2006－BIO－6 (4)
2006／9／15

# 染色体異常に対する混合木モデルの紹介とその改良

山本 幸生,　　大羽 成征,　　石井 信

奈良先端科学技術大学院大学 情報科学研究科
〒 630–0192 奈良県生駒市高山町 8916–5

## 概 要

近年、がん細胞における染色体異常蓄積のモデルとして混合木モデルが提案されている。このモデルによれば、各症例で起こっている染色体異常イベントのデータから木構造の因果関係を推定することができる。しかし従来のモデルには観測ノイズに起因する偽陰性の悪影響を受けやすいという問題があった。そこで我々はイベント推定における偽陰性の影響に対してロバストな因果関係を推定するべく混合木モデルの改良を行った。また人工データにもとづいて従来のモデルのノイズの影響を検証し、さらに偽陰性ノイズを考慮した改良の効果を確認した。

# Mixture oncogenetic trees model for chromosomal alteration and its improvement

Kosei YAMAMOTO,　Shigeyuki OBA,　and　Shin ISHII

Graduate School of Information Science,
Nara Institute of Science and Technology
Takayama 8916–5, Ikoma-shi, Nara, 630–0192 Japan

## ABSTRACT

The mixture oncogenetic trees model is a recent model for accumulation of chromosomal abnormality in cancer cells. The original model could not rule out the possibility that there is a harmful effect of false negatives on observation data in which the tree structure of the causal relationship between chromosomal abnormality events is estimated. We improved the mixture oncogenetic trees model to provide more robust estimation of the causal relationship against the effect of false negatives in observation data. Simulations using artificial data confirmed that the improvement works well.

## 1　導入

正常な細胞の細胞分裂では、まず遺伝子をコードしている染色体が複製され、続いて複製された染色体が均等に 2 分された後、最終的に核と細胞質が分裂する。この一連の流れを細胞周期と呼ぶ。がん細胞の第一の特徴はこの細胞周期が乱れていることであって、そのせいで細胞が過剰な増殖を起こしたり染色体複製が不均一になったりしてしまう。またがん細胞の第二の特徴として、細胞の異常が個体に悪影響を与えないための細胞自死システムの異常が挙げられる。この二つの理由から、がん細胞内染色体の異常が蓄積されてゆき、それがさらに様々なシステム異常の原因となってゆくのだと考えられている。

染色体異常の具体例として、細胞に含まれる染色体の量が一部断片において通常よりも多くなったり (増幅)、少なくなったり (欠損) といった様子

が観測されている [1]。これらの染色体異常は未知の因果関係の一連の流れに従って起こっていると考えられており、これを解明することには、がんの病理学的解明臨床での治療法選択などにつながる大きな意義がある。しかし特に腫瘍として発見されるがんにおいては、染色体異常を測定する時点まで異常の蓄積は近づいており、過去に起った染色体異常のイベント間の因果関係を知ることはできない。そこで、同種のがんに関する複数の症例の染色体異常にもとづく因果関係の推定が試みられてきた。

直腸がんにおいて 3 つの遺伝的イベントが順番に起こるという経路モデルが提案されている [2]。経路モデルでは、遺伝的イベントはがんの進行に併せて 1 番目のイベントが起こると 2 番目のイベントが起こる確率が増加するというような、直線的な因果関係を仮定している。またこのモデルの拡張として、Desper ら (1999) は 1 つのイベントが 2 つ以上のイベントの原因となる事を許した木状の因果構造を持つモデルを提案した [3]。

木モデルはその特別な場合として先に述べた経路モデルを含むなど高い説明力を持っているが、因果関係上流のイベントが起きているときのみ下流イベントが起りうるという順序関係を厳密に要求するために、必ずしも全症例を説明できない。症例の中でこのモデルに適合しないものを表現するため、隠れイベントを考慮した distance-based モデル [4] や、複数の木構造を推定する混合木モデルが提案された [5] 。とくに混合木モデルでは複数の木構造のうちの一つを星状にするアイデアが重要である。星状の構造は木構造の特別な場合であって、全ての遺伝的イベントが独立に一定確率で起こると仮定しており、どのようなイベントパターンを持つ症例も有限の確率で表現できる。混合木モデルは、主要な木構造で説明できない症例を星状モデルに担当させることでロバストな推定を実現した。

