

課題名 ヒトゲノムにおける広義の遺伝子発見研究

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. 研究目的

ヒトゲノムの完成配列の公開とともに、ゲノム中に潜むタンパク質遺伝子の総数が見積られた。その数は、ショウジョウバエのものとほぼ同じ 22,000 個ほどであり、ヒトの複雑さを説明するには余りに少ないものであった。この見積りは妥当だろうか？ この報告では、既知の転写産物との配列類似性によって遺伝子が同定されているため、網羅性には未だ疑問が残る。一方で、この報告により、「ヒトのタンパク質遺伝子数は 5 万～10 万」という従来の仮説も否定された。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、選択的プロモータ、選択的スプライシング、RNA 干渉などが注目されている。

本研究課題には、大きく二つのねらいがある。第一に、ヒトのタンパク質遺伝子の総数を再評価し、未だゲノム中に潜む未知のタンパク質遺伝子を網羅的に発見することである。そして、第二に、遺伝子の転写/発現調節の仕組みを解明するための情報処理技術を確立することである。これらの研究成果は、データベースあるいはソフトウェアの形で、速やかにインターネット上で公開する。

1-2. 研究開発概要

本研究課題では、タンパク質遺伝子、偽遺伝子、機能性 RNA 遺伝子の発見法を開発するとともに、それらを組み合わせたヒトゲノムのアノテーション法を開発する。アノテーション法のヒトゲノムへの適用の結果は、データベースにまとめ、インターネット上で公開する。

タンパク質遺伝子の発見法では、転写産物による発見法、*ab initio* 発見法、ゲノム比較による発見法を考察する。転写産物による遺伝子発見をゲノム全域に対して行なう場合、膨大な数の配列アラインメントを行なう必要があり、その計算量の削減が必須の研究課題となっている。ここでは、精度の犠牲を最小限に抑え、その計算量を大きく削減する試みを行なう。*ab initio* 発見法は、未知遺伝子の発見に有効であると言われているが、偽陽性も数多く検出してしまう。ここでは、未知遺伝子の発見に関する *ab initio* 法の性能を評価するとともに、偽陽性の検出を抑える技法を考察した。ゲノム比較法は、潜在的に高い予測精度を持つと考えられているが、その研究は始まったばかりであり、手法としての完成度は低い。ここでは、*ab initio* 法で培われた遺伝子発見のための統計情報をゲノム比較の枠組みで評価することで予測精度の改善を目指す。さらに、3 種以上のゲノム配列の比較による遺伝子発見法についての考察を行なう。

最近になって、幾つかの研究グループから体系的な偽遺伝子の発見が報告された。ここでは、それらとの比較を通して、現行の偽遺伝子発見法の評価を行なうとともに、発見結果からの生物学的

な知識の発見を試みる。機能性 RNA 遺伝子の発見法は、その研究の歴史は比較的古いものの、ゲノム全域への適用に関する性能は、全くの未知数である。ここでは、計算量と予測精度の両面について、あらゆる技法を考察し、実用的な発見法の確立に勤める。

本研究課題では、以上の手法を組み合わせたヒトゲノムのアノテーション法を開発するとともに、その適用結果をデータベースにまとめて公開する。データベースは、ヒトゲノムのビルドの進捗に合わせて更新する。更新においては、以前のアノテーションとの比較を通して、我々のアノテーション法の信頼性を考察する。

1-3. 研究成果概要

転写産物によるタンパク質遺伝子発見法の高速化を実現するために、アラインメントにおけるイントロンの存在の類推とイントロン部分のアラインメントの枝刈りを効果的に行なう枠組みを考案した。その結果、予測精度を犠牲にすることなく、3 倍ほどの高速化が実現できた。

本研究課題では、複数の *ab initio* タンパク質遺伝子発見法を組み合わせることで、感度の犠牲を抑えつつ、偽陽性の検出も抑えることに成功したが、さらに、この方法が発見した未知遺伝子の実験的な検証を行なった。その結果、半数以上の遺伝子について実際の転写産物が観察された。また、このアルゴリズムを *ab initio* 法とゲノム比較法の組み合わせに応用したところ、予測精度の有意な改善が観察された。

ゲノム比較によるタンパク質遺伝子の発見では、アラインメントされたコドンのペアやシグナル配列の塩基ペアについての統計量を導入したところ、予測精度の有意な改善が観察された。また、3 種以上のゲノム配列の比較による予測精度の改善を試みたところ、一定の成果は得られたが、信頼性の高いアラインメントを実用時間で計算する技法の開発、比較する種の組み合わせごとに行なうパラメータの調整技法の開発などに課題が残った。

