Vol.2011-MPS-86 No.10
Vol.2011-BIO-27 No.10
2011/12/1

転写因子とそのシスエレメント配列の共進化解析

本郷 沙妃[†] 權 娟大[††] 宮崎 智[††]

転写因子がシスレメント配列と呼ばれる塩基配列を認識し，結合することによって，遺伝子の転写を制御している．本研究では，転写因子とシスエレメント配列の結合部位に着目し，両者には一方の配列の変化に伴い他方の配列も変化するという共進化の関係性について，配列解析による解明を試みた．まず，結合に関与する転写因子配列の進化距離を計算し進化系統樹を作成した．また該当する転写因子に認識されるシスエレメント配列の距離を情報量によって計算しクラスタリングを行った．両者のクラスタリングのトポロジーを比較することにより，両者の関係性を探索した．

Co-evolutionary analysis of transcription factors and cis-element sequences

Saki Hongo[†], Yeondae Kwon[††] and Satoru Miyazaki[††]

Transcription factors regulate gene transcription by recognizing and binding the nucleotide sequences, called cis-elements. In this study, we try to clarify a relationship between transcription factors and cis-elements by focusing on binding sites of them, assuming that there exists a co-evolutionary relationship that change in one sequence causes change in another. First, we calculated the evolutionary distances of transcription factor binding sequences and created a phylogenetic tree. Next, we performed hierarchical clustering using the entropy-based distances of the cis-elements regulated by their corresponding transcription factors. Then, we analyzed a relationship between transcription factors and cis-elements by comparing the topology of the clustering results.

1. 背景・目的

遺伝子は生物の遺伝情報を担う主要因子と考えられており，mRNA へ転写されタンパク質へ翻訳されることにより機能を発現する．この最初の過程である転写は，転写因子が遺伝子の上流または下流にあるシスエレメント配列を認識・結合することによって制御されている．転写は情報伝達において大変重要なプロセスであり，遺伝子の発現に大きく関わっていることから，転写因子とシスエレメント配列の関係性の解明により,疾患遺伝子により効果的に働く新薬開発への一助につながる.本研究では，転写因子とシスエレメント配列の結合部位に着目し，両者は一方の配列の変化に伴い他方の配列も変化するという仮説のもとで，両者の共進化性の有無についての考察を試みた．

従来の配列解析手法は，配列比較により，塩基，アミノ酸の置換，変異率などを基にした配列類似性を評価することが出来る．しかし，シスエレメント配列は 4～20 個の塩基からなる非常に短い配列であり，同じ転写因子から認識されるシスエレメント配列の間に相同性がないにも関わらず転写因子との結合に関与する特定の塩基があること，またシスエレメント配列の長さがそれぞれ異なるということから，従来の手法によるシスエレメント配列の解析は困難である．そこで，情報量の概念を応用したシスエレメント配列の解析がなされている[1]．本研究では，シスエレメント配列間の比較を可能にするため情報量の概念を導入し，シスエレメント配列群のクラスタリング結果と転写因子群のクラスタリング結果を比較することにより，転写因子の結合部位とシスエレメント配列間における関係性の解析を目的とする．

2. 準備

2.1 転写反応

遺伝子の発現は，転写因子と呼ばれるタンパク質が，遺伝子の上流または下流に存在する特徴的な塩基配列に結合・解離することで制御されている（図 1）[2]．この塩基配列は「シスエレメント（cis-element）」と呼ばれ，4～20 塩基程度の短い配列である．このうち，転写を促進する因子が結合するシスエレメント配列をエンハンサー，転写を抑制する因子が結合するシスエレメント配列をサイレンサーと呼ぶ．転写因子には,転写そのものに関わる基本転写因子と,転写の調節を行う転写制御因子がある．単に転写因子と示す場合には転写制御因子を指す場合が多く，本稿でも転写制御因子を転写因子と呼ぶこととする．

† 東京理科大学大学院 薬学研究科 薬科学専攻
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.2011-MPS-86 No.10
Vol.2011-BIO-27 No.10
2011/12/1

