確率モデルによる配列解析

津本　周作
島根大学医学部医療情報学講座
平成16年7月21日

Outline

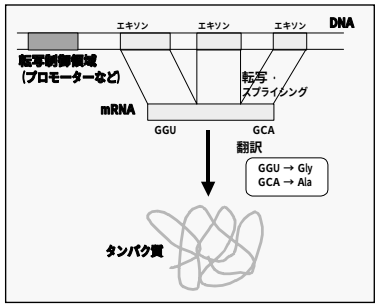
① 分子生物学の基本事項
② Bioinformaticsに必要な確率統計の基礎
③ 配列アラインメントとスコア関数
④ 隠れマルコフモデル
⑤ EMアルゴリズム

ゲノム研究と情報科学

- ゲノム解析計画の急速な進展
  - 既に数十種の微生物の全塩基配列が決定
  - 線虫（1000細胞, C.elegans）他、ショウジョウバエも決定
  - ヒトも概要配列が公開済み
- 情報解析の必要性
  - DNA配列⇔プログラムのオブジェクトコード
  - 意味の解析が必要
- DNA配列：A,C,G,Tの4文字からなる文字列
  - 様々な情報技術が利用可能

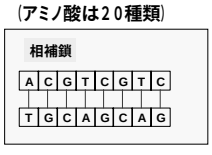
遺伝子と蛋白質

- 遺伝情報の流れ
  - DNA⇒RNA⇒タンパク
- 遺伝子
  - DNA配列中で直接的に機能する部分
- ゲノム
  - 染色体全体（半数体）
  - 遺伝情報の総体
- タンパク質
  - アミノ酸（20種類）の鎖



DNAとアミノ酸

- DNAはA,C,G,Tの4文字の並び
- DNAは二重ラセン構造⇒相補鎖
- 塩基：DNA1文字、残基：アミノ酸1文字
- DNA3文字がアミノ酸1文字に対応
  （アミノ酸は20種類）

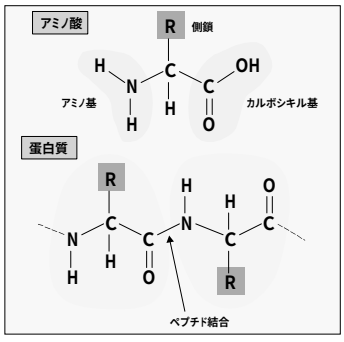


コード表

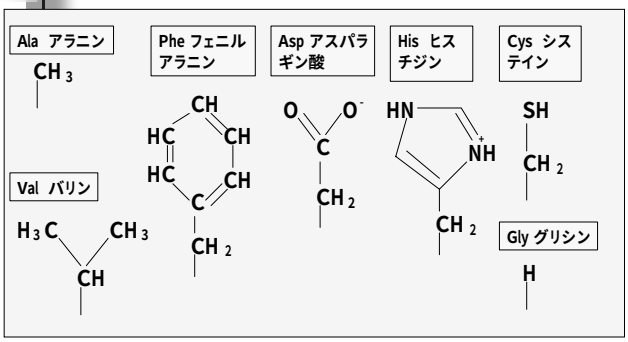
1文字目	2文字目 T	2文字目 C	2文字目 A	2文字目 G
T	TTT F TTC F TTA L TTG L	TCT S TCC S TCA S TCG S	TAT Y TAC Y TAA stop TAG stop	TGT C TGC C TGA stop TGG W
C	CTT L CTC L CTA L CTG L	CCT P CCC P CCA P CCG P	CAT H CAC H CAA Q CAG Q	CGT R CGC R CGA R CGG R
A	ATT I ATC I ATA I ATG M	ACT T ACC T ACA T ACG T	AAT N AAC N AAA K AAG K	AGT S AGC S AGA R AGG R
G	GTT V GTC V GTA V GTG V	GCT A GCC A GCA A GCG A	GAT D GAC D GAA E GAG E	GGT G GGC G GGA G GGG G

アミノ酸と蛋白質

- アミノ酸：20種類
- 蛋白質：アミノ酸の鎖（短いものはペプチドと呼ばれる）
- 側鎖によって性質が異なる．



側鎖の例



アミノ酸コード表

Ala	A	アラニン	Leu	L	ロイシン
Arg	R	アルギニン	Lys	K	リシン
Asn	N	アスパラギン	Met	M	メチオニン
Asp	D	アスパラギン酸	Phe	F	フェニルアラニン
Cys	C	システイン	Pro	P	プロリン
Gln	Q	グルタミン	Ser	S	セリン
Glu	E	グルタミン酸	Thr	T	トレオニン
Gly	G	グリシン	Trp	W	トリプトファン
His	H	ヒスチジン	Tyr	Y	チロシン
Ile	I	イソロイシン	Val	V	バリン

