綜　　説

大量配列情報からの知識抽出

奥　田　修二郎
新潟大学大学院医歯学総合研究科
バイオインフォーマティクス分野

Knowledge Elucidation from Large - scale Sequence Information

Shujiro OKUDA

Niigata University Graduate School of Medical and Dental Sciences

要　旨

世界で初めて全ゲノムが解読された生物種はインフルエンザ菌で1995年のことである．現在では，2000生物種以上のゲノム配列が公開されていることに加え，高速シーケンサーの登場により，ますます多くの配列情報を扱う必要が生じている．大量配列時代において，生命医科学分野における計算機の役割は非常に大きくなりつつある．ゲノムからメタゲノムに至るまで，様々な切り口で生命システムを解析するバイオインフォマティクスについて紹介する．

キーワード：バイオインフォマティクス，ゲノム，メタゲノム，次世代シーケンサー

はじめに

ワトソンとクリックによってDNAの二重らせん構造に関する論文がNature誌に発表されたのは，今からちょうど60年前の1953年である．それから現在に至るまでに，データベースに登録される遺伝子配列の数は，1億7千万配列（2013年12月現在）にまで膨れ上がり，新規のシーケンシング技術の発展とともに，今後ますます大量のDNA・遺伝子配列の情報が世に送り出されてくる時代になる．情報処理を駆使しない限り必要なデータを得ることが出来ないほどにデータ量が多い中，如何に大量の情報の中から効率的に知識を抽出するかという命題がかせられたのが，バイオインフォマティクスという分野と言ってもよい．ここでは，大量配列情報からの知識発見という観点から，バイオインフォマティクスがこれまでに何を実現し，今後どのように発展していくかについて議論してみたい．

Reprint requests to: Shujiro OKUDA
Niigata University Graduate School of Medical and Dental Sciences,
1 - 757 Asahimachi - dori, Chuo - ku,
Niigata 951 - 8510, Japan.

別刷請求先：〒951 - 8510　新潟市中央区旭町通1 - 757
新潟大学大学院医歯学総合研究科
バイオインフォーマティクス分野　　奥田修二郎

ゲノム配列からの知識抽出

ゲノム配列が最初に決められたのは，バクテリアの一種であるインフルエンザ菌（*Haemophilus influenzae*）[1]である．約180万塩基のゲノムDNAの中に，1,600ほどの遺伝子がコードされている．インフルエンザ菌ゲノム決定の約2年後（1997年）には，大腸菌のゲノム配列[2]が発表され，460万塩基のゲノム中に約4,000の遺伝子が存在することが判明した．これらのゲノム解析においては，「遺伝子はその生物を構成する部品であり，その生物のすべての部品としての遺伝子（ゲノム）が判明することで，様々な生命現象が解明される」と期待された．単なるDNA配列の中で，どの領域が遺伝子をコードする部分かを同定するための遺伝子探索技術の開発[3]，その遺伝子配列がどのような機能を担っているかを配列の相同性から推測する技術[4]など，ゲノム配列の解析にとって必要な技術が開発されており，現在でも日進月歩で進歩を続けている．しかしながら，大腸菌の場合，同定された4,000個もの遺伝子のほぼ半数は機能を推測することすら出来ない遺伝子だということもわかった．つまり，ゲノム配列を決めたことにより，何をしているか解らない遺伝子が増えたのである．こういった背景もあり，機能未知遺伝子の機能推定という問題は，バイオインフォマティクス分野において活発に議論されたテーマの一つであった．その中でも最もわかりやすくよく利用された技術が，いわゆるゲノムコンテキスト法と呼ばれるものであり，ゲノム上での遺伝子の並び方の情報を利用する方法である．原核生物の場合，同じ代謝系に関与する遺伝子群がゲノム上で隣接してコードされ，同一mRNAに転写されるというオペロンと呼ばれる転写システムが利用されている．つまり，ゲノム上でオペロンのように近接にコードされた遺伝子群は，同じ系で機能する可能性が高いということを遺伝子の機能予測に利用する方法である．また，複数の遺伝子が進化の過程で融合し一つの遺伝子になる現象を利用する方法（ロゼッタストーン法[5]）や，同じ遺伝子（オーソログ）を持つ生物種のパターン（系統プロファイル）が似た遺伝子同士は，相互作用の関係を持ちやすい[6]と言ったような考えを駆使し，遺伝子の機能推定の精度向上が図られた．これらは，遺伝子というピースを如何にしてゲノムという形に仕上げるかというパズルを解いているようなもので，計算機科学が生命医科学分野に入り込むのに非常に都合のよいテーマであったと言える．

