Vol.2010-MPS-81 No.19
Vol.2010-BIO-23 No.19
2010/12/17

遺伝子転写制御領域を利用した遺伝子群の共通性探索について

赤坂 志津[†]　權 娟大[††]　宮崎 智[††]

マイクロアレイを用いた生化学的実験によって見つけられた共発現する遺伝子を転写因子群から分類する手法を考える．本研究では，シスエレメントの共起性を利用して，共発現遺伝子の共通性を見いだし，転写制御の観点から関連性の最も高い遺伝子をグループ化することを目的とする．まず，散在したシスエレメント情報からシスエレメント群を遺伝子ごとにシスモジュールとして統合し，それらのデータを格納した Cis-Module Database（Cis-DB）を構築する．次に，Cis-DB より得られた転写因子の共起性情報に対して，相互情報量を用いた遺伝子のクラスタリングを行う．実験の結果，転写制御機構に関連性の高いと考えられる遺伝子群をグループ化することができた．

Classification of Gene Groups Based on Transcriptional Regulation Region and its application

Shizu Akasaka[†], Yeondae Kwon[††] and Satoru Miyazaki[††]

In microarray-based gene expression analysis, it is difficult to identify transcriptional factors (TFs) involved in co-expression only by biochemical experiments. Our aim is to group genes having strongest connection in view of cis-element pattern related to co expressed genes. First, we grouped together cis-element per gene as cis-modules and constructed Cis-Module Database (Cis-DB). Then, we applied mutual information criteria to co-occurrence of TFs obtained from Cis-DB. As an experimental result, we can cluster gene groups which are involved in a same transcriptional regulation mechanism.

1. 背景・目的

2003 年にヒトゲノムプロジェクトが終了し，全ヒトゲノム配列が解読された．現在までに，多くの研究者が個々の遺伝子機能の解明へ向け，様々な研究を行ってきた．我々は，特定の条件下で有意に発現変動している遺伝子群を見いだしたいが，膨大な数の遺伝子に対して，それぞれ遺伝子発現の変動量を調査するのは効率的ではない．そこで，一度に数万個の遺伝子発現情報を得られる手法として，マイクロアレイデータ解析が近年着目されている．特定の条件下で得られたマイクロアレイデータに対して適切な統計解析手法を適用することによって発現変動の見られる遺伝子を抽出できれば，何らかの共通性を持つ遺伝子群の機能解明に向けた大きな手がかりになる．

特定の条件下で共発現する遺伝子群が存在する場合，それらの遺伝子群は共通の転写因子によって制御されることが予想される．しかし，マイクロアレイを用いた遺伝子発現解析においては遺伝子の共発現パターンが大量に見出されるため，生化学的実験のみによって共発現に関与する転写因子群を特定することは困難である．そこで，遺伝子の上流に存在し，転写因子が認識するシスエレメントと呼ばれる特定の塩基配列パターンに着目する．遺伝子の転写制御に関わるシスエレメントの出現パターンに共通性がある場合，遺伝子の転写制御機構にも共通性があることが予想される．本研究では，制御因子であるシスエレメントの共通性を利用して，共発現している遺伝子の共通性を見いだし，最も関連性の高い遺伝子をグループ化することを目的とする．

2. 準備

2.1 転写反応

遺伝子発現にとって重要な転写反応は，転写因子と呼ばれるタンパク質が遺伝子の上流又は下流に存在する特定の塩基配列を認識することで促進もしくは抑制される（図 1）. この特定の配列はシスエレメントと呼ばれ, 4～20bp 程度の短い配列である．特に, 真核生物の転写制御においては, 複数の制御因子が作用することによって, 個々の因子が単独で発揮するよりも発現量が上がるなど，強い効果を発揮することが分かっている[1]．このように，ある遺伝子の発現において相乗効果を及ぼしあうシスエレメント群をシスモジュールと呼ぶ．転写制御に関わる因子の種類やメカニズムなどは遺伝子特異的と言われており，様々なシグナルに応じた遺伝子発現の制御に関与している．

