2007－BIO－9 (4)
2007／6／14

相互情報量の応用による未知シスエレメント配列の予測

西浜 睦子，松林 航， 宮崎 智
東京理科大学大学院 薬学研究科 生命情報科学研究室

遺伝子の発現は転写因子がゲノム中のシスエレメントと呼ばれる配列に特異的に結合することによって制御されているため，遺伝子制御を理解するためには，ゲノム中の転写因子結合部位を特定することが重要となる．しかし，転写因子は様々なパターンのシスエレメント配列に結合するため，これまでの生化学的手法のみでは，全ゲノム中における転写因子結合部位を明らかにすることは困難な状況にある．そこで本研究では，実験的に実証された転写因子結合部位のデータベース「JASPAR」に登録されている１０７件の多細胞生物のシスエレメントデータを取得し，シャノンの情報量の概念を応用することでシスエレメントを網羅的に解析すると共に，ゲノム中の転写因子結合部位の配列パターンを予測することを試みた．

Prediction of unknown *cis*-element sequences by applying mutual information

Mutsuko Nishihama, Wataru Matsubayashi, Satoru Miyazaki
Department of Pharmaceutical research, Tokyo University of Science

Since the gene expression is regulated by binding transcription factors to the sequence called *cis*-element specifically in genome, to understand gene control, it is important to identify binding sites of transcription factors. However, transcription factors bind to many pattern of *cis*-regulatory sequences. Thus we have a hard time of identifying all binding sites of transcription factors on whole genome only by conventional biochemical technique. Therefore in this study, we get 107 *cis*-element sequence data of multicellular organism form the JASPAR database compiled experimentally verified transcription factor binding sites, and by applying the concept of Shannon's information measurement, we tried to analyze *cis*-elements cyclopaedically and to predict sequence patterns of transcription factors binding sites on genome.

1. はじめに

生命の設計図であるといわれているDNAは，主にアデニン (A)，チミン (T)，グアニン (G)，シトシン (C) の 4 種の塩基で構成される塩基配列である．DNAにより暗号化されている遺伝情報の最も重要な役割は，生物が生き，形作るために必要な様々なタンパク質合成を指令することである[1]．細胞はDNAの担う遺伝情報をRNAへと転写し，さらにそれを鋳型として翻訳することによりタンパク質合成を行う．遺伝子の発現は，転写因子と呼ばれるタンパク質が，遺伝子の上流又は下流に存在する，「シスエレメント(cis-element)」と呼ばれる特徴的な配列に結合することで，適時適所で，適切な遺伝情報が，適量発現されるように制御されている．したがって，遺伝子の発現を制御する転写因子がゲノム上のどこに結合し，どのように制御しているのかを明らかにすることは生命現象の謎を解く一助となると考えられる．

シスエレメントは 4 ～22 塩基程度の非常に短い塩基配列であるが，多様な配列である．転写の機構を解

明するには，この転写因子とシスエレメントの結合の規則性を明らかにすることが重要であるが，どの転写因子がどのシスエレメントと結合するという明確な規則性はまだ解明されていない．また，配列や機能が分かっている遺伝子であっても，その転写制御に関わるシスエレメントは分かっていないことが多く，ヒトで約22,000存在するといわれている遺伝子の，全てのシスエレメントを生化学的実験手法のみで明らかにすることは困難であると考えられる．そこで本研究では，シスエレメント配列の数量化とコンピュータを駆使した解析法を研究すると共に，未知のシスエレメント配列パターンを予測する手法を考案することを試みた．

2. シスエレメント配列の網羅的解析

シスエレメント配列の多様性を特徴付けるために，これまで，各々の転写因子のシスエレメント配列パターンはプロファイル形式や正規表現などで表されて来た．このような方法では，個々のシスエレメント配列パターンの特徴を見ることは可能だが，シスエレメント群間での特徴を見ることは出来ない．

