# 結晶化環境における pH 値を考慮した SVM によるタンパク質結晶化の予測

片岡 義雅[†1]　野口 保[†2]　百石 弘澄[†3]　小林 大輔[†4]　山名 早人[†5]

†1 早稲田大学 基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1
†2 産業技術総合研究所 生命情報工学研究センター 〒135-0064 東京都江東区青海 2-4-7
†3, 4 早稲田大学 基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
†5 早稲田大学 理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

E-mail: [†1, †3, †4, †5]{y.kataoka, hyaccoku, d_kobayashi, yamana}@yama.info.waseda.ac.jp
[†2]noguchi-tamotsu@aist.go.jp

**あらまし** タンパク質立体構造決定は，その機能を解明する上で重要である．タンパク質立体構造決定に広く用いられている方法として X 線結晶解析があるが，タンパク質の結晶が必要であり，最適な結晶化条件を発見するための実験に多大な時間やコストが必要となる．このため，タンパク質の結晶化を予測する研究が進められている．タンパク質の結晶化はその溶解度に依存しており，溶解度は pH と深い関係を持つが，既存研究では考慮されていない．本研究ではタンパク質の pH を結晶化の要因として考慮し，特徴量のひとつとして pH，学習器として SVM を用いた結晶化予測を行うことにより，pH を使用しない予測手法に比べ，最大で 0.75%の精度向上が見られ，この差異は T 検定によって統計学的に有意であることが確認できた．

**キーワード** タンパク質結晶化，pH，SVM

# Prediction of Protein Crystallization Using SVM by Considering the pH Value of the Crystallization Circumstances

Yoshimasa KATAOKA[†1]　Tamotsu NOGUCHI[†2]　Hiroto HYACCOKU[†3]

Daisuke KOBAYSHI[†4]　Hayato YAMANA[†5]

†1 School of Fundamental Science and Engineering, Waseda University 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555, Japan

†2 Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology
AIST Tokyo Waterfront Bio-IT Research Building, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

†3, 4 Graduate School of Fundamental Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, Japan

†5 Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, Japan

E-mail: [†1, †3, †4, †5]{y.kataoka, hyaccoku, d_kobayashi, yamana}@yama.info.waseda.ac.jp
[†2]noguchi-tamotsu@aist.go.jp

**Abstract** Determining the protein structure is important to understand the protein function. X-ray crystallography, which is the most popular method to determine the protein structure, needs a protein crystal. However, the experiments to find the optimal circumstances require huge amounts of time and also cost high. Therefore, there are many researches about the prediction of protein crystallization.

Protein crystallization depends on the solubility, and the solubility has deep involvement with the pH value but no researches are using the pH value as a feature. In this paper, we consider the pH value as a factor of protein crystallization and use it as one of the feature vector of the Support Vector Machine. As a result, there were 0.75% increase in accuracy and this result is statistically significant according to the result of the T-test.

**Keyword** Protein Crystallization, pH, SVM

## 1. はじめに

タンパク質は水に次いで人体内に数多く存在する物質であり，その機能を解明することは創薬分野をはじめとする様々な場面で有益である．タンパク質の機能は各々に固有な立体構造に依存しており，立体構造を決定することは機能解明へと繋がる．しかし，立体構造決定は非常に難しいプロセスである．現在ではタンパク質のアミノ酸配列データが NCBI (National Center for Biotechnology) [1]，EBI (European Bioinformatics Institute) [2]，TargetDB (Target Database for Structural Biology) [3]などのデータベースに 1000 万個以上登録されている一方，立体構造を含むデータを扱う PDB (Protein Data Bank) [4]では 6 万個程度の立体構造データしか登録されていない．タンパク質の立体構造を決定する方法として，現在では主に Nuclear Magnetic Resonance(核磁気共鳴法，NMR)，X 線結晶解析，電子顕微鏡の 3 種類が用いられている．X 線結晶解析は NMR や電子顕微鏡に比べて高速，低コスト，高分解能で立体構造が決定できるため，タンパク質の立体構造を解明する構造ゲノミクス分野では X 線結晶解析が用いられることが多く[5], PDB に登録されている立体構造が決定されたデータの 85%以上が X 線結晶解析によるものである．X 線結晶解析では NMR において必要としないタンパク質の結晶が必要不可欠であるが,すべてのタンパク質が結晶化するわけではない．また結晶化するタンパク質でも温度や pH が変わることで結晶化しなくなる場合があるため，最適な結晶化条件を発見するための生化学実験は多大な時間と費用を要する．これより，タンパク質が結晶化するかどうかを予測することは結晶化実験の効率化，そして実験におけるコスト削減へと繋がる．

