ヒトゲノムにおける広義の遺伝子発見研究

京都大学大学院情報学研究科
矢田 哲士

Research on broad sense gene finding in the human genomes

Tetsushi Yada
Graduate School of Informatics, Kyoto University

ここでは、ヒトゲノムに対する広義の遺伝子発見研究、すなわち、タンパク質遺伝子、偽遺伝子、機能性RNA遺伝子の発見研究について報告する。我々は、まず、各々の研究課題に関する遺伝子発見プログラムを新たに開発し、さまざまなベンチマークデータを用いてそれらの性能を評価した。さらに、それらのプログラムを用いたゲノムアノテーションプロトコルを確立し、ヒトのゲノム配列に対して適用した。我々のゲノムアノテーションは、新しく開発されたゲノムブラウザ HAL（Human genome Annotation Library: http://hal.genome.ist.i.kyoto-u.ac.jp/）から公開されている。以上の研究により、現在の遺伝子発見技術の可能性と限界が明らかになり、幾つかの生物学的な発見がもたらされた。

We introduce here our research activities on broad sense gene finding for the human genomes. Our activities include protein coding gene finding, pseudo gene finding and functional RNA gene finding. We have newly developed computer programs (gene finders) for each research subject and have evaluated their performance by using various types of benchmark data. Moreover, we have established a protocol for genome annotations by using them and have applied them to the human genome sequences. Our genome annotations have been provided from a newly developed genome browser named HAL (Human genome Annotation Library: http://hal.genome.ist.i.kyoto-u.ac.jp/). These research activities have clearly shown possibilities and limitations of current gene finding technique and have brought several biological discoveries.

1．はじめに

ヒトゲノムの完成配列が公開され、そこに潜む遺伝情報の解読に多くの研究者が挑戦している。その先駆けとなる研究は、やはり、Nature誌で発表されたヒトを形作るタンパク質遺伝子の網羅的な探索であろう。その報告によると、ヒトゲノムから発見されたタンパク質遺伝子の総数はショウジョウバエとほぼ同じ22,000個ほどであり、ヒトの複雑さを説明するには少なすぎる数であった。そこでヒトの複雑さを説明する鍵として、遺伝子の転写や発現の調節に注目が集まり、それを実現する仕組みとして、機能性RNA遺伝子や選択的スプライシングが注目を浴びている。

一方で、上述のNature誌の見積りは妥当なのだろうか？ Nature誌の報告では、これらのタンパク質遺伝子は、配列との類似性、なかでも転写産物との配列類似性によって同定された。そのため、同定された遺伝子に擬陽性が含まれることは少ないが、網羅性には大きな疑問が残る。す

なわち、ヒト遺伝子の組織特異的な発現や時期特異的な発現、さらに転写産物の収集における実験感度を考慮すると、これまでに収集された転写産物とは全く配列類似性を示さない遺伝子や弱い配列類似性を示さない遺伝子が、未だゲノムの中に数多く潜んでいると考えられる。実際、その後の網羅的な再探索により、およそ1,500個のタンパク質遺伝子が新たに同定されている。

そこで本研究課題では、既存手法では発見できないタンパク質遺伝子を主たる対象として、異なるアプローチに基づいた予測精度の高い遺伝子発見プログラムを組み合わせ、網羅的で信頼性の高いタンパク質遺伝子の発見に挑戦した。まず従来法を上回る信頼性を示す様々なアプローチのタンパク質遺伝子発見プログラムを用意し、それらを組み合わせた遺伝子発見プロトコルの確立に注力した。また、遺伝子発見の結果を整理・格納・表示・検索するデータベースの開発を行ない、さらに、タンパク質遺伝子のカタログ化を側面から支援するために、偽遺伝子の発見やRNA遺伝子の発見にも果敢に取り組んだ。

2. 研究開発の成果

2.1　配列類似性に基づくタンパク質遺伝子の発見研究

当期間において3つの研究課題に取り組んだ。第1に、複数生物種間でオーソロガスな遺伝子調節領域の多重配列アラインメント法を検討した。遠縁のゲノムで保存される領域を感度よく検出するため、多重シード法や動的計画法の適用限界を検討した。

