統計数理（2002）
第 50 巻 第 1 号 3–15

特集「DNA 配列の統計解析」
［総合報告］

配列のホモロジーと統計情報を併用した真核生物遺伝子構造の予測

後藤 修†

（受付 2001 年 12 月 6 日；改訂 2001 年 12 月 19 日）

要　　旨

近年大量のデータが作り出されているゲノム配列から，様々な有用情報を引き出すための第一歩は（特に蛋白質をコードする）遺伝子を同定することである．しかし，真核生物の遺伝子のほとんどはイントロンにより分断されているため，遺伝子領域を探索し，正確な遺伝子構造，すなわちエキソン・イントロンの配置，をゲノム配列から予測することは未だ難しい問題である．転写制御領域や翻訳領域（エキソンのうち蛋白質をコードする部分）における k−塩基出現頻度の偏りや，エキソン・イントロン境界配列の特徴を数値化し，ニューラルネット，判別分析，あるいは隠れマルコフモデルを用いて総合的に判断する方法がこれまでに開発され，かなりの成功を収めている．しかし，正確にエキソンを予測できる精度は塩基レベルで約 75%とされ，いっそうの性能向上が望まれている．筆者は，上記のゲノム配列に関する統計情報に加え，既知のアミノ酸配列や cDNA 配列との相同性を併用して，真核生物の遺伝子構造をより精度よく予測する方法を開発してきた．本稿では，この方法を中心に，最近の真核生物遺伝子構造予測法の進展について概説した．

キーワード：遺伝子構造予測，エキソン・イントロン，スプライシング，ゲノム情報，
配列ホモロジー，アラインメント．

1. はじめに

2001 年はじめ，ヒト全ゲノムの大部分（約 94%）を解読したドラフト配列が公表され，社会的にも大きな関心を集めた（Lander et al.（2001）; Venter et al.（2001））．真核生物としてはパン酵母（*Saccharomyces cerevisiae*），線虫（*Caenorhabditis elegans*），ショウジョウバエ（*Drosophila melanogaster*），シロイヌナズナ（*Arabidopsis thaliana*）に続くもので，多数の細菌，古細菌のものとあわせて，生物分類上重要な枝それぞれにつき，代表的な生物種のゲノム情報がほぼ出揃ったことになる．

ゲノム配列から様々な有用な情報を引き出すための第一歩は（特に蛋白質をコードする）遺伝子を同定することである．この問題は「遺伝子発見」（gene-finding）問題と呼ばれ，この 10 年ほどの間に非常な進展を見た（高木（1997）; Claverie（1997）; Burge and Karlin（1998））．しかし，現在も発展段階にあり，とくにヒトの遺伝子数についてはまだ様々な議論がある（Hogenesch et al.（2001））．仮にゲノム上のある領域に遺伝子の存在が推測できたとしても，正確な遺伝子

† 産業技術総合研究所 生命情報科学研究センター：〒135–0064 東京都江東区青海 2–41–6

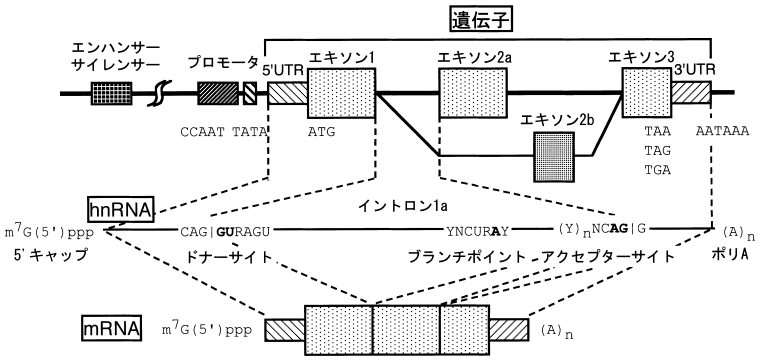

図 1. 真核生物遺伝子の模式図．エキソン 2 は選択的スプライシングに従うとした．直接の転写産物である hnRNA は 5′ キャップおよび 3′ ポリ A 付加という修飾を受け，スプライスされて mRNA に成熟する．短い塩基配列はそれぞれの機能部位のコンセンサス配列を表す．

構造，すなわちエキソン・イントロンの配置，をゲノム配列のみから予測することは未だ難しい課題である（Birney et al.(2001)）．正確な遺伝子構造予測に基づいて正確なアミノ酸配列を知ることは，その蛋白質の構造や機能を理解し，また分子系統学的な知見を得るために必須の要件である．

蛋白質をコードする典型的な真核生物遺伝子の模式図を図 1 に示した．遺伝子の定義として転写制御領域を含めることもあるが，ここでは転写開始部位から転写終了部位までをひとつの遺伝子とした．現実には複数の転写開始部位や転写終了部位が存在することも多く，選択的スプライシングと相まって事態を複雑にしている．転写開始部位から翻訳開始部位までの間を 5′UTR（5′Untranslated Region），翻訳終了部位から転写終了部位までを 3′UTR とよぶ．これらの領域は翻訳の効率や mRNA の安定性などに関与すると考えられ，アミノ酸配列には翻訳されない．また，直接の転写産物である hnRNA がリボゾームで翻訳される mRNA に成熟するまでの間に，イントロンはスプライセオゾームという RNA・蛋白複合体によって除去される．我々の主な目的は，UTR 領域やイントロンをのぞいた翻訳領域（翻訳エキソン）を予測することである．そのための方法としては，統計情報だけに基づくもの（しばしば *ab initio* 法とよばれる）と，既知の配列との相同性を利用したもの（簡単のためホモロジー法とよぶ）とに大別される．最近，その中間的なものや，いくつかの方法の結果を総合的に判断するものも開発されている（矢田（2001））．3 章で詳しく述べる我々の方法は相同性に偏った中間法といえるだろう．その紹介の前に，これまでに多くの研究がなされてきた代表的な統計的方法について次章で概観する．

