

課題名 ヒト遺伝子の転写・発現の多様性解明を目指した基盤データベースの開発

代表研究者 京都大学大学院情報学研究科 助教授 矢田 哲士

1. 代表研究者による成果概要報告

1-1. 研究開発のねらい

ヒトゲノム配列決定プロジェクトにおける最大の興味は、ヒトを形作るタンパク質遺伝子の網羅的な探索である。先頃、配列決定コンソーシアムが Nature 誌で行なった報告によると、ヒトゲノムから発見されたタンパク質遺伝子の総数は 22,287 個であった。この数は、ショウジョウバエにおけるタンパク質遺伝子の総数とほぼ同じであり、ヒトの複雑さを説明するにはあまりに小さい数である。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多様性に注目が集まり、その仕組みの解明がポストシーケンス時代の極めて重要な研究課題となっている。

ところで、このヒト遺伝子の総数は妥当なものだろうか？ 結論から先に述べると、この数はヒト遺伝子数の下限値として捉えるべきである。Nature 誌によると、これらの遺伝子は、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、ここで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。

ヒト遺伝子の転写や発現における多様性の仕組みを解明するためには、その基盤となる遺伝子カタログデータの整備が必須である。本研究課題では、異なるアプローチに基づく高精度の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する。そこでまず、従来法を上回る信頼性を示すさまざまなタイプのタンパク質遺伝子の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力を注ぐ。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う。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 遺伝子の発見にも取り組む。

本研究課題では、独自に開発した遺伝子発見プログラムを積極的に活用する。そのため、ここで開発されたデータベースは、他の類似プロジェクトからは提供することができないユニークなデータで構成される。また、各遺伝子発見プログラムの高い信頼性とそれらの組み合わせ的な活用によって、このデータベースには、未知の遺伝子が豊富に含まれていることが期待される。

1-2. 研究開発の成果

本研究課題では、網羅的で信頼性の高いタンパク質遺伝子の発見研究に取り組んだ。そこでまず、従来の遺伝子発見プログラムを信頼性の面で大きく上回るさまざまなタイプの遺伝子発見プログラム(転写産物による遺伝子発見プログラム、ab initio 遺伝子発見プログラム、ゲノム比較による遺伝子発見プログラム)を開発した。さらに、それらを組み合わせた遺伝子発見プロトコルの確立に力を注ぎ、そのプロトコルによる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った。

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 遺伝子の発見に取り組んだ。

□ 転写産物による遺伝子発見プログラム

ゲノム上の未知の遺伝子を同定するために、既知の転写産物(cDNA 塩基配列、相同タンパク質のアミノ酸配列)の情報が利用できれば、信頼性の高い予測が可能となる。そのための計算機プログラム AIn を開発した。AIn の基本方針として、ゲノム上のイントロン候補領域をあらゆる組み合わせで読み飛ばしながら転写産物との最良アラインメントを行うスプライストアラインメントアルゴリズムを採用した。参照配列がアミノ酸配列の場合には概念的な翻訳を同時に行い、アミノ酸レベルでの類似性を基準とする。これを効率よく行うために、翻訳情報を内包した塩基配列表現法 TRON(Translating codON)コードを考案した。上記の類似性情報に加え、AIn は ab initio 遺伝子予測に用いられるコーディングポテンシャル、スプライシングシグナルなどの統計情報、イントロンの長さの分布を最適化の基準に含めた。そのため、遠縁の相同アミノ酸配列を参照配列とした場合にも高い予測精度を保つことができた。また、長大なゲノム配列データへの適用を鑑み、省メモリアルゴリズムの実装を併せて行なった。

AIn の性能を評価するために、配列決定コンソーシアムが Nature 誌における遺伝子発見で中心的に利用した転写産物による遺伝子発見プログラム GeneWise との比較を行なった。さまざまなベンチマークデータを用意したところ、全てのデータセットについて AIn の優位性が明瞭に示された。実行速度についても、AIn は GeneWise に比べ約一桁速いことが示された。したがって AIn は、高い実用性を備えているものといえる。

□ Ab initio 遺伝子発見プログラム

既知遺伝子との配列比較に頼ることのない ab initio 遺伝子発見プログラムは、網羅的な遺伝子の発見には必須である。しかし、ab initio 遺伝子発見プログラムは数多くの偽陽性エキソンを検出してしまうため、実用性に乏しいと言われてきた。そこで、複数の遺伝子発見プログラムの解析結果を組み合わせ、遺伝子を発見する汎用的なアルゴリズムを開発し、そのアルゴリズムを計算機プログラム DIGIT(Digit Integrates Gene Identification Tools)にまとめた。DIGIT は、ひとつの遺伝子発見プログラムが検出したエキソンより、幾つかの遺伝子発見プログラムが同時に検出したエキソンを尤もらしいと考える。DIGIT のモデルパラメータは、ヒトの ab initio 遺伝子発見プログラムを組み合わせるために最適化されている。

