構造生物学研究の進展状況とその問題点

岩崎 愛（指導教員：由良 敬）

1 はじめに

ゲノムとは、生物を構成するタンパク質の設計図が書き込まれている生体高分子である。多くの生物において、ゲノムの実体は DNA であり、DNA は 4 種類の塩基が組み合わされて構築される鎖状高分子である。ヒトゲノムの場合、その長さは約 30 億塩基に達する。塩基の並び順（塩基配列）は、生物種ごとに異なっており、塩基の並び順そのものが、暗号化された設計図となっている。1980 年代初頭にヒトのゲノムを「解読」するプロジェクトが開始され、2001 年にゲノム全塩基配列が決定された。ゲノム塩基配列決定の技術は、このプロジェクトによって向上し、現在ではさまざまな生物種のゲノムが決定されるようになった。

ゲノム塩基配列には、タンパク質をコードする部分（Open Reading Frame(ORF)）とタンパク質の発現を制御する部分がある。タンパク質は 20 種類のアミノ酸がくみ合わされて構築される鎖状高分子であり、アミノ酸の並び順が DNA に暗号化されて記されている。4 種類の塩基で 20 種類のアミノ酸をコードしていることより、連続する 3 個の塩基がひとつのアミノ酸をコードしていることになる。よって、何かの原因によって塩基が変化すると、コードされているアミノ酸が変化することがある。

ヒトには、約 10 万種のタンパク質が存在するといわれている。ゲノムから転写翻訳されたこれらのタンパク質は、様々な機能をもち、生命活動の基盤を形成している。タンパク質には、生体中で発生する様々な化学反応の触媒としてはたらく酵素、生体の形を形成する構造タンパク質、外界の情報に対応して生体の状態を変化させる情報伝達タンパク質などがある。これらのタンパク質が適確に役割を果たせるようにしているのが、タンパク質の発現を制御するゲノムの部分である。

ゲノム塩基配列に変化が生じると、これらのタンパク質に変化が生じ、適切な役割を果たせなくなり、疾患にいたる場合がある。ゲノムに見つかる変異は、疾患に直接関係するわけではなく、ゲノム情報が転写翻訳されたタンパク質の異常が疾患と直接関係する。

ゲノムが決定されただけでは、その遺伝子がどのような機能をもっているかは分からない。そのことから、ゲノム塩基配列の読み取りが完了した後、ゲノムにコードされているタンパク質の機能解析が生命科学分野における重要な研究となった。

タンパク質が機能するためには、三次元構造を形成する必要がある。タンパク質が構造をとって機能することから、タンパク質の三次元構造を知ることはタンパク質の機能をしる上で重要であり、タンパク質の三次元構造を測定する実験が数多く行われている。現在までに明らかにされたタンパク質の三次元構造の情報は、生体高分子立体構造データをあつめたデータベースである Protein Data Bank(PDB)に登録されている。

タンパク質の立体構造を実験的に決定するには、タンパク質を大量に発現させ、結晶を作成し、X 線を用いて原子の位置を明らかにする必要があり、多くの時間と手間を要する。そのため、すべてのタンパク質の立体構造を実験的に決定することは現実的ではない。そこでコンピュータによるタンパク質の三次元構造予測が行われている。現在、精度よく構造予測をする方法としてホモロジーモデリングがある。これは、構造既知のタンパク質と類縁関係にあるタンパク質の構造を、コンピュータで構築する方法である。この方法を用いることで、全生物がもつ全タンパク質の立体構造を明らかにすることができ、その情報にもとづいて、全タンパク質の機能解析を行うことができるようになる。

それでは、現在のタンパク質立体構造データは、全タンパク質の構造解析（予測）を行うために十分な量になっているのだろうか（構造充填度は何割か？）。十分でない場合は、いつごろ十分になると予想されるだろうか。さらには、どのような順番で実験データを産出すれば、効率よく機能解析をすすめることができるようになるだろうか。これらのことを明らかにすることを目的に、ゲノム塩基配列が明らかになっているすべてのバクテリアとヒトがもつすべてのタンパク質のアミノ酸配列と、既知タンパク質立体構造すべての総当たり解析を行った。