パスウェイからの知識抽出

仮にすべての遺伝子の機能推定ができた場合，それであらゆる生命現象を説明できるだろうか．実際の生命は，複雑なタンパク質間の相互作用によって成り立っているため，遺伝子単位での機能だけが理解できても，システムとしての生命を理解できたことにはならない．そこで，タンパク質間相互作用を対象に解析する必要が出てくる．いわゆる代謝系は，よく調べられた相互作用であり，様々な機関によってデータベース化が行われている．その一つであるKEGG（Kyoto Encyclopedia of Genes and Genomes）[7]というデータベースでは，遺伝子・ゲノムだけでなく，それらの関係を表すパスウェイなどの情報がデータベース化されている．このKEGGを利用すると，ゲノム配列を決定した後，どの遺伝子がどのような機能を持つか，さらにそれらの遺伝子がどのような代謝系を担うかを推定することが出来る．ゲノム中の複数の遺伝子群が同じパスウェイ上で連続して酵素反応を担っていることがわかれば，その反応系を持つと推測できることになる．あらゆる代謝系の情報をつなぎあわせたグローバルなパスウェイマップを利用すると，あるゲノムが持つ反応をすべて俯瞰して解析することも可能である．そのためのツールとしてKEGG Atlas[8]やiPath[9]などが開発されている．最近ではパスウェイよりさらに意味がはっきりしたより細かな反応単位としてのモジュールのデータベースも作成されており，ある生物種のゲノムが決まった場合に，迅速にパスウェイ・モジュールレベルでの細胞機能の推定が可能である．パスウェイを解析することで，単なる

1遺伝子の機能ではなく，細胞としての機能を理解できることになる．

配列解析のパラダイムシフト

2000年代中盤になって，DNAシーケンサーの世界に大きな変革が起きた．いわゆる「次世代シーケンサー」と呼ばれるDNAシーケンサーの販売が開始されたのである．次世代と呼ばれる所以は，配列決定の方法論がそれまでと異なることはもちろんであるが，出力されるDNA配列の量という点で完全にそれまでと次元が異なっていたのである．一度のランで得られる配列長は最大でも数千塩基という状況だったものが，突然，数M（メガ）あるいは数G（ギガ）塩基という状況になり，完全に配列解析の方法が変わってしまった．これらの次世代シーケンサーの最大の特徴は，出力する大量のDNA配列が数十から数百塩基程度の短い断片（リード）であるということである．大量のショートリードは，その相同性領域を探しつなぎ合わせる（アセンブル），あるいは，相同性領域をリファレンスのゲノム配列にマッピングするなどして解析に用いる．これらの過程はすべて計算機で処理する以外には対処できないくらいの大量情報が得られる時代になったのである．ある程度安価で，大量のDNA配列情報を解析できることから，この技術の応用範囲は非常に広く，医学・生物学分野全般にわたって，新たな分野を創出しつつある．例えば，1,000人分のヒトゲノムを決定し，そのゲノム配列の多様性をすべて調べてしまう1,000人ゲノムプロジェクト[10]や，遺伝病の原因遺伝子探索では，疾患を持つ家系全員のゲノム・エキソーム領域をシーケンス[11]すれば，必ず原因の遺伝子変異が見つかるはずだ，という具合に力技で解析できるようになったのである．このように大量のDNA配列が得られるようになったことで，DNA配列解析の世界ではパラダイムシフトが起きているのである．

