多重ゲノム配列アラインメントに基づく機能情報の抽出

●後藤 修
京都大学大学院情報学研究科

＜研究の目的と進め方＞

多重配列アラインメントは、遺伝子やタンパク質の機能、配列、構造、進化の間の関係を明らかにする上で中心的な役割を果たすバイオインフォマティクスの技法である。主にタンパク質のアミノ酸配列を対象として長年の研究成果が蓄積しているが、現在でもなお活発な研究が進められている。一方、近年の革新的な塩基配列決定技術の進歩に伴って、多数の生物種のゲノム塩基配列が決定されるようになってきた。ゲノム配列を対象とした多重配列アラインメントを求めることができれば、個々の遺伝子やタンパク質を対象とした場合と同様に、ゲノム配列に書き込まれた機能や構造に関する有用な知見が得られるものと期待される。具体的には、未知の遺伝子の発見やその正確な内部構造の同定、複製、転写、あるいはスプライシングなどの分子機構に関与する制御シグナルの同定などに貢献できるものと考えられる。

しかしながら、ゲノム配列の多重配列アラインメントを行うには多くの困難を伴うこともよく知られている。第一に、ゲノム配列は一つの遺伝子に比べ 10^4 ～ 10^6 倍も長いことが挙げられる。標準的な２配列間アラインメントの手法によれば、計算時間は配列の長さの二乗に比例するので、二つのゲノム配列を比較するだけでも膨大な計算時間を要することになる。また、ゲノム配列は進化の過程で転移、逆位、重複などの大規模な再編成をしばしば伴う。共線形性を仮定する通常のアラインメント手法では、これらの再編成を取り扱うことが困難である。仮に、ゲノム配列全域でなく、プロモータ領域などその一部分だけをアラインメントの対象とする場合でも、アミノ酸配列に比べ信頼性の高い結果を得ることが困難である。それは、アミノ酸配列には科せられる立体構造生成に関する制約を受けないため、塩基配列がより高い自由度の下で進化するためである。さらに、繰り返し配列の挿入などを考慮しなければならないことや、アミノ酸が 20 種類存在するのに対し塩基が 4 種類に限られることも塩基配列のアラインメントをいっそう困難にしている。

本研究課題では、これらの困難を乗り越えるための方法を考案し、原核生物から哺乳動物に至るさまざまな大きさのゲノム配列の比較解析を現実的な計算機資源の下で可能とするソフトウェアの開発を具体的な目標とする。我々はまた、真核生物のゲノム配列上に存在する遺伝子を発見し、その正確な内部構造を推定する方法の開発にも取り組んでいる。我々の方法は、ゲノム配列間あるいはゲノム配列と既知の転写産物配列とのアラインメントを利用したものであり、従来の統計的手法に比べ高い精度が期待できる。

＜ 2008 年度の研究の当初計画＞

本研究は、いくつかの段階的な過程を経て実現される。まず、２つの長大なゲノム DNA 配列どうしのアラインメントから、両ゲノム配列上の対応関係（シンテニー）を明らかにする。一方、同一ファミリーに属する遺伝子群をゲノム横断的に探索し、それらの内部構造（エキソン･イントロンの配置）を精度よく求める。両者の結果を併合することにより、複数ゲノム配列から効率よく相同領域（オーソロガス領域）を同定することができる。前者の目的には、前研究期間において新規開発した CGAT 法を適用する。本研究の中心課題はゲノム配列相同領域の多重アラインメント法の確立である。重複、転座、逆位など複雑な再編成を伴うゲノム配列に適合させるために CGAT 法の改良を試みる。特定された制御領域については、長い欠失・挿入の存在を考慮した多重アラインメント法である PRIME を適用する。このようにして得られた多重配列アラインメントから、特によく保存されたモチーフを抽出する。本計画ではオーソロガス領域に含まれる保存領域に着目するため、一般的なモチーフ抽出法に比較して偽陽性を削減できるものと期待される。一方、アミノ酸配列の保存性に基づくゲノム横断的な遺伝子予測法である FamilyWise に cDNA/EST 配列の情報を付加し、選択的転写開始・選択的スプライシングを含むより包括的な遺伝子予測法を確立する。