2. 統計的方法に基づく遺伝子発見

2.1 統計情報

翻訳領域の予測のために有効な統計情報として利用されているものは次の 4 つのカテゴリーに分けられる．(1) エキソンやイントロンの長さなどの総体的な量，(2) 翻訳領域の塩基配列の特徴，(3) 転写開始・終了，翻訳開始，スプライシングの 5′，3′ 境界，ブランチポイント（図 1）など，細胞内分子機構にとって重要な部位周辺の塩基配列パターン，(4) 翻訳領域でないことの有力証拠として，(CA)n などの単純反復配列や Alu, L1 などの散在配列の存在．なお，これらの性質はゲノム全体を通して一定であるとは限らず，その領域の平均的な G+C 含量な

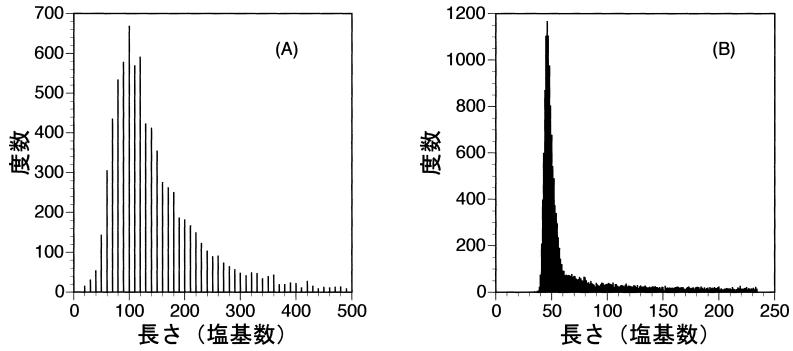

図 2. 線虫遺伝子のエキソン（A）およびイントロン（B）の長さの分布.

どによって影響を受けることも知られている.

（1）エキソン，イントロンの長さ

図 2 に線虫遺伝子のエキソンとイントロンの長さの分布を示した．イントロンの長さが長い領域では，ほぼ幾何分布に従う．しかし短い領域では明瞭な下限があるとともに，鋭い分布の山が認められる．他の生物種でも傾向としては同様であるが，下限値とピークの位置は生物種に依存して，ある幅だけ右側へ移動している．これらの観察はイントロンの長さが複数の機構によって支配されていることを示唆するが，詳細は不明である．一方，図 2 では区別していないが，最初あるいは最後のエキソンと内部エキソンとで分布に差が見られる．両端のエキソンの分布がかなり広いのに対し，内部エキソンの分布域は狭い傾向が見られる（Burge and Karlin (1997)）.

（2）コーディングポテンシャル

翻訳領域における塩基配列の特徴は通常「コーディングポテンシャル」と呼ばれる量として表現される．もっとも標準的には，6 文字塩基出現頻度の偏りを利用する．コドン使用頻度に生物固有の偏りがあり（Nakamura et al.（2000）），また翻訳領域には 3 塩基の周期性がある（Shepherd（1981））というよく知られた事実がこの背景にある．コドンの 1, 2, 3 文字目からそれぞれ始まる 5 次のマルコフモデルや，3 塩基を単位とした 1 次のマルコフモデル（ダイコドン−dicodon−モデル）を用い，ゲノム全体の塩基配列から得られた類似のモデルを対照とした対数オッズによってコーディングポテンシャルを定義する.

（3）境界シグナル

図 3 にスプライシングのドナーサイト（イントロンの 5′ 端）およびアクセプターサイト（イントロンの 3′ 端）近辺の塩基配列パターンを例示する．ほとんどのイントロンは GT で始まり AG で終わるという規則に従う．ただし，0.5〜1%のイントロンが GC-AG，ごく少数が AT-AC 末端をもつ．それぞれの境界を基準にゲノム配列を並置（align）した後，各塩基の出現頻度と平均的な値との対数オッズ（log odds = lod）を求めた．不変な GT(GC)/AG 部位以外にも，5′, 3′ 境界近辺の各部位は平均的な塩基組成から大きくずれていることが見て取れる．各部位でもっとも多く用いられる塩基を並べると，5′: CAG/**GT**RAGT, 3′: (Y)nNC**AG**/G ，というコンセンサス配列が得られる．ここで，太文字：不変部位，R=A or G, Y=C or T, N=A,C,G,T，/ は境界を表す．各部位が独立であると仮定すれば，この lod 値を直接用いるか（Stormo（1990）），線形判別（飯田（1995））を用いるかして重み行列（weight matrix）を導くことができる．実際には部位の間にかなり強い相関が観察され，より複雑な処理が必要である．ここでも 1 次ないし 2 次のマルコフモデル（Zhang and Marr（1993）; Salzberg（1997））が主に使われるが，さら

