DNA を機能区間に分類する
線形時間区間推定アルゴリズムの実装

2015/12/22

森下研究室 鈴木裕太

区間推定問題

- 一次元的に並んだデータ(DNA !)を、特徴によって複数種の領域(区間)へ分割する問題
  - 遺伝子領域の予測
    - reference genome の anotation
  - CNV(copy number variation)領域の予測
    - short-read 特有の問題
  - Understanding epigenetic features
    - ChromHMM など

演習の内容

- 区間推定問題を解くアルゴリズムを一つ実装する
  - HMM or 最小長制約分割
- 結果を吟味するための手法を考察する
  - 遺伝子領域・CpG islands との相関
  - MDL基準の紹介

- 各アルゴリズムの詳細は、講義資料や過去の演習の資料も参照

www.gi.k.u-tokyo.ac.jp/~moris/2013_2_DNA_range_decomposition.pdf

www.gi.k.u-tokyo.ac.jp/~moris/201211MethylationAdaboost.pdf

メチル化状態の区間推定

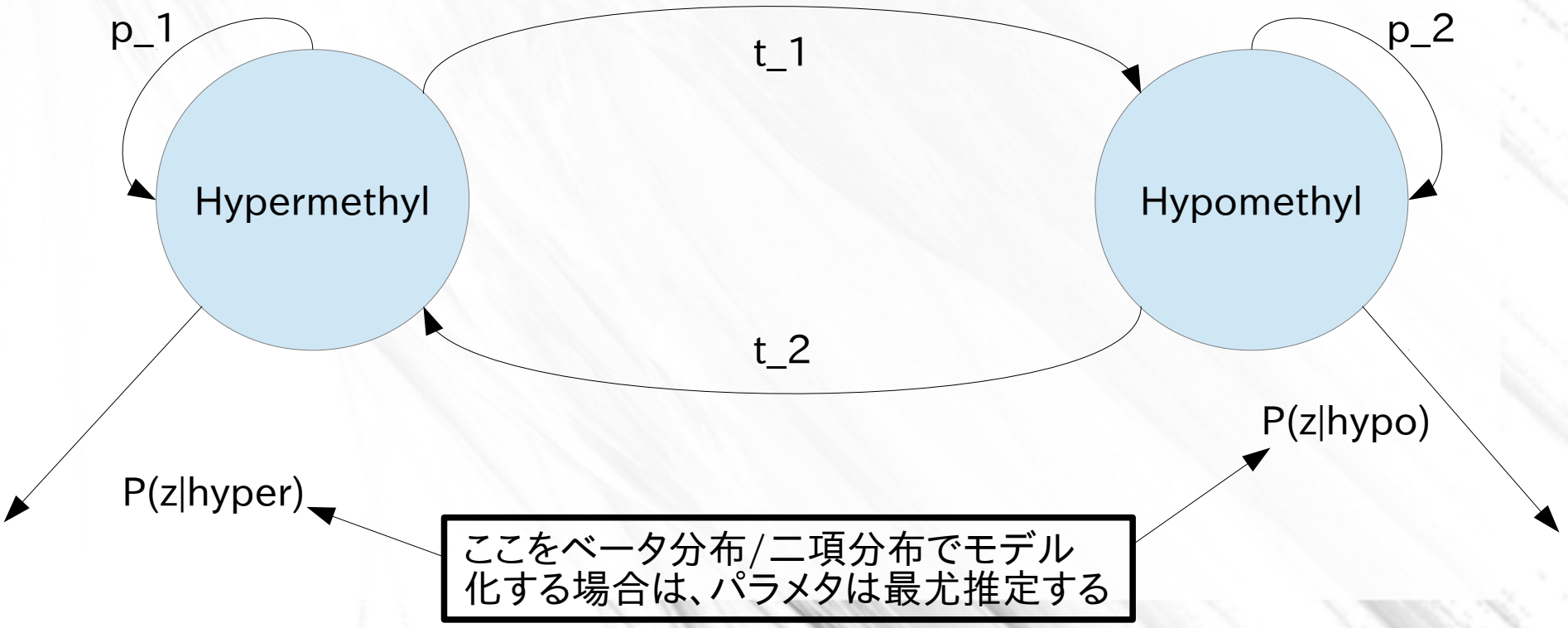
- DNAメチル化、今回は特にCpGメチル化を考察する
- 微生物・植物・動物はCpG以外でもメチル化を様々に利用している
  - 制限酵素システム
  - トランスポゾンの活動抑制
  - 転写制御（～分化と発生）
  - 環境への応答（～老化・ガン化）
- 脊椎動物DNAはほとんどのCpGがメチル化している中に、機能的な低メチル化領域をもつ
- Bisulfite-seq データから高/低メチル化領域を見つけることが典型的な問題になる

問題の定義（の例）

- 問題
  - 入力：bisulfite-seq での unconverted C の割合を示した実数(Double)の列
  - （or 入力：二種のリードの数: (Int, Int) の列）
  - 出力：低メチル化領域のリスト
- モデル
  - 高メチル化/低メチル化の2つの隠れ状態
  - それに対応した二つの [0,1] 上のベータ分布
  - （or 二つの二項分布）

