

## タンパク質の立体構造に基づく相互作用構造の推定

由良 敬<sup>a,b</sup>**A Trial to Predict Interactions between Proteins and Biomolecules Based on Their Three-dimensional Structures**Kei YURA<sup>a,b</sup>*<sup>a</sup>Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan, and <sup>b</sup>CREST, JST*

(Received June 23, 2008)

A vast amount of DNA sequence data, protein three-dimensional (3D) structure data, and RNA expression data have been produced by the efforts of genome sequencing, structural genomics, and omics projects, and we are at the stage where comprehensive views of cell activity and molecular mechanisms of life can be deduced. But in reality, we are inundated with massive amounts of data and are still in the process of finding ways to fully utilize the data. In this report, I would like to present our observations on the growth of protein 3D structure data and our effort to deduce the functions from the 3D structures. We found that the 3D structure of quite a high proportion of proteins derived from genome sequences can be now predicted and methods to predict the functions from 3D structures are in high demand. The methods we have developed can be used to predict some functions, namely RNA and ligand interfaces, based on those 3D structures and DNA sequences with relatively high accuracy. The methods enable predictions that are accurate enough to help with deducing the atomic structures of the complexes.

**Key words**—computational biology; homology modeling; prediction; protein-ligand interaction; protein-ribonucleic acid interaction; protein three-dimensional structure

**1. はじめに**

ヒトゲノム塩基配列の読み取り完了により、ヒトに関するすべての生命情報が手に入ったかのように思われている。2008年4月には、DNA二重らせん構造を見出したJames D. Watson個人の全塩基配列が読み取られた。<sup>1)</sup>非常に近い将来には、個人のゲノムが千ドルで読み取られる日が訪れると言われている。ゲノム全塩基配列が個体の設計図であることから、このことは本格的なテーラーメイド医療や遺伝子診断の幕開けが近いことを意味する。Landerら<sup>2)</sup>が2001年に報告したヒトゲノム全塩基配列と個人ゲノムを比較すれば、その人の塩基配列のどの部分が平均的な配列から異なっているかが明らかになり、そのことが病気などの原因になってい

る可能性があるからである。Watsonの全塩基配列には少なくとも20個の変異がみついている。ところが、この変異が個体にどのような結果を及ぼすのかを知るすべがない。<sup>3)</sup>ゲノム塩基配列からの情報抽出はこれからの問題である。つまりヒトゲノム塩基配列の「解読」はやっと始まったばかりと言える。<sup>3)</sup>情報抽出の研究を進めることと、そのための人材養成が急がれる所以である。

情報抽出が難しいのは、DNAの塩基配列が直接生物個体を構築している訳ではないことに由来する。DNA塩基配列は個体を構築するタンパク質のアミノ酸配列情報とタンパク質及びRNAの発現制御情報を持っているに過ぎない。よってDNA塩基配列から、タンパク質とRNAにまつわる情報を抽出しなければならぬが、これが想像以上に難しいことがゲノム全塩基配列の決定により明らかとなった。ゲノムのどの部分が転写されるのか？転写翻訳されたタンパク質はどのような機能を果たすのか？タンパク質やRNAはどのようにして機能するの

<sup>a</sup>お茶の水女子大学大学院人間文化創成科学研究科 (〒112-8610 東京都文京区大塚 2-1-1), <sup>b</sup>科学技術振興機構・CREST

e-mail: yura.kei@ocha.ac.jp

本総説は、日本薬学会第128年会シンポジウムS12で発表したものを中心に記述したものである。

か？ これらの質問に答えるために、様々な細胞の状態での全転写物の測定や、全タンパク質の立体構造決定が行われることになった。しかしゲノム塩基配列とこれらのデータを組み合わせることで、「解読」ができるわけでもない。データ間の関係を明らかにし、データに基づく様々な推論とその検証、及び正しい推論に基づくデータの拡張によって、ゲノムの「解読」がやっと始まる。ここに計算生物学の役割がある。本稿では、われわれが展開している計算生物学、特にタンパク質の立体構造情報を中心とした計算生物学を紹介する。

## 2. タンパク質立体構造情報はどこまで分かったか

タンパク質は立体構造を形成して機能する。タンパク質を構成するアミノ酸残基が、適確な空間配置を取ることによって、他の分子と特異的な相互作用が可能となり、生物機能を発揮する。よって、タンパク質の機能を理解するためには、立体構造情報が不可欠である。それでは現在われわれはどれだけのタンパク質立体構造情報を持っており、いつになれば、全タンパク質の立体構造が分かるのだろうか？

