ゲノムの中にあらわれた待ち行列

豊泉　洋

バイオインフォマティックスを初めて学んだ待ち行列研究者の苦闘とゲノム情報の解析に待ち行列を使った遺伝子解析を解説する．

キーワード：遺伝子工学，相同性検索，遺伝子発見，待ち行列，Lindley equation

1．論より証拠

論より証拠である．ゲノムの中に待ち行列が現れるなんて信じられないという人もいるだろうが，図1を見ていただきたい．待ち行列理論を学んだ人には，見慣れた待ち時間の時間変化の図だ．これが，インターネット上のパケットの遅延時間や工場の製造工程での遅れを表しているのであれば不思議なことは何もない．

実はこのデータ，大腸菌O-157のアミノ酸配列[1]を元にしている．酸性のアミノ酸が蓄積されている部分に大きな「待ち時間」が当たるように簡単な細工を施している．単純にデータを待ち時間に置き換えるだけなら，工夫次第で図1のように見せることもたやすい．しかし，図1では，驚くべきことに，待ち時間が大きく蓄積している期間（長い全稼働期間）が大腸菌のタンパク質をコードしている部分に相当している．このデータから，表1のような遺伝子[2]を特定できるのである．

なぜこのようなことが可能なのか，疑問を持たれた方は，研究の経緯を説明しながらの謎解きにおつきあい願いたい．

2．きっかけ

以前から，ゲノムには興味があった．それは，人間が本来持つ，「自分とは何か知りたい」という単純な欲求からくるものであったかもしれない．それとも，「ヒトの遺伝子情報解読へ」，「遺伝子情報により新しい治療の可能性」，「生命の秘密が解き明かされる」といったセンセーショナルな見出しのせいかもしれない．

2000年から2001年にかけて，新聞やテレビのニュースの記事は，生物に関する基礎知識のない私にとっては，別世界のように思えた．

記事の中で，国際コンソーシアムのヒトゲノムプロジェクト[2]とセレーラ社[4]とのヒトゲノム解析競争の記事が目を引いた．セレーラ社の手法は「ショットガン法」と呼ばれる方法を使っていると解説されていた．後になって，セレーラ社の使っている方法は，ホールゲノムショットガン法と呼ばれるべきであり，ヒトゲノムプロジェクトもショットガン法を使っているということがわかった．とにかく，このときには，その「ショットガン法」という名前に魅了されていたのである．

ショットガン法，その名前からは「数打てば当たる」式の図を想像してしまう．私は，待ち行列や応用確率論の研究者である．「数打てば当たる」という論法であれば，確率論が使えて，自分にもわかるような理論かもしれないという淡い幻想を抱いたのである．ちなみに，以下ではショットガン法の研究については述べないが，ショットガン法も応用確率論の研究者の出番がある面白い研究対象である[9, 10]．

これが，その後3年をかけて，私の研究室の学生たちと一緒に待ち行列の研究者がバイオインフォマティクスの研究を始めるきっかけとなった．

3．待ち行列学者の初めてのバイオインフォマティクス

簡単に「ショットガン法」でヒトの遺伝子情報を解析するといっても，生物学の素人には，どこから手をつけていいのかわからない．手元にセレーラ社とヒトゲノムプロジェクトの遺伝子解析の論文[2, 4]を手に

とよいずみ　ひろし
会津大学
〒965-8580 会津若松市一箕町鶴賀

[1] ここで使われているのは，大阪堺市で大規模な食中毒を起こした悪名の高い大腸菌O-157のゲノムデータである．

[2] 例えば，ECs 0081 は putative acetolactate synthase III large subunit というタンパク質に相当する．