DIGIT の性能を評価するために、GENSCAN をはじめとする ab initio 遺伝子発見プログラムとの性能比較を行なった。さまざまなベンチマークデータを用意したところ、全てのデータセットについて DIGIT の優位性が明瞭に示された。すなわち DIGIT は、感度を犠牲にすることなく、偽陽性を大きく低減することに成功していた。このことは、DIGIT がヒトゲノムからの遺伝子発見における初の実用的な ab initio 遺伝子発見プログラムであることを示している。現在は、DIGIT の感度を改善するために、EST 配列のゲノムへの写像情報を取り込んだ遺伝子発見アルゴリズムの研究を行ない、予備的ではあるが、良好な結果が得られている。

□ ゲノム比較による遺伝子発見プログラム

転写産物による遺伝子発見では、偽陽性が検出されることは少ないが、全く未知の遺伝子を探索する

ことが原理上不可能である。一方、ab initio 遺伝子発見は未知の遺伝子を探索することが可能であるが、数多くの偽陽性を検出してしまう。そこで、二つないし複数の類縁ゲノム配列同士を比較することにより、偽陽性少なく未知遺伝子を同定する方法の開発に着手した。

ゲノム比較による遺伝子発見プログラム Phinal(PHase-INdexed ALignment) は、まず、ヒトゲノムとマウスゲノムのシンテニー領域の配列をアラインメントし、続いて、アラインメントで浮かび上がった保存領域から遺伝子の領域を取り出す。アラインメントでは、コドンの3文字目に置換が蓄積されやすいことを利用し、できるだけオルソログなコード領域がアラインされるように工夫している。また、遺伝子領域の取り出しでは、新しい指標として、コード領域のアラインメントと非コード領域のアラインメントに観察されるギャップの入り方の違いに着目している。

数百 Mb におよぶ哺乳動物の染色体どうしの比較を実用的な時間内に行うためには、従来の計算法では負荷が大きすぎる。そこで、次の4段階からなる階層的な戦略を用いて遺伝子を発見するプログラム Alngg を開発した。第1段階では、塩基の不連続一致パターンに合致するシードを高速に検索し、それらを種としてギャップを含まない初期クランプを求める。第2段階では、お互いに矛盾する初期クランプを取り除いたシンテニー表を作成する。第3段階では、シンテニーに含まれる近隣の初期クランプを集積して、エキソンに相当するクランプを作成する。最終段階では、これらのクランプを組み合わせて一つないし複数遺伝子の集合として整合性のあるクラスタを見いだす。クラスタ内部の構文解析により個々の遺伝子を同定する。

これらのプログラムは、いずれも開発途上のプログラムである。これまでのところ、さまざまなベンチマークデータについて従来法を大きく上回る予測精度が得られている。同時に、幾つかの問題点も明らかになった。Alngg は、両端のエキソンや単一エキソン遺伝子の検出が困難であった。また Phinal は、遺伝子を分割して検出する傾向が強いことが明らかになった。現在、プロモータ領域予測と組み合わせることや、複数のゲノムを同時に比較することにより、さらなる性能の向上を目指している。

□ 遺伝子発見プロトコール

網羅的で信頼性の高いヒトゲノムからの遺伝子発見を実現するために、さまざまなタイプの遺伝子発見プログラムを組み合わせたプロトコールを確立した。ヒトゲノムの半分近くは反復配列に覆われ、それらはタンパク質のコード領域とほとんど重複しないことが知られている。そこでまず、RepeatMasker を利用してゲノム配列に存在する反復配列をマスクし、この領域に遺伝子発見プログラムが誤って遺伝子を検出しないようにする。このマスクされたゲノム配列に対して、偽陽性エキソンを検出しないものから順番に遺伝子発見プログラムを適用する。すなわち、まず始めに Aln を適用して遺伝子を発見し、続いて Aln が遺伝子を発見できなかったゲノム領域に DIGIT と Phinal を適用する。DIGIT と Phinal の併用では、全体の60%に当たるエキソンが93%の特異度で正確に検出されることが示された。このことは、このプロトコールの確立によって、ヒトゲノムからの自動的な遺伝子発見が大きく実用に近づいたことを示している。現在、さらなる信頼性の向上を目指して、Alngg を取り入れたプロトコールの確立を考察している。

□ プロセス型偽遺伝子の発見

ヒトゲノム配列上に散らばる偽遺伝子を探索する方法を開発し、ヒト遺伝子のカタログ化に貢献する

ことを目的とした作業を行ってきた。ヒトゲノムには、細胞の mRNA が逆転写反応により cDNA 化し、ゲノムに再挿入されて生じた偽遺伝子(プロセス型偽遺伝子)が大量に存在する。このプロセス型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で網羅的に特定する方法の開発を進めた。まず、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、遺伝子以外の相同配列を検出した。次にプロセス型偽遺伝子と DNA 重複型偽遺伝子を識別するために、構造解析と進化解析を組み合わせた判別を行なった。以上の処理により、ヒトゲノム全体から 3,664 個のプロセス型偽遺伝子を取り出すことに成功した。最も多くの偽遺伝子が検出された遺伝子は keratin 18 (KRT18)(52 個)であり、次いで glyceraldehyde-3-phosphate dehydrogenase (GAPD)(43 個)、ribosomal protein L21 (RPL21)(38 個)と続いた。