そこで本研究では，シスエレメントを塩基単位で直接比較する配列類似度を基盤とする手法とは異なる，Shannon の情報量の概念を利用した新しいシスエレメント配列解析方法を考案した．この手法では，シスエレメントを構成するA，T，G，Cの4つの塩基のそれぞれの出現確率をもとにシスエレメントが持つ相互情報量[2]を正規化した，Entropy Evolutional Rate (EER)[3]を計算することによって，文字列であるシスエレメントを数量化する．

本研究では，情報量としてシャノンエントロピー[3]を用いた．シャノンエントロピーはある出来事（事象）が起こった際，その出来事がどれほど起こりにくいか（どれほど乱雑であるか）を表す尺度であり，事象系の乱雑さを表すことができる．相互情報量とは2つの事象において，一方の情報を得たときに，他方が何であるかについてどれくらいの情報が得られるかを表す量であると定義されている．

私たちは，各々のシスエレメント配列間の類似度をEER値として計算し，シスエレメント群間を比較した．現在のところ，どの転写因子がどのシスエレメント配列に結合するという，転写因子とシスエレメント配列の対応における明確な規則性はまだ解明されていない．そこで私たちは，転写因子のシスエレメントへの結合に関する規則性を探るために，シスエレメント配列の網羅的な解析を行った．

2.1 データの取得

データは，多細胞生物の転写因子結合部位データベースである The JASPAR database[4] から取得した．JASPAR では，実験的に実証された転写因子結合部位の配列が 123 種の転写因子ごとに，シスエレメント配列モチーフの保存状態を視覚的に表現した LOGO 形式と各位置における塩基出現頻度をわかるようにしたプロファイル形式[5]によって公開されている．そのうち，107 種の転写因子のデータが，プロファイルを作成するために用いられた実際の転写因子結合部位の配列を取得可能だった．そこで，ある1つの転写因子が結合するシスエレメント配列パターンのデータを1レコードとして，転写因子ごとにその結合するシスエレメント配列パターンを 107 レコードにまとめた．各々の転写因子のプロファイルを構成しているシスエレメント配列パターンは，6 配列～116 配列あり，最短で4塩基長，最長で22塩基長であった．また，これらのデータを構成する生物種は，*Antirrhinum majus, Arabidopsis thaliana, Drosophila melanogaster, Gallus gallus, Halocynthia roretzi, Homo sapiens, Hordeum vulgare, Mus musclulus, Oryctolagus cuniclus, Petunia hybrida, Pisum sativum, Rattus norvegicus, Xenopus laevis, Zea mays* の１４種である．

2.2 シャノンエントロピーの計算

各転写因子が結合する各々のシスエレメント配列について，シャノンエントロピーを計算した．任意の配列におけるシャノンエントロピー (S) を以下の式により与える．

$$S = -\sum_{i=A,T,G,C} Pi \log_2 Pi \quad \cdots (1)$$

このとき，Pi はシャノンエントロピーを計算しようとするシスエレメント配列中におけるA，T，G，Cの出現確率である．

シャノンエントロピーは乱雑さを表す尺度であるため，この値を計算することによって，その配列中における塩基の出現の偏りを知ることが出来る．シャノンエントロピーをシスエレメント配列に適用する場合，塩基は4種類であるため，シャノンエントロピーは0≦S≦2の値をとる．エントロピーの値が0に近ければ近いほど，その配列中における塩基の出現は大きく偏っていることを意味し，2に近ければ近いほど，その配列中では4つの塩基が均等に出現していることを意味する．

2.3 相互情報量の計算

次に，１０７レコードの各レコード内で，考えられる全ての2パターン配列間において相互情報量を計算した．任意の2配列X，Y間における相互情報量（I）を以下の式により与える．

$$I(X;Y) = \sum_{\substack{i=A,T,G,C \\ j=A,T,G,C}} P_{ij} \log_2 \left(\frac{Pij}{PiPj} \right) \cdots (2)$$

相互情報量は2つの情報源（X，Y）間の関連性の度合いを示すものであり，この場合はシスエレメント配列XとYにおける塩基の出現になんらかの従属関係があるのか，ないのかということを示す値になる．シスエレメント配列XとYの塩基の出現に全く関連がない場合，相互情報量は 0 になる．また，シスエレメント配列Xの塩基が決まれば，シスエレメント配列Yの塩基が完全に決まるという従属関係がある場合，その 2 配列間の相互情報量は最大値である2をとる．