染色体異常イベントの検出方法として、染色された染色体を顕微鏡下でしらべる FISH 法が有名である [6]。一方で近年、染色体異常の新しい検出方法としてアレイ CGH と呼ばれる手法が使われるようになってきた [7]。この手法は通常の細胞が必ず 2 コピーの染色体セットを持っているという事を利用して、多数 (数千から数万個) の染色体断片につけた蛍光色素の蛍光比で染色体の部分的な増減を判断する手法である。アレイ CGH 手法を使ったこれまでの研究は、多くの種類のがんでいくつかの一致した染色体変化のパターンが存在していることを示唆している [8]。

アレイ CGH 手法にもとづくイベント検出は一度に染色体の全領域にわたる調査ができる利点があるが、染色体上イベント検出において偽陽性と偽陰性のノイズを含む事がある。現状の混合木モデル [10][9] はイベント単位のノイズを考慮しておらず、わずかな偽陰性ノイズを含む症例であっても悪影響が大きい。我々は本稿で、データのノイズの影響を考慮した混合木モデルの改良型を提案する。この改良の結果、全ての症例の中でノイズと判断されて星状モデルに割り当てられる症例を減らし、主要木構造の推定のためにより多くの症例を使うことができるようになった、またこれにより木モデルの確率推定の精度を高める事ができることがわかった。

## 2 染色体異常に対する混合木モデル

各症例 $i = 1, \ldots, N$ において観測された遺伝的イベントを 2 値ベクトル $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{iM})$ で表す。その各要素は

$$x_{ij} = \begin{cases} 1, \text{症例 } i \text{ において } j \text{ 番目の} \\ \quad \text{イベントが起こっている} \\ 0, \text{起こっていない} \end{cases} \quad (1)$$

の 2 値をとるものとする。ただし、0 番目のイベントは特別な null イベントとし、常に $x_{i0} = 1$ が成り立つこととする。null でない各イベントとしては、染色体部位の増加や欠失、変位などが想定されている。

### 2.1 木モデル

ここで染色体上イベント間の因果関係が木構造をなしていると仮定し、これを表す木構造を $\mathcal{T} = (V, E, r, P)$ で表現する。図 1 はその例を示す。ここで、$V$ は全イベントを表現するノードの集合、$E$ はノード間をつなぐリンクの集合、$r$ は null イベント、$P$ は直接の因果関係を持つイベントペア間の条件付き確率を表す。(例えば親イベント $j_1$

と子イベント $j_2$ をつなぐリンク $e \in E$ について、$P(e) = \Pr(j_2|j_1)$ はイベント $j_1$ が起こった時の $j_2$ の起こる条件付き確率を表している)。

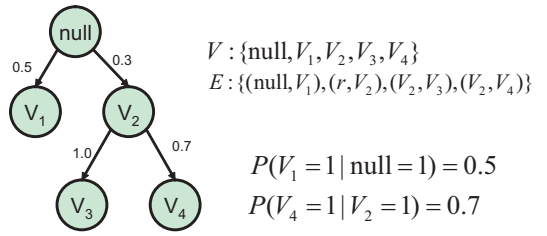


図 1: 5 つのノードを持つ木構造の例。矢印はイベント間の因果関係を表しており、各エッジについている数字は条件付き確率を表している。

いったん木構造が与えられると、全症例における全イベントの観測データ $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ にもとづく木構造 $\mathcal{T}$ の尤度 $L(\mathcal{T}|X)$ を計算する事ができる。

$$L(\mathcal{T}|X) \equiv \prod_{i=1}^{N} P(\boldsymbol{x}_i|\mathcal{T}) \tag{2}$$

