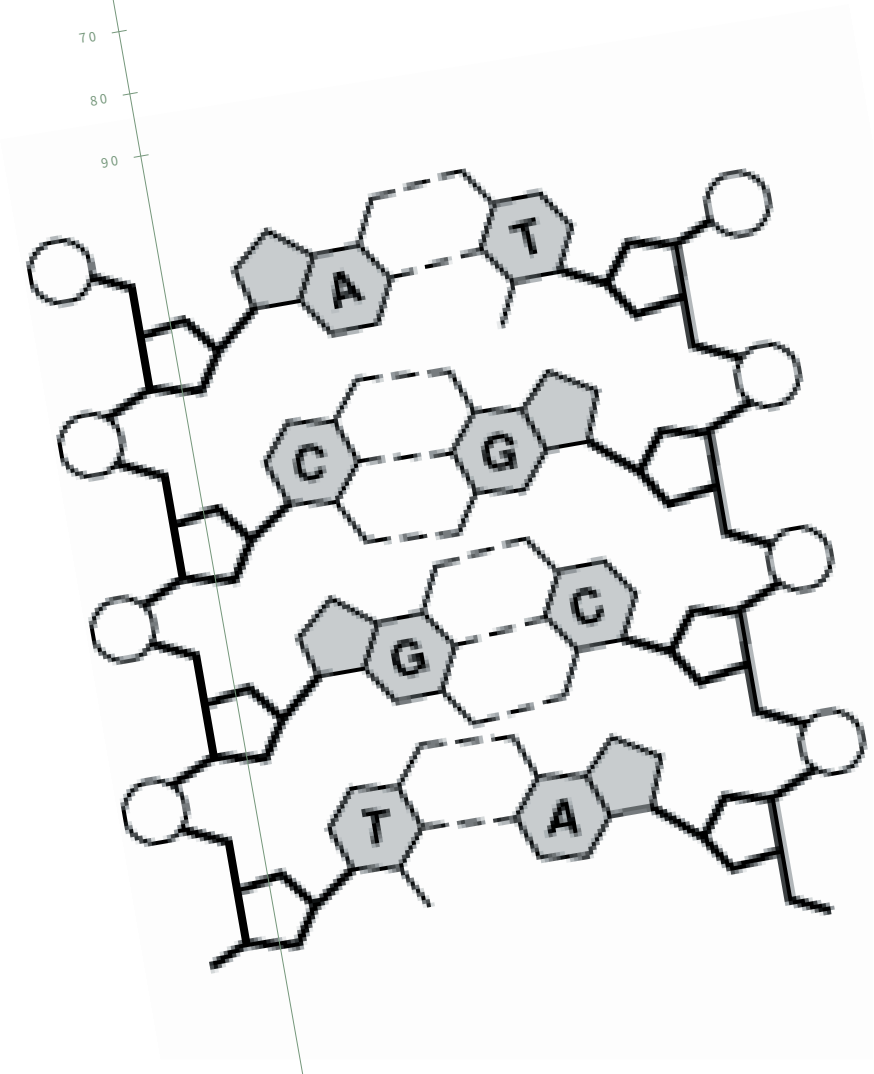
第4回

似たもの探し ―― 遺伝子やアミノ酸の機能は類似性・相同性から推定

人間を解明する競争は、ヒトゲノムプロジェクトによってヒトのゲノム配列解読が完了した後、中に隠されている遺伝子の同定とその機能の解析というポストゲノムにその主戦場が移された。ゲノムの海の中から遺伝子を見つけたあとは、それがどのような機能を持っているかを決定しなければならない。そのために重要なのが「似たもの探し」である。

浅田一憲 株式会社オープンループ、北海道大学大学院医学研究科
多田光宏 北海道大学遺伝子病制御研究所、株式会社ジェネティックラボ



進化は「へぼ職人」の間違いの産物

「自然は、職人であって発明家ではない」という有名な言葉がある。この言葉はF. Jacobという生物学者が1977年に『進化とへぼ職人』と題して、サイエンス誌に書いたものだ。生物すなわちDNAの進化においては、機能を持つDNA配列が突然「天才」や「天啓」によって発明されるものではなく、前作を模倣しようとする「へぼ職人」が間違いをやらかし、それが偶然、前作よりも優れていて、選択されるという過程が綿々と続いてきたことによるものだ。進化する配列は時とともにDNA塩基の置換・欠失・挿入といった間違い（＝変異）を蓄積してゆく。時には「動く遺伝子トランスポゾン」などの組み換えにより本棚と机を合体させ、学習机を作ってしまうこともある。実際に作られるDNA配列やアミノ酸配列には、ほとんど必ず似た配列が存在する。それがヒトのゲノムにはない場合でも、遠い祖先で分かれた他の動物種に存在するのだ。