突然変異とアミノ酸置換

- TTT: F(フェニルアラニン)
- TTC: F
- 三番目の文字がC,T間で置換しても，アミノ酸としては変わらない．
- TTA, TTG: L (ロイシン)
- 側鎖の大きさは異なるが，疎水性という性質は違わない．
- 以上から，TT*というコードは，3文字目が置換しても，タンパク質の性質を変えるものではない．
- ATT (I: イソロイシン),
- AAT(N: アスパラギン)では，疎水性と親水性でかなり異なるため，2文字目が置換すると，タンパク質の性質を変える可能性がある．

コード表

1文字目	2文字目 T		2文字目 C		2文字目 A		2文字目 G	
T	TTT	F	TCT	S	TAT	Y	TGT	C
	TTC		TCC		TAC		TGC	
	TTA	L	TCA		TAA	stop	TGA	stop
	TTG		TCG		TAG		TGG	W
C	CTT	L	CCT	P	CAT	H	CGT	R
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	Q	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	I	ACT	T	AAT	N	AGT	S
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	K	AGA	R
	ATG	M	ACG		AAG		AGG	
G	GTT	V	GCT	A	GAT	D	GGT	G
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	E	GGA	
	GTG		GCG		GAG		GGG	

Bioinformaticsにおける確率統計

- 重要なのはデータからのモデル (もしくはパラメータ) の推定
  - 最尤法
  - ベイズ推定
  - 最大事後確率推定 (MAP)
- 通常，データが持っている情報よりも多い量の情報(モデル)を推測することになる
  - このデータを満たしうるモデルが数多く存在しうる．
  - パラメータの推定: モデルの選択 → EM Algorithm

ベイズの定理

- 重要なのはデータからのモデル (もしくはパラメータ) の推定

$P(\theta, D) = P(\theta \mid D)P(D) = P(D \mid \theta)P(\theta)$

θ : パラメータ(モデル), D : データ

$P(\theta, D)$は，データDとモデルのパラメータθが同時に起こりうる確率

ベイズの定理 (2)

- 重要なのはデータからのモデル (もしくはパラメータ) の推定

$P(\theta, D) = P(\theta \mid D)P(D) = P(D \mid \theta)P(\theta)$

$P(D \mid \theta)$: 尤度 : モデルが成立した時のデータが起こる条件付き確率

$P(\theta \mid D)$: 事後確率 : データが観察された時にモデルが成立する条件付き確率

最尤推定

- $P(D|\theta)$ (尤度)
  - モデルパラメータ θ のもとでのデータ D の出現確率
- 最尤法
  - $P(D|\theta)$ を最大化する θ を選ぶ
- 例
  - コインを5回投げて、表が3回出た後、裏が2回出た
  - p(表)=a, p(裏)=1-a とすると$P(D|\theta)={}_5C_3a^3(1-a)^2$
  - a=3/5の時、$P(D|\theta)$ は最大
  - 一般に表が出る頻度を f とすると $a=f$ で尤度は最大

ベイズ推定とMAP推定

- ベイズ推定：尤度とモデル (パラメータ) の事前確率から、ベイズの定理により、事後確率を推定

$$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{P(D)}$$

ただし、$P(D)=\int_{\theta'}P(D|\theta')P(\theta')$ （θが連続値の時）

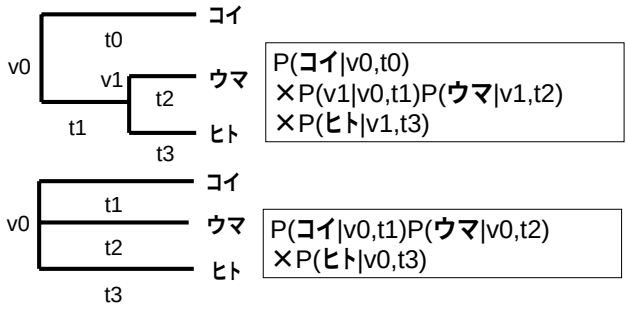
- 最大事後確率 (MAP) 推定
  - $P(D|\theta)P(\theta)$を最大化するθを計算
  - $P(\theta)$が一様分布なら最尤推定と同じ

