4.1 スコア行列

20 種のアミノ酸、進化の仮定で互いに置換されやすいペアとそうでないペアが存在する。このような非一様性をもたらす要因は、

1. DNA 塩基間の突然変異率の非一様性
2. 遺伝暗号表における非一様性
3. RNA/タンパク質レベルでの淘汰の非一様性

である。 もっとも簡単なスコア行列は、同一塩基/アミノ酸には 1 それ以外は 0 を値としてもつ単位置換行列 (Unitary substitution matrix/Identity matrix) である。 また遺伝暗号表における非一様性をとりいれたスコア行列としては、

$GCM_{ab} = 3-$ アミノ酸置換に最低必要な塩基置換の数

で定義される遺伝暗号行列 (Genetic code matrix) がある。しかし、淘汰圧は DNA 配列よりアミノ酸配列に作用するので、これらはアミノ酸間の分子進化的類似度として。

4.1.1 アミノ酸置換確率行列

実際の相同タンパク質において生じたアミノ酸置換の詳細なデータは、最初に Dayhoff によってまとめられた。Dayhoff et al.[3] 等は、アミノ酸置換をマルコフ過程としてモデル化し、使用した相同タンパク質におけるアミノ酸置換は定常状態にあり詳細均合が成立していると仮定、以下の手順でアミノ酸置換行列を計算した。

1. 相同タンパク質の系統樹を最大節約法により作成し, 各ノードに相当する祖先配列を推定する。
2. 多重置換が無視できるよう十分近縁なに生じた置換置換数行列 A を作成する。アミノ酸タイプ a と b 間の置換は a から b と b から a への 1/2 回づつの置換として A_{ab} と A_{ba} にカウントする。
3. 平均して 1 %の置換をもたらす進化時間 $t = 1PAM$ に対応するタイプ a から b への置換確率行列 $M_{ab}(t = 1PAM)$ を、A_{ab} とアミノ酸頻度 f_a から以下のように計算した。

$$M_{ab}(t) \quad = \quad [\frac{A_{ab}}{\sum_b A_{ab}}]^t, \qquad A_{aa} = [100 \sum_a \sum_{b(\neq a)} A_{ab}]f_a - \sum_{b(\neq a)} A_{ab} \tag{1}$$

また、アミノ酸置換における非一様性を示す指標として, 無作為置換における置換頻度との比の対数、つまり対数尤度比 (log-likelihood-ratio) を定義し、対数オッズスコア行列 () とした。対数オッズは進化時間に依存するが、$t = 250PAM$ に対応する対数オッズ、特に 1978 年に発表された 1572 個の置換から計算された $MDM78$ は現在もよく使用される。 Dayhoff et al.[3] は対数オッズの値から、物理化学的性質の類似したアミノ酸間ではより置換が生じ易いことを指摘した。

MDM78 はその後の配列データの増大にも係わらず長らく更新されなかったが、1992 年 Jones 等 [7] は多数の近縁配列から得られた 59190 個の置換に基づき、同様の方法で対数オッズスコア行列 () を計算した。低頻度のアミノ酸で MDM78 ものの違いは見られなかった、

ミトコンドリア DNA は、遺伝暗号が核 DNA と異なる。Adachi & Hasegawa[1] は, ミトコンドリア DNA にコードされているタンパク質のアミノ置換確率行列を系統樹の枝長とともに最尤法に基づき推定した。1PAM のアミノ置換確率行列を MDM78JTT のと比較すると、遺伝暗号の違いに起因すると思われる違いがれた。

置換確率行列はタンパク質進化におけるアミノ酸置換をモデル化したものであり、最尤法による系統樹作成においては、系統樹の尤度を計算するために不可欠なデータである。一方、このようにして計算された置換確率行列は比較的短い進化時間におけるアミノ酸置換の様相を記述するには適しているものの、250PAM のような進化時間にまで外挿した時にも良い近似であるかは [14]。