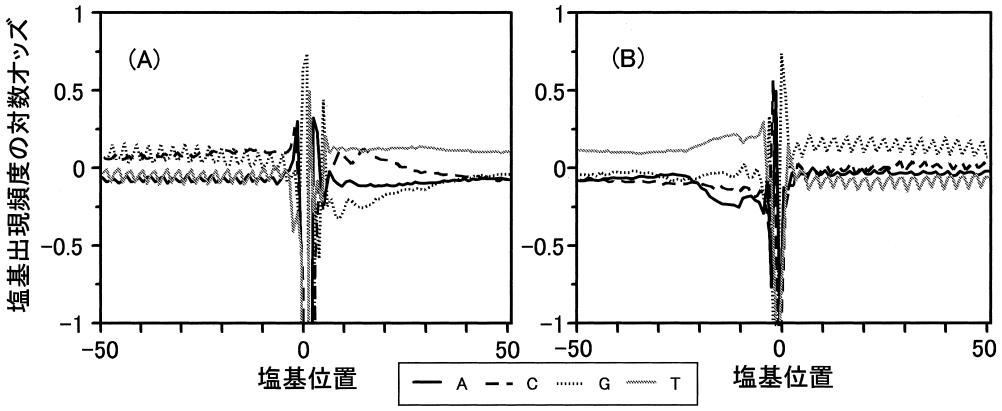

図 3. シロイヌナズナ，イントロン 5′(A) および 3′(B) 境界近辺の塩基出現頻度の偏り．境界を中心とした各位置での 4 種類の塩基の出現頻度と背景的塩基組成との比の対数を塩基ごとに異なる種類の線で表す．

に遠距離の相関を取り入れた方法，例えば，ニューラルネット（Brunak et al.(1991)），決定木（Burge and Karlin（1997)），ベイズネット（Cai et al.(2000)）などの手法も提案されている．ここではスプライシング部位を中心に述べたが，翻訳開始部位など他の機能部位の予測にも同様の方法が用いられる．

（4）反復配列

反復配列は統計情報と呼べないかもしれないが，有力な非翻訳領域の証拠となる．特にヒトの場合，ゲノムの実に半分以上が反復配列とみなされる．ゲノム断片の重複を除いて，反復配列には大きく 2 種類がある．単純反復配列あるいは低情報量（low complexity）領域は，1〜500 塩基の（しばしば不完全な）単位がタンデムに繰り返した領域を指す．一方，反復配列の大部分を占める散在配列は，トランスポゾンや逆転写の機構によって重複複製され，ゲノム全体に広がったものである．前者は低情報量領域を同定するアルゴリズムによって，後者は各種反復配列のコンセンサス配列との類似性によって検出できる（Smit and Green（1999)）．

2.2 ニューラルネット

前小節で，翻訳領域を予測するために有用な様々な統計量を概観したが，そのいずれからも決定的な法則性を導くことは出来ない．例えば，スプライシング境界のコンセンサスに一致する配列は翻訳領域や非翻訳領域内部にも多く存在し，どれが本物の境界であるかを見分けることは容易でない．様々な情報を総合して判断する方法の一つにニューラルネットがある．ある領域がエキソンであるか否かを判別するために Uberbacher and Mural（1991）はニューラルネットを用い，それまでの方法に比べ 20%以上も高い正答率を得た．これは領域毎の予測であるので，遺伝子全体のモデルを構築するには，ダイナミックプログラミング法を用いたさらに総合的な情報処理を必要とする（Snyder and Stormo（1995)）．

2.3 判別分析

すでに前節境界シグナルの項でも触れたが，線形判別は統計的に異なる性質を持つグループを分別するための一般的な手法として広く用いられている．2.1 節で述べた様々な統計量を要素とするベクトルを用い，本物のエキソンと偽エキソン（AG と GT ではさまれた ORF のうち本物のエキソンではないもの）を最大限分離するため，Solovyev et al.(1994）は線形判別を

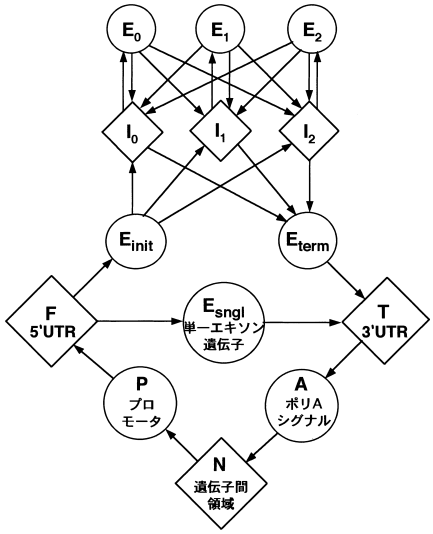

図 4. 真核生物遺伝子を表現する隠れマルコフモデル．遺伝子間領域（N）→ プロモータ領域（P）→ 5′UTR（F）→ 第一エキソン（E_{init}）→ イントロン（三つのフェイズのいずれか）→ · · · →最終エキソン（E_{term}）→ 3′UTR（T）→ ポリ A シグナル（A）→ 遺伝子間領域（N）という状態遷移が，複数のエキソンから成る一つの遺伝子構造に対応する．各状態は一般に異なる確率分布で塩基を出力する．状態間の遷移は境界シグナル強度に応じて確率的に起きる．Burge and Karlin (1997) を参考にした．

適用し，ニューラルネットより優れた結果を得た．その後，Zhang（1997）はより分別能力に優れた二次判別法を適用し，さらによい結果を得ている．

2.4 隠れマルコフモデル

上記のニューラルネットや判別分析ではあらかじめ可能なエキソン領域を設定し，本物か否かを判定した．真核生物遺伝子発見用の隠れマルコフモデル（Kulp et al.（1996））では先に領域設定を行わず，ゲノム上の各塩基は図 4 に示す状態のいずれかに属するものと考える．モデルの各状態は固有の確率で塩基を出力し，また状態間の遷移も確率的に起きる．塩基の出力確率は，例えば状態ごとに異なる 5 次のマルコフモデルに従うものと仮定する．また，状態遷移の確率は境界シグナルの強さに応じて調整される．このモデルをゲノム配列の特定の領域に適用したとき，最大の確率を与える状態の並びが予測される遺伝子構造となる．