すべてのタンパク質立体構造座標をアミノ酸配列情報のみから理論的に得ることは、いまだに難しい。タンパク質の立体構造予測は、まだほんの数例でしか成功していない。<sup>4)</sup>そこで、立体構造座標を得るためのタンパク質構造測定プロジェクトが、世界の多くのタンパク質構造決定研究者を巻き込んで進んでいる。しかし、このプロジェクトが目指していることは、全タンパク質の立体構造を実験的に測定することではない。共通祖先由来のタンパク質は立体構造が類似しており、この拘束条件の基では、タンパク質の構造予測はかなりの精度で行えるようになってきている。ホモロジーモデリングと呼ばれるこの手法は、構造の分かっているタンパク質を「鋳型」として、共通祖先由来タンパク質の立体構造を構築する。共通祖先由来のタンパク質とはいえ、アミノ酸配列はずいぶん変化しており、立体構造の細部もずいぶん異なっている。そこでコンピュータの中で、「熱くした」アミノ酸配列を「鋳型」に「流し込み」、「冷やしながら」構造を求めていく。共通祖先のタンパク質であっても、鋳型タンパク質と標的タンパク質のアミノ酸配列一致度が低ければ低いほど、ホモロジーモデリングは困難になる。つ

まり共通祖先由来のタンパク質であってもホモロジーモデリングが正確にできない場合もある。アミノ酸配列一致度 30%前後が、現在の限界と言われている。1つのタンパク質の立体構造を実験的に決定することができたときに、そのタンパク質と配列全体に渡ってアミノ酸配列の一致度が 30%以上のタンパク質は、立体構造を推定できる。計算生物学の手法で、測定データを増幅できることを意味する。よって、タンパク質の全立体構造情報を得るために、すべてのタンパク質の立体構造を決定する必要はない。ある試算に基づく、約 16000 個の独立なタンパク質立体構造を実験的に決定すれば、全タンパク質の立体構造座標情報を手に入れることができる。<sup>5)</sup>

われわれは、2003 年末のデータを使って立体構造既知率を測定した。<sup>6,7)</sup>2003 年当時に立体構造が判明していたタンパク質の全アミノ酸配列と、ゲノム塩基配列から推定される全アミノ酸配列の類似性検索を行うことで、2003 年時点で立体構造が分かるタンパク質の割合が判明する。その結果、例えば大腸菌の全タンパク質の 56%は、立体構造情報を(部分的にでも)得ることができ、ゼブラフィッシュの既知全タンパク質の 61%は、立体構造情報を(部分的にでも)得られることが判明した。ゲノム塩基配列が分かっている全生物での解析結果を平均すると、約 50%のタンパク質において立体構造情報を得られることが分かった。2008 年現在は 60%を超えているとわれわれは推定している。この数字が意味していることは、立体構造決定プロジェクトによって、タンパク質の機能研究のあり方が変わろうとしていることである。今までは、ある特定のタンパク質の機能を知る際には、立体構造を知らずに実験を進めていくのがふつうであったが、これからは立体構造を見ながら実験を設計することが普通になっていくであろう。



由良 敬

1966 年ロンドン生まれ。早稲田大学理工学部応用物理学科卒業。名古屋大学大学院理学研究科生物学専攻博士課程修了。1993 年名古屋大学理学部助手、2002 年日本原子力研究所計算科学推進センター研究員、2008 年お茶の水女子大学大学院教授。研究テーマは転写後翻訳前の生体高分子動態を計算生物学で解明すること。

われわれは上記の計算を 1999 年から行ってきた。<sup>8)</sup> その結果、立体構造が既知になっていくタンパク質の割合の年次変化を知ることができ、年次変化率一定を仮定すると、すべてのタンパク質の立体構造がいつの時点で明らかになるかが推定できる。ゲノム塩基配列から推定されるアミノ酸配列の中には、明らかに構造を取らない部位（同一アミノ酸残基の長い繰り返し部位など）が含まれている。これらを除くと、真性細菌の水溶性タンパク質の全立体構造は、2018 年には判明することが推定できた。古細菌の全水溶性タンパク質の立体構造は、2021 年に、真核生物の全水溶性タンパク質は 2031 年には、判明することになる (Fig. 1)。タンパク質の構造決定技術の躍進、及びホモロジーモデリングの精度向上により、これらの推定年代は早まることは間違いない。タンパク質の立体構造が分かっていることが当たり前になる時代は目の前に迫っている。

それではタンパク質の立体構造が判明すれば、そのタンパク質の生物機能がすぐに分かるのであろうか？ 残念ながらそうはいかない。タンパク質立体構造からそのタンパク質の機能、又は機能部位を推

定する方法は、現在も研究が盛んに進められている分野である。

### 3. タンパク質と RNA の相互作用推定

ゲノム塩基配列読み取り完了後、かなりの数の RNA 分子が細胞内でなんらかの役割を担っていることが分かってきている。ヒトのトランスクリプトーム解析によると、21037 座位由来の完全長 cDNA のうち、約 7% に当たる 1377 種は RNA 分子として機能している。<sup>9)</sup> これらの RNA 分子は、決して単独で機能している訳ではなく、様々な他分子、特にタンパク質と相互作用して機能するらしい。ヒトゲノムには RNA と相互作用するタンパク質が、少なくとも 1500 個コードされているという推定もあり、<sup>10)</sup> これらが上記の RNA 分子とともに活躍しているに違いない。

