ゲノムに関する基礎学習

永松　秀人

1　研究課題

昨年よりの研究課題であった，GA を用いた画像認識に関連して，GA を用いたパターン認識に関する文献の調査を行った．また，GA によるゲノムデータのマッチングを行うにあたり，基礎学習としてゲノム情報に関する調査を行った．

2　文献調査

GA を用いたパターン認識に関する文献調査を行った．以下に調査した文献の概要を示す．

- 安永守利　他，”Kernel Optimization in Pattern Recognition Using a Genetic Algorithm ”，GECCO 2000

  パターン認識処理の一手法であるカーネルマッチングにおいて，識別関数を作成するための核関数のカーネルのサイズをサンプルパターン毎に可変とする手法である．このカーネルサイズを GA により適応的に決定する．

- ”遺伝的プログラミングを用いた画像認識アルゴリズムの自動生成 ”，奈良先端化学技術大学院大学修士論文

  進化的計算の一手法である遺伝的プログラミングを用いて，識別に最適な前処理，特徴量抽出のオペレータを選択する．識別部に関しては，最近傍決定測を用いている．

3　ゲノム情報の基礎

3.1　ゲノム

ゲノムとは，各生物の持つ一揃いの遺伝情報の総体を表す抽象的な概念である．人間の場合，ゲノムは染色体の約半分に書かれている遺伝情報のことである．

ゲノムの実体は DNA であり，DNA は，A（アデニン），C（シトシン），G（グアニン），T（チミン）の 4 種類の塩基分子が糖やリン酸を介して連なってできた高分子である．この塩基の連なりは鎖と呼ばれ，4 種類の文字の配列として表現される．

3.2　遺伝子

DNA に書かれている情報は生物の設計図にすぎず，生命活動のための機能的な部品は，DNA の情報を元にしてタンパク質を合成することによって作られる．しかしながら，DNA 上の情報のすべてが部品を定義しているわけではなく，定義箇所はごく一部である．遺伝子とは，ゲノムの中で部品を定義している箇所のことであり，RNA 分子やタンパク質に翻訳される部分を指す．ヒトの場合，ゲノムサイズ（ゲノムの文字列の長さ）は 30 億あるのに対し，遺伝子の数は約 3 万程度である．

3.3　タンパク質

遺伝子は，タンパク質に翻訳されて機能を発現する．まず，DNA 配列情報の一部が mRNA にコピーされる（mRNA への転写の際には T は U（ウラシル）に置換される）．次に，この mRNA 上の文字列が 3 文字分を単位としてアミノ酸 1 つに翻訳され，これによってアミノ酸の連なりが生成される．Fig. 1 にその模式図を示す．そして，これらのアミノ酸が折りたたまれてタンパク質ができる．

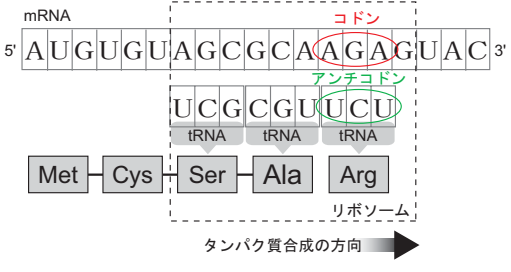

Fig. 1　タンパク質の合成

3.4　SNP

DNA を構成する個々の文字配列は，99.9%が一致し，残りの 0.1%がそれぞれの人で異なる．この 0.1%の文字配列の違いが人間の多様性を生みだすことになる．文字配列の違いは，文字の欠落や挿入などの種類があるが，約 90%が，ある文字が別の文字に置き換わる SNP（スニップ）と呼ばれるものである．この SNP を特定することが病理の解明などにつながる．

4　翌月への課題

ゲノムデータのマッチングには，BLAST（Basic Local Alignment Search Tool）と呼ばれる，DNA シークエンスを，公的データベース全体と比較するために用意されてたシークエンスアライメントプログラムがある．今後は，BLAST の調査と，引き続き GA を用いたパターン認識の文献調査を行う．