公募研究：2000年度

真核生物の比較ゲノム情報解析

●後藤　修

埼玉県立がんセンター研究所

〈研究の目的と進め方〉

ヒト全ゲノムのドラフト配列も公表され、線虫、昆虫、植物においても、それぞれの代表の全ゲノム配列が明らかになってきた。複数のゲノム配列の比較を通じて、ゲノム塩基配列に書き込まれた高次の情報を読み解くことが本研究の主たる目的である。すなわち、ゲノム配列上のタンパク質コード領域とそれ以外の機能部位を予測するための配列比較法を開発し、それを用いて、未知遺伝子の発見、正確な遺伝子内部構造（エキソン・イントロン配置）の予測、転写や複製制御シグナルの同定などへの応用を試みる。

〈研究開始時の研究計画〉

スプライシングシグナル定量化の基となるエクソン・イントロン境界データを大量に得るため、ゲノム配列とそれに由来するcDNAあるいはEST配列との比較を効率よく行うプログラムをまず作成する。得られた境界配列について種々の統計解析を適用し、高性能なスプライシングシグナルの導出を試みる。

ゲノム配列対アミノ酸配列間のアラインメントに基づく真核生物遺伝子構造予測に関しては、まず対象を線虫だけにしぼって、できるだけ多くの遺伝子を試すことにより、実際上の問題点を洗い出すとともに、パラメータの最適化やアルゴリズムの調整を行う。特に、現実のイントロンやエキソンの長さの分布に配慮できるようアルゴリズムの修正を図る。その後、生物種毎に固有な最適条件を探索する。また、多数の相同遺伝子が共存する多重遺伝子族について、それらの遺伝子構造を共同的に推測する方法を確立する。この方法は複雑な計算工程を伴うので、一連の工程を可能な限り自動化するための手続きを開発する。

〈研究期間の成果〉

上記研究計画の内容は、1年間の課題としては大きすぎるものであり、ここではその後の継続研究の成果も一部交えて報告する。

ゲノム配列とcDNAあるいはEST配列との比較には、当初の予想より多くの困難を伴ったが、各種ノイズに対して安定して動作するプログラムALNを作成した(0309031734)。これを用いて、当時完全長cDNA配列が得られていたヒト遺伝子について、選択的スプライシングを含む極めて精密な構造地図を作製した(Mizuno et al. GIW 2003, 412)。ヒト(Nagasaki et al. GIW 2003, 424)を含む6真核生物種について選択的スプライシング、選択的転写開始の網羅的発見と型分類を行い、各生物種に固有の特徴を抽出することができた(0602101752)。また、スプライシング境界部位を精度よく予測するための統計的手法を開発した(Yamamura & Gotoh, GIW 2003, 426)。

既知のアミノ酸配列との相同性を利用した真核生物遺伝子構造予測に関して、各生物種へのパラメータの最適化、イントロン長の分布を考慮したスコア体系の確立、使用メモリー量の削減など当初の計画に沿ってALNプログラムの改良を行った(0304271021)。特にヒトにおいて、ENSEMBLデータベースの作成ツールとして著名なGeneWiseに比べ、顕著に高い予測精度をALNが示すことを実証した(Gotoh & Sugimori, GIW 2002, 398)。なお、ALNは入力配列の種類（核酸またはアミノ酸配列）を自動判別して適切な処理を行うよう設計されている。ALNはソースコードを公開するとともに、Web上での実行サービスを行っている(0309031734)。

多数の相同遺伝子の共同的構造予測法の開発も行い、他の予測法より有意に高い信頼性を持つことを確認した。現在、100%の自動化を目指した改良を進めている。

〈国内外での成果の位置づけ〉

我々の開発したALNプログラムは、統計情報と配列の相同性情報を統合した初めての真核生物遺伝子予測法として教科書(Handbook of Molecular Computational Biology, Aluru S. ed.)にも紹介されている。また、麹菌ゲノムアノテーション(Machida et al. Nature, 438, 1157)や、ゲノム上のGPCR(Ono et al. Gene, 364, 63)やチトクロームP450遺伝子の網羅的発見に利用され、その実用性が実証されている。特にヒトゲノム上の網羅的遺伝子発見プロジェクト(http://hal.genome.ist.i.kyoto-u.ac.jp/)で重要な役割を果たしている。

〈達成できなかったこと、予想外の困難、その理由〉

この研究課題は多くの要素的問題を含み、それらひとつひとつを解決していく必要があった。当初の目的を最終的にはほぼ達成することができたものの、そのためには予想を大幅に上回る時間が必要であった。ゲノム上の遺伝子密度、遺伝子あたりの平均エキソン数、イントロン長の分布などには生物種により大きな違いがあることが分かり、それらを考慮したパラメータ最適化手法の開発が未だ達成できていない。

〈今後の課題〉

ALNのWeb上での実行サービスは未だ試験的なもので、ヒトゲノムにしか対応していない。現在、様々な生物のゲノムに対応した拡張を進めており、近々改良版の公開を計画している。

高い予測精度を持つことを第一の目標としてALNを開発したため、他のプログラムに比べて多くの実行時間を要する。今後、現在の精度を保ったままALNをより高速化する必要がある。

研究開始時には、複数のゲノム配列比較法の開発も念頭に置いていた。この課題については現在準備段階であり、これからの取り組みを待たなければならない。

〈研究期間の全成果公表リスト〉

1) 論文／プロシーディング

1. 0304271021
後藤修, 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, 統計数理　50(1), 3-15 (2002).

2. 0602101752
Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M., and Gotoh, O., Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes, Gene, 364(1), 53-62 (2005).

2) データベース／ソフトウェア

1. 0309031734
汎用アラインメントプログラムＡＬＮ：http://www.genome.ist.i.kyoto-u.ac.jp/~aln_user/ALN/top.htm.