—Reviews—

バイオデータベースリテラシーと制御領域配列を利用した新規創薬ターゲットの探索

宮崎　智

Bio-database Literacy and Its Application with Cis-regulatory Modules to Find Novel Drug Target Proteins

Satoru MIYAZAKI

Department of Medicinal and Life Science, Faculty of Pharmaceutical Sciences, Tokyo University of Science, 2641 Yamazaki, Noda City 278–8510, Japan

(Received June 23, 2008)

We have expected Bioinformatics as tools to extract new knowledge from whole genome sequences of various organisms. In the post-genome era, to find some knowledge of the gene regulation including locations of cis-regulatory elements, modules and those combinations became one of the big challenges on Bioinformatics field. Because, it is difficult and inefficient to determine all possible combinations of cis-regulatory elements by bio-chemical approach. However, computational ways might allow us to find out all cis-elements within a time frame. In this review, we introduce the current status of public available databases on Internet comparing our original database for the cis-modules. We also explain our new mathematical measurement to characterize sequence patterns for cis-elements of each transcription factors and its application to predict the gene expression regulation network.

Key words——bioinformatics; cis-regulatory element; drug target gene search; bio-database

1. 公開データベースの現状

各種生物のゲノム配列の決定とインターネットがより一般的になってきたことが重なり，公的資金によるプロジェクトの成果としての生物学的データが多くインターネット上に公開されている．そうしたデータの多くは無償であり，利用制限がないものがほとんどであるが，データ形式が統一されている訳ではなく，プロジェクト期間の終了とともに閉鎖されるサイトがある．したがって，安定的な運用が期待できない場合があり，利用の観点からは問題が残っている．また，Google などの検索エンジンで，「遺伝子」などのキーワードを入れて検索した場合には，数十万件のサイトがヒットしてしまうなどの事態になっており，利用者に適切なサイトをうまく見付ける仕組みが急務となっている．

質の高いデータベースを見付ける 1 つの方法と Nucleic Acid Research が毎年 1 月に特集しているデータベース特集（http://nar.oxfordjournals.org/content/vol36/suppl_1/index.dtl）を参照する方法がある．また，筆者らは，NCBI の提供している文献データベース（Pubmed：http://www.ncbi.nlm.nih.gov/pubmed/）のアブストラクト全体に対して，そこに記述されているデータベース名とその引用回数をまとめている．この結果をみると，Pubmed でよく用いられているのは 190 種ぐらいのデータベースであることが分かる．逆に言えば，これらは，様々な生物系の研究者に利用されているという点において，第 3 者的な評価があり，信頼性の高いデータを提供しているデータベースであると言えよう．Table 1 は，これらの 190 種のデータベースの中で，特に，転写制御に係わる情報を提供しているものをまとめたものである．

公開データベースの現状を語る際に取り上げておきたいもう 1 つの事柄は実際のそれらのデータベースを利用するときの留意点である．われわれは，先の Nucleic Acid Research などの文献で利用できそうなデータベースとその URL を見付けることが可

東京理科大学薬学部生命創薬科学科（〒278–8510 野田市山崎 2641
e-mail: smiyazak@rs.noda.tus.ac.jp
本総説は，日本薬学会第 128 年会シンポジウム S12 で発表したものを中心に記述したものである．

Table 1. Useful databases related to gene regulation on INTERNET

データベース名	データ種	内　　容	対象生物種
EPD	配列データ	転写開始点｜プロモータ領域｜転写制御｜転写因子	真核生物
Regulon DB	配列データ	転写制御｜転写制御領域｜転写因子	大腸菌 K-12 株
EcoCyc	ネットワークデータ	転写制御｜転写制御領域｜ゲノム｜代謝パスウェイ｜酵素｜シグナル伝達パスウェイ	大腸菌 K-12 株
JASPAR	配列データ	プロモータ領域｜転写制御	
PlantCARE	配列データ	プロモータ領域｜転写制御｜エンハンサー領域｜リプレッサー領域	植物
CORG	配列比較データ	転写制御｜転写制御領域｜ゲノム	脊椎動物
EpoDB	統合データ	転写制御領域｜遺伝子発現｜タンパク質｜赤血球｜発生・分化	脊椎動物
PLMItRNA	配列データ	転写制御｜転写制御領域｜ミトコンドリア｜tRNA	植物
TRANSFAC	配列データ	プロモータ領域｜転写制御｜転写因子	
SCPD	配列データ	プロモータ領域｜転写制御｜転写因子	酵母
ooTFD	配列データ	プロモータ領域｜転写制御｜転写因子	

能であるが，実際にその URL にアクセスしてみると，文献上で見付けたデータベースがそのまま公開されていないことが多いのである．文献上で報告されるのは，あるプロジェクトがその成果の全体を取りまとめるために構築したデータベースである．しかし，公開サイトではそのデータベースがいくつかに細分されて複数のデータセットとなっている場合や，そのプロジェクトが自らのデータを解析するために使った外部のデータベースのコピーが含まれていることがある．そのために，文献によって得たデータベース名が，公開サイトではそのまま用いられていないことがあることに加え，オリジナル以外のデータもあり，利用者（特に初心者）は注意が必要である．

2. バイオインフォマティクスによる転写制御研究

2-1 シスエレメント配列構造の規則性と進化

ポストゲノム時代になって，研究者の興味が，ゲノム配列中の遺伝子探索から，転写制御のメカニズム探索や遺伝子ネットワークの予測など，遺伝子や分子間の相互作用に移ってきている．生化学的な実験を基に，転写因子が認識する塩基配列（シスエレメント）の解明が進行している．こうした配列は，JASPER[1,2] や TransFAC といったデータベースにまとめられて提供されている．

