

## 脊椎動物の比較ゲノム：遺伝子間領域の比較解析

隅山健太・斎藤成也

Kenta Sumiyama, Naruya Saitou  
 国立遺伝学研究所 集団遺伝研究部門

最近の脊椎動物のゲノム解析の進行により複数生物種間での比較が容易にできるようになり、タンパク質をコードする領域以外の遺伝子間領域にも進化的に保存されている機能領域が存在することが明らかになってきた。ゲノム重複などの影響を考え、適切な種を組み合わせて比較をすることで、より多くの機能領域を発見することができる。さらに、複数生物種の比較を統合することで、ペアワイス比較よりも詳細な機能領域のマッピングが可能になる。

### はじめに

脊椎動物のゲノムは、大半が“がらくたDNA (junk DNA, 用語解説)”と呼ばれる機能のない領域であり、遺伝子をコードする領域はわずかに過ぎないとされている。ヒトゲノムを例にとると、タンパク質をコードする遺伝子の数はたかだか25,000個であると推定されているが<sup>1)</sup>、1遺伝子あたりのエクソンの総塩基数を3,000個とすると、全ゲノムでは75MBとなる。これは3,000MBと推定される全ゲノムのわずか2.5%を占めるにすぎない。ヒトゲノムが特殊なのではなく、むしろ哺乳類ゲノムの典型だと考えられる。

では、ゲノムの残りの領域はすべてが機能を持たない、文字通りがらくたの領域なのだろうか？ ヒトゲノムの場合、膨大な直列反復配列領域や、散在反復配列のSINE (short interspersed nuclear element) とLINE (long interspersed nuclear element) が過半数を占めており、これらは特に積極的に機能を持たないがらくたであるようだが、それ以外のDNA領域には、単純な繰り返しではないユニーク配列がたくさんある。この中には、遺伝子重複のなれの果てである偽遺伝子がたくさん含まれているが、それら以外にも、由来の不明な配列が多数存在する。それらの塩基配列の中には、その生物を成り立たせるのに重要な機能を有するものが数多

く存在することが、近年少しずつわかってきた。これには、進化的に近縁な生物種間のゲノム配列の大規模比較が可能になり、高解像度で進化的な保存領域が同定できるようになってきたことが大きく寄与している。

本稿では、これからさらに急速に増大すると予想される、脊椎動物の近縁種ゲノムの比較解析の中でも、非コード領域の解析を中心とする脊椎動物のゲノム比較をどのように行うのか、また実際にどのようなことが見えるのかについて、筆者らの研究例を交えながら解説する。

### ◆ 1. 比較ゲノムから明らかになる遺伝子間の“進化的保存領域”

タンパク質をコードする遺伝子領域からはmRNAが発現されるので、発現したmRNAの塩基配列の全長を大規模に決定すれば、ゲノム配列中のどの部分がエクソンとなり、どの部分がイントロンとして切り出されるのかが明確にわかる。また、5'RACEや3'RACE (用語解説)により、転写開始点および終結点についても知見を得ることができる。さらに、このような既知の遺伝子構造のパターンを学習し、未知の配列中の遺伝子を予測することも盛んに行われている。

これに加え、ヒトゲノム配列やマウスなどのモデル生物ゲノム配列決定終了後に次々とモデル生物以外の近縁種ゲノム配列が決定されつつあり、その比較ゲノム手法により、“進化的に保存された領域”を機能部位の探索に用いることが可能になった。この進化的保存性を指標にした機能部位の探索法は、機能そのものが何であるかを知らずに機能部位を同定することができる、非常に強力な方法である。さらに、発現データやコンピュータによる予測などの独立した手法との組み合わせにより、信頼性の高い遺伝子発見ができるようになった。