新しい遺伝子を発見したとき、それがどのような働きをするのかを推定するために一番手っ取り早い方法は、すでに機能が判明している遺伝子の中から似たような

進化は「へぼ職人」の間違いの産物

親の遺伝子
……A T A G T A C C G A
子の遺伝子
……A T A G T T G C G A

「へぼ職人」の間違いが偶然に前作よりも優れていて、選択されるという過程＝DNAの進化

DNA塩基配列を持つものを探し出すことだ。塩基配列が似ているということは、作り出すたんぱく質が似ており、機能が似ているということを意味する。類似性・相同性を決定する「ホモロジー解析」は、バイオインフォマティクスにとって重要な技術なのだ。

ホモロジー解析

ではどのように2つの似た配列を比較するのか？ その基本としてアラインメント（整合化）という概念がある。たとえばGATCAGTAとGTTGATAという2つの配列があったとしよう。これらはそのまま比較したのではうまく合わせることができない。そこでこれらDNA配列には置換のほか、欠失や挿入が入っているとして、その部位に「—」（ギャップ）を適切に入れてやることにより、

G A T C G — T A
G — T T G A T A

のように一致している部分をうまく合わせることができる。これがアラインメントであり、それを行うアルゴリズムをダイナミックプログラミングという。

表：DP行列の例

	G	A	T	C	G	T	A
G	＼	—					
T			＼				
T				＼			
G					＼		
A						|	
T						＼	
A							＼

ダイナミックプログラミング（DP）の原理を単純化すると、上表のようなDP行列と呼ばれる行列を作成することと考えればよい。アラインメントは表の線（ポインター）をたどっていくことによって行う。すなわち「＼」があるところでは対応する塩基同士を合わせ、「—」があるところではギャップとして縦の配列に、「｜」があるところでは横の配列に「—」を挿入する。塩基同士の比較をそのまま進めれば一番似ている場合は次に進み、似ている部分が出てくるまでちょっと待ったほうがいい場合はどちらかにギャップを入れてやる。実際は、行列の各要素には「＼」や「—」ではなくスコアとよばれる数値が入る。このようにしてダイナミックプログラミングによって似ている配列を見つけ出すのである。

参考：DP配列の作成方法

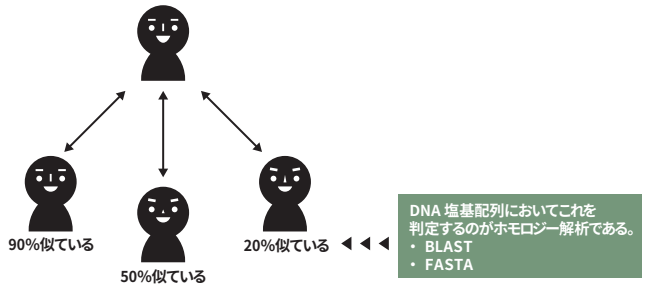
塩基同士を並べる場合、一致していれば「正」、不一致であれば「負」の得点を与え、ギャップが入る場合にはギャップペナルティーという「負」の得点を与える。それらの値を、計算するマス目の左斜め上、上、左のマス目の値に加え、そのうちの最大値として順次計算する。
*i*行*j*列の升目の値 $F(i,j)$ は以下の数式で計算する。

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

$s(xi,yi)$ は塩基の一致・不一致とその種類により与えられるスコアで、d はギャップペナルティーである。マス目の計算の際に、3つの中のどれが使われたかをポインターとして記録しておき、DP行列ができ上がった段階で、右下のほうからポインターをたどってアラインメントを決定する。アミノ酸配列の場合には、もとのDNAコドンの1置換でアミノ酸が変化したり、構造上似たアミノ酸の場合に高い得点を与え、遠いアミノ酸同士には負の値を与えるためにBLOSUM50またはPAM250アミノ酸置換行列というものを与えて、$s(xi,yj)$ を計算する。BLOSUM50でA、R、N……と並ぶのはアラニン、アルギニン、アスパラギン……といったアミノ酸の1文字表記である。境界条件の与え方により、配列全部を合わせる大域的アラインメント（Needlman-Wunschアルゴリズム）と一部だけ合わせる局所アラインメント（Smith-Watermanアルゴリズム）という2種があり、後者はインターネット上で利用できる。