2.4 Entropy Evolutional Rate (EER) の計算

相互情報量の大きさは，シャノンエントロピーの大きさに依存するため，解析する際に全てのシスエレメント群を等しく扱うことができない．そこで，相互情報量を正規化した値であるEER[9]を利用した．このような正規化した値を利用することで，シャノンエントロピーの大きさの違いに左右されない解析が可能となる．EER は（1）式と（2）式を用いた，以下の式により与える．

$$EER(X;Y) = \frac{1}{2}\left(\frac{I(X;Y)}{S(X)} + \frac{I(X;Y)}{S(Y)} \right) \cdots (3)$$

このときEERは，0≦EER≦1の値をとる．

比べる 2 配列間の EER 値が0に近ければ近いほど，シスエレメント配列XとYにおける塩基の出現には関係性がないことを意味し，EER が1に近いほど，シスエレメント配列XとYの塩基の出現には従属関係が存在することを意味する．私たちは全てのレコードについて，各レコード内で考えられる全ての 2 配列間における EER 値を計算した．よって，1 つのレコードから EER 値は ${}_m C_2$ 個（m は1レコード中のシスエレメント数）得られる．

2.5 頻度分布の作成

各々の転写因子が結合するシスエレメント配列パターン（各レコード）を網羅的に比較するために，各々のシスエレメント配列パターンから得られた EER 値を 0.1 の階級幅で頻度分布化した．各レコードによって得られる EER 値の個数は異なるため，縦軸はその階級に入る EER 値の個数を ${}_m C_2$ で割った相対値を示すようにした．例として転写因子ＡＧＬ３が結合するシスエレメント配列パターンのEER値頻度分布を図1に示す．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グラフは左寄りになる．

2.6 シスエレメントの階層的クラスタリング

作成した頻度分布の類似性をもとにユークリッド距

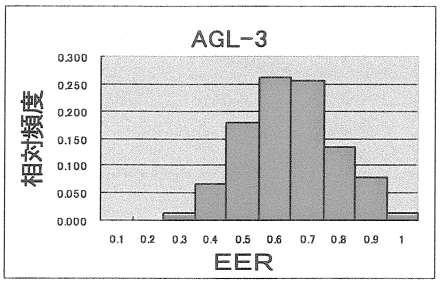

図 1. 転写因子 AGL3 の EER 頻度分布

離・ウォード法を用いて階層的クラスタリングを行った. 階層的クラスタリングを行うことによって, シスエレメント配列パターンの揺るぎの度合いが似ている転写因子を知ることが出来る. 私たちは各頻度分布の形状を, 頻度分布の各階級におけるＥＥＲ相対値 10 ポイントと隣接する階級間の傾き 9 ポイントの 19 次元ベクトルによって表すことにした. 比較する要素に頻度分布の階級間の傾きを加えることで, 頻度分布の形状がより類似しているものをクラスタリングすることが出来る[6].

階層的クラスタリングは, まず 1 つのクラスタに 1 つのレコードから得た頻度分布を含む, 107 個のクラスタがある初期状態から始める. この 107 個のクラスタについて, 19 の変数からクラスタ間のユークリッド距離行列を計算し, ユークリッド距離行列の中から最も類似性が高い 2 つのクラスタを合併して, 1 つのクラスタを作る. そして, 次々とクラスタを結合し, 最終的に 1 つのクラスタに合併されるまで繰り返すことで階層構造を作成する.

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る.

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

このとき, i は各階級における EER 相対値 10 ポイントと隣接する階級間の傾き 9 ポイントである. 従って, n ＝19 となる.

2.7 結果と考察

転写因子とシスエレメントの結合には, 転写因子のＤＮＡ結合ドメイン構造が関与するため[7], 私たちはクラスタリングを行った結果近いクラスタに入った頻度分布同士, つまりシスエレメント配列パターンの揺るぎの度合いが似ている転写因子同士ではＤＮＡ結合ドメイン構造も類似するのではないかと考えた. そこで, 得られたクラスタリングの結果を転写因子のＤＮＡ結合ドメイン構造に注目して評価した.

