3A1-03

ILP に基づく蛋白質一次構造からの機能予測における生成ルールの分析

Analysis of generation rule in function forecast from the primary protein structure based on ILP

田畑 雅也 Masaya Tabata　松井 藤五郎 Tohgoroh Mastui　大和田 勇人 Hayato Ohwada

東京理科大学大学院理工学研究科経営工学専攻
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

東京理科大学理工学部経営工学科
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

Genome DNA is analyzed, and being made a data base now. The data of the array of the protein is collected. It is important to clarify the function of the protein as the next stage. The function of the protein and the protein tertiary structure have strong ties. However, the data base is very huge. Then, the protein structure of the protein is forecast with data mining tool ILP. We analyzed rule and accuracy for four classes (All-alpha,All-beta,alpha/beta,alpha+beta).

1. はじめに

Turcotte らは, SCOP[2] の PDB エントリに格納れた二次構造の情報から, 帰納論理プログラミング (ILP; Inductive Logic Programming) を用いてフォールドを予測するための規則を獲得する方法を提案した. ILP[4]は構造的な知識を扱うことができるため, たんぱく質のような構造を有するデータの扱いに優れており, たんぱく質の二次構造予測や化合物の突然変異性の予測などに応用されている[5.6.7].

Turcotte らの実験によって, たんぱく質の二次構造とフォールドの間に関連性があることが明らかになった. しかしながら, たんぱく質の二次構造は, X 線解析や核磁気共鳴 (NMR) など高度に専門的な作業によって立体構造を調べることで明らかにされており, 二次構造が判明しているたんぱく質はそのフォールドが判明しているものと考えられる. すなわち, フォールドが未知であるたんぱく質は二次構造も未知であり, 新しく発見されたたんぱく質に対して Turcotte らの方法を適用することはできない.

佐伯らの研究[1]では, たんぱく質の一次構造から二次構造予測ツールを用いて二次構造を予測し, 予測された二次構造からフォールド予測ルールを ILP で学習する. これにより, 新しく発見されたたんぱく質に対してもフォールド予測を行うことをかのうとした.しかしながら、この研究では、全部で4つあるクラスのうち、All-αクラスでの実験のみであり、その All-αクラスに存在する173個すべてのたんぱく質に実験を行ったわけではない.

本研究では、All-α、All-β、α/β、α＋β すべてのフォールド度に対しその、結果を検証する.

2. 一次構造からのフォールド予測

従来研究では, Turcotte らの手法[3]をベースとして, 一次構造からフォールドを予測する. 一次構造の解析は比較的容易であり, 多くの生物の DNA 配列が解析され, データベースとして公開されている. 従来研究では, Web 上で公開されている二次構造予測ツールを用いて一次構造から二次構造を予測し, その結果を一階述語論理表現に変換し ILP でフォールド予測ルールを学習する. 従来手法と Turcotte らの手法の違いを, 図1に示す.

連絡先:田畑 雅也, 東京理科大学理工学研究科経営工学専攻, 神奈川県横浜市緑区東本郷 5-7-2, 090-8115-2826

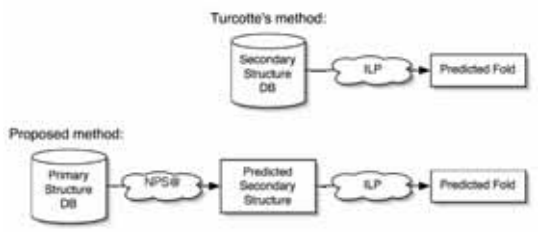


図1:Turcotte らと従来手法の違い

まず始めに, インターネットで公開されている一次構造のデータベース SCOP から, たんぱく質データバンク (Prottein Data Bank; PDB) エントリに格納されたアミノ酸配列を取り出す. PDB エントリには, 一次構造の配列だけでなく, タンパク質やその断片, 基質や阻害剤に結合したたんぱく質のアミノ酸の原子の三次元座標なども格納されている.

