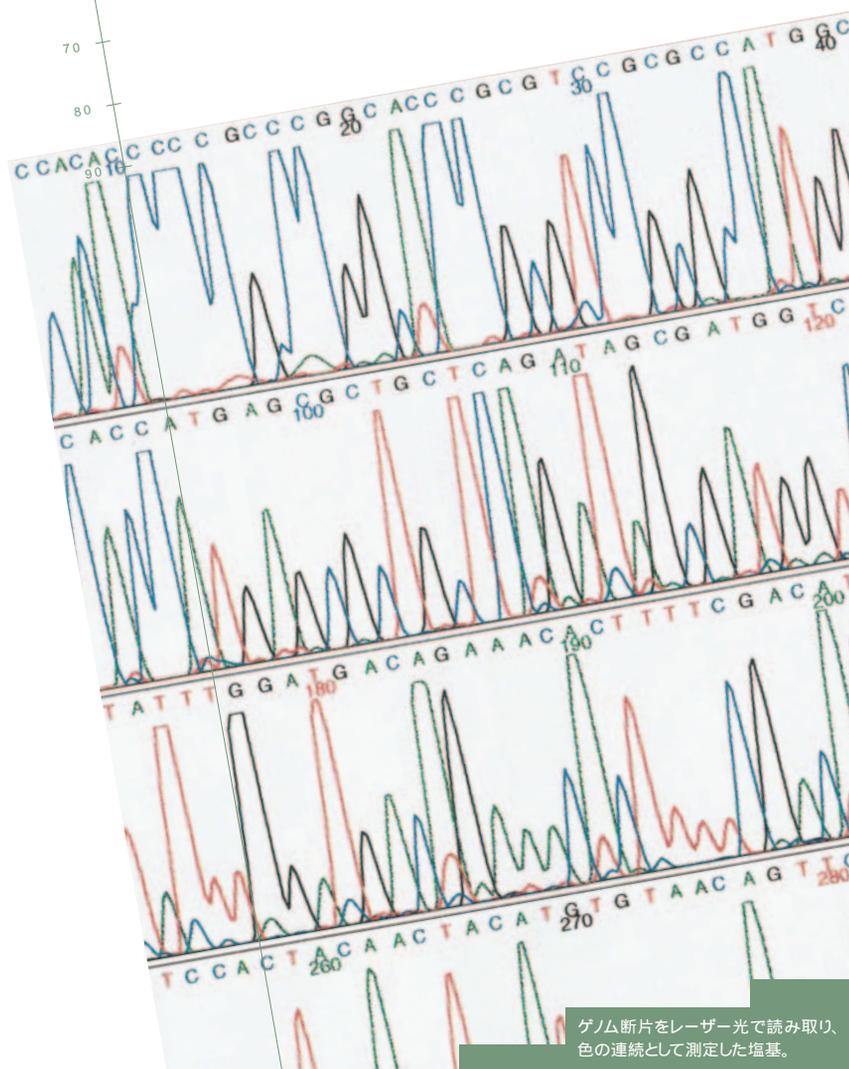


第3回

ゲノムの読み方、遺伝子の探し方

ヒトゲノム(遺伝情報を表す塩基配列)を読み解くことは、人間の仕組みそのものの解明にほかならない。神の創造物たる人体が、A・T・G・Cの4種類の塩基の組み合わせによる、言わばきわめてデジタル的な構造を成していることが明らかになるにつれ、多くの企業や国家が総力をあげてヒトゲノム情報の解読に挑戦した。このヒトゲノム解読のプロセスに欠かせない存在がIT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



ゲノム断片をレーザー光で読み取り、色の連続として測定した塩基。

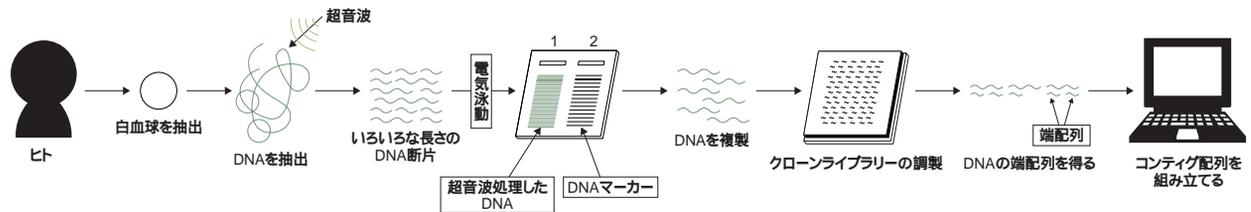
ヒトゲノムプロジェクト

ゲノム研究の第一歩は、兎にも角にもまずは塩基配列をすべて読み取ることにある。それを目的に1991年に米国の国家プロジェクトとして始まった「ヒトゲノムプロジェクト」はその後、英、仏、独などの欧州および日本が手を組む国際プロジェクトに発展した。当初、プロジェクトはヒトゲノム配列を“端から端まですべて”読み取ろうとしていたが、プロジェクト進行中に、米国の米セレーラジェノミクス社が、ゲノム中で“遺伝子が存在する部分だけ”を読み取るという独自の戦略を展開した。こうして遺伝子配列の特許をめぐり、公的機関と私企業が熾烈な競争を行う事態になった。

