卒業論文要旨

スコアマトリックスの統計処理を用いたアミノ酸配列のホモロジー検索

(知能情報システム学)　浦野静香

１．緒言

近年、ゲノムや cDNA の大規模な配列決定が行われ、その成果である大量の塩基配列およびアミノ酸配列データが、公共のデータベースに蓄積されている。配列情報から遺伝子機能を予測する試みはすでに多数行われている。その一方で、予測はやはり予測に過ぎず、最終的には実験的な検証が必要であることも事実である。機能予測を高精度に行うことができれば、機能実証のために必要な実験を、より効率的に行うことができる。

アミノ酸置換は構造的・機能的制約の中で保存的に生じる。アミノ酸置換が生じるときには、物理化学的に類似したアミノ酸に変化しやすい。この変化しやすさを得点化したものが、スコアマトリックスである。多数のスコアマトリックスが提案されており、その作成法もいく通りかある。そして、具体的なアミノ酸配列に対してどのスコアマトリックスを用いてホモロジー検索を行うかは、経験的に決められている。

本研究では、多数のスコアマトリックスを統計処理したものをホモロジー検索に用いることにより、これまで見落とされていた類似性を見出せる手法の開発を行った。なおホモロジー検索プログラムとしては、現時点で最も検出感度の高い方法とされている PSI(Position Specific Iterated)-BLAST(Basic Local Alignment Search Tool)を用いた。

２．方法

(１)開発環境

OS は Microsoft 社の Windows XP を用いた。プログラミング言語として、Microsoft 社の Visual C++ 6.0 を用いた。そして、パソコンは DELL OPTIPLEX GX260(CPU：Pntium４ 2.53GHz　メモリ：512MB)を使用した。

(２)PSI-BLAST プログラムとデータベースの入手

PSI-BLAST はホモロジー検索プログラムとしては、実質的に世界標準となっている BLAST の拡張版で、BLAST をギャップを取り扱えるように改良し、更に問い合わせ配列とデータベースに共にアミノ酸を用い、BLAST を数回繰り返して行うことにより一度の BLAST 実行ではヒットしなかったような、別の関連した配列を見つけ出すことができるプログラムである。

BLAST の開発元である米国 NCBI(National Center for Biotechnology Information)で開発された。では、ホームページ上[1]にてフリーでプログラム及びデータベースを提供しているので、そこから PSI-BLAST を含む BLAST プログラムとデータベースをダウンロードした。

(３)スコアマトリックスの統計処理

スコアマトリックスはデータベース検索において、各アラインメントに従って配列間の類似度計算するために用いられており、一般的に PAM(Point-Accepted Mutation)と BLOSUM(Blocs Substitution Matrix)の 2 系統がよく使われている。

BLOSUM シリーズは、約 2000 の保存されたアミノ酸パターン(ブロック)の大規模な集合の中で観察されるアミノ酸置換に基づいている。付加する番号は一致と見なすアミノ酸のパーセンテージを表している。

PAM シリーズは進化の間に相同なタンパク質配列中の１つのアミノ酸から別のアミノ酸に変化する確率をリストにまとめたものである。PAM とは進化距離の単位のことで、配列全体の１パーセントに点突然変異が起こる場合が 1PAM である。付加する番号はこの進化距離を意味し

ている。

研究で使用したスコアマトリックスは、NCBI の Web サイトで PSI-BLAST 上で使用されていた BLOSUM45,62,80 と PAM30,70 の 5 つである。

①スコアアマトリックスの平均と標準偏差を求め、各スコアマトリックスが平均値０、標準偏差１となるように変換した。以下ではこの処理をスコアマトリックスの標準化と呼ぶ。

②標準化した５つのスコアマトリックスの各値の平均値と標準偏差を求めた。

③②の各平均値と標準偏差をもとに正規分布に従う乱数を発生させて、各値を乱数で与えたスコアマトリックスを作成した。そして、この処理で作られる多数のスコアマトリックスを用いて、PSI-BLAST でホモロジー検索を行った。なお、この処理ができるように、PSI-BLAST プログラムを拡張した。

３．結果

Sequences producing significant alignments:	Score	E-Value
ref|NP_031968.1| erythropoietin [Mus musculus] >gi|119528|sp|P07...	327	6e-089
pir||A24902 erythropoietin precursor - mouse	326	1e-088
ref|NP_058697.1| erythropoietin [Rattus norvegicus] >gi|232064|s...	314	5e-085
gb|AAA41126.1| erythropoietin	307	5e-083
ref|NP_776334.1| erythropoietin [Bos taurus] >gi|1352379|sp|P486...	282	2e-075
sp|P07865|EPO_MACFA Erythropoietin precursor >gi|86625|pir||JQ01...	268	3e-071
sp|P01588|EPO_HUMAN Erythropoietin precursor (Epoetin) >gi|69711...	267	6e-071
sp|P33709|EPO_SHEEP Erythropoietin precursor >gi|2144695|pir||I4...	267	9e-071
emb|CAA26095.1| erythropoietin [Homo sapiens]	266	1e-070
sp|Q28513|EPO_MACMU Erythropoietin precursor >gi|2144694|pir||I8...	266	1e-070
pir||JC7699 erythropoietin - rabbit >gi|11527285|gb|AAG36961.1| ...	264	5e-070
gb|AAG36962.1| erythropoietin [Oryctolagus cuniculus]	264	5e-070
ref|NP_000790.1| erythropoietin [Homo sapiens] >gi|31230|emb|CAA...	263	1e-069
emb|CAB96416.1| erythropoietin [Sus scrofa] >gi|8920357|emb|CAB9...	262	3e-069
sp|P49157|EPO_PIG Erythropoietin precursor >gi|2136498|pir||I465...	261	6e-069

図.1 ecoli のホモロジー検索結果

統計処理を行ったスコアマトリックスを用いてホモロジー検索を行った結果を図 1 に示す。

網掛け部分が、NCBI のウェブサイトにあるスコアマトリックスを直接用いた検索結果と比較した場合に、スコアの値が変わり、順位が入れ替わるという変化があった部分である。

４．結言

統計処理を行ったスコアマトリックスをホモロジー検索に用いることで、これまでに得られていない検索結果を得られることが確認できた。しかし本研究で統計処理を行ったマトリックスの数は 5 つと少ないので、今後より多くのスコアマトリックスを扱えるようにし、使用できるマトリックスを任意に選択できるようにプログラムを拡張していく必要がある。また、ホモロジー検索プログラムは PSI-BLAST だけではなく、目的別に数多くのものがあるので、それらへの応用なども検討する必要がある。本法を用いることにより、スコアマトリックスを選択することなく、統計処理に用いたスコアマトリックスの情報を統合した結果が得られる。

[参考文献]

[1] http://www.ncbi.nlm.nih.gov