[†] 東京理科大学大学院　薬学研究科　薬学専攻
Graduate School of Pharmaceutical Sciences, Tokyo university of Science
[††] 東京理科大学　薬学部　生命創薬科学科
Faculty of Pharmaceutical Sciences, Tokyo University of Science

Vol.2010-MPS-81 No.19
Vol.2010-BIO-23 No.19
2010/12/17

転写制御メカニズムを解明するためには，どのようなシスエレメントが協調して転写制御に関与しているかといった情報が重要であるが，シスエレメントに関する網羅的な解析はこれまでになされていない．

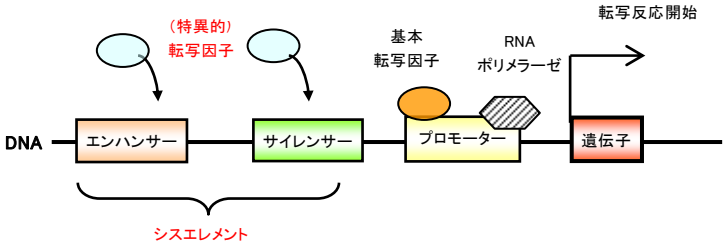


図 1　転写制御に関わる制御因子とその反応

2.2 シスエレメントの網羅的解析の必要性

遺伝子の発現制御には複数のシスエレメント（シスモジュール）が協調して働いている．したがって，遺伝子の発現制御メカニズムを解明するためにはシスモジュールを対象とした解析が必要である．

現在，多くのシスエレメント配列が研究され，それらの配列がデータベース上に登録されている．しかし，それらのシスエレメント情報をシスモジュールとして統合しているデータベースはほとんど存在しない．また，現在公共データベース上に登録されているシスエレメント情報には以下の二つの問題が存在する．一つ目は，制御因子（転写因子・シスエレメント）と遺伝子間の関連性が無いことである．

二つ目は，シスエレメント群（シスモジュール）の情報がデータベース上に散在していることである．以下，これらの問題について具体的に説明する．

（1）制御因子群とそれに制御される遺伝子群の関連性の欠如

既存の転写因子データベースとして，JASPAR[2]やTRANSFAC[3]などが挙げられ，これらのデータベースには転写因子とそれが認識するシスエレメント配列情報が記載されている．しかし，制御因子と遺伝子の間の関連性は不明である．したがって，遺伝子の発現制御に関わる因子と遺伝子との直接的な関連性が得られず，遺伝子発現機構を考える上では十分ではない．

（2）シスエレメント情報が散在して存在している現状

生化学的実験によって保証されたシスモジュール情報が公共データベースである国際塩基配列データベースに登録されている．同一の遺伝子に関するシスモジュールの情報が異なる研究者によって登録されていることがよくある（図 2 の Entry_1 と Entry2）．図 2 において，転写因子 SP1 と AP2 がそれぞれ認識するシスエレメント配列が Entry_1 の上流には存在するが，Entry_2 には存在しない．また，Entry_2 には NFKB が認識するシスエレメント配列が 3 箇所見られるが，Entry_1 では 2 箇所しか見られない．すなわち，各エントリが異なるシスエレメント情報を有することが分かる．

遺伝子ネットワークを構築する上では，上述したように網羅的なシスエレメント情報の解析が必要となる．そのため，本研究では，散在したシスエレメント情報からシスエレメント群を遺伝子ごとにシスモジュールとして統合し，シスモジュールデータベース（Cis-Module Database）を構築した．