＜ 2008 年度の成果＞

前年度に引きつづき、転写産物のゲノム配列へのマップとスプライスアラインメントの高効率化を推進した。昨年度開発・公開した我々のソフトウェア Spaln（801222000, 805072020）は、cDNA/EST 配列を問い合わせとしたが、本年度はこれをさらに発展させ、アミノ酸配列を問い合わせとすることも可能とした。これにより、ゲノム配列を検索し、対応する遺伝子領域内で問い合わせアミノ酸配列を鋳型としたスプライスアラインメントを実行することにより、極めて高速かつ高精度に当該遺伝子翻訳領域の遺伝子構造を同定することが可能となった（表 1，812221037）。FamilyWise の実装に向けて、重要な進捗が得られたことになる。

Spaln を実行した結果、ならびに以前開発した選択的スプライシングの自動分類法（606212000）を用いることにより、正常組織と腫瘍組織における選択的スプライシングの様式差について検討を行った。生検組織から調整した試料に基づき、特定の腫瘍組織で選択的に発現または非発現するスプライス変異体を同定することに寄与できた（812221135）。

ゲノム配列を粗視化することにより、段階的に対象を限定していく戦略を用いた我々の CGAT 法は、他に類を見ない独自のゲノム配列アラインメント法である。この１年間の改良により、計算速度に関してはかなりの改良を果たせた。今後、比較対象ゲノム配列間の類似度に対する依存性など、その有効性についてさらに詳しく検討する必要がある。

＜国内外での成果の位置づけ＞

既知のアミノ酸配列を鋳型とした真核生物遺伝子の構造予測法は、ゲノムアノテーションの主要な情報源として広く用いられている。現在最も多く用いられているプログラムは欧州分子生物学研究所が開発したGeneWsie[1]およびその後継プログラムであるExonerate[2]であろう。表１はヒトおよびマウス遺伝子それぞれ約500を用いた、Spalnを含む様々なプログラムの性能試験の結果を示す。SpalnCSはアミノ酸配列ではなくcDNA塩基配列の翻訳領域を鋳型とした場合の結果を示すが、他はすべてヒトのアミノ酸配列とマウスの対応するゲノム領域、あるいはその逆の組合せを用いた結果である。

表から明らかなように、Spalnの精密モード（SpalnFD）が最も高精度である一方、高速モード（SpalnQA）が最も高速であった。Exonerateの実行速度はほぼSpalnQAに匹敵するが、遺伝子あるいはエキソンレベルでの予測精度には大きな差異が見られた。実際、Spaln以外では最も高精度であったGeneWsie（global mode）に比べてSpalnQAの実行時間は1/300以下であるにも関わらず、有意により高い精度を示した。

Spalnは限られたゲノム領域ばかりでなく、ゲノム配列全体から対応領域を検索する機能を備えている。哺乳動物程度の大きさのゲノム配列に関して、検索とアラインメントとを連続して実行できる実用的なソフトウェアは他に存在せず、この点だけをとってもSpalnの優れた独自性が示される。なお、アミノ酸配列を問い合わせとしたゲノム塩基配列の検索には従来Blast[3]やBlat[4]が用いられてきた。しかし、問い合わせ配列と対象配列との間の配列一致度が70%を超えるなら、Spalnはこれらの標準的な検索プログラムより一桁以上高速に対象領域を検索できることも明らかとなった。

＜達成できなかったこと、予想外の困難、その理由＞

これまでの試みにより、ブロック分割法によるゲノム配列間アラインメントが予期通りの性能を示すことが確認できた。本年度はこの方法の多重ゲノム配列アラインメントへの拡張が最も重要な課題として研究を継続してきた。しかし、現時点ではまだ満足すべき性能を得るに至っていない。ゲノム配列間アラインメントに基づく遺伝子予測法については、塩基配列の物理化学的性質に基づく遺伝子上流領域予測などいくつかの拡張を試みた。ある程度の性能向上につながることは確認できたが、まだ際だった進展を得ることができていない。これらの課題は本質的な困難さを抱えており、いくつもの角度から検討を継続していくことが必要であると思われる。