4.1.2 遠縁な配列の間のアミノ酸置換

遠縁の配列を用いてアミノ酸置換行列を計算する試みは多数あるが、その中で BLOSUM 行列 [6] が最も広く使用されている。BLOSUM 行列は、多数の相同タンパク質からなるローカルアライメントをブロックと定義し、PROSITE データベースに登録された保存部位を含むブロックの全ての配列対における各座位のアミノ酸対の頻度から対数オッズスコア行列として計算される。アミノ酸対 a と b は 1/2 回の置換として置換数行列要素 n_{ab} と n_{ba} にカウントする。ただし、類似配列による偏りを避けるため、クラスター間で配列の一致度が閾値 60%以下になるようにクラスタリングし、クラスター内の配列からの寄与は同等としクラスターからの全寄与が 1 となるよう重み付けして置換数行列を計算する。

$$BLOSUM_{ab} \equiv \frac{2}{\log 2} \log \frac{p_{ab}}{\sum_c p_{ac} \sum_d p_{db}}, \qquad p_{ab} = \frac{n_{ab}}{\sum_c \sum_d n_{cd}} \tag{2}$$

ブロックデータベースを得るにもが必要、として単位置換行列を用い、反復してブロックデータベースを作成し自己無撞着な BLOSUM 行列を計算した。最後のクラスタリングにおける配列一致度の閾値の値により 250PAM 相当の BLOSUM45, 160PAM 相当の BLOSUM62, 120PAM 相当の BLOSUM80 等がある。

4.1.3 アミノ酸置換とアミノ酸指標

Dayhoff 等 [3] が指摘したように、配列比較から得られたアミノ酸置換行列はいずれも、物理化学的性質の類似したアミノ酸間ではより置換が生じ易いことを示す。Grantham[5] は体積, 親水性と組成からなる物理化学的指標と置換頻度の間に高い相関がみられることを定量的に示した。また Miyata et al.[11] は, 規格化された体積, 親水性座標空間でのユークリッド距離を指標として用い、指標の関数としてオッズ () をした。、そのを用いて、相同タンパク質におけるアミノ酸置換が物理化学的な性質を保存する傾向にあることは、タンパク質の立体構造を保持するような淘汰圧によるものであり、分子進化における中立説を支持する証拠の一つであることを示した。更にこれらの解析に基づき、体積・親水性指標の 1

次関数として置換の受容確率を定めアミノ酸の同義置換頻度と非同義置換頻度を配列比較から推定する方法 [12] を考案し、分子進化学における基本的な手法としてその有用性を示した。

4.1.4 アミノ酸指標とスコア行列

データベース [17] 化されているアミノ酸の各種指標を用いてアミノ酸置換のスコア行列を定義する試みはこれまで多数なされている。例えば最も簡単な方法では、次節で述べるスコアの条件を満たすよう

$$s_{ab} \quad = \quad I_{ab} - \sum_a \sum_b p_a p'_b I_{ab} +$$

としてアミノ酸指標 I_{ab} 比較する配列におけるアミノ酸の頻度 p_a, p'_b からスコア s_{ab} を定義する。いづれも遠縁な配列の間のアミノ酸置換スコアとしてしかし、アミノ酸指標が (スコアであるための条件である) 対数オッズに比例するとは限らない。

一方、 Miyazawa et al.[13] は, アミノ酸置換の適応度をタンパク質構造の安定性と見做し、アミノ酸 a と b 間の置換による適応度の低下 (δw_{ab}) を置換による構造エネルギーの増加の期待値 (δe_{ab}) から $\delta w_{ab} = \exp(-\delta e_{ab})$ と仮定し, コドン置換率行列 $R_{\alpha\beta}$ をコドン置換率 $m_{\alpha\beta}$ と適応度の積により評価した;

$$R_{\alpha\beta} \quad = \quad m_{\alpha\beta} \sum_{a,b} c_{\alpha a} c_{\beta b} w_{ab} \ ,$$