ここで $P(\boldsymbol{x}_i|\mathcal{T})$ は各パターンベクトル $\boldsymbol{x}_i$ が $\mathcal{T}$ から生成される確率である。サンプル $\boldsymbol{x}_i$ の中で起こっているイベントの集合を $S \subseteq V$ とする。木構造を構成する全エッジ $E$ の中で、実際におこったイベント $S$ の要素間をつなぐエッジだけを抜き出してきたものをエッジ集合 $E' \subseteq E$ とする。さらに可能な全てのエッジの中で実際に起こったイベント $S$ と起こらなかったイベント $V\backslash S$ とをつなぐエッジの集合を $S \times V\backslash S$ で表す。例えば図 1 の例で $\boldsymbol{x}_i = (11110)$ とすると、$S = \{r, V_1, V_2, V_3\}$、$E' = \{(r, V_1), (r, V_2), (V_2, V_3)\}$、$S \times V\backslash S = \{(V_2, V_4)\}$ となる。するとこの木構造 $\mathcal{T}$ からイベントベクトル $\boldsymbol{x}_i$ が生成される確率は以下のように表すことができる。

$$P(\boldsymbol{x}_i|\mathcal{T}) = \prod_{e \in E'} P(e) \cdot \prod_{e \in (S \times V\backslash S)} (1 - P(e)) \tag{3}$$

ところで $\mathcal{T}$ から生成不可能なパターン $\boldsymbol{x}_i$ があり得る。例えば図 1 の例では $\boldsymbol{x}_i = (11010)$ のようなイベントパターンは、$V_2$ が起っていないにも関わらず $V_3$ が起こっており $\mathcal{T}$ では説明できない。このような時 $P(\boldsymbol{x}_i|\mathcal{T}) = 0$ とする。この条件の問題について 3 節で解説する。

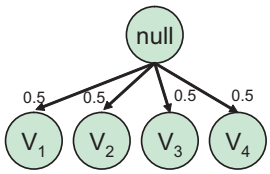


図 2: 5 つのノードを持つ星状構造の例。null 以外のイベントは null から起こるためイベント間の相関のないノイズを表現している。

### 2.2 混合木モデル

混合木モデル $\mathcal{M}$ は各パターン $\boldsymbol{x}_i$ が $K$ 個のツリー $\mathcal{T}_k, k = 1, \ldots, K$ のうちの一つから生成されたと考えるモデルである [5]。$k$ 番目のツリー $\mathcal{T}_k$ が選ばれる確率を混合比 $\alpha_k$ とするとき、混合木モデル $\mathcal{M}$ の尤度は以下のように定義される。

$$L(\mathcal{M}|X) = \prod_{i=1}^{N} \sum_{k=1}^{K} \alpha_k P(\boldsymbol{x}_i|\mathcal{T}_k) \tag{4}$$

モデルから説明できない症例パターンが無いようにするために $\mathcal{T}_1$ の構造を図 2 のような星状構造に固定するのが、混合木モデルで最も重要なアイディアである。星状構造では、null イベントが他の全てのイベントの上流になっており、生成不可能なパターンが存在せず $P(\boldsymbol{x}_i|\mathcal{T}) > 0$ が成り立つ。混合木モデルでは主要木モデルでは説明できないノイズ的症例を星状構造に集めることで、残りの症例モデルで安定した主要木構造を推定する。

なお主要木構造が複数ある状況を考慮できることも混合木モデルの特徴であるが、本稿では主要木構造 $\mathcal{T}_2$ の一つだけ、すなわち $K = 2$ である場合について考えることにする。

## 3 拡張混合木モデル

### 3.1 ノイズとその影響

真の主要木モデル $\mathcal{T}_2$ から生成されたパターンであっても、少しのノイズ (とくに偽陰性) のせいで $\mathcal{T}_2$ から生成され得ないパターンが観測されてしまうことがある。例えば図 1 の木構造から $\boldsymbol{x}_i = (11110)$ のようなパターンが生成されていたとして、イベント $V_2$ がノイズのために偶然に見落とされた場合 $\mathcal{T}_2$ から生成され得ない $\boldsymbol{x}_i = (11010)$ が観測される。このとき $\boldsymbol{x}_i$ に対応する $\mathcal{T}_2$ の尤度は $P(\boldsymbol{x}_i|\mathcal{T}_2) = 0$ となるため、前節で説明した混合木モデルでは、そ