タンパク質と RNA がどのように相互作用し、どのようにして機能を果たしているかを理解する基礎データはタンパク質と RNA の複合体立体構造である。複合体の立体構造解析は着実に進んでいるが、解析しなければならない数とその困難さを考えると、計算生物学による複合体構造の高精度な予測は不可欠になる。先に記したように、タンパク質の立

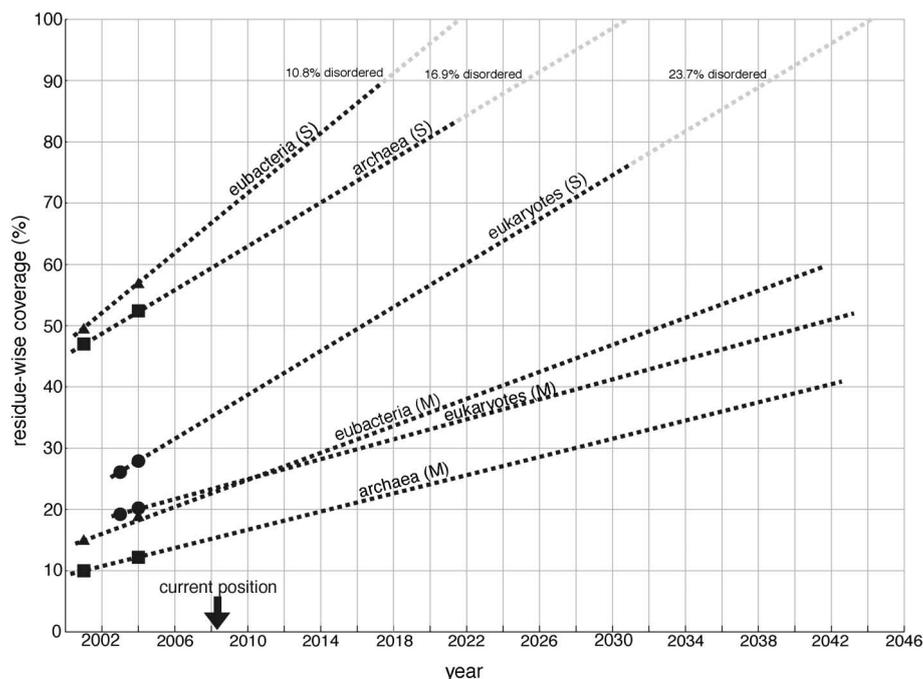


Fig. 1. Coverage of Whole Amino Acid Sequences Derived from Genome Sequences by Protein 3D Structures.

The horizontal axis is a year of calculation and the vertical axis is the coverage percentage. The calculation by the real data was done in the beginning of years 2000, 2003, and 2004, and they are indicated by solid marks. Dotted lines are extrapolation based on the result of two calculations in each category of species. (S) stands for soluble proteins and (M) stands for membrane proteins. Grey dotted lines indicate a predicted ratio of disordered region in whole proteins in each category.

体構造は近いうちにすべてが分かってしまう。よって設問は、タンパク質の立体構造が与えられたときに、どこにどのように RNA が相互作用するかを推定すること、つまりタンパク質と RNA の複合体座標を算出することである。われわれはこの設問に答えられるために研究を進めている。

タンパク質の立体構造が分かっているときに、RNA との界面を予測することは、簡単な問題と考えがちである。RNA は負電荷を帯びた生体高分子であることから、タンパク質表面の正電荷が集中している部分面が RNA と相互作用する面であると考えられる。この考え方に誤りはない。しかし RNA と結合するタンパク質の表面には正電荷が集中した部分面が散在する。またタンパク質表面の電位を正確に計算するのは、想像よりもずっと困難なことである。昨今のタンパク質立体構造解析結果を発表している論文には、当たり前のように表面電位の計算結果を示したきれいな図が掲載されているが、それらの計算においては、誘電率とイオン強度をある値に仮定している。タンパク質が置かれた環境によるこれらの値の違いが、結果にもたらす影響は小さくない。<sup>11,12)</sup> また先に示したように、多くのタンパク質立体構造はホモロジーモデリングで明らかになる。現在のホモロジーモデリングの精度では、アミノ酸残基側鎖の位置を高精度で推定することは容易ではない。側鎖原子の位置が正しくない立体構造モデルに基づく電位計算が、どの程度危ういかは想像するに難くない。

そこでわれわれは、精度が低いタンパク質の立体構造を用いても RNA 相互作用面が推定できる方法の開発を進めている。タンパク質と RNA との複合体構造は X 線結晶解析により少しずつ明らかにされており、RNA に結合するタンパク質のアミノ酸配列は大量に分かっていることより、物理化学的な手法ではなく、大量データから統計的に情報を抽出し、データが語る言葉に耳を傾ける態度を取った。<sup>13)</sup>

