バイオインフォマティクスによるゲノム配列解読のロゼッタストン探索（特集 数理科学の最先端）

著者	宮崎 智
雑誌名	理大科学フォーラム
巻	33
号	6
ページ	14-17
発行年	2016-06
URL	http://id.nii.ac.jp/1275/00002548/

バイオインフォマティクスによるゲノム配列解読のロゼッタストン探索

東京理科大学 薬学部 生命創薬科学科 教授　宮崎（みやざき） 智（さとる）

完全長ゲノム配列データのロゼッタストンは見つけられるか？

2003年に完全長ヒトゲノム配列決定の記事が世界中を駆け巡ってから15年が経過した。ヒトゲノム配列の決定時には，この成果が生命科学の進展を加速させ「夢」の医療や創薬が早々に期待できるとのアナウンスもされていた。

筆者は当時，ゲノムデータ収集を効率化すべく，国際塩基配列データベースの拠点であるDNA Data Bank of Japan（DDBJ）のデータベース構築局副局長の任についていたが，ヒトゲノム配列の将来について，現場の見方はいささか冷ややかであった。それに呼応するように，完全長ヒトゲノム配列データの記事の特集をしたNatureの編集記は，ヒトゲノム配列決定の成果を，「この成果によって，人類は，ようやく生命科学や医療，創薬への扉を開いた」とのみ記していた。なぜなら，完全長ヒトゲノムデータは，DNA分子を構成している４つの塩基の並び順と塩基長（ヒトの場合は30億塩基）を明らかにしたに過ぎないからである。

細胞内では，転写因子や翻訳に関わる，タンパク質を中心とした分子が，ゲノム配列データを「解読」し，適切な時期に，適切な量で，必要とされる分子を発現させている。塩基の並びの中の規則性が，この「解読」の仕組みに関わっていると推測できるものの，15年経過した今でさえ，そうした規則性についての知識が十分に得られているわけではない。すなわち，我々は「設計図」を発掘したものの，そこに書かれている内容を解読できないでいるのである。ロゼッタストンが見つかり，古代エジプト文字からギリシャ文字への翻訳が可能となったように，まさに「ゲノム文字」を我々の理解できる知識へ変換するための基盤を見つけることが必要である。

ロゼッタストンの解読には，言語学者だけでなく，物理学者のトマス・ヤングが関わっていたことは非常に興味深い。これをゲノム配列の解読に当てはめるなら，生命科学者とともに，コンピュータ科学者や数理科学者の関与がキーとなる可能性は高いと考えている。コンピュータによるデータ処理，機械学習や数理科学を駆使した解析が必修となるからである。

タンパク質遺伝子のロゼッタストン

ゲノム配列の中に，タンパク質の設計図である遺伝子がどのように書かれているのかを明らかにしたのは，遺伝暗号表の発見に因るところが大きい。遺伝暗号表は，塩基の３つ組が，１つのアミノ酸を指定するコードとなっており，20種類のアミノ酸について，それぞれ対応する３つ組を解読したものである。

この３つ組は「コドン」と呼ばれている。遺伝暗号表は，タンパク質遺伝子のロゼッタストンといっても過言ではない。しかも，タンパク質の先頭を示す開始コドンや終了を示す終止コドンがあることも判明した。塩基は４種類なので，３塩基の組み合わせの場合の

数は，４×４×４＝64種類となる。一方アミノ酸は，20種類なので，幾つかのアミノ酸は，複数のコドンを持つようになる。これは遺伝暗号表の縮退と呼ばれている。

ゲノム配列上でのタンパク質遺伝子領域は，開始コドンで始まり，終止コドンで終わるから，機械的に開始コドンと終止コドンで囲まれた領域を見つければ，それがタンパク質遺伝子の候補と考えることができる。最も簡単な遺伝子探索の数理モデルといえよう。この段階では，「単なる候補」であって，疑陽性（実際は遺伝子でないもの）も含まれている。一方，生化学の力を借りて，実際にタンパク質として使われている領域の具体例が特定されてくると，機械学習のアルゴリズムを用いて，先の遺伝子候補のうち，本物の遺伝子と偽物では塩基の並び方にどのような違いがあるのかを識別することができるようになる。これは，我々の言語をコンピュータで解析する例に似ている。26種類のアルファベットをランダムに発生させてその順番に並べた文字列は，全く意味のない英単語となっていく。しかし，実際の辞書をみれば判るように，英単語には，Eで始まる単語は多いが，Xで始まる単語は少ないといったように，我々の言語に従う「並び」には特別な傾向があると考えられる。機械学習のアルゴリズムを適用すると，自然言語の特性を解析するように，ゲノム配列上の遺伝子領域に出現する塩基文字の並びの特性を判別する関数が得られるのである。ゲノム配列を記号列とみなして，情報科学の観点から遺伝子領域を探索する手法は，代表的な「バイオインフォマティクス」である。

こうして，タンパク質遺伝子（タンパク質の設計図）がゲノム配列のどの位置にどれくらい記述されているのかが分かってきたわけだが，ヒトゲノムの場合，アミノ酸に対応する領域の合計は，全体の３％ほどである。残りの領域には何が書かれているのであろうか？　筆者らが着目している領域の１つに転写制御に関わる領域がある。遺伝子は，使われる時期や細胞種に応じて適切に発現の制御がされている必要がある。それが崩れた事例として，細胞の癌化があるわけである。したがって，ゲノム配列のどこかに，遺伝子の使われ方を制御するための情報（以下，転写制御領域と呼ぶ）があり，その文法があると仮定し，それを実証しようというわけである。

