Vol.2009-BIO-18 No.11
2009/9/18

生物種間の配列パターンの変化に着目した情報科学的手法によるシスエレメント配列解析

山﨑雄也† 鈴木智典†† 宮崎智†

†東京理科大学大学院 薬学研究科
††東京理科大学 薬学部

遺伝子の転写調節を果たす転写因子と，そのターゲットとなるシスエレメント配列との関係には未だ不明な点が多い．そこで本研究では，情報科学的手法を用いて *in silico* によるシスエレメント配列構造の解析を行い，転写因子のシスエレメント配列 認識における規則性を見出すことを目的とする．規則性探索の足がかりとしてシスエレメント配列の生物種の違いによる配列パターンの柔軟性の変化に着目して解析を行った．

Analysis of *cis*-element regularity by informatics approach considering variety of sequence among species

Yuya Yamazaki† Tomonori Suzuki†† Satoru Miyazaki†

†Graduate School of Pharmaceutical Science, Tokyo University of Science
††Faculty of pharmaceutical sciences Tokyo University of Science

Transcription is regulated by transcription factors (TFs) which binding to the specific nucleotide sequence called *cis*-regulatory sequences (*cis*-elements). The binding regularity of *cis*-elements and TFs has not yet been defined. In this study, we performed *in silico* analysis by an informatics approach to reveal the *cis*-element recognition regularity of TF. As the foothold of the analysis, here we especially focus on the variety of *cis*-element pattern depending on difference of species.

1. はじめに

生命の根幹には，アデニン (A), チミン (T), グアニン (G), シトシン (C) の 4 種の塩基を持つ DNA から構成されるゲノム配列が存在し，ゲノム配列上に存在する遺伝情報が適宜発現することで生命は形作られ，生命活動が維持されている．遺伝情報は DNA から mRNA へ（転写），さらにタンパク質へ（翻訳）伝達されることで実際に生命の中で発現する．

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．

転写因子が結合するシスエレメント配列は数から数十塩基の短い配列であるが，転写因子ごとにただ一つのパターンだけでなく，多様なパターンが存在し，転写因子とシスエレメントの結合における規則性には未だ不明な点が多い．この規則性を見出すことは，転写制御機構の解明や未知シスエレメント配列を探索する上で重要であると考えられる．

一般的に配列解析においては，自分が調べたい配列を，既知の配列群に問い合わせ，類似，近縁関係と予想される配列を見付出し，それらの配列の情報を元に未知の配列の機能や関係性を推測していくという相同性検索の手法が良く用いられている．

しかしながら，シスエレメント配列でとても短い配列であるがゆえに，相同性検索で比較するのは困難である．

転写因子のシスエレメント配列への結合に関する，明確な規則性が不明であり，その規則性を解明するにあたって既存の配列解析手法では達成が困難な現状から，本研究では，既存の方法に依らない，情報量を応用した解析手法を試み，転写制御に関わるシスエレメント配列の規則性を見出ことを目的とする．

2. 解析用配列レコードデータセットの作成

2.1 データの取得

解析対象となるシスエレメント配列のデータを転写因子データベースである JASPAR と TRANSFAC から取得し，そのデータを転写因子ごとに結合するシスエレメント配列群のレコードとしてまとめた．

2.2 データの整理

TRANSFAC のデータについては 1 つの転写因子につき複数の生物種が含まれてい

†東京理科大学大学院 薬学研究科
Graduate School of Pharmaceutical Science, Tokyo University of Science
††東京理科大学 薬学部
Faculty of pharmaceutical sciences Tokyo University of Science

たり,データに含まれている配列の長さが違ったり,ギャップが含まれる場合がある．そこで，データの除外や，配列の長さの揃えなどを行い，本解析に適するように処理した．また，データ内に重複データが含まれている場合は重複データを削除したものを 1 レコードとした．