不正サイコロのベイズ推定

- 公正サイコロと不正サイコロ
  - 公正：$P(i|$公正$)=1/6$
  - 不正：$P(6|$不正$)=1/2, P(i|$不正$)=1/10$ for $i\neq6$
  - $P($公正$)=0.99$, $P($不正$)=0.01$
- 6が3回続けて出た場合の事後確率

$$P(\text{不正}|666)=\frac{P(666|\text{不正})P(\text{不正})}{P(666)}$$
$$=\frac{(0.5)^3(0.01)}{(0.5)^3(0.01)+(\frac{1}{6})^3(0.99)}=0.21$$

モデルから尤度の計算 (2)



推定が必要なパラメータは上が6個，下が4個

確率モデルによる配列解析

- 配列データのレコード数は少ない
  - 例えば，前の例だと3レコード
- レコードが含んでいる項目数は多い
  - ヘモグロビンだと，184個のアミノ酸
- 変数の数は多いが，方程式の数が少ない
  - 不定要素が多い．(一意的に決定できない)
- 確率モデルでのパラメータ推定
  - 不定要素が多いため，当てはまるモデルが多数存在する．
  - 不定要素が多い中でのパラメータ推定

配列解析

① 異種の生物の配列解析: 遺伝子の対応づけの必要性 (アライメント)

② 対応づけが済んだあとの配列の解析

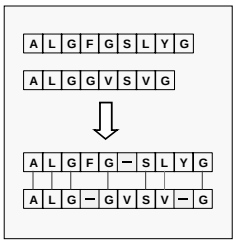
① 系統樹推定

② 隠れマルコフモデル

③ Neural Network 他

アラインメント

- 配列解析の最重要技術
- 2個もしくは3個以上の配列の類似性の判定に利用
- 文字間の最適な対応関係を求める (最適化問題)
- 2個の配列長を同じにするように、ギャップ記号 (挿入、欠失に対応) を挿入



アラインメント

- 遺伝子の欠失, 挿入は進化の過程で起こりうる →しかし, 単に欠失, 挿入だけでは, タンパク質の構造は変わらない.
- 遺伝子の配列から逆に欠失, 挿入の可能性を推定しよう.
- ABCDEF ABCDEF -BCDEF- -BCDEF-
- CDF -- CDE- AB-EEFG ABEEFG
- 文字間の最適な対応関係を求める (最適化問題)
- 2個の配列長を同じにするように、ギャップ記号 (挿入、欠失に対応) を挿入

アラインメント (例:ヘモグロビン)

> human
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
> gorilla
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
> horse
VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKV
KAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARHFGK
DFTPELQASYQKVVAGVANALAHKYH
> chicken
MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM
VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS
KDFTPECQAAWQKLVRVVAHALARKYH
> eastern gray kangaroo
VHLTAEEKNAITSLWGKVAIEQTGGEALGRLLIVYPWTSRFFDHFGDLSNAKAVMANPKV
LAHGAKVLVAFGDAIKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNIIVICLAEHFGK
EFTIDTQVAWQKLVAGVANALAHKYH

アラインメント (例:ヘモグロビン)

```
human
      MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
gorilla
      MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
horse
      -VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPK
eastern
      -VHLTAEEKNAITSLWGKVAIEQTGGEALGRLLIVYPWTSRFFDHFGDLSNAKAVMANPK
chicken
      MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM
       *: :  ***  : .**.**   : *.***.***:*****.***  **.**.. *::.**
```

スコア行列 (置換行列)

