分類子システムを用いた蛋白質膜貫通領域の識別[1]

3M－2

湯沢 巧　元木 達也[2]
新潟大学[3]

1 はじめに

蛋白質は20種類のアミノ酸がつながってできている。特に膜蛋白質には細胞膜を縫うようにはまり込んだ膜貫通領域と呼ばれる部分があり、この部分を（文字列としての）アミノ酸配列から予測しようという試みが分子生物学の分野で為されている[1]。一方、AI分野ではより単純化された「アミノ酸配列が膜貫通領域であるかどうかを識別する」問題がAIの応用問題として取り上げられてきた。これまでのAIアプローチとしては、決定木を用いたものを始めとして色々なものが考えられてきた[2,3,4,5]が、この論文では分類子システムを用いた試みを紹介する。

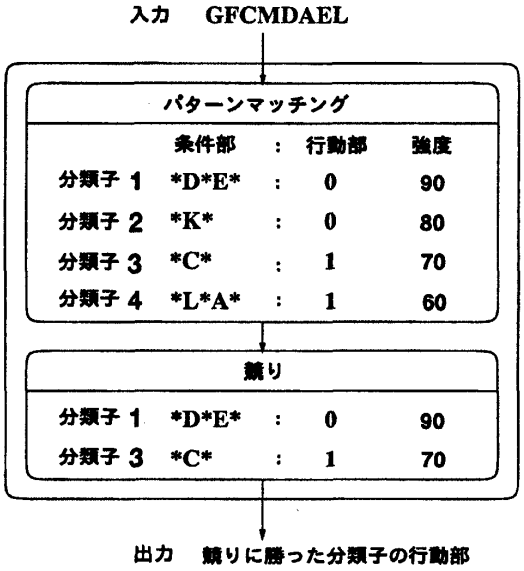
この実験では、PIRデータベースに登録されたアミノ酸配列の中で膜貫通領域とされている689個の部分を正例、膜貫通領域以外の部分から取り出した19256個の部分を負例とし、これらの一部を訓練事例、残りをテスト事例とした。これらの事例は、例えば次の様なものである。

正の例（膜貫通領域）
PLGFVKLQWVFAIFAFATCGSY
FFVTVAVFAFLYSMGALATYIFL
YWVIHSITIPMLFIAGWLFVSTGLA

負の例（非膜貫通領域）
KTGQAPGYSYTAANKNKGIIWGEDTLMEYL
CSQCHTVEKGGKHKTGPNLHGLFGRKTGQA
PKKYIPGTKMIFAGIKKKTEREDLIAYLKK

2 基本的な分類子システム

文献[2,3]に倣って、アミノ酸残基を表す20種類の文字、及び任意の文字列に置き換われるワイルドカードを表す＊記号を（重複を許して）並べたものを分類子の条件部とした。例えば、条件部が ＊R＊Q＊ の分類子は、RとQがこの順に現れるアミノ酸配列に対して適用可能となる。また、各分類子の行動部は調査対象のアミノ酸配列に対する判定結果を表す1と0とする。1の方が「膜貫通領域だ」という判定を表す。分類子システムの行う識別動作は次の図の様に表すことができる。



3 分類子システムの学習

訓練事例を用いて各分類子に付けた強度の調節を繰り返し、時折GA操作を用いて分類子の入れ換えを図る、という学習の流れは通常通りである[6]。この実験に固有の事柄を次に幾つか挙げる。

集団 初期の頃を除いて集団の大きさは20とする。初期集団は ＊:1 と ＊:0 の2つのデフォルト規則だけから成る。

突然変異 文献[3]で述べられている精密化の考えを用いて、例えば、分類子の条件部の＊が ＊A＊ に変わったり、この逆の変わり方をしたりする。

交叉 行動部の等しい個体同士で行う。親の遺伝子の共通部分で最も長い部分が保存されるように交叉点を揃えてから1点交叉を行う。

親1 ＊D｜＊KE＊　交叉→　子1 ＊D＊KE＊K＊
親2 　｜＊KE＊K＊　　子2 ＊KE＊

分類子の多様性の維持 分類子の行動部が一方に偏ることを防ぐため、行動部が0、1の分類子の集団中の比率は同じになるようにし、また、デフォルト規則の強度は70を下回らないように設定する。

