J04

自己組織化マップを用いた Gタンパク質共役型受容体の分類

（神奈川大理情報科学科）○中山 堯*、森 昭人
（神奈川大理生物科学科）大瀧丈二、山本晴彦

1. はじめに

G タンパク質共役型受容体（GPCR）のファミリーをそのアミノ酸配列データから予測するシステムを開発した。このシステムは自己組織化マップ［1］を用いて実装したもので、アラインメント済み配列データとアラインメントなし配列データの両方に対応しており、いずれも高い分類精度を与えることが示された。 GPCR は近年、創薬ターゲットとしての有用性が注目されており、多くの Orphan GPCR あるいは新規 GPCR を高精度で分類予測するシステムが手軽に利用できることは、この分野に強力なツールを提供することになると期待される。

２． GPCR の表現

GPCR データは、アラインメント済みのものとアラインメントなしのものが提供されている。アラインメント済みのデータは、長さがそろっているのでそのままベクトル表現できるが、アラインメントなしの生データは様々の長さの配列データであるので、それをベクトルとして表現することがまず問題となる。すなわち、自己組織化マップ(SOM)によって GPCR データを分類するためにはこれをある次元数の特徴ベクトルに変換する必要がある。本システムでは、アラインメントなしのデータについてはアミノ酸の５つの z–値による物理化学的特性値による配列表現から得られる自己共分散・相互共分散の行列を用いた[2]。実際に用いた自己共分散 $c_{d,d}$ と相互共分散 c_{d_1,d_2} の定義を下に示す：

$$c_{d,d}(l)=\sum_{i=1}^{n-l}\frac{(v_{d,i}-\bar{v}_d)\cdot(v_{d,i+l}-\bar{v}_d)}{(n-l)^p}$$

$$c_{d_1\neq d_2}(l)=\sum_{i=1}^{n-l}\frac{(v_{d_1,i}-\bar{v}_{d_1})\cdot(v_{d_2,i+l}-\bar{v}_{d_2})}{(n-l)^p}$$

ここで、l はラグ、n は配列長、$v_{j,k}$ は配列の k 番目アミノ酸の j 番目の特性 z_j の値、p は正規化次数を表している。 $c_{d,d}$ と c_{d_1,d_2} は l の関数であり、自己共分散・相互共分散行列の要素を与える。また、アラインメント済みのデータについては、z–値を割り当てる方法とアミノ酸の記号をそのまま用いる方法の２通りを検討したが、ここではアラインメントなしの場合について述べる。

３． 実験

3.1 GPCR データ

GPCRDB で公開されている 2004 年 2 月現在の Class A データの総数は 4297 でその内 Orphan が 181 である。Orphan を含む場合と含まない場合についてデータセットを訓練データとテストデータに分割した。分割は SOM の学習ごとにランダムにデータを選択して行うが、データの選択はファミリ毎に行い、各ファミリがほぼ２分されるように調整している。

3.2 SOM の学習

上記の訓練データを用いて SOM の学習を行い、その結果を用いてテストデータおよび Orphan データの分類を行った。学習の条件を以下に示す：

(1) SOM の構成：50×50 の正方格子（ニューロン数は 2500）
(2) 学習方式：バッチ SOM を用いて、収束するまで学習させる。
(3) 学習パラメータ：近傍は円または円の一部とし、近傍半径は初期値が 50 で繰り返し毎に 1 ずつ減少させる。近傍半径 0 の場合は着目ニューロン 1 個だけで近傍が構成される。
(4) 訓練データの選択：バッチ方式なので個々のデータの選択順序は問題とならない。
(5) 入力データの生成：最大ラグを 10ー100, p = 0.0, 0.5, 1.0, 2.0 の 4 通りの組み合わせについて、z-値に関する自己共分散と相互共分散から入力データを作成した。

3.3 学習結果

最大ラグを 10、正規化次数を 1.0 とした場合の SOM 学習結果を図 1 に示す。マップのニューロンは、GPCR の各ファミリに別々の色を対応させて、各ニューロンに対して最良一致する入力データ（訓練データ）の所属ファミリの個数をカウントし、個数が最大となるファミリの色を表示している。実際には、１つのニューロンが複数のファミ

*nakayama@info.kanagawa-u.ac.jp

リに反応する場合があるので、単一の色で表示するのは問題があるが、混色にすると同定しにくいので、表示は１つのニューロンに１つのファミリを割り当てた。従って、図の色別の領域はファミリの大まかな領域に対応している。

SOM は入力空間の類似パターンに対して、ニューロン空間の近接ニューロンが反応するように学習が進行する。従って、図の各領域がファミリごとにほぼクラスターとして抽出されているのは SOM によるファミリ識別の妥当性を定性的に示すものであると考えられる。他方、ファミリ領域の間で隣接関係や包含関係を持つ場合や、非連結領域となることに対しては、GPCR データの解釈が必要であると思われる。