2.1 で取得し，2.2 で一部データを除いた結果，322 レコードを得た．

3. データベース登録配列における問題点の解決

データベースへ登録されているデータは，主に予め配列を用意し，それらの配列に転写因子が結合するかどうかを確認する実験を元にして，転写因子と結合するとわかった配列のデータを元に，登録が行われている．しかし，それらの中には，実際に配列のどこに転写因子が結合しているかを確認していないものが多いため，実際には結合に関与していない部分が含まれている可能性が考えられ，実際に転写因子と関与している部分を比較するうえで，それらの部分が障害となる可能性が出てきた．

実際にレコードの中のシスエレメント配列群を各縦の列で区切り，その列の中で塩基のバラツキを見てみると，極端に塩基出現のバラツキが大きく，転写因子との結合への寄与が疑わしい部分が存在していることが多く確認された．そこで，これらの各列の塩基のバラツキを，シャノンエントロピーという概念によって数値として表し比較することにした．

3.1 シャノンエントロピーの定義

シャノンエントロピーは事象の乱雑さを表す尺度であり，本研究では配列パターン内における，4 種の塩基の乱雑さ，偏り具合を表す値となる．ある 1 列の配列から求められるシャノンエントロピーは以下の式で定義される．

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

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が，0 に近いほど塩基の出現が大きく偏っていることを意味し，値が 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．

3.2 配列パターンの各サイトにおける塩基の乱雑さの分布

塩基が極端にバラバラな部分の分布をみるために，以下の手順で配列パターンの各サイト（配列パターンを縦に区切った列）における塩基の乱雑さの分布を求めた．

i. 全レコードを縦に 1 列ずつ区切り，全ての列からエントロピーを計算．

ii. 全レコード内の列をグループ分けし，配列パターンの両端 3 列を，端から順に 1 列ずつ分けた①，②，③の 3 グループと，それ以外の内側の列の④の 4 つの列グループに分ける．

iii. 高いエントロピーの値の範囲を 4 種類（1.6～1.7,1.7～1.8,1.8～1.9,1.9～2.0 ）設定し，範囲内の値を持つ列が，4 列グループの内，どのグループに属するかを数える．

iv. 計数した値を元に，乱雑な列が配列パターンのどの位置に存在するかの度数分布を作成する．

i ～ iv の手順により求めた分布は以下の図である．

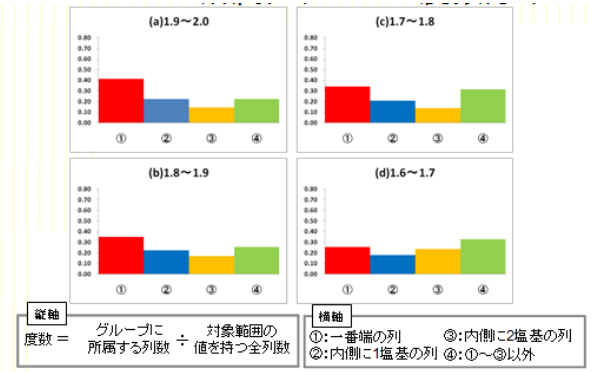

図 1. 配列パターン内の位置における塩基出現の乱雑さの分布

1.7～2.0 の間のエントロピーの範囲における(a),(b),(c)らの 3 つの度数分布ではより左側のグループ，すなわち配列のより端に近い位置ほど度数が大きく，エントロピーの高い列は配列レコードの端に近い方に分布する傾向があるということがわかった．

3.3 配列パターン端に存在する極乱雑部位の削除

解析の障害となるであろう，エントロピーの極端に高い site の削除を行った．

配列パターンの中で，エントロピーの高い site が他のパターンと関連がなく，たまたま 4 種の塩基が均等の確率で出現している結果エントロピーが高くなっているならば，その site の塩基の出現は他の配列パターンの塩基の出現と関わりがないと予想される．よってエントロピーが高い site の塩基の種類を元に，各塩基が所属する配列レコードを 4 つのグループに分けたとき，各グループが持つエントロピーには偏りがないと予想した．

