3C6-2

ILPを用いたBCL2ファミリーのフォールド予測

Fold prediction of BCL2 family using inductive logic programming

河村 真平[*1] 松井 藤五郎[*2] 賀屋 秀隆[*4] 大和田 勇人[*2] 朽津 和幸[*3*4]
Shimpei Kawamura Tohgorou Matsui Hidetaka Kaya Hayato Ohwada Kaduyuki Kutitsu

[*1]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

[*2]同 理工学部 経営工学科
Department of Industrial Administration, Faculty of Science and Technology

[*3]同 理工学部 応用生物科学科
Department of Applied Biological Science, Faculty of Science and Technology

[*4]同 理工学部 ゲノム創薬研究センター
Genome and Drug Research Center, Faculty of Science and Technology

Function prediction of protein and three-dimensional structure prediction of protein are closely related. We suggested three-dimensional protein structure prediction technique from the protein primary structure using ILP. However, that method used primary structure in protein database that three-dimensional structures are already known for rule learning. Between protein database that three-dimensional structure is already known a database that three-dimensional structure is unknown there is the case that notation of primary structure of protein is different with the same protein. Therefore there were problems that a three-dimensional structure prediction was difficult for the database how three-dimensional structure was unknown in a rule of conventional technique. Therefore solid structure suggests technique to make a rule of solid structure prediction use from an unknown database to be able to perform a prediction for the database how three-dimensional structure is unknown in this article and inspects the utility.

1. はじめに

近年，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．これまでに，いくつかの生物について，ゲノム DNA の全塩基配列が解明された．Turcotte らは 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方法を提案した [1]．ILP は構造的な知識を扱うことができるため，タンパク質のような構造を有するデータの扱いに優れており，タンパク質の２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．Turcotte らの実験により，タンパク質の２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，フォールドが未知である場合は２次構造も未知である．故に Turcotte らの手法は，新しく発見され１次構造のみしか解明されていないタンパク質に対して適用することができない．

そこで我々は，これまでにタンパク質の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を提案した [2]．これにより，新しく発見されたタンパク質に対してもフォールド予測を行うことを可能とした．

これまでの手法で用いられていた１次構造は、タンパク質データバンク (Protein Data Bank;PDB) エントリに格納されている１次構造であった．PDB エントリに格納されているタンパク質は X 線結晶解析法，NMR 法 (核磁気共鳴法) などによって実験的に３次構造が決定されたデータである．故にフォールドが未知なタンパク質に対しては，ルール学習に３次構造が実験的に決定されていない配列データも含んだ Swiss-Prot や，国際塩基配列データベース（INSD）を用いる必要が有る．

しかし，Swiss-Prot と PDB では格納されている１次構造データの表記に若干の差が有り，従来の予測ルールでは Swiss-Prot の格納データに適用する事が不可能である．そこで本研究ではフォールド予測ルールを Swiss-Prot 内の配列データを用いて作成する事により，Swiss-Prot に格納されている配列データにも本手法が有効かを実験し，検証する．

連絡先: 河村 真平，東京理科大学大学院 理工学研究科 経営工学専攻，千葉県 野田市 山崎 2641 ，j7406604@ed.noda.tus.ac.jp

2. ILP を用いた１次構造からのフォールド予測

2.1 ILP とは

ILP(inductive logic programming) は，論理プログラミングに従って帰納論理を行う枠組みである．ILP は，一階述語論理を扱う事が出来る事から，属性値の集合では表現出来ない関係表現を学習する事が出来る．ILP の最大の特徴は，一階述語論理で記述された背景知識 (background knowledge) を学習システムに与えられる事である．これより ILP システムに複雑なルールを学習させる事が出来る．

2.2 フォールドとは

フォールド（fold）とは SCOP (Structural Classification of Proteins) データベースにおける分類階層の一つである．３次構造のデータベースには，SCOP, CATH (Class,Architecture, Topology and Homologous superfamily) , DALI/FSSP などが存在する．これらは，異なった比較法に基づいて，タンパク質を分類している．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（図 1）．

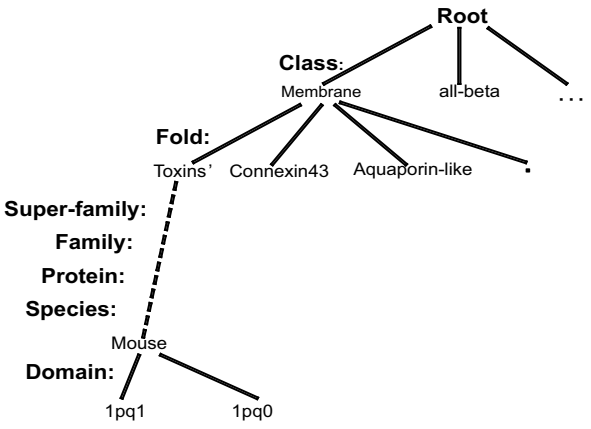

図 1: SCOP 階層図

図中のクラスの Membrane は Membrane and cell surface proteins and peptides，フォールドの Toxins 'は Toxins' membrane translocation domains の事を表している．フォールドは，クラスの下の階層で，Membrane and cell surface proteins and peptides クラスには Toxins' membrane translocation domains, Connexin43, Aquaporin-like などのフォールドがある．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．