＜今後の課題＞

上述のように、ブロック分割法に基づく実用的な多重ゲノム配列アラインメントプログラムの作成が今後の最も重要な課題である。そのためには、ブロックレベルでの多重アラインメント法を確立するとともに、塩基レベルでの多重アラインメント作成法の性能向上にも努める必要がある。比較ゲノムに基づく遺伝子発見法で現在最も困難な課題である遺伝子境界の推定問題に関しては、DBTSSの内容やSpalnの実行結果、あるいはヒストン修飾などの外部情報を組み入れることで現実的な解決を目指す。アミノ酸配列の保存性に基づくゲノム横断的な遺伝子予測法であるFamilyWiseをより信頼性の高いものとするため、遺伝子構造の保存性を考慮した遺伝子予測法ならびにアミノ酸配列レベルでの多重アラインメント法を開発する。これとSpalnの実行結果を融合することにより、選択的転写開始・選択的スプライシングを含むゲノム横断的でより包括的な遺伝子予測法を確立したい。

表１.様々なスプライスアラインメントプログラムの性能比較

Method	Human Gene (%)			Mouse Gene (%)			*CPU*
	Gene	*Esn*	*Esp*	*Gene*	*Esn*	*Esp*	(min)
Projector	51.3	93.8	87.0	58.5	94.6	90.3	-
GeneAlign	82.3	96.7	97.1	79.2	96.6	96.4	-
Exonerate	56.8	86.9	92.6	55.0	86.3	92.7	2.27
SpalnCS	76.6	96.0	96.5	76.8	96.1	96.3	3.17
SpalnQA	87.4	98.1	98.1	85.3	97.3	97.4	2.25
SpalnQD	88.0	98.1	98.2	86.0	97.4	97.6	2.53
Nap-AAT	60.1	88.1	92.4	61.9	87.4	92.2	39.18
SpalnFA	88.1	98.2	98.3	85.5	97.4	97.5	97.91
SpalnFD	89.0	98.2	98.4	86.4	97.5	97.8	131.73
ExoneE	65.8	93.3	93.4	64.6	93.2	93.5	315.46
GeneWise	73.7	94.7	96.4	72.9	94.2	96.3	774.68
GWgl	80.7	96.2	97.3	82.1	95.7	97.4	784.80

*Gene*は遺伝子レベルでの予測精度を，*Esn*および*Esp*はエキソンレベルでの感度と選択性をそれぞれ示す。*ExoneE* は“exhaustive”モードでの*Exonerate* を、*GWgl*は大域モードでの*GeneWise*を表す。812221037より転載。

参考文献

1) Birney, E., et al. (2004) *Genome Res.*, 14, 988-995.
2) Slater, G. S. and Birney, E. (2005) *BMC Bioinformatics,* 6, 31.
3) Altschul, S. F., et al. (1997) *Nucleic Acids Res.*, 25, 3389-3402.
4) Kent, W. J. (2002) *Genome Res.*, 12, 656-664.

＜成果公表リスト＞

1) 論文／プロシーディング

1. 805072020
   Gotoh, O.: A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence, Nucleic Acid Res., 36, 2630-2638 (2008) .
2. 812221037
   Gotoh, O.: Direct mapping and alignment of protein sequences onto genomic sequence, Bioinformatics, 24, 2438-2444 (2008) .
3. 812221054
   Ichinose, N., Yada, T., Gotoh, O., Aihara, K.: Reconstruction of transcription-translation dynamics with a model of gene networks, J. Theor. Biol., 255, 378-386 (2008) .
4. 812221135
   Ohnuma, S., Miura, K., Horii, A., Fujibuchi, W., Kaneko, N., Gotoh, O., Nagasaki, H., Mizoi, T., Tsukamoto, N., Kobayashi, T., Kinouchi, M., Okabe, M., Sasaki, H., Shiiba, K., Miyagawa, K., Sasaki, I.: Cancer-associated splicing variants of the CDCA1 and MSMB genes expressed in cancer cell lines and surgically resected gastric cancer tissues, Surgery, 145, 57-68 (2009) .

2) データベース／ソフトウェア

1. 801222000
   Spaln: A space-efficient and accurate tool for mapping and aligning a set of cDNA sequences onto a genomic sequence
   URL: http://www.genome.ist.i.kyoto-u.ac.jp/~aln_user/