ゲノム進化とマッピングの階層モデルと解析アルゴリズムの開発

東京大学大学院農学生命科学研究科
○岸野 洋久

Development of hierarchical models and statistical procedures for the inference of genome evolution and mapping genes

Hirohisa Kishino
Department of Agricultural and Environmental Biology
Graduate School of Agricultural and Life Sciences
University of Tokyo

We developed assorted statistical procedures for analysis of genetic data. Our hierarchical model of linkage analysis accounts for linkage phase uncertainty and population structure. To understand functional adaptation, we constructed a model relating protein structure to evolution. Conventional phylogeny inference procedures make systematic errors when sequences in different branches of an evolutionary tree change in a convergent fashion. Our novel procedure robustly handles convergent change. We also studied models in which variation over time in evolutionary rate is explicitly incorporated as a prior distribution. Using these models, co-evolution of genes is detected via correlated patterns of rate change and footprints of adaptive evolution via large rate changes. We applied techniques developed during this project to characterization of homoplasy in mammalian evolution, accelerated evolution in plants and drosophila, macroevolution of bacterial genomes, quasi-speciation of viruses, and fates of duplicated yeast genes.

1. はじめに

本プロジェクトは、ゲノムデータベースに内在する分布を階層モデルで記述し、ゲノムに刻まれた生物多様化と適応を高感度で検出するとともに、重要遺伝子をマッピングする手法を構築することを目的とした。階層モデルは、尤度に含まれるパラメータに分布を導入する。分布を記述する超パラメータについてさらに超事前分布を導入することにより、事前分布に対する不確実性を考慮に入れる。こうして、パラメータ、あるいはその分布を頑健に推定することが可能となる。

階層モデルに着目した背景には、申請者らが開発した、分子進化速度の確率変動モデルがあった。生物の多様化と適応進化の痕跡は、分子進化速度の変化となって露呈する(図1)。従って、分子進化のパターンと速度の変動を見ることにより、適応進化の痕跡を検出することが出来る。系統毎、遺伝子毎の分子進化速度はデータから推定される未知パラメータであるから、速度の変化や共進化はパラメータの間の分布や相関構造をモデリングした階層モデルを通して適切に表現される。階層モデルは適応進化の検出にとどまらず、遺伝子マッピングにも有効である。連鎖解析の階層モデルにより、linkage phase に対する不確実性を考慮に入れた解析が可能となり、相関解析の階層モデルでは分集団構造と連鎖不平衡を同時推定することが可能となる。