・Smith-Watermanアルゴリズム
Jump www.ch.embnet.org/software/FDFSW_form.html

ホモロジー解析とは？



ホモロジー解析の高速化

このようなアルゴリズムは最適な答えを導き出してくれるが、実用上困った問題がある。それは過去の多くの経験と同様、計算時間の問題である。こうしたアルゴリズムの計算にはn文字とm文字を比較するとして$O(nm)=O(n^2)$の計算時間がかかる。1000塩基くらいを問い合わせ配列として、10^8程度のデータベース内でアラインメントを行うと10^4秒程度（約3時間）もの時間がかかる。

これでは実用上困るので、それほど精度の必要のない解析では少し妥協してヒューリスティックなアルゴリズムが使用される。有名なのは、BLASTまたはFASTAと呼ばれ、最初に短い断片でアラインメント検索を行い、完全に一致するものが見付かる都度両端を伸長していき、長く延ばせたものの順に報告するものである。BLASTは開発当初ギャップを許さなかったが、現在ではギャップを許した解析が可能であり、Smith-Watermanアルゴリズムに遜色ない結果を出すようになった。多くの研究者がBLASTを利用しており、非常にポピュラーなアルゴリズムと言えよう。

もう1つは、「隠れマルコフモデル」と呼ばれる確率論的なアプローチだ。つまり確率的な進化の力がゲノムの上に働いているという考えのもとに塩基の変化を確率で捉えるものである。隠れマルコフモデルは信号とノイズが混在する音声認識において使われてきた。DNAのホモロジー解析では隠れた本当（あるいは仮想的）の配列（＝状態列）に対し、それが各々置換・欠失・変異を起こしたと仮定して、実際に観察されるシンボル列になる確率を計算し、もっとも確からしい元の状態列は何になるかを推定する。

遺伝子機能の決定はポストゲノム課題

このようなアラインメントという考えを導入することによって、塩基配列やアミノ酸配列同士のホモロジーをスコア値あるいは確率として客観的な数値によって知ることができる。もともとホモロジー解析は、進化上DNA配列同士が近い関係にあるか否かを遺伝子間距離として測定し、DNA進化系統樹を作成するために開発されたのだが、現在ではむしろ、インターネットで公開されている大きなゲノムデータベースを解析し、自分の把握している遺伝子の配列と似ているものを検索したり、配列から機能のある領域を見いだしたりして機能の未知な配列の機能を推定したりするために用いられている。

ゲノムプロジェクトが終了しても、大半の遺伝子は機能が未知のままであり、このようなホモロジー解析を駆使して、得られる手がかりから機能を決定してゆくという重要な仕事が残っている。

株式会社オープンループでは、BLAST解析を非常に高速に行う技術を発明し、現在、商品化に向けて開発中である。
Jump www.openloop.co.jp/press/

・BLAST
Jump www.ncbi.nlm.nih.gov/BLAST/
・FASTA
Jump www.ebi.ac.uk/fasta33/

参考文献：『バイオインフォマティクス―確率モデルによる遺伝子配列解析』リチャード・ダービン（著）／阿久津達也（訳）／医学出版（発行）／4757801009（ISBN）

［インターネットマガジン　バックナンバーアーカイブ］ ご利用上の注意

このPDFファイルは、株式会社インプレスR&D（株式会社インプレスから分割）が1994年～2006年まで発行した月刊誌『インターネットマガジン』の誌面をPDF化し、「インターネットマガジン　バックナンバーアーカイブ」として以下のウェブサイト「All-in-One INTERNET magazine 2.0」で公開しているものです。

http://i.impressRD.jp/bn

このファイルをご利用いただくにあたり、下記の注意事項を必ずお読みください。

- 記載されている内容（技術解説、URL、団体・企業名、商品名、価格、プレゼント募集、アンケートなど）は発行当時のものです。
- 収録されている内容は著作権法上の保護を受けています。著作権はそれぞれの記事の著作者（執筆者、写真の撮影者、イラストの作成者、編集部など）が保持しています。
- 著作者から許諾が得られなかった著作物は収録されていない場合があります。
- このファイルやその内容を改変したり、商用を目的として再利用することはできません。あくまで個人や企業の非商用利用での閲覧、複製、送信に限られます。
- 収録されている内容を何らかの媒体に引用としてご利用する際は、出典として媒体名および月号、該当ページ番号、発行元（株式会社インプレスR&D）、コピーライトなどの情報をご明記ください。
- オリジナルの雑誌の発行時点では、株式会社インプレス R&D（当時は株式会社インプレス）と著作権者は内容が正確なものであるように最大限に努めましたが、すべての情報が完全に正確であることは保証できません。このファイルの内容に起因する直接的および間接的な損害に対して、一切の責任を負いません。お客様個人の責任においてご利用ください。

このファイルに関するお問い合わせ先
株式会社インプレスR&D
All-in-One INTERNET magazine 編集部
im-info@impress.co.jp