第2の課題は以前からの継続で、複数ゲノム配列アラインメントに基づくヒト遺伝子翻訳エキソンの同定である。前研究期間で基本的な部分の開発を終え、Alnggプログラムとして実装している[1]。本期間ではより多くの情報を取り込むことにより、いっそうの精度向上を試みた。具体的な改良点は、(1) イントロン長の分布に基づくイントロン挿入スコア関数の導入、(2) そのスコア関数を含む各種パラメータ値のG+C含量依存性の導入、(3) アラインされた塩基ペアの統計量を用いたスコア関数の算出、である。従来はコーディングポテンシャルやスプライスシグナルをそれぞれのゲノム配列ごと独立に計算した後加算していた。改良点 (3) では、正しくアラインされた配列ペアを学習データとすることにより、塩基保存情報も考慮したより精度の高い統計情報が得られるようになった。ゲノム比較法を用いた公開プログラムの中で現在もっとも高精度とされるN-ScanとAlnggとの性能を、エキソンレベルでの予測精度で比較した結果を図1に示す。AlnggとN-Scanとでは感度と選択性のバランスが異なるが、平均値としてはほぼ同等であることが示された。

転写産物情報を利用した遺伝子予測法であるALNプログラム[2]では、第3の課題として主に計算速度の改善に努めた。対象配列と参照配列との間の類似度に強く依存するが、平均3倍ほどの高速化が予測精度を損なうことなく実現できた。ENSEMBLデータベースの既知遺伝子を用いた検証の結果は、同様の目的で一般に用いられるExonerateプログラムに比べALNの明らかな優位性を示す（図2)。

AlnggおよびALNの本年度中の一般公開に向けて現在ウェブサイトの整備に取り組んでいる。

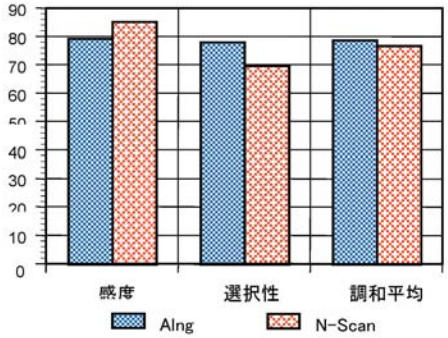


図1　Alnggのエキソン予測精度（%）

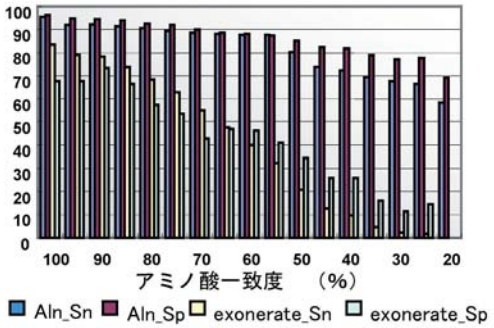


図2　ALNのエキソン予測精度（%）

2.2　統計情報に基づくタンパク質遺伝子の発見研究

複数のab initio遺伝子発見プログラムの予測結果を組み合わせて信頼性の高い遺伝子発見を行なうプログラムDIGITは、ベンチマークデータによる評価では、既存のab initio遺伝子発見プログラムが検出する数多くの擬陽性エキソンを取り除き、とりわけ遺伝子レベルの信頼性が大幅に改善されることが報告されている[3]。本研究開発では、DIGITのヒトゲノム配列への網羅的で体系的な適用を行いその実用性を検証した。DIGITを最新のヒトゲノム配列（Build36）の既知遺伝子（Ensemble Gene）以外の領域に応用し、未発見遺伝子を新たに3363個予測し、これらの予測遺伝子の多くは機能ドメインを持つことを確認した。DIGITを応用した最近の研究では、ヒト11番染色体でDIGITが新たに予測した未発見の遺伝子座65個のうち34個（52%）の遺伝子座で70個の転写産物の発現がRT-PCR法によって確認され、これらの新規遺伝子の多くが機能ドメインを持つことが報告されている[4]。これらの新規遺伝子うち、76%（26/34）はnested PCRでしか発現が確認されない極めて低発現な遺伝子で、61%（43/70）は組織特異的に発現する遺伝子であった。これらの低発現、組織特異的発現の遺伝子は過去の大規模なヒトcDNAプロジェクトでは発見できなかった遺伝子であり、DIGITを応用すると、他の手法では検出不可能な遺伝子を発見できることが実証された。ヒトの遺伝子カタログを完成させるには、このように高精度な遺伝子発見プログラムの予測結果を基に実験により発現解析を行う方法論を他の染色体にも応用することが必要である。また、新規遺伝子の実験による発現確認をより多くより効率的に行うには、より高精度な遺伝子発見プログラムが必要である。本研究開発では、ab initio遺伝子発見プログラム以外の予測結果とDIGITの予測結果を組み合わせてさらなる予測精度の向上に成功した。具体的には、EST配列のヒトゲノム配列へ写像データとDIGITの予測結果と組み合わせたDIGITest、比較ゲノム解析の手法を用いた遺伝子発見プログラムALNGGの予測結果とDIGITの予測結果を組み合わせたmetaDIGITを新たに開発して予測精度の評価を行い、予測精度の向上を確認した（表1）（表2）。