各々の転写因子が認識するシスエレメント配列は 1 パターンではなく，様々な配列パターンを認識することが明らかになっており，Fig. 1 に示したように，‘‘MYC’’ という名前の転写因子は「CACGTG」という配列をはじめ，「CACGTC」や「CACACG」など，様々なパターンの配列に結合する．そして，列毎に配列パターンを比較した場合，一応のコンセンサス配列は得られるものの，各列における塩基にはかなりのゆらぎがあり，これらの配列パターンの汎用的な共通性ルールは分かっていないと言える．また，現在では 1500 以上のタンパク質と DNA の複合体の構造が X 線結晶解析などの手法で解かれているが，これらの構造を詳しく調べてみても，相互作用しているアミノ酸残基と塩基のペアの間には厳密な対応関係はみられていない．また同じ相互作用ペアでも，その塩基とアミノ酸残基の空間的な位置関係は様々である．[3] すなわち，アミノ酸残基と

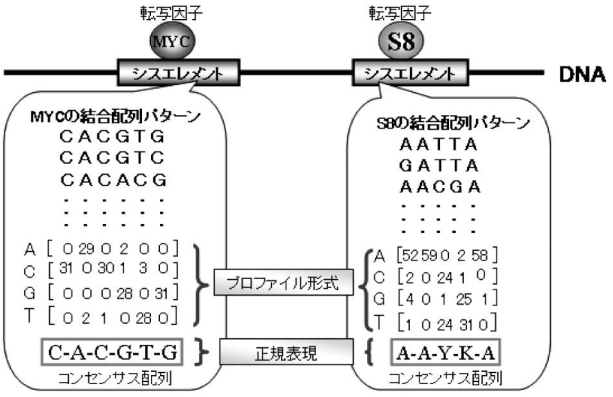

Fig. 1. Cis-regulatory Element and its Typical Data Fomat

塩基の相互認識にもかなりの冗長性と柔軟性があると言える．

転写因子のターゲットを予測するには，実験的に結合することが知られている塩基配列を集めて共通なパターン（コンセンサス配列）を見付け，それを配列モチーフあるいは重み行列という形で表現し，類似の配列を検索するという方法が現在最も広く用いられている．しかし，転写制御研究は比較的新しい分野であり，明らかとなっているシスエレメント配列パターンがまだ十分にはないため，これらの方法では予測精度が悪く，転写因子が結合するシスエレメントを特定することは困難な状態にある．また，シスエレメント配列はわずか数塩基から数十塩基程度の非常に短い配列であることも，ゲノム上に多数の予測結合部位を生み出す原因となっており，ターゲットとなる配列の予測を困難にしている．

われわれは，転写因子とそれらが結合するシスエレメント配列の間に十分な規則性を見い出していないが，転写因子とシスエレメントの間にある相互認識のルールがシスエレメントの文字列中に隠されている可能性に着目している．もし，シスエレメント配列に潜む規則性が明らかになれば，ゲノム上に存在する未知のシスエレメント配列を精度よく予測することが可能となり，ひいては転写制御機構の解明につながると考えられる．

1 つの方法として，情報量の概念を応用することでシスエレメント配列を数量化し，網羅的に比較・解析することで転写因子のシスエレメント配列認識における規則性を探ったので報告したい．

2-1-1. 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[1,2]（version 3.0）から取得した．JASPAR では 138 種の転写因子の結合配列パターンが LOGO 形式やプロファイル形式によって公開されている．そのうち 124 種の転写因子については，LOGO やプロファイルを作成する際に用いられた結合配列パターンが取得可能であった．JASPAR からダウンロードしたシスエレメント配列データは，124 種の転写因子それぞれに対して結合配列パターンが FASTA 形式でまとめられていた．取得したシスエレメント配列データは小文字英字と大文字英字で記されている（Fig. 2）．小文字英字を含めた部分は実験的に転写因子が結合することが明らかになった配列を示しており，大文字英字部分は配列パターン中で最も保存されている部分を示す．本研究では，転写因子が結合するための特に重要な情報を有していると考えられている，大文字英字で記された配列部分を解析に用いることにした．

取得した配列データは，重複を除き，大文字で表された塩基配列をまとめ，124 レコードとした．例えば，AGL3 という転写因子が結合するシスエレメント配列は，JASPAR データベース中に 97 配列存在していたが，そのうち大文字英字で記されている配列部分だけを取り出し，重複している配列を省くと 63 パターンの配列となる．そこで，その 63 パターンの配列を 1 つのテキストにまとめ，1 レコー