単純な隠れマルコフモデルでは，各状態が持続する確率は幾何分布になる．現実のイントロンやエキソンの長さの分布はこれより明らかに狭い（図 2 参照）．ある状態が持続する確率も取り入れた「一般化隠れマルコフモデル」（Kulp et al.（1996））や「セミ隠れマルコフモデル」（Burge and Karlin（1997））を用いることにより，現状ではもっとも精度の高い統計的遺伝子構造予測が可能となっている．

3. 相同性と統計情報に基づく遺伝子構造予測

統計情報による様々な真核生物遺伝子予測プログラムの性能を改めて評価すると，制作者が報告したものほどには精度が高くない場合が多い．テストデータの偏りや，過度に最適化した条件設定が原因であると思われる．より高い精度を目指すために，スプライシングを受けた後

の配列情報を何らかの形で利用することが考えられる．最も直接的には，mRNA に由来する cDNA 配列との比較を行う．また，相同性が認められる既知のアミノ酸配列との比較に基づく，より間接的な方法も考えられる．この章ではこれらの方法を紹介する．

3.1 cDNA 配列との比較

ゲノム配列とそれに由来する cDNA あるいは EST（細胞から採取した mRNA を無作為に逆転写し，シークエンサーを 1 回だけ通して読みとった cDNA の一種）配列とを比較すればもっとも直接的に遺伝子構造を決めることができる．そのために，例えば blastn（Altschul et al.（1990））などの局所配列アラインメントプログラムを利用することが一般に行われている．しかしながら，この方法にはいくつかの問題がある．第一に，イントロンを挟む境界領域の配列は，ある程度重複していることが多い．そのため，どこが実際の境界であるかを GT-AG 規則など別の基準で決めることになる．大量データに対してこれを人手で行うとエラーが生じやすい．第二に，とくに EST 配列には読みとり誤差が多い．欠失・挿入がある場合や，誤差が境界に近い場合にはエキソン・イントロン境界の決定が難しくなる．第三に，転写の向きが不明な場合があり，それを決定する基準が必要となる．最後に，多型や人為的な要因による配列の欠失・挿入もしばしば見出され，イントロンと区別する必要がある．

これらの問題点を考慮したプログラムがいくつか報告されている（Mott (1997); Florea et al.（1998); Chao（1999); O. Gotoh（unpublished））．アルゴリズムとしては，長いギャップを許して 2 配列間の最良アラインメントを求めるダイナミックプログラミング法（Gotoh（1990））を修正したものが用いられる（図 5（A））．単に GT-AG 規則を境界シグナルとして用いることもあるが，統計情報に基づくシグナル強度を考慮することもできる（O. Gotoh（unpublished））．

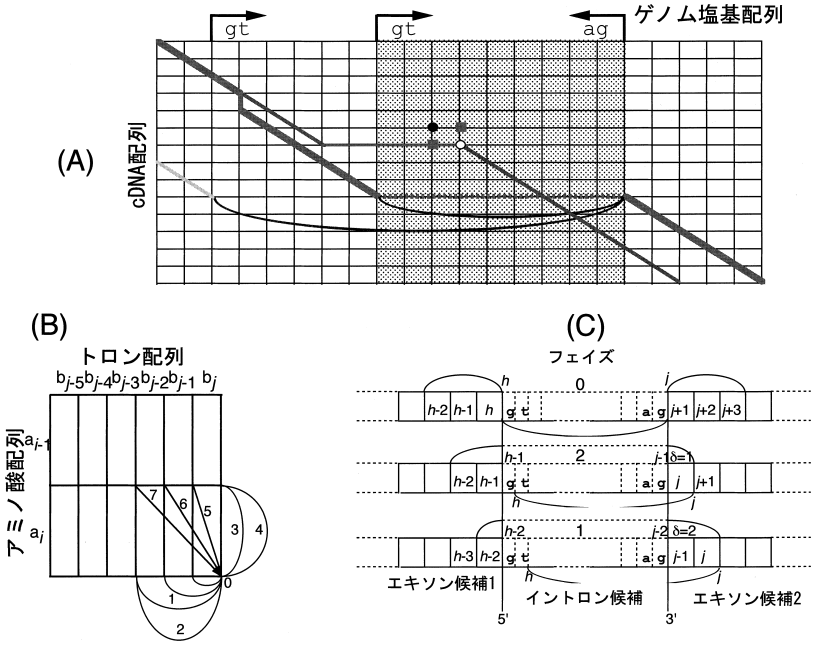

図 5. ダイナミックプログラミング法を用いた，ゲノム配列対 cDNA 配列（A）またはアミノ酸配列（B）のアラインメントパスを示す．後者の場合，3 通りのイントロンのフェイズを考慮しなくてはならない（C）．