タンパク質の立体構造データベース PDB<sup>14)</sup> には、RNA と相互作用した状態で立体構造が判明しているタンパク質が 86 種類存在した (2005 年末)。2008 年現在では、100 種類を超えている。これらの立体構造を用いて、RNA 相互作用面に現れるアミノ酸残基の傾向を調べた。各複合体立体構造におい

て溶媒接触表面積を計算し、コンピュータの中で RNA を消してしまい再び溶媒接触表面積を計算し、溶媒接触表面積に差が生じるアミノ酸残基を RNA と相互作用するアミノ酸残基と定義した。このような単純な計算からでも、大切な量がみえてくる。例えば、RNA 界面に存在するアミノ酸残基は、1 残基当たり  $40 \text{ \AA}^2$  を RNA に向けていることが分かる。また RNA 界面は  $1000 \text{ \AA}^2$  から  $3000 \text{ \AA}^2$  当りに分布していることも分かる。この計算では、複合体の構造から機械的に RNA を取り去った構造が、タンパク質が単体で存在しているときの構造であると仮定している。ここで示した傾向値をのちに複合体構造の予測用パラメータとして利用するに当たり、この仮定は問題になる。つまり、RNA とタンパク質が相互作用する際に、タンパク質が大きく構造変化をするのであれば、タンパク質単体の立体構造にこのパラメータを用いて RNA 界面を予測しても、正しく予測はできない。この問題を Ellis と Jones<sup>15)</sup> がすでに調べており、RNA に結合する際のタンパク質構造変化は、既存のデータを調べる範囲では、それほど大きくないことを確かめている。

RNA の界面に存在するアミノ酸残基の傾向を Fig. 2 に示す。グラフの縦軸は、ある種類のアミノ酸残基が RNA 界面に現れる割合をタンパク質表面に現れる割合で割った値の  $\log_2$  である。これはアミノ酸残基 1 つの情報なのでシングルレットの情報と呼んでいる。このような計算は古くからなされており、今までに報告されている傾向と Fig. 2 に示す傾向に大きな違いはない。<sup>16-20)</sup> 結果を概観すると、正電荷を持つアミノ酸が好まれており、負電荷を持つアミノ酸が嫌われる傾向にあるのは当然のことである。物理化学的な視点に立っとう少し詳細に結果をみると、2 つの疑問がわく。第 1 は、Lys と Arg の界面出現傾向にみられる有意な差である。Lys と Arg はどちらも側鎖に窒素原子を持つ正電荷を帯びたアミノ酸であり、RNA との界面に高い頻度で登場することは、理にかなっている。しかし Lys と Arg の出現傾向に有意な差があるのは不思議である。側鎖の大きさと構造のわずかな違いに、この有意な差を生み出す原因があるに違いない。第 2 は側鎖に芳香環を持つアミノ酸の違いである。DNA 及び RNA との相互作用面において、芳香環が塩基と並列に配列することで、 $\pi$  電子雲を共有し

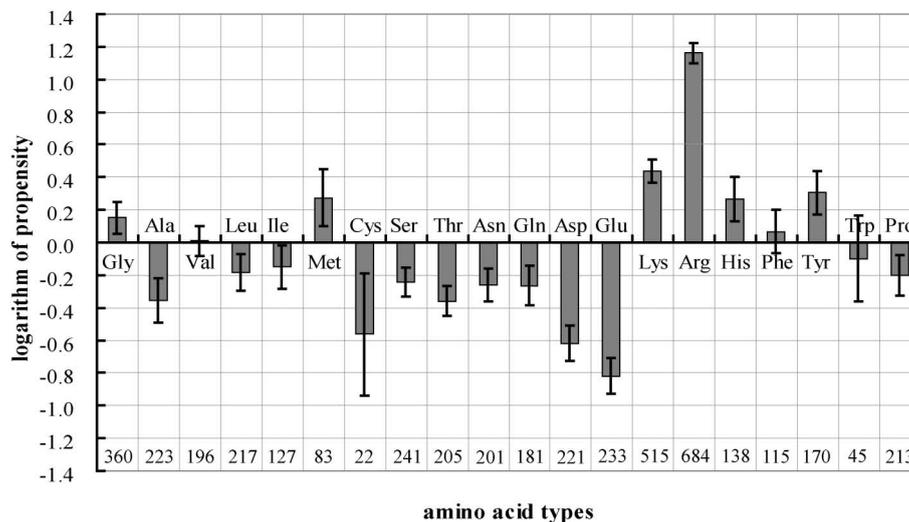


Fig. 2. Histogram of the Calculated Residue Singlet Interface Propensity in a Logarithm ( $\log_2$ ) Scale

A positive propensity indicates that a residue occurs more frequently in the interface than on the protein surface. An error bar for each propensity corresponds to standard deviations estimated from a bootstrap procedure with 1000 resamplings. The number given below the horizontal axis is the count of each amino acid type in the protein-RNA interfaces.