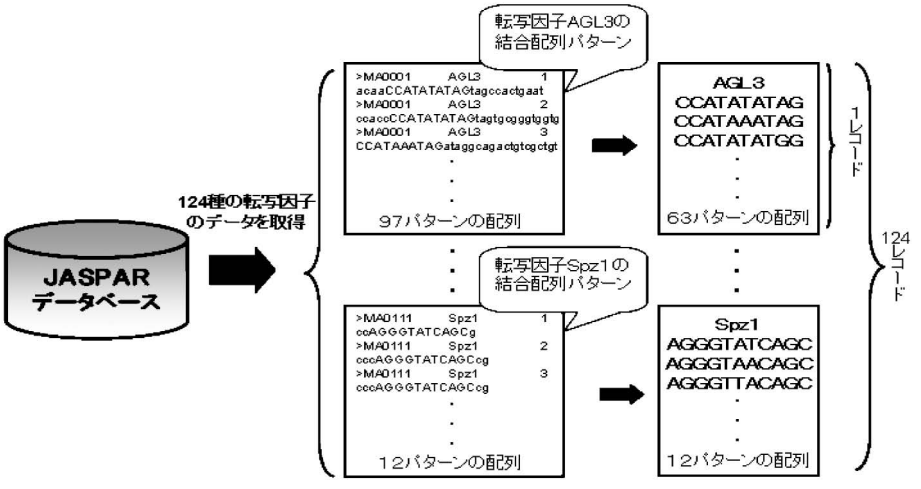

Fig. 2. Raw Data in JASPAR Database

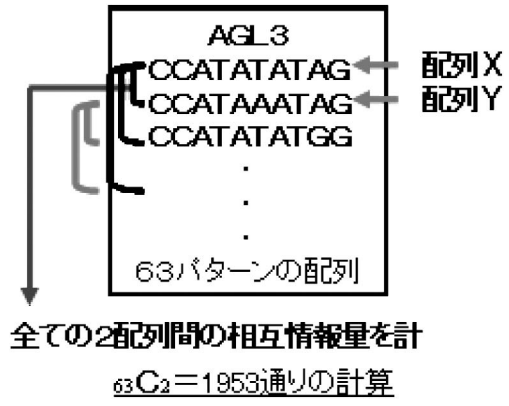

Fig. 3. Re-formatted Data from JASPAR and Trans FAC Database

ドとした（Fig. 3）．各々のレコードを構成しているシスエレメント配列パターンは，3–114 配列あり，配列の長さは 4–22 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*（キンギョソウ），*Arabidopsis thaliana*（シロイヌナズナ），*Drosophila melanogaster*（キイロショウジョウバエ），*Gallus gallus*（ニワトリ），*Halocynthia roretzi*（マボヤ），*Homo Sapiens*（ヒト），*Hordeum vulgare*（オオムギ），*Mus musculus*（ハツカネズミ），*Nicotiana sp.*（タバコ），*Petunia hybrida*（ペチュニア），*Pisum sativum*（グリーンピース），*Rattus norvegicus*（ドブネズミ），*Triticum aestivum*（コムギ），*Xenopus laevis*（アフリカツメガエル），*Zea mays*（トウモロコシ）の 15 種であった．

2-1-2. シスエレメントの数量化

2-1-2-1. シャノンエントロピーの計算　各転写因子が結合するシスエレメント配列それぞれについて，シャノンエントロピー[4]を計算する．任意の配列におけるシャノエントロピー(S)を以下の式により与える．[5,6]

$$S=-\sum_{i=A,T,G,C} P_i \log_2 P_i \qquad (1)$$

このとき，P_i 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 それぞれの出現確率をさす．

例として，「CCATATATAG」という配列のシャノンエントロピーの計算例を以下に示す．配列中の A, T, G, C 各塩基の出現確率(P_A, P_T, P_G, P_C)は，それぞれ

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

である．したがって，この配列のシャノンエントロピー(S)は，

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

となる．

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できる．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 の値を取る．エントロピーの値が 0 に近いほど，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．

2-1-2-2. 相互情報量の計算　次に，124 レコードの各レコード内で，考えられるすべての 2 パターン配列間において相互情報量[4]を計算した．例えば，Fig. 3 に示したような AGL3 のレコードの場合，63 パターンの配列があるため，すべての 2 配列の組み合わせは $_{63}C_2$＝1953 通り考えられ，そのすべての組み合わせについて相互情報量を計算した．計算はすべて Perl 言語でプログラムを組むことによって計算した．任意の 2 配列 X，Y 間における相互情報量(I)[5]を以下の式により与える．

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

このとき，P_i, P_j はそれぞれ配列 X，配列 Y における A，T，G，C それぞれの出現確率である．また，P_{ij} は各位置における配列 X と配列 Y の塩基の組み合わせ（A－A, A－T, … C－C）の出現確率である．以下にその計算例を示す．

配列 X：CCATATATAG

配列 Y：CCATGTGTAG

の相互情報量を求める場合，配列 X，配列 Y における各塩基の出現確率［P(X), P(Y)］は，それぞれ，

$$P(X_A)=\frac{4}{10}\quad P(X_T)=\frac{3}{10}\quad P(X_G)=\frac{1}{10}$$

$$P(X_C)=\frac{2}{10}$$

$$P(Y_A)=\frac{2}{10}\quad P(Y_T)=\frac{3}{10}\quad P(Y_G)=\frac{3}{10}$$

$$P(Y_C)=\frac{2}{10}$$

である．また，各位置における配列 X と配列 Y の各塩基の同時出現確率［P(X；Y)］は，

$$P(X_A\,;\,Y_A)=\frac{2}{10}\quad P(X_A\,;\,Y_G)=\frac{2}{10}$$

$$P(X_T\,;\,Y_T)=\frac{3}{10}\quad P(X_G\,;\,Y_G)=\frac{1}{10}$$

$$P(X_C\,;\,Y_C)=\frac{2}{10}$$

である．したがって，この配列 X，配列 Y の相互情報量(I)は，

$$I=\frac{2}{10}\times\log_2\left(\frac{\frac{2}{10}}{\frac{4}{10}\times\frac{2}{10}}\right)+\frac{2}{10}\times\log_2\left(\frac{\frac{2}{10}}{\frac{4}{10}\times\frac{3}{10}}\right)+\frac{3}{10}$$
$$\times\log_2\left(\frac{\frac{3}{10}}{\frac{3}{10}\times\frac{3}{10}}\right)+\frac{1}{10}\times\log_2\left(\frac{\frac{1}{10}}{\frac{1}{10}\times\frac{3}{10}}\right)+\frac{2}{10}$$
$$\times\log_2\left(\frac{\frac{2}{10}}{\frac{2}{10}\times\frac{2}{10}}\right)=1.571\cdots$$