Vol.2009-BIO-18 No.11
2009/9/18

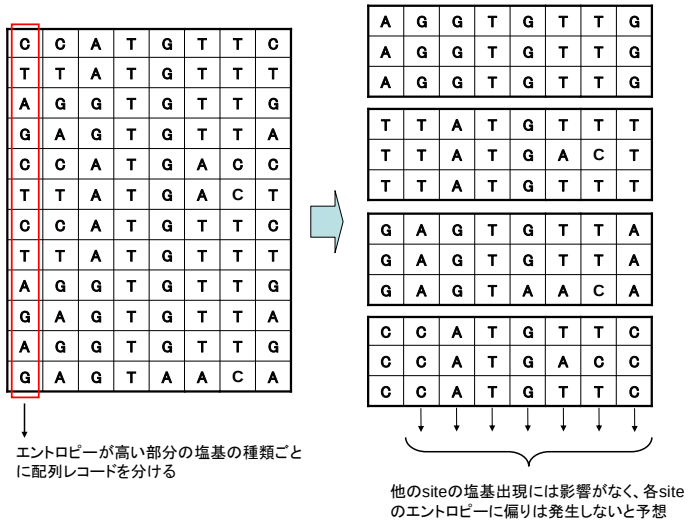

図 2　例：極乱雑 site の塩基の種類によるグループ分け

ここでは 3.1 で求めた配列レコードの各 site のシャノンエントロピーを利用する．

高エントロピーの閾値 Si を設定する．（今回は 1.9,1.8,1.7,1.6 の 4 通りについて試した）配列レコードデータセットにおける各レコードについて以下の操作を行う．

i. 配列レコードの一番外側の site について，Si 以上のエントロピーを持つものを調べる．次に，見つかった site の 1 つ内側の site について，Si 以上のエントロピーを持つかを調べる．これを内側へ 1site ずつずらしながら繰り返し，一致値以上のエントロピーを持ち，配列レコードの端から連続する削除候補 site の位置を抽出する．

ii. 削除候補 site が見つかった配列レコードについて，抽出した site 位置における塩基の種類(A,T,G,C)を元にして配列レコードを 4 つのグループに分ける．1 つのレコードに複数の削除候補 site がある場合は見つかった全ての site に関して同様のグループ分けを行う．

iii. 4 つに分けたグループについて，3.1 と同様に各 site のエントロピーを求め，各 site のエントロピーの合計値を求める．

iv. 4 つのグループにおけるエントロピー合計値についての標準偏差を求める．

v. ｉ～ⅳを全ての配列レコードについて繰り返して得られた標準偏差の平均値をとる．

vi. 各レコードの削除候補 site における標準偏差と，v で求めた標準偏差の平均値を比べ，平均値以上の値を持っていた削除候補サイトを削除 site に決定する．

vii. 各配列レコードについて削除 site を持つものはその site を取り除き，新しい配列レコードとする．

以上の操作によって配列レコードの修正を行った．今回は(Si=1.9,1.8,1.7,1.6)の 4 通りについて試したので，オリジナルの配列データセットと合わせて 5 つの配列レコードセットが存在する．

4. entropy evolutional rate を指標としたシスエレメント配列の解析

取り除くべき列を削除し，配列パターン洗練した後，レコードとしてまとめたシスエレメント配列パターンの情報を，「entropy evolutional rate」，略して EER という値によって数値化して比較した．

EER は 3.1 で定義されるシャノンエントロピーと，そのエントロピーの 2 つの事象での関連の度合いを表す値である相互情報量を組み合わせた値である．まず相互情報量と EER の定義について詳しく述べる．

4.1 相互情報量の定義

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

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,今回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の有無を示す値になる．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．

4.2 EER の定義

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[2] 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．