遺伝暗号表 $c_{\alpha a}$ はコドン α がアミノ酸 a に対応する場合は 1, 対応しなければ 0 の値をとる。置換による構造エネルギーの増加 (δe_{ab}) は残基間統計ポテンシャルにより見積もられた。このようにして計算された置換確率行列は Dayhoff 等の行列と相関を持ち、また対数オッズ行列は MDM78 に匹敵する類似配列検索能力をこの試みは、タンパク質の分子進化における致死的/中立的なアミノ酸置換の置換確率を、物理化学的の方法で定量的に説明しようとする試みと解釈することができる。

4.1.5 スコアの条件

まず、配列間で局所的に類似した領域のアライメントすなわちローカルアライメントの計算におけるスコアを考えよう。Smith & Waterman[15] は、正のスコア値を持つ領域を類似領域と定義し、最大のスコア値を持つ領域をもとめるアルゴリズムを提案した。意味あるローカルアライメントを得るには、スコアは以下の条件を満たさねばならない [8, 2]。

1.

2. $\sum_a \sum_b p_a p'_b s_{ab} < 0$

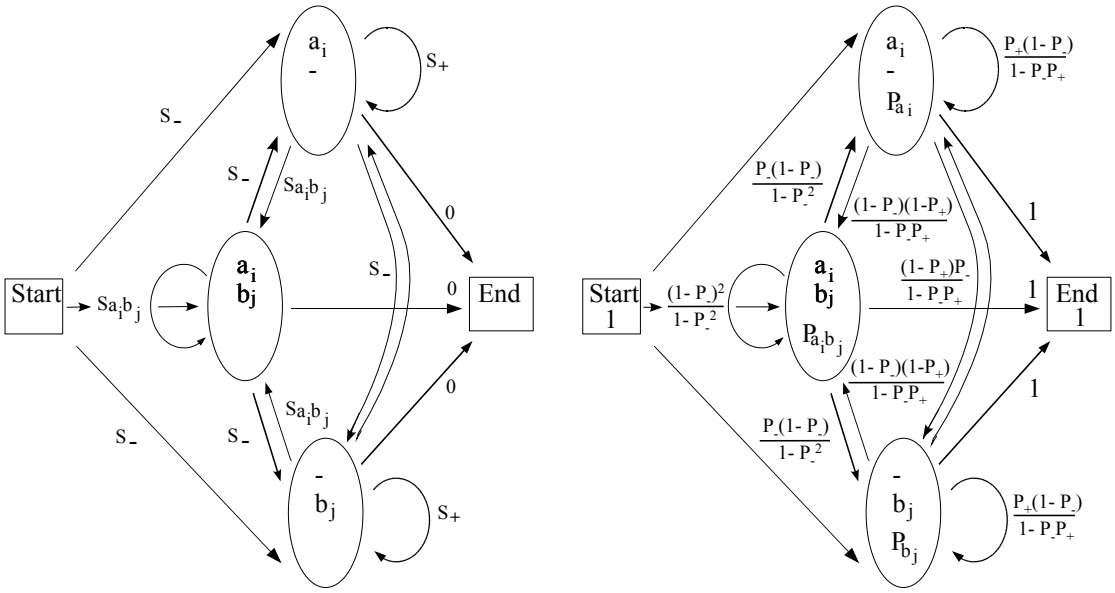
Karlin & Altschul[8, 2] は、ランダム配列の比較において、スコア最大のギャップ無しローカルアライメントにおける $a : b$ 対の対確率は以下の関係式を満たす p_{ab} に収束することを示した。

$$s_{ab} \quad = \quad \frac{1}{\lambda} \log \frac{p_{ab}}{p_a p'_b}, \qquad p_a = \sum_b p_{ab}, \qquad p'_b = \sum_a p_{ab}, \qquad \lambda > 0 \tag{3}$$

上式は、ローカルアライメント用のスコアは、目標対確率 () とランダム配列で期待される対確率 () の比の対数, つまり対数尤度比/対数オッズに比例するよう定義すべきであることを示している。ここで、スコアに適当な正実数 λ をかけてもギャップ無しローカルアライメントのスコア値も λ 倍されるだけでその相対的な関係は同じである、上式不定