機能の内容にかかわらず、機能の重要性に応じて進化的

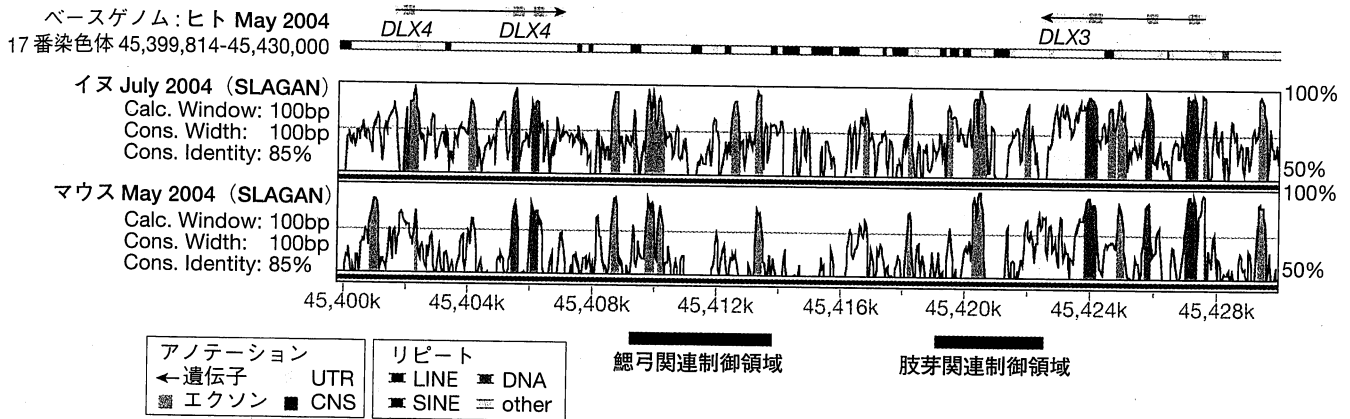


図1 VISTAを用いたヒト、イヌ、マウスのゲノム配列の比較の例(口絵参照)

ウィンドウのサイズは100bp. マウスの発生において鰓弓と肢芽特異的エンハンサー活性を示す領域を図中に重ねて示した. UTR ; untranslated region, CNS ; conserved nucleotide sequence.

に保存されるというその特性から、タンパク質コード領域ばかりでなく、イントロンや転写されていない遺伝子間領域 (intergenic region, 用語解説) にも、機能的に重要な多数の進化的保存領域が見いだされることが近年わかってきた。それらの中には、極保存配列<sup>2)</sup>と呼ばれる、通常のタンパク質コード領域よりもはるかに保存度が高い領域も見いだされる。こうした遺伝子間の高度保存領域は、おそらく遺伝子の発現調節を司る機能領域(エンハンサーやオルタナティブプロモーターなど, 用語解説)や、タンパク質をコードしないRNA 遺伝子などであり、中立進化する周辺領域の配列に比較して進化的制約度が高いために、保存領域として発見できるのだと予想される。特に高い保存度を示す領域は、初期の発生など進化的な制約が非常に強い生命現象に関わる遺伝子の発現制御などに関連しているのかもしれない<sup>3)</sup>。

## ◆ 2. 遺伝子間領域の保存領域を解析するための方法

ここでは、長大なゲノム配列の比較解析について説明する。種間でゲノム配列を比較するには、適切な整列(アライメント)を行うソフトウェアが必要である。塩基配列の整列をどれだけ正確にできるかが、後の解析すべてに大きな影響を与えるため、最も重要なステップといってもよい。さらに、整列結果を図示する手段も必要である。広く使われている定番のソフトウェアには、PipMaker<sup>4)</sup>とVISTA<sup>5)</sup>がある(関連web サイトURL 参照)。前者は局所的整列の集合体として全体の相似度を表現する方法であり、相同性(ホモロジー)の程度が直感的に理解しやすく、逆位や重複などがある場合にも使いやすい。後者は固定した塩基長のウィンドウを配列全体にずらして当てはめていったときの各所のスコアを表示する方法であり、指定したウィンドウサイズ

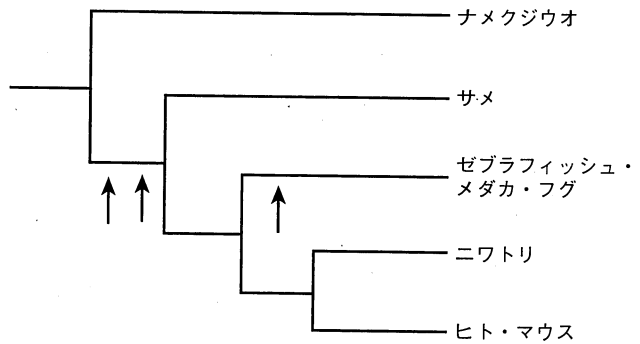


図2 脊椎動物の進化過程におけるゲノム重複時期のモデル 矢印はゲノム重複による遺伝子コピー数の増大を示す。

に依存して結果が変化するという難点はあるものの、原理的にシンプルで昔からよく使われている。どちらも有用であり、それぞれの長所がある。どちらを使うかは利用者の目的や好みによって選択すればよい。図1にVISTAの解析結果を示しているが、まったく同じ領域をPipMakerで解析した図を発表している<sup>6)</sup>ので興味のある方はその違いを見比べていただきたい。

