2010 年度第 6 回情報処理学会東北支部研究会(山形大学)
資料番号　10-6-A5-1

バクテリアの塩基配列における文字の含量を用いた解析

山形大学大学院理工学研究科　応用生命システム工学専攻　**小池 公洋、木ノ内 誠**

1.はじめに

近年多くの生物の全ゲノム配列が決定されている。2010 年 3 月現在、バクテリアでは 1400 種以上の全ゲノム配列が決定されている。決定された配列から生命現象を解明するために、情報学（インフォマティクス）の手法や技術が用いられるようになってきている。そのため、情報解析手法の研究と開発が盛んに行われている[1]。

解析手法の 1 つとして GC skew 解析がある。GC skew は 1 本鎖 DNA 分子における G 含量と C 含量の偏りを表わす指標で、(C の個数－G の個数)/(C の個数＋G の個数)の式で表わされる。生物のゲノム配列は A(Adenine)、T(Thymine)、G(Guanine)、C(Cytosine)の 4 種類の塩基からできている。ゲノム全体では A と T、G と C の量はほぼ等しいが、局所的な領域ではその量比に偏りが見られる。原核生物のうち真正細菌の多くの種ではゲノム中で明確にその傾向が逆転する個所が見られ、その個所が複製開始点・複製終結点と一致することが多いことが知られている[2]。

図 1 に、真正細菌の一種であり原核生物の代表的な生物である大腸菌の GC skew 解析を行った結果を示す。G が多い領域、C が多い領域があり、複製開始点・複製終結点と一致している。GC skew という現象が発生する原因には様々な説が考えられており、リーディング鎖とラギング鎖の異なる突然変異確率、コドン使用による変異のバイアスなどによるといわれている。しかし、現象が発生する根本的な理由は未だにわかっていない。また、図 2 に示すパイロコッカス菌のように GC skew を用いた解析では、複製開始点・複製終結点を予測できないバクテリアも多く存在する[2]。

GC skew では G と C の組み合わせで解析を行うことに特徴がある。本研究ではこの組み合わせを敢えて使わず、文字の含量を用いることによって解析を行い、全ゲノム配列から生物学的な情報の抽出を試みる。

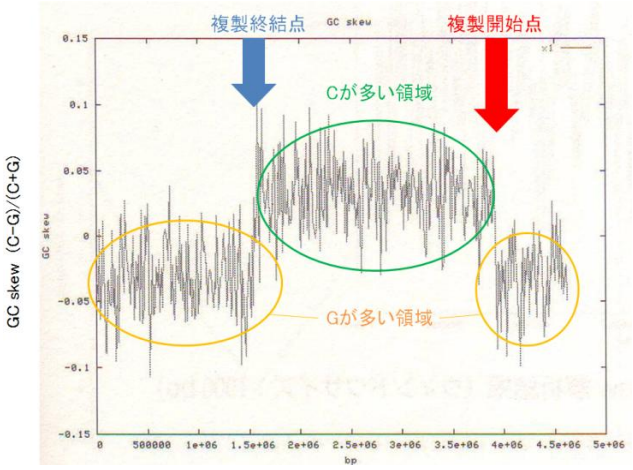


図 1　*Escherichia coli* str. K-12 substr. W3110 (大腸菌)の GC skew 解析(文献[2]85 ページを改編)

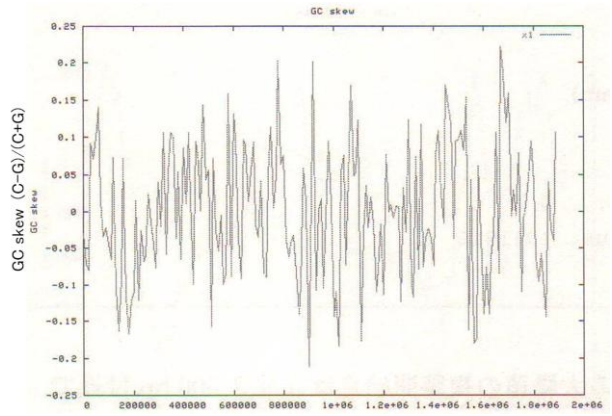


図 2　*Pyrococcus abyssi* GE5 (パイロコッカス菌)の GC skew 解析(文献[2]85 ページから引用)

