様式 C-19

科学研究費補助金研究成果報告書

平成 21 年　5 月 29 日現在

研究種目：基盤研究（C）
研究期間：2006～2008
課題番号：１８５００１２０
研究課題名（和文）　ＩＬＰに基づくたんぱく質一次構造からのフォールド予測
研究課題名（英文）　Fold prediction from protein of one-dimensional structure using inductive logic programming

研究代表者
大和田　勇人（OHWADA HAYATO）
東京理科大学・理工学部経営工学科・教授
研究者番号：30203954

研究成果の概要：本研究ではプログラム細胞死に影響を与える Bcl-2 ファミリーのたんぱく質を対象にして，シロイヌナズナの中から発見することを目的とした．たんぱく質の二次構造予測ツールを用いて二次構造を予測し，帰納論理プログラミング（ILP）を用いて予測された二次構造に基づいた背景知識から仮説を構築した．それを，Bcl-2 ファミリーたんぱく質のフォールド予測に適用した．実験の結果，提案手法は予測精度，特に再現率を向上させることが実証された．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	1,500,000	0	1,500,000
２００７年度	1,400,000	420,000	1,820,000
２００８年度	700,000	210,000	910,000
年度			
年度			
総　計	3,600,000	630,000	4,230,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：帰納論理プログラミング，バイオインフォマティクス，たんぱく質，フォールド予測，ドメイン，機械学習

１．研究開始当初の背景

(1) ゲノム解析の究極目標はゲノム DNA の塩基配列に刻み込まれている遺伝情報の完全解読であり，どんなたんぱく質が，いつ，どこで，どれだけ働いているかを解析することもその一つである．未知のたんぱく質の機能は，機能が明らかにされているたんぱく質遺伝子に全域で，しかもきわめて高い類似を有していることが示されれば，その機能は既知のたんぱく質と類似であると推測できる．生物学者が未知のたんぱく質の機能を調べる時に行う処理は，その配列と似たところのある既知のたんぱく質の一次構造から相同性検索ツールを利用して探したり，機能を特定できるような部分配列のパターンをモチーフ検索を用いて持っているかどうか調べたりする．しかし，そのたんぱく質の一次構造解析のスピードに比べ，その機能を確定するためには膨大な時間を要している．

(2) 過去に Turcotte らはたんぱく質データバンク(Protein Data Bank;PDB)エントリに格納された二次構造の情報から，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，たんぱく質の二次構造と

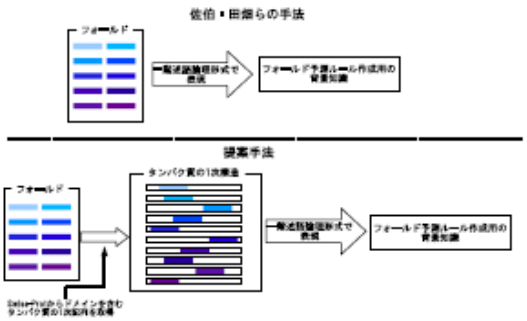

図１　たんぱく質一次構造の取得方法

フォールドの間に関連性があることを明らかにした．そして佐伯，田畑らは Turcotte らの手法を拡張し，たんぱく質のドメイン配列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 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．これにより，ドメインに対してもフォールド予測を行うことを可能とした．しかし，ドメインはたんぱく質立体構造の分類単位の一つであり，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，従来の手法でドメインから作成されたフォールド予測ルールでは一次構造しか分かっていないたんぱく質で解析，公開されているたんぱく質の一次構造に対してフォールドを予測する事が困難である．

２．研究の目的

本研究ではたんぱく質の一次構造からのフォールド予測を可能とする為に，フォールド予測ルール作成用の背景知識を，構造分類概念を考慮しつつたんぱく質の一次構造から作成する事で可能とする手法を提案する．

３．研究の方法

(1) ドメイン名の取得

Swiss-Prot 内のたんぱく質一次構造データを使用し，SCOP 内からはドメインのフォールド分類情報のみを使用し，目標のフォールドに登録されているドメインの名前のみを取得する．このドメイン分類データは(2)でのたんぱく質の一次構造取得時や，(4)での正事例作成時に使用する．また，同様に(4)で負事例を作成する為に，目標のフォールドを含むクラス以外の，クラスから同数ずつ，正事例の事例数と同じ数になるようにドメインの名前のみを取得する．