HMM による区間推定

- パラメタと出力確率の関数型を適当に決め、Viterbiアルゴリズムで隠れ状態列を推定する。



HMM の特徴

- 隠れ状態列の推定は線形時間
- 多くの分布に対してパラメタの推定も高速

- いくつかの前提を（本来は）吟味すべき
  - マルコフ性を満たす
  - 区間の長さが幾何分布に従う
    - これを緩和した HSMM などの拡張がある

最小長さ制約付き区間推定

- 所与の実数列 $r\_1, \ldots, r\_n$ と正数Lに対し、次の式

  $\Sigma\_{k=1\sim K}\ \Sigma\_{i \text{ in } I_k}\ r\_k$

  を最大にする区間の集合$\{I_k\}$を求めよ。

  但しKは(固定せず)自然数を動かし、

  各I_kの大きさはL以上となるようにせよ。

- (hint) 入力のサイズに関する動的計画法を用いる。
- (hint) $L=1$, $L=2$ のときは、どうなる？
- (hint) $O(nL)$ のアルゴリズムから、$O(n)$へ

区間推定問題の変種たち

# of Segments	Maximize	Length Const.	Density Const.	Complexity	Literature	Note
1	Sum	None	None	O(n)	Kadane, Bentley	
1	Sum	L<U	None	O(n)	Lin et al. (2002)	
1	Density	L<	None	O(n log L)	Lin et al. (2002)	
1	Density	L<U	None	O(n)	Goldwasser et al. (2004)	Uniform length
1	**Density**	**L<U**	**None**	**O(n)**	**Chung and Lu (2004)**	**Non-uniform length, Optimal, beat Goldwasser O(nlog(U-L))**
* (non overlapping)	Sum of Sum	L<	None	O(n)	Csuros (2004)	
non overlapping k	Sum of Sum	None	None	O(kn)	Brin	
non overlapping k	**Sum of Density**	**L<**	**None**	**O(n+k^2 L log L)**	**Liu and Chao (2006)**	**Best among known in this setting, beat Bergkvist (2005) O(nL + k^2L^2) and Chen et al. (2005) O(nLk)**
overlapping k	**Sum**	**None**	**None**	**O(n+k)**	**Brodal & Jorgensen (2007)**	**Optimal, beat Cheng et al. (2005) O(n+k log min{n,k})**
overlapping k	Sum	L<U	None	expected O(n log(U-L) + k)	Lin and Lee (2007)	Randomized
overlapping k	**Sum**	**L<U**	**None**	**O(n+k)**	**Liu and Chao (2008)**	**Independent & strict extension of Brodal and Jorgensen (2007) both relying essentially Frederickson's heap alg. Working on Tree**

課題1 / 低メチル化区間の検出

- 低メチル化区間を検出するアルゴリズムを実装する。例として、
  - HMM(パラメタは固定してしまっても良い)
  - K個の低メチル化区間の検出(昨年の資料/講義の資料参照)
  - N CpG sites 以上を含む高/低メチル化区間への分割
  - 示されていない詳細は適宜自由に埋めてください
- データは hub_8415_Roadmap_Human_2015 から好きな細胞腫を選んでください。
- bigWigはこちらのbigWigToWigで変換できます。
  http://hgdownload.cse.ucsc.edu/admin/exe/

課題2 / CpG islands の検出

- hg38 参照配列 から CpG islands を検出する。
- CpG islands の定義は以下のものがよく参照される。

  Gardiner-Garden, Frommer (1987) CpG islands in vertebrate genomes. http://www.ncbi.nlm.nih.gov/pubmed/3656447
- CGIs are regions with ≧ 200 bp, O/E CpG ≧ 0.6, %GC ≧ 0.5
- 上の論文では、100 bp 移動平均で O/E CpG, %GC を計算しておき、両基準を満たす位置が200 bp以上続く箇所を取ってきている。
  - 例えば O/E = #CpG / ([%C][%G] * 100)
  - 遺伝子上流配列のみを利用していて、もちろん全ゲノムではない
- 上述の基準に従った CpG islands を求める。
  - (or 100 bp 移動平均を使わず、可能な全部の区間を列挙する場合のアルゴリズムを考察してもよい。包含関係にある区間に対し最大の区間のみを報告する場合、計算量は改善できる?)
- or CGI 内部/外部の隠れ状態を考えて、課題1のアルゴリズムを修正して使っても良い

課題3 / 結果の考察

- 課題1,2で調べた

  a. 低メチル化領域　b. CpG islands　c. 遺伝子領域

  これらの関連を調べる。

  - (プロモータ領域の) CpG islands は低メチル化していることが期待されるが、それを確かめよう。
  - 低メチル化CGIを伴う遺伝子のリストを作ろう。
    - 遺伝子領域データは GENCODE を使う。
    - (どんな遺伝子が多いのか、GOとの比較)
- R や gnuplot を利用して、結果を要約・可視化できればなおよい。

MDL基準によるモデル選択

- 別の資料に基づき解説します。
- 必須の課題はありません。

課題の提出

- 提出するもの
  - ソースコード
  - 実行結果は一部分で良いです。
  - 簡単なレポートや図表があれば付けてください。
- 提出の方法
  - 私と森下先生を宛先にしてメールに添付
  - moris@cb.k.u-tokyo.ac.jp
  - ysuzuki@cb.k.u-tokyo.ac.jp
  - 2016/01/31締切
  - 質問はいつでも歓迎します