生物情報学: 生物学と情報学の出会い

(DNA/タンパク質配列解析のための確率モデル)

工学研究科　　宮澤 三造

TEL: 0277-30-1940, E-mail: miyazawa@smlab.cs.gunma-u.ac.jp

1. はじめに

ここ4半世紀における生物学分野における分子レベルでの解析技術の進歩は著しく、各種生物の遺伝情報の解析が以前とは桁違いに容易になった。生物の遺伝情報は、A,T,C,Gの4種類の塩基が鎖状に結合したDNAという分子における塩基の結合順序に暗号化されている。(生殖)細胞に含まれる遺伝子の総体をゲノムと言うが、ヒトだけでなくチンパンジー、マウスをはじめ数多くの生物種のゲノム塩基配列が既に解析され、現在までに解析された全ての生物種におけるDNAの総塩基長は，約30億塩基からなるヒトのDNAの約30倍、約1000億塩基にも上る。このような生物情報の爆発的増加が、その管理に情報学の手法を必要とする原因となった。現在では、管理だけでなく解析においても情報学の手法が必須となっている。

ヒトのゲノムを例にとると、タンパク質をコードする遺伝子配列は、総塩基数の4%程度にすぎず、30億塩基からなる配列上に約5万の遺伝子が散在して分布している。遺伝子の発現を調節している配列部位の同定、遺伝子位置の予測に塩基配列の統計的特長に基づく情報学的手法が用いられている。

一方、進化的に類縁な塩基/アミノ酸配列はその配列順序が類似していることから、類似な配列を塩基/アミノ酸配列データベース上で検索し、もし類似配列が発見されれば、その類似配列と同じ機能を所持していると推測できる。また、配列間の類似性に基づきDNA塩基配列と生物種間に生じた進化の道筋を明らかにすることも可能である。このように、類似配列のデータベース検索、類似性に基づく分子系統樹作成は、生物情報学の基盤をなす解析手法である。

さて、アミノ酸が鎖状に結合した高分子であるタンパク質は、生理的条件下では、各タンパク質に固有の立体構造に折り畳まれる。タンパク質は小分子や他のタンパク質と結合して機能し、タンパク質の立体構造と機能は不可分な関係にある。幸いなことに地球上の生物種に含まれるタンパク質の構造の基本型は数千であることが予想されている。そこで全ての基本型を明らかにすべく解析が世界中で進行中で、最近では1日あたり20以上ものタンパク質構造が報告されている。しかし、配列解析に比較すると時間と費用がかかるので、配列からの構造予測が期待されている。タンパク質は、温度の上下により、可逆的に立体構造の崩壊と折り畳みが生じるので、アミノ酸の配列順序が、物理化学法則の下、立体構造を規定していることは明らかである。それ故、配列からの構造予測は、半世紀前より研究されてきたが、最近になってようやく、既知の立

体構造データの解析に基づく帰納的方法により、かなりの精度で予測可能になりつつある。

ここでは、DNA/タンパク質配列解析で用いられている情報学的手法として代表的な確率モデルを簡単に紹介する。

2. DNA/タンパク質配列解析のための確率モデル

モデルMの下、データDが得られた時、仮説Hである条件付確率（事後確率）$P(H|D,M)$は

$$P(H|D,M) = P(D|H,M)P(H|M)/\sum_H P(D|H,M)P(H|M),$$
$$\hat{H} = \arg\max_H P(H|D,M) = \arg\max_H P(D|H,M)P(H|M),$$

である(Bayesの定理)。ここで、$P(H|M)$を事前確率という。最も確からしい仮説Hをもって仮説の推定$\hat{H}$とする。表に配列解析の基本手法におけるDとHを挙げた。

失するものとする。最良アライメント/分子系統樹推定におけるモデルパラメータであるコドン/アミノ酸間の置換率は、推定の質を向上させる上で重要で、各種の手法で評価されている。一方、タンパク質配列**a**を構造**C**に埋め込む場合は、アミノ酸間相互作用がモデルパラメータとなる。タンパク質の骨格構造の予測でも、配列-構造の適合性評価のため、相互作用ポテンシャルが肝要である。適合性評価のためのポテンシャルは、(1) 水との相互作用を含む、(2) 主鎖構造のみで評価、(3) 立体構造の安定性(変性状態との自由エネルギー差)を評価するものであり、現在知られている経験的原子間相互作用ポテンシャルよりも、粗い粒度の統計/知識ベースのポテンシャルが既知のタンパク質構造における原子/残基間距離分布を用いて評価され用いられている。

	D	H	M	モデルパラメーター(θ_M)
配列アライメント	$\{\boldsymbol{a}\}$	A	T, θ_M	置換率に基づく塩基/アミノ酸間類似度
分子系統樹推定	T	$\{\boldsymbol{a}\}, A$	θ_M	コドン/アミノ酸間置換率
タンパク質配列ー構造アライメント	$\boldsymbol{c}$	A	$\boldsymbol{a}, \theta_M$	アミノ酸間相互作用

$\boldsymbol{a}$: 配列, $\{\boldsymbol{a}\}$: 配列セット, A: アライメント, T: 系統樹, $\boldsymbol{c}$: タンパク質構造

アライメントAとは、複数の配列もしくは構造$(\alpha, \beta, \ldots)$の各要素の順番を変更することなく、個々の配列/構造要素間に任意に欠失(ー)を挿入し、整列させたものである。

$$A \equiv \begin{bmatrix} \cdots & \alpha_{i-1} & \alpha_i & - & - & \alpha_{i+1} & \cdots \\ \cdots & - & \beta_j & \beta_{j+1} & \beta_{j+2} & \beta_{j+3} & \cdots \end{bmatrix}.$$

配列が同一祖先に由来する時、突然変異により塩基/アミノ酸が置換もしくは挿入/欠

3. 終りに

タンパク質はDNA(遺伝子)と相互作用することで遺伝子発現を制御し、また互いに相互作用し生体反応を制御している。ポストゲノム時代を迎え、生物情報学は各種の情報を統合化してシステムとして生物を理解する方向へと発展しつつある。