## ◆ 3. 比較に用いる生物種はどのように選ぶべきか

種間で保存されている機能部位を探索する目的でゲノム配列を比較する場合、比較に用いる生物種の選択が非常に重要である。第一に、ゲノム重複(用語解説)などによる遺伝子重複に注意しなければならない。脊椎動物では、祖先の遺伝子数を $N$ とすると、哺乳類に至るまでに2回のゲノム重複を経験し $4N$ になっていると考えられている。硬骨魚類の系統ではさらにもう1回ゲノム重複が起こり、 $8N$ となっている(図2)。

ゲノム重複が起こると一時的に進化的な制約が緩んで再編成が起こるためなのか、検出できる保存領域が減少することが知られている<sup>7)</sup>。ゲノム重複を1回多く経験した硬骨魚類のゼブラフィッシュと哺乳類間の比較よりも、系統独自のゲノム重複が起きていないサメと哺乳類の方が、遺伝的により遠縁の関係にあるにもかかわらず、むしろ保存領域が多く見いだされるのである。このため、可能であれば、ゲノム重複を超えた比較を避けて、ゲノムの倍数性が同じ種同士を比較する方が、より効率的に保存領域が解析できる確率が高くなる。

次に、比較に用いる生物種があまりに遠い関係であれば、ほとんどのゲノム領域にわたって整列が不可能であることがある。例えば、ショウジョウバエとマウスのゲノム配列を比較しても、遺伝子間領域にはほとんど進化的な保存領域を見いだすことはできない。これに対して脊椎動物種間であれば、例えばフグとヒトのゲノム配列比較を行って遺伝子発現調節領域を探索することはずっと現実的になる。ただし、遺伝距離が離れていることに加え、先に述べたゲノム重複の影響もあって、進化的保存領域として検出されるのは、機能的にきわめて重要でほとんど塩基配列変化が許されない領域だけである。フグ・ヒトの間でこのような保存領域が1,400程度見つかった<sup>8)</sup>。比較する脊椎動物がさらに近縁種同士になっていけば、より多くの保存部位が見いだされるようになる。

図1は哺乳類種間での比較で、ヒト、マウス、イヌの *Dlx3/4* 遺伝子クラスター(用語解説)のゲノム配列をVISTAで解析したものである(隅山; 未発表)。*Dlx3/4* 遺伝子のエクソンが高いスコアを出しているのは当然として、そのほかにイントロンや、*Dlx* 遺伝子3'側遺伝子間領域に非常に多くの保存領域を示すピークが存在することがわかる。実際、この遺伝子間の保存領域は、発生過程の鰓弓や肢芽での発現制御をする機能領域であることが明らかになっている<sup>6), 9)</sup>。

このように、ヒトとマウスやイヌといった哺乳類の中でも系統的に大きく異なる生物種間のゲノム比較は非常に有用であるが、同時に2種だけの比較には限界もある。種間の保存度を指標にするこのような方法では、比較する領域内に種が分岐してから現在に至るまでに十分な量の塩基置換などの突然変異が蓄積していることが前提である。あまりにも小さな領域を比較しようとすると、その範囲に十分な塩基置換が期待できず、保存しているのかどうかかわからない。また、もし比較する2種が最近に分岐した生物であれば、機能のない領域であっても種間の違いはもともと非常に少なく、たとえ広い領域をとっても保存領域をあぶり出すこと自体が困難になる。