全 107 レコードを用いてクラスタリングを行った結果は図 2 に示す. 得られたデンドログラムを見ると, ZN-FINGER_GATA や ETS ドメインなど一部のドメインに関しては近隣にクラスタリングされているが, 多くの場合, 近隣にクラスタリングされることとＤＮＡ結合ドメイン構造との間に特には関係がないように思われた.

そこで, 生物種ごとに分けて再度クラスタリングを行い, その結果について考察した. 生物種ごとのクラスタリングでは, 近隣にクラスタリングされているものはＤＮＡ結合ドメイン構造が類似しているという例が, 全体でクラスタリングを行った時よりも多く見られた. 例として *Homo sapiens* のクラスタリングの結果を図 3 に示す.

次に生物種ごとでクラスタリングをした方が, 近隣にクラスタリングされることとＤＮＡ結合ドメイン構造の関係性を見やすくなる理由を探るために, ＤＮＡ結合ドメイン構造別にクラスタリングを行った. 例として MADS ドメインの結果を図 4 に示す. この結果は, あるＤＮＡ結合ドメイン構造の種類に着目したとき, そのＤＮＡ結合ドメイン構造で結合出来るシスエレメント配列の揺らぎの度合いは生物種によって異なることを示している. そのため, 生物種を区別せずに全データで行ったクラスタリングでは, 近隣のクラスタとＤＮＡ結合ドメイン構造の関係性がよく見えなかったと考えられた.