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

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．

4.3 解析対象の選択

データベースに登録されている転写因子には, DNA と結合することができる特徴的な部位である，DNA 結合ドメインという分類情報が登録されている．今回，その DNA 結合ドメインの一つである「zinc finger C2H2 ドメイン」を手始めの解析対象として選択し，そのドメインを持つ転写因子に関するレコードを選出し EER を指標として解析を行った．解析対象のレコード数は全部で 37 レコードである．

4.4 EER による配列レコードの数量化

4.4.1 配列内の同じ site 位置における塩基の出現に着目した EER の算出

解析対象の全配列レコードについて，レコード内の配列群から選び出せる全ての 2 配列間における EER 値を計算した．

この時エントロピーはレコード内の各配列における塩基の出現確率，相互情報量は任意の 2 配列間における共通の site での塩基の組合せの出現確率である．よって,1 つのレコードから EER 値は ${}_mC_2$ 個 (*m* は 1 レコード中のシスエレメント数)得られる．

4.5 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 ${}_mC_2$ で割った相対値を示すようにした．

4.6 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．

階層的クラスタリングでは，まず 1 つのクラスタに 1 つの要素だけが存在する初期状態から始め，クラスタお互いの類似度を比べて最も類似性の高い 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．

今回，階層的クラスタリングで比べる類似度として，各レコードから求められた EER から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ることとした．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．

階層的クラスタリング手法には，クラスタ間の距離の取り方の違いなどによって色々な手法が存在するが，今回は代表的なクラスタリング手法である，（ウォード法，最近隣法，最遠隣法，重心法，平均法，メディアン法）ら 6 種類の方法を用いて行った．

5. 解析結果の評価・考察

5.1 配列修正による影響の評価

3.3 で行った配列の修正がどのような効果をもたらしたかをクラスタリング結果から検証することとした．

クラスタの分離度をはかる指標として，クラスタの収束度合いとクラスタ間の距離によって計る Dunn’s Validity Index (Dunn), Davies-Bouldin Validity Index (DB) ,の 2 種類の指標を用いる．

Dunn，DB は次の式で表される．

$$Dunn=\min\left\{\frac{d(c_i,c_j)}{\max\{d(c_k)\}}\right\}\qquad DB=\frac{1}{K}\sum_{i=1}^{K}\max\frac{d(c_i)+d(c_j)}{d(c_i,c_j)}$$

ここで $d(c_i,c_j)$はクラスタ c_i と c_j との距離（inter cluster distance）である．$d(c_k)$はクラスタ内における距離（intra cluster distance）である．

Dunn と DB はある数に分かれたクラスタから求められる値であるため，4.6 でのオ

リジナルの配列レコードを用いたクラスタリング，高エントロピーの閾値 1.9,1.8,1.7,1.6 を設定して配列を削除した 4 種レコード群に関して行ったクラスタリング結果それぞれについて，クラスタリング結果を 2~9 個のクラスタ数が出来る段階で区切り，各段階での Dunn と DB を求めた．各指標は統計ソフト R を用いて求めた．

各指標はクラスタ内距離 3 種("complete","average","centroid")，クラスタ間距離 6 種("single", "complete", "average","centroid", "aveToCent", "hausdor")から 1 つずつ選択し，求められることから，距離の選択は全部で 18 通りである．

各レコードのクラスタリング手法の選択が 6 通り，クラスタ形成数が 2~9 の 8 通り，指標の距離の取り方が 18 通りなので，1 つの配列レコード群からの算出 EER によってクラスタリングした結果，1 つの手法につき，864 通りの指標の値が求められる．

配列の修正なしのレコードから求めた指標と，配列レコードを修正した 4 種のレコードから求めた指標を共通するクラスタリング手法間で比べ，修正したほうの指標が高くなっているかどうかを調べるため，修正前後で求めた 864 データの平均値に差があるかどうかを，全て同じ条件で求められた指標を対応する変数とする，片側の対応のある t 検定を行った．