2.方法

全塩基配列の中に A・T・G・C の文字の量がどれだけあるか調べる。 文字の実際の累積値と平均的な増加率との差を取ることでグラフを描く。グラフは横軸に塩基数、縦軸に文字の累積数をとる。実際の累積値と平均的な増加率との差をとることで、塩基配列を 1 次元グラフで表現し、そのグラフから生物学的な情報を読み取る。

全ゲノム配列データは NCBI[3]からコンプリートゲノムファイルを使用した。

3.結果と考察

3.1　1 文字の含量を用いた解析

大腸菌に対して解析を行った結果を図 3 に示す。グラフを見ると、G と C の値が対称的であり、GC skew と同じ特徴がはっきりと表れている。

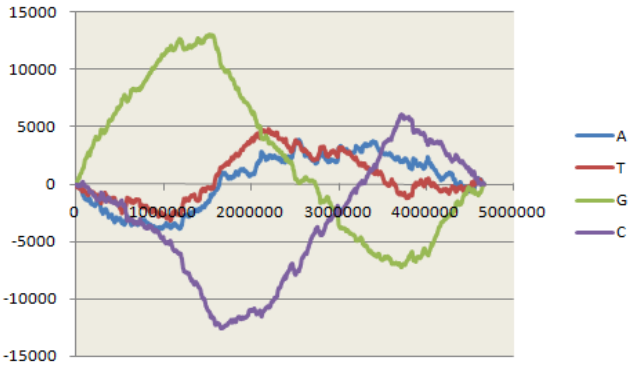
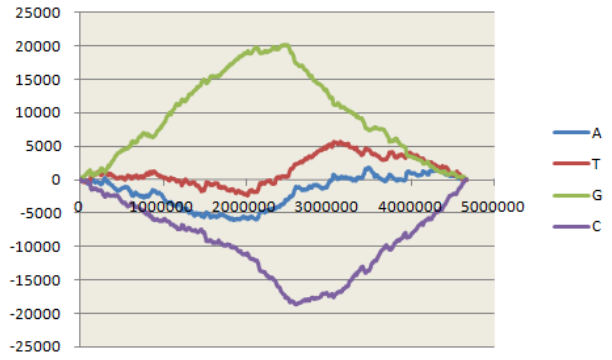
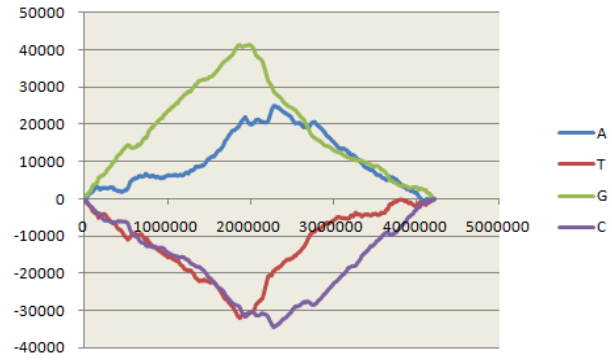


