様式Ｃ－１９

# 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月３１日現在

機関番号：８２６２６
研究種目：基盤研究（B）
研究期間：2010～2012
課題番号：22310124
研究課題名（和文）　大量シークエンシング時代に向けた新規配列比較法の開発

研究課題名（英文）　Development of novel methods for sequence comparison in the era of massive DNA sequencing data

研究代表者
後藤　修（OSAMU GOTOH）
産業技術総合研究所・生命情報工学研究センター・招聘研究員
研究者番号：40142111

研究成果の概要（和文）：新型シークエンサーが解読したDNA配列断片に対応するゲノム配列を高速に検索し、イントロンの存在を考慮したアラインメントを行うプログラム「Spaln」の開発を行った。Spaln は他の同類のプログラムより有意に正確でかつ計算速度的にも優れていた。中ー長（>100bp）のcDNA配列を問い合わせとしたRAN-Seq法として用いた場合、現在最も広く普及している方法では対応できない長さや高い誤読率の断片でも、Spaln は高い感度と精度でマップできることを確かめた。

研究成果の概要（英文）：We focused on middle-to-long DNA reads generated by new generation sequencers, and developed a computer program named “Spaln” that maps and spliced-align a set of transcripts onto reference genomic sequences. When tested on full length cDNA or protein amino acid sequences as the queries, Spaln was proven to be most accurate and reasonably fast compared with any other similar tools examined. When Spaln is used as a tool for RNA-Seq analysis for middle-to-long (> 100bp) cDNA reads, Spaln is shown to be much more sensitive and accurate compared with currently popular RNA-Seq tools.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| ２２年度 | 6,700,000 | 2,010,000 | 8,710,000 |
| ２３年度 | 3,700,000 | 1,110,000 | 4,810,000 |
| ２４年度 | 3,700,000 | 1,110,000 | 4,810,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 14,100,000 | 4,230,000 | 18,330,000 |

研究分野：生命情報学
科研費の分科・細目：
キーワード：　次世代シークエンサー、アラインメント、ゲノム、RNA-Seq、比較ゲノム、スプライシング、アルゴリズム、アセスメント

## 研究開始当初の背景

(1) 研究開始当初、DNA塩基配列決定法に関して第三の革新が進行していた。いわゆる次世代DNAシークエンサーが一度に解読できる塩基長が数10塩基に限られていたのに対し、一分子測定技術がちょうどその頃開発され、数千塩基に上る配列決定が可能となりつつあった。

(2) DNAシークエンサーから読まれた大量の断片に対応するゲノム配列上の位置を見つける操作は、通常マッピングと呼ばれる。研

究開始前１～２年の間に MAQ、SOAP、ZOOM、Bowtie など、マッピングのための様々なツールが提案されていた。しかし、これらは、数 10 塩基長の DNA 断片を解析することに特化していたため、より長い断片への適用が困難であった。

１．　研究の目的

(1) ゲノム配列上のイントロンの存在を考慮した転写産物とゲノム配列とのアラインメントはスプライスアラインメントと呼ばれる。第一の目的は、生物種ごとに特徴のあるスプライスシグナルを考慮した、正確なスプライスアラインメント法の開発である。

(2) さらに、中ー長の cDNA 断片を参照ゲノム配列にマップ・アラインする高速・高性能のソフトウェアを開発し、公開することが本研究課題の主たる目的である。

２．　研究の方法

(1) 申請者自身がすでに開発していたプログラム「Spaln」は EST, 完全長 cDNA, アミノ酸配列を問い合わせとして、参照ゲノム配列とのマッピングおよびスプライシングを考慮したアラインメントを行うためのツールである。まず、EST 配列を自身のゲノム配列にマップ・アラインすることにより、信頼性の高いエキソン・イントロン境界配列を抽出した。

(2) 得られた境界配列を集計し、5’および 3’スプライスシグナルやイントロン長などのスプライシングに関わる統計量を求め、それに基づく種特異的なスプライシングシグナルを取得した。

(3) 上記とは独立に取得した完全長 cDNA 配列を用いてスプライスアラインメント検定用のベンチマークデータベースを作成した。

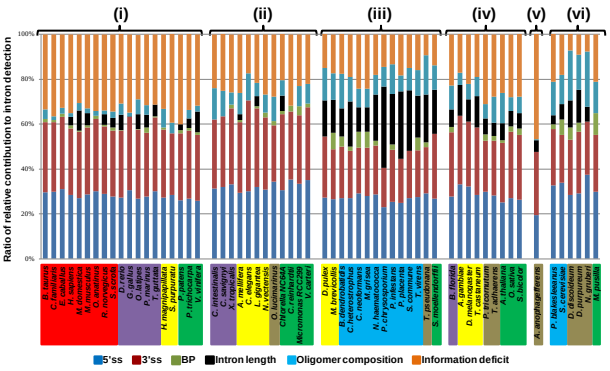
(4) 上記ベンチマークデータを用いて様々なスプライスアラインメントプログラムの性能を評価した。