例えば、マウスとラットとの間での保存領域を探すこと

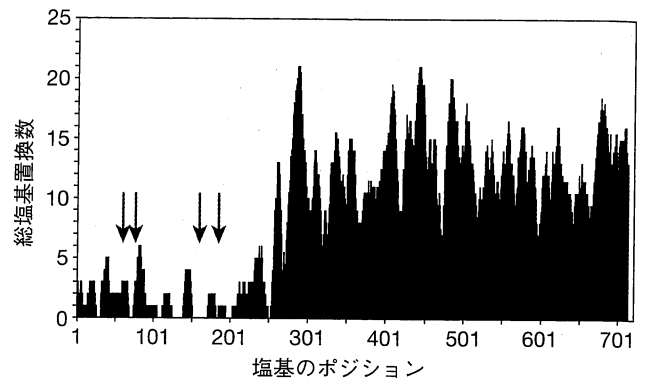


図3 *HOXC8* 遺伝子の早期エンハンサーにおける8哺乳類種を用いた進化保存モチーフの解析

矢印が実験で転写因子が結合することが示された場所。ウィンドウサイズは10bp。

は、ゲノム配列を用いることができるにもかかわらず、両者が進化的に近縁なので、おそらくかなり難しい作業になる。この問題を解決するための方法が、多種間でのゲノム配列比較である。

#### ◆ 4. 近縁多種間ゲノム比較による機能領域高精度マッピング

VISTAで描いた図1をもう一度見ていただきたい。ヒトとイヌ、ヒトとマウスのペアワイズ比較(用語解説)が示されているが、よく見ると両方の同じ位置のピークの高さが異なっている場合があることがおわかりいただけると思う。これは種に特異的な変化である可能性もあるが、大部分は遺伝的な距離が十分でないため、偶然に置換数がばらつくことも影響している。このような2種だけの解析では、ウィンドウの幅が100bpであり、これよりも小さな保存領域を検出することは難しいが、かといってウィンドウのサイズを小さくすると、今度は偶然に種間で一致してしまうような擬陽性のピークが多数生じてしまうことになり、検出の信頼度が落ちてしまうことになる。

そこで、より多くの種間比較を行い、各比較の合計塩基置換数の期待値を上げることによって、信頼できるウィンドウサイズを小さくしていくことができる。近縁種間のペアワイズ比較では限られた情報しか得ることができないが、生物種を増やして近縁・多種間で比較を行いそのデータを統合すると、非常に高精度で緻密な進化的保存領域をマップすることができるようになるのである。

具体的な例として、図3に哺乳類 *HOXC8* 遺伝子の制御領域の近縁多種間比較の結果を示した。近縁多種間での比較により、転写因子が結合する配列のサイズに相当する10bp以下の保存配列を検出することができた<sup>10)</sup>。この検出結果

は、生化学的な転写因子の結合実験の結果ともよく一致するものであった。

このような手法は、系統フットプリント法(phylogenetic footprinting)とも呼ばれる。この呼び名の提唱者でもあるモリス・グッドマンらは、合計の遺伝距離が250MYA (million years ago) 以上になるような種の組み合わせが望ましいと述べている<sup>11)</sup>。また、一応の目安として、統計的に有意に進化的保存があることを示すためには、統合された多種間比較ウィンドウ内に10以上程度の塩基置換が期待できるようなウィンドウサイズを設定するべきだろう。

さらにこのような系統フットプリント法によって、保存度の高いモチーフを抽出し、系統グループごとにその比較をすることで、系統特異的な保存モチーフを同定することができる。また、エンハンサーなどの非コード領域の機能領域が進化的にさかのぼっていった場合にどの程度保存されているのかを定量的に解析するために、系統フットプリント法により抽出されたモチーフがカバーする領域のサイズ変化を指標として解析する方法に発展させることができる。通常エンハンサーは転写調節因子などの結合配列などで構成されており、これらが突然変異で失われると重大な機能変化をもたらすため、機能的な制約からある一定の遅い進化速度を示すと考えられる。このため、系統フットプリント法で抽出された保存モチーフがカバーする領域のサイズは、機能的な制約が一定であれば比較に用いた生物種の進化的な時間、すなわち遺伝的な距離が大きくなるに従って、一定の速度で小さくなるのが期待される。ところが、生物の進化の過程で機能的な制約が大きく変化した場合、例えば機能が不要になった場合や、ある新しい機能を獲得した場合には、エンハンサー配列に不連続的に急速な進化が観測される可能性がある。筆者らはこの原理を用いたERA (evolutionary rigidity assay) という方法を開発し、*Dlx3/4* 遺伝子クラスターの鰓弓特異的エンハンサー配列に系統特異的な不連続的急速進化が起きていることを検出した(隅山; 未発表)。このような手法を用いることで、形態などの表現型進化とゲノムレベルでの非コード領域の進化とを対応させて解析することが、これからの比較ゲノム機能解析の重要なトピックとなると考えられる。

