2001 年 ベイジアンネットチュートリアル
Tutorial on Bayesian Networks (BN2001)
Tokyo, Japan, July 29, 2001.

遺伝子ネットワークと確率モデル
Genetic Networks and Probabilistic Models

有田 正規*
さきがけ 21
産業技術総合研究所 生命情報科学研究センター

Abstract: 大量データをもとにした生命メカニズムの解明はゲノム情報のテーマであり、ベイジアンネットを利用した研究は増えてゆくと思われる。本チュートリアルではベイジアンネットの入力データをつくる DNA マイクロアレイ技術とデータの前処理を解説する。またデータから遺伝子ネットワークを学習する際のアルゴリズムについて概観する。

1 はじめに

分子生物学は生命をオートマトンとして捉える学問である。教科書を開くと、細胞の核内で遺伝子 (DNA) から mRNA へ情報を転写、mRNA が核外へ移動、mRNA を編集、タンパク質への翻訳、タンパク質が相互に作用し遺伝子の発現を調節、と、一連の過程があたかも剛体運動のように記述されている [5]。しかし各プロセスが「どのように」動くかは記述されていても、「どうして」そう動くかがあまり書かれていない。いわば、オートマトンの具体的な動作例のみがわかっていて、仕様が未知のままになっている。今後「どうして」の部分が焦点となるのは必至であろう。

ゲノム情報は、分子生物学のデータを情報科学的視点から処理することで「どうして」の問いに答えようとしてきた。大量データからの学習が成功した例にはニューラルネットワークによるタンパク質二次構造予測 [20] と遺伝子シグナル予測 [18]、隠れマルコフモデルによる遺伝子コード領域予測 [9] を挙げることができる。また近年は、サポートベクターマシンによる分類が注目を浴びている [16]。ただ、これらのアプローチはある種のクラスタリングを行うもので、細胞内のメカニズムを出力するものではない。つまり「どうして」に対する回答は得られていなかった。

しかし、DNA チップや DNA マイクロアレイと呼ばれる技術が急速に進歩し、この状況は変化しつつある。遺伝子発現のデータが体系的に得られる見通しが立ってきたため、遺伝子の相互関係を大量データから学習する試みが始められた。本講演では、DNA マイクロアレイより得られるデータの特徴と前処理方法、そしてデータから遺伝子ネットワークを学習する際のアルゴリズムについて概観する。

2 DNA マイクロアレイ

DNA マイクロアレイは、細胞内の遺伝子発現を高速かつ大量に観測するための装置である [19, 21]。cDNA と呼ばれる遺伝子配列のコピーを稠密にスポットしたガラス表面に、細胞から抽出し蛍光標識した遺伝子転写産物を結合させることで、遺伝子発現の変化を測定できる。測定サンプルにより発現の絶対量は異なるため、正常例と異常例のそれぞれを異なる色で蛍光標識し、標準となるスポットで正規化した後に、二色の相対強度を遺伝子発現量として測定する [8]。(DNA チップもマイクロアレイとほぼ同義に用いられるが、ガラス表面への DNA の植え付け法が既知配列のスポッティングではなく、光リソグラフィーを用いた固相合成であることが多い。) 大量生産可能なプレート 1 枚で数千の遺伝子発現を一度に観測できることから、その利用は急速に拡大した。今後の分子生物学はより定量的になり、大量データからのモデリングが重要になると考えられている [4]。

DNA マイクロアレイのデータは、主に実験装置を持つ研究室でクラスタリング [3, 10]、自己組織化マップ [24]、主成分分析 [2] 等で解析されてきた。しかし、結果は遺伝子間の距離 (または類似度) の定義に依存して変化するため、その解釈は難しいとされる [4]。またこれらの手法は、発現パターンが似た遺伝子は機能的にも関連するという前提に基づいているが、これが生物学的に正しい保証はない。そのため、遺伝子の依存関係そのものをデータから学習する試みとしてブーリアンモデル

*〒 135-0064 江東区青海 2-41-6 tel: 03-3599-8080, e-mail: m-arita@aist.go.jp, URL: http://www.cbrc.jp

[1]、線形モデル [14]、ニューラルネットワーク [25] が提案されてきた。これらのモデルでは、実際に大量データを処理するには至っていない。初めて 800 遺伝子という大量データの解析に踏み込んだのがベイジアンネットである [12]。