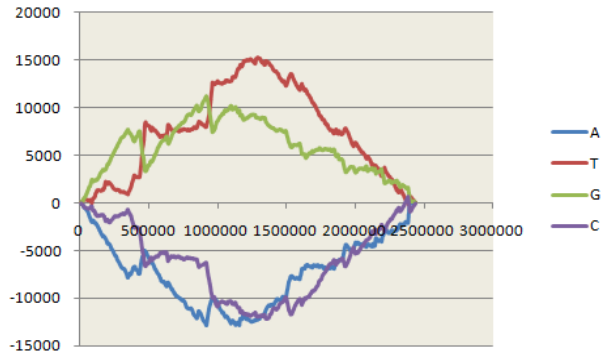
図 3　*E.coli*（大腸菌）の 1 文字の含量を用いた解析



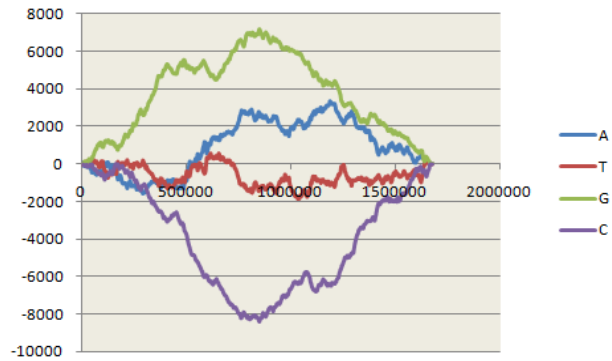
(a) *E.coli*（大腸菌）



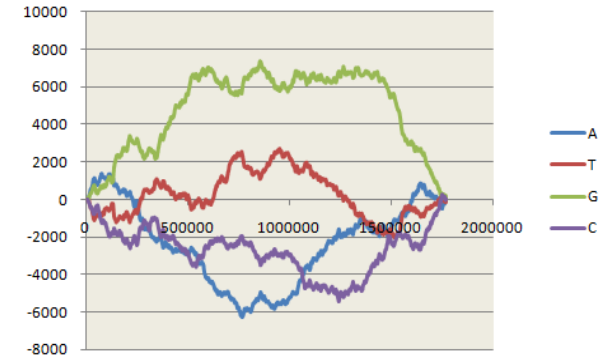
(b) *Bacillus subtilis* subsp. subtilis str 168 (枯草菌)



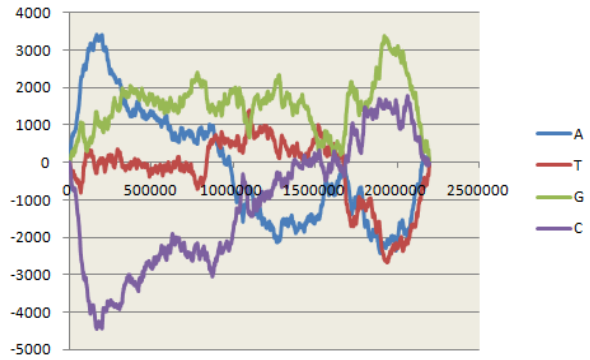
(c) *Synechococcus* sp. WH 8102 (シアノバクテリア)



(d) *Helicobacter pylori* 26695(ピロリ菌)



(e) *P.abyssi* (パイロコッカス菌)



(f) *Metallosphaera sedula* DSM 5348

図 4　1 文字の含量を用いた解析

図 3 では、登録配列の開始点を変えずにそのまま読み込み、グラフ化を行っている。しかし、ほとんどのバクテリアではゲノムが環状であり、登録配列の開始点は定められておらず、その基準は生物種によって異なる。そこで本研究では G の値が最小となる位置を読み込みの開始点とし、グラフを描く際には各文字で開始点がグラフの原点となるようにした。図 4(a)に G を基準として図 3 を書き換えた例を示す。このように開始点を統一することで、各バクテリアのグラフがより比較しやすいものとなる。

1 文字の含量を用いた解析をバクテリア 895 種に対して行った。これらをグラフの形から、いくつかのパターンに分類した。図 4(a)～(f)にそれぞれのパターンの代表的なバクテリアを示す。

図 4(a)において A と T は同期し、さらに G と C は対称的な値をとっている。このような形のグラフを本研究では「大腸菌型」とする。大腸菌型となるバクテリアは 895 種中 101 種であり、解析を行ったバクテリアの約 12%であった。

図 4(b)において A と G は同期し、また T と C が同期している。さらに A と T（G と C）は対称的な値をとっている。このような形のグラフを「枯草菌型」とする。枯草菌型となるバクテリアは 895 種中 144 種あり、解析を行ったバクテリアの約 16%であった。

図 4(c)において A と C は同期し、また T と G が同期している。さらに A と T（G と C）は対称的な値をとっている。このような形のグラフを「シアノバクテリア型」とする。シアノバクテリア型となるバクテリアは 895 種中 234 種であり、解析を行ったバクテリアの約 26%であった。

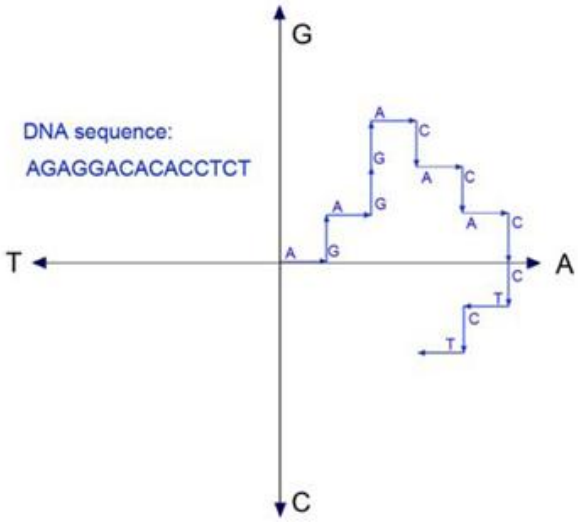


図 5　DNA Walk の模式図([3]から引用)