(5) より短い DNA、ペプチド断片に対しても十分な検出感度が得られるように Spaln を改良し、上記と同様の検定を行った。

４．研究成果

(1) 約 60 生物種につき、スプライシングに関わるシグナル強度の種特異性を調べた。脊椎動物、陸上植物、真菌類などはそれぞれに固有の特徴を示した。一方、図１に示すように、系統的に遠い生物種のスプライシングシグナルがより近い生物種のものより類似している場合も見られなど、スプライシングシグナルの複雑な進化過程が示唆された。

図１．スプライシングシグナルの種特異性




(2) 完全長 cDNA、その翻訳領域（CDS）、タンパク質アミノ酸配列という３種類の転写産物を問い合わせとした時のアラインメント精度を、動物、植物、菌類の代表的な生物種について検証した。図２にまとめた結果が示すように、我々の開発した Spaln は検討したほとんどすべての条件下で、最高の精度を示すことが確かめられた。計算速度的にも、Sim4cc を除く他のプログラムと同等かそれ以上であった。

図２a．cDNA を問い合わせとした場合のアラインメント精度
語末の X などはオプションの違いを表す。

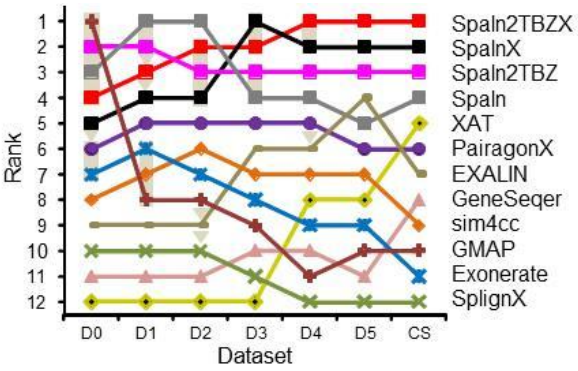
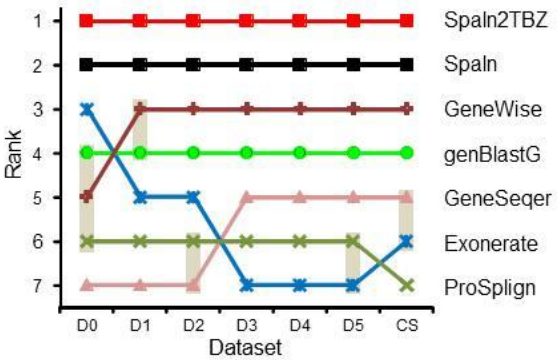


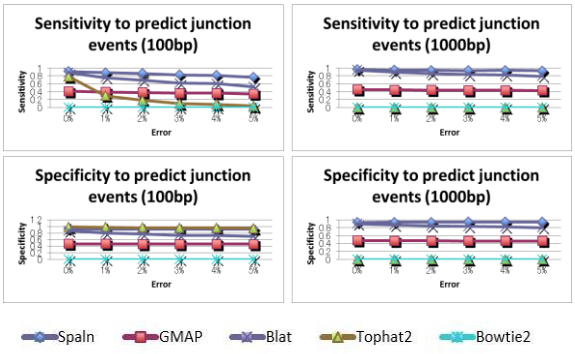

図２b．アミノ酸配列を問い合わせとした場合のアラインメント精度



(3) また、中ー長(>100bp)の長さを持つ cDNA 配列を効率よくゲノム配列にマップする RAN-Seq 法の開発も行った。いくつか試した中で、重なりを許した連続シードを用いる比較的単純な手法が最も高性能であることが判明した。これからの普及が見込まれる PacBio シークエンサーの出力に関し、シミュレーションデータと実データを用いた検証を行った結果、現在最も広く用いられている RAN-Seq 法である TopHat では全く対応できない長さや高い誤読率を持つリードでも、我々の方法は高い感度と精度でマップできることが確かめられた。図３はリード長が 100bp と 1000bp の場合における、各手法の感度、精度を表す。特にリード長が長い場合に Spaln の優位性が顕著であった。

一方で、Spaln は TopHat や Star などの接尾辞配列や Barrows-Wheeler 変換を用いる手法に比べ計算速度で劣るという欠点がある。並列化によりほぼコア数に比例する高速化が可能であるが、更なる高速化のために現在アルゴリズムの改良に取り組んでいる。

