Vol.2017-MPS-113 No.33
Vol.2017-BIO-50 No.33
2017/6/24

トピックモデルを用いた がんゲノムの変異シグネチャー解析

松谷太郎[1]　宇恵野雄貴[2]　福永津嵩[2,3]　浜田道昭[2,4]

概要：がんゲノムの変異パターンと，その背景にある変異源の分布は変異シグネチャー (Mutation Signature : MS) と呼ばれ，本研究では機械学習の手法を用いてこれを明らかにする．MS の推定は発がんメカニズムの解明の後押しになるなど重要な課題であり，先行研究では非負値行列因子分解や混合メンバーシップモデルを使った学習が行われていたが，MS の数が予測困難である等の問題点がある．本研究では MS ごとの変異の生成過程に対して潜在的ディリクレ再配置 (LDA) と呼ばれるトピックモデルを採用し，サンプルごとの体細胞突然変異からその背後にある生成モデルを推定する．学習に変分ベイズ法を用いることで，変分下限から MS 数を予測することが可能となり，シミュレーションベースではその推定に成功した．また, COSMIC データベースを用いた実データ解析にも着手している．

1. はじめに

発がんした細胞のゲノムには多数の突然変異が含まれている．従来のがんゲノムの研究では，がん細胞の増殖や生存にアドバンテージを与え，進化的な選択圧を受ける driver 変異に注目したターゲット型の研究が広く行われてきた。しかし，次世代シーケンサ技術の発展により大規模なゲノムデータが集積されつつある現在では，選択圧を受けない passenger 変異も，その背後に存在する変異源と深く関係しているとして解析の対象として注目されている [1].

これらの変異には，内因性, 外因性を問わずそれぞれに何らかの原因があると考えられる．例えば喫煙癖のあるがん患者には TP53 などのがん抑制遺伝子上で C が A になる変異が多く見つかることが知られている．この場合，喫煙癖を「変異源」，変異源に対して C が A に変化しやすい等の分布の情報を付加したものを「変異シグネチャー (Mutation Signature : MS)」と呼ぶことにする．ヒトが生まれてから発がんするまでの変異の蓄積は，複数の MS の作用の結果であると考えられ，このような MS の分類・及び解析は発がんメカニズムの根底的な解明や，早期診断におけるバイオマーカーに利用できる可能性があるとして期待されているものの，機械学習などの情報処理を用いた手法は未だ発展途中の段階にある．

1　早稲田大学 先進理工学部 電気・情報生命工学科
2　早稲田大学 先進理工学研究科 電気・情報生命専攻
3　日本学術振興会
4　産総研・早大 生体システムビッグデータ解析 オープンイノベーションラボラトリ

変異シグネチャーの作用によって表現される変異の蓄積データは，その変異源によってサンプル毎に大きく異なる．このように観測されるデータに大きな偏りがある場合，トピックモデルなどの，文書構造を持つデータに対して単語の共起性を元に生成モデルを構築する手法が有効であると考えられる．トピックモデルは，専ら自然言語処理などの分野において用いられ，複数の文書が与えられたとき，それぞれの文書中に現れる単語はその背後に存在するトピックに基づき生成されていると仮定し，文書ごとのトピック分布とトピックごとの単語分布を学習することによって元のデータを表現する．変異データに対してこのトピックモデルを適用したとき，文書をサンプル，単語を変異，トピックを変異源と見なすことで，トピックごとの単語分布である MS を学習する．

トピックモデル的な手法を用いた変異シグネチャーの解析の先行研究としては，非負値行列因子分解 (Nonnegative Matrix Factorization : NMF) を使ったもの [1][2] や，混合メンバーシップモデルを使ったもの [3] が報告されている．しかし, いずれもトピック数を既知のものとして与えねばならず，そのモデル選択も適切なトピック数を決めることが難しいことなどの改善点が残っていた．そこで，本研究ではトピックモデルの 1 つである潜在的ディリクレ再配置 (Latent Dirichlet Allocation : LDA) を用いた MS の解析を提案する．LDA を用いた解析においてもトピック数は既知のものとする必要があるが，学習を変分ベイズ法によって行うことで変分下限と呼ばれる指標をモデル選択に利用できることや，共役事前分布としてディリクレ分布を導入

Vol.2017-MPS-113 No.33
Vol.2017-BIO-50 No.33
2017/6/24

することによって過学習を回避できることが期待される．

2. 理論と方法

2.1 潜在的ディリクレ再配置 (LDA)

