単純反復配列を考慮した塩基配列データの類似性検定

――スペクトル解析による特徴量の抽出――

Similarity Test of Base Sequence Including Simple Sequence Repeat

――Feature Extraction in Spectral Domain――

井嶋　博　　南山泰宏　　来山顕久
Hiroshi IJIMA　　Yasuhiro MINAMIYAMA　　Akihisa KITAYAMA
（和歌山大学教育学部）

2014年９月30日受理

Abstract

同一種から得られた複数のDNA塩基配列が、全塩基配列の同じ領域のものかどうかを判別する、類似性の検定は遺伝子解析の前処理として重要である。しかし、数個のヌクレオチドが繰返し現れる単純反復配列がゲノム中に広く散在しており、このような反復配列は類似性検定における問題点となっている。これまで、塩基データを配列順に確認する手法が用いられてきた。この手法では対象となる塩基データの１端から塩基情報を走査し、反復配列が確認された場合これ以降の塩基データは切り捨てられる。このため、作業は複雑で、場合によっては有用な情報が削除されることがある。

本研究では、このような類似性検定手法として、従来行われているような塩基配列データそのものを走査的に確認するのではなく、信号処理分野でよく利用されているスペクトル解析法を用いることにし、周波数領域に変換したデータを基に指標を導出するといった新しい手法を提案する。これにより反復配列の有無にかかわらず解析的に相似性検定を行うことが可能となる。また、本研究の有効性を確認するため、実際に取得されているトウガラシの塩基配列データを用い、その相似性検定のための指標を計算した。

The similarity test of DNA sequences in a specific is important as a preprocessing for the genetic analysis. However, the sequence repeats in a whole genome disturbs this test. For this problem, conventionally, the sequence repeats are detected by using the step-by-step scanning and, a detected repeat is used as a marker. Then the sequence before the marker is only employed as data to be analysis. Then the process is complicated and important information in the part of the sequence might be ignored.

In this paper, we propose a new method of the similarity test for using the spectral analysis such that the feature value for the test is derived as the data in the frequency domain. By using this method, the analytic similarity test can be achieved without the influence of the sequence repeats analytically. The efficacy of the proposed method is illustrated by an example of sequences of the hot pepper.

１．はじめに

細胞内のデオキシリボ核酸（DNA）は生物の遺伝情報を担う重要な物質の一つであり、その遺伝情報を読取り解析する研究が盛んにおこなわれている。このDNAは、核酸の最小単位であるヌクレオチドが鎖状に結合したポリヌクレオチドを２本有し、それらが水素結合によってらせん状に形成されている（二重らせん構造）ことでもよく知られている。ヌクレオチドはそれに含まれる有機塩基によって４種類存在し、それぞれAdenine（A）、Guanine（G）、Cytosine（C）、Thymine（T）と呼ばれている。その４種類のヌクレオチドの配列パターンが生物の遺伝情報となり、この配列を単に塩基配列（sequence）と呼ぶことが多い。DNAの持つ遺伝情報は生物種、さらに個体間でも異なる。そのため塩基配列も同様に生物種個体間で差異がある。つまり塩基配列を解析することで種や個体の様々な情報を得ることできる[1]。

塩基配列の中には「反復配列」と呼ばれるものが含まれている。これは同じ配列のヌクレオチドが複数回現れる配列である。その中でも単純反復配列（simple sequence repeat; SSR）は、細胞核やオルガネラのゲノム上に存在する反復配列で、とくに数塩基の単位配列の繰り返しからなるものである。縦列型反復配列（short tandem repeat; STR）あるいはマイクロサテライト（microsatellite）とも呼ばれる。SSRはゲノム中に広く散在しており、遺伝子マーカーとして利用されている。この遺伝子マーカーとは、品種や個体間の遺伝子の塩基配列の異なる部分を利用した目印のことであ

り、特定の形質に連鎖した遺伝子マーカーは遺伝子診断等に利用されている。遺伝子マーカーとして利用されるSSRのことをSSRマーカーと呼ぶ。SSRマーカーを比較することで、種の判別だけでなく、個体の判別を行うこともできる。すでにDNA型判定や品種改良等、多くの分野で利用されている[2,3]。