帰無仮説　「指標（DB or Dunn）について，元の配列レコードデータセットのクラスタリング結果について求めた値と，配列修正後（閾値 4 通り）のデータセットのクラスタリング結果から求めた値に差はない」

有意水準　0.95

結果，以下の表の通りになった．

修正閾値		1.6	1.7	1.8	1.9
clustering method	single	<0.001*	<0.001*	<0.001*	1
	complete	0.006*	<0.001*	1	1
	average	0,017*	<0.001*	1	0.991
	centroid	<0.001*	<0.001*	<0.001*	1
	median	0.02*	<0.001*	0.629	0.498
	ward	0.758	0.991	1	0.79

表 1　評価に DB を用いた場合における検定での p 値

修正閾値		1.6	1.7	1.8	1.9
clustering method	single	0.971	1	1	0.264
	complete	0.99	0.999	0.314	0.991
	average	0.004*	0.185	0.029*	<0.001*
	centroid	1	1	1	<0.001*
	median	<0.001*	<0.001*	<0.001*	<0.001*
	ward	0.0354*	<0.001*	<0.001*	1

表 2　評価に Dunn を用いた場合における検定での p 値

5.2　クラスタリング結果の有用性（生物種集合の評価）

EER を用いてシスエレメント配列パターンを数量化し，クラスタリングを行った結果，どのようなクラスタがあつまっているかを調べるために，各レコードの配列の由来となる生物種に着目し，同種の生物種のラベルを付けたレコードの集合を評価した．

5.2.1 クラスタ内での同種レコードの集合の評価

全体のレコード数が N の樹形図で，k 番目のクラスタの中に，同種ラベル X のレコードが n 個含まれる場合，X の集まりの評価を以下で与える．

$$\frac{\sqrt{2}^{n-1}}{N}$$

同一種のラベルが多いほどこの評価値は高くなる．

複数のラベルが含まれる場合，各ラベル Y,Z・・・に同様の値を求め，すべての和をする．

k 番目のクラスタの同種レコードの集まり具合の値 A_k を 1 つのクラスタ内での同種レコード集まり具合の評価とし，以下の式で与える．

$$A_k = \sum \frac{\sqrt{2}^{n-1}}{N}$$

5.2.2 クラスタ内での生物種のペナルティ

進化的に近い生物種が含まれるクラスタの方が評価が高くなるように，クラスタ内に含まれる生物種の進化的隔たりによってクラスタの集合の評価にペナルティを加え

ることとした．

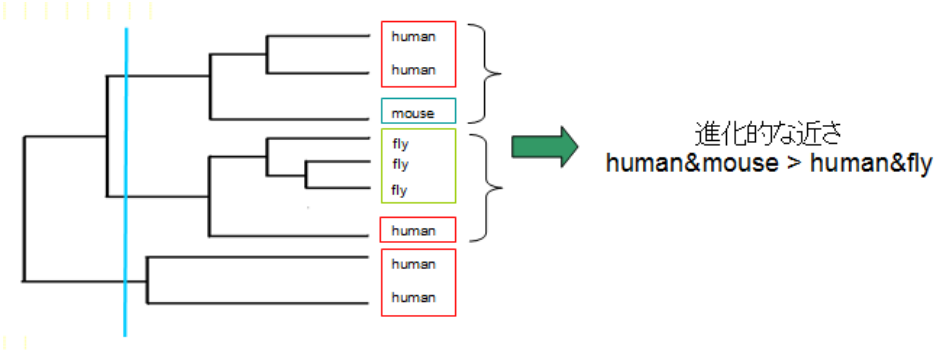

図 3　クラスタリング結果樹形図での生物種分布の例

● ペナルティの設定

生物種の組み合わせ		P_{spi_spj}※
yeast	fly	2.5
yeast	chick	4.5
yeast	mouse&rat	4.9
yeast	human	5.0
fly	chick	3.0
fly	mouse&rat	2.6
fly	human	2.5
chick	mouse&rat	0.4
chick	human	0.5
mouse&rat	human	0.1