となる．

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，2 つの系が共有している情報を表している．配列解析においてはシスエレメント配列 X と Y における塩基の出現における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，相互情報量は最大値である 2 を取る．相互情報量は，配列 X と配列 Y の間で共有されている情報の量であり，結合する転写因子が配列 X と配列 Y を「どの程度同じ配列としてみなしているのか」という指標になる．

2-1-2-3. エントロピー進化率（Entropy Evolutional Rate: EER）の計算　相互情報量を計算することによって，各転写因子が結合する配列の冗長度を数値化することはできたが，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，解析する際にすべてのシスエレメント配列を等しく扱うことができない．例えば，配列 A と配列 B，そして配列 C と配列 D の相互情報量を考えてみる．I(A；B)＝0.8，そして I(C；D)＝0.4 であるので，配列 C, D よりも配列 A, B の方が共有されている情報が多いと思われがちである．しかし，配列 C と配列 D のシャノンエントロピーはもともと小さいため，完全に情報が共有されている場合だとしても相互情報量の値が小さくなる場合がある (Fig. 4)．

そこで，相互情報量を正規化した値である EER[6–8] を利用した．EER は 2 つの情報源のエントロピーを足し合わせたものに対して，そのうちどの位を相互情報量が占めているのかという値を示す．このような正規化した値を利用することで，2 つの情報源の関連度合いを正しく評価し，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 Eq. (1) と Eq. (2) を用いた，以下の式により与える．

$$EER(X\,;\,Y)=\left(\frac{I(X\,;\,Y)}{S(X)+S(Y)-I(X\,;\,Y)}\right)\quad(3)$$

このとき EER は，0≦EER≦1 の値を取る．以下に配列 X と配列 Y の EER 計算例を示す．

配列 X：CCATATATAG

配列 Y：CCATGTGTAG

Eq. (1) に従って，配列 X，配列 Y それぞれのシャノンエントロピー(S)を計算すると，

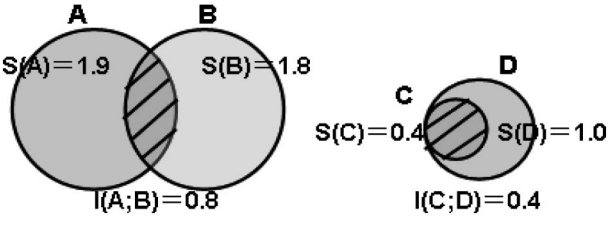

Fig. 4. Matual Eneropy and its Concept

$$S(X)=\left(-\frac{4}{10}\times\log_2\frac{4}{10}\right)+\left(-\frac{3}{10}\times\log_2\frac{3}{10}\right)+\left(-\frac{1}{10}\times\log_2\frac{1}{10}\right)+\left(-\frac{2}{10}\times\log_2\frac{2}{10}\right)$$
$$=1.846\cdots$$
$$S(Y)=\left(-\frac{2}{10}\times\log_2\frac{2}{10}\right)+\left(-\frac{3}{10}\times\log_2\frac{3}{10}\right)+\left(-\frac{3}{10}\times\log_2\frac{3}{10}\right)+\left(-\frac{2}{10}\times\log_2\frac{2}{10}\right)$$
$$=1.971\cdots$$

である．さらに，Eq. (2)により，配列 X と配列 Y の相互情報量(I)は，

$$I=\frac{2}{10}\times\log_2\left(\frac{\frac{2}{10}}{\frac{4}{10}\times\frac{2}{10}}\right)+\frac{2}{10}\times\log_2\left(\frac{\frac{2}{10}}{\frac{4}{10}\times\frac{3}{10}}\right)+\frac{3}{10}\times\log_2\left(\frac{\frac{3}{10}}{\frac{3}{10}\times\frac{3}{10}}\right)+\frac{1}{10}\times\log_2\left(\frac{\frac{1}{10}}{\frac{1}{10}\times\frac{3}{10}}\right)+\frac{2}{10}\times\log_2\left(\frac{\frac{2}{10}}{\frac{2}{10}\times\frac{2}{10}}\right)$$
$$=1.571\cdots$$

である．よって，配列 X と配列 Y の EER は Eq. (3)より，

$$EER(X;Y)=\left(\frac{I(X;Y)}{S(X)+S(Y)-I(X;Y)}\right)=\left(\frac{1.571}{1.846+1.971-1.571}\right)=0.6994\cdots$$

となる．

比較した 2 配列間の EER 値が 0 に近いほど，配列 X と配列 Y における塩基の出現には関連性がないことを意味し，EER が 1 に近いほど，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．そして，EER が 2 つの配列の関連度合いを示すことから，EER は転写因子のシスエレメント配列の認識に対する柔軟性の度合いを示していると考えることができる．

2-1-3. 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，転写因子毎に，それぞれの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 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，${}^{m}C_2$ 個と異なるため，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．シスエレメント配列パターン間で従属関係がみられるものが多い場合は，グラフは右寄りになり，従属関係があまりみられない場合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言える．