安定な構造になることが報告されている [例えば Ref. 21)]. Phe, Tyr, Trp は芳香環を持つ点で同じであるにも係わらず, RNA 界面に現れる傾向はだいぶ異なっており, その違いが何に由来するのかは分からない. いったん物理化学を離れ, 近年の大量データが物語ることを真摯に受け止めることで, 物理化学の問題が新たにみえてきている.

RNA 結合タンパク質の RNA 結合部位は面を形成しているので, あるアミノ酸残基が RNA と相互作用する残基であるならば, 空間的に隣に存在するアミノ酸残基も RNA 結合部位を形成していることが多いのは当然のことである. それならば, タンパク質の表面で隣り合う2つのアミノ酸残基 (アミノ酸残基ペア, 又は単にペア) が RNA 結合面にあるか, そうではないかによって, ペアの出現頻度は異なる可能性がある. ペアをなすアミノ酸残基を  $C\beta$  間距離が  $7.0 \text{ \AA}^2$  以内の2つの残基としたときに, RNA 界面におけるアミノ酸残基ペアの出現傾向を調べると Fig. 3 のようになる. このペアの情報をシングレットに対してダブルット情報と呼んでいる. この図は, アミノ酸残基ペアの RNA 界面に現れる傾向を, 同一種のペアがタンパク質表面に現れる傾向で割り算した数値を暖寒色のグラデーションで示したものである. このような解析はわれわれが初めて行った. ×印になっているペアは, 統計的に意味がある結果を得ることができなかったペアであ

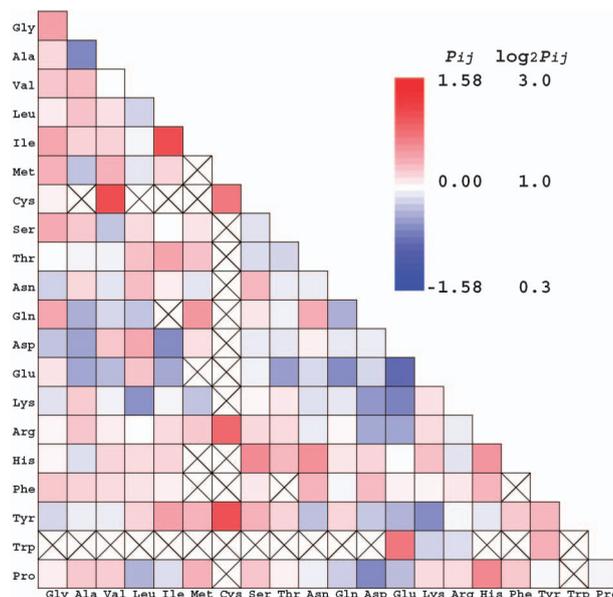


Fig. 3. A Graphical Matrix of the Residue Doublet Interface Propensities Color-coded in a Logarithm ( $\log_2$ ) Scale.

A value with a cross mark indicates that the data are not statistically sufficient to warrant the result.

る. Cys と Trp は RNA 界面に見い出されることは稀なため, これらの残基が関与するペアには×印が付いていることが多い. Lys-Arg ペアは, RNA 界面ではタンパク質の表面に比べるとやや好まれている傾向にある (薄い赤になっている). 同じ正電荷を持つ側鎖が隣り合って配置されるのは非常に不安定なはずだが, RNA 分子が常にその界面に存在す

るのならば、そのような配置を安定に保てるのかもしれない。同様に Lys-Lys ペアもやや好まれる傾向にある。しかし Arg-Arg ペアはむしろ嫌われる傾向にある。ここでも物理化学的に整合性の取れた説明は今のところできていない。アミノ酸残基単体の出現頻度において、負電荷を持つ Glu や Asp は嫌われていた。これらはペアになっても嫌われる傾向にあるのは、物理化学的に当然な帰結と思える。ところがアミノ酸残基単体では嫌われる傾向にあった Ile が、ペアになるとむしろ好まれる傾向に転じている。RNA と結合しているタンパク質立体構造において、Ile-Ile ペアをいくつか調べると、2つの Ile の側鎖が平らな平面構造を作り、その上に塩基が乗っている共通なパターンを見出すことができる。この相互作用の物理化学はこれからの問題であろう。

RNA と相互作用するタンパク質の RNA 界面にどのようなアミノ酸残基が存在するかが分かってくると、この傾向を利用して、界面が分かっている RNA 結合タンパク質の界面を推定することができるかもしれない。独立な情報として、タンパク質の機能を担っているアミノ酸残基は進化の過程で保存されることが分かっているので、<sup>22)</sup>ゲノム塩基配列から分かるたくさんの相同な RNA 結合タンパク質のアミノ酸配列を利用することでも、RNA 界面

