ILP を用いた BCL2 ファミリータンパク質の一次構造からのフォールド予測

河村 真平* 松井 藤五郎† 賀屋 秀隆‡ 大和田 勇人† 朽津 和幸‡

東京理科大学理工学研究科経営工学専攻* 同 理工学部 経営工学科† 同 応用生物科学科‡

1 はじめに

近年, タンパク質の機能決定に非常に大きな影響を与えるタンパク質の持つ立体構造の解析が, 積極的に行われている. しかし, タンパク質の立体構造の解析には多くの時間と費用がかかるといった問題点が有る. そこでタンパク質の 1 次構造からタンパク質の立体構造を予測する手法が注目を浴びている.

これまでも,Turcotte らにより SCOP の PDB エントリに格納された 2 次構造の情報から,ILP を用いてフォールドを予測するための規則を獲得する手法 [1] や, 田畑らによりタンパク質の 1 次構造から 2 次構造予測ツールを用いて 2 次構造を予測し, 予測された 2 次構造からフォールド予測ルールを ILP で学習するという方法が提案された [2].

しかし, これまでの手法で用いられていた 1 次構造の配列は, タンパク質データバンク (PDB) エントリに格納されているものであった.PDB エントリに格納されている一次配列は実験的に解明された「タンパク質構造」のデータである. タンパク質構造はタンパク質配列の一部であり、又、既に立体構造が解明された配列しか PDB には登録されないため, 従来の手法ではフォールドが未知なタンパク質配列に対してのフォールド予測を行っていないという問題点が有った.

そこで本研究ではフォールドが未知なタンパク質の 1 次構造からのフォールド予測を可能とする手法を提案する. 本論文ではさらに提案した手法の有効性を, 実験や, 得られたルールを分析する事で検証する.

2 フォールドとは

フォールドとは SCOP データベースにおける分類階層の一つである.SCOP データベースは, 専門家による構造類似性の定義に基づいて, クラス, フォールド, スーパーファミリー, ファミリーなど多数の階層レベルによってタンパク質を分類している. 例えば Connexin43 フォールドは細胞間結合装置の一つである, ギャップ結合の構成タンパク質の集合体であり, Aquaporin-like フォールドは細胞内の水分量を調節する水チャネルとして機能する重要なタンパク質の集合体である. このように,1 次構造しか判明していないタンパク質のフォールドを予測することは, タンパク質の機能を予測することと密接に関係している.

3 提案手法

本論文ではルールの学習に PDB 内の立体構造の配列データではなく, Swiss-Prot から入手したタンパク質配列データを使用する. PDB 内でフォールドに登録されている立体構造を所有するタンパク質を Swiss-Prot より入手し, ILP を用いてフォールド予測ルールを作成する. これにより, PDB の立体構造分類を活用し, タンパク質の 1 次構造からのフォールド予測が可能となると考える. 以下に詳細を示す.

STEP.1: SCOP より, ルールを作成したいフォールドのに所属する立体構造の登録名を取得する. 負事例として, 取得した正事例の数と同じ数の立体構造の登録名を, 正事例のフォールドが所属するクラス以外の全てのクラスから, 同数ずつ取ってくる. 例えば正事例の数が 50 個だとすると, 現在クラスは 11 個有るのでルール化したいフォールドが所属するクラス以外のクラス 10 個から, 5 個ずつ, 計 50 本を負事例とする.

STEP.2: 従来手法では, **STEP.1** で取得した立体構造の配列情報から 2 次構造を予測し, フォールド予測用のルールを作成していた. 本手法では 1. で取得した立体構造を含むタンパク質の 1 次構造を Swiss-Prot から取得する. これにより SCOP の構造類似性の定義を活かしつつも, タンパク質の 1 次構造を基にしたルール生成が可能になる.

STEP.3: **STEP.2** で取得した Swiss-Prot の 1 次構造の 2 次構造を, 2 次構造予測ツール SSpro[3] を用いて予測する.

STEP.4: **STEP.3** で予測した 2 次構造を, ILP 学習用の背景知識データに変換する. 背景知識データには, タンパク質の長さや 2 次構造 α ヘリックス, β シートの数, 各 2 次構造の位置, 長さ等が有る.

STEP.5: **STEP.4** で作成した背景知識データから ILP を用いて, フォールド予測用のルールを作成する.