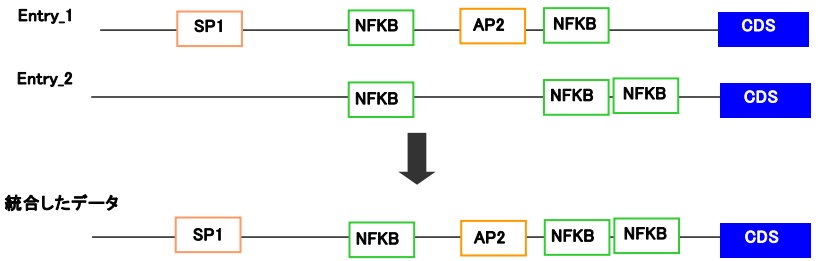


図 2　シスエレメント情報の統合方法

3. 方法

3.1 Cis-Module Databaseの構築

ヒト（Homo sapiens），マウス（Mus musclus），ラット（Rattus norvegicus），ショウジョウバエ（Drosophila melanogaster），酵母（Saccharomyces serevisiae）の 5 生物種を対象として，本データベースを作成した．

以下，具体的にシスエレメントデータベースの構築手法について述べる．なお，（ｉ）~（ⅲ）の数字は図 3 中の数字に対応している．

（ｉ）データ取得

INSDC（International Nucleotide Sequence Database Collaboration：国際塩基配列データベース）の一つである DDBJ（DNA Databank of Japan）を利用し，生化学的実験によって保証されたシスエレメント情報を含むレコードを取得した．なお，本レコード

Vol.2010-MPS-81 No.19
Vol.2010-BIO-23 No.19
2010/12/17

は，DDBJ が提供するキーワード検索システムである ARSA を利用して行った．同時に，Ensembl genome browser を用いて各生物種のゲノムデータを取得した．ゲノム配列を取得した理由は，個々の研究者が独立に登録した DDBJ のデータを，共通の遺伝子に注目して再編する基準として利用するためである．

（ii）DDBJ レコードのゲノム配列上での位置の特定

シスエレメント情報を持つ DDBJ レコードの中から，タンパク質コード領域（CDS）を抽出し（以下 CDS/DDBJ とする），同時にゲノム配列側からも CDS を抽出した（以下 CDS/Genome とする）．

続いて, CDS/Genome をテンプレートとして CDS 同士のアラインメントを行った[4]. 抽出した DDBJ レコードの CDS 領域がゲノム上でどこに位置するのかを特定することにより，CDS 上流領域に存在するシスエレメントのゲノム上の位置を把握できる．

（iii）上流配列の抽出と比較

CDS 同士のアラインメントによってヒットした CDS/DDBJ と対応する CDS/Genome より，各上流配列を抽出した．その後（ii）と同様にして，それぞれから抽出した上流配列同士を比較して，CDS 上流配列のゲノム上での位置を特定した．

以上の操作を経て，各遺伝子の上流配列情報を取得した．ゲノム上にこれらの情報を統合し，遺伝子毎にシスエレメント情報を統合したフラットファイルを作成した．

本データベースは Web 上で公開しており（http://www.pharmacoinformatics.jp/cis/），遺伝子上流におけるシスエレメントの分布が視覚的に確認可能である．各々のレコードを参照することで，各遺伝子の発現に関するシスエレメント，及び，シスエレメントが形成するモジュールを参照することができる．また，シスエレメントのゲノム上での位置や分布状態を調べることも可能である．このように，本データベースは遺伝子の発現メカニズムの全体像を反映している点で既存のデータベースと異なる．

