6L-1

ILP に基づく蛋白質一次構造からの機能予測における背景知識の改良

田畑 雅也　　松井 藤五郎　　大和田 勇人

東京理科大学大学院理工学研究科経営工学専攻
東京理科大学理工学部経営工学科

１　はじめに

現在コンピュータの技術の発展に伴い，ゲノム DNA の塩基配列の解析がすすめられている．Turcotte らは，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．

大和田研究室では，Turcotte らの研究[1]を基に，蛋白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 で学習する研究[2]を行なってきた．この研究で二次構造が未知である蛋白質のフォールド予測を可能とした．

本研究では，目的として，いかに導き出される分類ルールを，利用価値の高いものにしていくかと言う指標に基づき更なる実験を行なう．これは分類精度が高くてもその先にある，知識の発見に生かされなくては折角導き出したルールの価値も半減してしまうのでは，と言う考えからである．そこで背景知識に着目し目的の達成のためにこの改善を行なった．

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[1]をベースとして，一次構造からフォールドを予測する．これは二次構造が未知の蛋白質に対して Turcotte らの実験では適用することが困難であるとの考えからである．

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．

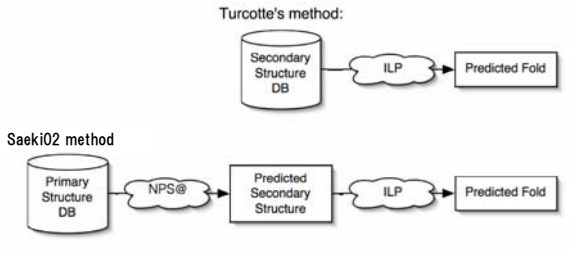

図１：従来研究と Turcotte らの研究

３　提案手法

蛋白質を構成する２０種類のアミノ酸のうち従来実験では，プロリン(Pro)に関する情報しか背景知識に与えられていなかった．蛋白質の構造は，アミノ酸の配列によって決定しているが，そのアミノ酸のなかにも，構造を生成する上で重要な役割をはたしているものが他にも存在する，そこで本論では，以下のものを背景知識に加えることにする．

①システイン(Cys)

システインの関与する，ジスルフィド結合によって，もう一つの別のシステインと反応して，形成する連結部分を「架橋」といい，蛋白質の構造を安定させる役割を担っている．

背景知識：has_syc(X)

ドメイン X はシステインを含んでいる．

②グリシン(Gly)

グリシンはアミノ酸の中で，最も単純な形を持っており，蛋白質の構造において，柔軟性を局所的に高める役割を担っている．

背景知識：has_gly(X)

ドメイン X はグリシンを含んでいる．

４　実験

4.1 実験方法

実験データ：ALL-α クラスフォールド 173 個
二次構造予測ツール：DSC
ILP システム：Progol 5.0

Imprving of background knowledge in function forecast from the primary protein structure based on ILP
Masaya Tabata　Tohgoroh Mastui　Hayato Ohwada
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

・正事例：fold('Globin-like', d1scta_).
ドメイン d1scta_は Globin-like に属する.

・負事例：:- fold('Globin-like', d1cei__).
ドメイン d1cei__は Globin-like に属さない.

実験 1 ：システイン、グリシンを背景知識に用いての実験
実験Ⅱ：プロリン，システイン，グリシン全てを背景知識に加えての実験
（背景知識増加によりパラメータを変更）

4.2 実験結果

図２のグラフは，All-αクラスの予測ツール DSC において，従来どおりの手法と背景知識に「has_cys」「has_gly」を用いての実験結果を比較したグラフである．図３は実験Ⅱの実験を従来道理の背景知識を用いたものと本実験を比較したものである．

以下に背景知識の変更を行なった実験で得られたルールの抜粋を記載する．

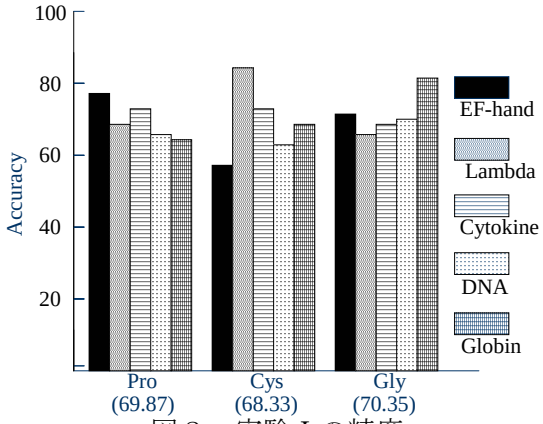

図２：実験Ⅰの精度

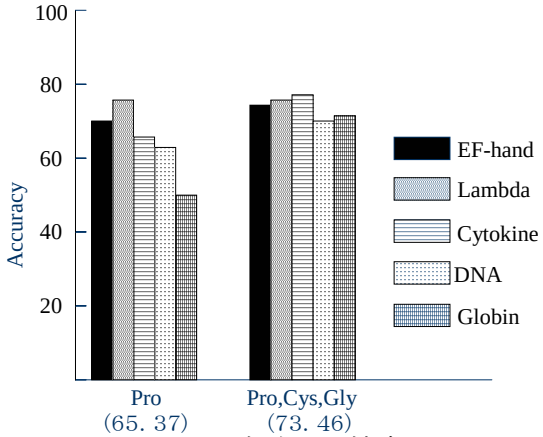

図３：実験Ⅱの精度

・実験Ⅰでのルール

fold('Globin-like',A) :- adjacent(A,B,C,4,h,e),has_cys(B).

このルールの意味は，「二次構造の２番目（A）は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（B）はシステインを含んでいる」というものである．

・実験Ⅱでのルール

fold('Globin-like',A) :-adjacent(A,B,C,1,h,h), has_gly(C).
fold('Globin-like',A) :- adjacent(A,B,C,5,h,h), has_pro(B).
fold('4-helical cytokines',A) :- adjacent(A,B,C,3,h,h), has_cys(B).

5 考察

実験Ⅰ（図２）の結果からシステインやグリシンを用いたそのフォールドの種類によって精度にばらつきがあることがわかる．しかし全体としては従来と比べ損傷のない結果が得られた．生成ルールの面では，グリシンを用いた実験ではあまりそれ自体が起用されることはなく、システインを用いた実験では数多くルールに用いられていた．

また，実験Ⅱ（図３）の結果では，精度の面では従来に比べその平均精度が向上し，また各フォールドごとの制度のばらつきもあまり見られなかった．生成されたルールを調べてみたところ，実験Ⅲでは has_pro に加え has_gly，has_cys の情報方が同程度ルールに起用された，このことから今回の実験でおこなった背景知識の変更により，例えば「この二次構造はシステインを有しており構造が安定している可能性がある」というように，生物分野へ新たな発見を提案しうる可能性があるのではないかと考える．

6 まとめ

今回、背景知識を改良したことによって、分類精度を従来に比べて 65.37%から 73.46%へ向上させることが出来た．

また，グリシン，システインを含んだ新たな知識の発見の可能性を見出せた．

参考文献

[1] 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).

[2] 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).