バイオインフォマティクスによる転写制御研究

⑴ シスエレメント配列構造の規則性と進化

真核生物においては，核内のDNA分子からRNA分子が作られる際にRNA合成酵素の結合を手助けする，転写因子と呼ばれるタンパク質群がある。

この15年の生命科学の研究によって，DNA側には，転写因子が結合する部位（シスエレメント配列）が存在し，ある転写因子が認識するシスエレメント配列は１種類ということはなく，認識する配列の文字の並びには「ゆらぎ」があることが判った。

例えば，“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」などの配列に結合できる。これらの「ゆらぎ」はさまざまであり，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いえる。また，現在では1,500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，これらの立体構造を詳しく調べてみても，その塩基とアミノ酸残基の空間的な位置関係から何らかの共通性を見出すこともむずかしい状況である。

我々は，バイオインフォマティクス手法によって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をシスエレメントの文字列中から見出すことに挑戦してきた。

JASPARデータベースでは，ヒトをはじめ

としてシロイヌナズナ，キイロショウジョウバエやマウスなど主要な実験生物種から，120種以上の転写因子と，結合に関わるシスエレメント配列が提供されている。各々の転写因子には，最低でも３種類以上，多いものでは，100種類以上のシスエレメント配列が登録されている。ある転写因子のシスエレメント配列の「ゆらぎ」の代表値を求めたり，「ゆらぎ」の群間の比較を行うために，情報論的シャノンエントロピーや相互情報量を使うことができる。例えば，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ことができる。

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

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。

「CCATATATAG」という配列は，A，T，G，C各塩基の出現確率（P_A, P_T, P_G, P_C）が，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，

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

$$= 1.8464\cdots\cdots$$

となる。

元来，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ようになる。さらに，以下で定義した相互情報量Iを計算すると，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知ることができる。また，相互情報量は，シスエレメント配列Xとシスエレメント配列Yの間で共有されている情報の量であるので，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。

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

ここで，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。

ある転写因子が認識するシスエレメント配列群のすべての組み合わせから，正規化した相互情報量（以下エントロピー進化率と呼ぶ）を計算して頻度分布を作成すると，シスエレメント配列の多様性を数学的に比較することができるようになるのである。

転写因子ごとに，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エントロピー進化率を0.1の階級幅で頻度分布化し，作成した頻度分布間の類似性をもと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実行してみる。

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群がどのようになっているのかを考察することができる。

私の研究室では，124種類の転写因子のシスエレメント配列から，転写因子のクラスタリング，転写因子のもつDNA結合ドメインの種類ごとのクラスタリングや生物種ごとのクラスタリングを実行している。

DNA結合ドメインを比較したクラスタリングでは，類似のDNA結合ドメインがまとまる傾向が見られた。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。しかし６種の生物種（*Arabidopsis thaliana*，*Drosophila melano-*

gaster，*Homo Sapiens*，*Mus musculus*，*Rattus norbegicus*，*Zea mays*）について，生物種別のクラスタ解析を行うと，いずれの生物種においても，すべての種をひとまとめにして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集まる傾向が見られた。

このことから，シスエレメント配列のパターンの進化は転写因子自体の変化よりも，種分化に影響を受ける可能性が示唆される。

⑵ 遺伝子上流領域や遺伝子間領域の配列構造

前節では，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。では，完全長ヒトゲノム配列でシスエレメント配列の存在位置はどうなっているのであろうか？　また，転写制御を受けると思われる遺伝子との相関性はあるのであろうか？

シスエレメントの分布を考える前に，そもそも，遺伝子上流配列（遺伝子の転写開始点より5′側にある領域）の塩基の出現パターンはどうなっているのであろうか？　筆者らは，EnsemblデータベースとENCODEデータベースに登録されたデータから独自のデータセットを抽出して，バイオインフォマティクスにより上記の疑問についても検討してきている。

ヒトの遺伝子配列約30,000件について，その上流配列（2000塩基）を取得して解析してみると，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も分かる。A，T，G，Cの出現確率がほぼ均等である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る。

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と上流配列をアライメント（お互いに類似している位置を探すこと）する。こうしてどのシスエレメント配列がどの上流配列に存在するのかが判るようになる。しかし，その配列がシスエレメントとしての機能を有しているかどうかは分からない。そこで，機械的にアライメントしたものを「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，数種類のシスエレメント様配列が繰り返し存在している場合や，数種のシスエレメント様配列が離散的に見つかる場合など，シスエレメント様配列の存在パターンも，各上流配列でかなりのばらつき（逆に言えば，特異性）があることが分かった。

また，各上流配列の各塩基の出現頻度を用いて，あるシスエレメント配列がその上流配列に見つかる確率の期待値と，先ほどのアライメント結果から計算できる出現確率（事後確率）を比較してみると，ほとんどすべてのシスエレメント配列において，事後確率が期待値の確率よりもはるかに小さく，この２つの確率の差について，有意水準５％における統計的仮説検定で有意差が認められた。すなわち，シスエレメントは，５～20塩基長の短い配列であるため，ゲノム上のいたるところで偶然に見つかる可能性が高いように思われがちであるが，実際には，必要な場所を選んで存在しているように思われるのである。さらに，ヒトゲノム配列の遺伝子間領域19,000件以上を用い，実際に機能しているシスエレメント（稼働シスエレメントと呼ぶ）を考察してみたところ，遺伝子間領域下流側の遺伝子のアノテーション（配列付加情報）には「生物学的プロセス」という単語が特異的に出現していることが示唆された。

ゲノム配列解読のロゼッタストンをより客観的に議論するために，ゲノム科学のための代数学を考案する必要があるのかもしれない。