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　６月　１日現在

機関番号：14401
研究種目：若手研究(A)
研究期間：　2010 年度　～　2012 年度
課題番号：22680023
研究課題名（和文）
次世代シークエンサーを用いた対立染色体配列の決定と発現解析法の精度向上
研究課題名（英文）
Accuracy improvement of allele and gene expression analyses by next generation sequencer
研究代表者
竹中　要一　（ TAKENAKA YOICHI ）
大阪大学･ 大学院情報科学研究科 ･ 准教授
研究者番号：00324830

研究成果の概要（和文）：
次世代シークエンサーから産生される DNA 断片を元に、対立染色体配列を決定する方法の研究を行い、有向非循環グラフ表現が有効である事を示した。また、遺伝子の発現解析手法の精度向上を目的とした研究を行った。細胞分化クロストークのモデル化、ベイズ推定手法の省計算量化、mRNA の反復ゲノムマッピングによる高精度アイソソフォーム発現量の同定、完全線形符号による DNA 符号化が有効である事を示した。

研究成果の概要（英文）：
I studied the method to determine the allele sequences of diploids from DNA fragments generated from next generation DNA sequencer and where the proposed algorithm used directed acyclic graph to express DNA subsequences. The study includes gene expression profile analyses such as a crosstalk model of cell differentiation, a stochastic method to reduce the time complexity of Bayes Network estimation, iterative genome mapping for accurate expression profile of each isoform and perfect hamming code.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	6,600,000	1,980,000	8,580,000
2011 年度	3,300,000	990,000	4,290,000
2012 年度	3,000,000	900,000	3,900,000
年度			
年度			
総　計	12,900,000	3,870,000	16,770,000

研究分野：　生体生命情報学
科研費の分科・細目：情報学･生体生命情報学
キーワード：生体生命情報学、ゲノム、生物情報学

１．研究開始当初の背景

研究開始当初，塩基配列シーケンサーの解析能力が指数関数的に向上しており，生物学・医学の各分野の研究に大きなインパクトを与える事が期待されている．具体的には，ゲノム情報に基づいた未知微生物遺伝資源ライブラリーの構築（メタゲノムプロジェクト）や，英米中が実施している１０００人のゲノムを解析するプロジェクト等が挙げられる．
シーケンサーを用いて対象となる生物の多様性を調べる研究では，シーケンサーから得られる短い塩基配列（タグ又はリードと呼ばれる）の対象生物ゲノムへのマッピングが最初に実施される．しかしながら，ゲノムへの

マッピングが成功するタグの割合（マッピング率）は６５％～８５％程度しかなく，マッピングされなかった１５％～３５％のタグは解析対象から除去されてしまう．この低いマッピング率は，シークエンス解析に係るデータの網羅性を棄損するだけでなく，シーケンサーや実験そのものへの信頼性も問題視される事となるため，マッピング効率の向上は重要な課題となっている．

２．研究の目的

研究代表者は，FANTOM(Functional Annotation of Mammal)プロジェクトにおいて，ヒト及びマウスの発現タグの解析を行ってきた．その経験を踏まえ，低いゲノムマッピング率は，培地として用いている血漿に含まれるＤＮＡ等のコンタミ（混入）だけでなく，マッピング対象をゲノム配列に限定しているゲノムマッピング法が問題であると考えている．すなわち，従来のゲノムマッピング法では染色体を複数組持つ生物対しても，タグと染色体一組分の情報である既知のゲノム配列との相同性に依拠し，対立染色体，対立遺伝子の存在を考慮してこなかった事が，低いマッピング率の一因だと考えている．そこで本研究では，次世代シーケンサーから得られたタグを基に対立染色体及び対立遺伝子の配列を決定する手法の研究を行う．タグのマッピングを行う際，ゲノムの配列情報に加え，本研究で得られる対立染色体，対立遺伝子の配列情報を用いる事によりマッピング率を向上させる事が可能であると考えている．

また，対立染色体，対立遺伝子の配列を決定する事により，ハプロタイプ解析能力が向上する．現在のハプロタイプ解析では，位置塩基多型（SNP）の組合せに終始しており，２本ある染色体上での組合せを考慮していない．具体的には，個体における遺伝子の２か所に SNP が存在したとし，それぞれ Aa,Bb で表わすとする．これが２本ある染色体において，(AB,ab)という組合せ存在するのか，(Ab,aB)という組合せで存在するかを判別することは現在のハプロタイプ解析では不明である（前頁図）．本研究の成果で得られる対立染色体，遺伝子の配列によって，SNP の染色体上での組合せが判明するため，ハプロタイプ解析能力が向上すると考えている．