2 手法

2.1 ゲノム塩基配列からのタンパク質アミノ酸配列情報の取得

バクテリアの既知ゲノム塩基配列から、ORF を抽出し、アミノ酸配列に翻訳した。同時にヒトゲノムにコードされているタンパク質のアミノ酸配列も取得した。これらのアミノ酸配列から、生体膜貫通部分と、構造不安定部分（天然変性領域）を予測した。生体膜貫通部分をもつタンパク質は膜タンパク質とよばれ、水中に存在するタンパク質（水溶性タンパク質）とは異なる性質をもっていることが分かっている。水溶性タンパク質とは異なり、膜タンパク質は立体構造の解析が難しく、PDB に登録されている膜タンパク質の数は水溶性タンパク質よりも圧倒的に少ない。よって水溶性タンパク質と膜タンパク質を区別することにし、膜貫通部位の有無によって、両タンパク質を判別することにした。膜貫通領域は TMHMM をもちいて予測した。

天然変性領域は、X 線結晶解析によって立体構造を測定することが非常に困難な部分であることから、その部分は今回の解析から除外することにした。天然変性領域は、DisEMBL を用いて予測した。

2.2 類似配列検索

天然変性領域以外のアミノ酸配列を、PDB に登録されているタンパク質のアミノ酸配列すべてと比較して、類似のアミノ酸配列を検索した。検索には隠れマルコフモデルにもとづく PHMMER を用いた。ゲノムから取得したアミノ酸配列の全長にわたって類似のアミノ酸配列をもつ PDB エントリーは稀であったので、部分一致データを蓄積した。

2.3 時系列の測定方法

タンパク質の立体構造データは年々蓄積されていることより、上記の対応データは年々増加することが予想される。未来における変化を、現在までのデータから予測するには、過去の PDB のデータとゲノムから推定されたアミノ酸配列の対応関係を明らかにする必要がある。そのデータがあれば、未来に向かって外挿することができる。PDB のデータには各タンパク質のデータがいつ登録されたかが記載されているので、その情報を用いて、2000 年までさかのぼりながら上記の計算を繰り返し、対応関係を蓄積した。

2.4 データフォーマット

得られる情報は膨大な量になる。この情報を色々な側面で検索出来るようにする必要があるため、データはすべて XML 形式で格納することにし、以下の DTD を設計した（図１）。

```
<!DOCTYPE chromosome[
 <!ELEMENT chromosome (ID)>
 <!ELEMENT ID (Disorder,TM,pdb)>
 <!ELEMENT Disorder (s,e)>
 <!ELEMENT TM (s,e)>
 <!ELEMENT pdb (template,target)>
 <!ELEMENT template (s,e)>
 <!ELEMENT target (s,e)>
 <!ELEMENT s (#PCDATA)>
 <!ELEMENT e (#PCDATA)>
 <!ATTLIST chromosome id ID #REQUIRED>
 <!ATTLIST ID id ID #REQUIRED>
 <!ATTLIST ID length CDATA #REQUIRED>
 <!ATTLIST Disorder n CDATA #REQUIRED>
 <!ATTLIST Disorder program CDATA #REQUIRED>
 <!ATTLIST TM n CDATA #REQUIRED>
 <!ATTLIST TM program CDATA #REQUIRED>
 <!ATTLIST pdb id CDATA #REQUIRED>
 <!ATTLIST pdb program CDATA #REQUIRED>
 <!ATTLIST pdb date CDATA #REQUIRED>
 <!ATTLIST pdb side CDATA #REQUIRED>
]>
```

図 1: ゲノムとタンパク質立体構造の対応データを記述する XML の DTD

3 結果

3.1 バクテリア由来タンパク質の構造充填度

バクテリアでの計算結果を図 2 に示す。現在、2,700 種超のバクテリアでゲノムが決定されており、このゲノムデータに対して、手法に記した計算を実行することができた。その結果、全タンパク質のうち 62%（アミノ酸残基単位で測定）が立体構造既知（実験的に明らかになっているタンパク質と類似の配列をもっている）ことが分かった（図 2）。PDB への登録データに付随する登録年代にもとづき、過去の PDB を再現し、過去の PDB を用いた場合に、上記と同じ計算を行ったところ、図 3 が得られた。バクテリアの水溶性タンパク質は、2008 年あたりから割合の増加率が鈍化していることがわかった。2008 年以降の増加率が維持されると仮定して、グラフを未来に向かって外挿すると、2032 年にバクテリア全タンパク質の三次元構造がわかることが予測できる。しかし、増加率は鈍化している傾向にあるため、飽和し 100%に達しない可能性があることがわかった。ところが、膜タンパク質は、立体構造既知のアミノ酸残基数の割合の変化が急であることがわかった。