LDA は，文書数 M と各文書における単語数 n_d，単語の種類数 V が観測データから与えられ，トピック数 K を解析者が与えたとき，文書中の各単語 $\boldsymbol{w}$ は潜在変数であるトピック $\boldsymbol{z}$ に従って出力されると仮定する確率モデルである (図 1)[4]．

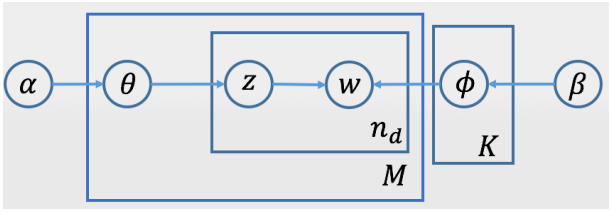

図 1　LDA のグラフィカルモデル

- $\boldsymbol{\theta}_d (1 \leq d \leq M)$ は文書ごとのトピック分布（多項分布）で，ハイパーパラメータ $\alpha_k (1 \leq k \leq K)$ に従うディリクレ分布から生成される．
- $\boldsymbol{\phi}_k (1 \leq k \leq K)$ はトピックごとの単語分布（多項分布）で，ハイパーパラメータ $\beta_v (1 \leq v \leq V)$ に従うディリクレ分布から生成される．

また，これらの生成過程は以下の式 (1)，式 (2) のように定式化される [5]．

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha}),\ \boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{\beta}) \tag{1}$$

$$z_{d,i} \sim \mathrm{Multi}(\boldsymbol{\theta}_d),\ w_{d,i} \sim \mathrm{Multi}(\boldsymbol{\phi}_k)\ (1 \leq i \leq n_d) \tag{2}$$

2.2 変分ベイズ法 (VB)

LDA の学習方法としては，主にギブスサンプリングと変分ベイズ法 (Variational Bayes： VB) が用いられている．本研究では，変分ベイズ法の理論の中で扱われる変分下限をモデル選択のための指標として利用するため，VB を使用した．

LDA の一般的な目標は，適切な潜在変数とパラメータの組 $\{\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}\}$ を推測することによって，観測データ $\boldsymbol{w}$ をよく表現することである．しかし実際のところ事後分布 $p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ を解析的に解くことは不可能であるため，因子分解可能という仮定の元で，式 (3) に表される近似事後分布 $q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})$ が真の分布と最も近くなるように学習を行うことになる．

$$q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{d=1}^{M} \prod_{i=1}^{n_d} q(z_{d,i}) \prod_{d=1}^{M} q(\theta_d) \prod_{k=1}^{K} q(\phi_k) \tag{3}$$

近似事後分布の良さの指標として，まず考えられるのは真の分布との KL 距離であり，これを最小化するような近似事後分布を選択すれば良いように思われる．しかし，この距離は事後分布 $p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ を陽に含むため，導出することができない。そこで z について周辺化した対数周辺尤度 $\log p(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ に関する，式 (4) の性質を用いると，式 (5) のように近似事後分布に関する最適化問題を定式化することができる．

$$\begin{aligned} &\mathbb{KL}[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) \,||\, p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})] \\ &\quad = \log p(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) - F[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})] \end{aligned} \tag{4}$$

$$q^*(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \underset{q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) \in \mathcal{Q}}{\arg\max}\ F[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})] \tag{5}$$

- $F[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})]$ は対数周辺尤度の下限という意味で「変分下限 (Variational Lower Bound： VLB)」と呼ぶ．KL 距離は非負の値をとり，対数周辺尤度は近似事後分布と独立であるため，変分下限の最大化と KL 距離の最小化は同値となる．
- $\mathcal{Q}$ は因子分解可能な近似事後分布の集合を表す．

変分下限はベイズの定理を用いて，式 (6) のように展開することができる．

$$\begin{aligned} &F[q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})] \\ &= \int \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) \log \frac{p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})} d\boldsymbol{\theta} d\boldsymbol{\phi} \\ &= \int \sum_{d=1}^{M} \sum_{i=1}^{n_d} q(z_{d,i}) q(\boldsymbol{\theta}_d) q(\boldsymbol{\phi}_k) \log p(w_{d,i} \mid z_{d,i}, \boldsymbol{\phi}) p(z_{d,i} \mid \boldsymbol{\theta}_d) d\boldsymbol{\theta} d\boldsymbol{\phi} \\ &\quad - \sum_{d=1}^{M} \sum_{i=1}^{n_d} \sum_{k=1}^{K} q(z_{d,i} = k) \log q(z_{d,i} = k) \\ &\quad - \sum_{d=1}^{M} \mathbb{KL}[q(\boldsymbol{\theta}_d) \,||\, p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha})] \\ &\quad - \sum_{k=1}^{K} \mathbb{KL}[q(\boldsymbol{\phi}_k) \,||\, p(\boldsymbol{\phi}_k \mid \boldsymbol{\beta})] \end{aligned} \tag{6}$$