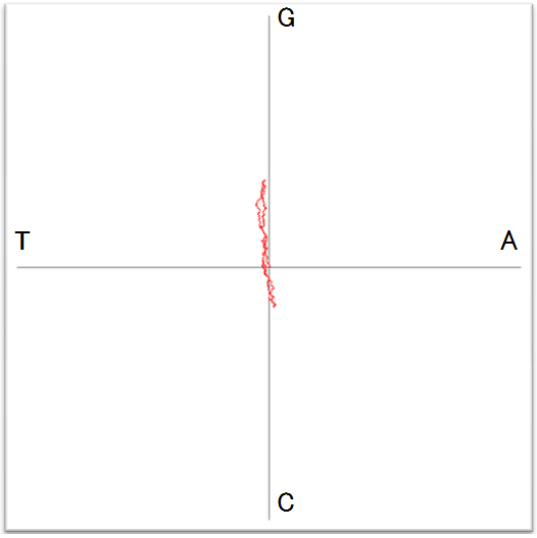


図 6　*E.coli*(大腸菌)の DNA Walk

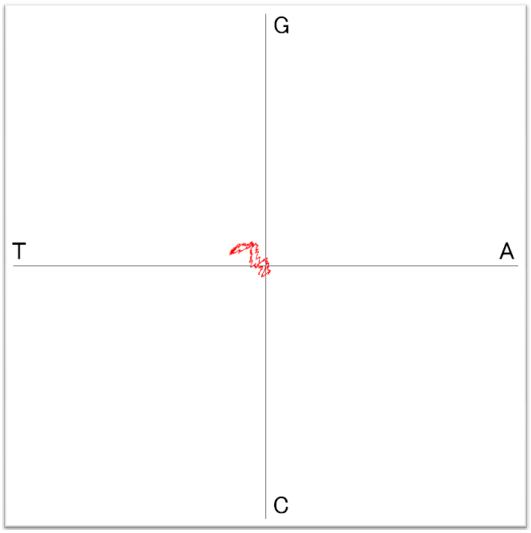


図 7　*P.abyssi* (パイロコッカス菌)の DNA Walk

図 4(d)において G と C が対称的な値をとっているが、A や T は他の文字との同期がみられない。このような形のグラフを「ピロリ菌型」とする。ピロリ菌型となるバクテリアは 895 種中 174 種であり、解析を行ったバクテリアの約 19%であった。

図 4(e)または(f)のように、図 4(a)～(d)のどのグラフのパターンにも分類出来ないものを「その他」とする。その他のバクテリアは 895 種中 242 種であり、解析を行ったバクテリアの約 27%であった。

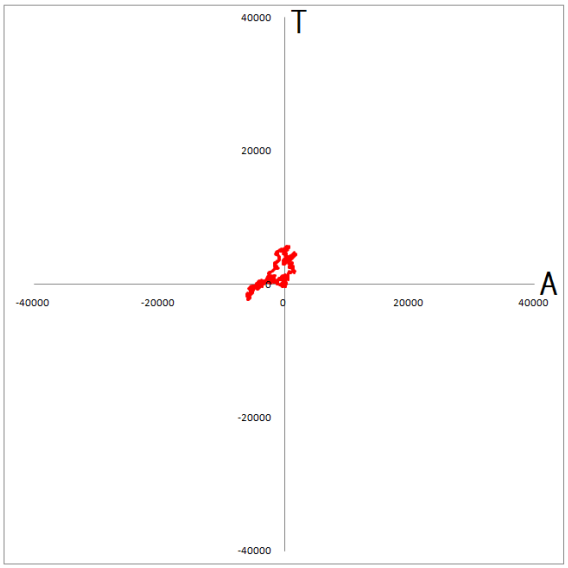
3.2　塩基配列の 2 次元の可視化

ここまでの解析ではバクテリアに対して 1 文字の含量を用いた解析という今までにない解析方法で、全ゲノム配列からの有用な生物学情報の抽出を試みた。その結果全ゲノム配列をグラフで表すことによって、グラフの形で分類できるというバクテリアの特徴を発見した。しかしこれらの解析から、より文字と文字の関係性を明らかにすることが重要であると考え、新たな解析方法を提案する。