SSRマーカーをDNA型判定等に応用するためには、１つの生物種において遺伝子地図上の様々な座位のSSRマーカーを開発する必要がある。１座位毎にSSRマーカーに多型があるか否かを検定し、その結果を統合して判定するためである。DNAの塩基配列はプライマーを用いPCR法で調査することが一般的である。この手法では２つのプライマー中の塩基配列データ、400～600程度の文字列が検出される。検出された塩基配列データは全塩基配列のどの領域のものであるかは分からない。つまり検出したSSRを含む塩基配列データがどの領域のものであるか、またそれぞれが異なる領域のものであるか否か、ということを調査する必要がある。その際に行われる操作が「類似性検定」である。そして集められたSSRを含むそれぞれの全塩基配列の異なる領域の塩基配列データからSSRマーカーが開発される。

上段の作業を行う際に問題となっているのが、どのように塩基配列データの類似性を検定していくか、である。通常の塩基配列データと異なりSSRを含む塩基配列データはこの作業が難しい。その理由はSSRの数塩基の単位配列の繰り返しという単純な配列にある。採取した膨大な塩基配列データの中には同様の単位塩基であるSSRを持つデータが多数あり、それらは数十～数百個の塩基が同じ配列を示すこととなる[4]。類似性検定を行ったとしてもSSR領域の一致による検定結果への依存が非常に高くなる。つまり、全塩基配列上の異なった領域の塩基配列データ同士であっても、同じ単位塩基であるSSRを持つものであれば、「類似性が高い」、同じ領域のものであるという検定結果がでてしまう。この問題を解決するため、反復配列を含む塩基配列の新たな類似性検定手法の開発が求められている。

現在一般的な類似性検定手法として「BLAST」というものがある。これは塩基配列データを集積したデータバンクを用意し、ある塩基配列と類似性が高い位置を検索し、またどの程度類似性があるかを調査することができるプログラムである。このプログラムは、反復配列には対応しておらず、反復配列の部分を目印としその前後を調査することで類似性検定を行っている[5]。その作業を行うためには反復配列を見分ける事のできる専門知識を持ちBLASTを使いデータ処理を行うことができなければならない。また非常に手間もかかる。そのため「BLAST」の問題点である反復配列に対するデータ処理の難しさ、手間を無くすことのできる新しい類似性検定手法が求められている。

本研究では、信号処理分野でよく用いられるスペクトル解析手法を塩基配列データの解析に応用させ、新しい類似性検定を提案する。スペクトル解析は、時系列データ配列をその周波数成分のエネルギー配列に一意に変換する手法であり、これにより反復配列に相当する周波数成分が他の配列成分と分離させることが期待できる。

本論文の構成は以下の通りである。２．において塩基配列の数値化手法とスペクトル解析手法を用いた類似性の検定手法を提案する。実際に得られたトウガラシの塩基配列データに対する検定結果を３．で述べ提案した手法の有効性を確認する。４．において本研究の結言を述べる。

２．解析手法の提案

塩基配列の類似性検定のための指標を、スペクトル解析を用いて導出するためには、各塩基の数値化が必要である。これを踏まえ、本研究では以下のような手順での類似性検定手法を提案し、本論文においては手順３および手順４について述べる。

手順１．塩基配列データの取得
手順２．各塩基の数値化
手順３．配列データのスペクトル解析
手順４．検定のための指標の導出
手順５．類似性検定の実施

２．１　塩基配列データの数値化

塩基配列を数値化する手法としては、各塩基を表１のように大きさの等しい複素数に割り当てる手法が提案されている[6,7]。本研究でもその手法を用いることにする。

表１．各塩基の数値化（$i^2=-1$）

塩基	数値
Adenine	$1+i$
Guanine	$-1+i$
Cytosine	$1-i$
Thymine	$-1-i$

この数値化に基づいて、各塩基に対する数値を複素平面上に示したものが図１である。各塩基の数値は、複素平面上の原点からの距離がすべて等しく対等な関係となっていることがわかる。またDNAの二重らせん構造では、塩基AdenineとThymine、またGuanineとCytosineが対となって結合されている。よってこの結合を切り離して取り出される２つの配列は、常に等価であり、この関係を相補鎖と呼ぶ。従来のアルゴリズ