３．研究の方法

研究の期間中、DNA シーケンサーから得られる短い塩基配列の集合から対立染色体、対立遺伝子の配列を決定手法の確立を目指す。また、マッピングされた結果に基づき、対立染色体、対立遺伝子を考慮することにより各種解析手法の精度･能力が向上する事を検証する。対立染色体、対立遺伝子の配列を同定する方法として、DNA の部分配列を頂点とする非循環有向グラフを構成するモデルを提案した。頂点の重みはその DNA 配列がショートリード中に存在した数を表す。辺の向きは、DNA 配列の 5'->3'の向きを表現し、辺の有無は２頂点の DNA 配列がショートリード中に存在した事を表し、存在した数を辺重みとする。このグラフから対立染色体･対立遺伝子に対応するパスを２つ見つける事で配列の同定を行う。また、同定を行った後は、マッピングにより対立遺伝子の発現量測定を行い、本研究後半の発現解析手法の入力と成す。発現解析手法として、細胞分化時の遺伝子発現系列解析、遺伝子制御ネットワーク推定を対象としたベイズネットワーク推定法、mRNA から得られたショートリードを用いたアイソフォームを対象とする。またこれら解析は長い計算時間を要するため、並列化、分散化による実計算時間の短縮化を図る。

４．研究成果

対立遺伝子･染色体の配列決定手法として、下図に示す、有向非循環グラフ表現を用いたアルゴリズムが有効である事を明らかにした。

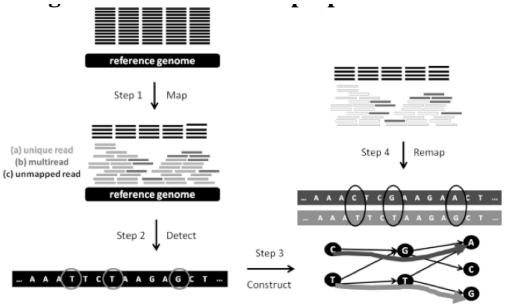

図 1　対立遺伝子･染色体配列の決定手法

次に、DNA 配列を表現する方法として、ACGT を元とするガロア拡大体 GF(4)上の完全線形符号の受信語と見なし、誤り訂正を行った後の語を用いる方法を提案した。本手法は既存マッピング手法を代替するための新しいアルゴリズムの核となりうると考えている。

また、発現解析手法として以下の成果を得た。細胞分化時の遺伝子発現系列解析手法では、細胞分化系列ごとに遺伝子を頂点とするグラフを作成し、遺伝子を介した細胞分化系列間の遷移をスコア化する事により、動的計画法の拡張によってクロストーク遺伝子が高精度で同定可能な事を示した。本アルゴリズムの概要を次ページ図２に示す。

ベイズネットワーク推定手法としては、解空間の大きさが頂点数Ｎに対して３のN^2乗のオーダーであるのに対し、探索空間 0(N^3 で)効

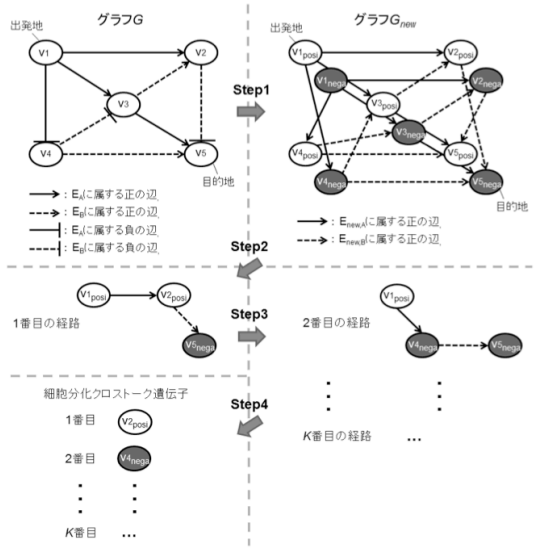

図 2　細胞分化系列における発現解析手法

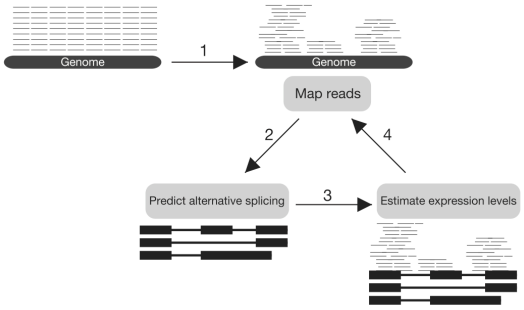

図 3　アイソフォームと発現量の同時推定法

率よく探索可能である事を明らかにした。本研究の内容は、国際会議 Asia Pacific Bioinformatics Conference において最優秀論文賞を受賞した。