2-1-4. クラスタ解析　各転写因子のシスエレメント配列認識に対する柔軟度を比較するために，作成した頻度分布の類似性を基にユークリッドの距離・ウォード法による階層的クラスタリングを行った．階層的クラスタリングとは，個体間の類似度あるいは非類似度（距離）に基づいて，最も似ている個体から順次に集めてクラスタを作って行く方法で，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の認識に対する柔軟性の度合いが似ている転写因子同士を知ることができる．そこでクラスタ解析を 124 レコードのデータすべてを用いて行った．また，DNA 結合ドメインの種類毎や生物種毎でもクラスタ解析を実行した．

各頻度分布の形状を，頻度分布の各階級における EER の相対値 10 ポイントと隣接する階級間の傾き 9 ポイントの合計 19 次元ベクトルによって表した．Figure 5 における頻度分布では，EER の相対値 10 ポイントは，〇の部分を示しており，階級間の傾きは実線で示した部分を指す．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できる．[9]

ここで，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．

$$D(a,b)=\sqrt{\sum_{i=1}^{n}(a_i b_i)^2} \qquad (4)$$

このとき，i は各階級における EER の相対値 10 ポイントと，隣接する階級間の傾き 9 ポイントを示す．したがって，$n=19$ となる．

2-1-4-1. シスエレメント配列構造の進化系統関係　JASPAR から 3 レコード以上のデータが得

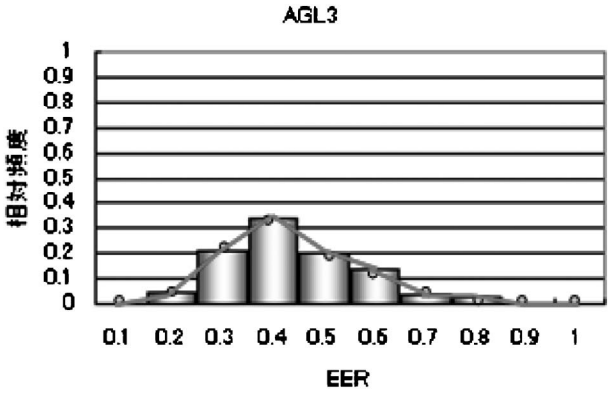

Fig. 5. Frequent Distribution of EER

られた生物種 8 種（*Arabidopsis thaliana*, *Antirrhinum majus*, *Zea mays*, *Drosophila melanogaster*, *Pisum sativum*, *Rattus norvegicus*, *Mus musculus*, *Homo sapiens*）について，生物種別に相対頻度分布を作成し，各階級値と傾きをベクトルとして分子系統樹を作成した．

転写因子が持つ DNA 結合ドメインの種類とシスエレメント配列の認識パターンの関連性を考察するために，DNA 結合ドメイン名をラベルにして，すべての頻度分布を用いたクラスタ解析を行った．その結果を Fig. 6 に示す．作成されたデンドログラムをみると，一部のクラスタでは類似した DNA 結合ドメインがまとまる例がみられた．しかし，すべてのクラスタにおいて DNA 結合ドメインとシスエレメント配列の認識パターンに関連性を示唆できるには至らなかった．ところで，同種の DNA 結合ドメインを持つ転写因子でも，認識するシスエレメント配列の長さやパターンが全く異なっている．そこで，同種の DNA 結合ドメインが認識するシスエレメント配列パターンには，何か規則性がないか解析することにした．5 個以上のデータが存在している 11 種の DNA 結合ドメイン（bHLH, bHLH-ZIP, bZIP, ETS, FORKHEAD, HMG, HOMEO, MADS, NUCLEAR RECEPTOR, REL, ZN-FINGER C2H2）に対して，種類毎にクラスタ解析を行った．一例として bHLH のクラスタ解析結果を Fig. 7 に示す．この結果をみると，いくつかのドメインに関しては，同一の生物種のシスエレメント配列パターンが近隣にクラスタリングされる傾向がみられた．bHLH の例では，*Homo Sapiens* 同士がクラスタを作り，その上位に *Mus musculus* のクラスタが形成されている．こうしたことより，各 DNA 結合ドメインが許容できるシスエレメント配列の冗長度は，生物種によって異なることが考えられた．

次に，6 種の生物種（*Arabidopsis thaliana*, *Drosophila melanogaster*, *Homo Sapiens*, *Mus musculus*, *Rattus norbegicus*, *Zea mays*）について，生物種別のクラスタ解析について解説する．*Homo Sapiens* の解析結果を Fig. 8 に示す．生物種毎にクラスタリングを行った場合，全体でクラスタリングを行った場合よりも，DNA 結合ドメインが類似するもの同士が近隣にクラスタを作り易い傾向がみられた．

以上の結果より，生物種や DNA 結合ドメインの違いによって，転写因子が認識できるシスエレメント配列のゆらぎの許容度には差異があると考えられ，シスエレメントの配列パターンを生物種や DNA 結合ドメインによって特徴付けられる可能性が示唆された．

3. 新規創薬ターゲット分子の予測に向けて

3-1. 遺伝子上流領域の配列構造　前章では，個々のシスエレメントパターンの特徴を「ゆらぎ」の面から考察した．本章では，全ゲノム配列上でのシスエレメント配列の分布について報告する．シスエレメントの分布を考える前に，遺伝子上流配列の塩基の出現パターンについて解説しよう．ヒトの完全長 cDNA と全ゲノム配列を基に，ヒトの遺伝子マップを作成し，解析している H-invitational データベース（http://www.h-invitational.jp/）を利用して，ヒトの遺伝子配列約 30000 件についてその上流配列（2000 塩基）を取得して解析を行っている．するとこれらの配列中の GC 含量は意外に低いことが分かる．また，A,T,G,C 各々の塩基の出現確率について調べると，上流配列でかなりのばらつきがあることが分かる．30000 件の配列において，A, T, G, C の出現確率がほぼ均等であると思われるものは，半数程度であり，残りの半数については，どれかの塩基の出現確率が極端に高くなる傾向がみられた（Fig. 9）．次に，先の JASPAR データベースに登録されているシスエレメント配列を上流配列にマップしてみる．これらは単に，シスエレメントと上流配列のアライメントを行っただけであるので，その配列がシスエレメントとしての機能を有してい