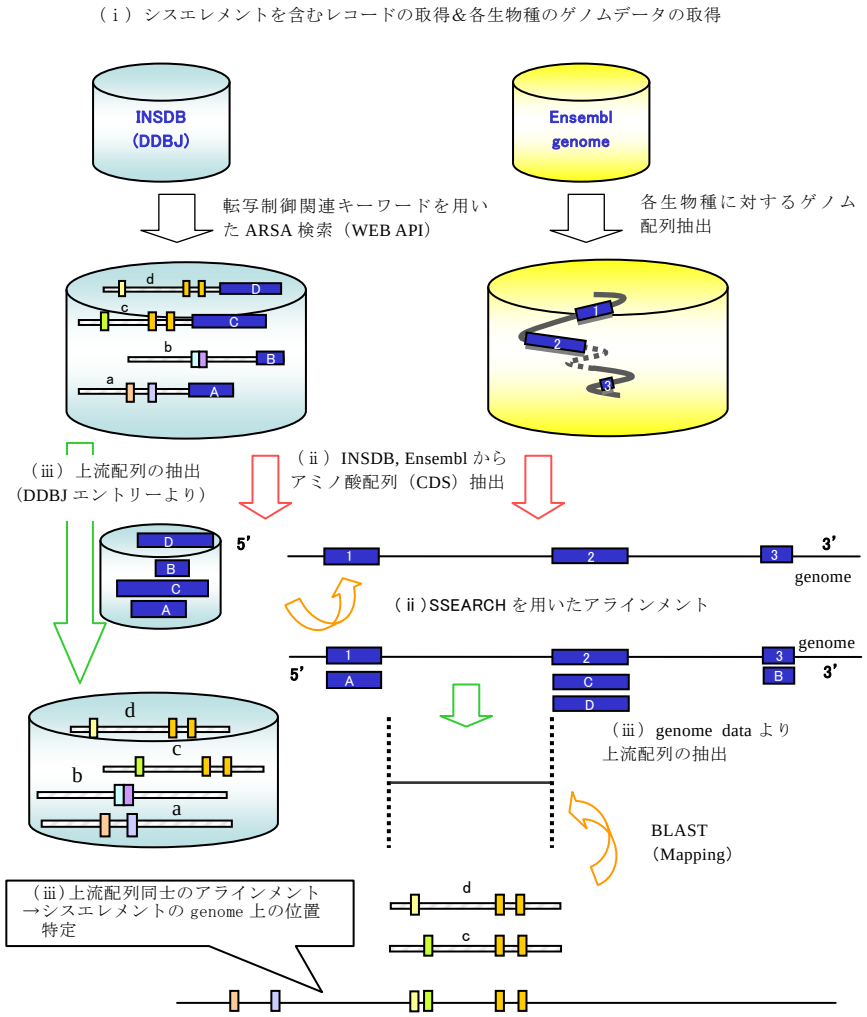


図 3　Cis-Module Database の作成手順

Vol.2010-MPS-81 No.19
Vol.2010-BIO-23 No.19
2010/12/17

3.2 遺伝子のグループ化の概要

ある条件下でマイクロアレイを用いた遺伝子発現解析により，共発現する遺伝子群が抽出された場合に，転写制御の観点から遺伝子をグループ化する手法について以下に示す．共発現が確認された各遺伝子の上流配列中に，図 4 で示すように各遺伝子の複数のシスエレメントパターンが見い出されたとする．各遺伝子の上流配列中に存在するシスエレメントパターンの中で，共通して存在する複数のシスエレメントがあった場合（図 4 では gene A, D, Q に対して CE1 と CE3 が共通），それらのシスエレメント群が遺伝子の共発現に寄与している可能性が予想される．

遺伝子の共発現に寄与しているシスエレメント群が把握できれば，それらのシスエレメントを認識する転写因子（図 4 ではそれぞれ TFa と TFb）も間接的に把握可能であり，これを元にした遺伝子のグループ化が可能となる．