近似事後分布は因子分解可能の仮定を置いているため，$q(\boldsymbol{z})$，$q(\boldsymbol{\theta})$，$q(\boldsymbol{\phi})$ について関係のある項を式 (6) よりそれぞれ抜き出して，変分法によって最大化するような更新式を実行すれば良い（式 (7)，(8)，(9)）．

$$\begin{aligned} &q(z_{d,i} = k) \\ &\propto \frac{\exp\left[\Psi(\xi^{\phi}_{k,w_{d,i}})\right]}{\exp\left[\Psi(\sum_{v'=1}^{V} \xi^{\phi}_{k,v'})\right]} \frac{\exp\left[\Psi(\xi^{\theta}_{d,k})\right]}{\exp\left[\Psi(\sum_{k'=1}^{K} \xi^{\theta}_{d,k'})\right]} \end{aligned} \tag{7}$$

$$q(\boldsymbol{\theta}_d \mid \boldsymbol{\xi}^{\theta}_d) = \frac{\Gamma(\sum_{k=1}^{K} \xi^{\theta}_{d,k})}{\prod_{k=1}^{K} \Gamma(\xi^{\theta}_{d,k})} \prod_{k=1}^{K} \theta_{d,k}^{\xi^{\theta}_{d,k} - 1} \tag{8}$$

$$q(\boldsymbol{\phi}_k \mid \boldsymbol{\xi}^{\phi}_d) = \frac{\Gamma(\sum_{v=1}^{V} \xi^{\phi}_{k,v})}{\prod_{v=1}^{V} \Gamma(\xi^{\phi}_{k,v})} \prod_{v=1}^{V} \theta_{k,v}^{\xi^{\phi}_{k,v} - 1} \tag{9}$$

Vol.2017-MPS-113 No.33
Vol.2017-BIO-50 No.33
2017/6/24

- $\Psi(\cdot)$ をディガンマ関数，$\Gamma(\cdot)$ をガンマ関数とする．
- $\boldsymbol{\xi}^{\theta}$，$\boldsymbol{\xi}^{\phi}$ はそれぞれ $q(\boldsymbol{\theta})$，$q(\boldsymbol{\phi})$ をディリクレ分布と見なしたときのパラメータになっており，以下のように更新する．

$$\begin{aligned}\xi^{\theta}_{d,k} &= \mathbb{E}_{q(\boldsymbol{z}_d)}[n_{d,k}] + \alpha_k \\ &= \sum_{i=1}^{n_d} q(z_{d,i}=k) + \alpha_k \qquad (10)\\ \xi^{\phi}_{k,v} &= \mathbb{E}_{q(\boldsymbol{z}_d)}[n_{k,v}] + \beta_v \\ &= \sum_{d=1}^{M}\sum_{i=1}^{n_d} q(z_{d,i}=k)\delta(w_{d,i}=v) + \beta_v \qquad (11)\end{aligned}$$

2.3 ハイパーパラメータの学習

これまでの議論は全て事前分布を既知のものとして扱ってきたが，実際にはハイパーパラメータも観測されたデータに基づいて学習できる方が望ましい．ハイパーパラメータの学習においても，目的関数には変分下限を用いるのが自然であり，本研究では固定点反復法を用いて学習を行う．

固定点反復法では関数 $f(x)$ に対して $x = f(x)$ を満たす不動点 x を初期値 x_0 から反復計算を行うことで探索する．ディリクレ分布のハイパーパラメータ推定では，変分下限の更に下限をとってそれを最大化するような固定点を見つけるのが目的となる．詳細は省略するが $\boldsymbol{\xi}^{\theta}$，$\boldsymbol{\xi}^{\phi}$ の更新後，$\boldsymbol{\alpha}$ についてそれぞれ関係のある項のみ取り出したとき変分下限は式 (12) のように下限をとることができる [5]．

$$\begin{aligned}&F[q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})]\\ &\geq \sum_{d=1}^{M}\left[-b_d\sum_{k=1}^{K}\alpha_k + \sum_{k=1}^{K} a_{d,k}\log\alpha_k\right] + (\text{const.}) \qquad (12)\end{aligned}$$