表1: ab initio遺伝子発見プログラムとDIGIT、DIGITestの予測精度の比較（ヒト22番染色体を使用）

Program	Sensitivity (Gene)	Specificity (Gene)	Sensitivity (Exon)	Specificity (Exon)
FGENESH	15.0 (69 / 461)	7.9 (69 / 874)	73.0 (2857 / 3913)	49.4 (2857 / 5781)
GENSCAN	9.5 (44 / 461)	5.5 (44 / 804)	73.6 (2879 / 3913)	43.4 (2879 / 6628)
HMMgene	14.1 (65 / 461)	4.3 (65 / 1524)	65.3 (2554 / 3913)	37.0 (2554 / 6903)
DIGITv0	6.7 (77 / 461)	18.8 (77 / 410)	68.5 (2682 / 3913)	71.6 (2682 / 3747)
DIGITv1	14.1 (65 / 461)	13.5 (65 / 481)	72.0 (2819 / 3913)	66.7 (2819 / 4226)
DIGITest	17.6 (81 / 461)	16.0 (81 / 507)	76.3 (2987 / 3913)	67.1 (2987 / 4451)

表2: DIGIT、ALNGG、metaDIGITの予測精度の比較（ヒト18番、22番染色体を使用）

Program	Sensitivity (Gene)	Specificity (Gene)	Sensitivity (Exon)	Specificity (Exon)
DIGITv1	6.3 (79 / 1260)	9.9 (79 / 801)	61.1 (4751 / 7781)	63.7 (4751 / 7454)
ALNGG	7.1 (89 / 1260)	20.1 (89 / 443)	37.1 (2884 / 7781)	79.0 (2884 / 3649)
metaDIGIT(hsp)	6.7 (85 / 1260)	14.1 (85 / 603)	64.0 (4976 / 7781)	69.2 (4976 / 7189)
metaDIGIT(hsn)	7.9 (100 / 1260)	12.8 (100 / 780)	66.1 (5147 / 7781)	66.5 (5147 / 7743)

2.3　ゲノム比較によるタンパク質遺伝子の発見研究

PHINAL[5]はヒトとマウスのゲノムDNAアライメントを入力として、ヒトのタンパクコード遺伝子を予測するプログラムである。タンパクのコード領域は種間で高度に保存されているが、高頻度な同義置換のために特にアミノ酸レベルでの保存度が高い。このことを利用したアミノ酸マッチングのスコアと、またアライメント上での種間の読み枠のずれを検出するための新たな指標（phase-index）を導入することで、PHINALでは特異性の高い遺伝子予測を実現していた。本研究では、アミノ酸マッチングスコアを同義・非同義置換の頻度の違いを加味したスコアへと拡張し、またスプライス部位や開始・終止コドンのスコアリングにも比較情報を用いることで更なる精度の向上を目指した。

種間でアミノ酸が一致するのは、コドンが完全に一致する場合と同義置換が起こった場合とがある。高度に保存された領域では、たとえ非コード領域であっても、見かけ上のコドンが完全に一致する割合は高くなる。一方で、全塩基置換に占める同義置換の割合はコード領域の方が有意に高いため、この頻度をスコアに導入することで保存された非コード領域をコード領域と区別できると期待される。

スプライス部位や開始・終止コドンなどは比較的短い配列をもとに判別することになるため予測の特異性は必ずしも高くない。ここではこれらのシグナルの予測モデルを確率モデルにより構築し、各シグナルのスコアをヒトとマウス配列からの同時確率（対数オッズスコアの和）とすることで精度の向上を果たした。

以上のモデルを統合することで、エキソンレベル（境界、読み枠が完全一致）での感度を落とすことなく特異性の向上を果たした。さらに遺伝子レベル（全エキソンが一致）では感度、特異性とも大幅に改善されている。さらなる精度の向上を目指し、マウスに加えてイヌゲノムとのアライメントを用いて同様のモデルを構築し予測を行った。イヌはマウスに比べてヒトとの相同性が高い（全体で80%以上）ため、ペアワイズアライメントに基づいたコード領域の予測精度はマウスほど高くはないが、3種のゲノムを用いることで予測特異性の更なる改善が見られた。以上の改良により、より実用性の高い遺伝子発見プログラムが構築できたと考えられる。