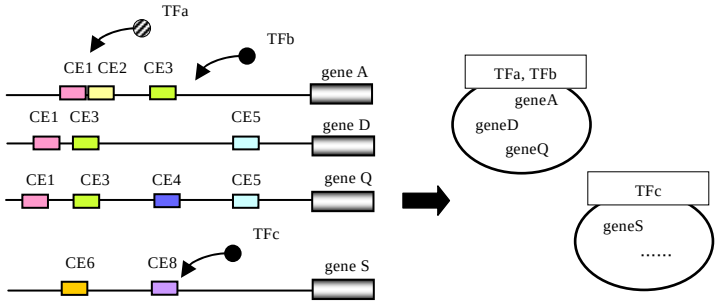


図 4 遺伝子グループ化の模式図

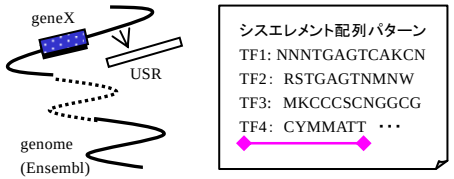
しかし，シスエレメントパターンの有無を機械的に探索した段階では，遺伝子の共発現に寄与している転写因子群が予想されたに過ぎず，シスエレメント様配列，すなわち，その条件下で実際に転写制御に関与していない擬陽性のヒットも多数含まれている可能性がある．このような擬陽性のヒットを除くために，本研究で作成した Cis-Module Database（以下 Cis-DB とする）を利用できる．Cis-DB は，生化学的実験によって保証されたデータに基づき遺伝子ごとに発現制御に関わる因子の情報を統合したものであり，各遺伝子に対して，発現変動に関与する転写因子群，またはシスエレメント群の共起性情報を得ることができる．

しかし，Cis-DB には転写因子・シスエレメントの共起性情報が十分に登録されていないため，シスエレメントの擬陽性のヒットを十分に削除できない可能性がある．そこで，本研究では，マイクロアレイを用いた遺伝子発現解析において，発現変動したと判断された遺伝子群を解析対象データとして，転写因子データベースである TRANSFAC に登録されている制御因子情報を利用し，遺伝子のグループ化を試みた．

3.3 遺伝子グループ化の手法

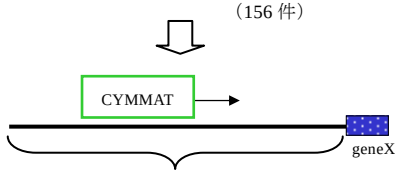
(1) 遺伝子の上流配列に対するシスエレメントパターン解析

本研究では，遺伝子グループ化のデータセットとして，カロリー制御条件下で 31099 遺伝子を対象としたラットの GeneChip を解析した結果，発現変動していると判断された 52 遺伝子を選定した．まず，これらの 52 遺伝子について上流配列（その遺伝子の開始点から上流 2000bp～下流 200bp の配列，以下 USR とする）の取得を試みた．その結果，重複遺伝子などを削除し，42 件の遺伝子に対して遺伝子 USR を取得した（図 5-ⅰ）．次に，転写因子データベースである TRANSFAC より，シスエレメント配列と転写因子の情報を抽出した．生物種をラットに限定して，制御因子情報を抽出した結果，156 件の転写因子とそれが認識するシスエレメントパターンが得られた（図 5-ⅱ）．次に，42 件の各遺伝子の USR を対象として，TRANSFAC に登録されている全 156 件のシスレメントパターンの存在の有無を機械的に探索した（図 5-ⅲ）．最後に，各遺伝子についてシスエレメントパターンを探索した結果を統合し，図 5-ⅳに示すようなシスエレメントの共起性テーブルを作成した．



（ⅰ）42 遺伝子の上流配列（USR）取得

（ⅱ）TRANSFAC より抽出したシスレメントパターン（156 件）



(ⅲ) シスエレメントパターンの機械的探索

	TF1	TF2		TF156
gene 1	1	1	…	0
gene 2	0	1	…	0
gene 3	0	0	…	0
……			…	
gene 42	0	1	…	1

（ⅳ）TF 共起性テーブル

作成した表は各遺伝子の上流配列中にシスエレメントパターンが存在する場合は 1，存在しない場合は 0 が示されている（遺伝子 42 件×転写因子 156 件）．

この例の場合，gene1 の上流配列中には TF1，TF2 が認識するシスエレメントパターンは存在するが，TF156 を認識するシスエレメントパターンは存在しないことを示す．