転写制御メカニズムの解明には，転写因子とシスエレメント配列の関係性を解析する必要がある．この解析を行う上で，例えば，ある転写因子が認識するシスエレメント配列のデータ，またそれらが制御する遺伝子とその機能，シスエレメント配列のゲノム上の位置等の情報が必要となるが，複数のデータベースを駆使しても，全ての情報を取得することは困難である．

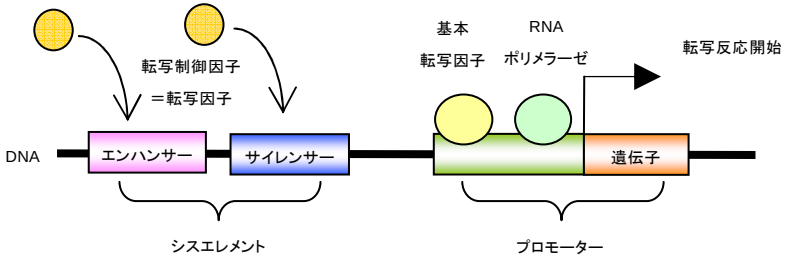

図 1　転写制御に関わる制御因子とその反応

2.2　転写因子とシスエレメントの関係性

各々の転写因子が認識するシスエレメント配列は，完全に決定された塩基の組み合わせの 1 パターンではなく，様々な配列パターンであることが明らかになっている．転写因子の結合部位のアミノ酸残基とシスエレメント配列の塩基の相互作用にはかなりの柔軟性があると言われているが[3]，その規則性は未だ明確になっておらず，転写因子とシスエレメント配列の対応関係には未知の部分が多い．

転写制御の際，転写因子はシスエレメント配列に接近し結合に関与している（図 2）．転写制御が遺伝子の情報伝達において重要なプロセスであることから，転写因子の結合部位（ドメイン）におけるアミノ酸配列とシスエレメント配列は共進化の関係にあると予想される．

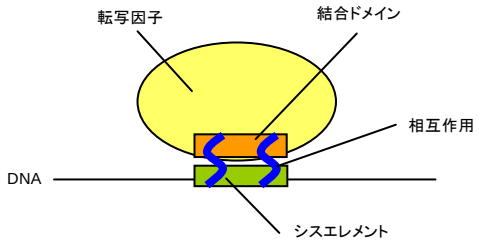

図 2　転写因子とシスエレメント配列の相互作用の模式図

2.3　情報量を用いたアプローチ

未知の配列を解析する手法として相同性検索が広く用いられている．相同性検索とは，未知の配列と類似度の高い配列群を集め，対象配列と比較することにより考察を行う手法である[4]．シスエレメント配列は一般的に 4～20 塩基の短い配列であり，その配列中には 4 種類の塩基（A，T，G，C）の組み合わせのパターンしか出現しない．また，シスエレメント配列間で相同性が低い部位に転写因子との結合に寄与するサイトがある場合や，相同性が高い部位にも関わらず転写因子との結合に寄与するサイトがない場合が報告されている．このことから，シスエレメント配列間の比較に一般的な解析手法である相同性検索を用いることは困難である．また，ClustalX[5]のような配列整列化ツールでは，スコアリングやクラスタリングの際に長さの異なる複数のシスエレメント配列群をひとまとめに扱うことは難しい．

転写因子データベース JASPAR[6]，TRANSFAC[7] に登録されているシスエレメント配列の多くは，実験により転写因子との結合が確認された塩基配列である．つまり，登録塩基配列中の全塩基が結合に関与しているかは不明であり，登録配列は結合に関与しない塩基も含めて余分に長く登録されている可能性がある．このことから，転写因子が認識するシスエレメント配列として，JASPAR や TRANSFAC に登録されているデータそのままを解析に用いることは不適である．

以上の問題点を踏まえ，本研究ではシスエレメント配列を解析するために，DNA やアミノ酸配列の解析に確率論的な捉え方を導入した方法[1]を用いることにした．情報科学の分野では，情報の価値の尺度として情報量（エントロピー）を定義している[8]．この場合の情報とは曖昧さを減らすものと考えられ，「その事象が発生する確率の度合」で表される．起こりにくい（確率が低い）事象ほどそこに含まれている情報の量が大きいと考える．関連研究の結果により，ある転写因子が複数のシスエレメント配列パターンを認識する場合，該当するシスエレメント配列間に配列ベースでは類似性は乏しいが，情報量を用いて各々の配列パターンを数量化し計算するとある特定の等しい値になることが，複数の転写因子において示唆された．そこで，シスエレメント配列を配列ベースではなく，情報量の概念を応用することで転写因子に認識されるシスエレメント配列パターンを数量化し，網羅的に比較・解析することとした．