表3: Gene prediction accuracies for the human chromosome 22

	Exon		Gene	
	Sn	Sp	Sn	Sp
PHINAL (HS-MM-CF)	0.73	0.80	0.28	0.34
PHINAL (HS-MM)	0.74	0.78	0.27	0.32
PHINAL (old version)	0.74	0.68	0.20	0.16
TWINSCAN	0.74	0.61	0.12	0.10
SGP2	0.71	0.60	0.17	0.11

2.4　プロセス型偽遺伝子の発見研究

○偽遺伝子探索の潮流

偽遺伝子とは、機能遺伝子のコピーで機能をもたない配列、および機能を失った遺伝子の総称である。真核生物の偽遺伝子は、生成機構から（1）DNA重複型偽遺伝子と（2）プロセシング済み偽遺伝子の2種類に区別されている。どちらの種類の偽遺伝子も何らかの起源遺伝子のコピーであるので、遺伝子／偽遺伝子の判別問題は繊細である。往々にして、予測遺伝子群に偽遺伝子が混在しており、また機能遺伝子が偽遺伝子と誤審される場合もある。このような理由から、全ゲノム解読後のゲノム解析において、偽遺伝子予測が活況を帯びている[6]。

○予測パイプラインの確立

本研究開発では、このプロセシング済み偽遺伝子をヒトゲノム配列の構築型番（ビルド）に依存せずに体系的、網羅的に特定する方法の開発を進めた。

本手法は①ヒトゲノム配列に対して、Ensemblプロジェクトの転写産物配列を相同性検索し、莫大な数の遺伝子類似配列を検出、②プロセシング済み偽遺伝子とDNA重複型偽遺伝子を識別するための計算処理をおこなう。最初に、UCSCのBLATサーバーと、BLATの公開ソースコードでは検索結果が異なるため、後者の実行条件を検討し、ローカルサーバーでも安定した検索結果が得られるようにした。次に、プロセシング済み偽遺伝子とDNA重複型偽遺伝子を識別するために考案した手順を実装したプログラム（ProsPect）を、相同性検索結果の一次データセットに適用する。

○最新ビルドへの適用

このパイプラインを用いて、最新のゲノムビルド・転写産物配列からプロセシング済み偽遺伝子を予測した。まず、ヒトゲノム配列（ビルド35）に対して、Ensembl プロジェクトの転写産物配列（b35）をタンパクコード配列に絞って相同性検索し（33,869 配列）、遺伝子以外の相同配列を検出した。この一次データをProsPectで処理し、プロセシング済み偽遺伝子配列を予測した。得られた配列総数は3,011であった。

○精度・感度評価

ProsPectの予測結果を評価するため、他のグループによる偽遺伝子の予測結果との比較検討をおこなった（下表）。

表4: Borkグループとの比較

	Bork グループ（全予測数）	ProsPect との共通予測数
Processed	15,926	1,051（34.9%）
Non-Processed	1,718	109（3.6%）
Ambiguous	1,045	54（1.8%）
合計	18,689	1,214（40.3%）

表5: PseudoPipeとの比較

	PseudoPipe（全予測数）	ProsPect との共通予測数
Processed	8,094	973（32.3%）
Duplicated	2,871	171（5.7%）
Ambiguous	3,516	61（2.0%）
FP	3,546	33（1.1%）
合計	18,027	1,238（41.1%）

表6: HOPPSIGENとの比較

	HOPPSIGEN（全予測数）	ProsPect との共通予測数
Processed	4,078	547（18.2%）

他グループとの共通予測数から判断すると、ProsPectが誤ってプロセシング済み偽遺伝子と判定したDNA重複型偽遺伝子の割合は9-15%程度である。このことから、ProsPectによる2種類の偽遺伝子の識別精度はおよそ90%と見積もられる。

一方、他グループによるプロセシング済み偽遺伝子の全予測数は、4,078から15,926とまちまちであるが、仮にPseudoPipeによる8,094をゲノム中の全数として採用すると、ProsPectの検出感度は全体の12%程度である。これは、ProsPectが予測配列と起源遺伝子間の分子系統情報を維持するために網羅性を犠牲にしているのが一因である。この点は現在改良中である。

