

プロジェクト名： 高速相同性発見手法を用いたゲノム解析とその応用

1. プロジェクト体制

研究代表者

[国立情報学研究所] 宇野 毅明

共同研究者

[国立遺伝学研究所] 小出 剛

(研究協力者)

[国立遺伝学研究所] 梅森 十三

2. 研究目標

- 1) 高速相同性検索を用いたDNA配列のアセンブルシステムの確立とアライメントソフトの開発

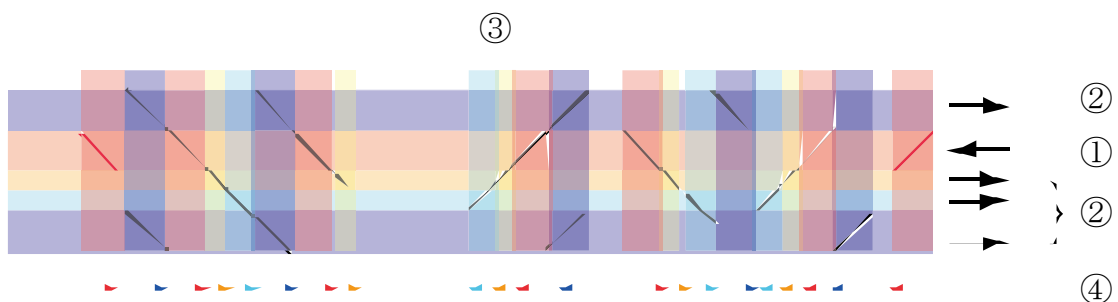
「相同領域発見手法」により検出された相同配列の結果から直接アセンブリを行う方法の確立、さらには自動化プログラムの作成を目指す。これにより、作業を効率化できるだけでなく、高速でかつ繰り返し領域に強い、新たなアセンブリプログラムの開発が期待できる。

- 2) 難解読領域の機能の探索

高速相同性検索アルゴリズムを用いて、これまでに解析が困難であった領域、つまり特定塩基の偏りや、繰り返し配列が高頻度に存在する領域のゲノム構造を解析する。

3. 平成 20 年度の研究進捗

- 1) 本年度は、富士通 SSL に開発を依頼して高速相同検索アルゴリズムの Web サービスを始めた。以下で述べる、大域的な構造を得ることを中心とした、マニュアル操作による解析を行いやすくするためのプロトタイプが完成した。今後はこのようなマニュアル型の解析への効率化を目指すために試用を行っていく。また、局所的、あるいは大量の小規模構造を機械的に処理するという目標に関しては、高速相同検索アルゴリズムを利用した、繰り返し配列に強いアセンブリソフトの開発を行っている。
- 2) 多くの繰り返し領域で構成される第 13 番染色体の *Genic1* 領域(1.6Mbp)について「相同性



- ① コアエレメントを決め、同じ色の帯を付ける
- ② コアエレメントの周りの領域で、ある一定以上の長さを持った共通配列を見つけ、同様に色を付ける
- ③ 縦の列に同じ配列がある場合は同じエレメントとみなし、横の列と同様に同じ色をつける
- ④ 各色につき横列一つを選び、そこからエレメントを抽出する

(図 1) 共通エレメントの決定法

高速検索アルゴリズム」を応用し、繰り返し領域の構造を1次元的に表現する方法を開発中である。目標としては、a)繰り返しを構成するエレメントを正確に知ること、b)そのエレメントが何回繰り返されているかを調べること、c)これらの情報の全体構造を視覚的により見やすくすることが目的である。このモデルとして、「相同性高速検索アルゴリズム」を用いて2次元的に描かれた繰り返し構造の二次元図を用いて、図1に示した方法で、マニュアル（手作業）でエレメントの抽出を行った。その結果を図2に示した。

この結果、マニュアルで行ったため、やや正確性に欠けるが、*Genic1* の繰り返し構造の概要を1次元的に図示することに成功した。また、大きく分けて7つのエレメントを抽出し



(図2) 第13番染色体 **Genic1** 領域の繰り返し構造

た。現在は、これらのエレメントとその塩基配列を自動抽出する方法を検討している。

4. 次年度以降の研究展開

上に示した通り、「高速相同領域発見手法」で描いた2次元の図を元に、繰り返し配列構造をマニュアルで検出し、解析することができた。現在は正確な繰り返しエレメントの自動抽

