周波数解析の手法を用いたホモロジー検索の一検討

A Study on Homology Search Using Frequency Analysis Technique

松村 岳敏 甲藤 二郎
Taketoshi Matsumura Jiro Katto

早稲田大学大学院 理工学研究科
Graduate School of Science and Engineering, Waseda University

１．まえがき

従来，ゲノム配列のホモロジー検索は，BLAST や FASTA といった文字列検索ベースでの実装が中心であった．しかし，近年文献[1]にあるような周波数解析の手法を用いたゲノム配列の解析手法が発表されている．本研究では，文献[1]を参考に，周波数解析の手法を用いてホモロジー検索を実装し，検証してみた．

２．今回の相違点

ホモロジー検索とは，相同性検索とも呼ばれ，性質が既知の配列と似ている部位を検索して，もし似ている部位が存在すれば同じ性質を持っているだろうと推定する手法である．

従来のゲノム配列でのホモロジー検索は，ゲノム配列が文字列配列で表されることから，文字列検索を元にした手法である BLAST や FASTA といった手法が主流であった．これは単純な逐次パターンマッチングをかけていきながら類似度を計算していき，最終的に類似度が一番高い部分を残すという手法である．

本研究では，文献[1]にある手法で，核酸配列として表記されているゲノム配列の次数を減らして DFT をかける．まず，以下のような変換を考える．

$$X_r[n]=\frac{\sqrt{2}}{3}(2u_T[n]-u_C[n]-u_G[n])$$

$$X_g[n]=\frac{\sqrt{6}}{3}(u_C[n]-u_G[n])$$

$$X_b[n]=\frac{1}{3}(3u_A[n]-u_T[n]-u_C[n]-u_G[n])$$

この式で，もともとの A,G,C,T で表される核酸配列を r,g,b の 3 変数に置き換えている．さらに，これらの 3 式に DFT をかける．この実数項と虚数項の 2 乗和である Power をとり，その Power の誤差の２乗和を比較することによって構成比が似ているかを検出する．これで構成比が似ていると判定されたら，改めて文字列ベースのパターンマッチングを行い類似判定を行う．

この提案手法の利点は，加減演算だけでなく，積和演算も含まれるというところにある．これによって DSP などのアーキテクチャ上ではより高速な実装が可能になる．

３．実験

ゲノム配列のデータベースである GenBank より，ヒトβ，ヒトζ，ウマα，サケα，ラットβと，5種類のヘモグロビンのデータを取り出して比較してみた．探索する配列として，

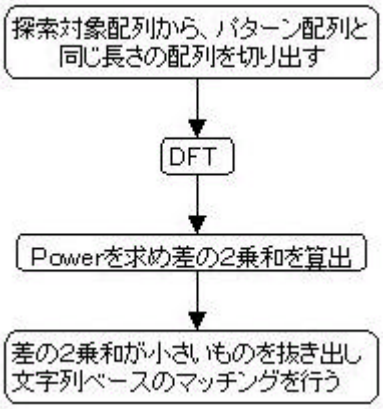

図１ 提案手法のフローチャート

ヒトのヘモグロビンβのうち，鎌形赤血球貧血を引き起こす遺伝子部分の先頭 60 塩基分を用いた．参考としてこの異常ヒトβでも実験してみた．

表１ 実験結果

配列	ＢＬＡＳＴ	提案手法
異常ヒトβ	正確に検知	正確に検知
正常ヒトβ	類似部を検出	類似部を検出
ヒトζ	検出せず	検出せず
ウマα	類似部を検出	違う場所が第１候補(第３候補で検出)
サケα	検出せず	検出せず
ラットβ	類似部を検出	類似部を検出

以上の結果より，実験結果としてはそれほど従来手法に比べて精度で劣るわけではなさそうである．

４．まとめ

本研究によって，従来の手法と比較して全く新しい方式でのホモロジー検索の実装とその有効性が確認できた．今後は更なる高速化と精度の向上が目標である．

参考文献

[1] Dimitris Anastassiou, BIOINFOMATICS Vol.19, no.121073-1081, 2000