４．寄稿

集団遺伝学における確率モデル

生体高次情報系　助教授　能登原　盛弘

１．はじめに

私が学生の頃（１９７０年代）は、生物学と数学はお互いにほとんど無縁の学問という認識が一般の方々には多く、実験をすること無く数学やコンピューターを使って理論だけで研究をするなどと言うと変り種で怪訝な顔で見られていた様に思います。しかし、現在では、バイオと情報科学との関係は今後無くては成らないほどまでに発展し、この２５年程の変化は予想できないほどのスピードだったと思います。私自身は大学院学生の時代から集団遺伝学の数理的なモデルの勉強や研究をしてきましたが、平成１４年４月より、当研究科で研究と学生の指導に専念出来ることに成り、この３年ほどは自分の浅学さを少しでも取り戻すため、学生に戻ったつもりで勉強することができました。一部昨年の清水昭信先生の寄稿論文と重なる部分もあるかも知れませんが、私自身の研究及び関連した話題について御紹介したいと思います。

私が大学院学生として勉強させていただいた研究室は教授、助教授、助手の先生方が全て理論物理出身で、九州大学へ来られてから生物学の研究を始められたばかりで、先生方と学生が一緒に勉強しているという雰囲気でした。教授は基礎物理学研究所から来られた松田博嗣先生、助教授は宮田隆先生（（現）生命誌研究館顧問）、助手は石井一成先生（（現）名古屋大教授）、郷通子先生（（現）お茶の水女子大学学長）で５年先輩に高畑尚之さん（（現）総研大副学長）、１年上に五條堀孝さん（（現）遺伝研生命情報 DDBJ 研究センター長）その他にも現在各大学や研究機関で活躍中の方が先輩、後輩の院生としておられたので、ほんとに豪華な研究室だったと思います。このような先生方や先輩と今でも親しくお付き合いできるのは大変幸福なことと思っています。私の指導教官をしていただいた松田博嗣教授は集団遺伝学の確率過程を使った理論的研究をされていましたので、私もそちらの方へ興味が向いていったように思います。当時、まず J.F.Crow 博士と木村資生先生共著の"An Introduction to Population Genetics Theory"をテキストとして勉強を始めました。集団遺伝学の理論について詳しくまとめられた優れた教科書で、今でもその価値は変わっていないと思います。しかし、確率過程についてもう少し系統的に詳しく知りたいということで、数学科の学生や先生方にセミナーをお願いしながら、勉強をしていました。また少し進むと国立遺伝学研究所の木村先生、太田朋子先生や丸山毅夫先生の論文などを通じて、最新の研究につい

ても知るようになりました。外国では S.Karlin, G.A.,Watterson G.M.Malecot などの論文を読んでいました。清水昭信先生や志賀徳造先生（（現）東京工大教授）とお会いしたのもこの頃でした。

２．中立説と拡散モデル

例えば血液型を支配している遺伝子座を考えるとＡ，Ｂ，Ｏの複数の対立遺伝子が集団中に存在しています。このような現象を多型と言います。各対立遺伝子（Ａ，Ｂ，Ｏ）の遺伝子頻度は一般に時間と共に変化することが予想されます。この変化は自然選択や突然変異によることもありますが、各生物集団の個体数が有限であるために起こる確率的な変化があります。極端な例として、離れ小島で生活する数十人から成るヒトの集団を考えてみましょう。何十世代も経つと全ての住民が親戚縁者で、祖先を辿ると共通な一人の御先祖様に到達すると予想されます。その御先祖様がＡ型の血液型であるならば、その子孫である島の住民は全てＡ型ということになります。ただし、それはたまたまＡ型の子孫が島を占めるようになったというだけで、先祖の集団が全員同じように健康な人ばかりであれば、どの御先祖様も同じ確率で同様なことが起こると考えられます。すなわち、Ｂ型、Ｏ型の御先祖様の子孫が島の住民として占有していた可能性もあるわけです。これは極端な例ですが、生物集団の個体数は有限なので、突然変異が無い場合には長い時間の後には、遺伝的多様性は次第に減少し同様な現象が起こると言う訳です。ある特定の対立遺伝子が集団の全てを占めるようになることを遺伝子の固定と言います。固定に到るまでの遺伝子頻度の変化を追跡すると、ブラウン運動のようにランダムな動きをする確率過程として表現されます。これは遺伝子頻度の機会的変動あるいは遺伝的浮動(random genetic drift)と呼ばれています。もちろん自然選択や突然変異によっても、遺伝子頻度は変化しますが、分子進化で重要な役割をはたしているのは、この遺伝的浮動であるというのが、木村資生先生(1968)によって唱えられた中立説です。この説が出される数年前に Lewontin&Hubby (1966)や Harris(1966)によって電気泳動法を用いてショウジョウバエやヒトの種々の蛋白質について予想以上に遺伝的多型が存在しているということが発見されました。そしてこの多型現象を自然選択と突然変異だけで説明することは困難ということが分かってきました。中立説であれば、突然変異によって生じた対立遺伝子が集団に広がるチャンスは他の遺伝子と同等にある訳ですし、その途中の段階を見れば、複数の対立遺伝子が集団中に共存できることが自然に説明できます。 またヘモグロビンなど種々の蛋白質のアミノ酸配列を色々な生物間で比較すると二つの生物種が分岐してからの時間にほぼ比例して異なるアミノ酸の数が増えてくることが同様に知られています。これも中立説の立場では、突然変異と生物種内でのその遺伝子の固定が繰り返しおこれば、ほぼ時間