2.3 1 次構造からのフォールド予測

タンパク質の配列からその立体構造を予測する事は，タンパク質の機能を解明する上で非常に重要である．Turcotte らは，タンパク質の 2 次構造から，ILP を用いてフォールドを予測するためのルールを獲得する手法を提案し，2 次構造とフォールド間に関連性が有る事を明らかにした．しかし，タンパク質の 2 次構造を明らかにするためには，タンパク質の立体構造を解析する必要があり，Turcotte らの手法を適用する事は出来ない．そこで田畑らは Turcotte らの手法に基づき，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学習するという手法を提案した．しかし，この手法で使用している 1 次構造は PDB に格納された 1 次構造であるため，Swiss-Prot に格納されている 1 次構造に対して正確な予測が出来ないという問題点が有る．この問題点については次節で詳細を述べる．

3. 提案手法

3.1 従来手法の問題点

従来は PDB に格納されたタンパク質の 1 次構造から，2 次構造を予測し，ILP を用いてフォールド予測ルールを作成した．しかし PDB に格納されているタンパク質とは，すでに 3 次構造が解明されているタンパク質である．1 次構造からのフォールド予測は，3 次構造が解明されていないタンパク質からフォールド予測ルールを作成し，そのルールを用いてタンパク質の機能を予測する事を目的としている．そのため，予測する対象のタンパク質は 3 次構造が解明されている SCOP や CATH に格納されている物ではなく，タンパク質の配列のみが格納されている Swiss-Prot 等のタンパク質を用いる必要が有る．又，SCOP に格納されているタンパク質配列と Swiss-Prot に格納されているタンパク質配列とでは，同じタンパク質の配列を示していても，両データベースで配列表記が異なる場合が有る．

```
>Apoptosis regulator Bcl-X from Mouse(PDB)
MSQSNRELVVDFLSYKLSQKGYSWSQFSDVEENRTEAPEETEAERETPSAINGNPSWHLA
DSPAVNGATGHSSSLDAREVIPMAAVKQALREAGDEFELRYRRAFSDLTSQLHITPGTAY
QSFEQVVNELFRDGVNWGRIVAFFSFGGALCVESVDKEMQVLVSRIASWMATYLNDHLEP
WIQENGGWDTFVDLYG

>Apoptosis regulator Bcl-X from Mouse(Swiss-Prot)
MSQSNRELVVDFLSYKLSQKGYSWSQFSDVEENRTEAPEETEAERETPSAINGNPSWHLA
DSPAVNGATGHSSSLDAREVIPMAAVKQALREAGDEFELRYRRAFSDLTSQLHITPGTAY
QSFEQVVNELFRDGVNWGRIVAFFSFGGALCVESVDKEMQVLVSRIASWMATYLNDHLEP
WIQENGGWDTFVDLYGNNAAAESRKGQERFNRWFLTGMTVAGVVLLGSLFSRK
```

図 2: 両配列の違い

その例として図 2 でそれぞれの配列を示す．両配列共にネズミの Bcl-X の配列である．上の配列が SCOP から，下の配列が Swiss-Prot から取得してきた物だが，これより 4 列目に違いが有る事が分かる．故に，PDB に格納されているタンパク質配列から学習したルールを用いては，Swiss-Prot から用いたタンパク質配列から正確に予測できないという問題点が有る．実際，TAIR で公開されているシロイヌナズナのタンパク質データベースは，Swiss-Prot 内のタンパク質データが使用されているため，従来手法にて学習したルールの適用は非常に難しい．

3.2 提案手法

本論文ではルールの学習に PDB 内のタンパク質ではなく，Swiss-Prot から入手したタンパク質配列データを使用する．PDB 内でフォールドに分類されているタンパク質と，同じタンパク質を 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，さらに Swiss-Prot タイプの表記のタンパク質に対しても的確な分類が実行可能となる．以下に詳細を示す．

1. SCOP より，ルールを作成したいフォールドの 1 次構造を取得する．ルール作成時には正事例としてこれらの 1 次構造を使用する．負事例として，取得した正事例の数と同じ数の 1 次配列を，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．

2. 従来手法では，1. で取得した 1 次構造から 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 1 次構造のタンパク質と同じタンパク質の 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，Swiss-Prot 側の 1 次構造を基にしたルール生成が可能になる．

3. 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．

4. 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，以下の様な物が有る．

   - len(D,L). タンパク質 D の長さは L である．

- nb_alpha(D,N). タンパク質 D のαヘリックスの数は N である.
- sec_struc(D,S). タンパク質 D は 2 次構造 S を含んでいる.
- sst(S,N1,N2,_,N3,L,_,_,_,_). 2 次構造 S は前から N1 番目の位置に有り，同型の 2 次構造としては前から N2 番目である. また，1 次配列の N3 番目から始まっており，その長さは L である.