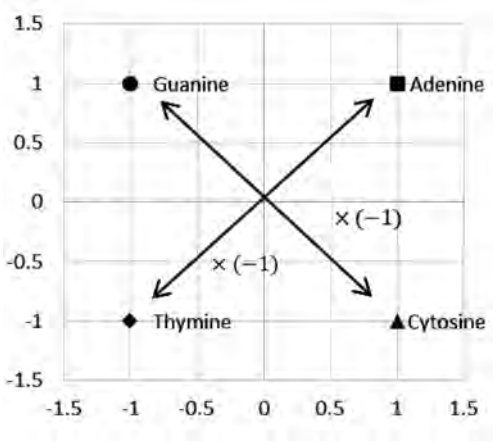

図１　複素平面上における各塩基の値

ムにおいては、この相補鎖を確認するプロセスを別途用意する必要があった。しかし、本研究で提案する上記の数値化においては、各塩基の値に－１を乗ずることで相補鎖の関係にある塩基に対応する値に変換される。またスペクトル解析の際には、値の絶対値操作がされることから、相補鎖の問題は自動的に解決されることになる。

以上により数値化された各塩基の値を、配列順に番号$n=1,2,\cdots,N$を与え

$$x(n)\quad n=1,2,\cdots,N$$

と表記することにする。但しNは配列に含まれる塩基の総数である。つまり、あるj番目の塩基の値は

$$x(j)=\begin{cases}1+i & (\mathrm{A})\\ -1+i & (\mathrm{G})\\ 1-i & (\mathrm{C})\\ -1-i & (\mathrm{T})\end{cases}$$

のいずれかで表現されることになる。

２．２　塩基配列データのスペクトル解析

２．１の処理により数値化された塩基配列データ$x(n)$($n=1,2,\cdots,N$)に対して離散フーリエ変換は次の式で与えられる

$$X(\omega)=\sum_{n=1}^{N}x(n)\,e^{-i\omega n} \qquad (1)$$

但し、ωは正規化角周波数であり、実際には０から２πの範囲を定義域として計算を行う。

(1)式によって得られた$X(\omega)$に対して絶対値を取った

$$P(\omega)=|X(\omega)| \qquad (2)$$

をピリオドグラム(Periodogram)と呼ぶ。このピリオドグラムはデータのスペクトル表現の一つとして用いられるが、データ$x(n)$が不規則な挙動を示す場合、得られたピリオドグラムも不規則な振舞いを示す。そこでデータ$x(j)$に対して一定の区間$[j,j+M]$($j=1,2,\cdots N-M$)で取り出した複数のデータについてピリオドグラムを計算し、それらの期待値をとるウェルチ法(Welch's method)を本研究では用いることにする[8]。

図２はデータ$x(n)$についてピリオドグラムおよびウェルチ法によって得たスペクトルである。ウェルチ法の方がピリオドグラムより滑らかな表現となっていることがわかる。

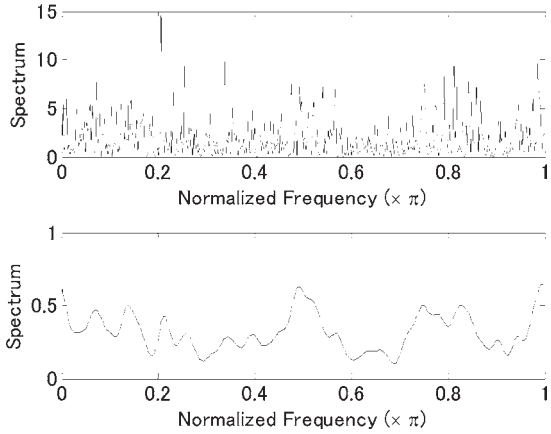

図２　塩基配列データのスペクトルの例(ピリオドグラム(上図)、ウェルチ法(下図)、トウガラシの塩基配列を用いた)