- 残基間 (アミノ酸文字間) の類似性を表す行列
  - PAM250, BLOSUM45 など

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A  5 -2 -1 -2 -1 -1 -1  0 -2 -1 -2 -1 -1 -3 -1  1  0 -3 -2  0
R -2  7 -1 -2 -4  1  0 -3  0 -4 -3  3 -2 -3 -3 -1 -1 -3 -1 -3
N -1 -1  7  2 -2  0  0  0  1 -3 -4  0 -2 -4 -2  1  0 -4 -2 -3
D -2 -2  2  8 -4  0  2 -1 -1 -4 -4 -1 -4 -5 -1  0 -1 -5 -3 -4
C -1 -4 -2 -4 13 -3 -3 -3 -3 -2 -2 -3 -2 -2 -4 -1 -1 -5 -3 -1
Q -1  1  0  0 -3  7  2 -2  1 -3 -2  2  0 -4 -1  0 -1 -1 -1 -3
E
G
H
I
L
K
M
F
P
S
T
W
Y
V
```

BLOSUM50 スコア行列 (置換行列) の一部分

ギャップ無しアラインメントのスコア

$$S = \log \frac{P(x,y \mid M)}{P(x,y \mid R)} = \log \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \cdot \prod_i q_{y_i}}$$

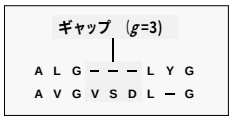
$$= \log \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} = \sum_i s(x_i, y_i) \quad \text{但し}, s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

$x_1\ x_2 \bullet\bullet\bullet x_n$
$y_1\ y_2 \bullet\bullet\bullet y_n$

- スコア: $P(x,y|M)$と$P(x,y|R)$の対数オッズ比
- $P(x,y|R)$: ランダムモデルRにおいて、配列x,yが独立に生起する確率
- q_a:文字aのRにおける出現確率
- $P(x,y|M)$: 一致モデルMにおいて、配列x,yが生起する確率
- p_{ab} :文字ペアa,bのMにおける出現確率

ギャップペナルティ

- 線形スコア
  - $\gamma(g)=-gd$
    (gはギャップの長さ、dは定数)
- アファインギャップスコア
  - $\gamma(g)=-d-e(g-1)$ (d:ギャップ開始ペナルティ, e:ギャップ伸長ペナルティ)
- ギャップの確率的解釈
  - $P(gap)=f(g)\prod q_{xi}$ → $\gamma(g)=\log(f(g))$
  - ギャップ長の確率の対数



ペアワイズ・アラインメント

- 配列が2個の場合でも可能なアラインメントの個数は指数オーダー
- しかし、スコア最大となるアラインメント(最適アラインメント)は動的計画法により、$O(mn)$時間で計算可能 (m,n:入力配列の長さ)

入力配列
AGCT, ACGCT

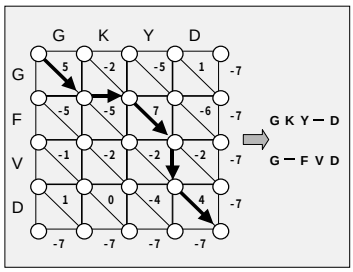
アラインメント

		最適アラインメント		
AGCT -	AG - CT	A - GCT	- AGC - - T	• • •
ACGCT	ACGCT	ACGCT	AC - - GCT	
スコア -3	1	3	-5	

(同じ文字の時1、違う文字の時-1,ギャップ1文字-1)

動的計画法による最適アラインメント

最長経路計算による最適アライメント



DP (動的計画法)による最長経路の計算

$$F(0,j)=-jd,\quad F(i,0)=-id$$
$$F(i,j)=\max\begin{cases}F(i-1,j-1)+s(x_i,y_j)\\F(i-1,j)-d\\F(i,j-1)-d\end{cases}$$

