高次元データからの最大エントロピー法による極大連結コアグラフの抽出とその性能評価

水産総合研究センター 中道礼一郎, 東京大学 岸野洋久, 東京海洋大学 北田修一

1. はじめに

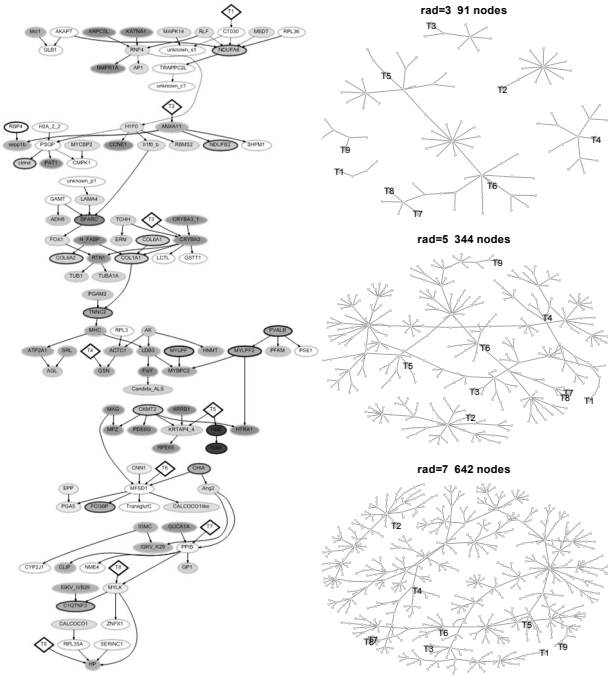
生物現象は多くの遺伝・環境因子が複雑に絡んだネットワークを形成し、小数のトリガー因子の変動が全体の形質に大きな変化をもたらすことから、現象を総合的に理解するために、トランスクリプトームやプロテオーム等の膨大な情報の中から目的にかなった遺伝子および、遺伝子間、遺伝子・環境間の相互作用を適切に抽出することが重要である。本研究では、遺伝子の発現と形質の関連を尤度ベースの有向グラフモデルでとらえ、コアとなる部分を最大エントロピー法で抽出する手法を提案し、大西洋サケ初期発生期の遺伝子発現カスケードの推定を題材に、モデル選択で全遺伝子の関連を推定するアップローチとの性能比較を行う。

2. 方法

本研究で提案する手法において、遺伝子発現カスケードのグラフ構造は AIC によって評価され、最大エントロピー法にもとづく遺伝的アルゴリズムによって、関心のある形質に直接相関を持つコアグラフと、そこに連結される周辺グラフの探索を行う。S/N 比向上のため、時系列による発現変動や先行研究データベース等から取得した外部情報により探索空間を適切に制限し、目的の形質に密接に関連する極大連結有向グラフとして抽出する。比較対象として、Chow & Liu アルゴリズム[1]により、モデル選択で全遺伝子の関連を推定し、形質と連結する遺伝子ノードを、形質ノードからの距離を段階的に増やして可視化する。

3. タイセイヨウサケ初期発生期の遺伝子発現ネットワーク

タイセイヨウサケを受精直後から孵化し稚魚が浮上するまでの 10 段階においてサンプリングし、全体組織において全ゲノムの遺伝子の発現量変動を計測した公開データ(GSE25938 [2])から、初期発生期の発現ネットワークを推定する。左図が最大エントロピー法による極大連結有向グラフで、右図が全遺伝子の関連を推定してから、形質ノードからの距離を 3・5・7 ステップとして可視化したものである。極大連結有向グラフでは、グラフ中の形質ノードの周辺のローカル構造から、形質に直接関与する遺伝子とその階層構造を容易に読み取ることができる。これにより膨大なデータを単純に集約するのではなく、情報量に見合って目的変数と関連する変数を最大限寄せ集めることにより、合理的な解釈を提供する枠組みを構築できる。



引用文献

[1] Chow & Liu (1968) Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, IT-14 (3): 462-467

[2] Jantzen et al. (2011) A 44K microarray dataset of the changing transcriptome in developing Atlantic salmon (*Salmo salar* L.). BMC Research Notes 4:88