タンパク質の結晶化予測において，現在では機械学習が用いられる場合があるが[5][6], 機械学習を行う場合，結晶化するタンパク質と結晶化しないタンパク質のデータが必要である．結晶化するタンパク質のデータは一般的に X 線結晶解析で立体構造決定されたタンパク質を使用する．一方，結晶化しないタンパク質はデータ選びが難しい．結晶化しないと実験で立証されたタンパク質を使用するのが理想的ではあるが，こういったデータは未だ十分に存在しない[7]．タンパク質は実験環境の変化によって結晶化の可否が変わるためである．そのため，既存研究ではある実験環境において結晶化しないという前提の下, NMR のみで立体構造決定されたタンパク質を結晶化しないタンパク質として学習を行っている[5][6].

機械学習による結晶化予測の手法としては，CRYSTALP[5]，CRYSTALP2[6]などが提案されている．これらの手法では機械学習に使用する特徴量としてタンパク質のアミノ酸組成，等電点 (pI)，疎水性などが用いられている．CRYSTALP では学習器として単純ベイズ，特徴量としてタンパク質のアミノ酸配列における各アミノ酸，ジペプチドの出現確率を使用している．CRYSTALP2 では特徴量として CRYSTALP で使用しているものに加え，トリペプチドの出現確率，等電点，GES の疎水性指標[8]を使用し，学習器としては SVM を使用している．しかし，これらの手法ではタンパク質の結晶化において重要な要因となる pH について考慮されていない．pH とは物質の酸性，アルカリ性の度合いを示す 1~14 の値である．タンパク質は pH が変化することでアミノ酸残基の状態が変化するため，溶解度も変化する[9]．タンパク質の溶解度は pH が pI 付近を取る場合が最も低く，pH が pI から離れると増加する．タンパク質の結晶化は溶解度に大きく依存するため,結晶化は pH と深い関係を持つと考えられる．これより，pH と pI の差を特徴量として使用することは有用であると考えられる．今までは機械学習をする上で十分な量のデータが公開されていなかったために pH は考慮されていなかった[10]．しかし近年はデータ量が増加し，機械学習を行う上で十分な量のデータが存在するようになった．

そのため，本研究では pH と pI の差分を特徴量のひとつとして使用する．ここで，特徴として使用する pH 値は結晶化に成功した実験環境下，または結晶化に失敗した実験環境下での値とする．その他の特徴量としてはアミノ酸配列と Zvelebil らの真理表[11]を組み合わせることによって得られるアミノ酸配列の物理化学的性質，GES の疎水性指標を使用する．また，学習器として SVM を用いる．

本稿の構成は以下の通りである．第 2 節では関連研究について述べる．そして第 3 節で提案手法，第 4 節で提案手法の評価実験について述べ，第 5 節でまとめを述べる．

## 2. 関連研究

タンパク質の結晶化予測は様々な特徴量を用いて行われている．本節では CRYSTALP，CRYSTALP2 について説明する．

### 2.1. CRYSTALP

Chen らはアミノ酸配列のみを特徴量として考慮した予測を行う CRYSTALP[5]と呼ばれるシステムを開発した．特徴量としては，アミノ酸配列において，20 種類のアミノ酸残基の出現確率，そして 2 つの連続したアミノ酸ペアであるジペプチド，インターバル P (P=1~4) を置いたアミノ酸ペアの出現回数を使用した．アミノ酸残基を $A_i$，$A_j$，インターバルを - で表すとす

ると，P (P=1~4) を置いたアミノ酸ペアとは，$A_i$-$A_j$，$A_i$--$A_j$，$A_i$---$A_j$，$A_i$----$A_j$なるアミノ酸ペアのことである．ここで，インターバル部分は 20 種類のアミノ酸のどれが当てはまってもいい．これら 2020 次元の特徴量を，CFSS[12]と 10 クロスバリデーションを用いた特徴選択によって最終的に 46 次元まで削減し，単純ベイズの学習器を用いて予測を行った．CRYSTALP では，配列長が 46 以上 200 以下のタンパク質のみを学習データとして使用したため，配列長が 46 以上 200 以下のタンパク質しか予測できない．これは結晶化しないタンパク質として使用している NMR で立体構造決定されたタンパク質は比較的配列長が短いものが多いからである．