4 実験

4.1 実験手法

提案手法の有効性を検証する為に, 本論文ではの三種類の実験を行った. 実験 1 では田畑らによる PDB 内の立体構造の配

Fold prediction of BCL2 family from first structure using inductive logic programming

Shimpei KAWAMURA*, Tohgoroh MATSUI†, Hidetaka KAYA‡, Hayato OHWADA†, and Kazuyuki KUCHITSU‡

Department of Industrial Administration, Graduate school of Science and Technology, Tokyo University of Science*, Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science†, Department of Applied Biological Science, Faculty of Science and Technology, Tokyo University of Science‡

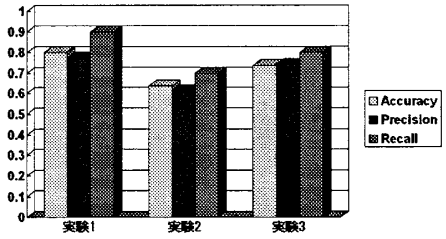

図1 実験結果:実験1ではAccuracyは0.8, Precisionは0.772, Recallは0.9, 実験2ではAccuracyは0.64, Precisionは0.62, Recallは0.7, 実験3ではAccuracyは0.74, Precisionは0.74, Recallは0.8となった

表1 従来手法, 提案手法それぞれで作成されたルール分析表.

	実験1, 実験2	実験3
総ルール数	21	28
adjacent	21	28
coil	9	20
has_pro	2	2

列データを元にフォールド予測ルールの作成し, そのルールを元に立体構造の配列データからフォールド予測を行う. 田畑らの手法をそのまま再現した形である. 実験2では田畑らの手法によって作成したルールで, タンパク質の1次構造の配列データに対してフォールド予測を行う. この実験で田畑らの手法でどれだけタンパク質の1次配列に対して正確にフォールド予測が出来るのかを検証する. 実験3では提案手法により作成された, タンパク質の1次構造から作成されたルールを用いて, タンパク質の1次配列に対してフォールド予測を行う. 実験2の結果と実験3の結果を比較する事で, 本手法の有効性を検証したい. 以下に本実験で使用したデータの詳細を述べる.

■実験1:トレーニングデータとして正事例にSCOP内でBCL2ドメインを含むToxins' membrane translocation domains (以下Toxins')フォールドに分類されている立体構造の配列データ40本, 負事例にToxins'フォールドが所属するクラス以外の10個のクラスからそれぞれ立体構造の配列データを4本ずつ, 計40本を使用. テストデータとしても同じデータを使用.

■実験2:トレーニングデータは1.で使用した物と同じ物を使用. テストデータにはトレーニングデータで用いた立体構造を含むタンパク質を,Swiss-Protから取得し使用.

■実験3トレーニングデータとして2.で使用したSwiss-Protのデータを使用. テストデータとしても同じデータを使用.

2次構造予測ツールにはSSpro,ILPシステムはGKS[4]を使用. ルール及び精度導出には5-fold cross varidationを用いた.

4.2 実験結果

実験結果を図1に示す. 各結果の三本の棒グラフは左からAccuracy, Precision, Recallを表している. 又, 本実験により作成されたルールの分析結果を表1に示す.

5 考察

従来手法のまま, フォールド予測ルールをSwiss-Prot内のデータに適用した結果は, 実験1と実験2を比較する事で分かる.Accuracy,Precision,Recall全てが低下している. この事から, ルールや背景知識作成の際に, 立体構造の配列データベースであるSCOPとタンパク質1次構造データベースのSwiss-Protでの配列の違いがフォールド予測に悪影響を与えている事が分かる. さらに実験2と実験3を比較すると, Accuracy,Precision,Recall全てで実験3の結果が実験2の結果を上回っていた. これより提案手法を用いてSwiss-Prot内のタンパク質の配列データからルールを作成する事で,SCOP内の構造分類の定義を活かしつつ, タンパク質の1次構造の配列データに対してフォールド予測ルールがより正確に作成出来たという事が分かる. さらに表1から分かるように, 従来手法で作成されたルールに比べて, 提案手法で作成されたルールの方が述語「coil」の出現数が多い事が分かる.adjacentとcoilの二つの述語を同時に持つルールは,Turcotteらの先行研究によって得られたルールでも多く見られる物であり, 提案手法の方がフォールドを正確に予測出来るルールを作成出来たと考えられる.

6 結論

従来手法ではタンパク質の1次構造データからのフォールド予測を行っていなかった. その為, 本論文ではSCOP内の立体構造データからではなく,Swiss-Protのタンパク質データの1次構造データからフォールド予測ルールを作成するシステムを作成し, それにより作成出来たルールの有効性を実験を通して検証した. 実験を通して従来手法がではタンパク質の1次構造に対して立体構造予測を行えない事に対し, 提案手法ではそれが可能になった事, またルールを分析する事で従来よりより有用なフォールド予測ルールが作成出来た事が検証出来た.

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] 田畑, 松井, 大和田: 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. 人工知能学会 (2005)

[3] J. Cheng, A. Randall, M. Sweredoski, P. Baldi, SCRATCH: a Protein Structure and Structural Feature Prediction Server, Nucleic Acids Research, vol. 33 (web server issue), w72-76, 2005

[4] Fumio Mizoguchi and Hayato Ohwada: Using inductive logic programming for constraint acquisition in constraint-based problem solving. In Proceedings of the 5th International Workshop on Inductive Logic Programming, pages 297.322, 1995.