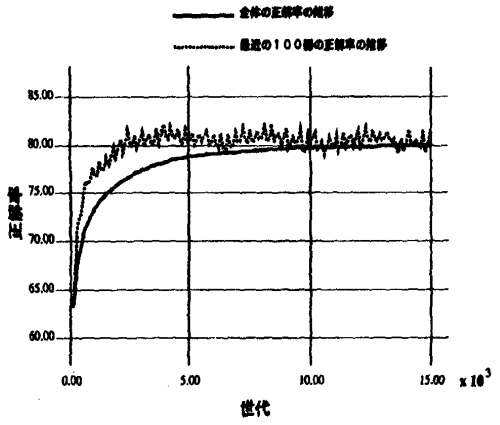
[1] Using Classifier Systems to Recognize Transmembrane Domains in Proteins
[2] Takumi Yuzawa and Tatsuya Motoki
[3] Niigata University

4 学習結果

正事例50個、負事例100個を訓練事例として用いて、15000回の強度調節/300回のGA操作を繰り返して得られる分類子システムの例とその正解率を次に示す。

学習結果		分類子が適用され正解を出す割合			
分類子	強度	正訓練例	負訓練例	正テスト例	負テスト例
*E*R* : 0	95.9	0 / 0	42 / 42	0 / 37	8859 / 8859
*R*P* : 0	95.3	0 / 0	29 / 29	0 / 13	3377 / 3377
E : 0	89.6	0 / 7	19 / 19	0 / 54	4672 / 4672
* : 1	76.6	43 / 43	0 / 10	535 / 535	0 / 2248
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
合計（正解率）		86.00%	90.00%	83.72%	88.26%

また、この様な学習を100回行った時、訓練時の平均正解率の推移、最終分類子システムの平均正解率は次の様になる。



	訓練事例	テスト事例
正事例	76.74%	72.82%
負事例	83.97%	82.40%

5 更なる実験

(1) 分類子システムの学習を繰り返し行って逐次的に1つの決定リストを得ることにした場合、得られた決定リストの正解率は平均的に次の様になる。

	訓練事例	テスト事例
正事例	94.02%	84.39%
負事例	91.44%	85.67%

(2) 親水性指標に基く"インデキシング"[2,3]を行った場合、得られた分類子システムの平均正解率は次の様になる。

	訓練事例	テスト事例
正事例	86.68%	84.38%
負事例	88.77%	88.00%

(3) インデキシングを行ってから決定リストを構成した場合、得られた決定リストの正解率は平均的に次の様になる。

	訓練事例	テスト事例
正事例	96.68%	92.94%
負事例	90.42%	89.20%

(4) 各分類子の条件部を $*\alpha*$（但しαの長さは5以下で、αの中では$*$記号の代わりに任意の1文字に置き換わるワイルドカード記号#を許す）という形のものに変更した場合、得られた分類子システムの平均正解率は次の様になる。

	訓練事例	テスト事例
正事例	77.90%	74.96%
負事例	83.30%	82.05%

6 まとめ

決定リストを構成したりインデキシングを行ったりすることにより正解率を上げることができた。実験結果から、分類子システムの学習を通して文字列の集合上の特徴パターンの抽出が可能であることが示された。今後の課題としては、他の方法との比較、膜貫通領域以外のデータへの適用などが挙げられる。

謝辞 この実験で用いたデータを分けてくださった九州大学の宮野悟教授（現東京大学）にお礼を申し上げます。

参考文献

[1] 中村、中井: バイオテクノロジーのためのコンピュータ入門, コロナ社,1995.

[2] S.Shimozono,et al.:Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI, 情報処理学会論文誌 vol.35,No.10, pp.2009-2018,1994.

[3] H.Arimura,et al.:Protein Motif Discovery from Positive Examples by Minimal Multiple Generalization over Regular Patterns,*Proc.Genome Infomatics Workshop 1994*, pp.39-48,Universal Academy Press,1994.

[4] S.M.Weiss,et al.:Transmembrane Segment Prediction from Protein Sequence Data,*Proc.First International Conference on Intelligent Systems for Molecular Biology*, pp.420-428,AAAI Press,1993.

[5] J.R.Koza:Evolution of a Computer Program for Classifying Protein Segments as Transmembrane Domains Using Genetic Programming,*Proc.Second International Conference on Intelligent Systems for Molecular Biology*,pp.244-252,AAAI Press,1994.

[6] D. E. Goldberg:Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley,1989.