(2) たんぱく質の一次構造の取得

背景知識作成データに Swiss-Prot よりたんぱく質の一次構造データを使用する．SCOP で分類されたドメインの PDB 内のアミノ酸配列を使用するのではなく，フォールド内のドメインを含むたんぱく質の一次構造を Swiss-Prot から取得し，一階述語論理形式で表現し，背景知識として使用する(図 1)．Swiss-Prot 内のたんぱく質データからたんぱく質の一次構造を取得し，背景知識作成用のデータとして使用する．同様の事を(1)で得られた正事例，負事例作成用に得たドメイン名全てに対して行う事で，背景知識作成用に使用するたんぱく質一次構造データを取得する．

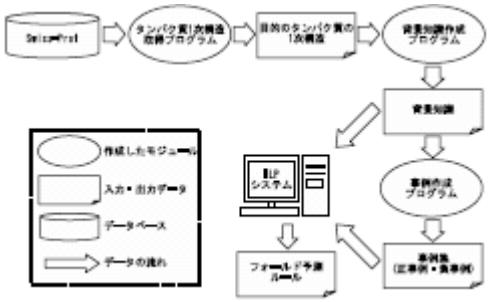

図２　作成したモジュールの流れ

(3) 一次構造からの二次構造の予測

(2)で取得したたんぱく質の一次構造から，さらに背景知識用に二次構造を予測する．一次構造からの二次構造予測には SSpro を用いる．SSpro とは双方向反復ニューラルネットワーク(BRNN)技術に基づく多種のたんぱく質二次構造予測の為のツールである．本手法ではたんぱく質の一次構造と，そこから SSpro を用いて予測された二次構造から次の(4)で背景知識を作成する．

(4) 背景知識の作成

(2)で取得したたんぱく質の一次構造と(3)でたんぱく質の一次構造から予測した二次構造を利用して，フォールド予測ルール作成用の背景知識を一階述語論理形式で表現する．たんぱく質一次構造に関する背景知識にはこの一次構造のたんぱく質の名前，たんぱく質の一次構造の長さに関する情報を使用，予測された二次構造に関する背景知識にはたんぱく質にあるαヘリックス，βシートの数，たんぱく質内の二次構造の位置，二次構造がアミノ酸プロリン，システイン，グリシンを含んでいるかどうかに関する情報を使用した．以上の情報から背景知識を作成し，(5)でフォールド予測用のルールを作成する．

(5) フォールド予測ルールの作成

(4)でたんぱく質の一次構造と，そこから予測した二次構造から作成した背景知識，そして事例から(5)では ILP システム GKS を用い，フォールド予測用のルールを作成する．

(6) モジュールの作成

(1)～(5)の処理を自動化するため，Ruby にて以下のようなモジュールを作成した．モジュール内の流れを図 2 に示す．

たんぱく質一次構造取得プログラムでは，入力されたたんぱく質の名前のリストから自動的に Web 上の Swiss-Prot データベースからたんぱく質の一次構造を取得する．背景知識作成プログラムでは取得した一次構造から SSpro を用いて二次構造を予測し，それらから先ほど説明した背景知識を作成する