- $a_{d,k} = (\Psi(\mathbb{E}[n_{d,k}] + \hat{\alpha}_k) - \Psi(\hat{\alpha}_k))\,\hat{\alpha}_k$
- $b_d = \Psi\left(n_d + \sum_{k=1}^{K}\hat{\alpha}_k\right) - \Psi\left(\sum_{k=1}^{K}\hat{\alpha}_k\right)$

この下限を最大化するような $\boldsymbol{\alpha}$ は式 (12) の右辺を微分して 0 になるような値となり，その更新式は式 (13) になる．$\boldsymbol{\beta}$ についても同様の操作を行ったとき，式 (14) のような更新式で停留点を求めることができる．なお，この固定点反復法で求められる停留点は必ずしも不動点に収束するとは限らない．

$$\alpha_k = \frac{\sum_{d=1}^{M}[\Psi(\mathbb{E}[n_{d,k}+\hat{\alpha}_k]) - \Psi(\hat{\alpha}_k)]\hat{\alpha}_k}{\sum_{d=1}^{M}\left[\Psi(n_d + \sum_{k=1}^{K}\hat{\alpha}_k) - \Psi(\sum_{k=1}^{K}\hat{\alpha}_k)\right]} \qquad (13)$$

$$\beta_v = \frac{\sum_{k=1}^{K}[\Psi(\mathbb{E}[n_{k,v}+\hat{\beta}_v]) - \Psi(\hat{\beta}_v)]\hat{\beta}_v}{\sum_{k=1}^{K}\left[\Psi(\sum_{v=1}^{V}\mathbb{E}[n_{k,v}] + \hat{\beta}_v) - \Psi(\sum_{v=1}^{V}\hat{\beta}_v)\right]} \qquad (14)$$

2.4 VB-LDA

2.2，2.3 に挙げた更新式を元に今回実際に用いた計算手順を Algorithm1 に示す．

Algorithm 1 Variational Bayes for LDA

Require: サンプル数 M，単語数 n_d，トピック数 K
1: $q(\boldsymbol{z})$ と $\boldsymbol{\alpha}, \boldsymbol{\beta}$ をランダムに初期化する．
2: 初期化したパラメータを用いて $\boldsymbol{\xi}^{\theta}, \boldsymbol{\xi}^{\phi}$ を求める式 ◁(10)(11)
3: 以下反復
4: $q(z_{d,i})$ $(1 \leq d \leq M,\ 1 \leq i \leq n_d)$ を更新 ◁ 式 (7)
5: $q(\theta_d)$ $(1 \leq d \leq M)$ を更新 ◁ 式 (8)(10)
6: $q(\phi_k)$ $(1 \leq k \leq K)$ を更新 ◁ 式 (9)(11)
7: α_k, β_v $(1 \leq k \leq K,\ 1 \leq v \leq V)$ を更新 ◁ 式 (13)(14)
8: 反復終了

反復の終了条件は式 (6) に示される変分下限の値の各イテレーション毎の差が十分に小さくなったとき，あるいは反復回数が 1000 回を超えたときとした．

またトピック数 K，すなわちシグネチャーの数は解析者が与える必要があるが，実データ解析におけるシグネチャー数の決定には生物学的な専門知識を要する．しかし，VB-LDA では反復終了時の変分下限の値を比べて最も大きかったときのトピック数 K^* が最適トピック数であると決めることができる．

2.5 変異の単語表現

COSMIC データベース[*1]には，一塩基置換の他にも挿入や欠失 (indel)，連続置換を含めた変異が登録されている．だが，それらの変異は一塩基置換と比べてどのような単語として扱うか，つまり特徴の選択が難しいため今回は一塩基置換のみを扱うことにする．

一塩基置換は置換前の塩基が A, C, G, T の 4 種類，置換後の塩基が置換前の塩基を除いた 3 種類存在するため $4 \times 3 = 12$ 種類の変異が考えられる．そこで，例えば置換前にシトシンだった塩基がチミンに変異した場合，[C>T] と表記することにする．しかし，DNA は構造上，相補鎖と塩基対を形成しているため，対合している塩基のどちらが原因で変異が起こったのかを決めることはできない．例えば，[C>T] という一塩基置換が起こったとき，その相補鎖では同時に [G>A] という一塩基置換が起こっており，その 2 つを差別化することができない．これが原因となって実際に区別することのできる一塩基置換は $12 \div 2 = 6$ 種類となる．今回は {[C>A]，[C>G]，[C>T]，[T>A]，[T>C]，[T>G]} をその 6 種類として選んだ．