行列からの経路の復元は、$F(m,n)$からmaxで=となっている$F(i,j)$を逆にたどることに行う(トレースバック)

局所アラインメント

- 配列の一部のみに共通部分があることが多い
  ⇒共通部分のみのアラインメント

$$F(i,j)=\max\begin{cases}0\\F(i-1,j-1)+s(x_i,y_j)\\F(i-1,j)-d\\F(i,j-1)-d\end{cases}$$

		H	E	A	W	G	E	H
		0	0	0	0	0	0	0
G		0	0	0	0	0	1	0
A		0	0	0	1	0	0	0
W		0	0	0	0	2	1	0
E		0	0	1	0	1	0	2
D		0	0	0	0	0	0	1

AWGE
AW - E

(スコア:ギャップ-1、置換-1、一致1)

アラインメント (例:ヘモグロビン)

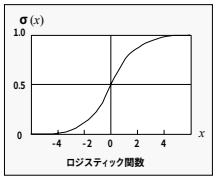
```
human
     MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
gorilla
     MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
horse
     -VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPK
eastern
     -VHLTAEEKNAITSLWGKVAIEQTGGEALGRLLIVYPWTSRFFDHFGDLSNAKAVMANPK
chicken
     MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM
      *: :  ***  : .**.**   : *.***.***:*****.***  **:**.. *::.**
```

スコアのベイズ論的解釈

- $P(M|x,y)$をベイズの定理に基づき導出
- $P(M)$:2個の配列に関連性がある事前確率
- $P(R)=1-P(M)$:ランダムモデルの事前確率
- 以下の式より尤度比にオッズ比を掛け合わせたものを0と比較すべきであることがわかる

$$P(M\mid x,y)=\frac{P(x,y\mid M)P(M)}{P(x,y)}$$
$$=\frac{P(x,y\mid M)P(M)}{P(x,y\mid M)P(M)+P(x,y\mid R)P(R)}$$
$$=\frac{P(x,y\mid M)P(M)/P(x,y\mid R)P(R)}{1+P(x,y\mid M)P(M)/P(x,y\mid R)P(R)}=\sigma(S')$$

但し、$S'=\log\left(\frac{P(x,y\mid M)}{P(x,y\mid R)}\right)+\log\left(\frac{P(M)}{P(R)}\right)$, $\sigma(x)=\frac{e^x}{1+e^x}$

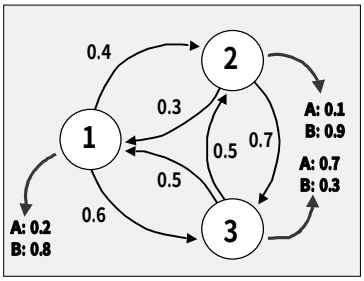


スコア行列の導出

- 基本的には頻度の比の対数をスコアとする
- BLOSUM行列
  - 既存のスコア行列を用いて多くの配列のアラインメントを求め、ギャップ無しの領域（ブロック）を集める
  - 残基がL%以上一致しているものを同一クラスタに集める
  - 同じクラスタ内で残基aが残基bにアラインされる頻度A_{ab}を計算
  - $q_a=\Sigma_b A_{ab} / \Sigma_{cd} A_{cd}$, $p_{ab}=A_{ab} / \Sigma_{cd} A_{cd}$ を求め、$s(a,b)=\log(p_{ab}/q_a q_b)$ としたのち、スケーリングし近傍の整数値に丸める

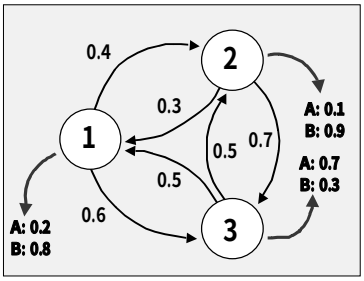
隠れマルコフモデル(HMM)

- HMM=有限オートマトン＋確率
- 定義
  - 出力記号集合 Σ