図 5 シスエレメントのパターン解析

(2) 2 遺伝子遺伝子間の相互情報量計算

表 1 はシスエレメントパターン解析により得られた転写因子の共起性テーブルのイメージであり，上流配列中に転写因子が認識するシスエレメントパターンが存在する

Vol.2010-MPS-81 No.19
Vol.2010-BIO-23 No.19
2010/12/17

場合は”1”，存在しない場合は”0”とした．

本研究では 2 つの遺伝子間の発現に関する関連性，すなわち制御メカニズムの共通性の度合を評価するために，相互情報量（Mutual Information，以下 MI とする）を採用した[5]．MI とは，2 つの確率変数間の依存の程度を示す尺度であり，本研究においては，この値が大きいほど，両遺伝子の発現制御機構に強い関連性があることを示す．MI は式(1)で定義される．

$$MI(X,Y)=\sum_{\substack{i=0,1\\ j=0,1}} P_{XY(i,j)}\log_2\left(\frac{P_{XY(i,j)}}{P_{X(i)}P_{Y(j)}}\right) \quad (1)$$

ここで，X，Y は，比較対象とする 2 つの遺伝子を示し，$P_{X(i)}$，$P_{Y(j)}$は，遺伝子 X と Y それぞれが転写因子によって独立に制御される確率，$P_{xy(i,j)}$は，遺伝子 X と Y が同時に制御される確率を示す．本研究では，式（1）をもとに，42 遺伝子に関して考えられるすべての遺伝子間（$_{42}C_2$＝861 通り）について相互情報量 MI を計算した．

表 1　転写因子共起性テーブル

	TF1	TF2		TF156
gene 1	1	1	…	0
gene2	0	1	…	0
gene M	0	1		1
……	…	…	…	…
geneN	1	1		0
……	…	…	…	…
gene 42	0	1	…	1

(3)　相互情報量を用いたクラスタリング

各遺伝子を転写制御の観点からグループ化するために，階層的クラスタリングを行う．本研究では，各遺伝子間の類似度の指標として相互情報量を用い，クラスタリング手法として 2 遺伝子間のユークリッド距離・Ward 法を選定した．

階層的クラスタリングとは，双方の類似度（あるいは非類似度である距離）に基づいて，最も似ている個体から順次集めてクラスタリングを行う手法のことである．本研究においては式(1)で計算した 861 通りの MI のそれぞれを一つのクラスタと考え，ある初期状態から次々とクラスタを結合し，最終的に一つのクラスタになるまで階層構造を作成していく．ここで MI は各遺伝子上流に存在する各転写因子が認識するシスエレメントパターンの存在の有無を起点として計算されたものであるため，共通の制御因子群による制御に関与する可能性の高い遺伝子同士をグループ化することが可能となる．また，情報量を用いてクラスタリングを行っているため，2 つの遺伝子間の制御機構の共通性の度合いを測ることができる．

4.　実験

機能未知の遺伝子を含めてクラスタリングを行うことにより，機能未知遺伝子の制御パターンの共通性から発現制御機構の予測を試みた．

図 6 は共発現が確認された 42 件の遺伝子に機能未知遺伝子（以下 unknown とする）を加えた 43 件の遺伝子に対してクラスタリングを行った結果である．各グループに属する遺伝子は，転写制御の観点から，ある条件下でカロリー制御に関わる遺伝子のネットワークの中で，それぞれ局所的なネットワークを形成していると考えられる．unknown が属しているグループ⑥の遺伝子には，”Ester hydrolase C11orf54 homolog”, “a neurotrophic factor that is involved in neuronal cell protection and cell survival”，”F-box and leucine-rich repeat protein 2”などの注釈が付いており，グループ内の各々の遺伝子の機能が異なっている．この結果から，遺伝子の機能が異なる場合であっても，転写制御機構に対する共通性を用いて，遺伝子転写制御において関連性の強い遺伝子群をグループ化することが可能であることが示唆された．unknown の遺伝子機能は不明であるが，グループ⑥に属する遺伝子と転写制御機構の観点から，共通性を持つことが予想される．

