論文題目　Estimation of Gene Network Model using Real-coded Genetic Algorithm
(実数値遺伝的アルゴリズムを用いた遺伝子ネットワークの推定)

氏　名　安藤　晋

近年の測定技術の進展は、遺伝子の発現過程を動的に、高い並列性を持って捉えることを可能にした。細胞内には代謝産物、たんぱく質、転写産物等による制御回路が存在する。その中で、DNA マイクロアレーによって観測することができる mRNA の時系列データは細胞内のメカニズムを理解するための鍵を提供しているといえるだろう。分子生物学ではさまざまな細胞内物質の関係をグラフ構造・ネットワークで表す手法が一般的であるが、発現データの増加とともに、遺伝子の関係に写像した遺伝子ネットワークの解析とモデル化が分子生物学における重要な課題となっている。

本論文はDNAマイクロアレーにより測定される遺伝子発現時系列データに基づいた遺伝子機能のネットワークモデル化を目標とする。現時点では既知でない遺伝子の相互作用をモデル化するため、制御等に用いられる一般性のある微分方程式系(S-system)を用いた。観測される遺伝子発現データをよく再現するようなモデルのパラメータを決定する問題はリバースエンジニアリングと呼ばれる。ここではパラメータの観測データとシミュレーション出力の差を赤池情報量基準で評価し、それ最小化する関数最適化問題と考える。これは多峰性・高次元な関数最適化問題であり、進化的計算アルゴリズムによって解くのに適した問題である。本論文では実数値遺伝的アルゴリズムを分布推定アルゴリズムの考え方を用いて拡張したアルゴリズムと Goldberg らによる Messy GA を拡張したアルゴリズムを組み合わせた解法を実装した。

しかしながら、現在の技術で得られる発現データは非常に短い(数十時点)時系列データであり、非常に大きな規模(数千の発現ノード)を持つモデルの決定には概ね不十分である。技術の進展には目覚しいものがあるが、本研究では以下の複合的な手法を用いて、まず 30 ノード(900 変数)程度の逆問題を解くことを目標とした。まずは逆問題をサブタスクに分割して一度に最適化するパラメータを減らす実装を行った。

さらに、生成されたモデルのうち、よりよく決定されたパラメータを抽出するためのロバストネス解析を利用した。これは GA の反復試行により生成される複数のモデルから各パラメータの統計値をとり、より安定して推定されるパラメータを選別する手法である。ただし、この手法では遺伝子ネットワークの局所的なモジュール性を仮定している。

もうひとつは、データベース等から得られる発現データ以外の知識を推定の過程に取り入れる手法である。ゲノムデータベースにはアノテーションやプロモータ、オペロンなどに関する知識がさまざまな確信度で蓄積されている。ここではデータベースの知識の信頼度を重みとして評価関数に反映させる。これは、探索のランドスケープ内に既存

知識に集約するようなアトラクタを作ることに相当する。

実験では、まず拡張したアルゴリズムを、高次元での関数最適化ベンチマーク問題での性能を確認した。本研究で利用したのは非直交分布の多峰性関数である Rastrigin 関数と Rosenbrock 関数である。目標とする問題のサブタスクと同じ 30 次元の関数を最適化した。遺伝的アルゴリズムを 100 回試行し、十分な性能が得られることを確認した。

続いて、遺伝子発現シミュレーションによって実数値遺伝的アルゴリズム、ロバストネス解析、知識導入な手法を検証した。ここでは 20 個の遺伝子からなる遺伝子ネットワークの S-system モデルを用意し、数値積分によって人工発現データを生成した。また、実際の観測条件を考慮し、正規分布の誤差を加えている。実験では発現データセット(20 ノード×100 時点の時系列データ)を異なるノイズパターンで 100 個用意した。時系列データのうち 20 点を赤池情報量基準でのフィッティングに用いた。

ノイズの分散を 0.6 とした実験では最適解として得られたモデルが正しいモデルと一致した。ノイズの分散を 1 とした場合には多数の局所解が得られた。これらの局所解のモデルからロバストネス解析による指標を算出すると高い感度での制御関係が推定できることが分かった。また、ターゲットモデルに関する知識を導入した実験では、20%程度の知識でネットワークの局所解の多くが削除され、推定の感度が向上することが示された。

シミュレーション実験をふまえ、大腸菌のパブリック発現データ(Stanford Microarray Database)の解析を行った。トリプトファンの過剰、不足条件下での大腸菌の約 200 遺伝子の発現を 5 つの時点で測定したものである。またこれらの遺伝子に関しては RegulonDB にてプロモータ、オペロンなどの情報が提供されている。実験では同オペロンに属する遺伝子に関する知識を評価関数に導入した。この中で、14 個の遺伝子トリプトファンに影響を受ける 14 遺伝子を選択し、それらを制御する遺伝子を全遺伝子の中から推定した。ロバストに推定されたパラメータのうち約 4 分の一を生物学的文献にて確認することができた。

今後は得られた結果から生物学的な仮説を抽出することを試みたい。また、課題としては間違った知識による影響の調査、ロバストネス解析の多峰性への対応、得られたモデルの安定点解析などが残っている。