，隠れマルコフモデルにおけるパラメータは，EM アルゴリズムで決定する手法が一般的であるが，本稿で提案する隠れマルコフモデルの場合，3 つの隠れ状態が完全に独立ではない（Loss と Gain の中間に Normal がある）ため，EM アルゴリズムを適用することができない．そこで，以下のような方法でパラメータの決定を行う．

(1) 遺伝子発現量の確率密度関数の平均と分散

ゲノムの状態が分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，求める．

(2) 状態遷移確率行列

以下の 2 種類の基準を用いて，パラメータの決定を行う．

(a) Loss, Gain 領域の長さの期待値

$$
\begin{aligned}
&(Loss\ \text{の長さの期待値})\\
&=\sum_{k=1}^{\infty} k\alpha(1-\alpha)^k\\
&=\lim_{n\to\infty}\sum_{k=1}^{n}\alpha k(1-\alpha)^k\\
&=\lim_{n\to\infty}\left\{\frac{1-\alpha}{\alpha}-(1-\alpha)^{n+1}(\frac{1}{\alpha}+n)\right\}\\
&=\frac{1-\alpha}{\alpha}\quad(\because 0<1-\alpha<1)\\
&\approx\frac{1}{\alpha}\quad(\because \alpha\ll 1)
\end{aligned}
$$

(b) 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与えられる.

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

一般に，bioinformatics の分野で用いられている解析ツールは，様々なパラメータの設定を必要とする. しかし，それらのパラメータはプログラム作成者側の視点に基づくものがほとんどで，ユーザである生物学者にとって解釈がしにくい. その点，本稿においてはパラメータと具体的な生物量を関連付けることによって，直観的にパラメータを解釈できるよう配慮されている.

3. 実　　験

Affimetrix 社[14]の発現マイクロアレイ（GeneChip U133AB）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）[12]の比較実験を行った. なお，6 個の検体に含まれる 22 種類の染色体に対して独立に解析を行ったため，$6\times 22 = 132$（本）の染色体に対する評価実験に相当する.

3.1 遺伝子発現情報と位置情報の対応付け

遺伝子発現情報と位置情報を対応付けるために以下のような処理を行い，GeneChipU133AB 上の 44,592 個の検査配列（プローブ）のうち，最終的に 12,485 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抽出（20,423 個）.
（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) 位置情報との関連付け.

(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）[8]で染色体上の位置を検索した. なお，1 つの遺伝子に対応するプローブが複数ある場合は，終端に近い方を採用した（12,485 個）. これは，染色体上の位置決定は，遺伝子配列の終端部側から行われているため，後ろ側の位置ははっきり決まるものの，先端部側の位置を正確に決定されていないと考えられるからである.

(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.

$$
\begin{aligned}
&(\ fold\,change\ \text{の対数値})\\
&=\log\frac{(\text{癌細胞での発現量})+C}{(\text{正常細胞での発現量})+C}
\end{aligned}
$$

3.2 正解セットおよび評価尺度

ゲノム上の特定の地点に関しては，欠損・増幅の状態を CGH アレイによって調べることができる. そこで，Vysis 社[15]の Genosensor 300 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. 6 検体に関する累計結果を**表 1** に示す. なお図 3 は，Loss/Normal/Gain と判定された地点の，近傍 1 Mbp（前後 500 kbp）に存在する GeneChipU133AB 上のプローブの fold change の対数値（3.1 節 (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 1 Mbp 以内に，5 個以上 3.1 節 (1) の閾値を満たす GeneChipU133AB のプローブが存在する 419 カ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.

$$precision = \frac{\text{正解数}\ (Loss/Gain)}{\text{予測箇所の数}\ (Loss/Gain)}$$

$$recall = \frac{\text{正解数}\ (Loss/Gain)}{CGH\text{アレイの判定数}\ (Loss/Gain)}$$

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

3.3 精度および再現率の評価

EIM において，*f-value* および *recall* が最大になるようにパラメータ調整を行った結果を**表 2** 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 7** は，Loss : Normal : Gain の占有比を 3 : 4 : 3 に固定した状態で，Loss および Gain の長

表 1 CGH アレイの結果とその近傍遺伝子の発現量
Table 1 Experimental results of CGH array and expression of genes in the neighborhoods.

	CGH アレイの判定数	プローブ数	*fc* 平均	*fc* 分散
Lose	230	693	−0.088	0.29
Normal	290	2,282	−0.0081	0.28
Gain	235	1,118	0.059	0.30

fc 平均：fold change の対数値平均
fc 分散：fold change の対数値分散

表 2 EIM による予測結果
Table 2 Results of predictions by EIM.

	recall	*precision*	*f-value*
EIM *f-value* Max	0.65	0.41	0.50
EIM *recall* Max	0.76	0.31	0.44

表 3 計算機の仕様
Table 3 Specification of the machine.

