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

京都大学大学院情報学研究科
○矢田 哲士

Development of the complete Human gene database toward the understanding of transcription and expression diversity.

Tetsushi Yada
Department of Intelligence Science and Technology,
Graduate School of Informatics,
Kyoto University

HAL (Human genome Annotation Library) is a database which provides novel protein coding genes in the human genomes. We newly developed three types of gene-finders, that is, similarity based gene-finder, *ab initio* gene-finder and comparative gene-finder. We applied them to the human genome sequences in a systematic manner and successfully identified thousands of candidates of protein coding genes. Our assessments including experimental verification clearly showed that these candidates contain novel true genes abundantly. We further assigned functional descriptions to the candidates and stored them in HAL database accompanying with their genomic loci. HAL provides graphical user interface to browse/search them effectively. HAL is now open to the public via the internet (http://hal.genome.ist.i.kyoto-u.ac.jp/).

1. はじめに

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 Nature 誌で行なわれた報告によると、ヒトゲノムから発見されたタンパク質遺伝子の総数はショウジョウバエとほぼ同じ22,287個であり、ヒトの複雑さを説明するには少なすぎる。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クエ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? 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、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。

そこで本研究課題では、既存手法では発見できない遺伝子を主なターゲットとし、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した。まず従来法を上回る信頼性を示す様々なタイプ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ー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、さらには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んだ。

2. 研究開発の成果概要

2.1 3 つのアプローチによるタンパク質遺伝子発見プログラムの開発およびゲノムへの適用

本プロジェクトでは、従来法を上回る信頼性を示す遺伝子発見プログラムとして、配列類似性に基づいた手法による Aln、*ab initio* 法による DIGIT および比較ゲノム法による Phinal を開発した。

- Aln[1]

Aln は、ゲノム配列を概念的に翻訳した Tron コードを用いてアミノ酸配列とのアライメントを行うことによる配列類似性に基づいた遺伝子発見プログラムである。より微弱な類似性をとらえるために、多重アラインメントから導かれる「プロファイル」を参照とすることも可能とし、境界シグナルの他に、コーディングポテンシャルもスコアに取り入れた。これらの結果 Aln は、Ensembl で用いられている GeneWise に比べて明らかに優れた予測精度を示した。GeneWise は参照配列として用いるアミノ酸配列の一致度が下がるにつれて予測精度が低くなるが、Aln では30～40%アミノ酸の一致があれば GeneWise の適用限界と同程度の予測精度が得られる。HAL プロジェクトではこの結果に基づいて、40%以上一致する予測結果全てを拾い上げているが、Ensembl プロジェクトでは、基準値が70%以上となっている。

- DIGIT[2]

DIGIT は *ab initio* 遺伝子発見プログラムで、多数の遺伝子発見プログラムの予測を組み合わせることに基づくアルゴリズムである。一般的に DIGIT は、多数の遺伝子発見プログラムにより同時に予測されるエキソンに高いスコアを与える。具体的にはファーストエキソンは FGENESH と HMMgene、インターナルエキソンは FGENESH と GENSCAN、ラストエキソンは FGENESH と HMMgene、シングルエキソンは GENSCAN と HMMgene の組み合わせを用いている。これらの組み合わせに基づいて、各既存プログラムのエキソンスコアを logistic 関数を用いて正解率に変換し、ベイズ推定を用いて共通の新しいエキソンスコアを導き出した。また、エキソン・イントロン構造をフレームを保証して表現する隠れマルコフモデル(HMM)を作成し、新しく導き出されたエキソンスコアに基づいて HMM の状態遷移確率を Baum-Welch アルゴリズムを用いて最適化した。これらの最適化により、DIGIT は既存の他の *ab initio* 遺伝子発見装置と比較して、エクソンレベルの specificity を sensitivity を低下させることなく著しく高めていることが確かめられた。

- Phinal[3]

Phinal は、ヒトとマウスの配列類似性に基づいた比較遺伝子発見プログラムである。まずシンテニー領域のゲノム配列をアライメントし、遺伝子候補はよく保存された領域から抽出される。保存領域からエキソン領域を効果的に抽出するために、コード領域における非同義置換のとりやすさを指標の一つとして用いた。配列の類似性に加えて、ヌクレオチドの挿入・欠失によりフレームシフトが引き起こされた場合に、読み枠を埋め合わせる形で他の挿入・欠失がしばしば見つけられるという観察結果に基づいた新しいインデックスも導入した。他の比較ゲノムによる遺伝子発見手法と比べて Phinal の予測精度は特異性が著しく高いことが確認された。