### 2.2. CRYSTALP2

Kurgan らは CRYSTALP で使用されたアミノ酸残基，ジペプチドに加えて，3 つの連続したアミノ酸から成るトリペプチド，Goldman，Engelman，Steitz (GES) の疎水性指標，等電点 (pI) を特徴量として利用した CRYSTALP2[6]と呼ばれるシステムを開発した．トリペプチドはジペプチド同様インターバルも考慮し，$A_iA_jA_k$，$A_iA_j$-$A_k$，$A_i$-$A_jA_k$，$A_i$-$A_j$-$A_k$の 4 つのトリペプチドを特徴量として使用した．ここで，CRYSTALP ではジペプチドの出現回数を利用したが，CRYSTALP2 ではジペプチド，トリペプチドの出現確率を利用した．GES の疎水性指標とは Goldman らによって定められた各アミノ酸の疎水性を表す値であり，表 1 の通りである．CRYSTALP2 では各アミノ酸配列において，すべてのアミノ酸残基について疎水性の値を足し合わせ，それを配列長で割ることによって疎水性の平均値を求め，それを特徴量として使用した．これらの特徴量を CRYSTALP と同様に，CFSS と 10 クロスバリデーションを用いた特徴選択によって最終的に 88 次元まで削減し，SVM を用いた予測を行った．CRYSTALP2 は CRYSTALP と違い，長さの制限がないため，配列長が 30 以上のタンパク質に関して予測が可能である．

**表 1 GES の疎水性指標**

| A | C | D | E | F | G | H |
|---|---|---|---|---|---|---|
| -1.6 | -2.0 | -2.1 | -2.6 | -3.7 | -1.0 | -3.0 |

| I | K | L | M | N | P | Q |
|---|---|---|---|---|---|---|
| -3.1 | -3.7 | -2.8 | -3.4 | -2.2 | -1.8 | -2.9 |

| R | S | T | V | W | Y |
|---|---|---|---|---|---|
| -4.4 | -1.6 | -2.2 | -2.6 | -4.9 | -3.7 |

### 2.3. 提案手法との関係性

CRYSTALP, CRYSTALP2 ではアミノ酸配列におけるアミノ酸残基の構成やタンパク質の疎水性を考慮した結晶化の予測を行っている．しかし，タンパク質の結晶化の一因である pH については，今まで十分な量のデータが公開されていなかったため，CRYSTALP，CRYSTALP2 では考慮されていない．近年，十分な量のデータが公開されてきたため，提案手法では pH を特徴量として使用した学習をすることで予測精度の向上を図る．

## 3. 提案手法

本節では提案手法について説明する．まず 3.1 項でデータセットから抽出する特徴量について述べる．その後 3.2 項で，特徴量ベクトルの削減について述べ，最後に 3.3 項で抽出した特徴量を用いた学習方法について述べる．

### 3.1. 特徴量

本項では，データセットから抽出する特徴量であるタンパク質の物理化学的性質，GES の疎水性指標，pH-pI の差分の，計 3 種類の特徴量について述べる．

#### 3.1.1. 物理化学的性質

タンパク質の物理化学的性質を特徴量として用いるにあたっては，Zvelebil らの真理値表[11]を使用する．Zvelebil らは各アミノ酸残基が持つ性質を真理値表で表した．Zvelebil らの真理値表を表 2 に示す．表の各行はアミノ酸残基の性質を表す．アミノ酸残基が持ちうる性質は全部で 10 個あり，それぞれの性質を持てば 1，持たなければ 0 と表記する．提案手法では，データセットに応じて，データセット中の全てのアミノ酸配列において，20 種類のアミノ酸残基の出現回数，もしくは出現確率を数え，それぞれの性質の出現回数，出現確率とする．例えば，「Positive (正電荷) 」の性質を持つアミノ酸残基は H，K，R であるため，それぞれの出現回数を合計した値が性質「Positive」の出現回数となる．アミノ酸残基を表す性質は 10 個であるため，10 次元の特徴量となる．出現確率も同様の方法で計算し，各性質の出現確率とする．

また，各アミノ酸残基の出現回数の他に，CRYSTALP や CRYSTALP2 で考慮されているジペプチドとインターバル P (P=1~4) を置いたアミノ酸ペアについても考慮した．インターバルを置いたアミノ酸ペアを考慮するのはタンパク質の折りたたみ構造に影響を及ぼすことがわかっているからである．ジペプチド，アミノ酸ペアの場合，2 つのアミノ酸残基があるため，それぞれの性質を考慮する．例えば，AK なるジペプチドが