2000年秋にヒトゲノムプロジェクトとセレーラ社が和解し、米クリントン大統領と英ブレア首相、セレーラ社のベンター社長の3者が配列解読終了を「前倒し」で共同宣言した。

しかし、ここで解読されたのはあくまでもかなりの空白部分を含んだドラフト配列のみであり、本来の目標だった完全解読は2003～2004年頃に達成される予定だ。なお、このヒトゲノムプロジェクトの公的成果はインターネット上で公開されている。

ヒトゲノムの塩基配列の調べ方(ショットガン法)



ゲノム配列の読み取りと遺伝子の検出

具体的にゲノムはどのように読まれるのか？ここでは「ショットガン法」という方法を紹介しよう(上図)。まず、数人のヒトの白血球から抽出して混合したDNA試料を用意する。それを超音波で裁断し、それらをクローニング(複製)して、ゲノムシーケンサーという装置にかけて塩基配列を読み取る。シーケンサーは、A・T・G・Cの各塩基に対応して蛍光色素で4色に色付けられたゲノム断片を1ベース(b)ごとに電気泳動という方法で分離して、レーザー光で色を読み取り、色の波の連続として塩基を測定する(左ページグラフ)。シーケンサーでは膨大な長さのゲノムを一度に読み取れないので、あとで多くの断片を正しい順序でつなぎ合わせる作業が必要になる。前工程のクローニングによって隣のゲノム断片と前後が一部重なる断片が用意してあるので、データの重なりをデジタル的に探し出すことはできる。このゲノムコンティングにIT技術が適しているのは言うまでもない。

このようにDNA分子を分子生物学的手法で解析して作成された塩基配列は、「物理地図」(Physical Map)と呼ばれる。染色体の「地図」には、このほかに「マーカー

」と呼ばれるCACACA.....などのわかりやすい塩基並びの部分を手がかりに、遺伝学的手法で染色体上の位置関係を割り出した「遺伝学的地図」(Genetic Map)がある。最終的にはこの2つの地図は統合されて「コンティグ地図」(Contig Map)となる。ヒトゲノムプロジェクトの最終目的はヒトのコンティグ地図を完成させることだが、この気の遠くなるような作業の積み重ねを経て、約10年でこの偉業を達成しようとしているのだ。

こうして読み取られたゲノムデータには、遺伝子が埋め込まれている。遺伝子はゲノムの約3パーセント、平均約100キロベースに1個の割合でしか含まれていないため、ゲノムデータ中の遺伝子を見付け出す作業は、大海原の中で島を発見するようなものだ。他の部分はジャンク配列と呼ばれる無意味な繰り返しや、死んだ遺伝子(偽遺伝子)が存在する。これらは進化の過程で使われてきたものと考えられるが、どのような働きをしていたのかは不明である。

次に、莫大なゲノムデータから遺伝子を見付け出す方法を解説しよう。遺伝子の塩基配列の始まりは、多くの場合ATGの順番で並ぶという特徴を持つ。これを「開始コドン」という。一方、終わりはTAA、

ベース.....塩基を意味する英単語「base」から、1本鎖DNAを数える単位としても使われる。なお、2本鎖DNAを数えるときの単位はベースペア(bp)。

参考URL
ゲノムアセンブリデータベース:
GenBank www.ncbi.nlm.nih.gov/genome/guide/human/
米国GenBank www.ncbi.nlm.nih.gov/GenBank/
ヨーロッパEMBL www.ebi.ac.uk/embl/
日本DDBJ www.ddbj.nig.ac.jp

エンジニアのための「ゲノム情報学」の招待

遺伝コード(遺伝暗号)と対応するアミノ酸

TTT } フェニルアラニン	TCT } セリン	TAT } チロシン	TGT } システイン
TTC } フェニルアラニン	TCC } セリン	TAC } チロシン	TGC } システイン
TTA } フェニルアラニン	TCA } セリン	TAA } 終止	TGA } 終止
TTG } フェニルアラニン	TCG } セリン	TAG } 終止	TGG } トリプトファン
CTT } ロイシン	CCT } プロリン	CAT } ヒスチジン	CGT } アルギニン
CTC } ロイシン	CCC } プロリン	CAC } ヒスチジン	CGC } アルギニン
CTA } ロイシン	CCA } プロリン	CAA } グルタミン	CGA } アルギニン
CTG } ロイシン	CCG } プロリン	CAG } グルタミン	CGG } アルギニン
ATT } イソロイシン	ACT } スレオニン	AAT } アスパラギン	AGT } セリン
ATC } イソロイシン	ACC } スレオニン	AAC } アスパラギン	AGC } セリン
ATA } メチオニン	ACA } スレオニン	AAA } アスパラギン	AGA } アルギニン
ATG } メチオニン	ACG } スレオニン	AAG } アスパラギン	AGG } アルギニン
GTT } バリン	GCT } アラニン	GAT } アスパラギン酸	GGT } グリシン
GTC } バリン	GCC } アラニン	GAC } アスパラギン酸	GGC } グリシン
GTA } バリン	GCA } アラニン	GAA } グルタミン酸	GGA } グリシン
GTG } バリン	GCG } アラニン	GAG } グルタミン酸	GGG } グリシン