図 3. 図 3. 中-長 DNA リードのシロイヌナズナゲノムへのマッピング感度と精度




５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 7 件）


① Gotoh, O. (2013) Heuristic Alignment Methods, Methods in Molecular Biology, (Russel, D. ed.) in press.
② Gotoh, O. (2012) Evolution of cytochrome P450 genes from the viewpoint of genome informatics, Biol. Pharm. Bull., 35 (6), 812-817, 査読有. DOI: 10.1248/bpb.35.812
③ Iwata, H. and Gotoh, O. (2012) Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features, Nucleic Acids Res., 40 (20) e161, 査読有. DOI: 10.1093/nar/gks708
④ Ueshima, T., Kawamoto, T., Honda, K.K, Noshiro, M., Fujimoto, K., Nakao, S., Ichinose, N., Hashimoto, S., Gotoh, O., Kato, Y. (2012) Identification of a new clock-related element EL-box involved in circadian regulation by BMAL1/CLOCK and HES1, Gene, 510 (2), 118-125, 査読有. DOI:10.1016/j.gene.2012.08.022
⑤ Ichinose, N., Yada, T., Gotoh, O. (2012) A large-scale motif discovery using a DNA Gray code and equiprobable oligomers, Bioinformatics, 28 (1), 25-31, 査読有. DOI: 10.1093/bioinformatics/btr606
⑥ Iwata, H. and Gotoh, O. (2011) Comparative analysis of information contents relevant to recognition of introns in many species, BMC Genomics, 12, 45, 査読有. DOI:10.1186/1471-2164-12-45
⑦ Nakato, R. and Gotoh, O. (2010) Cgaln: fast and space-efficient whole-genome alignment, BMC Bioinformatics, 11, 224, 査読有. DOI:10.1186/1471-2105-11-224


〔学会発表〕（計 14 件）


① Zeng, C., Iwata, H., Ichinose, N., Yata, T., Gotoh, O. Detection of splice junctions from mid-to-long RNA-seq reads by Spaln, GIW 2012, 2012/12/12-14, National Cheng Kung University, Tainan, Taiwan.
② Zeng, C., Iwata, H., Gotoh, O. Mapping middle-to-long RNA-seq reads onto genomic sequence, JSBi 2012, 2012/10/15-17, タワーホール船堀, Tokyo（東京都）.
③ Gotoh, O. and Morita, M. (2012) Gene-structure-aware multiple protein sequence alignment as a tool for assessment of predicted eukaryotic gene structures, JSBi 2012, 2012/10/15-17, タワーホール船堀, Tokyo（東京都）.
④ Morita, M. and Gotoh, O. (2012) The examination of enhanced diversity in the substrate recognition sites of P450, 2012/6/22-26, The Regional Museum of Natural Sciences - Turin, Torino, Italy.
⑤ Gotoh, O. and Morita, M. (2012) Comparison of two strategies for comprehensive identification of P450

genes on multiple eukaryotic genomes, Cytochrome P450 Biodiversity and Biotechnology, 2012/6/22-26, The Regional Museum of Natural Sciences - Turin, Torino, Italy.

⑥ 後藤修（2011）配列アラインメントのこれまでとこれから、CBI/JSBi 合同大会、特別講演、2011/11/10、神戸国際会議場，神戸（兵庫県）

⑦ Iwata, H. and Gotoh, O. (2011) Evaluation of Spliced Alignment Programs using Cross-Species Benchmark Datasets, JSBi 2010, 神戸国際会議場, Kobe（兵庫県）.

⑧ 後藤　修（2011）ゲノム上の分布から見た P450 遺伝子族の進化、日本生化学会、2011/9/24、京都国際会議場（京都府）

⑨ Nakato, R. and Gotoh, O. (2010) A Method for fast and space-efficient whole-genome sequence alignment, in Proceeding of the 5th International Conference “Genomics, Proteomics, Bioinformatics and Nanobiotechnology for Medicine” GPBNM-2010, 2010/5/31-6/5, St. Petersburg, Russia.

⑩ Komiyama, Y. and Gotoh, O. (2010) RDF Curator: A Novel Workflow that Generates Semantic Graph from Literature for Curation Using Text Mining, Proc. Biocuration 2010, International Society for Biocuration, 2010/10/11-14, AIST Waterfront Bio-IT Research Building, Tokyo（東京都）.

⑪ Komiyama, Y. and Gotoh, O. (2010) A Related Graph Generation System for Biomass Data and Its Evaluation by Comparison with Manual Curation, JSBi 2010, Fukuoka（福岡県）.

⑫ Iwata, H. and Gotoh, O. (2010) Evaluation of Spliced Alignment Programs Including an Extended Version of Spaln, JSBi 2010/12/13-15, 九州大学医学部百年講堂, Fukuoka（福岡県）.

⑬ Ichinose, N., Yada, T. and Gotoh, O. (2010), Memory Management of Pairwise Alignment Using Garbage Collection, JSBi 2010/12/13-15, 九州大学医学部百年講堂,, Fukuoka（福岡県）.

⑭ Iwata, H. and Gotoh, O. (2010) Comparative analysis of information contents relevant to intron recognition in many species, 3rd Asian Young Researchers Conference for Computational and Omics Biology, 2010/3/10-12, National Cheng Kung University, Tainan, Taiwan.

〔図書〕（計１件）
後藤修（翻訳）（2011）「バイオインフォマティクス」(A. ポランスキ・M. キンメル著)、シュプリンガージャパン。

〔その他〕
ホームページ等
http://www.genome.ist.i.kyoto-u.ac.jp/aligner.php

６．研究組織
(1)研究代表者
後藤　修（GOTOH OSAMU）
独立行政法人産業技術総合研究所・生命情報研究センター・招聘研究員
研究者番号：40142111

(2)研究分担者
市瀬　夏洋（ICHINOSE NATSUHIRO）
京都大学・情報研究科・助教
研究者番号：70302750