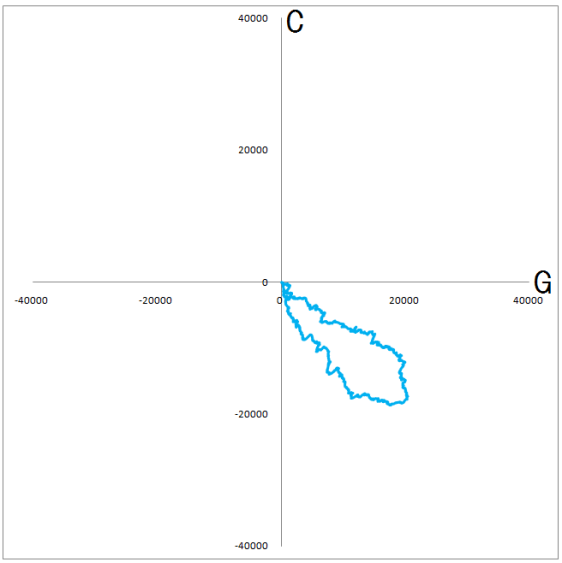
ここまで解析に用いた特定文字の累積値と平均的な増加率の差を取る方法は、塩基配列の傾向を強く表わすので継続して用いる。さらに、より文字と文字の関係を明らかにする方法として、塩基配列を 2 次元で可視化する表現方法を提案する。

塩基配列を 2 次元で可視化において、先行する解析方法として、DNA Walk がある[4]。DNA Walk は図 5 に示すように、平面の(0,0) の座標を始点とし、塩基配列に応じて A で右、T で左、G で上、C で下に 1 目盛りずつ進み軌跡を描くことによって、塩基配列を 2 次元で表現する。

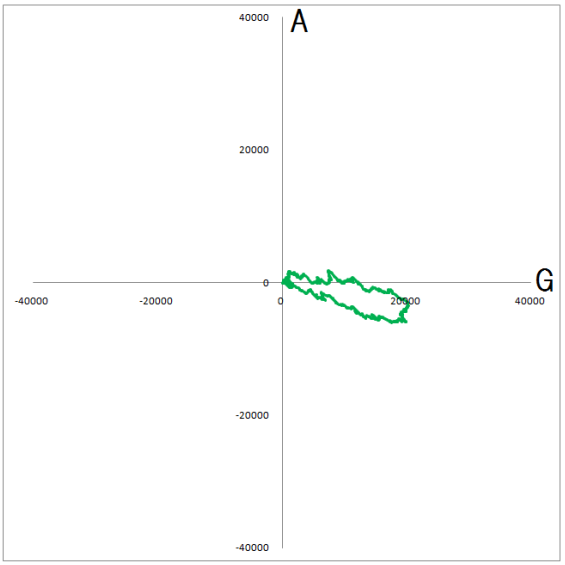
図 6 および図 7 に、実際に全塩基配列に対して DNA Walk を行った結果を示す。図 6 に示す大腸菌は GC skew に強く傾向が見られる生物種なので、軌跡は複製開始点から上に進み、複製終結点から下に戻ってくる。図 7 のパイロコッカス菌は図 2 からもわかるように GC skew に傾向が見られない生物種なので、軌跡ははっきりとした形にはならない。



