

#### A. 研究概要

機械学習やパターン認識の技術を基礎にして、Computational Biology の諸問題に取り組んでいる。主な研究内容は次になる。

##### (i) タンパク質間相互作用に関する機能部位の研究

DNA 配列は暗号化された生命の設計図であり、この設計図に基づき作られるタンパク質はそれぞれに特命を担った分子である。このような配列データを解析しその機能を理解することは、生命のパーツを理解する上で不可欠なことである。タンパク質の多くは、他のタンパク質と相互に結合することにより、その機能を発揮する。従って、これらの相互作用の機構等を理解することは、薬の設計等において重要な示唆を与えることになる。そこで行った研究の一つが、タンパク質  $A$  はタンパク質  $B_1, \dots, B_n$  は相互作用するという入力情報から、 $B_1, \dots, B_n$  のアミノ酸配列が共通に含む特徴的な配列を抽出することにより、 $A$  との相互作用に何らかの形で深く関与する（その多くは直接  $A$  とコンタクトする）部分配列を予測する手法の開発である。具体的には、ベイト (bait) とよばれるタンパク質  $P$  に共通に結合するプレイ (prey) とよばれる複数のたんぱく質の配列間で特徴的な部分配列のパターンを抽出する。パターンの選択方法は、その事後確率の最大化である。そして、得られたパターンをヒトの全タンパク質配列上で検索し、ヒットしたタンパク質の機能を Gene Ontology とよばれる知識データベースで評価し統計的有意性を算出した。Gene Ontology とは、遺伝子産物（転写産物とタンパク質産物）に機能に関する注釈をおこなうための用語の集合とその用語間の関係を定義したデータベースであり、生命機構に関する最も網羅的なデータベースの一つである。その結果、いくつかのタンパク質に対して信頼度の高い相互作用部位の特定に成功すると同時に、多くの信頼度の高いタンパク質間相互作用の候補を見つけることにも成功している。入力として用いるヒトのタンパク質間相互作用データは、共同研究者が独自の実験により作成したものであり、今後このデータからのさらなる発見が期待できる。

(ii) 次世代シーケンサーを用いた配列解析次世代シーケンサーが生成した大量の配列断片から、X 染色体の DNA メチレーションの状態を同定する計算手法を開発した。用いたデータは、男性と女性の白血球の一種である好中球の X 染色体である。その結果、好中球における遺伝子量補償（性染色体上にコードされている遺伝子の発現量が雄と雌の間で同じになるように調節されていること）の微調整には、DNA メチレーションに加えて、何か他の因子が関与していることが分かった。

##### (iii) 選択的スプライシング制御配列探索の研究

選択的スプライシングとは、一つの遺伝子から複数の相異なるタンパク質を生成する機構である。本研究では、相同配列から保存領域を発見する技法である phylogenetic footprinting を基礎に、相同配列が選択的スプライシングを行っているか否かの情報を利用してモチーフを探索するアルゴリズムを開発し、実データに適用した。その結果、選択的スプライシングを行う遺伝子配列に特異的に存在するパターンが偏在していることが明らかとなった。

##### (iv) ランダムサンプリングを用いた系統樹再構築アルゴリズムの開発

本研究では、対象とする各生物種の全タンパク質配列上でのオリゴペプチドの頻度を要素とするベクターから系統樹を再構築する手法の開発とその評価を行って

いる。近年、 $Q_i$ らにより、固定長の全オリゴペプチドの頻度からなるベクターから統樹を再構築する手法が提案され、その手法を 109 種類の微生物集合に適用し、基本的な分類群のほとんどが Bergey's Manual と大体一致するという報告がなされている。しかしながら、その方法では、オリゴペプチドの固定長を  $K$  とするとき、 $20$  の  $K$  乗次元 ( $K = 5$  の場合、 $3,200,000$ ) のベクトルを各生物種に対して生成する必要がある、大変膨大な計算資源を必要とする。実際、 $Q_i$ らは計算機実験を行うに当たりスーパーコンピュータを使用している。そこで、本研究では、固定長のオリゴペプチドの頻度ベクターからの系統樹再構築をパソコンでも行えるように、少量のオリゴペプチドをランダム・サンプリングする手法を提案している。アルゴリズムは次のようになる。少量のオリゴペプチドをランダム・サンプリングし、これらの頻度ベクトルを各生物種ごとに計算し、そのベクトルをもとに系統樹を構築する。これを独立に  $R$  回繰り返し、最後に  $R$  個の系統樹のコンセンサス木を 1 つ構築する。これを最終的な出力とする。計算機実験を行った結果、 $20$  の  $K$  乗の  $10\%$  以下のオリゴペプチドで  $Q_i$ らと同程度の質の系統樹がパソコン上での計算で得られた。従って、多くとも  $10\%$  程度の相異なるオリゴペプチドを用いれば、その選択方法によらず、進化系統樹に関する情報が得られることが分かった。

「マス・フォア・インダストリ」にかかわる H20, 21 年度の研究実績概要

研究概要 (i) で述べた「タンパク質間相互作用に関する機能部位の研究」において、様々なドメイン (近種間で保存されているタンパク質配列の部分領域であり何らかの機能を有していると考えられている) に対して、統計的に有意なそれらの部分領域を特定することに成功している。このような情報は未だデータベース化などされておらず、プロテインエンジニアリングや創薬の分野に貢献することが期待できる。

研究業績

1. Osamu Maruyama, Hideki Hirakawa, Takao Iwayanagi, Yoshiko Ishida, Shizu Takeda, Jun Otomo, Satoru Kuhara, Evaluating Protein Sequence Signatures Inferred from Protein-Protein Interaction Data by Gene Ontology Annotations, IEEE International Conference on Bioinformatics and Biomedicine, 417-420, 2008.

2. Yukio Yasukochi, Osamu Maruyama, Milind C. Mahajan, Carolyn Padden, Ghia M. Euskirchen, Vincent Schulz, Hideki Hirakawa, Satoru Kuhara, Xing-Hua Pan, Peter E. Newburger, Michael Snyder, and Sherman M. Weissman, X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils, Proceedings of the National Academy of Sciences of the United States of America (PNAS) 107, 3704-3709, 2010.

講演

1. Evaluating Protein Sequence Signatures Inferred from Protein-Protein Interaction Data by Gene Ontology Annotations, IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, USA, November, 2008.

2. Finding Protein Sequence Signatures from Protein-Protein Interaction Data Using Gene Ontology Annotations, 17th annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 8th European Conference on Computational Biology (ECCB), Stockholm, Sweden, June 27-July 2, 2009.

学位

博士・数理学（九州大学）

受賞歴

平成6年3月29日 電気通信普及財団第9回テレコムシステム技術学生賞

その他特記事項

- ・ IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2008, 2009.
- ・ International Workshop on Bioinformatics Research and Applications (IW-BRA) 2008, 2009, 2010.