## ◆ 5. 非コード領域で最も保存性の高い領域 ◆ “極保存領域”の進化解析

筆者らは極保存配列が哺乳類の進化でどのような重要性を持つのかを解明するため、新たに6種類の生物(クジラ、ゾウ、ジュゴン、アシカ、アザラシ、コウモリ)について極保存配列に対応する塩基配列を決定し、その進化的解析を進めている。極保存配列はヒト、マウス、ラットの3種のゲノム

ム配列中で200塩基以上にわたって同一の塩基配列を有している領域であり、481個が見つかった。筆者らはこれらのうちおよそ半分について塩基配列を決定したところ、他の哺乳類ではまったく同一というわけではないが、哺乳類のがらくたDNAにおける進化速度(塩基置換数)が年あたり塩基サイトあたりで $1\sim 5\times 10^{-9}$ 程度であるのに対して、その1/100の進化速度であり、哺乳類の中で確かに高度に保存されている配列であることを確認した(Kimら; 未発表)。

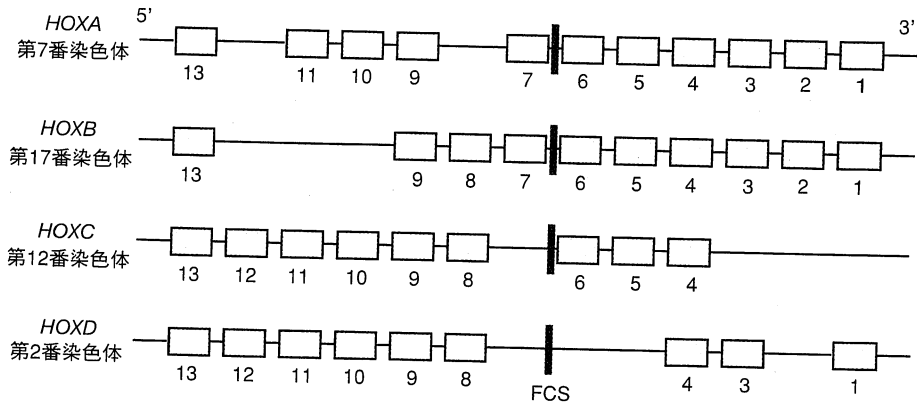
しかし、一部の領域においては、生物種の系統特異的に塩基置換が他の系統よりもずっと多く蓄積していることを見いだしている(Kimら; 未発表)。これは、その系統の祖先種で、一時的にその極保存配列の淘汰上の制約が弱まり、別の表現型になったことが考えられる。

## ◆ 6. パラログ間の比較ゲノム解析

種分化による配列変化の蓄積を用いた比較ゲノムが現在一般に行われている方法であり、本稿でもこれまでに主に種間ゲノム配列比較による解析方法を解説してきた。最後に、種間の比較ではなく、ゲノム内の重複イベントにより生じたセグメント間での比較解析の可能性について述べる。

脊椎動物のゲノムは過去に何度か全ゲノム重複を起こしていると考えられている。この結果生じたパラログ領域間で、非コード領域の比較解析を行うことは可能であろうか? 答えは、限定的な条件の下で可能である。すなわち、重複した領域間で機能的な制約が同等に維持されているような場合には、何らかの機能モチーフがそのまま維持・保存される場合があり、パラログ間比較により保存モチーフが検出されれば、それは重複領域間共通の保存的機能の存在を示唆することになる。脊椎動物のパラログ間比較で非コード領域に連続的な保存領域が検出された例は現時点では多くないが、その例の1つが*Hox* 遺伝子群の非コード領域に見つかったFCS (four cluster sequence) というモチーフである<sup>12)</sup>。哺乳類のゲノムには2回のゲノム重複に起因すると考えられる4つの*Hox* クラスタが存在するが、それらすべてのクラスタに共通に保存されている配列がFCSである。この配列は*Hox* パラロググループ6番と7番の間の遺伝子間領域に存在しプロモーターの近傍に存在することから、何らかの転写調節機能に関わっているようであるが、4つの*Hox* クラスタすべてに存在することから、クラスタ固有の特殊な発現調節機能に関連するのかも知れない。