のようなパターンは星状モデル $\mathcal{T}_1$ に吸収されてしまう。しかし $\boldsymbol{x}_i$ はイベント $V_2$ 以外に関しては真の主要木モデル $\mathcal{T}_2$ の特徴を保持しているのであって、これが $\mathcal{T}_2$ の推定に参加できないのは大きな情報のロスである。

### 3.2 ノイズを考慮したモデル

観測された染色体異常イベントのパターンベクトル $\boldsymbol{x}_i$ がノイズを含んでいる場合は以下のように定式化できる。

混合木モデルの各ユニットモデル $\mathcal{T}$ から確率的に生成された真のパターンを $\boldsymbol{y} = (y_1, \cdots, y_M), y_i \in \{0,1\}$ とするとき、染色体異常イベント $i$ が起こっているにもかかわらず、ノイズのためにそれが観測されないこと $(y_i = 1, x_i = 0)$ を偽陰性 (false negative; FN) と呼び、染色体異常イベント $i$ が起こっていないにもかかわらず、ノイズのためにそれが観測されること $(y_i = 0, x_i = 1)$ を偽陽性 (false positive; FP) と呼ぶことにする。また、偽陰性率を $\beta \equiv P(x_i = 0|y_i = 1)$、偽陽性率を $\alpha \equiv P(x_i = 1|y_i = 0)$ と書くことにする。

混合木モデルの各木モデル $\mathcal{T}$ が与えられているとき、$\mathcal{T}$ から真のイベントパターン $\boldsymbol{y}$ が得られる確率 $P(\boldsymbol{y}|\mathcal{T})$ は 2 章の式 (4) で定義されたとおりである。しかし、観測イベントパターン $\boldsymbol{x}$ が得られる確率 $P(\boldsymbol{x}|\mathcal{T})$ は観測ノイズを考慮することによって以下のようになる。

$$P(\boldsymbol{x}|\mathcal{T}) = \sum_{\boldsymbol{y}} P(\boldsymbol{x}|\boldsymbol{y}) P(\boldsymbol{y}|\mathcal{T}). \tag{5}$$

ただし和は全ての可能なパターン $\boldsymbol{y}$ に関してとる。ここで $P(\boldsymbol{x}|\boldsymbol{y})$ は真のパターン $\boldsymbol{y}$ にノイズが付与される過程を表し、以下のように計算できる。

$$P(\boldsymbol{x}|\boldsymbol{y}) = \beta^{k_{\mathrm{FN}}} \cdot (1-\beta)^{n_0 - k_{\mathrm{FN}} + k_{\mathrm{FP}}} \cdot \alpha^{k_{\mathrm{FP}}} \cdot (1-\alpha)^{n_1 + k_{\mathrm{FN}} - k_{\mathrm{FP}}} \tag{6}$$

ただし $n_0, n_1$ はそれぞれ $\boldsymbol{x}$ において観測されないイベント, 観測されたイベントの総数。$k_{\mathrm{FN}}, k_{\mathrm{FP}}$ はそれぞれ仮に $\boldsymbol{y}$ が真実であるとしたときに $\boldsymbol{x}$ に含まれると考えられる FN, FP の総数である。一般に偽陰性ノイズが root に近いノードで起こると生成され得ないパターンになってしまう可能性が大きく、悪影響が大きい。一方で偽陽性ノイズの影響はさほど大きくない。そこで以下では偽陽性数 $k_{\mathrm{FP}}$、偽陽性率 $\alpha$ が 0 である場合のみ考えることにする。このとき、

$$P(\boldsymbol{x}|\boldsymbol{y}) = \beta^{k_{\mathrm{FN}}} \cdot (1-\beta)^{n_0 - k_{\mathrm{FN}}} \tag{7}$$