	T		C		A		G		
T	TTT	**F** 14	TCT		TAT	**Y** 19	TGT	**C** 5	T
	TTC	(Phe)	TCC	**S** 16	TAC	(Tyr)	TGC	(Cys)	C
	TTA	**L** 11	TCA	(Ser)	TAA	**O** 23	TGA	**U** 22	A
	TTG	(Leu)	TCG		TAG	(Ter)	TGG	**W** 18	G
C	CTT		CCT		CAT	**H** 9	CGT		T
	CTC	**L** 11	CCC	**P** 15	CAC	(His)	CGC	**R** 2	C
	CTA	(Leu)	CCA	(Pro)	CAA	**Q** 6	CGA	(Arg)	A
	CTG		CCG		CAG	(Gln)	CGG		G
A	ATT	**I** 10	ACT		AAT	**N** 3	AGT	**J** 21	T
	ATC	(Ile)	ACC	**T** 17	AAC	(Asn)	AGC	(Ser)	C
	ATA		ACA	(Thr)	AAA	**K** 12	AGA	**R** 2	A
	ATG	**M** 13	ACG		AAG	(Lys)	AGG	(Arg)	G
G	GTT		GCT		GAT	**D** 4	GGT		T
	GTC	**V** 20	GCC	**A** 1	GAC	(Asp)	GGC	**G** 8	C
	GTA	(Val)	GCA	(Ala)	GAA	**E** 7	GGA	(Gly)	A
	GTG		GCG		GAG	(Glu)	GGG		G

図 6. トロンコード．この表の太字で示すトロンコードを 3 塩基（トリプレット）の中央の塩基に与える．最初と最後の塩基をトリプレットに対応づけるため，配列の直前直後には “A” があるものとみなす．23 文字からなるトロンコードは元の塩基配列情報を保持しつつ，仮想翻訳アミノ酸配列も表現する．

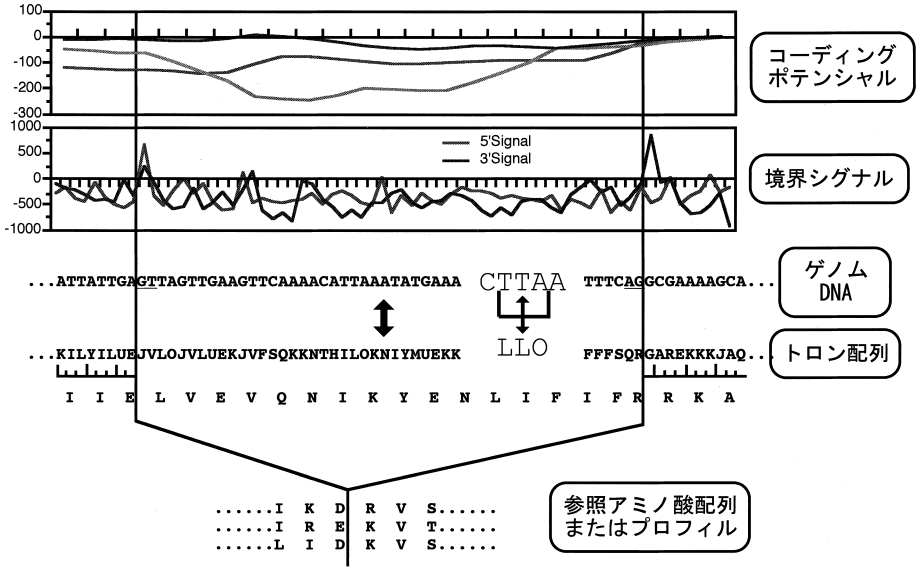

図 7. 統計情報と相同性を考慮した遺伝子構造予測．コーディングポテンシャルおよび境界シグナルが統計情報から得られ，参照アミノ酸配列またはプロフィルとの類似性と合わせて総合的にエキソンを予測する．最良のスコアを求めるには図 5 に示したダイナミックプログラミング法を用いる．

3.2 トロン（TRanslated codON）コード

一般的にいえることであるが，蛋白質の構造や機能に直接結びついたアミノ酸配列の方が塩基配列より分子進化の過程で変化を受けにくい．翻訳領域の塩基配列と，そこにコードされるアミノ酸配列の両方を同時に表現できる方法があれば，翻訳領域同士や翻訳領域とアミノ酸配列との比較が容易になる．「トロンコード」（図 6）はそのような目的のために標準遺伝暗号の

特性を考慮して考案された（Gotoh（2000））．すなわち，標準遺伝暗号では，対応するアミノ酸の物理化学的性質と最も深い関係を持つのはコドンの 2 文字目の塩基である（Crick（1968））．実際，セリン（TCN, AGY）と終止コドン（TAR, TGA）を除いて，どのアミノ酸に対するコドンも 2 文字目の塩基は固定している（図 6）．従って，通常の 20 文字に，第二のセリンと 2 種類の終止コドンに対応する文字を加えた計 23 文字を用意すれば，もとの塩基配列に関する情報を失うことなく，コードされるアミノ酸配列（および枠のずれた仮想翻訳配列）を表現することが可能となる（図 7）．トロン配列を通常の塩基配列，アミノ酸配列，別のトロン配列と比較するには，それぞれ 23×4, 23×20, 23×23 要素からなるスコア行列を用いる．

3.3 アミノ酸配列，プロフィルとの比較

DNA との比較に基づく遺伝子構造予測が適用できるのは，そのもの自身，またはごく近縁の遺伝子由来の cDNA 配列を参照配列とする場合に限られる．アミノ酸配列を用いれば，適用範囲が大幅に広がり，系統的に遠く離れた生物種由来の参照配列を利用することも可能である．さらに，単独のアミノ酸配列ではなく，多くの関連配列から得られる多重配列アラインメント（Gotoh（1999）），または，さらにそれから導かれるプロフィル（Gribskov and Veretnik（1996））を参照することで，一層の精度，感度の向上が見込まれる．