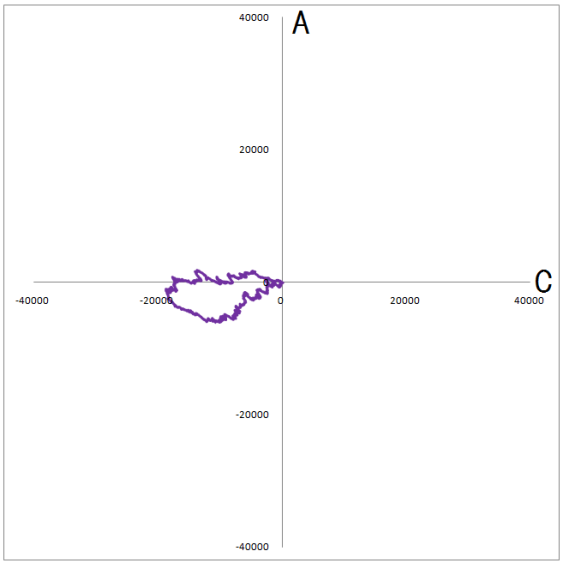
(a)　AT グラフ



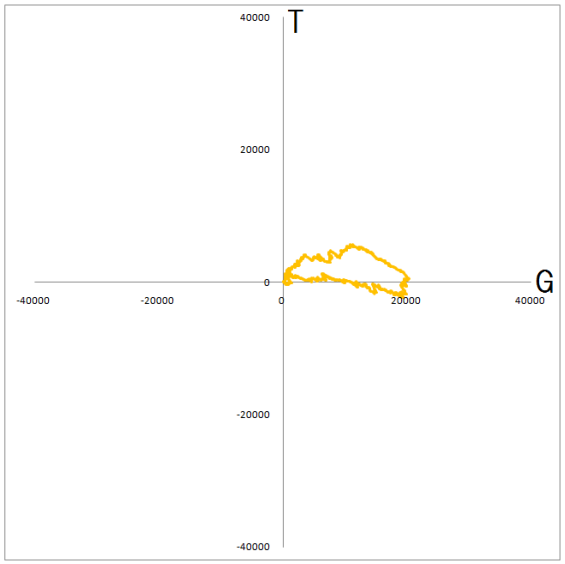
(b)　GC グラフ



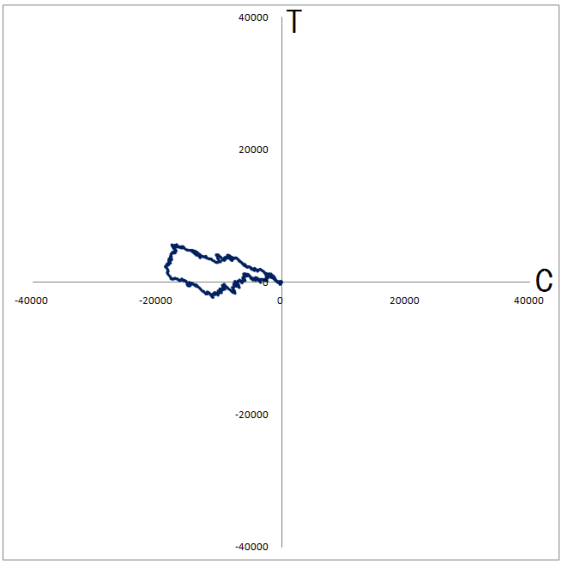
(c)　GA グラフ



(d)　CA グラフ



(a)　GT グラフ



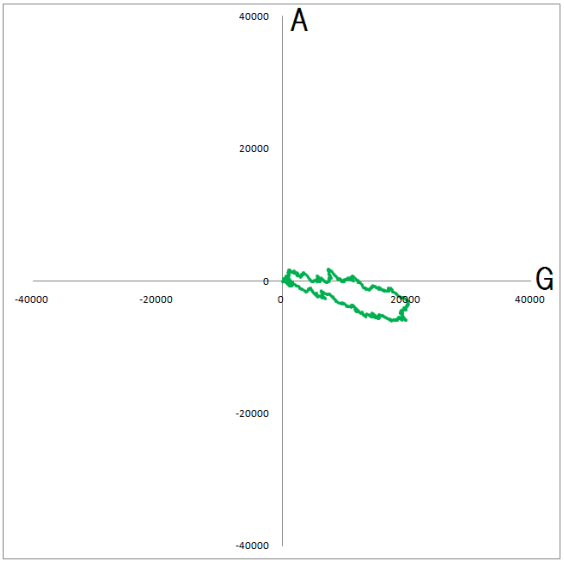
(b)　CT グラフ

(a)　図 7　*E.coli*(大腸菌)における 2 次元の可視化

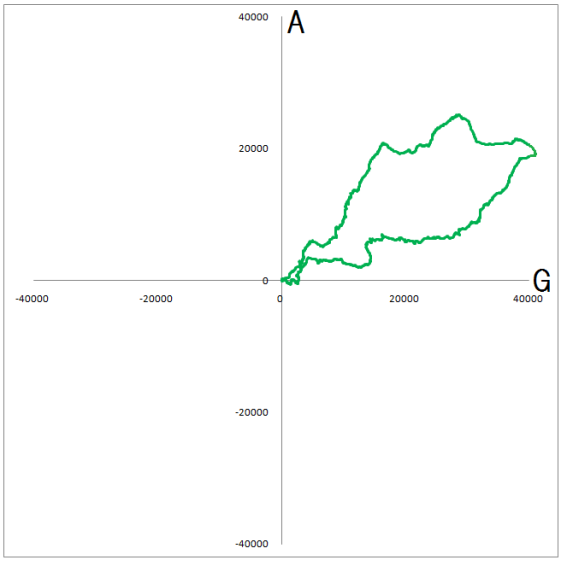
本研究で提案する 2 次元の可視化では、GC および AT を対にせず、特定文字の累積値と平均的な増加率の差を利用する。この値は 4 つの文字（A, T, G, C）のそれぞれに対してあるので、その中から 2 つを選びグラフの軸にする事によって、塩基配列を 2 次元で表現することができる。この方法で描いたグラフを図 7 に示す。2 つの文字の組み合わせで 6 通りのグラフ（AT, GC, GA, CA, GT, CT）が描ける。