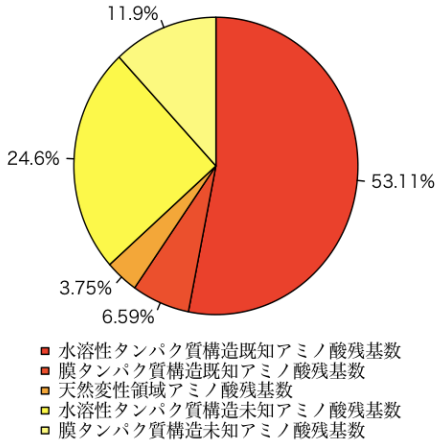

図 2：バクテリアにおける現在のタンパク質構造充填度

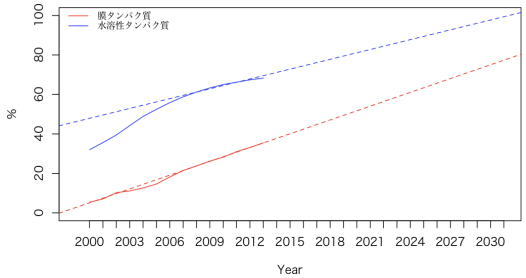

図 3:バクテリアにおけるタンパク質構造充填度の時系列

3.2 ヒトとバクテリアの構造充填度の比較

ヒトゲノムとバクテリアゲノムにおける構造充填度を比較すると、2008 年にバクテリアのグラフの傾きが鈍化するまでは、バクテリアの構造充填度の傾きが、ヒトの構造充填度の傾きよりも大きいことがわかった。ヒトにおいては立体構造既知のアミノ酸残基数の割合の変化がバクテリアに比べて少ないことが分かった。

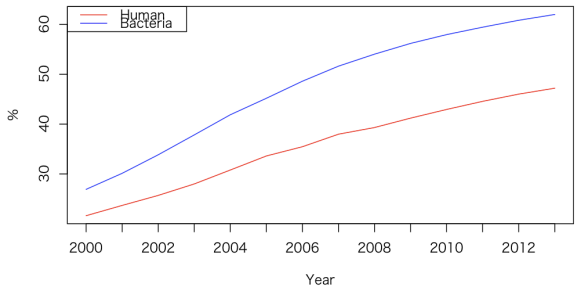

図 4:バクテリアとヒトの構造充填度の時系列比較

4 考察

バクテリアとヒトでの構造充填度を比較したところ、今現在の PDB では十分な構造情報がなく、バクテリアがもつタンパク質の 62%、ヒトの 47%のみが、立体構造と対応することがわかった。時系列のグラフより、構造充填度の伸び率が鈍化していることがわかる。割合の変化のグラフの傾きが小さくなっていることから、初期に比較的簡単に結晶化するタンパク質の解析がやり尽くされ、解析が難しいものが残っていると考えられる（Nature 443,382; 2006）。また、グラフの傾きが変化したタイミングでは、アメリカの構造解析プロジェクト「Protein Structure Initiative（PSI）」、日本のプロジェクトである「タンパク 3000」など大きなプロジェクトのフェーズの進行や終わりがあり、これらもタンパク質の構造充填度の伸び率鈍化の要因の一つであると推測できる。また、ヒトの方がグラフの傾きが鈍いことから、ヒト固有のタンパク質の構造の決定は難しいことが読み取れる。

5 今後の課題

本研究では、PDB 内の立体構造データの情報の充填度の推移を追い、その変化の原因について考察した。今回どのような順番で実験データを排出すれば、効率よく機能解析をすすめることができるようになるかについての提案が出来なかったので今後の課題としたい。また、バクテリアだけでなく、古細菌・真核生物全体に対しても同様の計算を行うことで、より詳しいタンパク質の構造充填度を求めたい。

参考文献

[1] Kei Yura, Akihiro Yamaguchi, Mitiko Go: Coverage of whole proteome by structural genomics observed through protein homology modeling database, Journal of Structural and Functional Genomics June 2006, Volume 7, Issue 2, pp 65-76

[2] RCSB Protein Data Bank, http://www.rcsb.org/pdb/

[3] HMMER, http://hmmer.janelia.org/

[4] TMHMM, http://www.cbs.dtu.dk/services/TMHMM/

[5] DisEMBL, http://dis.embl.de/