3. 実験

3.1 実験方法

提案手法に従って, ①二次構造予測ツールを用いてたんぱく質の一次構造から二次構造を予測し, ②予測された二次構造から背景知識を構築し, ③正例･負例と構築した背景知識から ILP システム用いてフォールド予測ルールを学習した. 本実験では, ILP システムとして Progol 5.0 を使し, all-α ,all-β, α/β、α＋β クラスを対象とした.

(1) データセット

評価実験には, Turcotte らが用いた all-α クラスく 173 個、all-β クラス 233 個 α/β クラス 216 個, α＋β クラス 187 個を使用した.

(2) 評価方法

評価には訓練例の数が非常に少ない時に行われる leave-one-out を使用した. Leave-one-out は, 与えられた例集合からひとつの例を取り出し, それをテスト用の例, 残りを訓練用の例

として学習とテストを行うことを，すべての事例について一つづつテスト用の例となるようにして全体の性能を評価するという方法である．評価尺度には，精度（Accuracy）を用いた．

3.2 実験結果

まず，二次構造予測ツール DSC，PHD，GOR4，HNN，SIMPA96 をそれぞれ用いて二次構造を予測し、そこから背景知識を作り出す．次に，これらの背景知識を用いて学習したときの精度を図2，3，4，5に示す．佐伯の従来研究では、このときに Turcotte らの分類精度と比較するため、All-αクラスでの実験において高い精度を示した二次構造予測ツール DSC を起用して、未知のたんぱく質に、生成ルールを適用し、評価を行った.その結果、従来研究では Turcotte らの分類精度より良い精度を出し、なおかつその生成ルールを比較してみた結果、Turcotte らのルールと遜色ないという結果を得ている．

また従来研究では、二次構造予測ツール DSC が all-αクラスにおいて最も高い精度を示しており、従来実験中の結果では最も有効な二次構造予測ツールであるという結果を示している．しかしながら、本実験により all-α以外のフォールドクラスにも実験を行ったところ、つねに DSE が高い精度を示しているとはいいがたいことが、判明した.図2，3，4，5、で示した結果を、その各フォールドごとの精度の平均値を求めてたものを、図6に示す．また今回行った All-βのフォールドクラスでのルールを例に取り下に示す.今回下に示すのは All-βクラスの Lipocalins フォールドに関する分類ルールである．

fold(’ Lipocalins’,X) :-
adjavent(X,A,B,1,e,h), coil(C,D,2).

このルールは二次構造予測ツールＰＨＤから得られたもので、このルールの分類精度（Accuracy）は 81.4 であった．このルールの意味は，「二次構造の 1 番目（A）はβストランドであり，2 番目（B）はαヘリックスであり，1 番目のもの（A）と 2 番目のもの（B）の間にはコイルが 2 つ存在している」というものである．同様のルールを DSC により獲得されたルールから探すと，

fold(’ Lipocalins’,X) :-
adjavent(X,A,B,2,h,e), unit_len(B,lo)

というものが見つかった．二次構造予測ツール DSC からえられた，このルールの分類精度（Accuracy）は 69.7 であった．このルールの意味は，「二次構造の 2 番目（A）αヘリックスであり 3 番目（B）はβストランドである，2 番目のもの（B）の長さは lo である」というものである．PHD で得られたルールは DSC に比べ遜色なく信頼性もありこの場合であるなら分類精度を上回る結果でもある．

