論文の内容の要旨

論文題目　Finding Optimal Models For Gene Networks
(遺伝子ネットワークにおける最適なモデルの探索)
氏名　オット、ザーシャ

近年、人間などの高等生物の遺伝情報を完全に解読することが可能になってきている。遺伝情報の単位である遺伝子がDNAに暗号化されており、**転写**と**翻訳**によって発現する。転写によって遺伝子がRNAへと複製されてから、RNAを鋳型に蛋白質が翻訳によって作られる。RNAと蛋白質の分子が細胞の中の多様な機能を実現しており、生命の現象の基本となっている。全遺伝子の転写と翻訳を調整することによって、細胞の成長、細胞の分裂、環境条件の変化に対応することなどが可能になる。その調整を果たしているのもRNAと蛋白質であるので、遺伝子の発現と遺伝子の調整がサイクルを形成し、それぞれの遺伝子の発現レベルは相互に調整される。その様な発現レベルから見た依存関係の全貌をここでは**遺伝子ネットワーク**とよぶ。

DNAマイクロアレイ等の新しい技術の開発により、細胞の殆どの遺伝子の発現レベルを同時に測ることが可能になってきている。その様な一連の実験で得るデータから遺伝子ネットワークについての情報を抽出する方法は、ポストゲノム時代の最優先のニーズの一つである。遺伝子ネットワークの情報を入手できれば、細胞システムの理解が深まり、分子生物学、医学、製薬等への貢献をもたらすと期待される。そのため、遺伝子ネットワークの発現データからの推定が生物情報科学の主なテーマの一つとなっている。

しかし、遺伝子ネットワークの探索問題はNP困難な問題であり、探索空間の大きさは超指数関数的である。ブルートフォース的な方法の適用では、９個の遺伝子のネットワークの場合でも実現できない規模の計算量となる。これらの問題を避けて、ヒューリスティクなアルゴリズムを適用すると探索の結果の良さが不明になる。そのため、今までは確実に最適な遺伝子ネットワークのモデルを得ることは不可能だった。

遺伝子ネットワークが注目を集める問題であるのに、今までの研究ではネットワークの推定について、本質的な評価が行われなかった。推定の評価を行う際、問題になるのは、実際の遺伝子ネットワークの知識が断片的であることやネットワークのどの部分がどういう実験の条件で働くかという不確かさ等である。それゆえ、発現データから推定される遺伝子ネットワークのモデルが有意であることの証明は今までになかった。

本研究ではネットワークの探索空間の解析を行い、指数関数的な時間で超指数関数的な探索空間の中の最適なネットワークを見付け出すことのできるアルゴリズムを開発した。これにより、遺伝子数が２０個の場合でもこのアルゴリズムを現実的な時間で利用することが可能となった。さらに生物学的に妥当な制約条

件のもとでは、遺伝子数が３５個前後の時でも最適なモデルの探索が行える。また、実際の発現データでは有意に違わない発現のパターンを見せる遺伝子が多いということを考慮して、パターンが等しいといえる遺伝子をグループにまとめることができる。そういった本質的な部分を**ネットワーク要素**とよぶ。遺伝子ネットワークをネットワーク要素から構成して１００個前後の遺伝子を扱って、最適なモデルの探索が現実的に可能であることを明らかにする。

したがって、遺伝子数が限られた場合においては、発現データから遺伝子ネットワークについて情報を得ることができるというメリットをこのアルゴリズムはもっている。このアルゴリズムが発現データの種類にもこれまで提案されている統計学的な遺伝子ネットワークのモデリングの方法にも依らないので、一般的に利用することが可能である。ヒューリスティクなアルゴリズムの不確実な結果を避け、こうした統計学的なモデリングの方法のそれぞれに対して最適なモデルの探索を行うことにより、モデリングの方法の適切さを評価することができる。同様に、異なる実験の方法を適用して得たデータのそれぞれに対して最適なモデルの探索を行えば、遺伝子ネットワークを発見するための実験方法の良さも評価でき、データの良い組み合わせ方も解析できる。

ところが、最適なネットワークとは、モデリングに対応したスコア関数に関して最適なネットワークに過ぎない。ほぼ等しいスコアで構造の異なるネットワークが数多く存在しうる。その理由に依り、データにはそもそもネットワーク全体の情報は無くて、全ネットワークの部分的な情報しかないという可能性を十分に考慮するべきである。遺伝子数またはネットワーク要素数が１０個のときでもネットワークの候補の数が $4.17 \cdot 10^{18}$ の規模であるので、最適ネットワークといっても、生物学的に真の遺伝子ネットワークとは違うことが多々ある。

本研究ではその問題に取り組んで、二段階の方法を開発した。一つ目に、上述のアルゴリズムの理論を更に展開させて、最適なモデルから始めてネットワークをスコアの順で数え上げることができるようにした。数え上げの行えるネットワーク要素数は上に述べた規模とほぼ同じであるし、m が２万のときでも一番スコアの良い m 個のネットワークを現実的に数え上げられる。二つ目に、スコアの良いネットワークを比較して、共通な部分を取り出せる幾つかの方法を適用する。この様な共通部分を**ネットワークモチーフ**とよぶ。研究の成果として、実際のデータを使って推定したネットワークから抽出したネットワークモチーフの方が最適なネットワークより生物学的な知識と有意に一致していることがわかった。数え上げとネットワークモチーフ抽出からなる方法により、枯草菌と大腸菌を用いて、本研究で初めてネットワーク推定に基づいた遺伝子ネットワークの比較を行った（図１と図２を参照）。その解析の結果も紹介する。