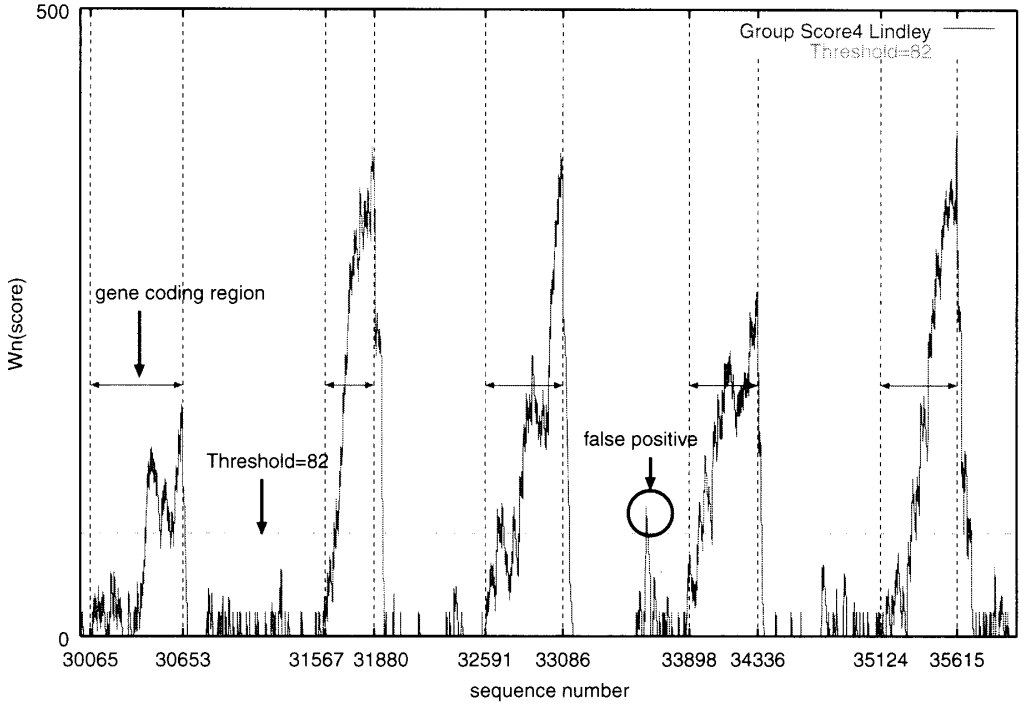

図1　アミノ酸配列にLindley equationを適用した結果：データのピークは酸性のアミノ酸が蓄積している部分に相当し，矢印（↔）が示す期間は，その部分にタンパク質がコードされている部分を表す[12]

表1　大腸菌O157：H7 Sakaiの遺伝子情報

コード名	長さ	開始	終了
ECs0081	1767	30065	30653
ECs0086	942	31567	31880
ECs0089	1488	32591	33086
ECs0092	1317	33898	34336
ECs0095	1476	35124	35615

入れても，書いてあることがわからない．「塩基」，「アミノ酸」，「RNA」，「たんぱく質」，「染色体」，「原核生物」，どれもこれも聞いたことはあるが，イメージと相互の関連性が頭に浮かばない．インターネットなどの通信ネットワーク関連の性能評価が主なフィールドで，その技術用語に慣れ親しんでいる身にとっては，まったく異質な言語，文化に放り込まれたような気がした．しかし言語というのは不思議なもので，様々な解説書や記事に目を通していくうちに，だんだんと言葉の概念が身に付いていくものである．

そうやって身につけた新しい言語でいうと，生物の遺伝子情報は，細胞内にある染色体上に4種類の塩基配列の形でコードされているということになる．これをメッセンジャRNAが転写し，三つの塩基の組から20種類のアミノ酸に解読し，最終的に，たんぱく質を合成する．異なるコードからは，異なるたんぱく質が生成される．これが最終的に，細胞の異なる機能の実現や種の違い，個体の違いに反映される[3]．

生物学の専門家ではない私にとって，塩基配列データは，4種類の塩基や20種類のアミノ酸が遺伝情報の列の形でコードされていると考えるのが一番わかりやすい．コードされた遺伝子情報は安定しているわけではなく，突然変異や交叉により，親から子に受け継がれるときに変異する．したがって，遺伝子情報は，文字列空間上を動く確率過程だと思うことができる．

バイオインフォマティクスに確率論を応用するというのは，わかりやすい解析アプローチであり，たくさんの研究者が取り組んでいる．その分野をわかりやす

[3] もう少し詳しくいえば，染色体はDNAと呼ばれる巨大な分子によって構成されている．塩基（nucleodite）と呼ばれる分子が二重らせん構造をとって，結合することによってDNAを形作っている．DNAを構成する塩基は4種類あり，それぞれアデニン（A），チミン（T），グアニン（G），シトシン（C）と呼ばれる．これらの塩基が三つ一組となり，20種類のアミノ酸および遺伝子の読み出し開始，終了のマークなどをコーディングしている．例えば，「ATG」という配列は，アミノ酸の一種である「メチオニン（M）」に相当する，またこのメチオニンはRNAが転写を開始するマークにもなっている．DNA上に三つの塩基としてコードされたアミノ酸の情報は，メッセンジャRNAに転写され，アミノ酸を部品として作った巨大分子であるたんぱく質を作り出すのに利用される．