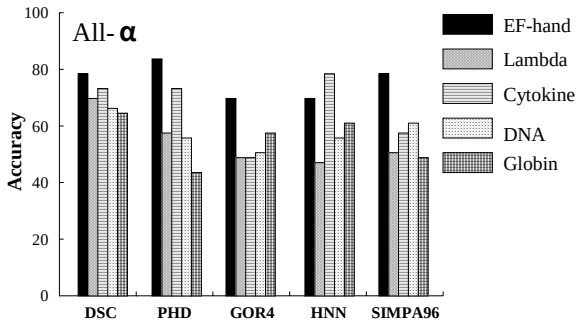


図2：All－αクラスの精度

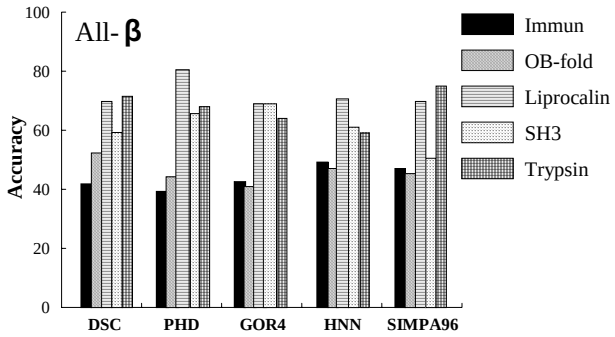
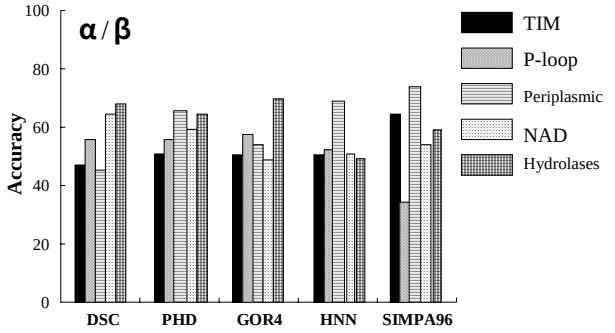


図3：All－βクラスの精度



図：α/βクラスの精度

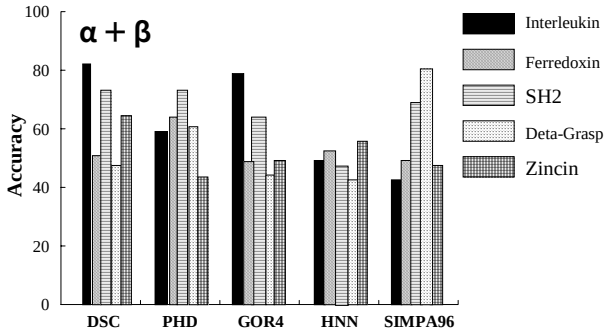


図2：α＋βクラスの精度

4. 考察

4.1 ルールを調べた結果

・述語 adjacent と coil を組み合わせたルールが多かった．
・ドメインの長さを用いたルールが少なかった．

前者の特徴は，Turcotte らの手法によって得られたルールにも見られるものである．したがって，二次構造予測ルールは，これらをある程度正しく予測しているということが推測できる．反対に，Turcotte らの手法で獲得されたルールには多く使われていた，たんぱく質の長さについての条件が，提案手法で獲得されたルールにはそれほど現れなかった．ドメインの長さは与えられたドメインの配列情報から調べることができるため，二次構造予測ツールの予測誤差が原因であるとは考えられない．このことから，ドメインの長さを用いてルールを作成するよりも，その他の知識を用いてルールを作成した方が評価が高くなったと考えられる．

4.2 全4つのフォールドの精度評価

図6の結果から、all-αクラスの平均精度が他のものより若干ではあるが、上回っていることが見受けられる.このことからも、二

次構造予測ツールの精度においては、all-αクラスのものが高いという評価ができる.

またこの結果より、二次構造予測ツールと実験で起用しているたんぱく質フォールドとの特性を、見出すことが出来る.従来実験で最も優れているとされている DSC の結果を見てみると、その分類精度が、非常にばらつきが、あることが解る、つまり DSC においては All-αクラスへの予測に非常に優れている、ということがみうけられるが、他の All-β、α/β、α＋β、のクラスに対しては特出して有効であるとはいえない.実際、図6を見てみると All-α以外のフォールドクラスではむしろ PHD での精度のほうが高い数値を示すことも多い.