ゲノム配列対 cDNA 配列の場合に比べ，翻訳後の配列を基準に比較を行う際には読み枠（frame）の存在を意識する必要が生じる．さらに，第 2 章で述べた統計情報の知識も考慮に入れることが望ましい．そのため，基本となるアルゴリズムに次のような拡張がなされた（図 5（B）, 7）．（1）ゲノム配列を前節で述べたトロンコードにあらかじめ変換しておき，翻訳レベルでの比較を容易にする．（2）3 通りの読み枠ごとにコーディングポテンシャルを計算し，スコアに加算する．（3）フレームシフトを引き起こす（3 の倍数でない長さの）塩基の欠失，挿入には特別のペナルティを課す．（4）エキソン・イントロン境界では境界シグナルをスコアに加算する．（5）そのときイントロン挿入部位の前後で読み枠を保存する．二つのコドンの間，コドンの 1 文字目の後，コドンの 2 文字目の後に挿入されるイントロンを，それぞれフェイズ 0, 1, 2 イントロンとよぶ．フェイズ 1 および 2 イントロンの場合，スプライシング後のコドンを合成してから翻訳しなくてはならない．（6）エキソン，イントロンの長さに応じてスコアを調整する．極端に短いエキソン，イントロン候補を予測から排除するために特に有効である．単独の参照配列の代わりにプロフィルを用いても，アルゴリズムに大きな変更は必要ない．その際，内部に含まれるギャップに対応した「一般化プロフィル（Gotoh（1994））」を用いることで厳密性と効率性が保たれる．

表 1 に，相同性を主に利用する既報の遺伝子構造予測プログラムを掲げた．Procrustes（Gelfand et al.（1996））はあらかじめエキソン候補を列挙しておき，様々な組み合わせの中から最良のスコアをもつものをダイナミックプログラミング法で求める．GeneWise（Birney and Durbin（1997））などそれ以外の方法は，隠れマルコフモデルを基礎とするが，イントロンを特殊な挿入と見なす配列アラインメントと基本的に同等である．

4. 予測の評価

通常，既知の遺伝子構造をどれだけ正確に再現できるかによって，真核生物遺伝子構造予測の性能を評価する．ここでは相同性を基準にしているので，一つの遺伝子が占めるゲノム配列上の範囲は既知であるとの（現実には必ずしも満たされない）仮定の下での評価を試みる．他の多くの予測問題と同様に，以下で定義される感度（Sn），精度（Sp），相関係数（CC）を定量的な基準として用いる．

表 1. ホモロジーに基づく遺伝子構造予測プログラム.

プログラム名	タイプ a)	文献	公開サイト
Est_genome	G×C	Mott, 1997	www.well.ox.ac.uk/~rmott/est_genome.shtml
Sim4	G×C	Florea *et al.*, 1998	globin.cse.psu.edu
Calign	G×C	Chao, 1999	iubio.bio.indiana.edu/soft/mol/bio/align/calign.c
Procrustes	G×P	Gelfand *et al.*, 1996	www-hto.usc.edu/software/procrustes
Nap	G×P	Huang & Zhang, 1996	www.cs.mtu.edu/faculty/huang.html
GeneWise	G×P	Birney & Durbin, 1997	www.sanger.ac.uk/Software/Wise2
Aln	G×P	Gotoh, 2000	www.cbrc.jp/~gotoh/
GeneSeqer	G×P	Usuka & Brendel, 2000	ftp.zmdb.iastate.edu
Glass+Rosetta	G×G	Batzoglou *et al.*, 2000	www.theory.lcs.mit.edu/crossspecies
Pro-Gene	G×G	Novichkov *et al.*, 2001	www.anchorgen.com/pro_gen/pro_gen.html

a) 比較の対象として用いる配列の種類：cDNA (G×C), アミノ酸 (G×P), または他のゲノム DNA (G×G).

$$Sn = TP/(TP + FN) \tag{4.1}$$

$$Sp = TP/(TP + FP) \tag{4.2}$$

$$CC = (TP \times TN - FP \times FN) \Big/ \sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)} \tag{4.3}$$

遺伝子構造予測では，塩基，境界，エキソン，遺伝子など，異なるレベルごとに評価を行うことが一般的である．例えば，個々の塩基レベルでは，TP は真のエキソンに含まれることが正しく予測された塩基数，FP は誤ってエキソンとして予測されたエキソン外の塩基数，TN は正しく予測されたエキソン外の塩基数，FN は誤ってエキソン外と予測された塩基数である．

統計情報だけに基づく *ab initio* 予測プログラムの性能評価に関してはすでにいくつかの報告がなされてきたが（Burset and Guigó (1996); Claverie (1997); Burge and Karlin (1998); Rogic et al.(2001)），ホモロジー法に関しての評価は比較的新しい（Guigó et al. (2000)）．ヒト遺伝子を対象とした結果は次のように要約できる．(1) 参照配列との類似性が高ければ（ホモロジー検索プログラム blastx の P 値が 10^{-100} 以下），ホモロジー法（GeneWise, Procrustes）は非常に高い感度・精度をもつ．(2) 類似性が下がるに従いホモロジー法の感度が低下し，*ab initio* 法（GeneScan）に劣るようになる．(3) 一方，精度に関してはホモロジー法（特に GeneWise）が類似度によらず *ab initio* 法に勝る．(4) ゲノム配列上の遺伝子領域が未知の場合には，ホモロジー法の結果があまり変わらないのに反し，*ab initio* 法の性能（特に精度）が大幅に悪化する．