２．３　類似性検定のための指標の計算

前節の手法により得られるスペクトル解析を用い、２つの塩基配列データの類似性の指標y_{12}を次式より得ることにする。

$$y_{12}=\frac{1}{T_1T_2}\int_0^{2\pi}|P_1(\omega)-P_2(\omega)|d\omega \qquad (3)$$

但し、$P_1(\omega)$および$P_2(\omega)$はそれぞれ、各塩基の配列データ$x_1(n)$および$x_2(n)$のスペクトルでありT_1およびT_2は各データの総エネルギー

$$T_\alpha=\int_0^{2\pi}P_\alpha(\omega)d\omega\quad(\alpha=1\text{ or }2)$$

である。(3)式によって得られる指標y_{12}は２つの塩基配列に含まれる各周波数成分の差の合計とみなせることから、２つの配列が同じであればこの指標の値が小さくなり、異なる配列では値が大きくなることが分かる。

３．解析結果

塩基配列データについてはトウガラシのゲノムDNAを由来としたSSRを含む塩基配列データを用意した。類似性があると判定されるべきデータを10、類似性がないと判定されるべきデータを４そのうち２つは全く同じデータを用意した。これらの配列の中から２つを選択し、全ての組合わせ91個について指標を求めた。

ウェルチ法におけるデータの切り出しとして幅50のハミング窓[8]を用い、データの重複幅は10とした。

まず、解析結果の例として類似性があると判断される２つのデータに対するスペクトルの計算結果を図３に示す。もし、同じデータであれば全く同じ曲線を描

くこととなる。つまり、これら２つのスペクトル曲線の形状が似ていれば、類似性が高いといえる。しかし、２つの曲線の形状から直接類似性を評価することは困難であることから、(3)式に示した方法によりその指標を定量化した。

図４はピリオドグラムおよびウェルチ法を用い計算した類似性の指標を、１から91までの組合せ番号に対して示したものである。ピリオドグラムを用いた手法では、類似性が有ると示されるべき組合せの中には、十分値が小さく、正しい判断ができると予想されるものが含まれているが、多くの組合せは、類似性が無いと判断される組合せの数値との差がほとんど見られず誤検定が行われるリスクが高い。一方ウェルチ法を用いた手法では類似性が有ると判断される組合せについては小さな値を持ち、無いと判断される組合せについては大きな値にそれぞれ偏っており、適切なしきい値を設定することが出来れば、本研究で提案した指標は、類似性の検定に有効であると考えられる。さらに相補差の関係にある組合せについても、その他の類似性のある組合せと同様、低い数値を示しており、提案手法の有効性が確認できる。

