マイクロアレイデータ解析コンテスト（CAMDA 2007）への参戦記

岡田　吉史

元 細胞機能設計チーム 産総研特別研究員（3/31 退職）

現 室蘭工業大学 情報工学科 助教

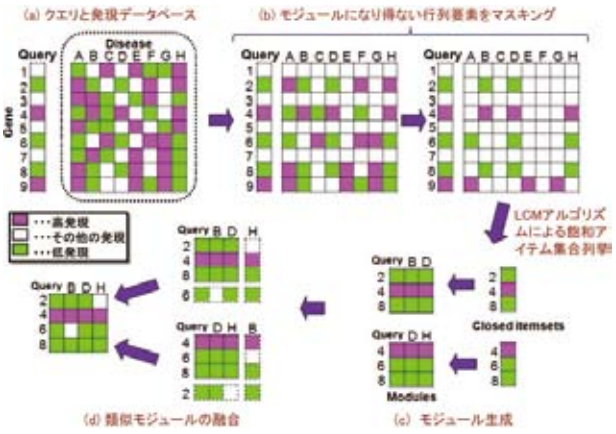
CAMDA（Critical Assessment of Microarray Data Analysis）[1]は、2000年から毎年開催されているマイクロアレイデータ解析コンテストで、１）参加チームによる課題データの解析と論文投稿、２）CAMDA審査員による評価、という２段階のプロセスを経て、特に優秀と判定されたチームには口頭発表の権利が与えられます。タンパク質立体構造予測コンテストであるCASPのように実験的に検証された解答がある訳ではなく、データ解析方法と得られた生物学的知見の両側面からの評価を受けます。

昨年１２月、私は細胞機能設計チームの藤渕チーム長と共に、スペインで行われたCAMDA2007に参加いたしました。我々のチームは、以前から開発を進めてきた遺伝子モジュール探索システムSAMURAI[2]を用いて、課題データの１つである「大規模アレイデータ」の疾病細胞発現データに焦点を当てて解析することにしました。我々は、まず、それらのデータを用いて、22,283遺伝子×2,899アレイからなるデータベースを構築しました。我々の目的は、そこから特定の疾病細胞で特異的に共発現する遺伝子の最小単位である「遺伝子モジュール」を網羅的に探索し、様々な疾病間の生物学的な関連を探ることでした。しかしながら、モジュール探索は遺伝子とアレイ（疾病種）の組み合わせ探索であるため、以前の探索法では現実的な時間内に計算を終了するには、せいぜい数千遺伝子×数百アレイのデータを処理するのが限界でした。そこで、我々は、SAMURAIをより大規模なデータベースに適用するため、問い合わせ遺伝子発現データ（クエリ）をもとに、モジュールとなり得ない遺伝子を除外することで探索空間を大幅に縮小する方法を考案しました[3]（図）。作成したデータベースからクエリとして１つずつアレイデータを抜き出してモジュール探索を行ったところ、ほとんどのクエリにおいて、わずか数秒～数分で遺伝子モジュールを網羅的に発見できることがわかりました。さらに、得られた全てのモジュールに対する機能解析の結果、ダウン症とハンチントン舞踏病で抽出されたモジュールが、ともに「細胞接着」や「細胞表面認識」に関わる機能で統計的有意に特徴づけられることを見出しました。興味深いことに、これらの機能を持つ分子は、ショウジョウバエやマウスにおいて神経回路の形成と障害に重要な役割を持つことが実験的に示されており、ヒトにおいても、ダウン症の知的障害やハンチントン舞踏病の痴呆症にも関連することが考えられます。以上の成果は、CAMDA審査員から高く評価され、口頭発表に選抜される結果となりました。CAMDAへの参加は、SAMURAIの実力を世界に知らしめる良い機会となったと思っています。

今後は、SAMURAIに、モジュールの発現パタンの時間的変化を検出する仕組みを導入し、例えば発生や分化、あるいは疾病の進行段階といった動的プロセスを発見するシステムへと進化させ、次期CAMDA（イタリア開催）への準備を進めていきたいと考えています。

Reference

1. http://camda.bioinfo.cipf.es/
2. Okada, Y., Fujibuchi, W. and Horton, P., "A biclustering method for gene expression module discovery using closed itemset enumeration algorithm", IPSJ Transactions on Bioinformatics, vol. 48, no. SIG5(TBIO2), pp.39-48, (2007).
3. Okada, Y. and Fujibuchi, W., "Mining a large-scale microarray database for similar gene expression modules to find distant relations between Down syndrome and Huntington disease", The 7th Int. Conf. for the Critical Assessment of Microarray Data Analysis (CAMDA 2007), Valencia, Spain, (2007).



図：SAMURAIの処理手順。
(a) クエリとデータベースの発現値は事前に離散化される。
(b) クエリで高発現または低発現の遺伝子のみを含むモジュールを探すため、モジュールを構成し得ない遺伝子を探索対象から除外する。
(c) モジュールを構成する遺伝子セットを高速に列挙するため、LCMと呼ばれる飽和集合列挙アルゴリズムを利用する。
(d) 類似のモジュールを融合する。