また本研究では置換変異が起こった場所と隣り合う塩基の種類の情報が，変異シグネチャーの分類に重要であるとして，5' 側と 3' 側の隣接塩基を単語の中に含ませた．例えば，上流の G と下流の T に挟まれた C が A に置換された場合，G[C>A]T と表現することにする．この場合，それぞれの隣接塩基の場合の数は 4 種類ずつであるため，総単語種類数は $4 \times 4 \times 6 = 96$ となる．

[*1] http://cancer.sanger.ac.uk/cosmic

Vol.2017-MPS-113 No.33
Vol.2017-BIO-50 No.33
2017/6/24

こういった単語の選び方の妥当性を示す例として，メチル化シトシンの脱アミノ反応などが挙げられる．3’ 側の G と隣り合う C（CpG 部位）はメチル化が起こりやすいことが知られており，またこのようにしてできたメチル化シトシンは脱アミノ化を起こしてチミンになりやすく，実際に COSMIC には X[C>T]G（X は任意の塩基）系列の変異の割合が大きいシグネチャーの存在が報告されている (図 2)．

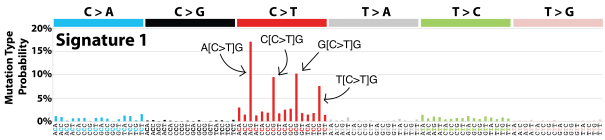

図 2　Signature1

3. 結果

3.1 シミュレーション

VB-LDA がどのような条件の下では適切に動くのかを確かめるために，条件を変えてシミュレーションを行った．報告されている変異シグネチャーのうち，Signature1 から Signature11 までの変異分布から各変異は発生するとして，サンプル数 M，1 サンプルに含まれる変異数 n_d，サンプルごとのトピック分布 $\boldsymbol{\theta}_d$ に対するハイパーパラメータ α_k $(1 \leq k \leq 11)$ をそれぞれ変化させてサンプルデータを生成し，2.4 の手順に従い解析を行った．

このとき元の分布を再現できているかどうかの指標として Jensen-Shannon 距離（JS 距離）を用いる（式 15）．

$$\mathbb{JS}(p_1 \,||\, p_2) = \frac{1}{2}\mathbb{KL}(p_1 \,||\, q) + \frac{1}{2}\mathbb{KL}(p_2 \,||\, q) \qquad (15)$$
$$\left(q = \frac{1}{2}p_1 + \frac{1}{2}p_2\right)$$

$\mathbb{JS}(p_1 \,||\, p_2)$ の値が小さいほど分布 p_1, p_2 は類似している．推測したトピックの変異分布 $\boldsymbol{\phi}_k$ $(1 \leq k \leq 11)$ との JS 距離を最も小さくする正解分布が，対応している Signature であると考え，以下の 3.1.1〜3.1.3 では，抽出 Signature 数と，それぞれの予測分布と Signature の平均 JS 距離を評価基準として用いる．この場合，抽出 Signature 数は 11 に近いほど良く，平均 JS 距離は小さいほど元の分布をよく表現できているといえる．

3.1.1 サンプル数を変化させたとき

$n_d \simeq 2000, \alpha_k \simeq 0.1$ で固定し，サンプル数を {20, 100, 1000, 5000} の間で変化させたときの比較を表 1 に示す．

表 1　サンプル数 M を変化させたときの結果

M	20	100	1000	5000
抽出 Signature 数	10	10	11	11
平均 JS 距離	0.0459	0. 0689	0.0201	0.0362

3.1.2 1 サンプルに含まれる変異数を変化させたとき

$M = 1000, \alpha_k \simeq 0.1$ で固定し，1 サンプルに含まれる変異数を {100, 200, 400, 2000} の間で変化させたときの比較を表 2 に示す．

表 2　1 サンプル辺りの変異数 n_d を変化させたときの結果

n_d	100	200	400	2000
抽出 Signature 数	9	9	10	11
平均 JS 距離	0.3335	0.2445	0.0542	0.0201

3.1.3 ハイパーパラメータ α を変化させたとき

$M = 1000, n_d \simeq 2000$ で固定し，ハイパーパラメータ α_k を {0.1, 0.5, 1, 10} の間で変化させたときの比較を表 3 に示す．