となる。

## 4 混合木モデルの推定

### 4.1 EM アルゴリズムの大枠

混合木モデルの尤度関数 $L(\mathcal{M}|X)$ を最大化するモデル $\mathcal{M}$ を求めるために EM アルゴリズムが提案されている [5][11]。各パターン $\boldsymbol{x}_i$ が各ツリー $\mathcal{T}_k, \quad k = 1, \ldots, K$ のどれに所属するかを示す変数を責任信号 $r_{ik}$ と呼ぶ。$r_{ik} \approx 1$ のときパターン $\boldsymbol{x}_i$ はツリー $\mathcal{T}_k$ に属する。ここで $\sum_{k=1}^{K} r_{ik} = 1$ である。EM アルゴリズムでは各ツリーモデル $\mathcal{T}_k$ を固定して $r_{ik}$ を求める E ステップと $r_{ik}$ を固定して各ツリーモデルを求める M ステップを収束するまで繰り返す。以下で各ステップを詳しく説明する。

### 4.2 M ステップ

各ツリー $\mathcal{T}_k$ に対応する症例に関する全てのイベントペア $(j_1, j_2), 0 \le j_1, j_2 \le M$ について共起確率を計算する。

$$\mathrm{Pr}(j_1, j_2) = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} x_{ij_1} x_{ij_2} \tag{8}$$

ただし $N_k = \sum_{i=1}^{K} r_{ik}$ とする。これにもとづいて木構造を推定する方法として一般的なものは maximum weight branching アルゴリズム [3] であり、ほとんどの場合に尤度最大の木構造が得られることが経験的に分かっている。maximum weight branching アルゴリズムでは、イベント間のエッジの重み $w$ を以下のように計算し、

$$w(j_1, j_2) = \log(\mathrm{Pr}(j_1, j_2) - \log(\mathrm{Pr}(j_i) + \mathrm{Pr}(j_2)) - \log \mathrm{Pr}(j_2) \tag{9}$$

木構造を構成するエッジの重みの合計が最大になるように木構造を構築する。同時に混合比パラメータを $\alpha_k = N_k/N$ で計算する。

$\mathcal{T}_1$ の星状構造については、特にエッジの重みを全エッジで共通の値 $b$ として以下のように求める。

$$b = \frac{1}{MN_1} \sum_{j=0}^{M} \sum_{i=1}^{N} r_{ik} x_{ij} \tag{10}$$

### 4.3 E ステップ

M ステップで更新された木構造 $\mathcal{T}_k$ から、各パターン $\boldsymbol{x}_i$ に対応する $\mathcal{T}_k$ の尤度 $P(\boldsymbol{x}_i|\mathcal{T}_k)$ を式 (3) のように計算できる。これを用いて以下のようにパターン $x$ の責任信号を更新する。

$$r_{ik} = \frac{\alpha_k P(\boldsymbol{x}_i|\mathcal{T}_k)}{\sum_{m=1}^{K} \alpha_m P(\boldsymbol{x}_i|\mathcal{T}_m)} \tag{11}$$

データに偽陰性を考慮した場合、木構造の尤度の計算式は (5) となるがその他のアルゴリズムは全く同様である。ここで真の偽陰性率 $\beta$ は不明であるので適当に与える。

正確には式 (5) では全ての可能な $\boldsymbol{y}$ について和をとらねばならないが、計算量の節約のため偽陰性の個数が 0 個もしくは 1 個のみである場合についてのみ和をとる近似を用いた。$\beta$ が十分に小さい値のときには偽陰性の個数が 2 個以上である確率は小さいため、近似の精度は十分であると考えられる。

## 5 数値実験

### 5.1 準備

アルゴリズムの評価のため人工データを以下のように用意する。

まず推定対象となる真のモデルとして 6 イベントからなる木構造 $\mathcal{T}^*$ を 3 種類用意した。(図 3)