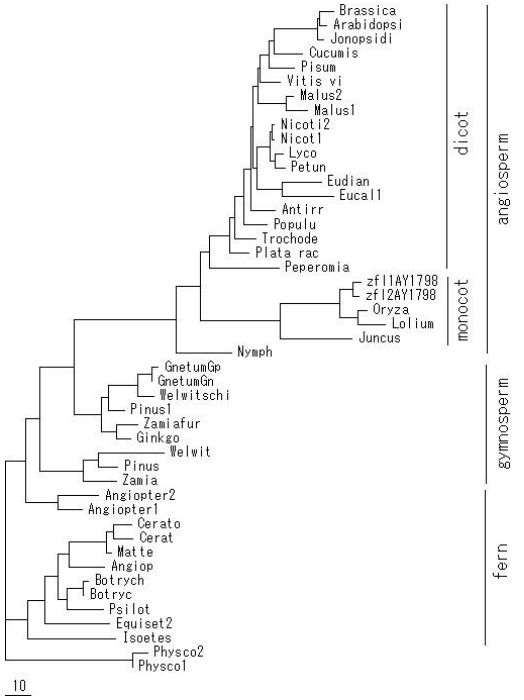

図 1　陸上植物における LFY 系統樹と被子植物へ到る系統、単子葉植物に至る系統での分子進化速度の加速化

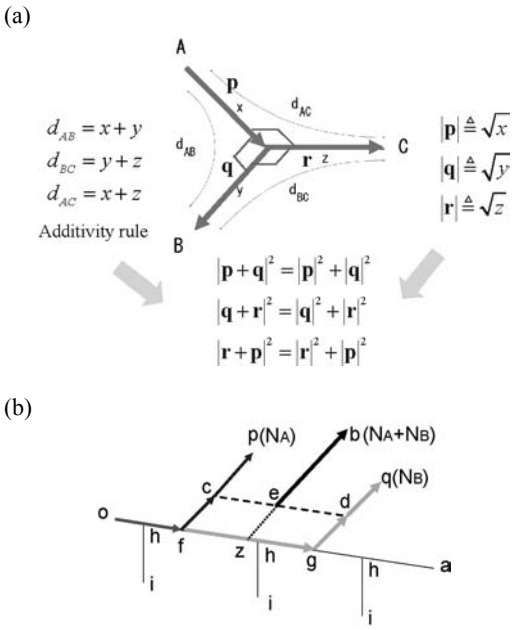

図 2　多次元ベクトル空間表現：(a)進化距離の加法則と進化ベクトルの直交性，(b) 組換えとベクトルの合成

2. 研究開発の成果概要

階層モデルを主軸として、上述のモデルに加え、遺伝子配列の解析のためのさまざまな統計モデルと解析手法を開発した。機能的な適応を理解するために、タンパク質進化に立体構造を関連づけたモデルを開発した。また、樹構造からの乖離として進化モデルの系統的な偏りを検出し、異なる系統間での収斂進化を考慮に入れて頑健に分子系統樹を頑健に推定する方法を開発した。開発された手法により、哺乳動物の進化における収斂進化、植物とハエにおける分子進化速度の加速、バクテリアゲノムのマクロ進化、ウイルスの準種分化、酵母における重複遺伝子の運命の非対称性を特徴づけた。

2.1　分子進化速度の確率変動と相関構造

分子系統樹の尤度は、各系統における期待置換数、すなわち進化速度と進化時間の積で表現される。進化速度に幾何ブラウン運動の事前分布を導入することにより、確率変動を記述する階層モデルを構築した。超パラメータから分子時計からのずれの大きさを測ることができる。複数の遺伝子配列に固有の超パラメータを用意することにより、進化速度の変動における相関の形で共進化を捉えることが可能となる[1-3]。また、同義置換・非同義置換の2変量速度変化モデルを通じて、背景要因を選択圧の変化、集団の大きさの変動、世代の長さの変動に分解することができる[4]。プログラムは *multidivtime*, *codonrates* として実現し、植物とハエ、哺乳類、あるいは広く多細胞生物における分子進化速度の加速の分析により、有効性が確認された[5-9]。

2.2　遺伝子重複とゲノム背景

ゲノム進化における遺伝子重複について、新機能化モデル、部分機能化モデルが提唱されている。し

かし、重複後いずれのコピーが機能を保持し、いずれが変異するか、予測するモデルは提案されていない。ショウジョウバエアミラーゼ多重遺伝子族における速度変化を観察する中で、我々は、ゲノム上局所的組換え率の低い部位に置かれたコピーが、変異を受ける傾向があるという仮説を提案した[10,11]。アミラーゼ遺伝子族の GC3含量の推移はこの仮説を支持していた。さらに、*Saccharomyces cerevisiae* ゲノムのパラログの解析からも仮説を支持する結果が得られた[12,13]。

2.3 立体構造制約下のタンパク質進化

機能的な制約から、アミノ酸を変える変異の多くは集団から排除される。タンパク質の構造を大きく変えず、関連するタンパク質との相互作用も損なわない突然変異は、集団に固定する可能性が高い。立体構造上の制約を受けながら変異を重ね、微細構造を変えて、活性やタンパク質の結合が進化していく。2アミノ酸残基間相互作用と親水性の2因子からなるポテンシャルを用いて、立体構造上の安定性を考慮に入れた遺伝子推移確率速度を記述した[14]。Annexin V, lysozyme では立体構造の制約は同程度に利いていることが確認された。後者においては正の淘汰圧が検出され、正の淘汰圧に関する情報と立体構造の制約の強さに関する情報は独立であることがわかった。

2.4 ウイルス遺伝子の系図と宿主適応、準種分化

遺伝子の系図に対する合体過程のモデリングを通して、擬似最尤法により、ウイルス分子進化速度と集団の大きさを同時推定する方法を開発した[15,16]。公開されたエイズ患者の潜伏期間における HIV-1 *env* 遺伝子の解析から、進化速度・多様化圧と集団サイズの間に負の相関関係を検出した。さらに知識ベースの構造予測により、gp120の V3領域における微細構造の変化を推定した。多次元尺度構成法により、感染後抗原決定基をはさむ領域の立体構造のダイナミックスを調べた。アミノ酸2残基の置換により感染時の構造と乖離した構造を生成したこと、多様性を維持しながら免疫系へ適応し、タイムリーに主タイプを移していく準種分化の実像を明らかにした。

2.5 バクテリアゲノムのマクロ進化

バクテリアの系統プロファイルからゲノムレパートリーの変化を辿ることを目的とし、遺伝子の得失モデルを開発した。COG の解析から、共生とともに遺伝子喪失速度が加速し、大量の遺伝子をふるい落としていることを確認した。また、得失の共時性を検出する統計手法を開発した。共時的な遺伝子の獲得による分類から、尿素に関わる遺伝子群、コバラミン生合成など、機能的に関連する遺伝子群が抽出された。同時に喪失した遺伝子の間には機能的な関連は見られず、水平伝播したゲノム断片は、それ自体で機能が完結することが求められていることが示唆された。