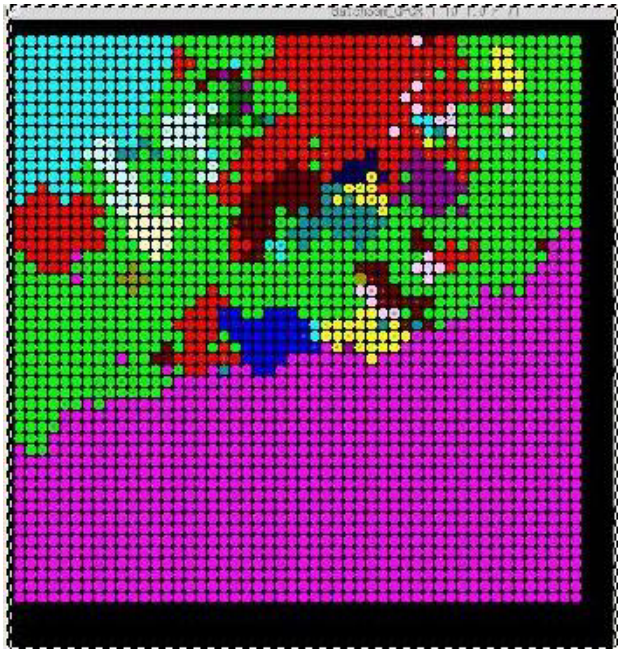

図 1　学習結果のマップ

3.4　分類精度

訓練データとテストデータについての分類精度を調べた結果を表１に示す。学習結果の SOM マップによって入力データが正しく分類された割合をここでの分類精度と定義する：

$$precision = \frac{Number\ \ of\ \ Data\ \ Correctly\ \ Classified}{Number\ \ of\ \ [Family]\ \ Data}$$

１つのニューロンに１つのファミリを対応させて生成されたマップをファミリマップと呼ぶことにする。上記の定義式の分子の項「正しく分類されたデータの数」とは、ファミリマップが生成された後で、すべての訓練データを入力した時に、マップのファミリ領域（のニューロン）がそのファミリのデータに対して（正しく）反応した回数を意味する。

SOM 学習で得られたマップに対して、訓練データあるいはテストデータを入力すると、各入力に対してニューロンが１つだけ反応する（反応ニューロン）。その入力データの所属ファミリおよび反応ニューロンのファミリラベルは既知であるから、その反応ニューロンをそのファミリの認識ニューロンとすることは妥当であろう。同一ニューロンが同一ファミリの複数データに反応する場合が多いほど、それをファミリの認識ニューロンとする妥当性は高くなると考えられる。ところが、一般には同一ニューロンが異なるファミリの入力データに対して反応する場合がある。その場合は、各ニューロンが反応するデータのファイミリをカウントしておき、最大カウントを与えるファミリをマップ上の領域としている。

表 1　分類精度

l	訓練データ			テストデータ		
	$p = 0.5$	$p = 1.0$	$p = 2.0$	$p = 0.5$	$p = 1.0$	$p = 2.0$
10	99.23	0.9923	0.9887	0.9779	0.9761	0.9660
20	0.9887	0.9875	0.9881	0.9761	0.9755	0.9684
30	0.9887	0.9875	0.9905	0.9737	0.9761	0.9624
40	0.9881	0.9881	0.9893	0.9737	0.9731	0.9707
50	0.9875	0.9881	0.9815	0.9672	0.9672	0.9612
60	0.9923	0.9815	0.9792	0.9773	0.9648	0.9510
70	0.9899	0.9851	0.9845	0.9743	0.9701	0.9558
80	0.9905	0.9875	0.9845	0.9743	0.9737	0.9534
90	0.9851	0.9798	0.9833	0.9696	0.9642	0.9516
100	0.9815	0.9833	0.9762	0.9576	0.9611	0.9504

４．　おわりに

本システムによってアラインメントなしのテストデータに対しても 98%程度の分類精度が得られた。アラインメント済みのデータの分類精度にほぼ匹敵する精度である。得られたマップについては、ファミリによっては分離領域や包含領域などが生じているので、それらの解釈が必要であるが、GPCR のファミリはサブファミリを含んでいたり、帰属が曖昧なデータが存在するので、それらは個々に検討することになる。逆に、この分類結果を用いて帰属を検証すべきデータが候補として得られるとも考えられる。

参考文献

[1] Self-Organizing Map, 3rd Edition, Springer, 2001.

[2] Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences, *Protein Science*,**11**, 795-805, 2002.