あった場合，A は「Hydrophobic (疎水性)」，K は「Positive」の性質を持つ．そのため，ジペプチド AK の性質は「Hydrophobic-Positive」となる．Zvelebil の真理値表では各アミノ酸残基において 10 個の性質を表しているため，2 つのアミノ酸残基を持つジペプチドの場合は 100 次元の特徴量となる．そして，インターバルを置いたアミノ酸ペアでもそれぞれ 100 次元となるため，インターバル P (P=1~4) を考慮すると 400 次元の特徴量となる．よって，アミノ酸残基，ジペプチド，インターバルを置いたアミノ酸ペアを考慮すると，合計で 510 次元の特徴量が得られる．

**表 2 Zvelebil の真理値表[11]**

| | Hyo | Hyi | Cha | Pos | Neg |
|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 | 0 |
| C | 1 | 0 | 0 | 0 | 0 |
| D | 0 | 1 | 1 | 0 | 1 |
| E | 0 | 1 | 1 | 0 | 1 |
| F | 1 | 0 | 0 | 0 | 0 |
| G | 1 | 0 | 0 | 0 | 0 |
| H | 1 | 0 | 1 | 1 | 0 |
| I | 1 | 0 | 0 | 0 | 0 |
| K | 1 | 0 | 1 | 1 | 0 |
| L | 1 | 0 | 0 | 0 | 0 |
| M | 0 | 1 | 0 | 0 | 0 |
| N | 0 | 1 | 0 | 0 | 0 |
| P | 0 | 1 | 0 | 0 | 0 |
| Q | 0 | 1 | 0 | 0 | 0 |
| R | 0 | 1 | 1 | 1 | 0 |
| S | 0 | 1 | 0 | 0 | 0 |
| T | 1 | 0 | 0 | 0 | 0 |
| V | 1 | 0 | 0 | 0 | 0 |
| W | 1 | 0 | 0 | 0 | 0 |
| Y | 1 | 0 | 0 | 0 | 0 |

Hyo:Hydrophobic, Hyi:Hydrophil, Cha:Charged, Pos:Positive, Neg:Negative

| | Aro | Ali | Tin | Sma | Pol |
|---|---|---|---|---|---|
| A | 0 | 0 | 1 | 1 | 0 |
| C | 0 | 0 | 0 | 1 | 0 |
| D | 0 | 0 | 0 | 1 | 1 |
| E | 0 | 0 | 0 | 0 | 1 |
| F | 1 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 1 | 1 | 0 |
| H | 1 | 0 | 0 | 0 | 1 |
| I | 0 | 1 | 0 | 0 | 0 |
| K | 0 | 0 | 0 | 0 | 1 |
| L | 0 | 1 | 0 | 0 | 0 |
| M | 0 | 0 | 0 | 0 | 0 |
| N | 0 | 0 | 0 | 1 | 1 |
| P | 0 | 0 | 0 | 1 | 0 |
| Q | 0 | 0 | 0 | 0 | 1 |
| R | 0 | 0 | 0 | 0 | 1 |
| S | 0 | 0 | 1 | 1 | 1 |
| T | 0 | 0 | 0 | 1 | 1 |
| V | 0 | 1 | 0 | 1 | 0 |
| W | 1 | 0 | 0 | 0 | 1 |
| Y | 1 | 0 | 0 | 0 | 0 |

Aro:Aromatic, Ali:Aliphatic, Tin:Tiny, Sma:Small, Pol:Polar

#### 3.1.2. GES の疎水性指標

提案手法では CRYSTALP2 と同様に，各アミノ酸配列において，すべてのアミノ酸残基における疎水性の値の総和を配列長で割り，平均を求めたものを特徴量として使用する．

#### 3.1.3. pH-pI の差分

提案手法では pH を独立した特徴量として使用するのではなく，pH と pI の差分を特徴量として使用する．これは pH と pI の差分がタンパク質の溶解度に影響を与え，提案手法で使用したデータセットではその傾向が顕著に出たためである．pI の計算には ExPASy Proteomics Server[16]を用いる．