2.6 分子レベルの収斂進化

並行進化や収斂進化は適応進化の重要な要素であるが、他方で分子系統樹の推定に偏りをもたらす。本プロジェクトでは、配列を多次元空間内でベクトル表現することにより、系統間の相関と系統樹を同時推定する方法を開発した[17]。ピタゴラスの定理により加法則が直交関係に一意的に変換され（図2a）、直交性からのズレとして収斂進化が検出される。哺乳類のミトコンドリアゲノムおよび22の核遺伝子を解析

した結果、幾つかの系統間で強い相関が検出され、かつゲノムレベルの解析を行っても相関は解消しないことがわかった。先行研究における遺伝子の間の矛盾が消失し、哺乳類進化の歴史を新たな視点で捉え直した。また、ウイルスの適応には同時感染による組換えが重要であるが[18]、定量的な推定は困難であった。ベクトルの合成として感度よく検出する可能性が示唆された(図2b)。

2.7 他殖性生物の QTL 解析、連鎖不平衡と集団構造

連鎖解析においては、linkage phase が不可欠な情報であるが、純系親を生成するのが困難な他殖性生物ではこれを一意に特定できない。そこで、ハプロタイプを確率的に生成する遺伝アルゴリズムにより、他殖性生物の連鎖解析の手法を開発した。また、階層モデルにより分集団構造を記述し、地理的標本に基づいて、メタポピュレーションにおける遺伝的分化と連鎖不平衡の大きさを同時推定する方法を開発した[19]。標本の採られた局所地域では Hardy-Weinberg 平衡が成り立っていると仮定する。数値的シミュレーションによって古典的方法と比較しつつ本方法の性能を評価し、アユのアイソザイムと人類のALDH2遺伝子の SNPs のデータにより有効性を確認した。

2.8 データベース解析

植物宿主の内部に侵入し根の細胞の間を練り歩く根瘤線虫について、植物の細胞壁を破り進入することに関与する遺伝子はバクテリアから水平伝播したという仮説に立ち、ゲノムレベルの相同性検索により系統的にはじき出した。その結果既知の7つの水平伝播遺伝子に加え、新たに7つの遺伝子が検出された。これらはマメ科植物の根に侵入して根粒を作る根粒菌から来ていた[20]。比較ゲノムを通じた家畜家禽の有用遺伝子の探索、および品種間の系統関係を調べるために、反復配列の抽出、重複領域の抽出を行った[21]。QTL の遺伝的変異の多くは転写制御にかかわるものと考えられている。牛肉の脂肪交雑の形成と関係が深いレチノイン酸(ビタミン A)をリガンドとする RXR と2量体を形成する核内受容体クラス1について、ヒトおよびマウスの全ゲノムシーケンスから応答配列の検出を行った。既知の遺伝子を超えて多数の応答配列がゲノム中に存在することが明らかとなり、また、それらは転写制御領域となりうるとされる領域外に多く存在した。データベースを整備し、広範な研究者が個々の応答配列が近傍遺伝子を転写制御しているかどうか、in vitro で確認実験を行うための基礎的なデータを提供する。また、2428のタンパク質ファミリーを解析し、遺伝子ネットワーク上多くの遺伝子と関わり、遺伝子産物が細胞内に存在し、複雑な機能を持つ遺伝子は保存性が高く、正の淘汰を通じた適応進化を受けにくいことがわかった[22]。

3. まとめ

当初計画は2節における1, 3, 4, 5, 7, 8の5項目でカバーされており、ほぼ目的は達成されたと考えている。2節2および6の2項目はプロジェクト成立後の新展開である。方法論の開発と仮説形成、データベース解析に勢力を注ぎ、ソフトウエア・データベースは可能な部分から公開してきた。一部未完の部分も残され、この点は反省しなければならない。本プロジェクトの着想は、永年にわたり培ってきた Dr. Jeffrey Thorne との信頼関係によるところが大きく、本助成で国をまたいでさらに緊密な共同研究ができたことに何よりも感謝している。ゲノムデータベースを階層モデルで記述するというアイデアは、プロジェクト申請時点では先駆的であったが、いまでは世界の流行にまでなった。むしろ、ベイズの枠組みで方法論に縛ら

れている感がある。ただ私たちの目指すものはそこにはなく、階層モデルの考え方を駆使してゲノムの多様性を自在に描写することであった。新たに共同研究に参加いただいた方々の力で、当初想定していなかった新たな視角が導入され、膨らみのある成果が醸成された。分子レベルの収斂進化や組換えの検出、ウイルス集団の準種分化の推定など、モデルを超えた探索的アプローチをも積極的に開発してきた。重複遺伝子の運命をそれらのゲノムにおける環境が左右するという新たな仮説を提示した。全体として当初計画を超えた成果が得られたと考えている。貴重な出会いのきっかけを作ってくださったこのプログラムに、改めて感謝している。