□ RNA 遺伝子の発見

ヒトゲノムに潜む RNA 遺伝子を発見するために、ここでは、配列比較による遺伝子発見と ab initio 遺伝子発見の両面からパイロット的な研究を開始した。配列比較による遺伝子発見では、隠れマルコフモデルを木構造上の確率モデルに拡張することで RNA 配列を比較することに成功した。また、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 RNA 配列の構造的アライメントを行う手法を開発した。ab initio 遺伝子発見では、確率モデル上のカーネル法を適用することによって、二次構造が未知の RNA 配列に対して潜在的な二次構造の対応関係を考慮しながら配列の比較と検索を行う手法を考案した。また、RNA 遺伝子の候補領域をゲノム配列から高速に絞り込むために、ステム候補列のアラインメント法を用いた配列比較・検索ソフトウェアを開発している。これまでのベンチマークテストでは、いずれの手法も良好な性能を示しており、RNA 遺伝子をゲノム配列から網羅的に探索するシステムの中核となるものと期待されている。

□ データベースの開発

以上の研究成果による解析結果を整理・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 は、NCBI が提供する最新のヒトゲノムデータに合わせて更新される。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。また、NCBI や Sanger 研究所で行なわれている遺伝子発見プロジェクトのようすを併せてブラウズすることができる。HAL は、アノテーションデータの GFF ファイルを用意すれば、簡単にその情報を取り込むことができ、今後のアノテーション情報の増加やユーザ独自のアノテーション情報の追加に容易に対応することができる。

1-3. まとめ

本研究課題で開発したデータベース HAL(Human genome Annotation Library)は、ヒトゲノムに潜む未知のタンパク質遺伝子の情報を提供するデータベースである。HAL は、NCBI が提供する最新のヒトゲノムデータに合わせて更新され、タンパク質遺伝子のゲノム上での位置や構造に加え、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る。また、それらの一次情報へのリンクも豊富に用意され、さらに、NCBI や Sanger 研究所で行なわれている遺伝子発見プロジェクトの

ようすを併せてブラウズすることができる。HAL は、アノテーションデータの GFF ファイルを用意すれば、簡単にその情報を取り込むことができ、今後のアノテーション情報の増加やユーザ独自のアノテーション情報の追加に容易に対応することができる。HAL は、<http://hal.genome.ist.i.kyoto-u.ac.jp/>より公開されている。

HAL を構築するために、我々は、3つのタイプのタンパク質遺伝子の発見プログラム、すなわち転写産物による遺伝子発見プログラム、ab initio 遺伝子発見プログラム、ゲノム比較による遺伝子発見プログラムを新たに開発し、それらを体系的にヒトゲノムに適用するプロトコルを確立した。さまざまなベンチマークテストや実験的な検証によって、これらのプログラムやプロトコルが高い信頼性と網羅性を持つことが確かめられたが、一方で、これらに内在する限界も明らかになった。この限界を打破するために、我々はプログラムやプロトコルの改良を検討するとともに、偽遺伝子の発見研究や RNA 遺伝子の発見研究に取り組んでいる。これらの取り組みは、従来のタンパク質遺伝子に着目した狭義の遺伝子発見研究から、偽遺伝子や RNA 遺伝子の発見を含めた広義の遺伝子発見研究に本研究課題が発展したことを意味している。

2. 統括の見解

転写産物を利用したものや ab initio 法によるもの、ゲノム配列比較によるものなど世界的に見ても優れた性能を示す遺伝子発見プログラムを開発した意義は大きい。またいくつかの遺伝子発見プログラムやこれらにより予測されたヒト遺伝子候補がデータベース化されて公開されており、様々な生物種のゲノム解読が完了しつつある現在においてライフサイエンス分野の研究者に対しても本研究開発の貢献が見込まれる。本課題ではゲノム配列解析を基盤とした遺伝子予測が進められたが、今後は違った観点からの遺伝子発見の挑戦も期待される。

3. 主な論文発表

国内 0 件、国外 10 件 主なものを以下に記す

1. Sogawa, K., Numayama-Tsuruta, K., Takahashi, T., Matsushita, N., Miura, C., Nikawa, J., Gotoh, O., Kikuchi, Y., Fujii-Kuriyama, Y. A novel induction mechanism of the rat CYP1A2 gene mediated by Ah receptor-Arnt heterodimer. *Biochem. Biophys. Res. Commun.* **318**, 746-755 (2004).
2. Yada, T., Totoki, Y., Takaeda, Y., Sakaki, Y., and Takagi, T. DIGIT: a novel gene finding program by combining gene-finders. *Pac. Symp. Biocomput.* **8**, 375-87 (2003).
3. Noguchi, H., Yada, T., and Sakaki, Y. A novel index which precisely derives protein coding regions from cross-species genome alignments. In *Proc. of Genome Informatics Workshop 2002*, pages 183-191 (2002).
4. Ohshima, K. and Okada, N. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenet Genome Res.* Single topic issue on Retrotransposable Elements and Genome Evolution. **in press**.
5. 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).