OS	WindowsXP
コンパイラ	gcc3.2
CPU	Xeon^{TM} CPU 3.06 GHz
メモリ	1.50 GB RAM

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* はつねに EIM における *f-value* の最大値を上回った．特に，Loss および Gain の長さが 5M (bp) のときには，*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 8** は，Loss および 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 x を 0.2 から 0.4 まで 0.05 刻みで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* はつねに EIM における *f-value* の最大値を上回った．また，*recall* を重視した解析を行う場合も，$x = 0.4$ に設定することで，*recall* = 0.84，*precision* = 0.51，*f-value* = 0.63 となり，EIM の *recall* の最大値を 8%上回り，かつそのときの *precision* も 20%上回った．

3.4 計算速度の評価

遺伝子発現量の解析ツールは，パラメータを変化させながら，何度も計算をやり直すという使い方が一般的である．このため，計算速度も解析手法を評価するうえで重要な指標となる．遺伝子数を n，サンプル数を m とした場合に，本手法の計算量は $O(mn)$ である．一方，EIM の計算量は $O(mn^2)$ である．$n = 12485$，$m = 6$ のときの計算時間は，EIM が 142.0 (s) であったのに対して，本手法は 6.8 (s) であり，約 21 倍の高速化が確認できた．なお，実験に使用した計算機は **表 3** のとおりである．

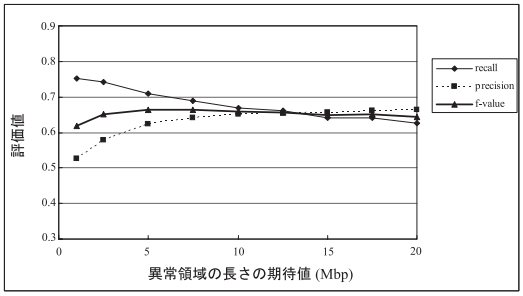

図 7 異常領域の長さの期待値と予測結果
Fig. 7 Prediction results by expectation value of the length of chromosomal abnormalities.

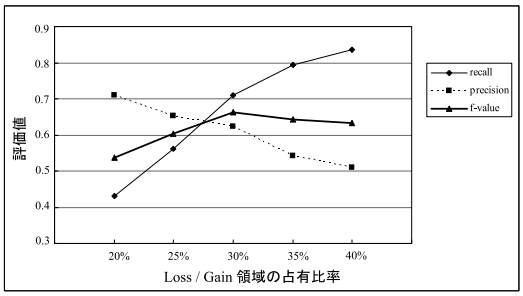

図 8 異常領域の占有比率の期待値と予測結果
Fig. 8 Prediction results by ratio of chromosomal abnormalities.

3.5 染色体異常領域の予測結果の可視化

図 9 は，Loss : Normal : Gain の占有比を 3 : 4 : 3，Loss および 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てある．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，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 10** は 1 番染色体の拡大図である．A の領域は，すべての検体で Loss と予測されているが，B の領域は 3 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．

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

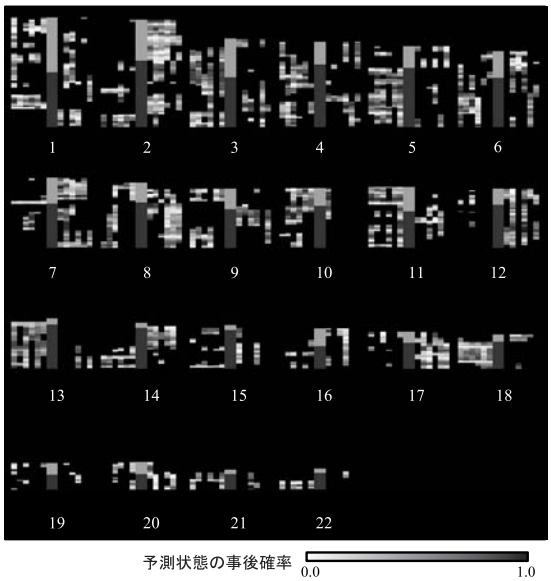

図 9 染色体異常領域の予測結果の可視化
Fig. 9 Visualization of prediction of chromosomal abnormalities.