4. 研究開発実施体制

代表研究者　岸野　洋久（東京大学大学院農学生命科学研究科）

研究開発題目

(1) ゲノム進化・マッピング階層モデル開発

グループリーダー　岸野洋久（東京大学大学院農学生命科学研究科）

(2) たんぱく質進化・ウイルス進化モデル開発

グループリーダー　Jeffrey L. Thorne（ノースカロライナ州立大学バイオインフォマティクス研究センター）

(3) QTL, 共進化データベース化

グループリーダー　和田康彦（佐賀大学農学部）

(4) 並行進化・収斂進化検出アルゴリズムの開発

グループリーダー　北添康弘（高知大学医学部附属医学情報センター）

5. 参考文献

[1] Kishino H, Thorne JL, Seo T-K, Kajitani Y (2003). Modeling of variable evolutionary rates to estimate divergence times and adaptive evolution. Proceedings of the Conference "Science of Modeling: the 30th Anniversary of Information Criterion (AIC)." Pp. 297-306.

[2] Thorne JL and Kishino H (2002). Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology*. 51: 689-702.

[3] Thorne JL and Kishino H (2004). Estimation of divergence times from molecular sequence data. (R. Nielsen, ed.) Statistical methods in molecular evolution. Springer Verlag.

[4] Seo T-K, Kishino H, and Thorne JL (2004). Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. *Molecular Biology and Evolution*. 21: 1201-1213.

[5] Aris-Brosou S and Yang Z (2003). Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa. *Mol. Biol. Evol.* 20: 1947-1954.

[6] Chujo A, Zhang Z, Kishino H, Shimamoto K, and Kyozuka J (2003). Partial conservation of LFY function between rice and Arabidopsis. *Plant Cell Physiology*. 44: 1311-1319.

[7] Sanderson MJ, Thorne JL, Wikström N, and Bremer K (2004). Molecular evidence on plant divergence times. *American Journal of Botany*. 91: 1656-1665.

[8] Wiegmann BM, Yeates DK, Thorne JL, and Kishino H (2003). Time flies, a new molecular time-scale for brachyceran fly evolution without a clock. *Systematic Biology*. 52: 745-756.

[9] Hasegawa M, Thorne JL, and Kishino H (2003). Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes & Genetic Systems*. 78: 267-283.

[10] Zhang Z, Inomata N, Yamazaki Y, and Kishino H (2003). Evolutionary history and mode of the *amylase* multigene family in *Drosophila*. *Journal of Molecular Evolution*. 57: 702-709.

[11] Zhang Z and Kishino H (2003). Genomic background drives the divergence at synonymous sites of duplicate *amylase* genes in *Drosophila*. *Molecular Biology and Evolution*. 21 222-227.

[12] Zhang Z and Kishino H (2004). Genomic background predicts the fate of duplicated genes:evidence from the yeast genome. *Genetics*. 166: 1995-1999.

[13] Zhang Z, Luo Z, Kishino H, and Kearsey MJ (2004). Divergence pattern of duplicate genes in protein-protein interactions follows the power law. *Molecular Biology and Evolution*. in press.

[14] Robinson DM, Jones DT, Kishino H, Goldman N, and Thorne JL (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*. 20: 1692-1704.

[15] Seo T-K, Thorne JL, Hasegawa M, and Kishino H (2002). Estimation of effective population size of HIV-1 within a host: A pseudo-maximum likelihood approach. *Genetics*. 160: 1283-1293.

[16] Seo T-K, Thorne JL, Hasegawa M, and Kishino H (2002). A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. *Bioinformatics*. 18: 115-123.

[17] Kitazoe Y, Kishino H, Okabayashi T, Watabe T, Nakajima N, Okuhara Y, and Kurihara Y (2004). Multidimensional vector space representation for correlated evolution and molecular phylogeny. *Molecular Biology and Evolution*. in press.

[18] Tan Z, Wada Y, Chen J, and Ohshima K (2004). Inter- and intralineage recombinants are common in natural populations of Turip mosaic virus. *J. General Virology.* 85: 2683-96.

[19] Kitada S and Kishino H (2004). Simultaneous detection of linkage disequilibrium and genetic differentiation of subdivided populations. *Genetics*. 167: 2003-2013.

[20] Scholl EH, Thorne JL, McCarter JP, Bird DM (2003). Horizontally transferred genes in plant-parasitic nematodes: A high-throughput genomic approach. *Genome Biology* 4: R39.

[21] Wada Y (2002) An exhaustive search for tandem repeats and long duplicated chromosomal regions within the mouse genome., *J. Anim. Genet*. 30: 3-10.

[22] Aris-Brosou S (2004). Determinants of adaptive evolution at the molecular level: The extended complexity hypothesis. Molecular Biology and Evolution. In press.