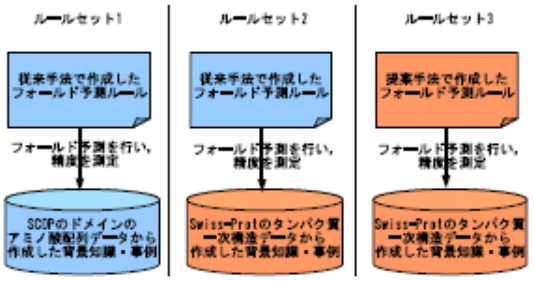

図３　実験で使用した３通りのルールセット

．事例作成プログラムは作成された背景知識から，正事例，負事例のリストを作成するプログラムである．以上のプログラムを作成し，本手法にかかる手間を大幅に削減した．

４．研究成果

(1) 実験

提案手法の有効性を検証する為に，本論文では三種類のルールセットで実験を行った(図 3)．１つ目のルールセットは，田畑らによる PDB 内のドメインのアミノ酸配列データを元にフォールド予測ルールを作成し，そのルールを元にドメインの配列データからフォールド予測を行う．佐伯，田畑らの手法をそのまま再現した形である．２つめのルールセットでは佐伯，田畑らの手法によって作成したルールで，たんぱく質の一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たんぱく質の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．３つめのルールセットでは提案手法により作成された，たんぱく質の一次構造から作成されたルールを用いて，たんぱく質の１次配列に対してフォールド予測を行う．ルールセット１とルールセット２の精度を比較する事で，従来手法のフォールド予測ルールがたんぱく質一次構造のデータに対して適応出来るかどうか，ルールセット２ の結果とルールセット３ の結果を比較する事で，本手法の有効性を検証した．

(2) 考察

① Toxins フォールドを用いた実験結果を図４に示す．ルールセット１から３の順番にそれぞれ Accuracy は 0.8, 0.69, 0.74, Precision は 0.77, 0.62, 0.74, Recall はそれぞれ 0.9, 0.7, 0.8 となった．ルールセット１と２の結果から従来手法で用いていたドメインのアミノ酸配列から作成したフォールド予測ルールはたんぱく質のフォールド予測には非常に不向きで有る事が分かった．又，ルールセット２と３の結果から，提案手法で作成したフォールド予測ルールの方が，従来手法のルールより Accuracy，Precision では精度の上昇がわずかだが，Recall においては大幅な上昇が確認出来る．やはりこのフォールドでも従来手法より正確なたんぱく質の一次

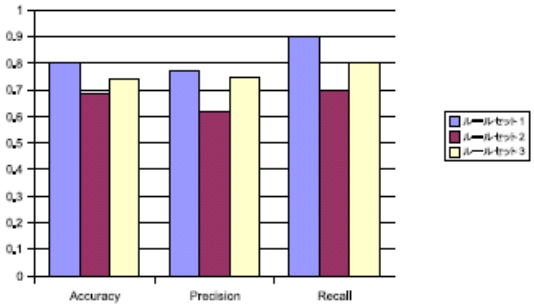

図４　フォールドから作成した背景知識を用いた，フォールド予測ルールの精度構造

データからのフォールド予測が可能になっている事が分かった．

② 次に作成されたルールを比較してみる．以下に Toxins フォールドから，提案手法によって獲得出来たルールの一部を具体的に示す．

```
fold('Toxins', A) :-adjacent(A, B, C, 5, h, h),
                    adjacent(A, D, E, 9, h, h),
                    coil(D, E, 3).
fold('Toxins', A) :-adjacent(A, B, C, 3, h, h),
                    adjacent(A, D, E, 5, h, h),
                    coil(D, E, 2).
fold('Toxins', A) :-adjacent(A, B, C, 3, h, h),
                    adjacent(A, D, E, 5, h, h),
                    coil(D, E, 5).
fold('Toxins', A) :-adjacent(A, B, C, 6, h, h),
                    adjacent(A, D, E, 9, h, h),
                    coil(D, E, 4).
```

全てのルールが“adjacent”と“coil”を組み合わせた物になっている事が分かった．従来手法で二次構造のデータとたんぱく質のフォールドの関係を ILP を用いてルール化した際も，“adjacent”と“coil”を一緒に含んだルールを多く観測していたので，フォールドをある程度正しく予測しているのではないかと考えられる．これらのルールを用いて，Swiss-Prot 内のたんぱく質の一次構造データに対してフォールド予測を行った結果，Accuracy は 69%，Precision は 67%，Recall は 75%であった．

同様のルールを従来手法で作成した場合，以下のルールが作成された．

```
fold('Toxins', A) :-adjacent(A, B, C, 1, e, h),
                    adjacent(A, D, E, 3, h, h),
                    adjacent(A, F, G, 7, h, h).
fold('Toxins', A) :-adjacent(A, B, C, 2, h, h),
                    adjacent(A, D, E, 3, h, h),
                    coil(B, D, 4).
fold('Toxins', A) :-adjacent(A, B, C, 7, h, h),
                    coil(B, C, 2).
```

提案手法と比較すると，１つのルール内に“adjacent”と“coil”を同時に含んだルールが少ない事が分かった．又，これらのルールを用いて，Swiss-Prot 内のたんぱく質の一