表 3　ハイパーパラメータ α_k を変化させたときの結果

α_k	0.1	0.5	1	10
抽出 Signature 数	11	10	9	1
平均 JS 距離	0.0201	0.1236	0.0489	0.0585

3.1.4 VB-LDA の扱えるデータ

3.1.1，3.1.2 の結果から，VB-LDA が真の分布をよく近似した事後分布を推測するにはある程度のサンプル数と変異数が必要であることが分かる。また，式 (1) のように生成されるサンプルごとのトピック分布 $\boldsymbol{\theta}_d$ に偏りがあるとき，ハイパーパラメータ α_k の値は十分に小さく，3.1.3 の結果から，そのようなデータセットに対してのみ LDA が有効であることが分かる．

適当なデータ（$M = 1000, n_d \simeq 2000, \alpha_k \simeq 0.1$）でトピック数を 11 としたとき，11 個の Signature は全て抽出され，近似分布 $\boldsymbol{\phi}_k$ は元の分布をよく推定することができていた（図 3）．

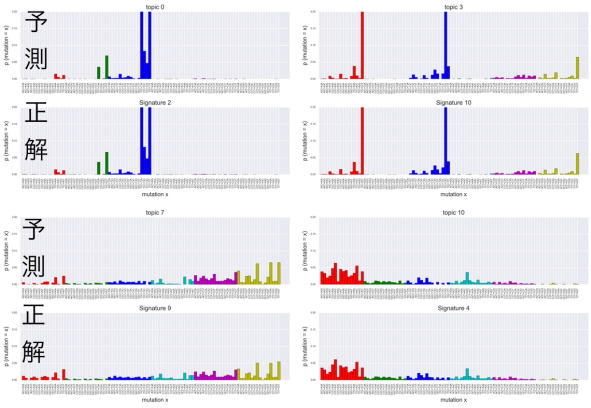

図 3　予測分布と正解分布の比較 *2

*2　紙面の都合上，推定された 11 個の単語分布のうち 4 つのみを表示している．

Vol.2017-MPS-113 No.33
Vol.2017-BIO-50 No.33
2017/6/24

3.1.5 変分下限を用いたモデル選択

VB-LDA を用いた解析では 2.4 で述べたように変分下限を用いたモデル選択が期待される．適切なデータ（$M = 1000, n_d \simeq 2000, \alpha_k \simeq 0.1$）を用いてトピック数を 2 から 50 まで変化させたとき，それぞれの反復終了時の変分下限を比べたところ，正解トピック数である 11 付近で最大化されていることが分かった（図 4）．

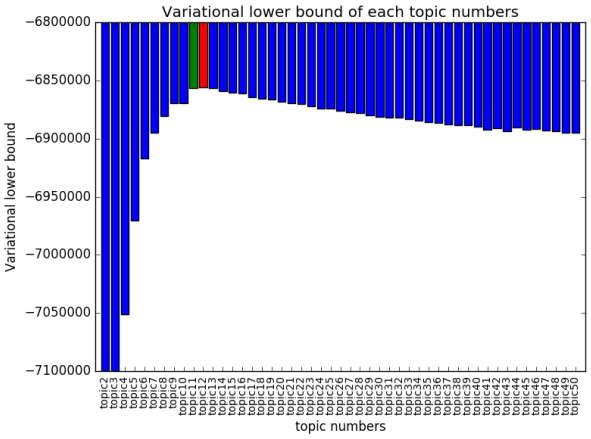

図 4　トピック数に対する変分下限の変化（シミュレーション）*3

3.2 実データ解析

COSMIC に登録されているデータの中には 1 つのサンプルに対して 1 つの変異しか登録されていないものなども多く，3.1.2 からも分かるように変異の共起性から推論を行う VB-LDA にとってこれらのデータは解析の障害となる．以下の解析では安定した解析が行えるよう，全データの中から 400 個以上変異が登録されているサンプルのみを用いて解析を行っている．また，全てのデータを用いて解析を行ったとき，「がん細胞のゲノムである」という性質以外，共通の特徴を持たず生物学的な解釈が難しくなってしまうため，原発巣の部位ごとにデータを分けて解析した．

3.2.1 肺がんゲノムの解析

肺がんゲノム（サンプル数 175）に対して学習を行い，トピック数を 2 から 20 まで変化させたときの変分下限を比較したところ，$K = 6$ で最大化された（図 5）．