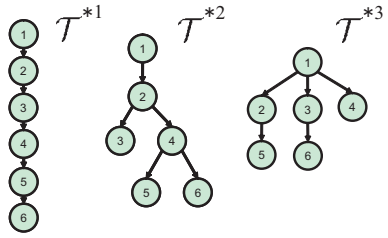


図 3: 人工データの基として用意した 3 種類の木構造。イベント間の条件付き確率は全てのエッジにおいて $P(e) = 0.75$ とした。

各モデルに基いてそれぞれ真のイベントパターンデータ 50 症例ぶんをランダム生成した。さらに、生成された人工データの各症例イベントに対してあらかじめ決めておいた FN 率 $\beta^*$ に従ってランダムに $1 \to 0$ の変換を加えることで観測データを作った。なお、$\beta^* = 0, 0.025, 0.05, \ldots, 0.2$ の各場合を比較した。

こうして作った人工データにもとづいて混合数 $K = 2$ の混合木モデルを推定した。$\mathcal{T}_1$ は星状構造に固定し、$\mathcal{T}_2$ は主要木モデルを推定した。FN を考慮する場合には適当な FN 率 $\beta$ を与える必要があったが、これには $\beta = 0, 0.025, 0.05, \ldots, 0.2$ の各場合を比較した。$\beta = 0$ を与えた場合は従来法と同等である。

推定されたモデルの精度は二種類の基準で比較した。第一は主要木モデル $\mathcal{T}_2$ の混合比 $\alpha_2$ の値である。本来は一つの木構造から得られたデータであるので、$\alpha_2$ が 1 に近いほど推定の精度が良かったことになる。第二は各エッジの確率推定のエラーである。図 3 のとおり、全エッジにおいて真の条件付きイベント生起確率は 0.75 である。この推定値と真値の間の平均二乗誤差も推定精度の基準とした。なお、木構造の構造推定はここでは考慮せず、正しい構造が推定できているものとした。それぞれの条件のもとで乱数種を変えながら 100 回の実験を行い、得られた結果の平均と標準偏差を示した。

### 5.2 実験 1: データに FN を加えたときの悪影響

観測データに含まれる FN の割合 $\beta^*$ と従来法の混合木モデルに基く推定精度の関係を調べた。図 4 の上段にその結果を示す。横軸は $\beta^*$ の値。3 本の曲線はそれぞれ、真の構造 $\mathcal{T}^*$ が $\mathcal{T}^{*1}$,$\mathcal{T}^{*2}$,$\mathcal{T}^{*3}$ であった場合の推定結果に対応する。図 4 の左側では主要木モデルの混合比、右側では条件付き確率の推定誤差を示している。

データに FN を加えた場合には全てのモデルで主要木モデルの混合比が減少し、条件付き確率の誤差が上昇しており推定精度が悪くなっている事がわかる。3 つのモデル間の違いを混合比でみると、主要木の構造がノイズを担当する星状構造と大きく異なる直線状の構造 $\mathcal{T}^{*1}$ の場合にノイズの