さらに，「転写因子が認識する配列」を対象とした解析を行うためには，転写因子データベースに登録されているシスエレメント配列を洗練する必要がある．そこで，本研究では，データベースに登録されているシスエレメント配列の中で，転写因子への結合の関与が低いと予測される塩基を，情報量を用いて選出し削除することにより，配列データの洗練を行った．具体的な手法については 3.2 節で説明する．

Vol.2011-MPS-86 No.10
Vol.2011-BIO-27 No.10
2011/12/1

3. 方法・実験

まず，転写因子のアミノ酸配列とそれが認識するシスエレメント配列を JASPAR と TRANSFAC から取得し，シスエレメント配列に関しては情報量を用いて配列の洗練を行った（3.2 節参照）．次に，転写因子群とシスエレメント配列群のそれぞれのクラスタリング結果を比較することにより，転写因子とシスエレメント配列間の関係性を解析した．実験手順の概要を図 3 に示す．

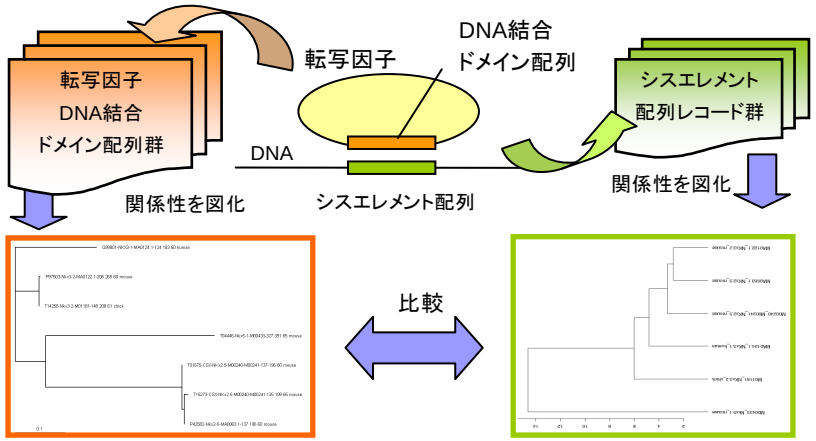

図 3　実験手順概要

3.1　配列データの取得

(1)　シスエレメント配列の取得

シスエレメント配列の取得に，転写因子データベース JASPAR と TRANSFAC を利用した．JASPAR からは，モチーフ配列と注釈された部分のみを抽出し，TRANSFAC からは，同じ ID をもつ長さの異なる配列群に関しては長さを合わせて余分な塩基を削除し抽出した．この作業は，3.2 節のシスエレメントの配列の洗練が同じ配列長でなければ行えないためである．また，一つの ID の配列レコードに複数の生物種に由来する配列が混在している場合は研究対象データから除外した．この作業は，転写因子のシスエレメント配列の認識の幅の他に，生物種の違いによる幅を除外するためである．

(2)　転写因子のドメイン配列の取得

上記で取得したシスエレメント配列には，該当するシスエレメント配列を認識する転写因子の関連情報も含まれている．TRANSFAC から取得したシスエレメント配列に関しては，対応する転写因子のアミノ酸配列を TRANSFAC のサブデータベースから取得し，JASPAR から取得したシスエレメント配列に関しては，UniProt[9]から対応する転写因子のアミノ酸配列を取得した．取得した配列に記されている注釈情報を用いて，シスエレメントとの結合に関与するドメイン配列のみを抽出した．

3.2　シスエレメント配列の洗練

転写因子データベースに登録されているシスエレメント配列は，生化学的実験により転写因子との結合が確認された塩基配列である．しかし，その配列中の全ての塩基が結合に関与しているかは不明であり，結合に関与しない塩基を含んだシスエレメント配列が登録されている可能性がある．そこで，本研究ではシャノンエントロピー[6]を指標に，シスエレメント配列の端に存在し転写因子との結合に関与しないと予想される塩基を取り除くことで，シスエレメント配列の洗練を行った．