そこで，$K = 6$ のときの単語分布を COSMIC に報告されている Signature と対応付けたところ，喫煙習慣を変異源とし，ベンゾ [a] ピレンの暴露とも関連がある [6]Signature4 に近いトピックを抽出できていることが分かった．また，紫外線光の暴露によって引き起こされると考えられ，皮膚がん患者によく見られる Signature7 に極めて類似したトピックも抽出された（図 6）．

*3　K=12(赤) で最大化され，K=11(緑) が次点で大きかった．

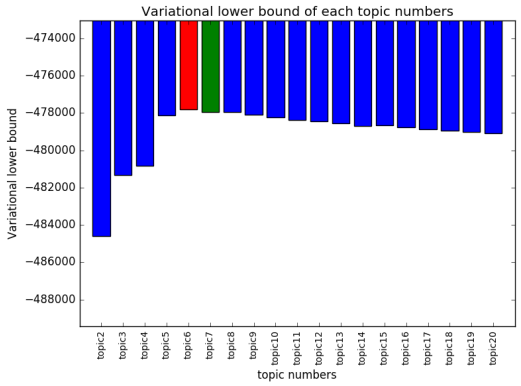

図 5　変分下限の変化（肺がん）

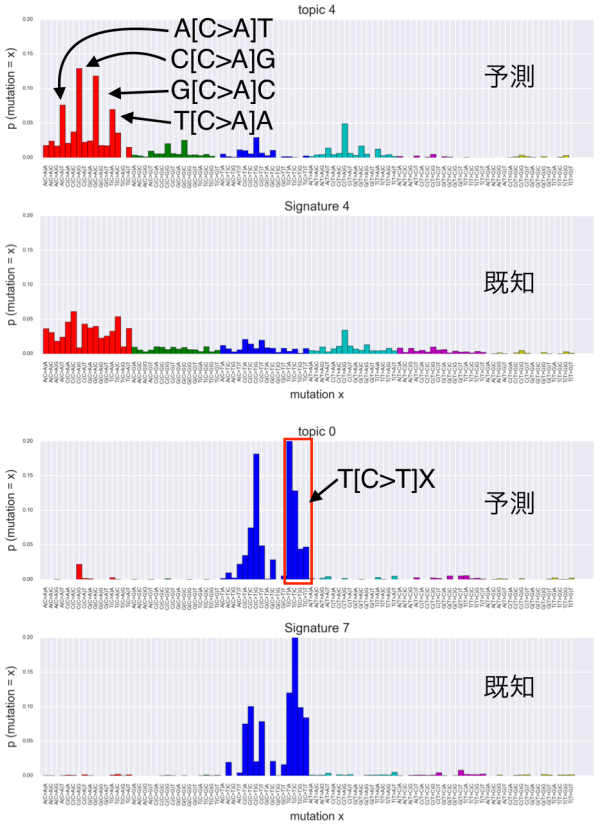

図 6　肺がんゲノムに見られる特徴的なトピックの単語分布

3.2.2 肝臓がんゲノムの解析

肺がんゲノムと同様に肝臓がんゲノム (サンプル数 64) についても学習を行い，トピック数を 2 から 20 まで変化させたときの変分下限を比較したところ，$K = 11$ で最大化された（図 7）．

$K = 11$ のときの単語分布を現在報告されている Signature と対応付けたところ，Signature12，Signature16，Signature22，Signature24 など肝臓がんに見られるといわれているシグネチャーに近い分布を持つトピックを抽出することができた．しかし，一方で Signature23 と分布が似ているトピックが 3 つ抽出されるなど分布推定，あるいはトピック数推定の精度に問題が見られた（図 8）．