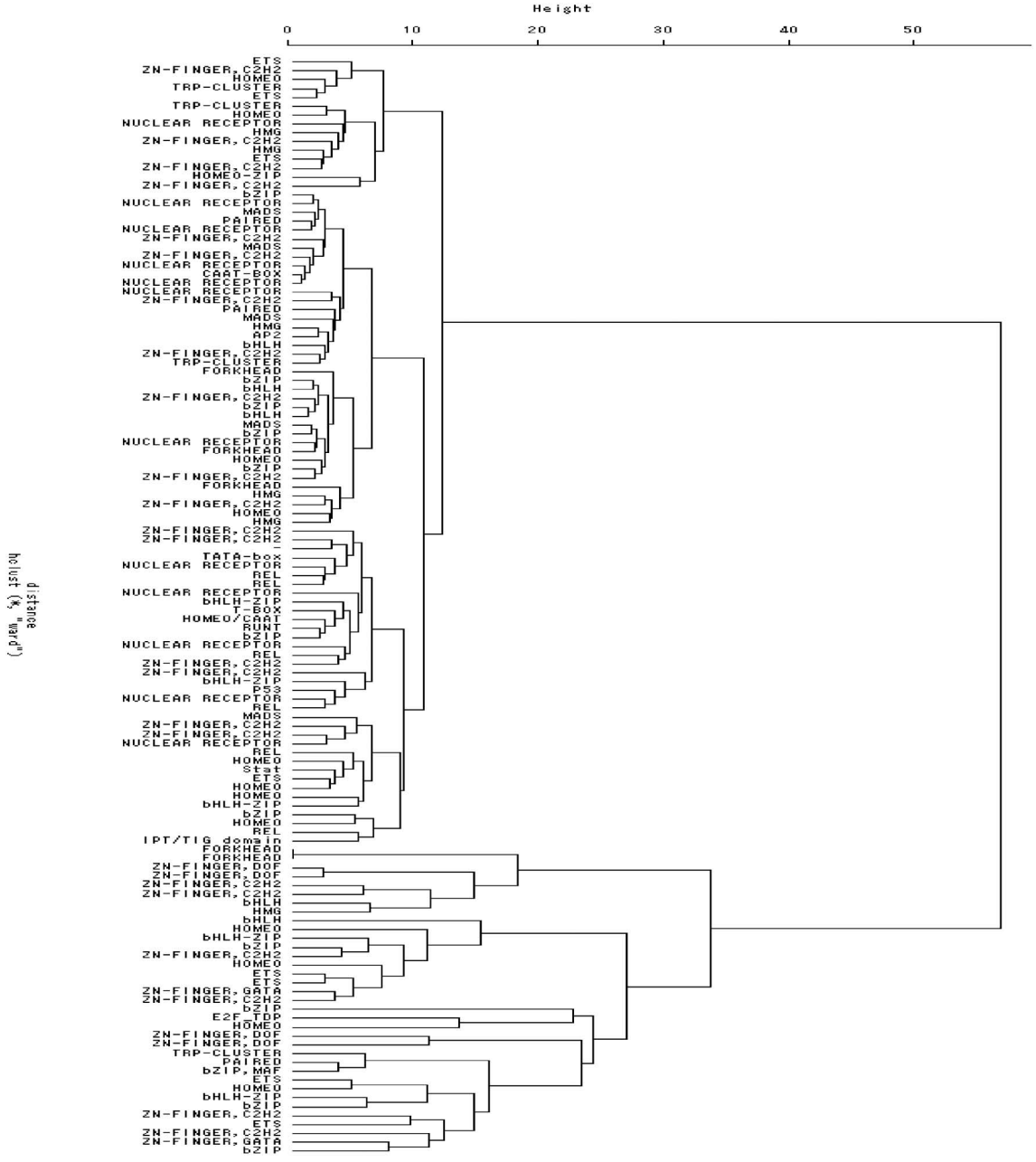

Fig. 6. Clustering of DNA Binding Domains of Transcription Factors

るかどうかは分からない．そこでここでは，マップされたシスエレメント配列を「シスエレメント様配列」と呼ぶことにする．結果をみると，ある遺伝子上流では，何種類もシスエレメント様配列がタンデムに存在している場合や，ある遺伝子配列上流では，数種のシスエレメント様配列が離散的に見付かる場合など，シスエレメント様配列の存在パターンは，各上流配列でかなりのばらつきがあることが分かった．逆にいえば，各々の遺伝子は，その上流配列の塩基構成が特徴的である可能性が示唆されているとも言える．

また，各上流配列の各塩基の出現頻度を用いて，あるシスエレメト配列がその上流配列に見付かる確率の期待値と実際にマッピングを行ったあとで，あるシスエレメントがマップされた事後確率を比較してみると，ほとんどすべてのシスエレント配列において，事後確率が期待値の確率よりもはるかに小さいということが分かった．この 2 つの確率の差について，有意水準 5% における統計的検定の結果，有意差が認められた．すなわち，シスエレメントは，