(1)　シャノンエントロピーの定義と配列解析への適用

シャノンエントロピー（Shannon Entropy: *SE*）は情報論的エントロピー，平均情報量とも呼ばれる値で，ある事象が起きる確率を元に，ある事象が持つ「情報の量」を定義する情報理論の考え方に従う値である．情報理論における考え方では，起こりにくい（確率が低い）事象ほどそこに含まれている情報の量が大きいと考える．配列を数量化するにあたって，*SE* を当てはめて考えると，1 配列中における各塩基（A，T，G，C）の出現確率によって，その配列が持つシャノンエントロピーを定義する．

ある 1 列の配列から求められるシャノンエントロピー *SE* は以下の式で表される．

$$SE = -\sum_{i=A,T,G,C} Pi \log_2 Pi$$

ここで，*Pi* は *SE* を計算する配列中における各塩基の出現確率である．*SE* は 0 以上 2 以下の値をとる．*SE* が 0 に近いほど配列中の塩基の出現が大きく偏っていることを意味し，2 に近いほど各塩基が均等に出現していることを意味する．例として，「CCATATATAG」という配列の *SE* の計算例を以下に示す．配列中の A，T，G，C の各塩基の出現確率（P_A, P_T, P_G, P_C）は，

$$P_A = \frac{4}{10} \qquad P_T = \frac{3}{10} \qquad P_G = \frac{1}{10} \qquad P_C = \frac{2}{10}$$

である．したがって，この配列の *SE* は，以下の値となる．

Vol.2011-MPS-86 No.10
Vol.2011-BIO-27 No.10
2011/12/1

$$SE=\left(-\frac{4}{10}\times\log_2\frac{4}{10}\right)+\left(-\frac{3}{10}\times\log_2\frac{3}{10}\right)+\left(-\frac{1}{10}\times\log_2\frac{1}{10}\right)+\left(-\frac{2}{10}\times\log_2\frac{2}{10}\right)=1.8464...$$

SE が 2 に近い値を持つので，配列中の各塩基が均等に出現していることが分かる．

(2) シスエレメント配列の洗練

シスエレメント配列における各サイトの塩基出現の偏り具合を数量化し，偶然に各塩基が均等に出現している高エントロピーのサイトを，転写因子のアミノ酸配列との結合に関与しない塩基とみなし削除した．本稿におけるサイトとは，一つの転写因子配列に認識されるシスエレメント群を並べ，1 塩基毎に区切った 1 つの縦列のことを指す．シスエレメント配列の洗練手順を図 4 に示す．

操作 1 : シスエレメント配列のサイト毎に *SE* を計算し，1.9≦*SE*≦2.0 をとるサイトを削除候補サイトとする．
操作 2 : 削除候補サイトにおける塩基の種類毎に 4 つのグループに分ける．
操作 3 : 分けたグループ毎に，全サイトの *SE* の合計を求め，値に偏りがあるか調べる．
操作 4 : 手順 3 の結果，*SE* 値の偏りが小さい時の削除候補サイトを転写因子との結合に関与しない塩基とみなし，各シスエレメント配列からそのサイトを取り除く．