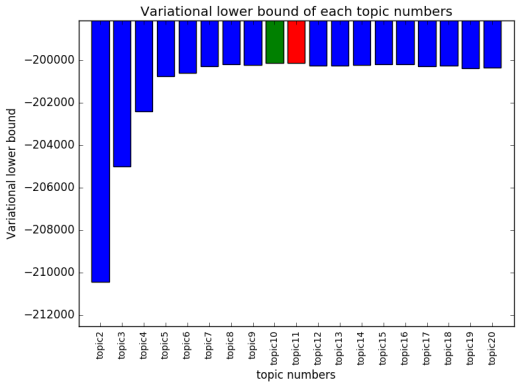

図 7 変分下限の変化（肝臓がん）

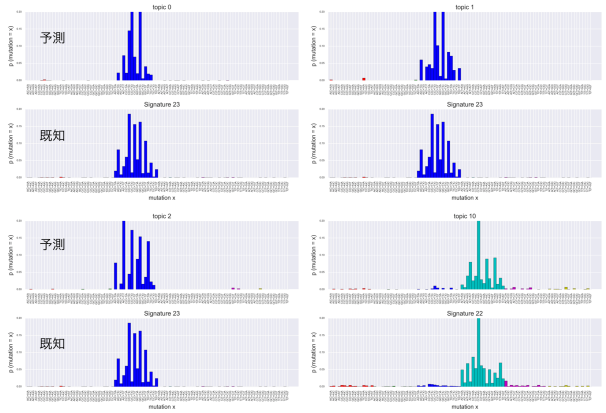

図 8 肝臓がんゲノムから抽出された一部のシグネチャーの単語分布

4. 考察・今後の展望

4.1 使用する単語の種類

今回用いた 96 種類の変異では上流と下流それぞれ 1 塩基までの情報のみを利用しているが，それぞれ 2 塩基先までの情報を用いた変異で解析を行うことも可能であり，この場合，変異は $4 \times 4 \times 6 \times 4 \times 4 = 1536$ 種となる．実際に乳がんのデータから YpTpCpXpX 部位（Y は「C か T」を示す）の C が T に置換されやすいシグネチャーの存在が報告されている [1]．

また，今回の実験ではデータとして一塩基置換のみを扱ったが，実際の突然変異には indel や連続置換なども多く存在する．これらの変異も含めて解析を行わなければ真のシグネチャーの分布を推定することはできないと考えられるが，挿入・欠失の中には数十塩基単位で起こるものも存在し，一塩基置換の場合と同じように隣接塩基も注目したとき，その種類は膨大な数になってしまう。また，稀にしか出現しない変異が多くなるとデータ構造がスパースなものとなり，トピック推定が難しくなるなどの問題が考えられるため，実際に挿入されたり欠失した塩基の情報については大まかな分類に留めるなど工夫するべきである．

更に一部のシグネチャーに見られる染色体上の特定部位に集中する高頻度変異などは，これまで論じてきたような単語の選び方では，その情報を抽出することができない．AID/APOBEC ファミリータンパクによって引き起こされ，乳がん患者のゲノムに見られる kataegis などがその代表例であり [7]，こういった変異の特徴を抽出するには変異同士の距離を用いるなど別のアプローチをとる必要がある．

4.2 モデル選択

VB-LDA における変分下限を用いたモデル選択は 3.1.5 の場合のように上手く行くときもあるが，3.2.2 のときのように変分下限が最大化されたトピック数付近での変動が殆どない場合もあり（図 7），トピック数に対して正則性があるとはいえない．また，3.2.2 においては Signature23 に近似されるトピックが 3 つ抽出される（図 8）など解釈性に乏しい結果が得られた．

このような問題を解決するためのモデル選択の手法として，近年提案された因子化情報量基準 (Factorized Information Criterion : FIC)[8] に基づく推論と Shrinkage Mechanism の利用が考えられる．FIC を用いた推論では，事後分布の計算において解析的に解くことのできない積分をラプラス近似することによって変分下限を別の形で導出する．

FIC を利用した EM アルゴリズムでは，異なる潜在変数（クラス）に対応する分布が似通っているとき，そのどちらかの負担率を貪欲に下げ，その結果殆ど機能しなくなったクラスを削除し正規化し直すことで解釈性に富む結果を得ることを可能とする．LDA では潜在変数がトピックに対応しているため，不要になったトピックが削除されることで 3.2.2 のときのような結果（図 7, 8）を回避できることが期待される．

参考文献

[1] Ludmil B. Alexandrov, et al. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports 3*, 246-259. (2013).
[2] S. Nik-Zainal, et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell 149*, 979-993. (2012).
[3] Yuichi Shiraishi, Georg Tremmel, Satoru Miyano, Matthew Stephens. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLOS GENETICS 11(12)*, e1005657. (2015).
[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993-1022. (2003).
[5] 奥村学，佐藤一誠．トピックモデルによる統計的潜在意味解析．初版． コロナ社． (2015).
[6] S. Nik-Zainal, et al. The genome as a record of environmental exposure. *mutagenesis 30(6)*, 763-770. (2015).
[7] Artem G Lada, et al. AID/APOBEC cytosine deaminase induces genome-wide kataegis. *Biology Direct 7:47*. (2012).
[8] Ryohei Fujimaki, Satoshi Morinaga. Factorized Asymptotic Bayesian Inference for Mixture Modeling. In *AISTATS 22*, 400-408. (2012).