(機能部位) を推定することができるはずである。そこで、ここで求めたアミノ酸残基の出現傾向値とアミノ酸残基の保存度を用いて、様々な RNA 界面推定法を構築した (Fig. 4)。いずれの方法も、立体構造が判明しているタンパク質において、表面に存在するアミノ酸残基に RNA と相互作用する可能性のスコアを与える。スコアが高ければ高いほど、そのアミノ酸残基が RNA との相互作用に関与している可能性が高くなる。Figure 4 には 9 種の方法 (式) が記載されている。  $S_x(S)$  とは、シングレットのみを用いた予測方法である。  $P_{i(x)}$  とはタンパク質の位置  $x$  に第  $i$  種 (20 種のアミノ酸のいずれか) のアミノ酸残基があるときのシングレットの値を意味する。この値の対数を取って、位置  $x$  にあるアミノ酸残基のスコアとしている。  $S_x(AS)$  とは、平滑化シングレットによる予測方法を意味する。これは  $S_x(S)$  を求めたのちに、ある位置  $x$  に存在するアミノ酸残基のスコアを、その周辺 7.0 Å 以内にある全残基のスコアの平均値として与える方法である。  $S_x(P)$  はアミノ酸残基の保存性のみを用いた予測法であり、  $V_i(x)$  とは、マルチプルアラインメントにおいて、タンパク質の位置  $x$  に第  $i$  種のアミノ酸が現れる割合 (0.0-1.0) を意味する。  $S_x(ASD)$  は、シングレットとダブルレットを用いた予測法であ

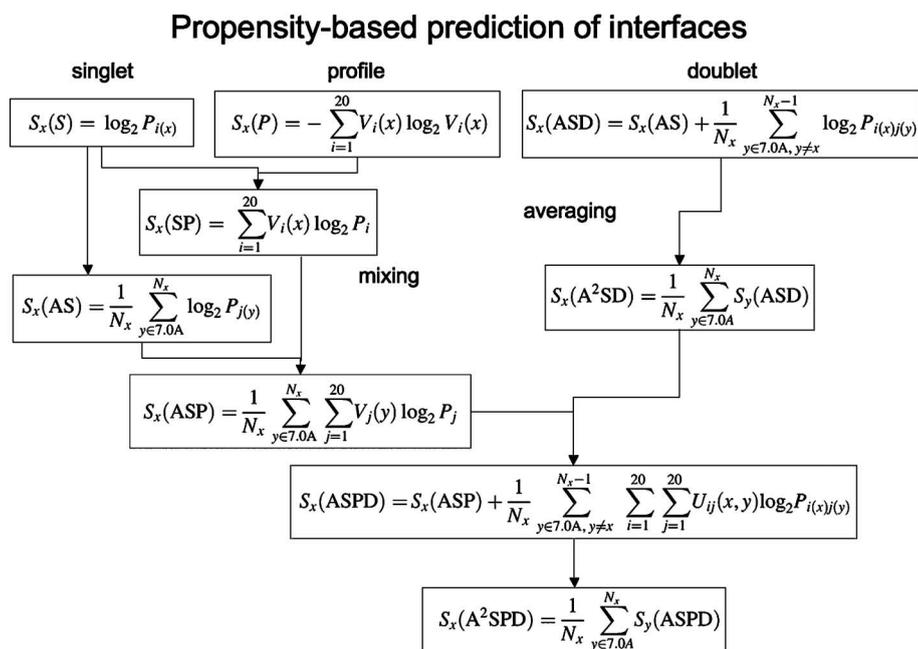


Fig. 4. Equations for Predicting RNA Interface Residues Based on the Propensities

The singlet and doublet propensities are derived from protein 3D structure data and the profile is derived from amino acid sequence multiple alignment. Read text for the detail.

り、式に現れる  $P_{i(x)j(y)}$  とは、タンパク質の位置  $x$  に第  $i$  種のアミノ酸残基が、 $x$  から  $7.0 \text{ \AA}$  以内の、ある位置  $y$  に第  $j$  種のアミノ酸残基があり、 $i$  と  $j$  のペアの出現傾向を意味する。以下  $S_x(SP)$ ,  $S_x(A^2SD)$ ,  $S_x(ASP)$ ,  $S_x(ASPD)$ ,  $S_x(A^2SPD)$  は上記の方法を組み合わせた予測法になる。

$S_x(AS)$  の考え方は、このような予測ではよく見掛ける方法である。Chou と Fasman によるアミノ酸配列からの二次構造予測<sup>23)</sup> も、同様の考え方で構築されている。20 種のアミノ酸になんらかのスコアを与え、実際のタンパク質 (アミノ酸配列) の各残基にそのスコアをアミノ酸残基種に応じて割り当て、注目している残基の周辺 (アミノ酸配列ならば前後数残基、タンパク質立体構造ならば、数  $\text{\AA}$  以内の全残基) のスコアを平均した値を、注目している残基の最終スコアにしているにすぎない。最終スコアが高い値を示している部位が予測しようとした性質を持っているアミノ酸残基とみなすことになる。