この解析方法の利点は平均的な文字の増加率を求める事によって、文字を対にして考える必要がなくなり、様々な文字の組み合わせに対して、2 次元で塩基配列を表現出来るところである。

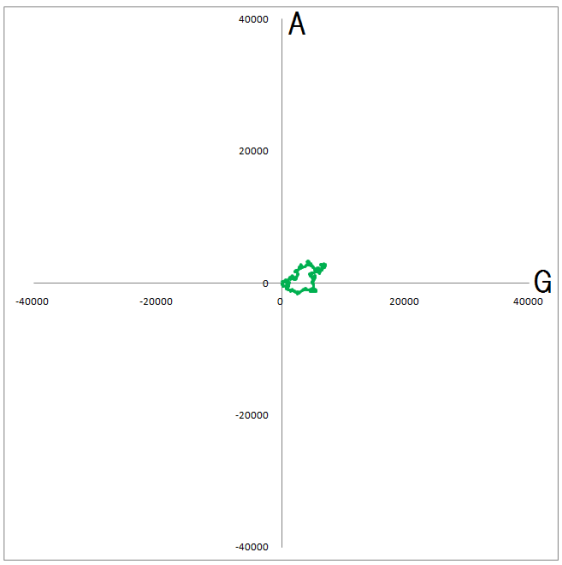
図 8 に 4 種のバクテリアの GA グラフを示す。図の(a)～(d)をみるとわかるように生物種によってグラフの形に違いが大きく表れる。GA グラフにおいて、それぞれの生物種について原点から一番遠い点をプロットし分布図を作ることによって、特徴の発見を試みた。図 9 に結果を示す。グラフに示す赤い点は Firmicutes 門に属するバクテリアである。他の生物種と比べて分布がグラフの第 1 象限に偏っている事がわかる。2 次元の GA グラフから分布図を作成する事によって、Firmicutes の特徴的な分布を発見できた。



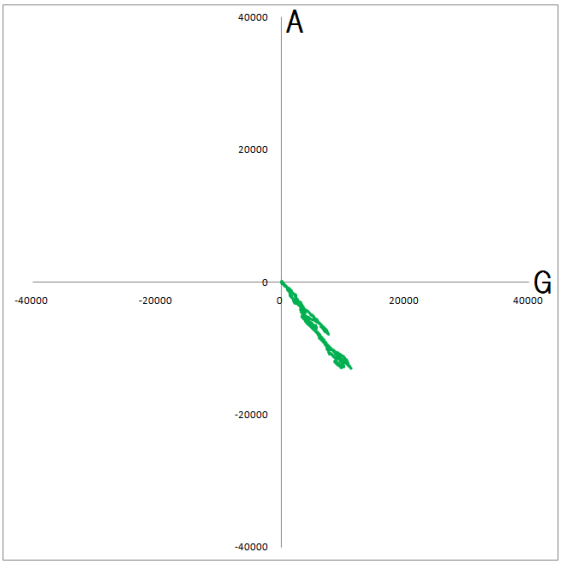
(a)　*E.coli*（大腸菌）



(b)　*B. subtilis* (枯草菌)