偽遺伝子の発見では、我々のグループを含む幾つかのグループの体系的なアノテーションを比較したところ、その差異が想像以上に大きいことが明らかになった。そしてそれが、遺伝子の相同配列の検索クエリ(cDNA 配列かアミノ酸配列か)に起因することを突き止めた(各々の妥当性については注意深い考察が必要)。一方、アノテーションされた偽遺伝子を精査したところ、RNA レベルでのエキソン混成とそれに続く LINE1 の転移機構に依存して誕生したと考えられる遺伝子を見い出した。

機能性 RNA 遺伝子の発見では、ゲノム中の RNA 遺伝子の領域を近似的に高速に検索する技法の開発、高精度に RNA 配列のアラインメントを行なうためのアルゴリズムやスコア体系の開発、ゲノム中の機能性 RNA 遺伝子を識別するカーネル関数の開発などを行ない、先駆的な成果を得た。しかし、計算時間、予測精度ともに実用レベルには未到達であると考えている。

2. 事後評価結果

2-1. 当初計画の達成度

○当初の研究開発計画から見た達成度等はどうか

タンパク質遺伝子のアノテーションは大いに改良されたが、プロセス型偽遺伝子の探索法についてはさらなる改良が必要と思われる。また、RNA 遺伝子探索法については、現在問題となっている非コード遺伝子をカバーできるものとはなっていない。

○ 当初計画では想定されていなかった新たな展開が生じたか

バイオインフォマティクスの側面からは新たな展開が見えないが、生物学的側面からは従来の見方では不十分である点が示唆されている。

2-2. 知的財産権、外部発表（論文等）等研究開発成果の状況

論文発表だけでなくホームページでの公開等が進んでいる。成果は多く得られているが、それらを体系的にまとめるには至らなかった。

2-3. 研究開発成果の公開による波及効果

○データベース・プログラムや論文、学会発表など、これまでの成果公開によるバイオインフォマティクス領域およびライフサイエンス分野全体への影響や効果はどうか。

アクセス数やベンチマークの依頼数など具体的なデータが示されていないので、公開による影響の評価は難しいが、論文や学会発表は多くなされている。遺伝子予測という古くて新しい問題には多くの研究者が関心を持っており、一定のインパクトはあったと思われる。

○国内外の類似研究開発の動向と比較して本課題の意義はどうか。

国内の本分野をリードする研究者を集結し、体系的に遺伝子発見技術の開発に取り組み、海外の既存手法にせまるシステムを構築できたことに意義はあるが、世界をリードするスタンダードとなるにはブレークスルー的な展開が必要となっている今日において、代表研究者らの更なる努力が期待される。

2-4. 成果の実用化の可能性及び成果から予想される波及効果

○今後、成果の生命科学、研究基盤、産業への貢献度はどう見込まれるか。

システムのもっと使いやすいものにする努力は必要と思われるが、貢献は今後の生物学者との研究成果を出すことが必須となる。その中でヒト遺伝子探索に対する貢献を検証していく必要があると思われる。

2-5. 総合評価

ヒトの遺伝子発見研究に対して独自の研究アプローチを示せたことは意義がある。しかしながらプロジェクト内の各課題は個別に羅列されたままになっており、目標に向けた体系的な集中が欲しかったのも事実である。また、機能性 RNA 遺伝子探索については時間的なことのみを問題点としてあげてあるが、問題の本質は類縁 RNA 探索という考え方にあり、それだけでは非コード遺伝子

の解析としては不十分であることを認識するべきであろう。遺伝子予測は古くからのテーマであるが、今も決定版のない新しいテーマである。今後このシステムを使って生物学的発見がなされることを期待する。

3. 主な論文発表等

1. Yamada, S., Gotoh, O., Yamana, H. Improvement in accuracy of multiple sequence alignment using novel group-to-group sequence alignment algorithm with piecewise linear gap cost, *BMC Bioinformatics*, accepted. (2006)
2. Matsui, H., Sato, K., and Sakakibara, Y. : Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures, *Bioinformatics* **21**, 2611-2617 (2005).
3. 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).

参考

論文発表 国内 0 件、海外 10 件

口頭発表 国内 1 件、海外 1 件

ポスター発表 国内 11 件、海外 4 件