筆者の開発した方法の性能を，約 300 の線虫遺伝子を対象として評価した結果を図 8 に示す（Gotoh（2000））．類似性の低下とともに正答率が低下するが，Guigó らの結果ほどには感度の低下が見られない．同じ遺伝子セットを対象とすると，GeneWise は遺伝子を断片化して予測する傾向が強いことが判明した．ヒト遺伝子に関する Guigó らの結果もこの傾向を反映しているものと推測される．現在，筆者の方法をヒト遺伝子の予測に適用して，その性能評価を試みているところである．

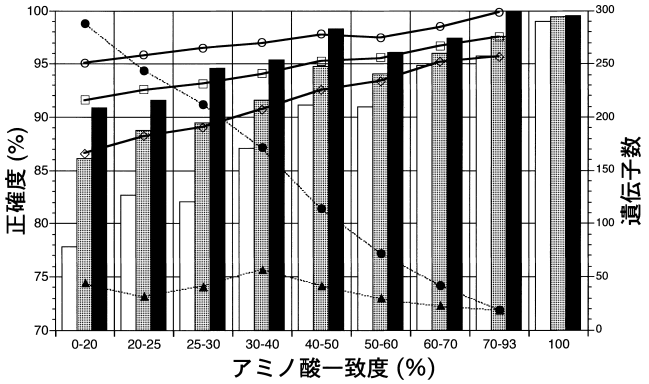

図 8. 約 300 の線虫遺伝子を対象として行った遺伝子構造予測結果の検証．横軸は参照配列と検査対象遺伝子産物とのアミノ酸の一致度を表す．棒グラフは表記した範囲の一致度をもつ組み合わせ（その数を▲で示す）ごとの結果を，折れ線グラフは表記以上の一致度をもつ組み合わせ（その数を●で示す）についての結果を示す．黒棒と○は塩基レベルでの相関係数，灰色棒と□は境界レベルでの正確度（精度と感度の調和平均），白棒と◇はエキソンレベルでの正確度を表す．

5. 今後の課題と展望

遺伝子発見はゲノム配列情報解析の第一段階にしかすぎず，それぞれの遺伝子産物の構造と機能の解明が求められている．そのためにも遺伝子構造予測に高い感度・精度が要求されるが現状は決して満足できるものでない．純粋に統計的な方法には限界があり，cDNA 配列や既知のアミノ酸配列情報を取り入れた融合的な方法が今後の主流となるであろう．また，この総説では詳しく触れることができなかったが，複数の相同な遺伝子の DNA 配列（Batzoglou et al.（2000); Novichkov et al.（2001））または予測翻訳配列（Gotoh（2000））を相互に比較することにより，遺伝子内部構造を予測することも非常に有望である．

重要な今後の課題として，プロモータ，エンハンサー，サイレンサーなど転写制御を司る領域の予測，および選択的スプライシングについての予測がある．多くの事例についての統計的な解析が，いまだ明確でないこれらの分子機構を解明する上で，有用な洞察を与えるに違いない．上に述べた比較ゲノム学的方法もまた大変有力な方針となるであろう．発現プロファイリングなどの情報と相互に連携した総合的な研究を推進していくことが，今後強く求められている．

謝　辞

原稿について有益なコメントを頂いた笠原直子氏（コンパックコンピュータ）に感謝いたします．

参 考 文 献

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J.（1990）. Basic local alignment search tool, *Journal of Molecular Biology*, **215**, 403–410.

Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. and Lander, E. S.（2000）. Human and mouse gene structure: Comparative analysis and application to exon prediction, *Genome Research*, **10**, 950–958.

Birney, E. and Durbin, R. (1997). Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison, *ISMB*, **5**, 56–64.

Birney, E., Bateman, A., Clamp, M. E. and Hubbard, T. J. (2001). Mining the draft human genome, *Nature*, **409**, 827–828.

Brunak, S., Engelbrecht, J. and Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence, *Journal of Molecular Biology*, **220**, 49–65.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA, *Journal of Molecular Biology*, **268**, 78–94.

Burge, C. B. and Karlin, S. (1998). Finding the genes in genomic DNA, *Current Opinion in Structural Biology*, **8**, 346–354.

Burset, M. and Guigó, R. (1996). Evaluation of gene structure prediction programs, *Genomics*, **15**, 353–367.

Cai, D., Delcher, A., Kao, B. and Kasif, S. (2000). Modeling splice sites with Bayes networks, *Bioinformatics*, **16**, 152–158.

Chao, K. M. (1999). Calign: Aligning sequences with restricted affine gap penalties, *Bioinformatics*, **15**, 298–304.

Claverie, J. M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences, *Human Molecular Genetics*, **6**, 1735–1744.

Crick, F. H. (1968). The origin of the genetic code, *Journal of Molecular Biology*, **38**, 367–379.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence, *Genome Research*, **8**, 967–974.

Gelfand, M. S., Mironov, A. A. and Pevzner, P. A. (1996). Gene recognition via spliced sequence alignment, *Proc. Nat. Acad. Sci. U.S.A.*, **93**, 9061–9066.

Gotoh, O. (1990). Optimal sequence alignment allowing for long gaps, *Bulletin of Mathematical Biology*, **52**, 359–373.

Gotoh, O. (1994). Further improvement in methods of group-to-group sequence alignment with generalized profile operations, *Computer Applications in the Biosciences*, **10**, 379–387.

Gotoh, O. (1999). Multiple sequence alignment: Algorithms and applications, *Advances in Biophysics*, **36**, 159–206.

Gotoh, O. (2000). Homology-based gene structure prediction: Simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps, *Bioinformatics*, **16**, 190–202.