メタゲノム解析

次世代シーケンサーによるDNA配列解析がもたらした新たな局面は，ヒトのゲノムに限ったものではなく，全く別の世界でもパラダイムシフトを起こしていた．それは，環境微生物を対象にした生態系の解析の分野であり，土壌や海洋だけでなく，ヒトに関係する部分で言うと腸内環境などの微生物叢も含まれる．これらの環境中に棲息する微生物は，極めて難培養なものが多く，そもそもどのような微生物が棲息しているのか把握することが困難であった．16S rRNAは様々な細菌でよく保存された遺伝子であることから，その配列だけを環境から取り出し種の同定に利用するなどして，環境微生物の解析は行われていた．しかしながら，ある環境にどのような種がいるかということ以上に理解することが困難であった．そこで，次世代シーケンサーを利用して，環境中の微生物が持つゲノムDNAをまるごとシーケンスするというメタゲノム解析が行われるようになった．とりわけ，ヒト腸内細菌は宿主である人の代謝系と密接に関係しており，病気や健康に影響を与えていることが示唆されていたため，ヒト腸内細菌を対象にしたメタゲノム解析は，世界中で注目されている．筆者が国際共同研究として参加したプロジェクト[12]では，2型糖尿病患者の腸内細菌叢のメタゲノム解析を行った．その結果，2型糖尿病の患者と健康な人の腸内細菌は，全く異なるコミュニティを形成していることが判明した．また，腸内細菌が持つ糖尿病特異的なマーカー配列を利用すると，糖尿病かどうかをよく判別できるということもわかり，2型糖尿病の治療・予防に貢献しうるということを示すことにつながった．現在では，腸内環境があらゆる疾患との関連で解析されるようになっており，今後ますます腸内細菌叢の研究が進展するものと思われる．

おわりに

これからのDNAシーケンサーは，より長いリードを精確に安価に決定できるものへと進化して

いくようである．数万から数十万塩基の超ロングリードのシーケンスが実現される日もそれほど遠くないと見積もられている．大量のショートリードの処理が大規模な情報処理を必要としていたのに対して，1本の長いDNA配列が出力されるようになれば，そのような情報処理は必要なくなることになる．つまり，来るべき超ロングリードの時代においては，バイオインフォマティクスの役割は，次のステップに移らないといけないことを意味している．しかしながら，そもそもバイオインフォマティクスという分野は，生命医科学分野において出力されるDNAを単に処理するという分野ではない．大量のDNAやタンパク質の情報の渦の中から医学・生物学にとって意味のある知識の抽出・整理から生命原理を追求するための学問分野である．生き物を使った実験ができないような壮大な世界を定式化することが求められる分野でもある．来るべき新たな時代においては，そういった大規模データから世界の誰も知り得ないような新しい物の見方を提案できるような仕組みを今から模索する必要がある．

参考文献

1）Smith HO, Tomb JF, Dougherty BA, Fleischmann RD and Venter JC: Frequency and distribution of DNA uptake signal sequences in the Haemophilus influenzae Rd genome. *Science* 269: 538－540, 1995.

2）Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado－Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B and Shao Y: The complete genome sequence of Escherichia coli K－12. *Science* 277: 1453－1462, 1997.

3）Salzberg SL, Delcher AL, Kasif S and White O: Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* 26: 544－548, 1998.

4）Moriya Y, Itoh M, Okuda S, Yoshizawa AC and Kanehisa M: KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35: W182－185, 2007.

5）Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D: Detecting protein function and protein－protein interactions from genome sequences. *Science* 285: 751－753, 1999.

6）Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96: 4285－4288, 1999.

7）Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T and Yamanishi Y: KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36: D480－484, 2008.

8）Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S and Kanehisa M: KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res* 36: W423－426, 2008.

9）Yamada T, Letunic I, Okuda S, Kanehisa M and Bork P: iPath2.0: interactive pathway explorer. *Nucleic Acids Res* 39: W412－415, 2011.

10）Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME and McVean GA: A map of human genome variation from population－scale sequencing. *Nature* 467: 1061－1073, 2010.

11）Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA and Shendure J: Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12: 745－755, 2011.

12）Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto JM, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K and

Wang J: A metagenome - wide association study of gut microbiota in type 2 diabetes. *Nature* 490: 55 - 60, 2012.