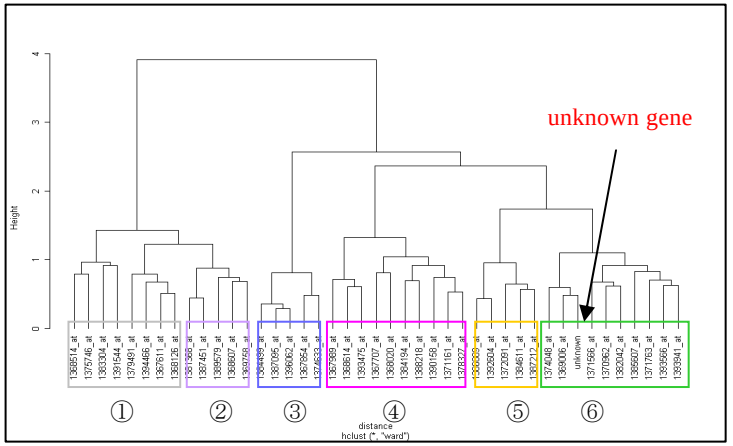


図 6　43 遺伝子のクラスタリング結果

Vol.2010-MPS-81 No.19
Vol.2010-BIO-23 No.19
2010/12/17

5. まとめと今後の課題

本研究の結果，転写制御機構の観点から共通性を持つ遺伝子同士は，そのグループ内で転写制御に影響を及ぼしあっている可能性があると考えられる．したがって，機能未知の遺伝子が存在する場合であっても，相互情報量を用いた遺伝子のクラスタリングを通じて共通の転写制御機構を持つ遺伝子のグループを見出すことが可能である．つまり，マイクロアレイなどを用いて遺伝子発現解析を行うことなく，遺伝子の制御機構の共通性から遺伝子間の転写制御ネットワークを予測することが可能であることが示唆された．

本研究では Cis-DB の情報量が少なかったため，暫定的に TRANSFAC の制御因子情報を利用して，各遺伝子の転写制御に関わり得るかということを機械的に探索したが，この方法では上流配列中にヒットしたシスエレメントパターンがシスエレメント様配列である可能性があり，また実際は転写制御に関わっていない擬陽性のヒットも多く含まれる可能性がある．したがって，今後，Cis-DB の充実を図ることが必要不可欠である．

更に，機能未知遺伝子も含めた遺伝子のクラスタリングを行ったが，機能未知遺伝子は総じて機能既知遺伝子との配列相同性がないものが多い．したがって，今後，今回の手法によって同じグループに属しているということが予想された遺伝子について，同様の制御機構を持つ遺伝子同士の配列相同性を調査する予定である．配列相同性の探索によって，配列に相同性がないことが確認された場合，遺伝子の配列相同性や機能に類似性がない場合であっても，転写制御の観点から遺伝子をグループ化する新たな手法を提案することにつながる．

また，今回対象とした 42 件の遺伝子のオーソログ遺伝子を異なる生物種，例えばヒトやマウスから抽出し，同様のクラスタリングを行う予定である．生物種間でクラスタを比較することによって，転写制御ネットワークに相違があるか否かを比較・検討したい．

謝辞 本研究のデータセットとしてラットのマイクロアレイデータを提供してくださった東京理科大学薬学部樋上賀一教授に心より御礼申し上げます．

参考文献

[1] Tuch B. B., Li H. and Johrnson A. D.: Evolution of eukaryotic transcription circuits, Vol.319, Science, pp.1797-1799 (2008).
[2] Sandelin A., Alkema W., Engström P., Wasserman W. and Lenhard B.: JASPAR: an open access database for eukaryotic transcriptional factor binding profile, Nucleric Acids Research, pp.D91-D94 (2004).
[3] Wingender E., Dietze P., Karas H. and Knüppel R.: TRANSFAC: a database on transcription factors and their binding sites, Nucleic Acids Research, Vol.24, No.1, pp.238-241 (1995).
[4] Smith T. F. and Waterman M.S.: Identification of common molecular subsequences, J. Mol. Biol, Vol.147, pp.195-197 (1981).
[5] Fuhrman S. and Somogyi R.: Reveal, a general reverse engineering algorism for inference of genetic network architecture, Pacific Symposium on Biocomputing, Vol.3, pp.18-29 (1998).