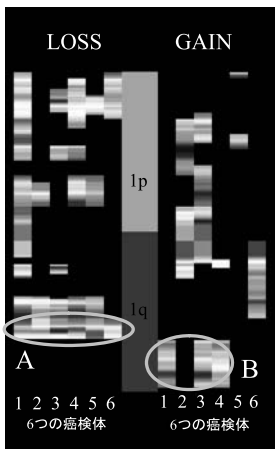

図 10 1 番染色体の拡大図
Fig. 10 Macrograph of chromosome 1.

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高速に予測できることを示した．

参 考 文 献

1) Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, Vol.286, No.5439, pp.531–537 (1999).
2) Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A.L., Brown, P.O. and Botstein, D.: Molecular Portraits of Human Breast Tumours, *Nature*, Vol.406, No.6797, pp.747–752 (2000).
3) Ramaswamy, S., Ross, K.N., Lander, E.S. and Golub, T.R.: A Molecular Signature of Metastasis in Primary Solid Tumors, *Nat. Genet.*, Vol.33, No.1, pp.49–54 (2003).
4) Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D.: Cluster Analysis and Display of Genome-wide Expression Patterns, *Proc. Natl. Acad. Sci. USA*, Vol.95, No.25, pp.14863–14868 (1998).
5) Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M.: Systematic Determination of Genetic Network Architecture, *Nat. Genet.*, Vol.22, No.3, pp.281–285 (1999).
6) Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F. and Pinkel, D.: Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors, *Science*, Vol.258, No.5083, pp.818–821 (1992).
7) Veltman, J.A., Schoenmakers, E.F., Eussen, B.H., Janssen, I., Merkx, G., van Cleef, B., van Ravenswaaij, C.M., Brunner, H.G., Smeets, D. and van Kessel, A.G.: High-throughput Analysis of Subtelomeric Chromosome Rearrangements by Use of Array-based Comparative Genomic Hybridization, *Am. J. Hum. Genet.*, Vol.70, No.5, pp.1269–1276 (2002).
8) NCBI. http://www.ncbi.nlm.nih.gov
9) Mukasa, A., Ueki, K., Matsumoto, S., Tsutsumi, S., Nishikawa, R., Fujimaki, T., Asai, A., Kirino, T. and Aburatani, H.: Distinction in Gene Expression Profiles of Oligodendrogliomas with and without Allelic Loss of 1p, *Oncogene*, Vol.21, No.25, pp.3961–3968 (2002).
10) Virtaneva, K., Wright, F.A., Tanner, S.M., Yuan, B., Lemon, W.J., Caligiuri, M.A., Bloomfield, C.D., de La Chapelle, A. and Krahe, R.: Expression Profiling Reveals Fundamental Biological Differences in Acute Myeloid Leukemia with Isolated Trisomy 8 and Normal Cytogenetics, *Proc. Natl. Acad. Sci. USA*, Vol.98, No.3, pp.1124–1129 (2001).

11) Fujii, T., Dracheva, T., Player, A., Chacko, S., Clifford, R., Strausberg, R.L., Buetow, K., Azumi, N., Travis, W.D. and Jen, J.: A Preliminary Transcriptome Map of Non-small Cell Lung Cancer, *Cancer Res.*, Vol.62, No.12, pp.3340–3346 (2002).
12) Kano, M., Nishimura, K., Ishikawa, S., Tsutsumi, S., Hirota, K., Hirose, M. and Aburatani, H.: Expression Imbalance Map: A New Visualization Method for Detection of mRNA Expression Imbalance Regions, *Physiol Genomics*, Vol.13, No.1, pp.31–46 (2003).
13) Durbin, R., Eddy, S., Krogh, A. and Mitchison, G.: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press (1998).
14) Affymetrix 社．http://www.affymetrix.com
15) Vysis 社．http://www.vysis.com/

(平成 15 年 9 月 25 日受付)
(平成 16 年 1 月 19 日採録)

（担当編集委員　石川 博，市川 哲彦，原 隆浩，
佐藤 聡，土田 正士）

加納　　真（正会員）

2000 年東京大学大学院工学系研究科修士課程修了．同年日本アイ・ビー・エム（株）入社．現在東京基礎研究所副主任研究員．

石川　俊平

2000 年東京大学医学部医学科卒業．2004 年医学博士取得．同年東京大学国際・産学共同研究センター助手．現在に至る．

油谷　浩幸

1980 年東京大学医学部卒業．同年東京大学附属病院医員．1988 年医学博士取得．1988 年マサチューセッツ工科大学ガン研究センター研究員．1999 年東京大学先端科学技術研究センター助教授．2001 年東京大学先端科学技術研究センター教授．2002 年東京大学国際・産学共同研究センター教授，現在に至る．