（様式第９　別紙２：公開版）

養成技術者の研究・研修成果等

１．養成技術者氏名：　　水野　政彦

２．養成カリキュラム名：バイオインフォーマテイクス（ゲノム配列解析）研究者養成カリキュラム

３．養成カリキュラムの達成状況

実験情報に裏打ちされた完全長ヒト遺伝子データセットを収集したが、他に公開されているデータセットに比べ、スプライス部位の信頼性が格段に高いという意味で、満足の行くものであり、これらに関する論文が投稿間近になっている。また、このデータを解析した結果、遺伝子構造においてはエクソン、イントロン長に特徴的な分布が見られ、スプライス部位周辺には興味深いコンセンサスが得られてきており、2 報目の投稿も含めた研究に発展する可能性が出てきた。既に理化学研究所、Aventis 社(Paris)、弘前大学、宮崎大学から本研究のデータの使用を打診されており、共同研究への発展も期待できる。論文発表と同時にデータベースとしてインターネットでも公開していく予定である。今後、本データが遺伝子予測やスプライス部位解析のための標準テストデータとなることを期待している。

４．成果

背景および目的

計算機による真核生物のゲノム配列からの遺伝子の予測率は 45%程度に留まっている。予測アルゴリズムの多くは既知の遺伝子配列および構造データをもとに予測を行っており、これらのデータに大きく依存することが指摘されている。ゲノムや EST（配列発現タグ）のシーケンスにより配列データは大量に手に入るようになったが、その精度については明確な指標がなく、遺伝子予測が基準とする学習データ中の配列の読み取りエラーや予測に過ぎないアノテーションを正解として扱うために起こる誤りが遺伝子予測に及ぼす影響はほとんど検討されてこなかった。そこで我々は実験的に確からしいヒト遺伝子の配列と構造を核酸配列データベース GenBank の精査により選択し、さらにそれらに EST 配列をマッピングすることで、全長性、スプライス部位という遺伝子構造を複数のデータによって確認し、高精度のヒト遺伝子データセットを構築した。

データ選択

核酸配列データベース GenBank (rel. 126) から 1) ***Homo sapiens***、2)核由来、3)完全長 mRNA 前駆体 (pre-mRNA) 配列（genomic DNA 配列）、4)実験で確認されたアノテーション（計算機予測を除く）、5) EST（one pass sequencing）を除く、の条件でデータを収集した。

結果および検討事項

我々はまずヒト遺伝子配列（830 個）を GenBank アノテーションの精査により選択し、それらに EST 配列（195,384 個）をマッピングして、全長性、スプライス部位という遺伝子構造を確認した。GenBank アノテーションでは全長と判断されたにもかかわらず、ヒト遺伝子（830 個）のうち 80 個は EST 配列との比較から部分長と判断され除いた。残った 750 個中で配列類似度 50%以上の偏りを除くと最終的に 723 個が残った。これらを Refseq 配列（19,755 個）と 5’端について比較したところ、723 個中 70.3%の全長性が支持された。

この配列セットから合計 11,982 個のスプライス部位が同定されたが、この内 GenBank アノテーションのみから発見されたものが 25.3%(3,032 個)に上った。また、定常的スプライス部位（2,245 個）、および選択的スプライス部位（4,167 個）が分類、同定された。この結果から 723 個中 67.4%がスプライス・バリアントを持つことが判明した。EST 配列はスプライス部位あたり平均 107.9 個マッピングされ、平均してその 83.3%がアライメント上その部位でスプライスされた。定常的、および選択的スプライス部位では、平均 26.4 および 201.2 個の EST 配列がマッピングされ、それぞれ平均 100% および 68.1%がスプライスされた。今後、この配列セットを基準データとして解析を進めることにより、従来以上に高精度の遺伝子予測、スプライス部位の配列解析・予測が期待できる。

５．成果の対外的発表等

（１）論文発表（論文掲載済、または査読済を対象。）

1) “Human Full-length Pre-mRNA Sequence Dataset for Computational Gene Prediction and Alternative Splicing Analysis”, Proceedings of Genome Informatics 2003, 14:412-413, Masahiko Mizuno, Osamu Gotoh and Makiko Suwa （査読あり、2003 年 12 月 14-17 日、横浜、パシフィコ横浜）

2) “Experimentally Verified Data Construction of Human Full-length Pre-mRNA Sequences and Identification of Constitutive and Alternative Splice Sites”, The Genome of Homo Sapiens, 68th Cold Spring Harbor Symposium on Quantitative Biology, Masahiko Mizuno and Makiko Suwa （査読あり、2003 年 5 月 28 日－6 月 2 日, NY, US）

（２）口頭発表（発表済を対象。）

1) 「計算機によるヒト遺伝子の選択的スプライス部位同定の精度評価」　産業技術総合研究所 2003 年度ライフサイエンス分野融合会議・生命工学部会バイオテクノロジー研究会合同研究発表会・講演会 （要旨集 p30） 水野　政彦、諏訪　牧子（2004 年 2 月 3-4 日、つくば研究センター共用講堂）

2) 「ヒト遺伝子構造および選択的スプライス部位の高精度同定」　産業技術総合研究所　生命情報科学人材養成コース　設立 1 周年記念シンポジウム　水野政彦、後藤修、諏訪牧子（2003 年 10 月 3 日、お台場、日本科学未来館）

3) “Data Collection of Human Full-length Pre-mRNA Sequences for Computational Gene Prediction and Alternative Splicing Analysis”, The Fifth International Workshop on Advanced Genomics (abstracts p117), Masahiko Mizuno, Osamu Gotoh and Makiko Suwa （査読あり、2003 年 6 月 26-27 日、横浜、パシフィコ横浜）

（３）特許等（出願番号を記載）

なし。