マイクロアレイ解析の問題点として常に挙げられるのは、1. 遺伝子発現は細胞内メカニズムの一面にすぎないこと、2. 数千の遺伝子に対し統計的に有意な結果を出せるほどデータが揃わないこと、3. データのノイズが非常に大きいことである。これらの疑問に対し、ベイジアンネットは以下のように対処できる。

1. マイクロアレイ解析は細胞集団の平均値を観察する技術であり、遺伝子発現の大局的な依存関係が観測される。遺伝子 A が B を活性化する、(または抑制する) という関係は、各遺伝子を変数とみなした条件付き確率モデルで表現できる。胞子形成 [13] や体節形成 [17] において鍵となる遺伝子の存在が知られているように、重要遺伝子のスクリーニングとしてベイジアンネットは有効である。
2. 観測する遺伝子数は数千あるが、データセットは多くても数十である。しかし、遺伝子ネットワークは非常に疎なグラフ構造を持つ。よって局所的な構造を学習することでデータの少なさをカバーできる。具体的には各遺伝子と相関の高いものだけを親とみなし、学習を開始する。また、学習結果の検証にはブートストラップ法を用いる。
3. アレイで見る発現とは配列にも依存した物理化学反応 (水素結合) であるため、実験の条件や計測法により発現量は大きく変化する。そのため、遺伝子発現のデータは発現レベルに応じて離散化する。マイクロアレイを用いた解析の多くで、データは -1, 0, 1 の 3 段階 (コントロールより非常に低い、同程度, 非常に高い) に分けられる。通常のしきい値は、発現量が数倍変化する点に設定する [8]。

3 ベイジアンネットの学習

大量データから効率よくネットワークを推定するには、構造を局所的に学習してゆかねばならない。このためのテクニックをいくつか紹介する [23]。ここではベイジアンネットの一般的な定義は省略する。

3.1 アウトライン

ネットワークの学習に重要なのは、候補となるネットワークの採点法と候補の探索法の二点である。大まかなアルゴリズムは以下のようになる。

入力:

- 訓練セット $D = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$
- 初期ネットワーク B_0
- ネットワークの採点関数 score
- 親頂点の数 k

出力: ネットワーク B

Loop 収束するまで (n = 1,2,...)
  /* ステップ 1. 親頂点を制限して探索 */
    foreach 確率変数 X_i
      D と B_{n-1} を使い、親頂点を k 個選択.
    選んだ頂点で、グラフ H_n を構成.
  /* ステップ 2. 採点関数で最適化 */
    score を最大化する $B_n = \langle G_n, \Theta_n \rangle$ を構成.
    (ただし $G_n \subset H_n$)
Return B_n

アルゴリズムが正しく動くためには、ループが収束する条件が必要になる。単純には、score が良くならない時に終了とする。また、親頂点の数を k に制限せず常に $H_i \subset H_{i+1}$ が成り立てば、score は減少しないので収束することは明らかである。

3.2 親頂点の選択

全ての頂点間に依存関係を仮定してネットワークを探索するのは計算量的に不可能である。そのため、各頂点には親頂点が k 個しかないという仮定を導入する。(これでも計算量は指数的に膨大な数になる。) 確率変数 X に相関のある頂点を選択する距離基準として一般的なのは、相互情報量である。

$$I(X,Y) = I(Y,X) = \sum_{x,y} \bar{P}(x,y) \log \frac{\bar{P}(x,y)}{\bar{P}(x)\bar{P}(y)}$$

ただし $\bar{P}$ はデータ中に観測された確率 (頻度) である。相互情報量は X と Y の依存度の大きさを示す。

しかし、相互情報量の大きい順に親頂点を定数個しか選択しないと、間接的に相関の大きな親を選択する可能性があり良い結果を与えない [12]。そのため、距離として相対エントロピー (Kullback-Leibler divergence) を計算せねばならない。

$$D_{KL}(P(X)||Q(X)) = \sum_X P(X) \log \frac{P(X)}{Q(X)}$$

相対エントロピーとは分布 Q(X) を真の分布 P(X) の代わりに用いたときの 'ずれ' の度合いである。具体的にはネットワーク B が与えられたとき変数 X, Y の距離を

$$I(X,Y|B) = D_{KL}(\bar{P}(X,Y)||P_B(X,Y))$$

とする。遺伝子数を考えると、実際は $k = 3$ 程度の探索しかできないと思われる。ここで $P_B(X, Y)$ はネットワーク B における分布だが、X, Y の値をサンプリングして近似することができる。