く説明している参考書もいくつか出版されている[3, 9]．まずは，そのような専門書を読み進んでいくことになる．すると，バイオインフォマティクスでは，遺伝子の機能を同定するために，相同性検索が大事で，そこに，音声などの情報処理によく使われている確率的な手法である隠れマルコフ法[3]が盛んに使われていることがわかった．「ショットガン法」への未練は残るが，応用確率論の研究者として「マルコフ」という言葉に条件反射してしまった．

4. 相同性検索とは？

染色体上にコードされた文字列がわかっただけでは不充分である．その文字列がどんな生物学的機能を実現するのかを知ることが重要である．そのための一つの方法として相同性検索がある．

相同性検索とは，膨大な遺伝子配列データの中から，遺伝子の機能として似たものを選びだすという検索技術である．その例を石川・金久[5]p. 65の例にならって示そう（表2参照）．上段の6本の文字列は，1列目が未知の配列で，後の5本がそれぞれ別の種類のレトロウィルスのある酵素が持つアミノ酸の配列[4]である．今，この未知の配列unknownが他の五つの配列と似たような遺伝子機能をコーディングしたものかを判断したいとする．同じアミノ酸や性質の似たアミノ酸が縦に同じ位置になるように，gap（ここでは，「—」で表している）を入れ調整する．アライメントと呼ばれる処理である．縦方向に揃った文字で構成される配列を，そのアライメントのコンセンサス配列（consensus sequence）と呼ぶ（表2の下段参照）．

ここでは，Hで示されるヒスチジンというアミノ酸2個と，Cで示されるシステイン2個が縦に揃っており，コンセンサス配列と見ることができる．また，コンセンサス配列のなかに見られるパターンが，アライメントされた配列群の遺伝的機能を特徴づけるものと判断できるとき，そのパターンを，配列モチーフ（sequencemotif）とか，単にモチーフと呼ぶ．Cが2個とHが2個で特徴づけられるパターンは，「亜鉛の指」（ジンクフィンガー，zinc finger）と呼ばれる有名なモチーフである．したがって，生物学的機能には，このunknownの配列は，他のレトロウィルスの酵素と同じ性質を持つであろうことが予想される．しかし，このようなunknownの配列がまったく偶然に他の五つの酵素の配列と同じ文字列を持つ可能性もある．この可能性を定量的に評価するために確率論的な考え方が必要となる．

5. 確率過程と待ち行列研究者の壮絶な戦い

確率論や確率過程論の特異な一分野として，待ち行列理論は位置する．確率論を応用して，システムの性能を評価する手法を研究するという分野である．その応用先は，伝統的な電話ネットワークの設計だけに留まらず，最近はインターネットやセキュリティ，さらにはファイナンスの理論にまでもその適用範囲を広げてきている．

待ち行列理論の歴史は，システムへの入力である到着過程との戦いの歴史でもある．歴史的にみれば，Poisson過程や独立な指数分布の列のように比較的穏やかで，取り扱いのやさしい確率過程を入力としたシステムの解析が最初に行われた．有名なErlang-B式やJackson networkの積形式解などは，Poisson過程やMarkov性を利用することにより，容易に解析ができることを利用して導出されている．

しかし，システムへの入力は気まぐれである．いつもPoisson過程になるとは限らない．もし，システムへの入力がPoisson過程と異なっても，待ち時間や呼損率などの性能評価指標が大きく異ならなければ，Poisson過程が入力の場合だけ考えればよい．しかし，残念ながら，Poisson過程とは大きく異なる結果をもたらす確率過程の存在がよく知られている．インターネットのトラヒックデータがそのような性質の悪い長距離相関を持つ可能性も指摘されている．

待ち行列の研究者は，Poisson過程とは異なる入力を持つシステムを扱うために，様々な理論，テクニックを磨いて，性能評価を行ってきた．到着過程の背後に位相を考えることによりMarkov性を持ち込むMatrix-geometric法，サンプルパスを用いた解析法，より一般的な確率過程で成立することを厳密に取り扱う枠組みを与えてくれる点過程法，異なる入力過程の関係を明らかにする確率順序，さらに，厳密な取り扱いが難しい場合に，性能評価指標の近似値を与えてくれる拡散近似，大偏差原理など，多種多様な手法が考案されてきている．