  - 状態集合 $S=\{1,2,\ldots n\}$
  - 遷移確率($k\rightarrow l$) a_{kl}
  - 出力確率 $e_k(b)$
  - （開始状態=終了状態= 0）

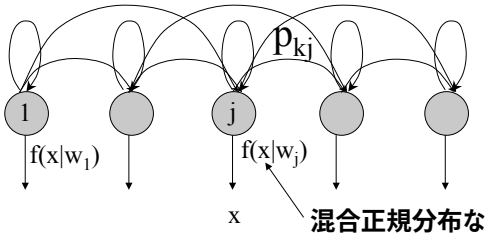


隠れマルコフモデル(HMM) (2)

- 音声認識について革命的な変化をもたらした.
- それまでは認識率は50%も出ればよい方.
- HMMによって，80%以上の正答率が実現した.



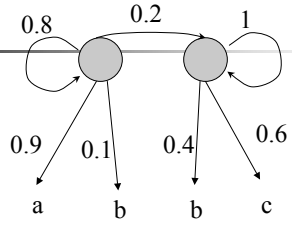
隠れマルコフモデルとは



外部からは状態については不明

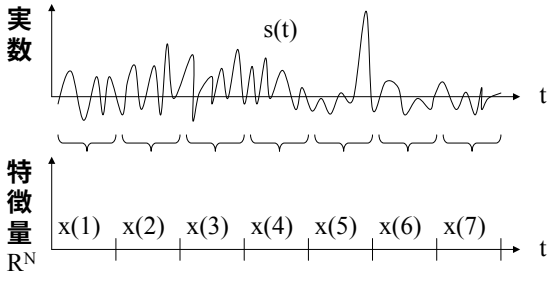
$x_1 x_2 x_3 x_4 \ldots x_t \ldots$ (時系列)

例



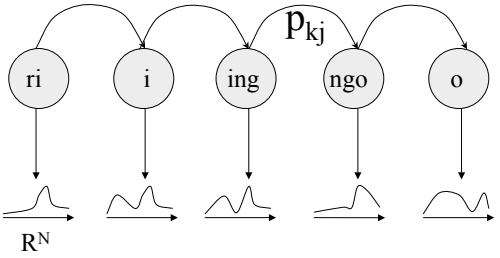
P(aaabccc)
= 0.9*0.8*0.9*0.8*0.9
  * (0.8*0.1 *0.2+0.2*0.4)*0.6*0.6*0.6

使い方（音声認識）



使い方（音声認識）

「りんご」＝「ri i ing ngo o」を発生するモデル



使い方（音声認識）

「りんご」 $p(x_1,x_2,\ldots,x_T \mid p_{kj}, w_j)$

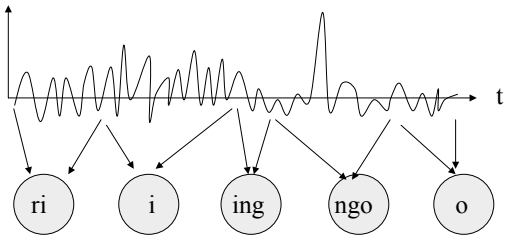
「ろんご」 $q(x_1,x_2,\ldots,x_T \mid p_{kj}, w_j)$

「りんぐ」 $r(x_1,x_2,\ldots,x_T \mid p_{kj}, w_j)$

特徴量を入力して発生確率の高い候補を認識結果として選ぶ。

利点は？

音声は、時間方向に伸縮がおおきい：
変化に対応できる　りーんご、 りんーご、 りーんーごー、



隠れマルコフモデルの意義

形式言語

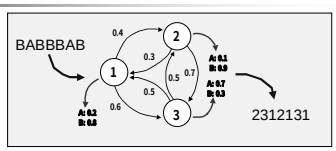
↓

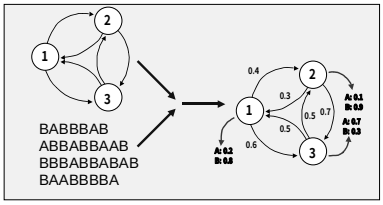
有限オートマトン
確率オートマトン
確率的な情報発生装置

実世界情報

HMMにおける基本アルゴリズム

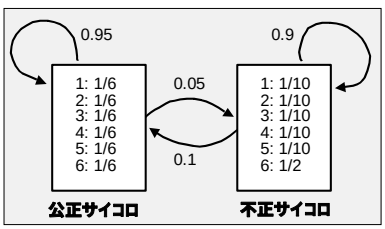
- Viterbiアルゴリズム
  - 出力記号列から状態列を推定
  - Parsing（構文解析）
- Baum-Welchアルゴリズム
  (EMアルゴリズム)
  - 出力記号列からパラメータを推定
  - Learning（学習）





時々いかさまをするカジノ

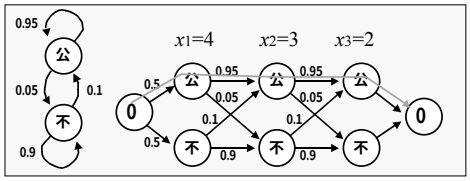
- サイコロの出目だけが観測可能、どちらのサイコロを振っているかは観測不可能
- サイコロの出目から、どちらのサイコロを振っているかを推定
- 6,2,6,6,3,6,6,6,
  4,6,5,3,6,6,1,2
  →不正サイコロ
- 6,1,5,3,2,4,6,3,
  2,2,5,4,1,6,3,4
  →公正サイコロ
- 6,6,3,6,5,6,6,1,
  5,4,2,3,6,1,5,2
  →途中で公正サイコロに交換



Viterbi アルゴリズム(1)

- 観測列（出力配列データ） $x=x_1 \ldots x_L$ と状態列 $\pi=\pi_1 \ldots \pi_L$ が与えられた時、その同時確率は

  $P(x,\pi)=a_{0\pi_1}\prod e_{\pi_i}(x_i)a_{\pi_i\pi_{i+1}}$　　但し、$\pi_{L+1}=0$
- x が与えられた時の、最も尤もらしい状態列は $\pi^*=\mathrm{argmax}_\pi\ P(x,\pi)$
- 例：どちらのサイコロがいつ使われたかを推定



$$\max_\pi P(x_1x_2x_3,\pi)=P(x_1x_2x_3,\text{公公公})=0.5\cdot\tfrac{1}{6}\cdot0.95\cdot\tfrac{1}{6}\cdot0.95\cdot\tfrac{1}{6}$$