3.3 Score の計算

訓練セット D に対し最適のネットワークを選ぶには、何らかの採点法が必要になる。最短記述長 (minimum description length; MDL) によるスコア等がよく用いられているが、これらのスコアを最大化するネットワークの探索は NP 困難であることがわかっている [7]。そのため、ネットワークを局所的に改善して準最適解を探索する。このときネットワーク全体の score 値は各変数に関する score 値の総和になっているため、キャッシュ等を用いて計算を簡略化する。

$$\mathrm{score}(G, D) = \sum_i \sum_{v \in X_i\text{親}} \mathrm{score}(X_i |\ v)$$

訓練データがネットワーク上の全ての変数値を与える (complete) 場合、MDL を一般化した以下の式で与えられる情報量基準は上記のように分解してスコア計算ができる。したがってネットワークのスコア計算は局所的に改善した部分のみの再計算で済む。

$$\begin{aligned} Q_I(B, D) \quad = \quad & \log(\text{与えられた事前確率分布}) + \\ & \sum_{ijk} N_{ijk} \log \frac{N_{ijk}}{N_{ik}} \\ & -f(N) \cdot (\text{モデル中の自由変数の数}) \end{aligned}$$

ここで N 訓練セット中のサンプル数、N_{ijk} はイベント $(X_i = k, X_i\text{の親} = j)$ の頻度である。MDL の場合は $f(N) = \log(N)/2$ となる。頂点に対する親の数が少ない方が $Q_I(B, D)$ が減るので、k 個ずつ選ばれた親頂点はここで減る可能性がある。これら基準についてのより詳しい内容については、文献 [6] が参考になる。

3.4 最適グラフの選択

次数 k のグラフでも、あるスコアを最適化するネットワークの探索は NP 困難になる [7]。そのため、サイズが 2 までの集合で尤度が高いものを選び、それらの集合和を親とするアプローチや、あらかじめ固定サイズの親集合を選んでから局所的な探索をおこなうアプローチが考えられる。前者の場合には、$\chi 2$ 乗検定により相互情報量を近似することが役立つかもしれない。

$$\chi^2(X, Y) \sim I(X, Y) N \ln(4)$$

後者の場合には、遺伝的アルゴリズムが応用できるだろう。

4 データへの応用と考察

Friedman らは、800 遺伝子の発現をベイジアンネットで学習させ、

`www.cs.huji.ac.il/labs/compbio/expression/`

で公開している。データは細胞周期データベース [22]

`http://cellcycle-www.stanford.edu`

より入手可能である。学習結果は、粗ではあるものの十分入り組んだネットワークで、他の頂点に対し大きな確信度で影響する頂点がごく僅か存在する。Friedman らはこれらの頂点が細胞周期に重要な遺伝子を含むとしているが [12]、遺伝子の生物学的な重要性と他の遺伝子への影響力は関係がないと思われる。(発現量に相関があるからといって、その遺伝子間関係が重要である必要はない。逆に、重要な遺伝子に大きな確信度で影響する遺伝子は重要であろう。) 正解がわからないため明確な判断は下せないが、確率モデルの学習結果だけから生物学的な知見を引き出すのは非常に難しい。

細胞周期のように有名なメカニズムには、生物学者によってある程度グラフモデルができている場合が多い。このような知識をもっと効率よく利用する枠組みを今後は発展させるべきであろう。例えば、入力として与える初期ネットワーク B_0 に既知のグラフモデルを用い、適切な事前確率分布を与えることは有効であろう。各遺伝子を頂点としたグラフは、例えば

`http://www.genome.ad.jp/kegg/`

より入手可能である。最適グラフの選択は学習アルゴリズムのボトルネックでもあるので、より '面白い' 結果がでると考えている。(しかし Friedman らの実装したアルゴリズムは上で紹介したものよりも雑な作りになっているため、より完成度の高いツールを用いて再学習することも重要である。)

生物学者が長年積み上げてきた実験事実を、数回のマイクロアレイ実験 (プラス機械学習!) で覆すことは難しいが、膨大な実験データを利用すれば、'発見' とまではいかないが '検証' を高い確率で行えるシステムが作成できると考えている。

参考文献

[1] T. Akutsu, S. Kuhara, O. Maruyama, S. Miyano, "Identification of gene regulatory networks by strategic gene disruptions and gene over-expressions," *Proc 9th ACM-SIAM SODA*, pp. 695–702, 1998.