全領域にわたる配列アライメントであるグローバルアライメント挿入、欠失、置換のみを考えその進化モデルを θ としよう。この進化モデルにおいて、長さ m と n の配列 $\mathbf{a}$ と $\mathbf{b}$ がアライメント A_l において祖先配列から進化した確率を $P(\mathbf{a}, \mathbf{b}, A_l|\theta)$ とする。この進化モデルと帰無仮説である無作為抽出のモデルにおける確率尤度の対数尤度比を考える。無作為抽出により配列 $\mathbf{a}$ と $\mathbf{b}$ が得られる確率は、個々の配列を得る確率の積 $P(\mathbf{a})P(\mathbf{b})P(\mathbf{a})$ は配列各座位のアミノ酸タイプの頻度 p_{a_i} の全長にわたる積に等しい。進化モデルとして隠れマルコフモデル [4] 図 4.1p_- はアミノ酸の欠失の確率, p_+ は欠失延長のである。 Needleman-Wunsch 法におけるギャップスコアはアフィンギャップペナルティーとし長さ k 個のギャップのスコアはとする。$S(..., A_l, ...)$ アライメント A_l のスコアとすると以下の式が成立する。 結局、式 (3) を満たすスコア このような確率モデルに基づく考察 ちなみに、パラメーター p_{ab}, p_-, p_+ の値最尤法により $\sum_{A_l} P(\mathbf{a}, \mathbf{b}, A_l|p_{ab}, p_-, p_+)$ を最大化するように決定する [16]。

以上、あらゆる配列にあてはまる一般的なスコアを考察してきたが、位置特異的スコア行列 (position specific sore matrix) のように特定の配列から計算されるスコアもある。これは配列プロファイルやギャップも考慮した隠れマルコフモデルによるプロファイルの定義/検索に使用される。



(a) (b)

図 4.1 一般アフィンギャップペナルティーの特別な場合に相当する (a) Needleman-Wunsch 法, (b) 隠れマルコフモデルにおける状態遷移図

References

[1] Adachi, J. and Hasegawa, M. (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA . *J. Mol. Evol.* 42:459-468.

[2] Altschul, S. F. (1991) Amino acid substitution matrices from an information theoretic perspective . *J. Mol. Biol.* 219:555-565.

[3] Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (Dayhoff, M. O. (ed.)) A model of evolutionary change in proteins , *Atlas of protein sequence and structure* 1978, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington D.C. 3453521978

[4] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison (eds.)) , Biological sequence analysis.Cambridge University Press 1998

[5] Grantham, R. (1974) Amino acid difference formula to help explain protein evolution . *Science* 185:862-864.

[6] Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks . *Proc. Natl. Acad. Sci. USA* 89:10915-10919.

[7] Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences . *CABIOS* 8:275-282.

[8] Karlin, S. and Altsul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes . *Proc. Natl. Acad. Sci. USA* 87:2264-2268.

[9] Kosiol, C., Holmes, I. and Goldman, N. (2007) An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.* 24:1464-1479.

[10] Le, S. Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25:1307-1320.

[11] Miyata, T., Miyazawa, S. and Yasunaga, T. (1979) Two types of amino acdi substitutions in protein evolution . *J. Mol. Evol.* 12:219-236.

[12] Miyata, T. and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nuleotide sequences and its application . *J. Mol. Evol.* 16:23-36.

[13] Miyazawa, S. and Jernigan, R. L. (1993) A new substitution matrix for protein sequence searches based on contact frequencies in protein structures . *Protein Eng.* 6:267-278.

[14] Russel, R. B., Saqi, M. A. S., Sayle, R. A., Bates, P. A. and Sternberg, M. J. E. (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* 269:423-439.

[15] Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences . *J. Mol. Biol.* 147:195-197.

[16] Thorne, J. L., Kishino, H. and Felsenstein, J. (1992) Inching toward reality: an improved likelihood model of sequence evolution . *J. Mol. Biol.* 34:3-16.

[17] Tomii, K. and Kanehisa, M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins . *Protein Eng.* 9:27-36.

[18] Whelan, S. and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein fam ilies using a maximum-likelihood approach. *Mol. Biol. Evol.* 18:691-699.