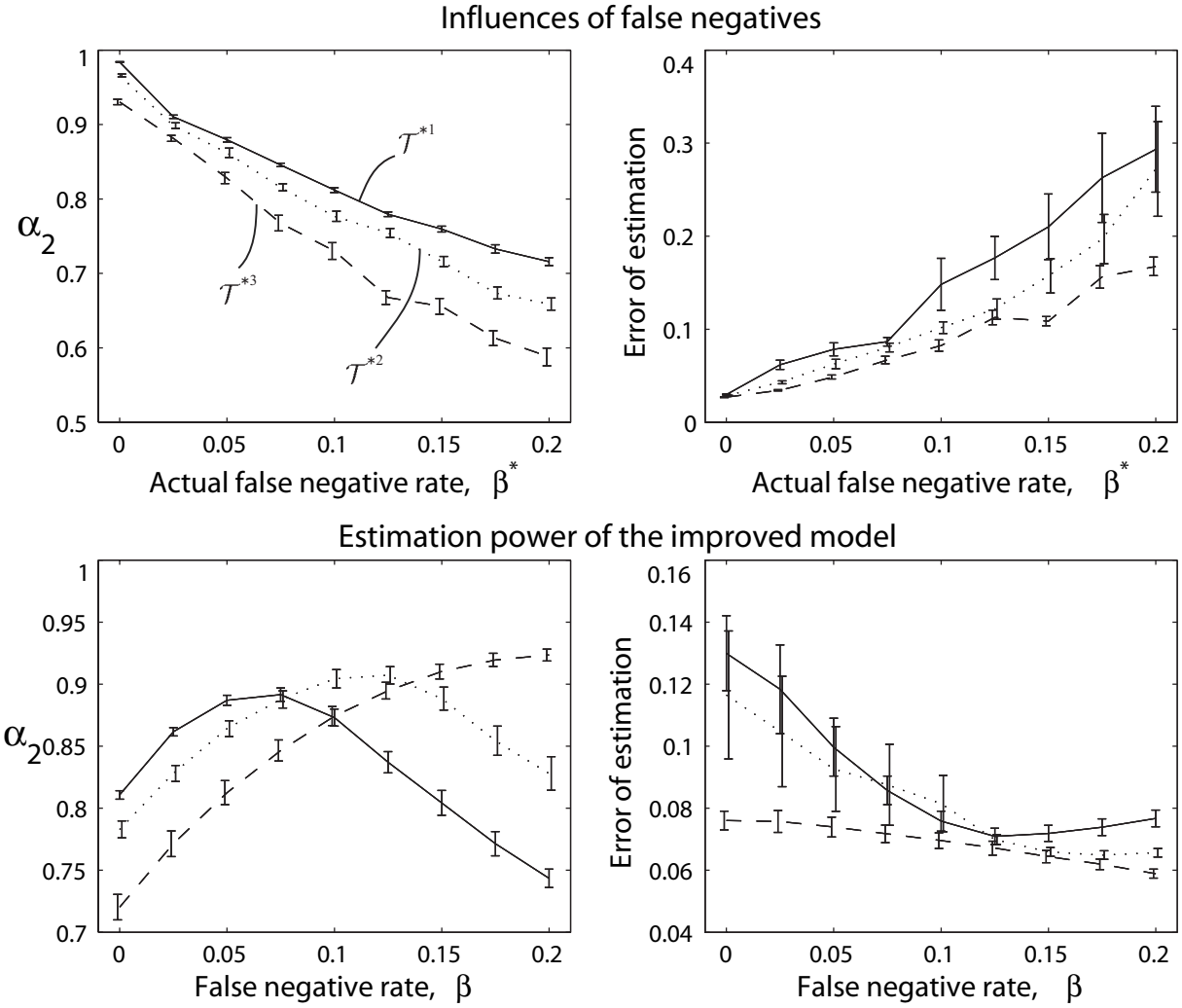


図 4: 上段は実験 1, 下段は実験 2 の結果を示す。各左段は混合木構造の推定結果として得られた混合比 $\alpha_2$、右の図は推定された条件付き確率と真の値との誤差を示している。実線はモデルとして $\mathcal{T}^{*1}$ を与えた場合、点線と波線はそれぞれ $\mathcal{T}^{*2}$ と $\mathcal{T}^{*3}$ を与えた場合。エラーバーは 100 試行中の標準偏差を示している。

影響が小さく、星状構造に近い $\mathcal{T}^{*2}$、$\mathcal{T}^{*3}$ ほど影響が大きくなっていることがわかる。対照的に条件付き確率の推定誤差は、推定するモデルの構造が直線に近いほど FN の影響を受けやすく、星状構造に近いほど影響が小さくなっている。また両者ともに FN を大きくするに従って標準偏差の値が大きくなっており、ノイズとしての FN が推定の安定性に大きな悪影響をもたらしていることがわかる。

### 5.3 実験 2: 改良手法の性能

観測データに含まれ偽陰性ノイズを $\beta^* = 0.1$ で固定した場合について、改良手法の性能を調べた。改良手法では偽陰性率 $\beta$ を適当に決める必要があるので、それを様々に変えながら結果を比較した。なお、$\beta = 0$ の場合が従来法に対応する。