逆にいえば，これらの壮絶な戦いが必要なほど，待

[4] レトロウィルスは自分の遺伝子情報を宿主のDNAに刷り込んで増殖する．この例の酵素はエンドヌクレアーゼと呼ばれるもので，DNAを切る働きをする．また，例えばHTLVはヒトT細胞白血病ウィルスである．

表2　アミノ酸配列の相同性検索の例

unknown : ILDFHEKLLHPGIQKTTKLFGETYYFPNSQLLIQNIINECSICNLAK
MMULV : LLDFLLHQLTHLSFSKMKALLERSHSPYYMLNRDRTLKNITETCKACAQVN
HTLV : LQLSPAELHSFTHCGQTALTLQGATTTEASNILRSCHACRGGN
RSV : YPLREAKDLHTALHIGPRALSKACNISMQQAREVVQTCPHCNSA
MMTV : IHEATQAHTLHHLNAHTLRLLYKITREQARDIVKACKQCVVAT
SMRV : LESAQESHALHHQNAAALRFQFHITREQAREIVKLCPNCPDWGS

unknown : IL-DF----HEKLLHPGIQKTTK-LF--GET-YY-FPNSQLLIQNIINECSICNL-AK
M－MULV : LL-DFL--LHQ-LTHLSFSKM-KALLERSHSPYYMLNRDRTL-KNITETCKACAQ-VN
HTLV : LQLSPA-ELHS-FTHCGQTAL-T-LQ--------GATTTEA--SNILRSCHACRG-GN
RSV : YPLREAKDLHT-ALHIGPRAL-S-KA-------CNISMQQA--REVVQTCPHC-N-SA
MMTV : IH-EAT-QAHT-LHHLNAHTL-R-LL-------YKITREQA--RDIVKACKQCVV-AT
SMRV : LE-SAQ-ESHA-LHHQNAAAL-R-FQ-------FHITREQA--REIVKLCPNCPDWGS
Consensus : ---------H----H---------------------------------C--C-----

ち行列システムは，入力となる確率過程に敏感に反応し，異なる様相を我々に見せてくれる．特に，周辺分布などの局所的な性質や自己相関関数などの大域的性質の双方に依存して，待ち時間分布は様々に変化する[5]．ということは，待ち行列システムというのは，本質的に入力となるデータ（確率過程）の分類に使えると期待できる．

6．現れてきたLindley equation

さて，話を遺伝子に戻そう．

相同性検索では，機能が不明なアミノ酸配列（以下では，ターゲットと呼ぶ）に対して，膨大なアミノ酸配列のデータの中から，遺伝的に似た性質を持つ配列をより分ける必要がある．しかし，節4でみたように同一の生物学的機能を実現する配列であっても，突然変異などの影響によって，局所的には異なる配列となる可能性がある．通常は，ターゲット配列上のあるアミノ酸と対応する部分が似た性質を持つアミノ酸の場合には高得点を与え，異なる場合は低い得点を与えるといったスコアリング法を使い，配列全体で高得点となるアミノ酸配列を探索し，未知の配列の機能を同定する．

しかし，未知のアミノ酸配列では，配列中のどの部分に遺伝機能が仕組まれているかわからない．したがって，ターゲットの配列上で任意の部分のスコアの合計値を求め，その最大値を探索し，これが高い場合には，ターゲットの該当する部分と似た遺伝子が発見できたことになる．配列上の各部分でのスコアを加算し，その最大値を求めるということは，実は，Loynes variableと呼ばれ，待ち行列の世界ではおなじみのLindley equation[6]という待ち時間を計算する漸化式によって求められる値と等価であることが知られている（例えば，文献[7]参照）．すなわち，相同性検索は，待ち時間を求め，評価する問題に帰着できる．実際に，待ち行列理論を用い，どの程度のスコアになればターゲットと同じ生物学的機能を持っているといえるかを定量的に求めることもできる[1, 6, 11, 12]．