ProsPectの全予測配列中の6割以上のものが、他グループの予測と重ならない。ProsPectと他グループの顕著な違いは、ProsPectが相同配列検索のクエリに転写配列を用いる点である（他グループは全てタンパクコード領域）。ProsPect独自の予測配列中に、未知の遺伝子様配列が存在する可能性があるので、引き続き精査したい。

○生物学的知識の発見

ヒトゲノムのプロセシング済み偽遺伝子を精査したところ、興味深い偽遺伝子配列を見出した[7]。これは、この偽遺伝子探索プロトコルの適用が、生物学的新知見の発見に結びついたひとつの成功例であり、本研究開発の有効性と今後の更なる発展の可能性を示している。

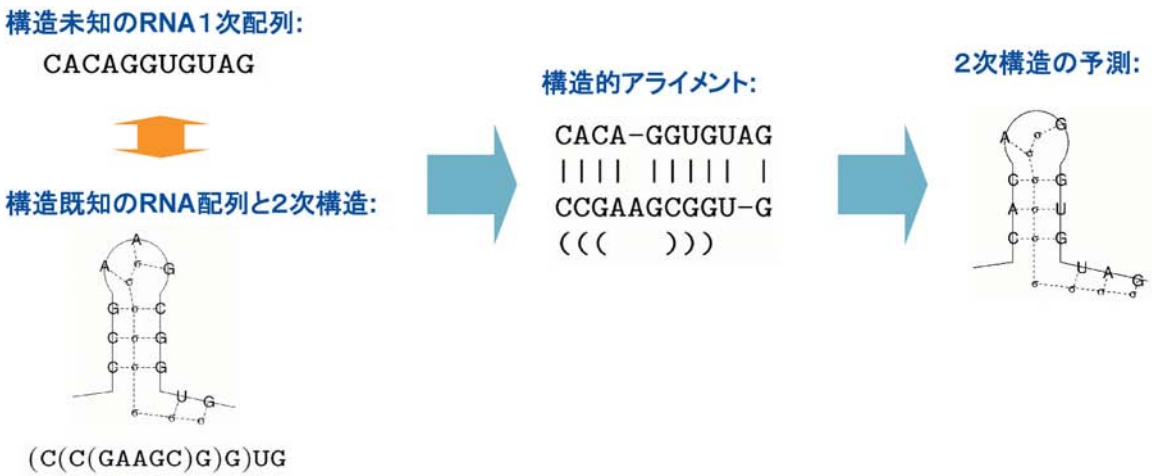
【参考】

ヒトゲノムのプロセシング済み偽遺伝子を精査したところ、興味深い偽遺伝子配列を見出した。エキソンシャッフリングは、ゲノムの不等交叉やレトロトランスポゾンによるトランスダクション（挿入近傍配列を他の座位へ運搬）によって生じると考えられている。trans-splicing（異種mRNA間のスプライシング）に関する最近の研究の進展により、新しいエキソンシャッフリングのメカニズムが提案されている。それは、RNAレベルでのエキソン混成と、それに続くLINE1

(L1）の転移機構（cDNA合成とゲノムへの挿入）に依存した新しい遺伝子の形成である。大島等は、このような遺伝子再構成の新機構により誕生したと考えられるヒト遺伝子を見出した。この遺伝子は、異なる2種類の遺伝子のRNAがスプライシング過程で連結し、L1の転移機構により転座した構造を持つ。新遺伝子は精巣特異的発現を示し、細胞内局在性やキナーゼ活性が親遺伝子とは変化している。またユビキチン化タンパク質との結合活性を示すことから、新たなユビキチン受容体タンパク質であると考えられる。これは偽遺伝子の機能に関する新たな事例の発見であり、偽遺伝子を重視する本研究開発の意義を裏付けている。

2.5　配列比較によるRNA遺伝子の発見研究

非コードRNA領域の網羅的探索と機能の同定を行うために確率文法を用いたシステムの開発を行った。具体的には、現在バイオインフォマティクスの分野において広く用いられているペア隠れマルコフモデルという手法が、ペアワイズアライメントや比較ゲノムに対して強力な解析手段を与えるものとして用いられているので、ペア隠れマルコフモデルを木構造上に拡張したPHMMTSという確率モデルを使用する。次にこの確率モデルを効率よく計算するシステムを動的計画法に基づいて開発を行った。さらにこのシステムを用いて、転移RNA配列とリボゾームRNA配列に対して、計算機実験を行った。その結果、RNA配列のアライメントや2次構造予測の問題に対して非常に精度の高い解が計算されることが証明できた。