に比例してアミノ酸配列の相違が蓄積していくことが説明されます。蛋白質によってその構造や機能を維持するために重要な部分が多いほど制約が強く変わりにくくなり変化の速度（分子進化速度）は遅くなります。進化速度は幾つかの測り方がありますが、１００個のアミノ酸サイト当たり１個のアミノ酸の置換が起きるのに要する平均時間Ｔで表すと、例えばヘモグロビンでは$T = 0.83 \times 10^7$年、フィブリノペプチドでは1.1×10^6年,ヒストンＨ４は10^9年となります。フィブリノペプチドはフィブリノーゲンが血液凝固のためフィブリンに変化するときフィブリノーゲンから切り出されるキャップのような部分で、切り出されると、どんな機能も持たないと考えられています。従ってあまり自然選択すなわち機能的制約がなく生じた突然変異の多くを蓄積していると考えられ進化速度が速いタンパク質です。ヒストンＨ４は核のＤＮＡと結合し遺伝情報の発現を調節するという重要な機能を持っていると考えられています。従って機能的制約が非常に強く変わりにくいことになります。例えば牛とエンドウの１０５個のアミノ酸からなる配列の間にわずか２個のアミノ酸の違いしか存在しません。このような進化速度の遅いタンパク質は突然変異体は有害で集団から速やかに消失してゆくと考えられます。従って、中立説は自然選択を全く無視している訳では無く、もちろん有害な変異を排除する自然選択の存在は認めています。さらに、このような中立説の考えがアミノ酸やＤＮＡ塩基配列の相違を用いた系統樹の作成を可能にしています。

集団遺伝学の理論的研究は上記のように、遺伝子頻度の時間発展を考える確率過程、特に拡散過程を用いた研究が中心になされてきました。１９５０年代から６０年代に純粋数学での Feller 等による拡散過程の研究と平行して木村先生の拡散過程を用いた集団遺伝学の研究が進んでいったということは驚くべきことだったと思います。このような集団遺伝に現れる拡散過程の研究はその後確率論の研究者によっても大変興味を持たれ、現在でも測度値確率過程(Measure-valued process)として純粋数学の問題として発展し研究されています。

３．ＤＮＡ解析技術の進歩と遺伝子系図の研究

他方、１９８０年代からもう一つのアプローチの方法が注目され研究されるようになりました。Kingman, Hudson, Tajima によって独立に１９８２年から８３年にかけて出されたCoalescent モデルと呼ばれる方法です。これは遺伝子の系図を考え祖先を辿って生物集団の進化を考える方法です。先程の拡散過程は時間を正の向きに発展方程式的に考えるのに対して、時間を逆向きに辿る方法です。Coalescent モデルの定着した日本語訳はまだ無いようですが、名詞形の Coalescence は合体、癒着などという意味なので、祖先の系図を遡るという意味も含めて合祖モデルと呼ぶことにします。集団から任意に幾つかの遺伝子を取り出しその祖先遺伝子を辿って行くと、この名前の様に、次第に幾つかの遺伝子は祖先を共有し十分時

間を遡ると最終的にはそのサンプル遺伝子の一つの共通祖先に到達します。N人の孤立したヒト集団があって、ある遺伝子について n 本のＤＮＡサンプルを取り出したとすると共通な一つの共通祖先に到達するまでの平均待ち時間は$4N(1-\frac{1}{n})$世代になります。１世代を平均して２０年とすると n を十分大きく取ると約４Ｎ世代＝８０Ｎ年となります。この時のNは有効個体数と呼ばれるもので、実際の個体数よりはかなり小さな値になります。拡散過程モデルは多次元になると扱い難い点がありますが、それと比較して合祖モデルの良い所は直接サンプル遺伝子に関する種々の情報を提供してくれ、さらにコンピューター・シミュレーションも容易であるという点にあります。n 個のサンプルを取り出したとき、それが共通祖先に到達するまで、各合祖が起こるまでの待ち時間の分布が求まります。さらにその系図の上に突然変異（普通ポアソン過程で生じさせます）が起こったとするとサンプル遺伝子間で、特にＤＮＡ塩基配列を比較してどの程度異なっているか、その分布も計算できます。この結果と実際のサンプルを比較してモデルの検定や有効個体数、突然変異率などの統計的推定などがなされています。このような理由で、この２０年ほど前からは実験、理論両面からこの合祖モデルが勢力的に研究されています。