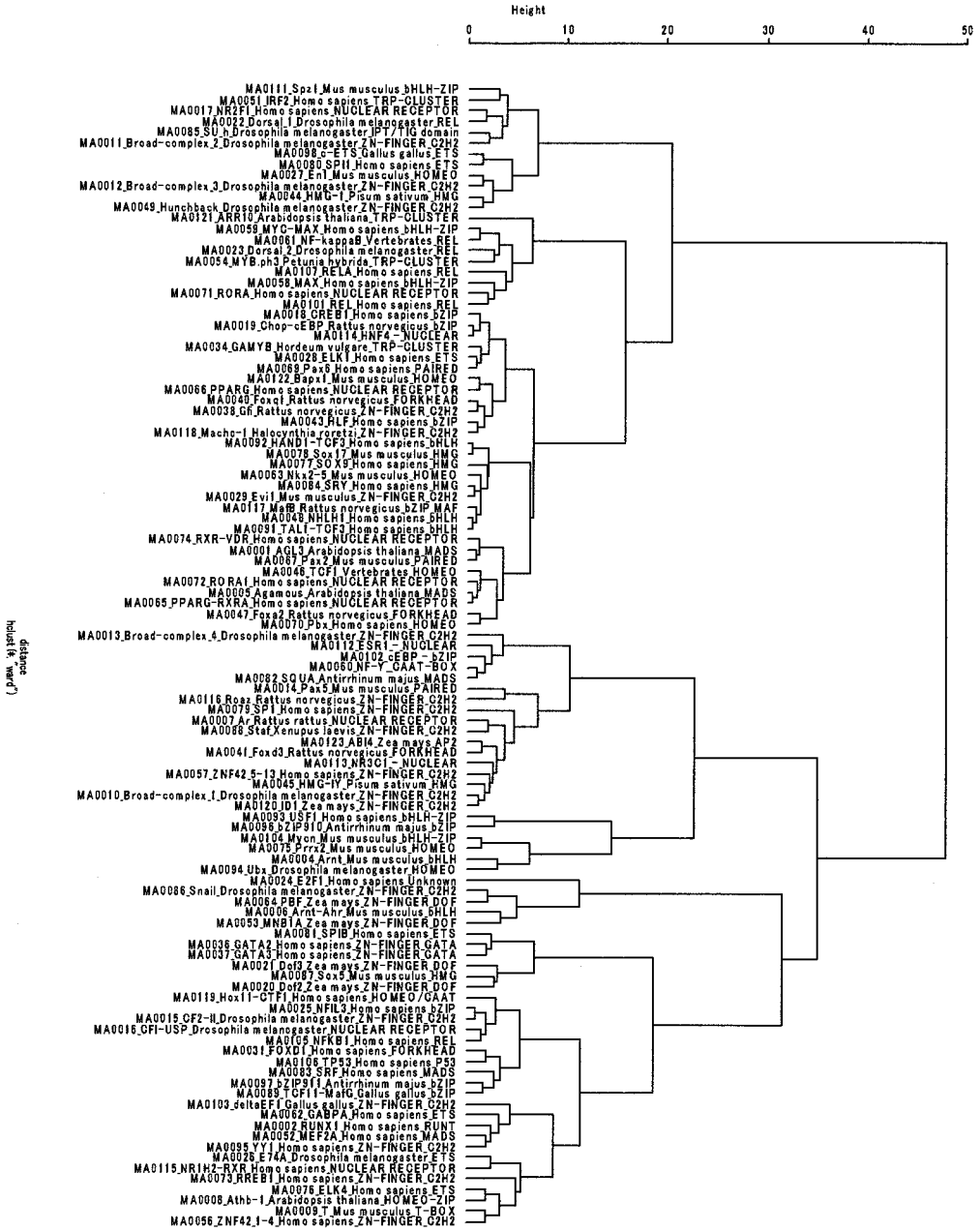

図 2. 全データを用いたクラスタリング結果. ID 番号__転写因子名__生物種__DNA 結合ドメインを明記している.

3.未知シスエレメント配列の予測手法の開発

配列や機能が分かっている遺伝子であってもその転写制御に関わるシスエレメントは分かっていないことが多く, 制御に関わる全てのシスエレメントを生化学的実験手法のみで明らかにすることは困難であると考えられる. そこで私たちは, ある転写因子のシスエレメント配列パターンを数量化した際に得られた EER 値が

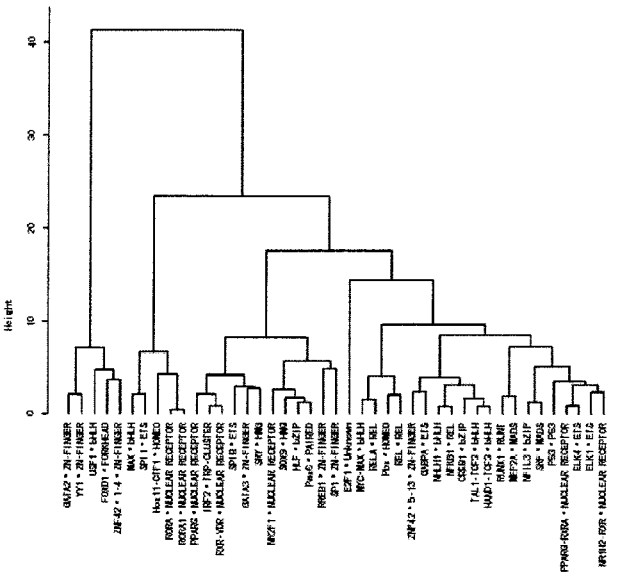

図 3. Homo Sapiens データのみのクラスタリング.

転写因子名__DNA 結合ドメインを明記している.

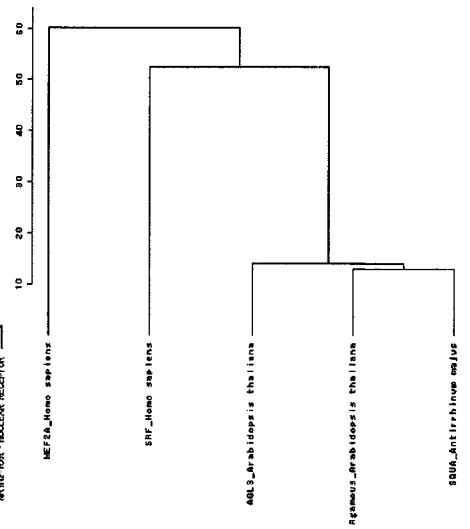

図 4. MADS ドメインのクラスタリング.