この結果から、二次構造予測ツールを起用する面において、高い精度を求めるのであるならば、そのフォールドクラスにあった二次構造予測ツールを用いて実験を行ったほうが精度の向上が見込めるということがわかった.

また、SIMPA96においてはクラスの変更による差があまり見受けられないので、分類精度の差をあまり作ることなく、評価したい場合には適しているといえる.

5. 終わりに

本手法は, 正しい二次構造から ILP を用いてフォールド予測ルールを学習した Turcotte らの手法に基づき, 二次構造予測ツールを用いてたんぱく質の一次構造から二次構造を予測し, 予測された二次構造から ILP を用いてフォールド予測ルールを学習するというものである.

複数の二次構造予測ツールを組み合わせて全4クラスの 5 つのフォールドに対して予測ルールを学習し, その精度比較した結果, all-αクラスにおいて、予測ツールである DSC を用いたとき予測精度が最も高くなることが分かった. Turcotte らの手法によって獲得されるフォールド予測ルールと比較すると, 正しい二次構造が分かっているときには Turcotte らの手法で獲得されたルールの予測精度の方が高いが, 二次構造が未知のとき, すなわち, 二次構造予測ツールで予測しなければならないときは提案手法の予測精度の方が高い. このことから, 新しく発見された二次構造が未知のたんぱく質に対してそのフォールドを予測するのには, 従来手法よりも提案手法の方が優れている.

また、すべてのフォールドクラス、において、実験を行ったことにより、クラスごとの二次構造予測ツールの精度比較が行えた. それにより、二次構造予測ツールの特性と各フォールドクラスとの関連性を示すことが出来た.

今後は, all-alpha クラス以外のたんぱく質について更なる検証が必要であり, 二次構造以外の背景知識を加えることによってより精度の高いフォールド予測ルールを獲得することなどが課題である.

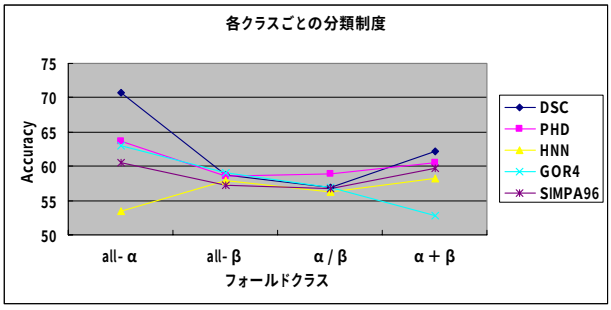


図6：各クラスごとの分類精度の平均

参考文献

[1] 佐伯康吏, 松井藤五郎, 大和田勇人. たんぱく質の一次構造からの機能予測. 情報処理学会第 66 回全国大会講演論文集, Vol. 4, pp. 541−542 (2004).

[2] LoConte, L., Ailey, B., Hubbard, T.J.P., Brenner, S.E., Murzin, A.G. and Chothia, C. SCOP: astructural classification of proteins database. Nucl.Acids.Res. 28:257–259 (2000).

[3] M. Turcotte, S.H. Muggleton, and M.J.E. Sternberg. Automated discovery of structural signatures of protein fold and function. Journal of Molecular Biology, 306:591-605 (2001).

[4] Muggleton, S. and De Raedt, L.D. Inductive logic programming: Theory and methods. Journal of Logic Programming 19/20:629-679 (1994).

[5] Muggleton, S., King, R. and Sternberg, M.J.E. Protein secondary structure prediction using logicbased machine learning. Protein Eng. 5:647–657 (1992).

[6] King, R.D., Muggleton, S., lewis, R.A. and Sternberg, M.J. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. Proceedings of the National Academy of Science 89:11322–11326 (1992).

[7] Hirst, J.D., King, R.D. and Sternberg, M.J.E. Quantitative structure-activity relationships by neural networks and inductive logic programming. I. The inhibition of dihydrofolate reductase by pyrimidines. Journal of Computer Aided Molecule Design 8:405–420 (1994).