表 3　生物種間ペナルティ Pspi_spj (i≠j)

異なる生物種間に存在するホモログ転写因子の進化距離を参考にし，表 3 のように生物種間のペナルティを設定した．生物種間の進化的隔たりが大きいほどペナルティが大きくなる．

● クラスタ内のペナルティの計算法

ある位置で樹形図を区切った時にできるクラスタのうち，k 番目のクラスタのペナルティ Pk を以下の式で定義する．

$$P_k = \sum \left(\frac{P_{spi\_spj}}{N} \times R \right)$$

ここで N は全体のレコード数，R はペナルティの組み合わせ対象となる生物種 spi と spj のクラスタ内に含まれるレコード数の比である．$C_{spi} > C_{spj}$ の時「$R = C_{spj} / C_{spi}$」，$C_{spi} < C_{spj}$ の時「$R = C_{spi} / C_{spj}$」である．

5.2.3 クラスタ内での生物種の集合の評価

k 番目のクラスタの生物種の集合評価 Q_k を 5.2.1 のレコード集合評価から，5.2.2 のクラスタ内ペナルティ値を引いた以下の式で定義する．

$$Q_k = Ak - Pk$$

樹形図全体の評価 Q を各クラスタ評価 Q_k の総和である以下の式で表す．

$$Q = \sum Q_k$$

5.2.4 生物種集合評価値の比較

5.2.3 の評価値 Q によりクラスタリング結果を評価した．評価値 Q は 2~9 個クラスタが作成される各クラスタの切断段階で求めた．そして，乱数によって，ランダムにどのクラスタに所属するかを決定し，作成したクラスタリングと，今回クラスタリングした結果から算出した評価の値の比較を行った．

ランダムなクラスタリングは 5 回行った．

(a)	(b)	(c)				
		org	1.6	1.7	1.8	1.9
wa	2	1.170	0.784	0.799	1.326	0.822
wa	3	0.738	0.411	0.653	0.527	0.713
wa	4	0.560	0.490	0.265	0.508	0.364
wa	5	0.524	0.454	0.245	0.246	0.331
wa	6	0.461	0.391	0.167	0.264	0.303
wa	7	0.315	0.396	0.193	0.161	0.266
wa	8	0.085	0.456	0.295	0.131	0.269
wa	9	0.223	0.140	0.311	0.187	0.211

(a)	(b)	(c)				
		org	1.6	1.7	1.8	1.9
ce	2	2.502	3.003	3.469	3.469	3.469
ce	3	2.273	2.774	2.273	3.327	3.327
ce	4	2.132	2.906	2.405	3.166	3.237
ce	5	1.795	2.766	2.265	1.972	1.696
ce	6	1.707	2.281	2.177	1.884	1.463
ce	7	1.723	2.193	2.193	1.547	1.474
ce	8	1.357	2.101	1.608	1.563	1.314
ce	9	1.373	1.518	1.624	1.451	1.042

(a)	(b)	(c)				
		org	1.6	1.7	1.8	1.9
si	2	3.469	3.829	3.469	3.469	3.469
si	3	2.273	3.600	3.600	3.327	1.925
si	4	1.936	3.459	2.405	2.132	1.784
si	5	1.795	2.766	1.818	1.795	1.696
si	6	1.811	2.610	1.834	1.681	1.463
si	7	1.203	2.126	1.695	1.521	1.474
si	8	1.205	2.142	1.608	1.537	1.314
si	9	1.045	2.117	1.257	1.451	0.962

(a)	(b)	(c)				
		org	1.6	1.7	1.8	1.9
av	2	2.502	2.502	2.502	2.502	3.469
av	3	2.273	2.273	2.273	2.273	1.925
av	4	1.728	2.405	1.557	1.728	1.784
av	5	1.271	2.161	1.471	1.244	1.395
av	6	0.699	1.700	1.558	1.300	0.569
av	7	0.702	1.716	1.592	0.913	0.514
av	8	0.769	1.202	0.698	0.918	0.448
av	9	0.786	1.236	0.662	0.661	0.245