転写因子名__生物種名を明記している.

その転写因子が結合し得るシスエレメント配列の揺らぎの度合いを示すと考え, 未知のシスエレメント配列を予測することは出来ないかと考えた. そして, EER を利用して予測に必要な閾値を設定し, ある転写因子の1つの既知シスエレメント配列から, その他の結合し得る未知シスエレメント配列を予測する手法の開発を試みた.

3.1 データセットの作成

2 つのシスエレメント配列の網羅的解析の結果, 生物種によって転写因子が結合するシスエレメント配列の揺らぎの度合いが異なると考えられたので, 2.1 で取得した 107 レコードのデータの中から, 今回は生物種が *Homo Sapiens* のレコードのみを抽出して用いることにした. *Homo Sapiens* のレコードを抽出した結果, 43 レコードが得られた. そして, 得られた 43 レコードを, 任意の 38 レコードが含まれる学習用データセットと 5 レコードが含まれる検証用データセットに分けた.

3.2 閾値の設定

EER から予測に必要な閾値を求める. 学習用データセットの各レコード内において, 考えられる全ての 2 シスエレメント配列間の EER 値を計算した. この各レコードから得られた EER 値を平均した値を, 今回私たちは閾値として設定した. この閾値は, 各転写因子が結合するシスエレメント配列パターンが平均してどの程度揺らぎを持っているのかということを意味する. そのため, ある転写因子が結合する基準となるシスエレメント配列 (基準配列) と同配列長のランダムな配列 (ランダム配列) を用意し, その 2 配列間の EER を計算するとき, 基準配列とランダム配列の間の塩基の揺らぎが閾値よりも少なければ (つまり EER 値が閾値よりも高ければ), その転写因子は比較したランダム配列にも結合し得ると解釈される.

3.3 シスエレメント配列の予測

次に, シスエレメント配列の予測を実行するために, 検証用データセットの 1 レコードから任意の 1 つのシスエレメント配列 (基準配列) を抽出し, 抽出した配列と同じ配列長のランダムな配列を用意した. そして, 基準配列と各々のランダム配列との 2 配列間における EER 値を算出し, 設定した閾値以上になるランダム配

列をその転写因子が結合し得る配列として，予測リストに書き出した．この予測手法にしたがって，検証用データセット中の 5 レコード全てに対して，それぞれシスエレメント配列の予測を行った．

3.4 結果と考察

今回，予測を実行した結果，各検証用データレコード中のシスエレメント配列が，多数予測出来ていた．しかし，検証用データレコード中には存在するが，きちんと予測出来なかったシスエレメント配列も多数見られた．また，予測リストの中には，検証用データレコード中に存在しないシスエレメント配列も多数存在していた．これらは新規のシスエレメント配列とも考えることが出来るが，まだうまく予測配列の絞り込みが行われていないように思われた．

その原因として以下のようなことが考えられた．EER は相互情報量を正規化した値であり，この相互情報量は配列中における塩基の出現確率と 2 配列間における塩基の一致度合いを考慮したもので，塩基の出現位置は考慮していない．相互情報量を基盤とした EER は，個々のサイトを独立に扱って計算しているが，それぞれの転写因子ごとのシスエレメント群を特徴付けるためには，サイト間の関係を考慮する必要があるかもしれない．もうひとつの可能性は，シスエレメント群内のばらつきが一定でも，それに含まれる配列パターンの詳細をみると，あるばらつきを指標とした際に考えられるすべての配列パターンが含まれているわけではないことである．ここで提案した EER は，シスエレメント群内での配列の多様度の上限あるいは下限を示すことができるが，その範囲以内に含まれる配列パターンの「質」を規定する能力を有していない．これを補うためには，たとえば，2 配列間の配列類似度を再考した判別量を定義できる可能性がある．例えば，‘ATGCATGC’と‘ATGCATGC’，‘TACGTACG’，‘ATGCTTGC’の 3 パターンの配列比較を検討してみる．(ATGCATGC, ATGCATGC) の 2 配列間の EER を 2.4 で示した（3）式を用いて計算すると EER＝1，(ATGCATGC, TACGTACG) の場合も EER＝1，(ATGCATGC, ATGCTTGC) の場合は EER＝0.848 となる．つまり，配列‘ATGCATGC’とそれぞれの配列の EER は，