7．Lindley equationで遺伝子発見ができるのか？

配列中の特定の部分と類似した遺伝子の探索（相同性検索）に待ち行列理論が使えることがわかった．しかし，節5でも見たように，待ち行列は，データの解析にもっと幅広く使える可能性がある．待ち行列は，入力データの局所的な性質や大域的な性質を待ち時間の形で我々に提示してくれる．遺伝子が配列データの中に埋まっていたとしても，遺伝子部分が他の部分と異なる性質を持てば，対応する待ち時間の違いを使って，埋もれた遺伝子そのものを発見できる．この予想に基づき，実際に遺伝子検出に待ち行列を使ってみた

[5] 例えば，到着間隔の平均が大きくなれば待ち時間は当然小さくなるが，同じ平均到着間隔であっても，大きな到着間隔の後に大きな到着間隔，小さな到着間隔の後に小さな到着間隔が起こりやすいとき（独立であるよりも自己相関が高い場合）に，一般的に大きな待ち時間となる．

[6] Lindley equationはそのシステムの待ち時間の変化をつかさどる，いわば，待ち行列理論のニュートン方程式に当たるものである．

表3　Escherichia coli O 157：H 7 Sakaiの配列上の各アミノ酸への配分スコア

性質	非極性	非電荷	塩基性	酸性	Stop
score	-2	-2	-3	20	-50

結果が図1だ．

図1は，表3のスコア配分に基づいて，大腸菌のアミノ酸配列のスコアをLindley equationで処理した結果である．酸性アミノ酸に特に大きなスコアを与え，その他のアミノ酸に負のスコアを与えている．また，タンパク質をコードしている配列は，明示的に「終わり」のマーク（stop codon）があるため，このstop codonにはさらに絶対値の大きな負の値を与えている．この結果，大腸菌のアミノ酸配列上で，特に大きな酸性アミノ酸が集積している部分をLindleye quationによってあぶりだすことができる．したがって，少なくとも酸性アミノ酸が特徴的に集積する遺伝子であれば，発見することができる．すなわち，待ち行列モデルを使って，遺伝子の特徴をとらえることに成功したのである．

8．最後に

今回は紙面の都合もあり，「ショットガン法」や「隠れマルコフ法による遺伝子発見」といった応用確率論屋が楽しく仕事ができるであろう場所を詳しく紹介することができなかったが，興味を持たれた方は，ぜひ，解説書を一読していただきたい．また，遺伝子発現の相互関係をネットワークの形で取り扱い，その確率的な振る舞いを調べるという研究も現れてきている[8]．

待ち行列の数学モデルは，その数学的な取り扱いの難しさと同時に豊かな表現力を内包している．この性質は，遺伝子工学だけではなく，もっと様々な現象の解析にも役に立ち，今後も我々の知的好奇心を満足させてくれることであろう．

謝辞　会津大学在学中にゲノムデータの解析に協力し，各種のデータを提供していただいた情報工房株式会社の土屋大介氏に感謝します．

参考文献

[1] S. F. Altschul: Amino acid substitution matrices from an information theoretic perspective, *Journal of Molecular Biology*, Vol. 219, pp. 555-565, 1991.

[2] The Genome International Sequencing Consortium: Initial sequencing and analysis of the human genome, *Nature*, Vol. 409, pp. 860-921, 2001.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison: *Biological sequence analysis*, Cambridge University Press, 1998.

[4] J. Craig Venter, et al.: The sequence of the human genome, *Science*, Vol. 291, pp. 1304-1351, 2001.

[5] Mikito Ishii and Hiroshi Kanehisa: Moji wo hikakushi naraberu, In *Human Genome Project and Knowledge Management*, Bifuukan, 1995.

[6] S. Karlin and S. Altschul: Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc. Natl. Acad. Sci. USA*, Vol. 87, pp. 2264-2268, 1990.

[7] L. Kleinrock: *Queueing Systems*, Vol. 1, John Wiley and Sons, 1975.

[8] J. Paulsson: Summing up the noise in gene networks, *Nature*, Vol. 427, pp. 415-418, 2004.

[9] J. C. Setubal and J. Meidanis: *Introduction to Computational Molecular Biology*, PWS Publishing, 1996.

[10] M. T. Tammi: The principles of shotgun sequencing and automated fragment assembly, http://web.cgb.ki.se/student/sfa.pdf, 2003.

[11] H. Toyoizumi and Y. Tanioka: An application of queueing theory to bioinfomatics, In *Proceedings of the queueing symposium*, 2003.

[12] D. Tsuchiya: Gene sequence analysis using a Lindley equation, *Technical report, U. of Aizu*, 2004.