次にアイソフォームの同定法として、ゲノムマッピングの修正暫時行いつつ再マッピングする手法（図３）が、アイソフォーム配列の同定だけでなく、アイソフォームの発現量推定にも有効である事を明らかにした。

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　６件）

① Ryo Araki, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: An estimation method for a cellular-state-specific gene regulatory network along tree-structured gene expression profiles, Gene, Volume 518, Isusue 1, Pages 17-25., (May 2013). 査読有

② Tomoshige Ohno, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: A method for isoform prediction from RNA-Seq data by iterative mapping, IPSJ Transactions on Bioinformatics, Vol.5, pp.27-33 (April 2012). 査読有

③ Yukito Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda：An Estimation Method for Inference of Gene Regulatory Network Using Bayesian Network with Uniting of Partial Problems , BMC Genetics Vol. 13(suppl 1), (Jan. 2012) 査読有

④ Tomoyoshi Nakayama, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda : Inference of S-system Models of Gene Regulatory Networks using Immune Algorithm, Journal of Bioinformatics and Computational Biology, Vol. 9(suppl 01) (Dec. 2011) 査読有

⑤ Yoichi Takenaka, Shigeto Seno and Hideo Matsuda :Perfect Hamming code with a hash table for faster genome mapping, BMC Genomics Vol. 12(suppl 3):S3 doi:10.1186/1471-2164-12-S3-S8 (30 Nov. 2011). 査読有

⑥ 吉澤陽志、瀬尾茂人、竹中要一、松田秀雄, 細胞分化クロストークのモデル化と細胞分化クロストーク遺伝子の推定手法, 情報処理学会論文誌　数理モデル化と応用(TOM), Vol. 4, No. 4, pp. 59-68 (Nov., 2011). 査読有

〔学会発表〕（計　２３件）

① 奥田華代, 竹中要一, 大野朋重, 瀬尾茂人, 松田秀雄: Improvement of the Accuracy of Mapping by Composing Alleles, 情報処理学会 第 75 回全国大会, IB-6, 東北大学川内キャンパス 宮城(2013 年 3 月 6 日)　査読無

② 渡邉之人, 瀬尾茂人, 竹中要一, 松田秀雄 :　複数時系列遺伝子発現プロファイルを利用した遺伝子制御ネットワーク推定の精度向上手法 ,　情報処理学会研究報告　第 92 回数理モデル化と問題解決 2013-MPS-92, No. 12, 佐賀県武雄市　武雄市文化会館 (2013 年 02 月 27 日)　査読無

③ Ryo Araki, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: An estimation method for a cellular-state-specific gene regulatory network along tree-structured gene expression profiles, Proceedings of the 2012 International Conference on

Genome Informatics (GIW2012), Tainan, TAIWAN, (Dec. 13th, 2012) 査読有

④ Yukito Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: Bayes-based inference of gene regulatory network for multiple time series gene expression profile, Joint Conference on Informatics is Biology, Medicine and Pharmacology (日本バイオインフォマティクス学会　２０１２年年会　生命医薬情報学連合大会), Tower hall Funabori, Tokyo (Oct. 29th-31st 2012) 査読有

⑤ Tomoyoshi Nakayama, Yoshiyuki Kido, Hiromi Daiyasu, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: Estimation of Dynamic Gene Regulatory Networks for Cell Differentiation by Splitting Time Course Data, Joint Conference on Informatics is Biology, Medicine and Pharmacology (日本バイオインフォマティクス学会　２０１２年年会　生命医薬情報学連合大会), Tower hall Funabori, Tokyo (Oct. 29th-31st 2012) 査読有

⑥ Masakazu Sugiyama, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: Comparison of Gene Expressions measured by RNA-seq and Microarray for Transcriptome Analysis of Adipose Tissues, Joint Conference on Informatics is Biology, Medicine and Pharmacology (日本バイオインフォマティクス学会　２０１２年年会　生命医薬情報学連合大会), Tower hall Funabori, Tokyo (Oct. 29th-31st 2012) 査読有