表 1　実験 1 で使用した背景知識の述語数

述語名	PDB/正	PDB/負	Swiss-Prot/正	Swiss-Prot/負
fir_struc	40	40	40	40
has_cys	62	97	123	68
has_gly	206	191	242	252
has_pro	112	134	173	118
len	40	40	40	40
nb_alpha	40	40	40	40
nb_beta	40	40	40	40
sec_struc	884	888	1168	973
sst	884	888	1169	973
unit_t	884	888	1169	973

表 2　実験 2 で使用した背景知識の述語数

述語名	PDB/正	PDB/負	Swiss-Prot/正	Swiss-Prot/負
fir_struc	56	56	56	56
has_cys	7	98	36	210
has_gly	119	181	155	286
has_pro	39	172	77	198
len	56	56	56	56
nb_alpha	56	56	56	56
nb_beta	56	56	56	56
sec_struc	950	858	690	1549
sst	950	858	690	1549
unit_t	950	858	690	1549

次構造データに対してフォールド予測を行った結果，Accuracy は 38%，Precision は 33%，Recall は 25%を記録した．この事からも提案手法で作成したルールは，従来手法では適応出来なかった，たんぱく質の一次構造データに対するフォールド予測の適応を可能にしたと言える．

③ 従来手法と提案手法で作成した各実験の背景知識の述語数を，表 1，表 2 で示し，比較した．どちらも事例毎に 1 つの述語が出来る“fir_stru”，“len”，“nb_alpha”，“nb_beta”以外は，従来手法の PDB 内のドメインデータから作成した背景知識述語数(表左 2 列) と，提案手法の Swiss-Prot 内のたんぱく質 1 次配列データから作成した背景知識述語数(表右 2 列)とでは大きく差が出ている事が分かる．

この差を考慮しつつルールを作成した事で，従来手法のフォールド予測ルールでは適応出来なかったたんぱく質一次構造の予測が可能になったのではないかと考える．Recall の上昇は，こういった背景知識の違いを考慮しルール作成を行った結果，新しいルールが作成出来，ルールの網羅性である Recall を上昇させる事を可能にしたのではないかと考えられる．したがって提案手法は従来手法より，たんぱく質の一次構造に対するフォールド予測ルールに有効な背景知識を作成出来たと考える．

(3) 成果

以上の実験から本手法の有効性が示せた．特に Recall の上昇が顕著であることから，従来手法の背景知識では表現しきれなかったたんぱく質の一次構造の特徴を，提案手法で表現する事が可能になったと考えられる．

本論文で作成した，全ての BCL2 たんぱく質を含む Toxins フォールドの予測結果と，その予測ルールが表現した特徴は，たんぱく質の機能解析の実験対象を大幅にしぼる事が出来，シロイヌナズナの BCL2 発見に役立つ事が期待される．又，シロイヌナズナ以外にも現在，多くの生物のたんぱく質データが提案手法で扱った一次構造である FASTA 形式で公開されている．本手法では従来手法では適応出来なかったこういった形のデータに対するフォールド予測を可能にする事が出来たといった点でも，非常に有効であったと考える．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

Predicting Protein Folding from Primary Structure Using Inductive Logic Programming. Tohgoroh Matsui, Hayato Ohwada, Kazuyuki Kuchitsu. Proceedings of the Workshop on Knowledge, Language, and Learning in Bioinformatics(KLLBI 2008), pp.36--45(2008). 査読有．

〔学会発表〕（計 2 件）

① ILPを用いたBCL2 ファミリーのフォールド予測．河村真平，松井藤五郎，賀屋秀隆，大和田勇人，朽津和幸．2007 年度人工知能学会（第 21 回）全国大会講演論文集，3C6-2 (2007)．査読無．

② 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．第 70 回情報処理学会全国大会講演論文集，第 4 分冊，5ZJ-3 (2008)．査読無．

６．研究組織

(1)研究代表者

大和田　勇人（OHWADA HAYATO）
東京理科大学・理工学部経営工学科・教授
研究者番号：30203954

(2)研究分担者

松井　藤五郎（MATSUI TOHGOROH）
東京理科大学・理工学部経営工学科・助教
研究者番号：90366443

(3)連携研究者