別の例として、ゲノム重複で生じたと考えられる*Dlx3/4* クラスタと*Dlx5/6* クラスタの遺伝子間領域間にも保存配列が検出されている<sup>9)</sup>。この配列は*Dlx* 遺伝子の発現を制御するエンハンサー配列の一部であり、*Dlx* タンパク質の結



HHOXA	ATAAATCCGTTGTTGTTT-ATGAAAATTTACAACCTTTGCAATACAAC--TTTATGAGTTG
HHOXB	ATAAATTCCTTGTGTTTATGAAAATTTACAACCTTTGTGATAGAAG--TTTATGAGTTG
HHOXC	-----GTTGTTTGCAGAAAATTTACAGCTGAGTAATAAAG--TTTACGATCGA
HHOXD	-TAAATTCCTTTTGTGTTGAAGAAAATTTACAACCTTGGTAATAGACCTTTTATGACCTA
	: : : : : * : * * * * : *
HHOXA	TTCCGCCCTTCCATTGGCCGCTGTCGGTCATGTGGA--TGAGAACCCTGAACATGAACCTT
HHOXB	TTGAATCCAGCGATTGGCCGCGCCGGTTCATGTGGCCGGGC-GATCGTGAACATGAACCTT
HHOXC	CTCACAGTTGGATTGGCCACAAGAAGTCATGTGGA--TTCCATCCATGAACCTGAACCTT
HHOXD	TTTGGCGTCGTCATTGGCTGCGGCTGGTTCATGTGCAGGCGCCGCTCGCCTTACCGCCTT
	: * * * * * : : : : * * * * * : : * * * * * : : * : : : : * : : : * * *
HHOXA	TTTTATAATTTCCCTTGGCAGAATAGAGCCGCATTCTTTTG
HHOXB	TTTA-TCATTTCCCTGGTGGTTATAATGCAGCATTCTTTTG
HHOXC	TTT-----
HHOXD	TTTCTGATTTCCCCGCGAATTTCCCCCGCATTCT----
	***

図4 パラログ間の非コード領域保存配列の一例，ヒトHOXクラスターFCS配列のアラインメント  
HOXのA, B, C, D各クラスターに保存している配列が，HOXパラロググループ6番と7番の間に存在する。

合配列を含む領域が含まれている。このDlx結合配列はDlxクラスターの発現調節機構を維持するために不可欠な配列であるために保存されていると考えられる。

脊椎動物の全ゲノム重複の時期は大まかにわかっており、重複後に十分な変異が蓄積していることが確かであるためにこのような解析が可能であるが、ごく最近の重複の場合には変異が十分に蓄積していないためあまり有効な手段ではないことは種間比較の場合と同様である。重複の時期を推定できることは重要な条件の1つとなる。

全ゲノム重複の他に過去の一時期に重複したことが明らかである配列の一例が、散在型の反復配列である。驚くべきことに、散在型反復配列であるSINE配列がゲノムに挿入後エンハンサーとなって種間保存領域として検出されることが報告されている<sup>13)</sup>。さらに、このような機能制約によると思われる保存されている散在型反復配列は例外的ではなく、ゲノム中に数多く検出されることがわかってきた<sup>14)</sup>。過去の一時期に一斉に増幅した反復配列は一種のパラログ配列であると見なすことができ、これらのパラログ配列間でも

し共通に保存されている配列があれば、それは何らかの機能を持つことを示唆する。これは一例に過ぎないが、筆者らは散在型反復配列の一種MIR配列の1座位に種間保存モチーフ配列を発見し、そのモチーフ配列が他の座位のMIRにも保存されている例があることを確認している。これが偶然の結果なのか、あるいは機能的な制約の結果なのかは今後の解析を待つ必要があるが、今後このような例の報告が増え、その機能を探索することがごく普通に行われるようになるだろう。

**おわりに —今後の展望—**

ゲノムデータの蓄積により、その機能がまだ不明なものも含めて、タンパク質をコードする領域以外の機能部位を、近縁多種間比較により詳細に探索することが可能になってきた。これからはこのような機能未知のデータが大量に蓄積されていく時代になっていく。その機能を知るためのハイスループットの実験系が必要になるだろう。それによ