上に挙げた 9 種類の予測方法の精度をテストした結果、 $S_x(A^2SPD)$  が一番精度の高い予測ができることが判明した。この予測を実行するためには、タンパク質の立体構造と相同タンパク質のマルチプルアランメントが必要となる。それらを用いると約 80% の精度で RNA 相互作用面に存在するアミノ酸残基を見い出すことができた。この 80% の意味するところは以下の通りである。 $S_x(A^2SPD)$  の方法で、あるアミノ酸残基が RNA 界面を構成すると予測された場合、10 個中 8 個は本当に RNA 界面に存在するアミノ酸残基と確認することができ、2 個は界面に存在するアミノ酸残基とは確認できなかった (specificity が 80%)。ただし、この予測方法で RNA 界面に存在するアミノ酸残基すべてを見い出せた訳ではない。実験的に分かっている全界面アミノ酸残基の 10% を見い出すことに成功し、残り 90% は、相互作用残基と推定することができなかった (sensitivity が 10%)。同じ評価を  $S_x(ASPD)$  に行くと、sensitivity が 10% の際に specificity が約 70%、 $S_x(SP)$  では sensitivity が 10% の際に specificity が約 50% となった。これらの RNA 結合面推定方法は、(<http://cib.cf.ocha.ac.jp/>) で利用可能である。

高い精度で RNA と相互作用するアミノ酸残基を予測できる方法は、いろいろな場面で有用である。例えば、立体構造が判明している RNA 結合タンパ

ク質の RNA 結合面を実験的に確かめる場合には、全くランダムに点突然変位を導入して、RNA との親和性の変化を測定するのは大変だが、 $S_x(A^2SPD)$  又は  $S_x(ASPD)$  の方法で相互作用部位と予測された残基に、実験的に変位を入れて RNA との親和性を測定すれば、70% から 80% の実験では、RNA との親和性が下がって、結合部位を効率よく同定することができる。また RNA 結合タンパク質と RNA の複合体構造をコンピュータで推定する場合には、 $S_x(A^2SPD)$  又は  $S_x(ASPD)$  の方法で RNA 結合部位と推定されたアミノ酸残基と RNA との間に擬似的な引力を導入することで、両者が接近する拘束条件にできる。全く制約なくタンパク質と RNA をドッキングさせることは、今でも困難を伴う計算である。われわれはこの方法を用いて、タンパク質と RNA の複合体構造を推定する世界コンテスト<sup>24)</sup> にも挑戦をしている。

#### 4. タンパク質と低分子の相互作用推定

タンパク質の立体構造が X 線結晶解析や NMR によって判明した場合、あるいはホモロジーモデリングによって推定できた場合に、そのタンパク質が RNA と結合するタンパク質であるならば、タンパク質表面のどの残基が RNA と相互作用するかを比較的高精度で推定する方法を説明してきた。われわれは、まったく同じ方法をタンパク質と低分子の相互作用部位を推定することにも適用できるのではないかと考えており、現在手法の開発に取り組んでいる。

タンパク質の立体構造を決定する際に、タンパク質が低分子と相互作用した状態で X 線結晶解析される場合が非常に多い。タンパク質立体構造データベース PDB<sup>14)</sup> において、タンパク質ではなく低分子に注目してデータを整理し直すと同一の低分子が様々なタンパク質と相互作用していることが分かる。<sup>25,26)</sup> ATP, GTP, FAD などの核酸が関与する低分子や、金属イオン、糖などが様々なタンパク質との複合体として立体構造データベースに格納されている。ビタミンの大部分の立体構造がタンパク質と相互作用した状態で、既に判明していることは、あまり知られていないことかもしれない。<sup>27)</sup> ある低分子に着目して、その低分子と相互作用するアミノ酸残基種を同定することで、RNA 界面残基予測と同様の高精度予測法が構築できるであろう。この方法

ができれば、タンパク質への低分子ドッキング計算に貢献できると考えている。低分子と相互作用すると推定されたアミノ酸残基と低分子との間に、仮想的な相互作用を導入することで、*in silico* ドッキングの効率を圧倒的によくできる可能性がある<sup>28)</sup>。

### 5. まとめ

ゲノム塩基配列から推定されたアミノ酸配列を持つタンパク質の大部分の立体構造が X 線結晶解析、NMR 及びホモロジーモデリングによって明らかになるときが近づいてきている。そのときに、タンパク質立体構造の座標データが無意味にならないように、構造バイオインフォマティクスの研究は精力的に進められている。われわれのグループもこの動きの中で、ここに概観した研究活動を展開している。構造バイオインフォマティクスの研究に従事する人材には、学際的な知識が要求される。要求される知識には分子生物学や物理学、化学、情報科学などが含まれ、大学及び大学院に設置されている既存の教育カリキュラムでは対応ができない幅の広さである。これからの薬学や分子生物学の発展のためには、このような学際教育をしっかりとやっていくことが不可欠であろう。