(c)　*H.pylori* (ピロリ菌)



(d)　*Synechococcu* sp. WH 8102 (シアノバクテリア)

図 8　GA グラフの例

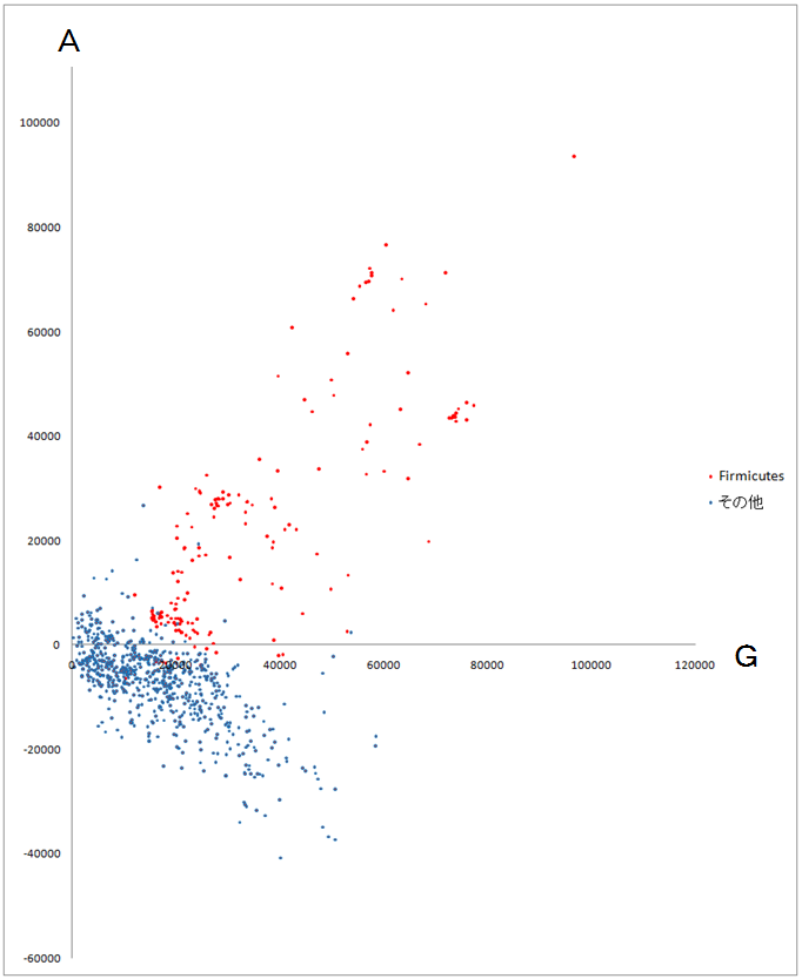


図 9　Firmicutes とその他の生物種との GA グラフを用いた分布図

4.むすび

本研究ではバクテリアに対して、特定の文字の累積数と平均的な増加率の差を取る解析方法により 1 次元で塩基配列を表現し、全ゲノム配列からの有用な生物学情報の抽出を試みた。

全ゲノム配列をグラフで表すことによって、グラフの形で分類できるというバクテリアの特徴を発見した。さらに、塩基配列を 2 次元で表現する事によって、Firmicutes の他のバクテリアにはない特徴を発見した。1 次元の解析でも Firmicutes は図 4 の(b)に示す「枯草菌型」として特徴を表わしていた。ただし、1 次元の解析ではグラフの形を判断することにより分類を行っていた。一方、2 次元の可視化では図 9 に示す分布図のように、より客観的に解析結果を見ることが出来る。

参考文献

[1] http://www.chart.co.jp/subject/joho/inet/inet09/inet09-1.pdf.
[2] 片山敏明 他, オープンソースで学ぶバイオインフォマティクス, 東京電機大学出版局, 2009.
[3] NCBI, http://www.ncbi.nlm.nih.gov/.
[4] Poptsova MS, Larionov SA, Ryadchenko EV, Rybalko SD, Zakharov IA, Loskutov A, Hidden chromosome symmetry: *In silico* transformation reveals symmetry in 2D DNA walk trajectories of 671 chromosomes, PLoS One, 4(7): e6396, 2009.