[2] O. Alter, P.O. Brown, D. Botstein, "Singuler value decomposition for genome-wide expression data processing and modeling," *Proc Natl Acad Sci USA*, vol. 97, no. 18, pp. 10101-10106, 2000.

[3] A. Ben-Dor, Z. Yakhini, "Clustering Gene Expression Patterns," *Proc 3rd ACM RECOMB*, pp. 33-42, 1999.

[4] A. Brazma, J. Vilo, "Gene expression data analysis," *FEBS letters*, vol. 480, no. 1, pp. 17-24, 2000.

[5] T.A. Brown, "Genomes," BIOS Scientific Publishers, 1999. (邦訳「ゲノム：新しい生命情報システムへのアプローチ」村松正實 監訳, メディカルサイエンスインターナショナル, 2000.)

[6] E. Castillo, J.M. Gutierrez, A.S. Hadi, "Expert systems and probabilistic network models," Springer, 1997.

[7] D.M. Chickering, "Learning Baysian networks is NP-complete," *Learning from Data: Artificial Intelligence and Statistics V* (D. Fisher, H.J. Lenz, Ed.), Springer, 1996.

[8] J.L. DeRisi, V.R. Iyer, P.O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, pp. 680-686, 1997.

[9] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, "biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids," Cambridge University Press, 1999. (邦訳「バイオインフォマティクス」阿久津達也, 浅井潔, 矢田哲士訳, 医学出版, 2001.)

[10] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci USA*, vol. 95, no. 25, pp. 14863-14868, 1998.

[11] N. Friedman, D. Geiger, M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131-163, 1997.

[12] N. Friedman, M. Linial, I. Nachman, D. Pe'er, "Using bayesian Network to Analyze Expression Data," *Journal of Computational Biology*, vol. 7, pp. 601-620, 2000.

[13] A.D. Grossman, "Genetic networks controlling the initiation of sporulation and the development of genetic competence in Bacillus subtilis," *Annual Review of Genetics*, vol. 29, pp. 477-508, 1995.

[14] P. D'Haeseleer, X. Wen, S. Fuhrman, R. Somogyi, "Linear modeling of mRNA expression levels during CNS development and injury," *Proc 4th Pacific Symposium on Biocomputing (PSB)*, vol. 4, pp. 41-52, 1999.

[15] D. Heckerman, D. Geiger, D.M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197-243, 1995.

[16] T. Jaakkola, M. Diekhans, D. Haussler, "Using the Fisher kernel method to detect remote protein homologies," *Proc 7th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pp. 149-158, Menlo Park, California, AAAI Press, 1999.

[17] P.A. Lawrence, "The Making of a Fly: the genetics of animal design," Blackwell Science, 1992.

[18] K. Nakai, "Protein sorting signals and prediction of subcellular localization," (review) *Advanced Protein Chemistry*, vol. 54, pp. 277-344, 2000.

[19] *Nature Genetics*, "The Chipping Forecast," vol. 21 supplement, pp. 1-60, 1999.

[20] B. Rost, "PHD: predicting one-dimensional protein structure by profile-based neural networks," *Methods in Enzymology*, vol. 266, pp. 525-539, 1996.

[21] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, "Quantitative monitoring of gene expression pattern with a complementing DNA microarray," *Science,* vol. 270, pp. 467-470, 1995.

[22] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273-3297, 1998.

[23] J. Suzuki, "A construction of bayesian networks from databases on an MDL principle", *Proc 9th UAI*, Morgan Kaufmann, pp. 266-273, 1993.

[24] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc Natl Acad Sci USA*, vol. 96, no. 6, pp. 2907-2912, 1999.

[25] D.C. Weaver, C.T. Workman, G.D. Stormo, "Modeling Regulatory Networks with Weight Matrices," *Proc 4th Pacific Symposium on Biocomputing (PSB)*, vol. 4, pp. 112-123, 1999.