TAG、TGAのいずれかだ(終止コドン)、3塩基のトリプレットからなるコドンは、このように開始や終止を表すコードになるだけでなく、合成するアミノ酸の種類をも指令する。4種類の塩基を2ビット文字と考えると、ゲノム配列の中から、開始コドンで始まって終止コドンで終わる文字列を探す作業になる。これもITがもっとも得意とするタイプの作業である。生命があまりにもデジタルでできていて驚くばかりだ。

しかし、開始コドンで始まって終止コドンで終わる文字列を見つけても、それが必ずしも遺伝子であるとは限らない。もし塩基配列が完全にランダムであれば、それぞれの終止コドンは確率的に64コドンに一度の割合で出現するはずだが、本当の遺伝子はずっと長い。そのため、文字列が本物の遺伝子かどうかはふるいにかけて判定する。このほか遺伝子の検索には多くのアルゴリズムも存在するが、現在のアルゴリズムには90パーセント程度の信頼性しかなく、こちらも最終的には実際の実験(ウェット実験)で確認する必要がある。コンピュータの予測が正しいかどうかを伝統的な実験で確認できるところも、IT技術者にとっては嬉しいことである。

たんぱく質への翻訳

セントラルドグマによって遺伝子はRNAに転写され、そしてたんぱく質(アミ

ノ酸からなる重合体物質)に翻訳される。遺伝子の中にあるたんぱく質を指令する部分はORF(Open Reading Frame、開いた読み枠)と呼ばれ、開始コドンで始まって終止コドンで終わる性質を持ち、その間のコドンの並びによって合成されるアミノ酸が決まる。また、遺伝子の読み枠と読み枠の間にはイントロンという挿入配列があるため、各遺伝子は連続していない。

塩基にはA・T・C・Gの4種類あるので、コドンは $4^3 = 64$ 種類という計算になる。そこから3つの終止コドンを引いた61種類のコドンが20種類のアミノ酸を指定する(上図)。このようにして指定されたアミノ酸が結び付いてたんぱく質が合成される。

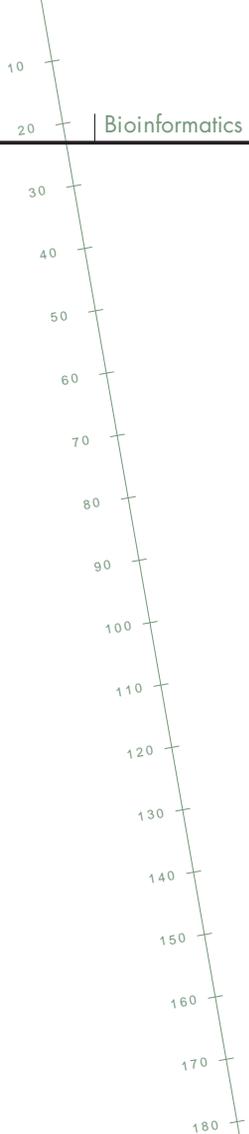
ゲノム配列からゲノムアノテーションへ

ゲノム配列決定後を「ポストゲノムシーケンス」(略してポストゲノム)と呼ぶが、ポストゲノムにまず達成しなければならない課題は、遺伝子の位置と構造の決定、遺伝子の機能・医学生物学的意味といった「ゲノムアノテーション」(注釈付け)である。遺伝子の構造機能が明らかになり、その配列の個人による違い(Single Nucleotide Polymorphism、SNPs、「スニップス」と呼ぶ)と疾患の関係、薬剤の対象としての意義が明らかになれば、特許を取った配列には高価な商品としての価値が発生することは間違いない。

代表的なアルゴリズムとしてはGeneFinder、Genieなどがある。
www.bioscience.org/urlists/genefind.htm
genome.cse.ucsc.edu/goldenPath/hgTracks.html

セントラルドグマ……「ゲノムに含まれる遺伝子がRNAという酵素に転写され、それがたんぱく質に翻訳されて、そのたんぱく質が生体を動かす」という遺伝情報の流れ。

参考文献:『ヒトの分子遺伝学』
Tom Strachan, Andrew P. Read(著)/村松正實(訳)





[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の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