Kingman 等によって導入された合祖モデルはシンプルで扱いやすいモデルですが生物集団を一つの任意交配集団、すなわち、九州の人も東京の人も北海道の人もお互いの距離に関係なく結婚し子供を儲けているというモデルです。合祖モデルはかなり長い時間スケールで遺伝子の系図を見ていますので、そのスケールで見ると同じ日本人集団内では少しの距離は無視でき任意交配集団と見ることができるかもしれません。しかし、世界の人類集団全体を考えたとき、これを任意交配集団と見なすことには無理があります。１９８７年の Cann 等によるミトコンドリアＤＮＡを使った系図及び共通祖先の年代推定の発表以来、現在の人類集団のアフリカ起源説は有力になっていますが、人類集団の遺伝的多様性からその起源を推定することは、系図を用いる研究として魅力あるテーマだと思います。

DNA 塩基配列について遺伝的多様性を図る統計量としては、塩基多様度(nucleotide diversity)と塩基多型度(nucleotide polymorphism)という二つの量があります。塩基多様度は n 本のＤＮＡ配列を集団からサンプルした時、２本づつペアで比較して異なる塩基サイトの数を求めそれを全てのペアの組み合わせの数$\frac{n(n-1)}{2}$で平均し、さらにＤＮＡ塩基サイトの数で割って１塩基当たり２本のＤＮＡ配列を比較したときの異なるサイトの割合を表します。普通πという文字でよく表されます。ある一つの塩基サイトを見たとき、n 個のサンプル間で２個以上の塩基が含まれるときそのサイトは多型的なサイトと呼ばれます。塩基多型度は n 本のサンプル DNA 塩基配列間で、このような多型サイトの割合を表す統計量です。この二

つは集団遺伝学で多用される重要な統計量で、この二つの統計量の分布からデータの様々な統計的推測や検定がされています。ヒト集団では塩基多様度は全人類集団では平均して約 $\pi = 0.001$ ですが、アフリカ内では遺伝的多様性は高く、コーカソイド、モンゴリアンなど非アフリカ系では低い傾向があります。$\pi = 0.001$ とはヒト集団からランダムに選んだ二人のヒトのＤＮＡ塩基配列を比較したとき、１０００塩基に１ヶ所ぐらい異なるということを意味しています。現在、このような塩基座位ごとの多様性は SNP(single nucleotide polymorphism)と呼ばれ、種々の遺伝的疾患や薬剤の感受性の個人差の原因ではないかと医療や薬学の関係から注目されているものです。SNP は現在は平均して１０００塩基に１個程度で見つかっていますが、データの蓄積とともに最終的には数百塩基に１個程度まで増える予想です。

このような地理的な構造を考慮に入れた合祖モデルは１９８０年代の終わりごろから、Takahata(1988), Tajima(1989)等によって２分集団モデルなどが出されていましたが、１９９０年に私は任意個数、任意集団サイズの分集団構造と一般的な移住パターンを含むモデルを提案し、以後その性質を研究し少しずつ論文として発表しました。n 個のサンプル遺伝子を n 個の粒子のように考えると時間を過去に遡るにつれ、これらの粒子が分集団(colony)間を移住したり、共通な祖先に合祖しながら最終的には共通な一つの祖先(一つの粒子)に到達します。その共通祖先までの待ち時間の分布や、系図の上に突然変異を生じさせ遺伝的多様性の分布などを考察することができます。これらの方程式を解析的に解くことはかなり困難で、一般にはコンピューターの手を借りなければいけません。修士論文研究として梅田高呂さんにコンピューター・シミュレーションをしてもらったおかげで、限られた特殊な場合ではありますが、全体のイメージが分かってきました。分集団間の移住率が高い場合や逆に非常に弱い場合には一般的なモデルで解析的に解を得ることが可能なので、今後数学とコンピューターの両者を旨く利用した研究ができればと思っています。