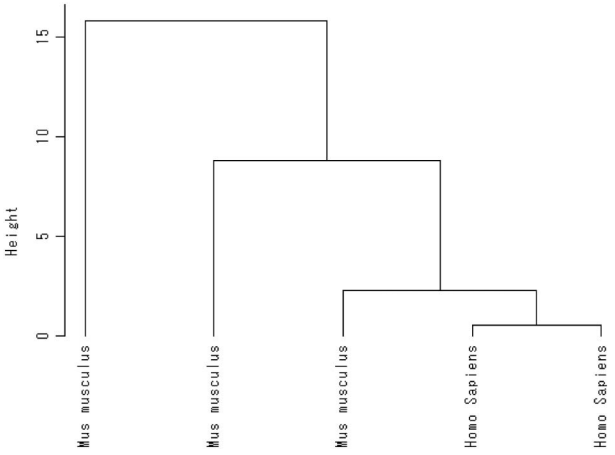

Fig. 7. Clustering of Species by Cis-elements for bHLH Domain

5–20 の短い配列であり，ゲノム上の至るところで偶然に見付かる可能性が高いように思われがちであるが，実際には，配列長から予測されるランダム性はさほど高くなく，必要な場所を選んで存在しているように思われる．

3-2. シスエレメントのパターンによる局在性予測 前節で示唆されたようなシスエレメント配列の存在パターンの制約性から，遺伝子上流配列による，遺伝子あるいはそれにコードされたタンパク質の機能予測について提案してみよう．前述した H-invitational データベースでは，予測された遺伝子について，その遺伝子がコードしているタンパク質