次に、RNAの機能予測とゲノム上での発見においては、その2次構造とともにシュードノット構造を考慮することが正確な予測には不可欠となる。木接合文法と呼ばれる形式文法をペア確率文法に拡張することにより、シュードノット構造も考慮した構造的アライメントを行う手法を提案した。さらに、シュードノットデータベースを用いて計算機実験を行い、既存の構造予測手法と比べて予測制度が非常に優れていることを示した。

一方、配列比較を行うためのアライメントの精度を向上させるためには、精度の高いアライメントアルゴリズムの設計と信頼度の高いスコア行列の計算が必要とされる。とくに、RNA配列の2次構造を考慮した構造的アライメントにおいては、塩基間のスコアだけでなく、塩基の相補的ペア間のスコアを必要とするという従来のアライメントにはない特徴がある。今までは、非常に

アドホックなスコア行列を用いていたが、この問題を解決するために、条件付き確率場を用いて信頼度の非常に高いスコア行列を計算する方法を開発した。そして、すでに構造的アライメントが行われているRNA配列のデータベースから訓練データを取得して、本手法を適用することにより既存の提案されているどのスコア行列よりも信頼度の高いスコア行列を計算することに成功した。

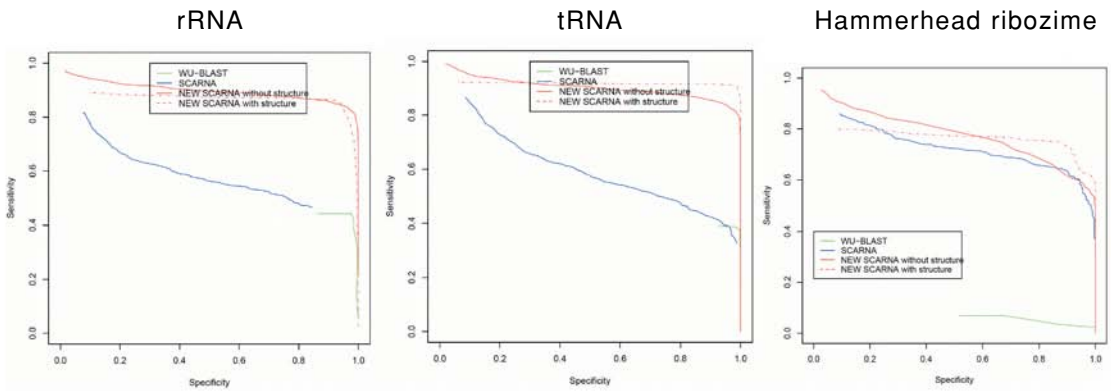
最後に、膨大なゲノム配列中から機能性RNA配列を発見して同定するためには、確率文法だけを用いて識別するには限界があり、より強力な識別手法が必要とされる。サポートベクターマシンを用いて、機能性RNA配列の識別と探索を行うために、新しくカーネル関数（ステムカーネル）を提案して、RNAファミリーの識別に関する計算機実験を行い、その有効性を示した。

2.6　RNA遺伝子のab initio的発見研究

ゲノム配列上のRNA遺伝子を、既知RNA遺伝子配列の二次構造情報と配列情報と総合した局所検索によって発見する手法を開発した。ステム候補のアラインメントによる、二次構造を考慮したRNA配列の比較・整列ソフトウェアScarnaを改良し、確率的スコアリングに基づく非コードRNA検索システムを実現した。

検索配列と検索対象配列共にMcCaskillのアルゴリズムにより、各塩基のペアーが塩基対を形成する確率を表現する塩基対確率行列を計算する。計算された塩基対確率上をウィンドウサーチすることによってステム候補を抽出し、ステムの候補は一塩基毎に分割して、さらに、5’部分と3’部分が分割される。それを、ステムコンポーネントと呼ぶ。2つの配列から抽出されたステム候補どうしを動的計画法でアライメントする[14]。整合性の取れないステムを除去した後、配列相同性のみを考慮してステム以外の塩基領域のアライメントを行う。アライメントスコアがある閾値以上ならば、そこをヒット領域と見なし、スコアとヒットした検索対象の配列上の位置を出力する。アライメントスコアは文献[14]とは異なり、RIBOSUMのみによりスコアリングされる。

Rfam SEEDを用いて、5S rRNA, tRNA, Hammer head ribozymeよりデータセットを作成し、計算機実験を行なった。二次構造を与えた場合、与えなかった場合のそれぞれについて、配列相同性50％～75％の場合のROC曲線を以下に示す。いずれの場合も、スコアリングの改良により、sensitivity、specificityとも向上させることが出来た。