合祖モデルはその他にも種々の方向へ拡張発展がされていますが、1988 年に Hudson 等によって自然選択を含む合祖モデルが出されました。ヒトのゲノムでは３万個ほどの遺伝子があると推定されていますが、一般に遺伝子座に突然変異が生じるとき、多くは有害な変異、中立な変異、それから非常に稀ですが有利な変異として生じます。自然選択が生物集団の進化にどのような働き、どのような影響を与えたのかという問題は非常に興味のある問題です。ゲノム上のどの領域に自然選択を受けている遺伝子があるのか、自然選択の影響の有無を調べるのに、実は中立な遺伝子がまた有効であることが分かってきました。すなわち、自然選択を受けている遺伝子が別の中立な遺伝子と染色体上の近い位置に連鎖しているとき、この中立な遺伝子は自然選択を受けている遺伝子に連鎖しているため、それに引きずられて親から子へと伝えられて行きます。中立な遺伝子にはほぼ時間に比例して突然変異が蓄積します。従って逆に生じた突然変異の数を調べることによってどのくらいの時間が経過したのか判断

できます。先程も述べたように有害な突然変異は集団から比較的速く消失して行きます。と言うことは時間を遡って見ると、有害な突然変異の起源祖先遺伝子までの待ち時間は短くなり、詳細は省略しますが、その遺伝子の近傍では集団の有効個体数が減少した様な系図となります。すると連鎖した中立なゲノム領域に蓄積した突然変異の数は減少します。すなわちサンプル遺伝子のＤＮＡ塩基配列を比較したとき、遺伝的多様性が低い中立な領域があると、その近くに有害な変異を持って自然選択を受けている遺伝子座が存在する可能性が高いと判断できる訳です。

現在ヒトゲノムを始めとして色々な生物種のゲノム解析が進んでいます。さらに種から個体レベルでの遺伝子の塩基配列の相違まで明らかにされつつあり、一つの生物種内に存在する遺伝的多様性について多くのデータが蓄積しつつあります。これは学問的興味として進化論的な意義があるだけでなく地球上の生物種の保全という視点からも貴重なデータだと思います。ゲノムの遺伝子配列の個体差を比較すると、遺伝的多様性が高いゲノム領域や低い領域があり、遺伝的多様性に関して一様でないことが分かっています。先程のヒト集団の塩基多様度が0.001というのは平均の意味であって、領域によって$\pi = 0.00036$から0.0046まで大きなばらつきがあることが分かっています。これはゲノム上のそれぞれの領域に様々な自然選択が働いている結果と考えられ、現在盛んな研究が進められているテーマです。詳細は省略しますが中立な遺伝子（遺伝子Ａ）が、ある自然選択を受けている遺伝子座（遺伝子Ｂ）に連鎖しているとき、遺伝子Ａが遺伝子Ｂのどの対立遺伝子に連鎖しているかによって遺伝子Ａを分類すると、その遺伝子の系図を表すモデルは地理的な構造と類似した合祖モデルになることが分かります。遺伝子座Ｂに起こる突然変異や遺伝子座Ａ，Ｂ間での組み換えが分集団間の移住の様な働きをします。このように、地理的な構造あるいは自然選択の存在下で対立遺伝子の型による分割構造を考慮にいれた合祖モデルは構造を持つ合祖モデル(structured coalescent model)と呼ばれています。遺伝子系図的には地理的構造の影響はゲノム全体に現れますが、自然選択の影響はその遺伝子を含む近傍のゲノム領域に特異的に現れます。そして、この二つの相乗効果で各生物種の遺伝的多様性は複雑な様相を呈してきます。今後のゲノム解析に伴うデータの蓄積とともに、その理論的解析の一つの方法として上記のモデルの研究を今後も進めて行きたいと思っています。

参考文献

1. 分子進化遺伝学　　根井正利　著　　培風館　（１９９０年）

2. ゲノムからみた生物の多様性と進化　　五條堀　孝　編　　シュプリンガー

3. Cann,R.L.,Stoneking,M. and Wilson,A.C. (1987) Mitochondrial DNA and human

evolution. Nature 325: 31-36.

4. Harris,H. (1966) Enzyme polymorphisms in man. Proc. Royal Soc.Lond. B. 164, 298-310
5. Hudson,R.R. and Kaplan,N.L. (1988) The Coalescent process in Models with Selection and Recombination. Genetics 120, 831-840.
6. Kimura,M. (1968) Evolutionary rate at the molecular level. Nature 217, 624-626.
7. Kingman,J.F.C. (1982) On the genealogy of large populations. J.Appl.Prob. 19A, 27-43.
8. Lewontin,R.C. and Hubby,J.L. (1966) A molecular approach to the study of genic heterozugosity in natural populations. Genetics 54, 595-609.
9. Notohara,M. (1990) The coalescent and the genealogical process in geographically structured population. J. Math. Biol. 29, 59-75.
10. Takahata,N. (1988) The coalescent in two partially isolated diffusion populations. Genet.Res. 58, 167-175.
11. Tajima,F. (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105, 437-460.
12. Tajima,F. (1989) DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics 123, 229-240.