てグローバルな転写制御機構などの遺伝子発現プログラムの解明などが劇的に進むことが予想される。こうした知識の蓄積によってゲノム進化が生物進化に果たす役割が分子レベルで理解できるようになるだろう。

## 文献

- 1) International Human Genome Sequencing Consortium: Nature 408, 708-713 (2004)
- 2) Bejerano, G. et al.: Science 304, 1321-1325 (2004)
- 3) Sandelin, A. et al.: BMC Genomics 5, 99 (2004)
- 4) Schwartz, S. et al.: Genome Res 10, 577-586 (2000)
- 5) Frazer, K.A. et al.: Nucleic Acids Res 32 (Web Server issue) W273-279 (2004)
- 6) Sumiyama, K. et al.: Proc Natl Acad Sci USA 99, 780-785 (2002)
- 7) Chiu, C.H. et al.: Genome Res 14, 11-17 (2004)
- 8) Woolfe, A. et al.: PLoS Biol 3, e7 (2005)
- 9) Sumiyama, K. et al.: Proc Natl Acad Sci USA 100, 4030-4034 (2003)
- 10) Sumiyama, K. et al.: Genomics 71, 260-262 (2001)
- 11) Gumucio, D.L. et al.: Mol Phylogenet Evol 5, 18-32 (1996)
- 12) Kim, C.B. et al.: Proc Natl Acad Sci USA 97, 1655-1660 (2000)
- 13) Bejerano, G. et al.: Nature 441, 87-90 (2006)
- 14) Mikkelsen, T.S. et al.: Nature 447, 167-177 (2007)

## 関連 web サイト URL

- ・ PipMaker  
<http://pipmaker.bx.psu.edu/pipmaker/>
- ・ VISTA  
<http://genome.lbl.gov/vista/index.shtml>

## 用語解説

### ◆がらくたDNA(junk DNA)

単純反復配列など特に機能的な意味を持たないと考えられている、現在のところ機能不明の配列。哺乳類を中心とする脊椎動物ではゲノム中でかなりの割合を占める。

### ◆RACE法

Rapid amplification of cDNA ends の略。mRNA の一部分の配列がわかっている場合に、cDNA の既知領域と5'あるいは3'側との間でPCR法により未知領域の配列を知る方法。cDNA ライブラリーを作製する必要がない方法である。

### ◆遺伝子間領域(intergenic region)

ゲノム中のタンパク質をコードする遺伝子構造(エクソンとイントロンを含む)の外側の領域。タンパク質をコードしないrRNAやtRNA以外のRNA遺伝子や、遺伝子の発現調節に関わるエンハンサーやサイレンサー、さらに大規模なクロマチン構造の制御に関わる因子などがあると期待されるが、機能の解析が難しいため全容は明らかになっていない。

### ◆エンハンサー

遺伝子発現を増強させる作用を持つシスに働く因子。組織特異的な発現などを制御する。

### ◆オルタナティブプロモーター

同一遺伝子のプロモーターが複数存在し、組織特異性などにより選択的に使われているもの。エンハンサーの多様性ととともに、高等脊椎動物の遺伝子機能の多様性を生み出す機構の1つと考えられる。

### ◆ゲノム重複

生物のゲノムが倍加することにより遺伝子重複を起こすこと。遺伝子の重複により各遺伝子にかかる機能的な制約による淘汰圧が緩むために遺伝子の進化速度が上がり、新しい進化を起こすきっかけになる可能性がある。

### ◆Dlx3遺伝子

ショウジョウバエの*distal-less*遺伝子の脊椎動物ホモログ遺伝子で、哺乳類に6つある*Dlx*遺伝子の1つ。*Dlx3*のノックアウトマウスは胎盤の形成不全で胎生期に致死となる。鰓弓、四肢、外生殖器、乳腺、胎盤などに発現し、これらの付属器官の形態形成を制御すると考えられる。遺伝子自体は脊椎動物の起源よりも古くから存在することから、その転写制御領域の進化が脊椎動物の付属器官の形態進化をゲノムから理解するうえで鍵となる。

### ◆ペアワイズ比較

塩基配列を比較するときに、1:1で2種類の配列を比較することをペアワイズ比較と呼び、複数の配列を同時に比較することを配列の多重比較という。多重比較を行うことで、ペアワイズの比較よりも多くの情報を得ることができる。