⑦ Tomoshige Ohno, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: Transcript-Type Dependent Normalization of Expression Levels in RNA-Seq Data for Non-Coding RNA Analysis, Joint Conference on Informatics is Biology, Medicine and Pharmacology (日本バイオインフォマティクス学会　２０１２年年会　生命医薬情報学連合大会), Tower hall Funabori, Tokyo (Oct. 29th-31st 2012) 査読有

⑧ Yoichi Takenaka, Shigeto Seno, Hideo Matsuda: All the 1+3n one-mismatch sequences of n-mer DNA are involved in 22.2+0.00879n strings of Perfect Linear Code words on DNA, International Conference on Intelligent Systems for Molecular Biology (ISMB2012), U44, Longbeach CA USA (Jul 17th, 2012) 査読無

⑨ Yoichi Takenaka, Shigeto Seno, Hideo Matsuda: Perfect linear code reduces the solution space of genome mapping from 1+3n to 22.2 + 0.00879n to find one-mismatch for n-mer short reads, Special Interest Group Meetings of High-Throughput Sequencing at International Conference on Intelligent Systems for Molecular Biology (ISMB2012), Longbeach CA USA (Jul 13rd-14th, 2012)査読無

⑩ Yukito Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda : An Estimation Method for Inference of Gene Regulatory Network Using Bayesian Network with Uniting of Partial Problems , The tenth Asia Pacific Bioinformatics Conference (APBC2012), A3.1,Melbourne, Australia (Jan. 17th, 2012) 査読有

⑪ Tomoyoshi Nakayama, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda : Inference of S-system Models of Gene Regulatory Networks using Immune Algorithm, Proceedings of the 2011 International Conference on Genome Informatics (GIW2011), Busan, KOREA, (Dec. 6th, 2011). 査読有

⑫ Yoichi Takenaka, Shigeto Seno and Hideo Matsuda :Perfect Hamming code with a hash table for faster genome mapping, Asia Pacific Bioinformatics Network's 10th InCoB－1st ISCB Asia Joint Conference 2011 (InCoB/ISCB Asia 2011), Kuala Lumpur MALAYSIA, 3.2, (Nov. 30th 2011). 査読有

⑬ Yoichi Takenaka, Tomoshige Ohno, Kayo Okuda, Shigeto Seno, Hideo Matsuda: Sequence determination and expression estimation of alleles from RNA-Seq data ,InCoB/ISCB Asia 2011, Kuala Lumpur MALAYSIA, No. 138, (Nov. 30rd, 2011)査読無.

⑭ Daisuke Ueta, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: Mutation-aware clustering with error correction for short-read genome mapping, The 2011 Jointa Conference of CBI&JSBi, Kobe, JSBi-27, (Nov. 8, 2011)[oral & poster presentation] 査読有

⑮ Yukito Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: A Method for

Inference of Gene Regulatory Network based on Bayesian Network with Uniting of Partial Problems , ISMB/ECCB 2011, Vienna, Austria, N13, (July 17,2011)査読無.

⑯ Tomoshige Ohno, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: A Directed Graphical Gaussian Model for Inferring Gene Regulatory Networks, ISMB/ECCB 2011, Vienna, Austria, X16, (July 18,2011)査読無

⑰ Kiyoshi Yoshizawa, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: Revealing regulatory relationships of crosstalk with multiple time-series gene expression profiles, Proceedings of the 2010 Annual Conference of the Japanese Society for Bioinformatics (JSBi2010), P017, Fukuoka (December 13, 2010)査読無.

⑱ Daisuke Ueta, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: A method for detecting structural variants from massive paired end genome sequences by mapping signatures, Proceedings of the 18th Annual International Conference on Intelligent Sysmtes for Molecular Biology (ISMB2010), U041, Boston (July 11, 2010) 査読無.

⑲ Tomoshige Ohno, Motokazu Ishikawa, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: An Improved RNA-Seq Analysis Method for Isoform Prediction by Iterative Mapping, Proceedings of the 2010 International Conference on Genome Informatics (GIW2010), No.8, Hangzhou (December 16, 2010)(Best Poster Award). 査読有

６．研究組織

(1)研究代表者

竹中 要一 (TAKENAKA YOICHI)

大阪大学･大学院情報科学研究科･准教授

研究者番号：00324830

(2)研究分担者

なし　　(　　)

研究者番号：

(3)連携研究者

なし　　(　　)

研究者番号：