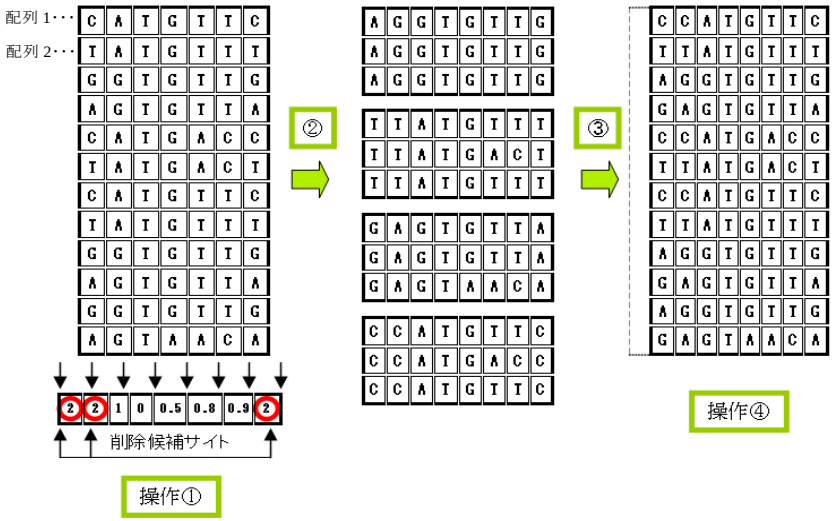

図 4 シスエレメント配列の洗練手順

3.3 クラスタリング

3.3.1 シスエレメント配列の階層的クラスタリング

シスエレメント配列間の類似度の指標として，エントロピー進化率を用い，階層的クラスタリングを行った．この操作により，シスエレメント配列パターンの揺らぎの度合が似ているレコードを集めることが出来る[10]．クラスタリングの手法として，群平均法を選定した．

(1) エントロピー進化率 (Entropy Evolutional Rate: *EER*)

エントロピー進化率 *EER* は 2 つの情報源間の関連性の度合を示す相互情報量を正規化したものである．*EER* の応用として，遺伝子配列の分類や，系統樹作成の際の距離や指標として使われることがある[11]．そのため，本研究でもシスエレメント配列の解析にエントロピー進化率を用いることとした．

2 つの情報の関連度合を調べるためには，相互情報量を計算すれば十分である．しかし，相互情報量は 2 つの情報源の大きさに依存するので，複数の相互情報量の大きさを比較する場合には，正規化した値 *EER* を用いる必要がある．同様に配列長が異なるシスエレメント配列同士を比較する場合も *EER* を導入する必要がある．

エントロピー進化率 *EER* は以下の式で計算される．

$$EER=\frac{I(X,Y)}{SE(X)+SE(Y)-I(X,Y)}$$

EER は 0 以上 1 以下の値をとる．式中の *X* と *Y* は情報源を表し，*X* と *Y* がより関連しているほど値が大きい．*I*(*X*,*Y*)は相互情報量であり，*X* と *Y* が共有している情報量を意味する．*SE*(*X*)は *X* の情報量の大きさを表す．

(2) 階層的クラスタリング

・EER の計算と頻度分布作成

転写因子ごとに，各々が結合するシスエレメント配列パターンから得られた *EER* 値を頻度分布化した．*n* 個の配列が含まれるレコードから 2 配列ずつ選び *EER* を計算し，${}_nC_2$ 個の *EER* を得た．*EER* は 0 以上 1 以下の値をとることから，0.1 の階級幅で 0 から 1 の値を 10 段階に分類し，各階級に当てはまる *EER* 値の個数を *EER* の総数 ${}_nC_2$ で割った相対値を縦軸に取り，頻度分布を作成した（図 5）．

シスエレメント配列パターン間で従属関係が見られるものが多い場合はグラフは右寄りになり，従属関係があまり見られない場合はグラフは左寄りになる．*EER* がシスエレメント配列の冗長度を表すことから，この頻度分布は転写因子のシスエレメント配列認識に対する柔軟度を表したものであるといえる．