Viterbiアルゴリズム(2)

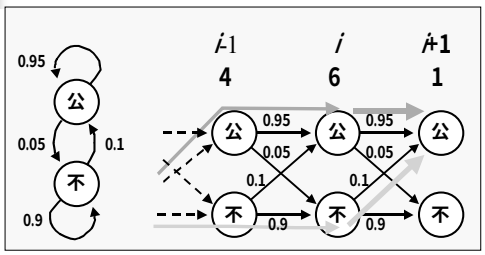
- x から、$\pi^*=\mathrm{argmax}_\pi\ P(x,\pi)$ を計算
- そのためには $x_1 \ldots x_i$ を出力し状態 k に至る確率最大の状態列の確率 $v_k(i)$ を計算

$$v_k(i)=\max_\pi a_{0\pi_i}\prod_{j=1}^{i-1}e_{\pi_j}(x_j)a_{\pi_j\pi_{j+1}}$$

- $v_k(i)$ は以下の式に基づき動的計画法で計算

$$v_l(i+1)=e_l(x_{i+1})\ \max_k(v_k(i)a_{kl})$$

Viterbiアルゴリズム(3)



$$v_{\text{公}}(i+1)=\max\{e_{\text{公}}(1)\cdot0.95\cdot v_{\text{公}}(i),\ e_{\text{公}}(1)\cdot0.1\cdot v_{\text{不}}(i)\}$$

EM(Expectation Maximization) アルゴリズム

- 「欠けているデータ」のある場合の最尤推定のための一般的アルゴリズム

x：観測データ、y：欠けているデータ、
θ：パラメータ集合
目標：$\log P(x|\theta)=\log\sum_y P(x,y|\theta)$ の最大化

- 最大化は困難であるので、反復により尤度を単調増加させる（θ^t より θ^{t+1} を計算）
- HMMの場合、「欠けているデータ」は状態列

EMアルゴリズムの導出

$$\log P(x|\theta)=\log P(x,y|\theta)-\log P(y|x,\theta)$$

両辺に $P(y|x,\theta^t)$ をかけて y についての和をとり、

$$\log P(x|\theta)=\sum_y P(y|x,\theta^t)\log P(x,y|\theta)-\sum_y P(y|x,\theta^t)\log P(y|x,\theta)$$

右辺第1項を $Q(\theta|\theta^t)$ とおくと、

$$\log P(x|\theta)-\log P(x|\theta^t)=$$

$$Q(\theta|\theta^t)-Q(\theta^t|\theta^t)+\sum_y P(y|x,\theta^t)\log\frac{P(y|x,\theta^t)}{P(y|x,\theta)}$$

最後の項は相対エントロピーで常に正なので、

$$\log P(x|\theta)-\log P(x|\theta^t)\geq Q(\theta|\theta^t)-Q(\theta^t|\theta^t)$$

よって、$\arg\max_\theta Q(\theta|\theta^t)$ をみたす θ^{t+1} をとれば尤度は増大

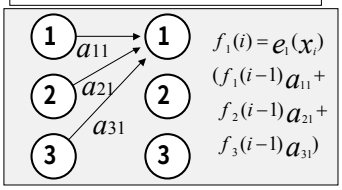
EMアルゴリズムの一般形

1. 初期パラメータ θ^0 を決定。$t=0$ とする。
2. $Q(\theta|\theta^t)=\Sigma P(y|x,\theta^t)\log P(x,y|\theta)$ を計算。(Expectation)
3. $Q(\theta|\theta^t)$ を最大化する θ^* を計算し、$\theta^{t+1}=\theta^*$ とする。$t=t+1$ とする。(Maximization)
4. Q が増大しなくなるまで、2，3を繰り返す。

前向きアルゴリズム

- 配列xの生成確率 $P(x)=\Sigma P(x,\pi)$ を計算
- Viterbiアルゴリズムと類似
- $f_k(i)=P(x_1\ldots x_i,\pi_i=k)$ をDPにより計算

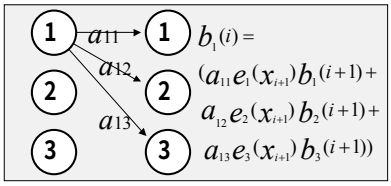
$$f_0(0)=1,\ f_k(0)=0$$
$$f_l(i)=e_l(x_i)\sum_k f_k(i-1)a_{kl}$$