**謝辞** ここに紹介した研究は、多くの方々のご指導とご協力によって初めて進めることができました。ホモロジーモデリングに関しては、お茶の水女子大学郷通子学長のご指導と長浜バイオ大学山口晶大氏の協力に、タンパク質と RNA の相互作用予測は、日本原子力研究開発機構郷信広先生のご指導と元奈良女子大学 Oanh T. P Kim 博士の協力に、国際コンテスト CAPRI への参加は、大阪大学蛋白質研究所中村春木先生と東京大学医科学研究所木下賢吾先生の協力に、タンパク質と低分子の相互作用研究は、高エネルギー加速器研究機構若槻壮市先生の激励と長浜バイオ大学塩生真史博士の協力に、それぞれお礼申し上げます。

### REFERENCES

- 1) Wheeler D. A., Srinivasan M., Egholm M., Shen Y., Chen L., McGuire A., He W., Chen Y.-J., Makhijani V., Roth G. T., Gomes X., Tartaro K., Niazi F., Turcotte C. L., Irzyk G. P., Lupski J. R., Chinault C., Song X.-z., Liu Y., Yuan Y., Nazareth L., Qin X., Muzny D. M., Margulies M., Weinstock G. M., Gibbs R. A., Rothberg J. M., *Nature*, **452**, 872–876 (2008).
- 2) Lander E. S., *et al.*, *Nature*, **409**, 860–921 (2001).
- 3) Wadman M., *Nature*, **409**, 788 (2008).
- 4) Dodson E. J., *Nature*, **450**, 176–177 (2007).
- 5) Vitkup D., Melamud E., Moulton J., Sander C., *Nat. Struct. Biol.*, **8**, 559–566 (2001).
- 6) Yura K., Yamaguchi A., Go M., *J. Struct. Funct. Genomics*, **7**, 65–76 (2006).
- 7) Go M., Yura K., Shionyu M., ‘Frontiers of Computational Science’ eds. by Kaneda Y., Kawamura H., Sasai M., Springer-Verlag, Berlin, 2007, pp. 75–80.
- 8) Yamaguchi A., Iwade M., Suzuki E., Yura K., Kawakita S., Umeyama H., Go M., *Nucleic Acids Res.*, **31**, 463–468 (2003).
- 9) Imanishi T., Itoh T., Suzuki Y., O’Donovan C., Fukuchi S., *et al.*, *PLoS Biol.*, **2**, e162 (2004).
- 10) Keene J. D., *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 7018–7024 (2001).
- 11) Sheinerman F. B., Norel R., Honig B., *Curr. Opin. Struct. Biol.*, **10**, 153–159 (2000).
- 12) Stawiski E. W., Gregoret L. M., Mandel-Gutfreund Y., *J. Mol. Biol.*, **326**, 1065–1079 (2003).
- 13) Kim T. P. O., Yura K., Go N., *Nucleic Acids Res.*, **34**, 6450–6460 (2006).
- 14) Berman H., Henrick K., Nakamura H., Markley J. L., *Nucleic Acids Res.*, **35**, D301–D303 (2007).
- 15) Ellis J. J., Jones S., *Proteins: Struct. Funct. Bioinfo.*, **70**, 1518–1526 (2008).
- 16) Jones S., Daley D. T., Luscombe N. M., Berman H. M., Thornton J. M., *Nucleic Acids Res.*, **29**, 943–954 (2001).
- 17) Treger M., Westhof E., *J. Mol. Recognit.*, **14**, 199–214 (2001).
- 18) Kim H., Heong E., Lee S. W., Han K., *FEBS Lett.* **552**, 231–239 (2003).
- 19) Allers J., Shamoo Y., *J. Mol. Biol.*, **311**, 75–86 (2001).
- 20) Ellis J. J., Broom M., Jones S., *Proteins: Struct. Funct. Bioinfo.*, **66**, 903–911 (2007).
- 21) Nakanishi K., Ogiso Y., Nakama T., Fukai S., Nureki O., *Nat. Struct. Mol. Biol.*, **12**, 931–

- 932 (2005).
- 22) Lichtarge O., Sowa M. E., *Curr. Opin. Struct. Biol.*, **12**, 21–27 (2002).
- 23) Chou P. Y., Fasman G. D., *Biochemistry*, **13**, 222–245 (1974).
- 24) <http://www.ebi.ac.uk/msd-srv/capri/>
- 25) Yamaguchi A., Iida K., Matsui N., Tomoda S., Yura K., Go M., *J. Biochem.* (Tokyo), **135**, 79–84 (2004).
- 26) <http://hetpdbnavi.nagahama-i-bio.ac.jp/>
- 27) Yura K., “Syokuhinokagaku” Chap. 3.5, eds. by Kaminogawa S., Tanokura M., Tokyo Kagaku Dozin Co., 2005, pp. 87–95.
- 28) Brewerton S. C. *Curr. Opin. Drug Discov. Devel.*, **11**, 356–364 (2008).