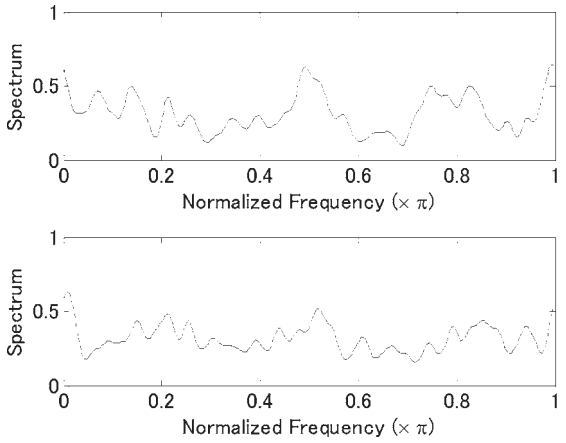

図３　２つの塩基配列データのスペクトル（データB_E02(上図)、データM_D01(下図)）

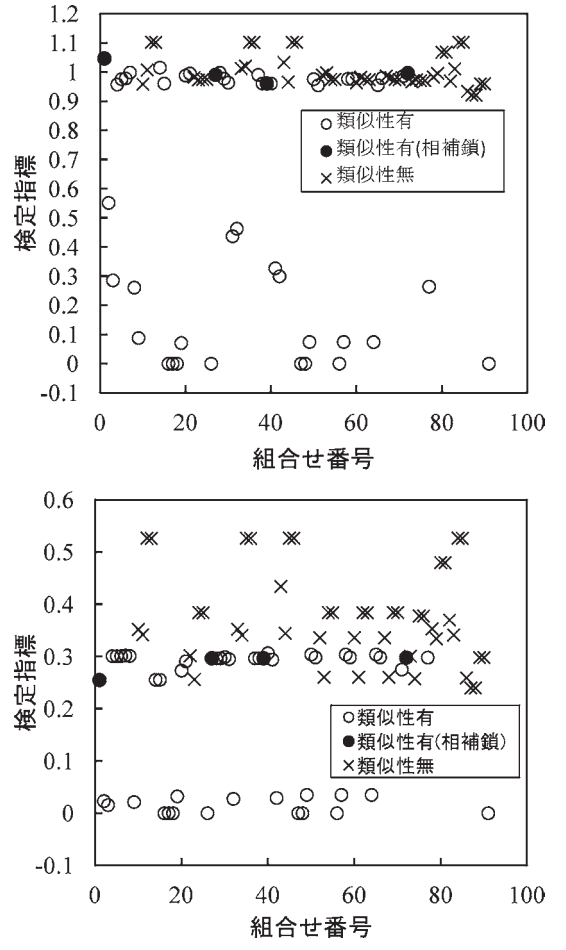

図４　２つの塩基配列の組合せに対する類似性の指標（ピリオドグラム(上図)、ウェルチ法(下図)）

４．まとめ

本研究では、スペクトル解析手法を利用し、反復配列を考慮した塩基配列データの類似性検定手法の開発を目的とした。塩基配列データの数値化としては、複素数を用い、塩基データの配列順を時間軸と見たときの値の変化を信号としてみる手法を用いた。数値化した２つのデータに対してスペクトル解析を実施し、その結果の類似性を定量化する方法を提案し、類似性検定のための指標として用いることにした。

提案した手法の有効性については、実際に取得されているトウガラシの塩基配列データを用いることでその有効性を確認した。その結果、スペクトル解析の一つであるウェルチ法を用いた手法では反復配列を持つ多くのデータで期待される解析結果を得ることができた。また、本研究で提案した手法は、２つの塩基配列のデータ長が異なる場合や配列の抽出箇所がずれている場合においても用いることが可能であることから、今後さらに精度を高めることで、類似性検定手法の発展につながることが期待できる。

参考文献

(1) C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Long-Range Correlations in Nucleotide Sequences, Nature, vol. 356, pp.168-171, 1992.

(2) 平尾他,　SSRマーカーを利用したスギ精英樹のクローン識別, Japanese Forestry Society, pp.202-205, 2006.

(3) 近藤禎二他, SSRマーカーによるスギ在来品種サンブスギ, ニュウカワスギと精英樹との関係解明,第56回日林関東支部論文集, pp.139-140, 2005.

(4) 小笠原他, イネにおいて発現する反復配列の検出と組織間での差異, 育種・作物学会北海道談話会会報, 52, pp.41-42, 2011.

(5) H. Fukuoka, T. Nunome, Y. Minamiyama, I. Kono, N. Namiki and A. Kojima, read2Marker: a data processing tool for microsatellite marker development from a large data set, BioTechniques, vol. 39, pp. 472-474, 2005.

(6) R. F. Voss, Evplution of Long Rance Fractal Correlations and 1/f Noise in DNA Base Sequences, Physical Review Letters, vol. 68, no. 25, pp. 3805-3808, 1992.
(7) 佐藤・吉原・山森・安永, 1/fゆらぎによる生物進化の解析, Memoirs of the Faculty of Eigineering, Miyazaki University, vol. 35, pp. 263-268, 2006.
(8) A. V. Oppenheim and R. W. Schafer, Digital signal processing, Prentice-Hall, Englewood Cliffs, N.J, 1975.