2.2 偽遺伝子および RNA 遺伝子の探索手法の開発

タンパク質をコードする遺伝子に加えて、プロセス型の偽遺伝子を探索するプログラム ProsPect4) も合わせて開発した。この手法は下記の事実に着目したものである。複数のエキソンから成る遺伝子とのアライメントでは、転写産物はエキソン数に対応して複数のアライメント・ブロックに分断されるが一方、プロセシング済み偽遺伝子とのアライメントでは、転写産物は全く分断されないか、あるいはプロセシング済み偽遺伝子の生成後に生じた別の挿入による擬似的な少数の分断しか見られないことが多い。さらに平成16年度からは、タンパク質をコードする遺伝子に加えて RNA 遺伝子を探索するアルゴリズムも合わせて開発に取り組んでいる。

2.3 ゲノムからのタンパク質遺伝子発見プロトコル

HAL プロジェクトでは、上述の3つのタンパク質遺伝子発見プログラムを下記のプロトコルに従って適用することで、完全なヒト遺伝子カタログ構築を目指している。Ensembl の予測結果は非常に正確なので、Ensembl プロジェクトによって予測されていないゲノム領域を探索範囲と設定する。まず、3種の遺伝子発見手法の中で最も偽陽性の少ない Aln を探索範囲のゲノム配列に適用する。次に、遺伝子が Ensembl または Aln によって予測されていないゲノム領域に DIGIT および Phinal の両方を適用する。DIGIT および Phinal の予測精度は Aln ほど高くはないため、同一ゲノム領域に対し DIGIT と Phinal の両方を適用し、それらの結果を比較検討することで予測精度の向上を図っている。ヒトの22番染色体配列を用いたテストでは、両プログラムによって共通に予測されたエキソンの特異性は93%に達する。

2.4 HAL データベースの開発

上述のプロトコルに基づき、最新の NCBI Human Genome Assembly Build34および Ensembl リリース23.34e.1を用いてタンパク質遺伝子、偽遺伝子の予測を実施した。その結果3,465の遺伝子が Aln によって予測され(9,781転写産物)、2,200の遺伝子が DIGIT によって予測された。また、4,868の遺伝子が Phinal によって予測された。これら予測遺伝子は、Ensembl や NCBI のような他のプロジェクトによる予測遺伝子とともに、InterPro による機能アノテーションや発現プロファイルデータへの対応付けが施され HAL データベースに格納された。HAL データベースは WWW 経由でインタラクティブなユーザインターフェースを通して利用可能である(http://hal.genome.ist.i.kyoto-u.ac.jp/)。 HAL ウェブサイトでは、ゲノムの一領域を拡大してその領域内に含まれる遺伝子、EST 情報、GC の内容、CpG islands、マーカー情報が俯瞰できる Genome Viewer および個々の遺伝子に対する詳細な情報を提供する Gene Viewer から主に構成されている。また、ユーザが目的のデータに容易にたどりつけるように、各種検索機能、ホモロジー検索機能、ユーザデータのアップロード機能も合わせて提供されている。

3. まとめ

本プロジェクトでは、ヒトゲノム配列中に潜むタンパク質をコードする遺伝子を網羅的に探索することを目標とし、3つの異なった手法に基づいた遺伝子発見プログラムを開発、改良を進めた。また、これらプログラムを効果的にゲノムに適用するためのプロトコールも合わせて開発し、ゲノム全体に適用することで信頼度の高い遺伝子候補の抽出に成功した。これら候補に対する InterPro を用いた機能予測結果、あるいは試験

的に実施した実験による発現の確認を通してもこれら予測遺伝子の信頼性は高いと言えるであろう。今後はアルゴリズム・プロトコルの更なる改良に加え、RNA 遺伝子の発見などにも力を入れていく予定である。

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) 遺伝子発見プロトコルの確立とデータベース開発
　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) 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）

(5) 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）

(6) 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. 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)

[2] 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)

[3] 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)

[4] 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)