$$P(x)=\sum_k f_k(L)a_{k0}$$



後向きアルゴリズム

- $b_k(i)=P(x_{i+1}\ldots x_L|\pi_i=k)$ をDPにより計算
- $P(\pi_i=k|x)=f_k(i)b_k(i)/P(x)$

$$b_k(L)=a_{k0}$$
$$b_k(i)=\sum_l a_{kl}e_l(x_{i+1})b_k(i+1)$$
$$P(x)=\sum_k a_{0l}e_l(x_1)b_l(1)$$



HMMに対するEMアルゴリズム (Baum-Welchアルゴリズム)

A_{kl}: a_{kl}が使われる回数の期待値　　x^j: j番目の配列

$E_k(b)$: 文字bが状態kから現れる回数の期待値

$$A_{kl}=\sum_j \frac{1}{P(x^j)}\sum_i f_k^j(i)a_{kl}e_l(x_{i+1}^j)b_l^j(i+1)$$
$$E_k(b)=\sum_j \frac{1}{P(x^j)}\sum_{\{i|x_i^j=b\}} f_k^j(i)b_k^j(i)$$

パラメータの更新式

$$\hat{a}_{kl}=\frac{A_{kl}}{\sum_{l'}A_{kl'}}\qquad \hat{e}_k(b)=\frac{E_k(b)}{\sum_{b'}E_k(b')}$$

Baum-WelchのEMによる解釈

$P(x,\pi|\theta)=\prod_{k=1}^{M}\prod_b [e_k(b)]^{E_k(b,\pi)}\prod_{k=0}^{M}\prod_{l=1}^{M}a_{kl}^{A_{kl}(\pi)}$　および

$Q(\theta|\theta^t)=\sum_\pi P(\pi|x,\theta^t)\log P(x,\pi|\theta)$　より、

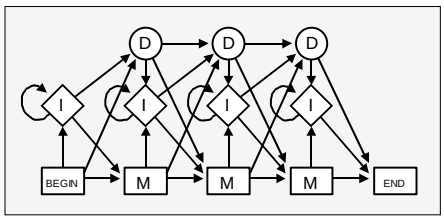
$$Q(\theta|\theta^t)=\sum_\pi P(\pi|x,\theta^t)\times\left[\sum_{k=1}^{M}\sum_b E_k(b,\pi)\log e_k(b)+\sum_{k=0}^{M}\sum_{l=1}^{M}A_{kl}(\pi)\log a_{kl}\right]$$
$$=\sum_{k=1}^{M}\sum_b E_k(b)\log e_k(b)+\sum_{k=0}^{M}\sum_{l=1}^{M}A_{kl}\log a_{kl}$$

ここで$\sum_i p_i\log q_i$は$q_i=p_i$の時、最大より、

$$e_k(b)=\frac{E_k(b)}{\sum_{b'}E_k(b')},\quad a_{kl}=\frac{A_{kl}}{\sum_{l'}A_{kl}}$$

プロファイルHMM(1)

- 配列をアラインメントするためのHMM
- 一致状態(M)、欠失状態(D)、挿入状態(I)を持つ

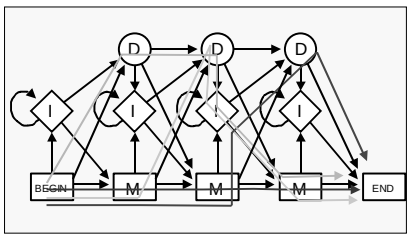


プロファイルHMM(2)

マルチプルアラインメント

	M M ・ ・ ・ M
こうもり	A G ー ー ー C
ラット	A ー A G ー C
ネコ	A G ー A A ー
ハエ	ー ー A A A C
ヤギ	A G ー ー ー C

プロファイルHMM

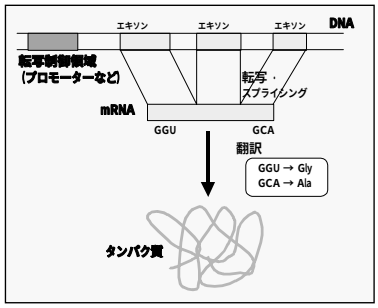


隠れマルコフモデルの適用

- マルチプルアライメント
- エキソン, イントロン, プロモーターの検出
- スプライス部位の検出

- モチーフの検出

(タンパク質の機能単位: 例えば, ヘモグロビンの鉄結合部位)



まとめ

- DNA, アミノ酸配列における突然変異
- 突然変異の蓄積を確率過程としてとらえる
- 適切な確率モデルを構築して:
  - 配列の類似性
  - 配列の規則性 を検出する.
  - そこでは, ベイズ推定, 確率過程が重要と思われる.
- 配列アラインメントとスコア関数
- 隠れマルコフモデル
- EMアルゴリズム