### 3.2. 特徴量ベクトルの削減

3.1 項の方法を用いて，タンパク質の物理化学的性質 (510 次元)，GES の疎水性指標 (1 次元)，そして pH-pI の差分 (1 次元) の，合計 512 次元の特徴量を算出した後，算出した特徴量を使用した機械学習を行い，結晶化を予測する．ここで，機械学習は特徴量の数が多すぎると過学習を引き起こしてしまうため，提案手法では特徴量ベクトルの次元削減を行う．提案手法ではすべての特徴量を考慮した次元削減を行うため，次元削減には主成分分析を用いる．ここで，主成分分析はすべての特徴量を用いて行う場合と，特徴量を分割して行う場合では結果が異なり，学習結果も異なる．提案手法では 3 つの特徴量を使用するため，以下の 3 通りの主成分分析を行い，最も結果が良い主成分分析の方法を採用する．

- 512 次元すべてを使用した場合 (①)
- pH-pI の差分以外の 511 次元を使用した場合 (②)

- 物理化学的性質の 510 次元のみを使用した場合(③)

②，③の場合は，主成分分析をした後に，主成分分析に使用していない特徴量を結合する．

### 3.3. SVM による学習

3.2 項で述べた主成分分析によって得られた特徴量を使用し，タンパク質の結晶化予測を行う．学習にはガウシアンカーネルに基づく SVM を用いる．SVM を用いる際には，カーネル，ソフトマージンのパラメータを決定する必要があるが，それぞれのパラメータを 0.01，0.05，0.1，0.5，1，5，10，50，100，500 と変化させ，学習を行う．

## 4. 評価実験

本節では，提案手法の評価実験について示す．まず，4.1 項で実験データの概要について説明する．次に 4.2 項で提案手法による実験，4.3 項で提案手法に対する比較手法による実験について説明する．そして 4.4 項で T 検定による実験について述べ，最後に 4.5 項で実験結果を示す．

### 4.1. 実験データ概要

#### 4.1.1. 実験に使用するデータの抽出

結晶化するタンパク質のデータをポジティブデータ，結晶化しないタンパク質のデータをネガティブデータと定義する．ポジティブデータは TargetDB において，X 線結晶解析に使用することができる結晶が生成できたときに付けられる「Diffraction-quality Crystals」のラベルが付けされたタンパク質の中で，PDB へのリンクがあるもの，ネガティブデータは PDB に登録されたタンパク質の中で，NMR でのみ立体構造決定されたものとした．2 つのデータセットを非冗長なデータセットとするため，CD-HIT[13][14]を用いて，アミノ酸配列が 30%以上類似するタンパク質の集合については，代表となるタンパク質を残し，類似するその他のタンパク質をすべて除去した．これをデータセット間でも行い，ポジティブデータセット，ネガティブデータセットを類似度が 30%未満の非冗長なデータセットとした．

また，PDB に登録されているタンパク質には，精製しやすくするためにヒスチジンタグが付けられたものがある．これは本来のタンパク質の性質を損なってしまう．よって，Carson ら[15]によって構造に影響があると証明されたヒスチジンタグをすべてのタンパク質から除去した．さらに提案手法では配列長を考慮した学習を行う．これは結晶化しないタンパク質として使用している NMR で立体構造決定されたタンパク質は比較的配列長が短いものが多く，学習する際に 3.1 項で説明した特徴量の性質以外に，タンパク質の配列長が大きく影響してしまうためである．よって提案手法では配列長の制限として，以下の 2 通りのパターンを作成した．

- 配列長が 30 以上のタンパク質のみを使用する場合
- 配列長が 46 以上 200 以下のタンパク質のみを使用する場合

これは CRYSTALP，CRYSTALP2 で考慮されている配列長である．この結果，PDB から得られたデータの数は表 3 のようになった．

**表 3 PDB から得られたデータ数**

| 配列長 | ポジティブデータセット | ネガティブデータセット |
|---|---|---|
| 30 以上 | 2899 | 2708 |
| 46 以上 200 以下 | 1345 | 2376 |

一般に，タンパク質は pH-pI の差分が小さいほど安定した結晶ができると言われている[9]．配列長が 30 以上のタンパク質を使用する場合のポジティブデータセット，ネガティブデータセットにおいて，pH-pI の差分とデータ量の関係は図 1，図 2 のようになった．

図 1，図 2 より，pH-pI の差分は，結晶化しないタンパク質データよりも結晶化するタンパク質データの方がピーク時に小さい値をとることが確認できる．両者に明らかな傾向の違いが認められるため，考案した特徴量が結晶化に関係していることが確認できる．