Gribskov, M. and Veretnik, S. (1996). Identification of sequence pattern with profile analysis, *Methods in Enzymology*, **266**, 198–212.

Guigó, R., Agarwal, P., Abril, J. F., Burset, M. and Fickett, J. W. (2000). An assessment of gene prediction accuracy in large DNA sequences, *Genome Research*, **10**, 1631–1642.

Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G. and Cooke, M. P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes, *Cell*, **106**, 413–415.

Huang, X. and Zhang, J. (1996). Methods for comparing a DNA sequence with a protein sequence, *Computer Applications in the Biosciences*, **12**, 497–506.

飯田陽一 (1995). 文字列の中にルールを見付ける ── スプライシングシグナルの定量化,『ヒトゲノム計画と知識情報処理』(美宅成樹, 金久 實 編), 99–139, 培風館, 東京.

Kulp, D., Haussler, D., Reese, M. G. and Eeckman, F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA, *ISMB*, **4**, 134–142.

Lander, E. S. et al. (2001). Initial sequencing and analysis of the human genome, *Nature*, **409**, 860–921.

Mott, R. (1997). EST_GENOME: A program to align spliced DNA sequences to unspliced genomic

DNA, *Computer Applications in the Biosciences*, **13**, 477–478.

Nakamura, Y., Gojobori, T. and Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: Status for the year, *Nucleic Acids Research*, **28**, p. 292.

Novichkov, P. S., Gelfand, M. S. and Mironov, A. A. (2001). Gene recognition in eukaryotic DNA by comparison of genomic sequences, *Bioinformatics*, **17**, 1011–1018.

Rogic, S., Mackworth, A. K. and Ouellatte, F. B. (2001). Evaluation of gene-finding programs on mammalian sequences, *Genome Research*, **11**, 817–832.

Salzberg, S. L. (1997). A method for identifying splice sites and translational start sites in eukaryotic mRNA, *Computer Applications in the Biosciences*, **13**, 365–376.

Shepherd, J. C.(1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification, *Proc. Nat. Acad. Sci. U.S.A.*, **78**, 1596–1600.

Smit, A. F. A. and Green, P. (1999). http://repeatmasker.genome.washington.edu/

Snyder, E. E. and Stormo, G. D. (1995). Identification of protein coding regions in genomic DNA, *Journal of Molecular Biology*, **248**, 1–18.

Solovyev, V. V., Salamov, A. A. and Lawrence, C. B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames, *Nucleic Acids Research*, **22**, 5156–5163.

Stormo, G. D. (1990). Consensus patterns in DNA, *Methods in Enzymology*, **183**, 211–221.

高木利久 (1997). タンパク質コード領域予測,『ヒューマンゲノム計画』(金久 實 編), シリーズ・ニューバイオフィジックス, 11 巻, 33–45, 共立出版, 東京.

Uberbacher, E. C. and Mural, R. J.(1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach, *Proc. Nat. Acad. Sci. U.S.A.*, **88**, 11261–11265.

Usuka, J. and Brendel, V. (2000). Gene structure prediction by spliced alignment of genomic DNA with protein sequences: Increased accuracy by differential splice site scoring, *Journal of Molecular Biology*, **297**, 1075–1085.

Venter, J. C. et al.(2001). The sequence of the human genome, *Science*, **291**, 1304–1351.

矢田哲士 (2001). 遺伝子発見プログラム DIGIT, 蛋白質核酸酵素, **48**, 2580–2585.

Zhang, M. Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis, *Proc. Nat. Acad. Sci. U.S.A.*, **94**, 565–568.

Zhang, M. Q. and Marr, T. G. (1993). A weight array method for splicing signal analysis, *Computer Applications in the Biosciences*, **9**, 499–509.

Proceedings of the Institute of Statistical Mathematics Vol. 50, No. 1, 3–15 (2002)

Prediction of Eukaryotic Gene Structures Based on Combined Information of Sequence Homology and Statistical Features

Osamu Gotoh

(Computational Biology Research Center (CBRC),
National Institute of Advanced Industrial Science and Technology (AIST))

Draft sequences of complete human genome were made publicly available in February 2001. This follows publication of virtually complete genomic sequences of yeast, nematode, fruit fly, and cress. Since many bacterial and archaeal genomes have been sequenced, we already had the basic information about at least one organism in representative phylogenetic branches. The first step toward extraction of any useful information from the blue-prints of life is to identify exact structures of individual genes. Biased distributions of k-tuple oligonucleotides in coding and non-coding regions are useful in deriving "coding potentials". In addition, specific sequence patterns around transcriptional, translational, and splicing boundaries can be converted to numerical signals that help to delineate exonic and intronic regions. Several mathematical methods, including neural networks, discriminant analyses, and hidden Markov models, have been developed to assemble various lines of information into predicted genes and gene structures. These intensive efforts have dramatically improved the prediction quality based on such statistical information in the last decade. However, the success rate for correctly predicting an exon is reported to be only about 75% at the nucleotide level, so there is still considerable room of improvement. We have taken a slightly different approach. In addition to the statistical information mentioned above, we incorporate sequence-homology information to more accurately locate coding regions conserved between the target gene and one or more reference cDNA or protein sequence. The most likely gene structure is inferred by optimizing an objective score by means of a dynamic programming algorithm. To assess the performance of our method, we compared the predicted gene structures with known structures of about 300 *C. elegans* genes. The results indicate that the percentage of correctly predicted exons exceeded 90%, which was significantly better than those obtained by other methods.

Key words: Gene-structure prediction, exon-intron organization, splicing, genome informatics, sequence homology, alignment.