出を目指しているが、非常に困難を極めている。これは、目で見ただけでは繰り返し配列の範囲をおおよそで掴むことができるが、自動化を試みた場合に、どの領域までが繰り返しなのかを判断するのが難しいためである。従って、この作業を完全に自動化することは難しく、マニュアルと組み合わせながら、正確な繰り返し配列の検出を行う必要がある。

このような複雑な構造を取る領域は、セントロメアやヘテロクロマチンのように、そのゲノム構造が生物学的に重要な機能を持っていると考えられるため、解析が可能となる意義は大きい。従って、*Genic1*領域以外に存在する複雑な領域を検索し、再び高速相同性検索アルゴリズムを用いて、反復配列や逆位配列などのゲノム上の特殊な構造について解析する。また、*Genic1*を含めたこれらの特殊な領域を、ラットやヒトと比較することにより、ゲノム構造の進化や機能について、新たな知見を生む可能性も期待できる。

5. 研究経費

平成18年度実績： 千円

平成19年度実績： 5,000 千円

平成20年度実績： 8,000 千円

6. 平成〇〇年度の研究成果

(1) 主要成果物（知見、成果物、知的財産権等）

(2) 成果発表

<論文発表>

[学術論文]

- 1) Takeaki Uno, "An Efficient Algorithm for Finding Similar Short Substrings from Large Scale String Data", Lecture Notes in Artificial Intelligence 5012, pp. 345-356, 2008. 2008年05月某日
- 2) Takeaki Uno and Hiroki Arimura, "Ambiguous Frequent Itemset Mining and Polynomial Delay Enumeration", Lecture Notes in Artificial Intelligence 5012, pp. 357-368, 2008. 2008年05月某日
- 3) Takeaki Uno, "An Efficient Algorithm for Solving Pseudo Clique Enumeration Problem", Algorithmica, Springer New York, 2008. 2008年10月某日
- 4) Takeaki Uno, "Genome Homology Visualization by Short Similar Substring Enumeration", RIMS Workshop on Acceleration and Visualization of Computation for Enumeration Problems, invited talk, 2008. 2008年09月某日
- 5) Hiroki Arimura, and Takeaki Uno, Mining Maximal Flexible Patterns in a Sequence, Lecture Notes in Computer Science, Springer, 4914/2008, pp. 307-317. 2008年01月某日
- 6) Shin-ichi Minato, Takeaki Uno, and Hiroki Arimura, "LCM over ZBDDs: Fast Generation of Very Large-Scale Frequent Itemsets Using a Compact Graph-Based Representation", Lecture Notes in Artificial Intelligence 5012, pp. 234-246, 2008. 2008年05月某日
- 7) Liu Y.-H., Takahashi A., Kitano T., Koide T., Shiroishi T., Moriwaki K., Saitou N.: Mosaic genealogy of the *Mus musculus* genome revealed by 21 nuclear genes from its three subspecies. *Genes and*

Genetic Systems, 83(1):77-88, 2008.

8) Takahashi, A., Shiroishi, T., Koide, T.: Multigenic factors associated with a hydrocephalus-like phenotype found in inter-subspecific consomic mouse strains. *Mammalian Genome*, 29:333-338, 2008.

9) Shigeta Y., Kasai S., Han W., Hata H., Nishi A., Takamatsu Y., Hagino Y., Yamamoto H., Koide T., Shiroishi T., Kasai K., Tsunashima K., Kato N., Ikeda K.: Association of morphine-induced antinociception with variations in the 5' flanking and 3' untranslated regions of the μ opioid receptor gene in 10 inbred mouse strains. *Pharmacogenet. Genomics*, 18:927-936, 2008.

10) Takahashi, A., Nishi, A., Ishii, A., Shiroishi, T., Koide, T.: Systematic analysis of emotionality in consomic mouse strains established from C57BL/6J and wild-derived MSM/Ms. *Genes, Brain and Behavior*. 7:849-858, 2008.

[会議録]

[解説・総説]

[研究ノート]

[その他]

<会議発表等>

[招待講演]

[一般講演]

<著書等>

(3) その他の成果発表

・梅森十三、「遺伝的不適合による神経発達異常」、第 31 回日本神経科学大会 (Neuroscience 2008)、口頭発表、平成 20 年 7 月 9 日

・梅森十三、「遺伝的不適合による神経発達異常」、若手クロストーク、平成 20 年 11 月 17 日—18 日