‘ATGCATGC’ ＝ ‘TACGTACG’ ＞ ‘ATGCTTGC’

となる．しかし，配列‘ATGCATGC’との類似性は，

‘ATGCATGC’ ＞ ‘ATGCTTGC’ ≫ ‘TACGTACG’

である．ATGCATGC を基点にしてシスエレメントの配列予測をした際には，配列類似性が高い ATGCTTGC を選択できるような工夫が必要であろう．

4. おわりに

本研究では，文字列であるシスエレメントを Shannon の情報量に基づいた EER を利用して計算し，数量化することで解析を行った．あるシスエレメント配列パターンの EER は，そのシスエレメント配列パターンにおける揺らぎの度合いと考えることができる．私たちは，転写因子はある規則性を持ったシスエレメントに結合すると考え，シスエレメント配列パターンの EER を詳細に解析することで，その規則性を探った．シスエレメント配列パターンの EER を頻度分布化し，その頻度分布の形状を比較すると，頻度分布の形状と転写因子の持つ DNA 結合ドメイン構造との間の関連性はあまり見られなかった．しかし，生物種別に比較すると，頻度分布の形状と転写因子の DNA 結合ドメイン構造に関連が見られるドメイン構造が，全体で比較した場合よりも多く存在した．このことは，生物種によって，転写因子のある DNA 結合ドメインが認識するシスエレメント配列の揺らぎの度合いは異なることを示していると考えられた．

また，私たちは EER が配列間の揺らぎの度合いを示すと考え，EER から閾値を設定することで，シスエレメント配列において許される揺るぎの範囲を設定し，未知のシスエレメント配列パターンの予測を試みた．しかし，予測を実行した結果，EER という考え方のみではうまく予測配列を絞り込むことは出来なかった．その原因として，EER は 2 配列中における塩基の出現確率を考慮したものであり，必ずしも配列類似性を反映して

いるわけではないということが考えられた.

今後は, EER を利用したシスエレメント配列の解析と予測手法をさらに改良したいと私たちは考えている. その第一歩として, 比較する2配列間で同一位置における塩基が一致しているときが多い場合は, 2 配列間における相互情報量が高くなるように, 相互情報量を補正したいと考えている. それは, ゲノム配列においては機能上重要となる塩基は保存される傾向にあるため, シスエレメント配列パターン内の2配列を比較する際にも, 2 配列間で一致している塩基と一致していない塩基では, 情報の重みが異なると考えられるためである. また, 閾値の設定方法についても再度検討したいと考えている.

謝辞

本研究を遂行するにあたり, 熱心にご指導くださいました, 東京理科大学薬学部　生命情報科学研究室の宮崎　智教授に深く感謝致します.
また, 研究の相談に乗って下さった, 東京理科大学薬学部　生命情報科学研究室の皆様に心より感謝致します.

文献

1) 宮野悟,江口至洋,金久實,高木利久,中井謙太 (2006) 「バイオインフォマティクス事典」 共立出版株式会社
2) Shannon C.E. (1948) The Bell System Technical Journal, Vol27,pp.279-423, 623-656.
3) Ohya M. Trans. IEICE. Vol. E72 No.5 pp 556-560
4) Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B(2004) Nucleic Acids Res. 1;32(Database issue):D91-4
5) Wasserman WW, Sandelin A.(2004) NatRev Genet;276-87
6) Shoudan Liang, Stefanie Fuhrman, Roland Somogyi, Receal,(1988) Pacific Symposium on Biocomputing, 3:18-29
7) Mandel-Gutfreund Y Margalit H ; (1988) Nucleic Acids Res.26:2306-2312