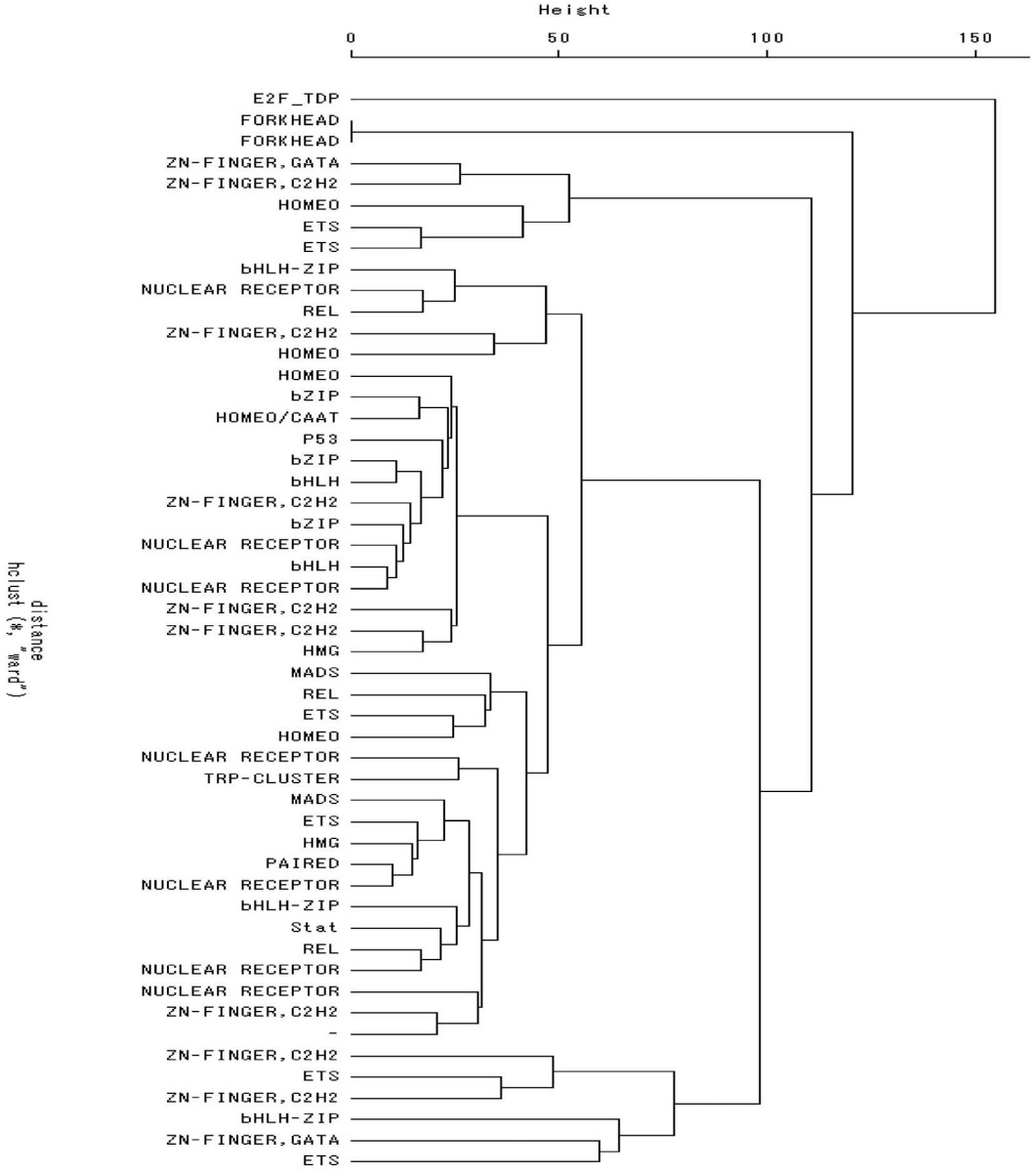

Fig. 8. Clustering of DNA Binding Domains in Human

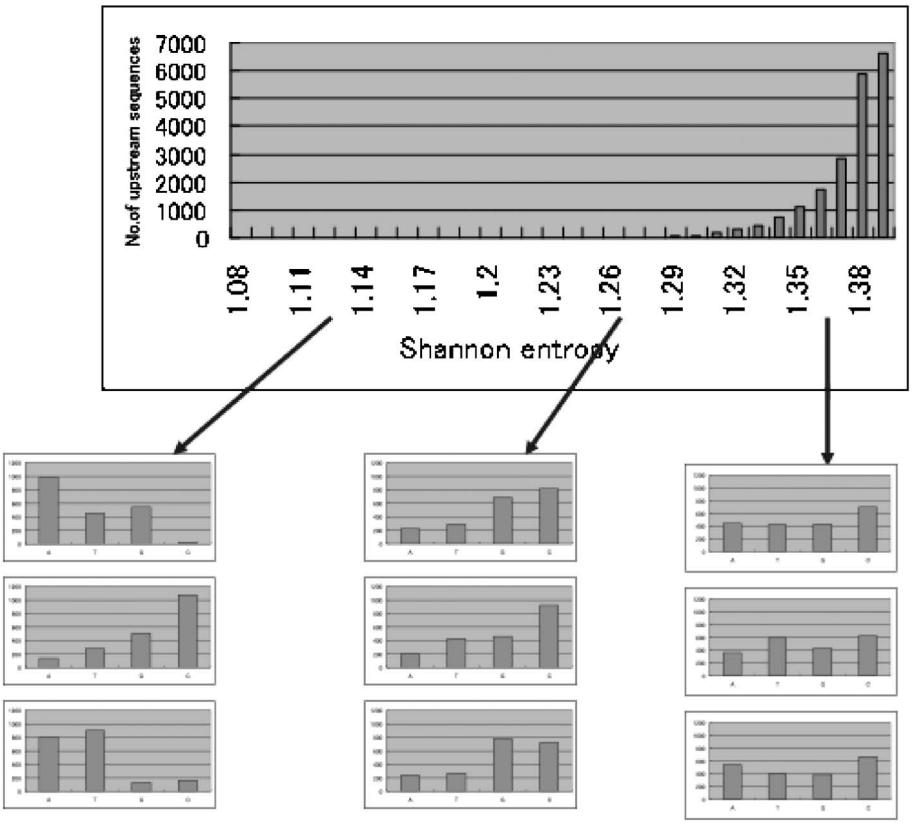

Fig. 9. Frequent Distribution of Upstream Sequences by Shannon Entropy

Table 2. Examples of cis-regulatory element sequence and protein localization

Cis element	Transcription factor	Localization in a cell								
		Cytoplasm	Cytoskeleton	ER	Ext cell matrix	Golgi	Mitochondria	Nucleus	Peroxisome	Plasma memb
tgacctttgcccag	COUP-TF	0	0	0	1	0	0	0	0	0
ggagacaccatt	HLF	0	0	0	0	0	0	0	0	1
attaattaggtcag	RO Ralfa-2	0	0	0	0	0	0	1	0	0

の局在化情報も持っている．こうした局在情報を持つ 3830 個の遺伝子について，局在性と上流配列中のシスエレメントの間の関係をまとめてみた．Table 2 は，3830 遺伝子の上流に見付かったシスエレメントからそれを認識する転写因子について，下流遺伝子にコードされたタンパク質の局在性をまとめたものである．この Table 2 の 2 行目の第 2 列をみると，例えば，ミトコンドリアに移行するタンパク質では，その遺伝子の 61 個で上流に AML-1 に認識されるシスエレメントがあることが分かる．局在性とシスエレメントの間には，特別な関係があるように見受けられないが，この調査の中で，その出現頻度が極めて低く，局在場所が 1 対 1 に対応しているシスエレメントが 3 種あることが判明した．このことから，ただちにシスエレメントを用いて局在性予測を行うことはできないが，非常に稀にしかみつからないシスエレメントがあり，それらは下流タンパク質の局在場所の判別の指標となる可能性があることが分かる．また本稿では触れていないが，上流配列中のシスエレメントの分布によって遺伝子ネットワークを予測する試みが行われてきている．3-1. 節で述べたように，上流配列の塩基組成にかなりの差があることから考えると，シスエレメントの有無を指標にした遺伝子ネットワークの予測法の開発が

おおいに期待できると思われる.

REFERENCES

1) Sndelin A., Alkema W., Engstrom P., Wasserman W. W., Lenhard B., *Nucleic Acids Res.*, **32**, D91–D94 (2004).
2) Wasserman W. W., Sndelin A., *Nat. Rev. Genet.*, **5**, 276–287 (2004).
3) Sarai A., Kouno H., *Seibutubuturi*, **47** (3), 160–166 (2007).
4) Shannon C. E., *Bell Syst. Tech. J.*, **27**, pp. 279–423, 623–656 (1948).
5) Ohya M., *Trans. IEICE*, **E** (72), 556–560 (1989).
6) Ohya M., Sato K., Rep. Math Phys., **46**, 419–427 (2000).
7) Ohya M., *Densi Johothusingakukaishi*, **71** (3), 295–297 (1988).
8) Miyazaki S., Sugawara H., Ohya M., *Genes Genet. Syst.*, **71**, 323–327 (1996).
9) Michaels G. S., Carry D. B., Askenazi M., Fuhrman S., Wen X., Somogyi R., *Pac. Symp. Biocomput.*, **3**, 42–53 (1998).