Vol.2011-MPS-86 No.10
Vol.2011-BIO-27 No.10
2011/12/1

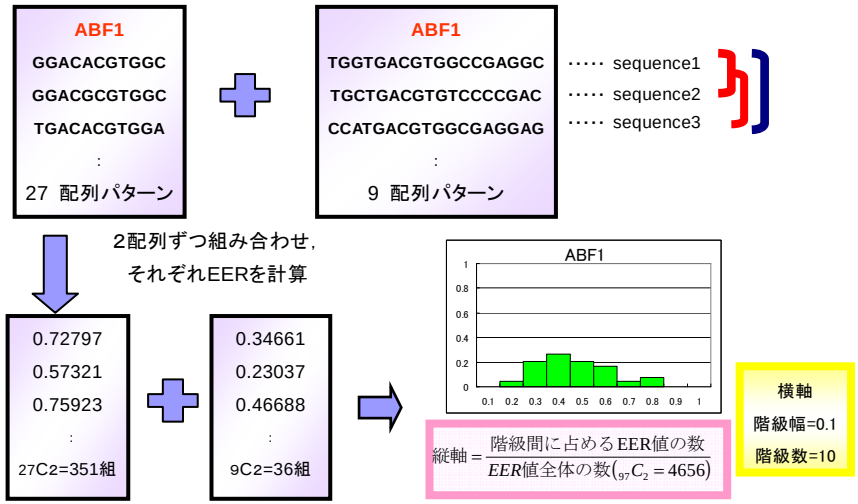

図 5　*EER* の計算と頻度分布作成

・クラスタリング手順

本研究では，クラスタリングに用いる非類似度として，各レコードから求めた *EER* より作成した頻度分布間におけるユークリッドの距離を用いた．

ユークリッドの距離はある対象間の距離の取り方の一つであり，一般的な平面，空間上の距離としても用いられているものである．平面では 2 次元を要素として距離が求められるが，その距離を n 次元の空間に当てはめるのがユークリッドの距離である（図 6）．

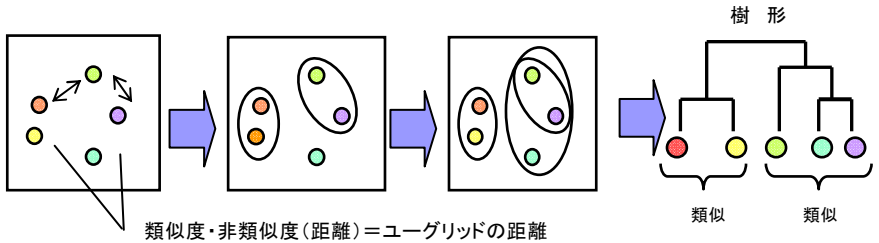

図 6　ユークリッド距離の定義式

本研究において，頻度分布 a と頻度分布 b 間のユークリッド距離 *D* は以下の式により与えた．今回，*i* は各階級における *EER* の相対値 10 次元が当てはまり，n＝10 となる．実際のクラスタリング処理は統計ソフト R を用いて行った．

$$D(a,b)=\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2}$$

3.3.2　転写因子の系統樹作成

クラスタリングに用いたシスエレメント配列の各々の配列に対応する転写因子について，結合部位のアミノ酸配列における相同性を元に配列の関係性を図化するために系統樹の作成を行った．マルチプルアラインメントと系統樹の作成には ClustalX を用いた．

4　結果・考察

転写因子のドメイン配列のクラスタリング結果と，それが認識するシスエレメント配列のクラスタリング結果のトポロジーを比較することにより，両者の関係性の解析を試みた．解析には，NKx ファミリーの homeo ドメイン，Fox ファミリーの Forkhead ドメイン，Maf ファミリーの bZIP ドメインの 3 種類の転写因子とそれらの転写因子が認識するシスエレメント配列を用いた．同じドメインを持つ同じファミリー内で，複数の生物種をもつデータという規準で解析用データを選定した．本稿では，NKx ファミリーの homeo ドメインの解析結果に着目する（図 7）．

転写因子のクラスタリング結果（図 7，左）から，Nkx の機能の近いもの同士が近隣に集合していることが分かる．このことから、転写因子のドメイン配列は生物種を超えて機能を保存して変異していることが分かる．一方，シスエレメント配列は生物種毎にクラスタリングされ（図 7，右），おおよそ種の系統関係を反映していることが分かる．

転写因子のクラスタリング結果とシスエレメント配列のクラスタリング結果のトポロジーの相関からは，homeo ドメインの機能獲得の過程とシスエレメント配列の形成が完全に対応していないことが示唆された．すなわち，転写因子の結合部位のアミノ酸配列が種を超えて，タンパク質の機能を保存するように変異しているのに対し，シスエレメント配列は種特有の規則性を持った塩基を保存するように種内で留まる変異をしていると予想される．このことから，転写因子のアミノ酸配列の進化と比較して，シスエレメント配列の変異の許容範囲が大きいと考えられる．また，二つの生物種において，ある転写因子の遺伝子がオーソログの関係にあったとしても，その転写因子の発現制御を受ける遺伝子群には必ずオーソログの関係にあるとは限らない．すなわち，転写因子間と遺伝子間にオーソログの関係がない場合には，転写因子とシス