さらに、上述の方法を任意に大きな遺伝子ネットワークの場合にも適用できるアプローチを提案する。ここで、広く認知されている遺伝子ネットワークの一般的な構造、すなわち密度の高い構成成分が存在し、その構成成分同士はまばらにつながっている構造を前提する。この前提のもとで生物学的に意味のある探索空間の部分を選択してから、部分探索空間に最適なモデルの探索アルゴリズム、数

え上げのアルゴリズム、モチーフ探索アルゴリズムが適用できることを示す。その部分探索空間を選択するために、生物学的な既知の知識を適用できる。既知の知識が十分でない場合、クラスタリングまたはヒューリスティクなアルゴリズムを利用できる。ネットワーク要素数が n のとき、部分探索空間を探索する時間は $O(n)$ になるので、このアプローチにより要素数の制約を解決できる。この方法を利用して生物学的な知識と有意に一致する推定ができることを示す。

本研究で幾つかの手法を使って、推定したネットワークのモデルを生物学的な知識と厳密に比較した。いずれの手法でも推定が有意に知識と一致することが明らかになったので、研究の成果として遺伝子ネットワーク推定の意義を示すことができた。

また、時系列データからの推定に微分方程式を使うという特殊ケースについて得た結果も含める。人工的な発現データを作って、データのエラーの大きさと測定の数とは、推定の良さにどの様に影響を与えるかについて調べた。さらに、微分方程式のスコア関数と BNRC というスコア関数をそれぞれ同じデータに対して利用した結果を議論する。また、微分方程式を使った遺伝子ネットワーク推定問題の特殊ケースの複雑さを決めるために、その問題がある定義のもとで NP-困難であることを証明する。

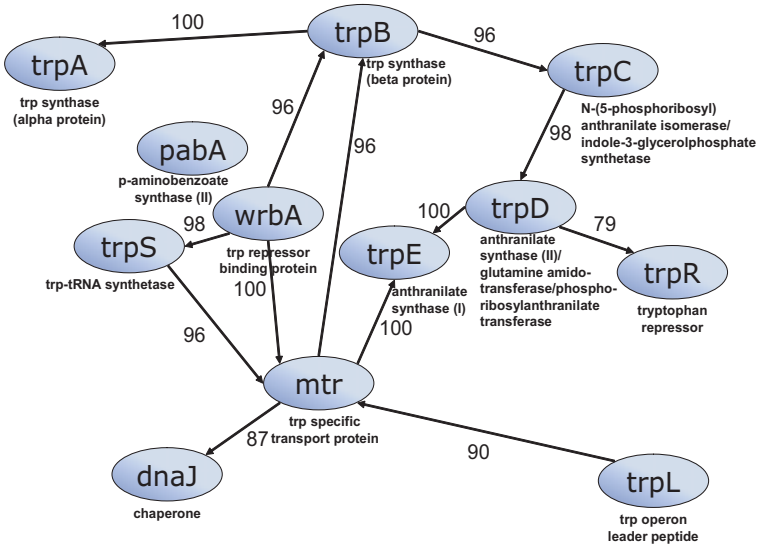

図 1: 大腸菌の遺伝子ネットワーク推定から抽出したネットワークモチーフ.

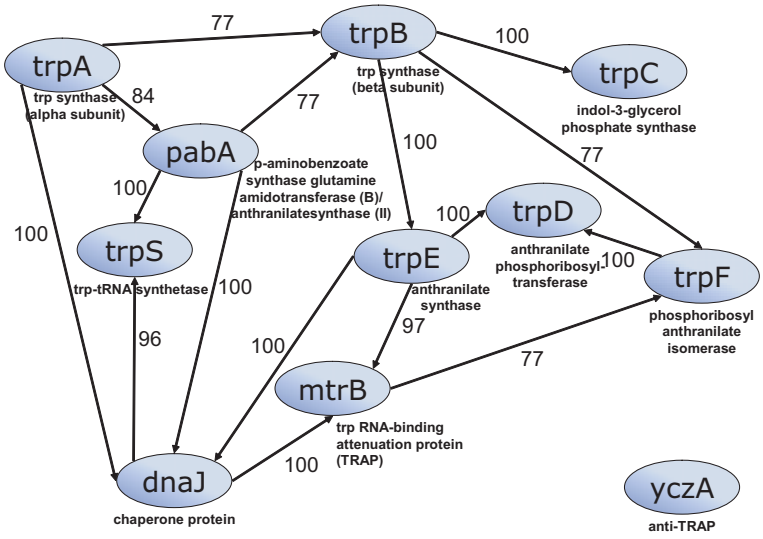

図 2: 枯草菌の遺伝子ネットワーク推定から抽出したネットワークモチーフ.