2.7　タンパク質遺伝子の発見プロトコルの確立

ここでは、転写産物による遺伝子発見プログラムAln、ab initio遺伝子発見プログラムDIGIT、ゲノム比較による遺伝子発見プログラムPhinalを用いてヒトゲノムから網羅的に遺伝子を発見するプロトコルの確立を実施した。最終的なプロジェクトにおける遺伝子発見プロトコルは以下の通りである。

まず始めに、Ensemblプロジェクトが遺伝子を発見することができなかったゲノム領域にAlnを適用し、遺伝子を発見する。この時、スプライシングのバリアントを併わせて検出する。そして、Alnが遺伝子を発見できなかった領域を取り出し、その領域にDIGITとPhinalを適用する。

上記プロトコルにより予測されたタンパク質遺伝子の精度を検証するために、ヒトゲノムBuild35に対するAln, DIGIT, Phinalによる予測結果を、Ensemblアノテーションの Build35と36の差分と比較した。その結果、Aln の予測結果のうち611Transcripts（401 Gene)、DIGIT の予測結果のうち234 Transcripts（Gene)、Phinalの予測結果のうち180 Transcripts（Gene）が、Ensemblアノテーションの Build35と36の差分とoverlapをした。これらはそれぞれ、各予測結果の7.1%、14.1%、4.3%（いずれもTranscriptsレベル）であり、またEnsemblの差分が約1,500遺伝子であることを考えるとその約4割を予測できていたことになり、これらプログラムの予測精度の高さが改めて示される結果となった。

2.8　遺伝子発見データベースHALの開発

上記解析結果を整理・格納・表示・検索するデータベースHAL（Human genome Annotation Library）を開発し、http://hal.genome.ist.i.kyoto-u.ac.jp/から公開した。HALでは、発見された遺伝子のゲノム上での位置やGC含量、CpG アイランド、反復配列、マーカーなどの情報がグラフィカルに表示され、それらの一次情報へのリンクも豊富に用意されている。

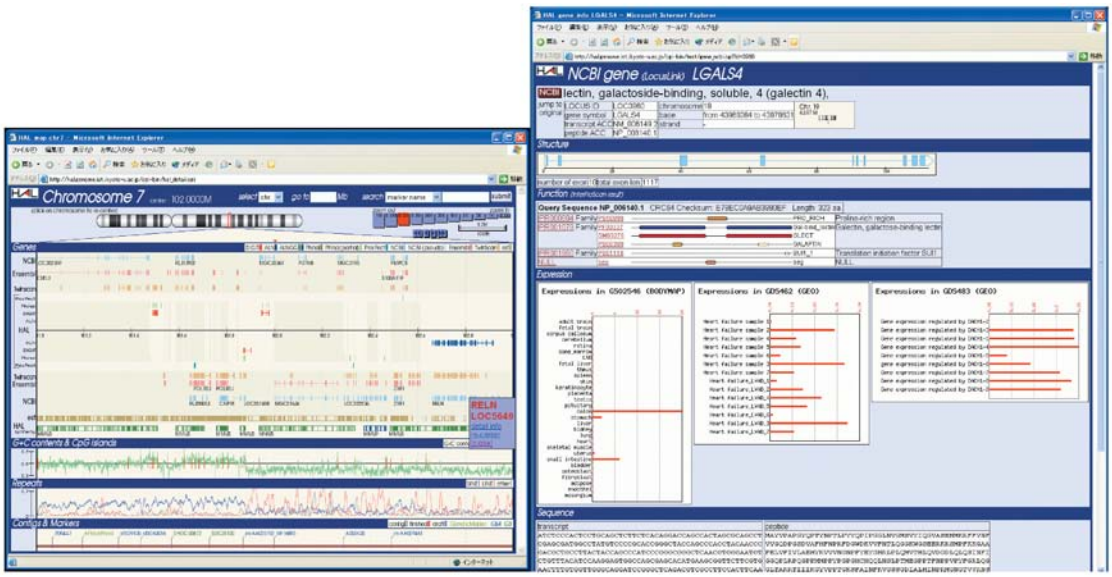


図1　HALの画面例（左：ゲノムビューア、右：遺伝子ビューア）

3. まとめ

本プロジェクトにより、ヒトゲノムに潜むタンパク質遺伝子を高い信頼性で発見するプロトコルが確立された。冒頭で述べたNature誌の発表以降に同定されたおよそ1,500個の新規遺伝子のうち、本プロトコルは約40％を予測していた。遺伝子の組織特異的な発現や時期特異的な発現、さらに転写産物の収集における実験感度を考慮すると、残り60％を疑陽性として片付けてしまうのは時機尚早というものだろう。