図 4 の下段にその結果を示す。横軸は $\beta$ の値である。主要木モデルの混合比の値は従来法よりも改善が見られたが、性能がピークを示す $\beta$ の値はモデルの構造によって異なった。$\mathcal{T}^{*2}$ では $\beta = 0.1$ 付近で混合比が最大になっており、改良手法により FN ノイズの影響を減少できたと考えられる。最も星状構造に近い $\mathcal{T}^{*3}$ では、仮定する FN を大きくすればするほど混合比が大きくなっており、推定する対象の構造が星状構造に似ている場合には $\beta$ の適切な選択が困難であると考えられる。条件付き確率の推定については、全てのモデルで誤差と標準偏差が減少しており改良手法の有効性が確

認できた。特に $\mathcal{T}^{*1}$ では $\beta = 0$ から $\beta = 0.1$ まで誤差が減少しその後上昇している事から、真の木構造が直線上に近く root から見て深い枝を含む場合に改良型手法が最も有効である事がわかる。

## 6 まとめ

本研究では偽陰性の悪影響を避けるために、データ中に偽陰性を考慮する事で混合木構造を改良した。人工データにもとづく数値実験の結果、従来の混合木モデルではデータの中に FN が含まれる場合に因果関係推定に使われる症例の割合が小さくなり、推定結果も悪くなる事を確かめ、さらに混合木アルゴリズムに FN の影響を考慮した改良手法を導入すると、データ損失が少なくなりさらに木構造の条件付き確率推定の精度が上昇する事を示した。

本稿では偽陽性を含めた場合や、遺伝的イベントの数や症例の個数が大きくなった時にその影響が混合木構造の推定にどう関わるかは考慮していない。また混合木アルゴリズムの中で、Desper ら (1999)[3] によって推定された木構造の尤度推定に対してのみノイズの影響を考慮したが、主要木構造の構築に対するノイズの影響の調査も今後の課題である。

## 謝辞

本研究の一部は、文部科学省特定領域科学研究費 (応用ゲノム) の支援を受けて実施されました。

## 参考文献

[1] Nowell,P.C. The clonal evolution of tumor cell population. *Science*, **194**, 23-28. (1976)

[2] Vogelstein, B., Fearon, E., Hamilton, S., *et al.* Genetic alterations during colorectal-tumor development. *N. Engl. J. Med.*, **319**, 525-532. (1988)

[3] Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C., and Schaffer, A. Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comp. Biol.*, **6**(1), 37-51. (1999)

[4] Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, Papadimitriou, C., and Schaffer, A. Distance-based reconstruction of tree models for oncogenesis. *J. Comp.Biol.*, **7**(6),789-803. (2000)

[5] Beerenwinkel, N., Rahnenfuhrer, J., Daumer, M., Hoffmann, D., Kaiser, R., Selbig, J., and Lengauer, T. Learning evolutionary pathways from cross-section dara. *J.Comp.Biol.*, **12**(6), 584-98. (2005)

[6] Khac, F.N., Waill, M.C., Romana, S.P., Radford-Weiss, I., *et al.* Identical abnormality of the short arm of chromosome 18 in two Philadelphia-positive chronic myelocytic leukemia patients with erythroblastic transformation, resulting in duplication of BCR-ABL1 fusion. *Cancer Genet. Cytogenet.*, **138**(1), 22-6. (2000)

[7] Kallioniemi, A., Kallioniemi, O.P., Sudar, D., *et al.* Comparative genomic hybridization for molecular cytogenic analusis of solid tumors. *Science*, **258**, 818-821. (1992)

[8] Forozan, F., Karhu, R., Kononen, J., Kallioniemi, O.-P. Genome screening by comparative genome hybridization. *Trend Genet.*, **7**, 85-90.(1997)

[9] Rahnenfuhrer, J., Beerenwinkel, N., Schulz, W.A., Hartmann, C., von Deimling, A., Wullich, B., Lengauer, T. Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics*, **21**(10), 2438-2446. (2005)

[10] Szabo, A., Boucher, K. Estimationg an oncogenetic tree when false negatives and positives are present. *Mathmatical Bioscience*, **176**, 219-236. (2002)

[11] Marina, M., Jordan, M. Learning with mixtures of trees *J. Machine Learning Res.*, **1**, 1-48. (2000)