(a)	(b)	(c)				
		org	1.6	1.7	1.8	1.9
co	2	2.502	3.003	2.502	0.910	1.290
co	3	0.612	0.583	2.117	0.631	0.449
co	4	0.538	0.588	1.455	0.665	0.470
co	5	0.647	0.553	0.555	0.510	0.397
co	6	0.548	0.569	0.571	0.274	0.372
co	7	0.564	0.580	0.686	0.183	0.309
co	8	0.563	0.444	0.370	0.240	0.347
co	9	0.553	0.425	0.335	0.235	0.248

(a)	(b)	(c)				
		org	1.6	1.7	1.8	1.9
me	2	2.502	3.469	3.469	3.551	3.469
me	3	2.273	3.327	3.327	2.355	3.327
me	4	1.579	3.237	2.132	2.127	3.237
me	5	1.570	2.543	2.148	1.567	2.543
me	6	1.339	1.707	2.281	1.635	1.463
me	7	1.341	1.723	2.193	1.247	1.474
me	8	1.357	1.857	1.608	0.554	1.131
me	9	1.356	1.449	1.624	0.663	0.905

表 4　各クラスタリング結果における評価値 Q の値
(a)クラスタリング手法（wa：ward 法, si：最近距離結合法, co：最遠距離結合法, ce：重心法, av：平均法, me：median 法）　(b)クラスタ切断数　(c)レコードデータセットの種類（org：元データ，　1.6~1.9：3.3 において，閾値をその数字に設定して修正を行ったデータセット）

(a)	(b)					
		1	2	3	4	5
2		0.549	0.600	0.489	0.478	0.478
3		0.479	0.262	0.177	0.215	0.174
4		−0.040	−0.087	−0.238	−0.063	−0.068
5		0.165	−0.238	0.100	−0.011	−0.361
6		0.143	0.053	−0.028	0.056	0.104
7		−0.078	−0.019	−0.135	0.081	0.177
8		−0.481	−0.757	−0.613	0.179	−0.282
9		−0.455	0.068	−0.147	0.176	−0.194

表 5　ランダムにレコードがどのクラスタに属するかを決定したクラスタリングでの評価値 Q　(a) クラスタ切断数　(b)試行回数

5.3　考察

5.1 の結果について，DB の値によって評価した場合では，修正閾値が比較的小さい方が有意な結果が多く観察でき，修正度が大きい方が評価値が良くなった例が多い．Dunn の値によって評価した場合では，修正閾値間での違いは良く分からないが，クラスタリング手法によって有意な例が多いかどうかが異なっていた．どの指標でみるかによって，結果が異なるので，今回の手法に関して配列の修正が有効であるかどうかは今後さらに吟味していく必要がある．

5.2 の結果については，ただランダムにレコードを順番に抽出し，作成したクラスタより求めた生物種の集まりの評価値 Q は，表 5 のように全て 1 以下の値となった．それらと表 4 の評価値 Q を比べてみると，表 4 における値の方が概ね高い値を示しており，EER を元にした類似度を指標としたクラスタリングの方が生物種の集まりが良くなっていることが分かる．

6.　おわりに

本研究では, 文字列であるシスエレメントの塩基配列パターンを，情報科学的手法によって, 数量化することで解析を行った. シスエレメント配列の解析にあったっては, 元々のデータの登録状況に揺らぎがあるため，データの信頼性に関して不安が付きまとう. 今回は EER を指標としたクラスタリング結果を生物種間の関係性から見てみたが，今回の視点とは別の視点からのアプローチを試みると共に，今後，解析に必要な配列データのさらなる洗練，数量化，考察に必要な情報の選択，改善に取り組んでいく．

参考文献

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)　Ohya M. Trans. IEICE. Vol. E72 No.5 pp 556-560