複数箇所のがん組織ゲノムデータを利用し
高精度な体細胞変異検出を行う階層ベイズモデルの開発

東京大学医科学研究所 森山卓也，山口類，井元清哉

1. はじめに

がん細胞で生じたゲノム変異(体細胞変異)をがんゲノムシークエンスデータから同定することは、がん研究において標準的かつ重要な解析である．がんは，様々な変異を獲得しつつ増大し，進化する．このがんの進化の過程は，例えば図 1(A)のように，頂点をがん細胞の状態，辺を獲得する体細胞変異とした木構造で表され，どの箇所(図 1(A)の 1-6)を採取するかで，有するゲノム変異は異なる．これは，変異の腫瘍内不均一性と呼ばれ，悪性度や治療効果に影響することが近年分かってきた．そこで、同一患者のがん組織から複数箇所を選び，それぞれの箇所のシークエンスデータを用いた不均一性の解析が注目を集めている．この不均一性の解析では，図 1(A)のような腫瘍の進化過程を表す進化系統樹の推定などが行われる．本報告では，複数箇所で得られたシークエンスデータから推定した，がんの進化系統樹情報を利用し，体細胞変異同定の高精度化を可能とする階層ベイズモデルの構築法について報告する．

2. 高精度化のアイデア

我々の手法では，シークエンスデータのノイズを，進化系統樹の情報を用い，確実にノイズと判定することで，偽陽性率を改善し，高精度化を測る．簡単のために，ノイズが起きたゲノム位置では，確率 0.5 で体細胞変異の同定ミスが生じるものとする．

進化系統樹を用いたノイズ検出が可能であることを示すため，ゲノム上の位置 m での体細胞変異に対して，N 箇所がん組織の変異パターンを導入する．これは，体細胞変異がN 箇所がん組織のどこで起きているかを N bit の 2 進数で示したものである．例えば，図 1(B)の d, e での変異は $110000_{(2)}, 001100_{(2)}$ というように表現できる．

ここで，ノイズと体細胞変異が，変異パターンにおいてどのように異なるのかを考える．N 箇所でがん組織を採取した場合，ランダムノイズのとる変異パターンは， 2^N 種類存在する．しかし，体細胞変異の場合は進化系統樹に従うため，取りうる変異パターンは高々 $2N-1$ 種類しか存在しない(図 1(A, B))．そのため，ランダムノイズが進化系統樹に矛盾のない確率は, $(2N-1)2^{-N}$ 程度である．よって N が十分に大きければ，ほとんどのランダムノイズが進化系統樹と矛盾するため，ランダムノイズ自体を検出することが出来る．

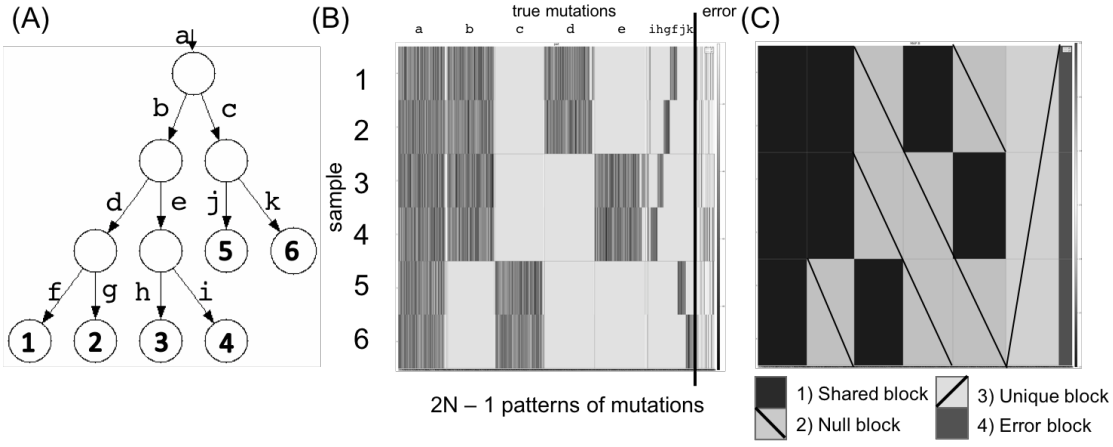

図 1 (A) 進化系統樹構造の例．辺 a-k は系統樹上での体細胞変異，葉 1-6 はがん組織の場所を示す．(B) (A)の系統樹を元にしたシミュレーションデータにおける，体細胞変異プロファイル．黒色は変異が存在することを示す．左側は真の体細胞変異，右側はランダムノイズ．(C)提案手法により推定されたブロックの構造.

3. モデル

我々は上記アイデアに基づいて，変異プロファイル上での体細胞変異とノイズの生成モデルを考え，それらに変異かノイズかを表す隠れ変数を置くような，階層ベイズモデルを構築した．体細胞変異では，図 1(C)にあるように， 1)複数組織に共通の変異が存在するブロック，2)変異が存在しないブロック，3)一つの組織にのみ変異が存在するブロックの３つが存在する．ノイズでは，4)ランダムノイズのみが存在するブロックのみが存在し，このブロックでは，進化系統樹を構築できない(Gusfield)．我々は，ノイズのような通常のものとは異なる列ブロックを含むバイクラスタリングの生成モデルを，Subset Infinite Relational Model(Ishiguro, et al.)をもとに構築した.

参考文献

D. Gusfield. Efficient algorithms for inferring evolutionary trees. Networks, 1991.
K. Ishiguro, N. Ueda, and H. Sawada. Subset infinite relational models. AISTATS, 2012.