Vol.2011-MPS-86 No.10
Vol.2011-BIO-27 No.10
2011/12/1

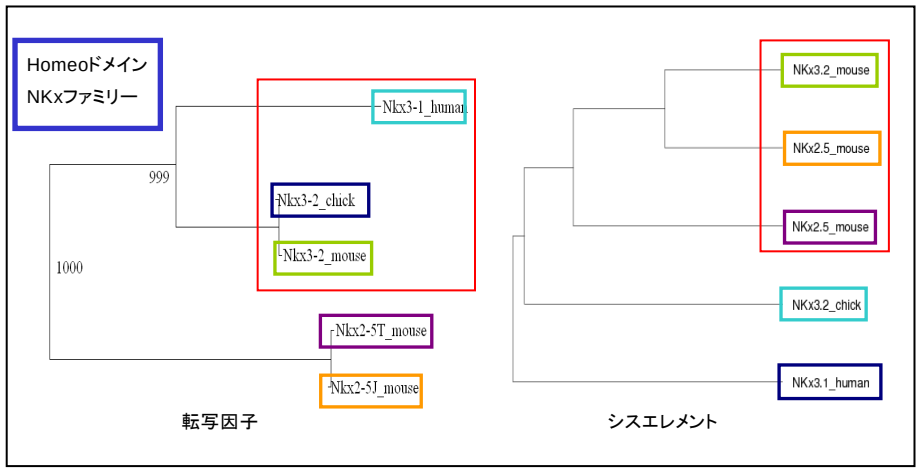

図 7　転写因子とシスエレメント配列のクラスタリング結果

エレメント間に共進化の関係がある可能性は低いと考えられる．以上のことから，転写因子とシスエレメント配列の共進化を解析するためには，両者間の対応に加えて，制御される側の遺伝子群との対応も明確にする必要があると考えられる．

そこで，本研究では今後の課題として，実験で考慮したドメインをもつ転写因子とその転写因子が発現制御する遺伝子群の組みを総合的に分類することにより，転写因子とシスエレメント配列間の共進化の有無を検討する予定である．

参考文献

[1] 山崎雄也，宮崎智，情報科学的手法によるシスエレメント配列の規則性探索, The 20th the Special Interest Group on Bioinformatics (SIG BIO2009)
[2] K. Birnbaum, P. N. Benfey, and D.E. Shasha, cis element/transcription factor analysis (cis/TF): a method for discovering transcription, Genome Res, **11(9)**, 1567-73, 2001.
[3] D. A. Papatsenko, V. J. Makeev, A..P. Lifanov, M. Régnier, A. G. Nazina, and C. Desplan, Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers, *Genome Res*, **12(3)**, 470-81 2002
[4] L. Ilie, and S. Ilie, Multiple spaced seeds for homology search, *Bioinformatics*, **23**, 2969-2977, 2007
[5] J. D. Thompson, T. J. Gibson1, F Plewniak, F Jeanmougin and D. G. Higgins, The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Research*, **25(24)**, 4876–4882, 1997
[6] A. Sandelin, W. Alkema, P. Engström, W. Wasserman, and B. Lenhard, JASPAR: an open access database for eukaryotic transcriptional factor binding profile, *Nucleric Acids Research*, **32**, D91-D94, 2004
[7] E. Wingender, P. Dietze, H. Karas, and R. Knüppel, TRANSFAC: a database on transcription factors and their binding sites, *Nucleric Acids Research*, **24(1)**, 238-241,1995
[8] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, **27**, 379-423 and 623-656, 1948
[9] UniProt: http://www.uniprot.org/
[10] P. Stegmaier, K. E. Alexander, and E. Wingender, Systematic DNA-Binding Domain Classification of Transcription Factors, *Genome informatics*s, **15 (2)**, 276-86, 2004
[11] S. Miyazaki, H. Sugawara, and M. Ohya, The efficiency of entropy evolution rate for construction of phylogenetic trees, *GENES & GENETIC SYSTEMS*, **71 (5)**, 323-327, 1996