偽遺伝子の発見についても、遺伝子の形成過程に関する新しい知見を見つけることができ、一定の成果が得られた。しかし、各研究グループが報告する偽遺伝子カタログには未だ大きな差異が存在するので、その理由を熟考する必要があろう。

一方、RNA遺伝子の発見では、有望な要素技術の幾つかを確立することに成功したが、ゲノム規模の探索には課題が残った。計算量の問題と疑陽性検出の問題である。これらの問題を解決するためには、ここで論じた方法とは異なるアプローチを検討してみる必要がありそうだ。なかでも、プロモーターの配列情報に基づいた遺伝子発見（標的遺伝子発見）は考察する価値がありそうだ。これまでの知見に従えば、プロモーターの配列情報のモデル化には、文脈自由文法のような計算量の大きなモデルは必要ない。しかし、モチーフ発見の難しさから、このアプローチは疑陽性を多く検出してしまう。最近の免疫沈降法や系統フットプリント法などのデータをうまく組み合わせれば、この困難を克服できるのではなかろうか。

4. 研究開発実施体制

代表研究者　矢田 哲士（京都大学大学院情報学研究科）

研究開発題目

(1) 広義の遺伝子発見研究
　グループリーダー　矢田 哲士(京都大学大学院情報学研究科)

(2) 配列類似性に基づくタンパク質遺伝子の発見アルゴリズムの研究
　グループリーダー　後藤　修(京都大学大学院情報学研究科)

(3) 統計情報に基づくタンパク質遺伝子発見アルゴリズムの研究
　グループリーダー　十時　泰(理化学研究所GSC)

(4) 比較ゲノムによるタンパク遺伝子発見研究
　グループリーダー　野口 英樹(東京大学大学院新領域創成科学研究科)

(5) 偽遺伝子の探索アルゴリズムの改良と適用
　グループリーダー　大島 一彦(長浜バイオ大学)

(6) 配列比較によるRNA遺伝子の発見研究
　グループリーダー　榊原 康文(慶應義塾大学)

(7) RNA遺伝子のab initio的発見研究
　グループリーダー　浅井　潔(東京大学大学院新領域創成科学研究科)

5. 参考文献

[1] Gotoh, O., Morita, M., Ichiyoshi, N., Yada, T.(2005)Discovery of protein coding genes through chromosome-to-chromosome sequence comparison, GIW 2005, P103.

[2] Gotoh, O. 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(2000)

[3] Yada, T., Takagi, T., Totoki, Y., Sakaki, Y., Takaeda, T. DIGIT: a novel gene finding program by combining gene-finders Pac Symp Biocomput , 375-387(2003)

[4] Taylor TD, et al. Human chromosome 11 DNA sequence and analysis including novel gene identification. Nature 2006, 440, 497-500.

[5] Noguchi, H., Yada, T., Sakaki, Y. A novel index which precisely derives protein coding regions from cross-species genome alignment. Genome Inform Ser Workshop Genome Inform 13, 183-191.(2002)

[6] Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y. and Okada, N. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol. 4:R74(2003)

[7] Babushok, D.V., Ohshima, K., Ostertag, E.M., Chen, X., Wang, Y., Okada, N., Abrams, C.S., Kazazian, H.H.-Jr. Novel Testes Ubiquitin Receptor, S5a-like, Arose by Exon-Shuffling in Hominoids.(Submitted)

[8] Ohshima, K. and Okada, N. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res. 110, 475-490(2005)

[9] Matsui, H., Sato , K., and Sakakibara, Y. : Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures, Bioinformatics, Vol.21, No.11, 2611-2617, 2005.

[10] Sato, K. and Sakakibara, Y. : RNA structural alignment with conditional random fields, Bioinformatics, Vol.21, Suppl.2, ii237-ii242, 2005.

[11] Sakakibara, Y. : Grammatical inference in bioinformatics, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.7, 1051-1062, 2005.

[12] Sakakibara, Y. : Learning context-free grammars using tabular representations, Pattern Recognition, Vol.38, No.9, 1372-1383, 2005.

[13] Sakakibara, Y., Asai, K., and Sato, K. : Stem Kernels for RNA Sequence Analyses, submitted, 2006.

[14] Tabei, Y., Tsuda, K., Kin, T. and Asai, K. SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. *BIOINFORMATICS* 22, 1723-1729(2006)