5. 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，本論文では以下の三種類の実験を行った.

1. SCOP 内の 1 次配列データから 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作成. そのルールの有効性をテストデータに SCOP 内の 1 次配列から作成した背景知識を使用し検証.
2. SCOP 内の 1 次配列データから 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作成. そのルールの有効性をテストデータに Swiss-Prot 内の 1 次配列から作成した背景知識を使用し検証.
3. Swiss-Prot 内の 1 次配列データから 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作成. そのルールの有効性をテストデータに Swiss-Prot 内の 1 次配列から作成した背景知識を使用し検証.

4.2 使用データセット

前節の実験では以下のデータを用いた.

1. トレーニングデータとして正事例に SCOP 内で Toxins' membrane translocation domains フォールドに分類されているタンパク質の 1 次配列 40 本，負事例に Toxins' membrane translocation domains 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タンパク質 1 次配列を 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.
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タンパク質と同じタンパク質を，Swiss-Prot から取得し使用.
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[4,5] を使用. ルール及び精度導出には 5-fold cross varidation を用いた.

4.3 実験結果

実験結果を図 3 に示す. 棒グラフの下に有るそれぞれの番号はそれぞれ実験番号を示し，三本の棒グラフは左から Accuracy，Precision，Recall を表している. 実験 1 では Accuracy は 0.8，Precision は 0.772，Recall は 0.9，となった. 実験 2 では Accuracy は 0.64，Precision は 0.62，Recall は 0.7，そして実験 3 では Accuracy は 0.74，Precision は 0.74，Recall は 0.8 という結果を記録した.

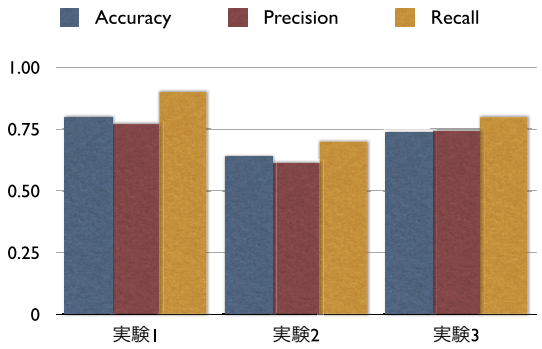

図 3: 実験結果 実験 1 は PDB 内のデータで学習，PDB 内のデータに対してルール適用 実験 2 は PDB 内のデータで学習，Swiss-Prot 内のデータに対して適用 実験 3 は Swiss-Prot 内のデータで学習，Swiss-Prot 内のデータに対して適用

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. SCOP 内のデータから作成したルールを，SCOP 内のデータではなく，Swiss-Prot 内のデータに適用することで，Accuracy，Precision，Recall 全てが低下している事が分かる. この事から，ルールや背景知識作成の際に，SCOP と Swiss-Prot での 1 次配列の表記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，Swiss-Prot 内のデータに対しての我々の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. 従来手法では低い精度を記録した，立体構造が解明されていないデータベースに対する ILP を用いたフォールド予測ルール作成だが，本手法を用いる事でその精度を向上させる事が可能となった.

6. まとめ

従来のタンパク質のフォールド予測ルールでは，Swiss-Prot 内のデータに対し精度の低い予測しか出来なかった. その為，本論文では SCOP 内のデータからではなく，Swiss-Prot のデータからフォールド予測ルールを作成し，その有効性を検証した. 本手法では，SCOP 内のデータと Swiss-Prot のデータ間の 1 次構造の表記の違いを考え，提案手法で新しいルールを作成した. その結果，従来手法のままでの Swiss-Prot のデータからのフォールド予測は精度が非常に低かったが，提案手法で作成したルールを用いる事で，その精度を向上させる事が出来た. この事から提案手法の有効性を示す事が出来た. また，本実験で作成した Toxins' membrane translocation domains フォールド予測ルールは，新しい BCL-2 ファミリータンパク質の発見につながる事が期待される.

参考文献

[1] M.Turcotte, S.H.Muggleton, and M.J.E.Sternberg.:Automated discovery of struc-

tural 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: Constrained relative least general generalization for inducing constraint logic programs. New Generation Computing, 13:335. 368, 1995.

[5] 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.