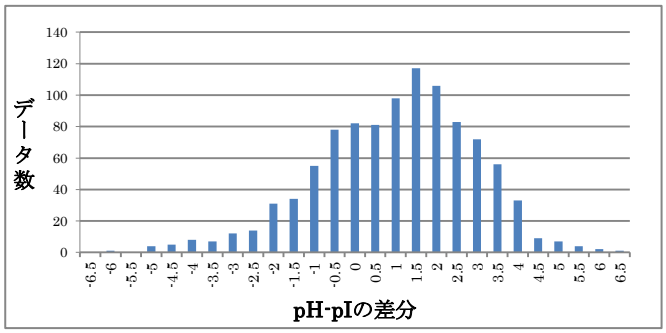


**図 1 ポジティブデータの pH-pI の差分**

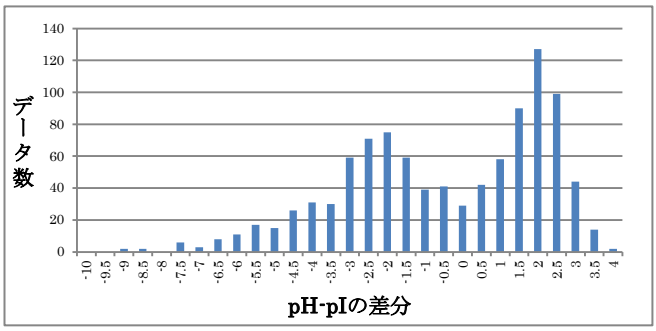


図 2 ネガティブデータの pH-pI の差分

#### 4.1.2. 学習用データと評価用データの作成

表 3 に示したデータセットから，ランダムサンプリングを行うことによって複数個のポジティブデータセット，ネガティブデータセットを作成し，それらを総当たりで組み合わせることによって学習用データセット，評価用データセットを作成する．各データセットのデータ数とデータセット数を表 4，表 5 に示す．

表 4 サンプリングしたデータ数

| 配列長 | ポジティブデータセット | ネガティブデータセット |
|---|---|---|
| 30 以上 | 500 | 500 |
| 46 以上 200 以下 | 400 | 400 |

表 5 作成したデータセット数

| 配列長 | ポジティブデータセット | ネガティブデータセット |
|---|---|---|
| 30 以上 | 5 | 5 |
| 46 以上 200 以下 | 3 | 5 |

配列長が 30 以上の場合の学習用データセットの作成について説明する．学習用データセットは表 5 で示したポジティブデータセット，ネガティブデータセットからあるデータセットをそれぞれ 1 つずつ選び，この 2 つを結合することによって作成する．次に評価用データセットについて説明する．評価用データセットは，学習用データセットとして使用しなかった残りのポジティブデータセット，ネガティブデータセットを総当たりで組み合わせることによって作成する．これより，1 つの学習用データセットに対して 16 個の評価用データセットが作成される．配列長が 46 以上 200 以下の場合にも同様の方法でデータセットを作成し，1 つの学習用データセットに対して 8 個の評価用データセットが作成される．学習用データセットは配列長が 30 以上の場合に 25 個，配列長が 46 以上 200 以下の場合に 15 個作成されるので，それぞれ 400 回，120 回の予測実験を行うことができる．複数回の予測実験から得られた予測結果を T 検定で評価することによって，特徴量として pH を使用する提案手法の有用性を統計学的に評価する．

### 4.2. 提案手法による実験

提案手法に使用する特徴量の抽出は 3.1 項で説明した方法によって行う．ここで，タンパク質の物理化学的性質において，配列長が 30 以上のタンパク質を使用する場合は各アミノ酸残基の出現確率，配列長が 46 以上 200 以下のタンパク質を使用する場合は各アミノ酸残基の出現回数を数える．アミノ酸配列長に上限・下限を設けた場合，各アミノ酸は似た配列長を持つため，学習する際に配列長の影響は受けない．しかし，配列長の上限を設けない場合，ポジティブデータセットの方が比較的配列長が長いタンパク質を多く含んでいるため，学習する際に配列長の影響を受けてしまう．よって配列長の上限を設けない場合は各アミノ酸残基の出現確率を考慮することにより，配列長の影響をなくす．

実験は 3.2 項で説明した①，②，③の主成分分析を行い，合計 6 通り行う．それぞれの配列長パターンにおいて最も良い結果を示した主成分分析の方法をその配列長パターンにおける主成分分析の方法として採用する．

### 4.3. 評価手法による実験

提案手法の比較として，pH-pI の差分を特徴量として使用しない場合の実験も行う．比較手法では特徴量として物理化学的性質，GES の疎水性指標のみを使用し，提案手法と同様の方法による学習を行う．比較手法では，以下の 2 通りの主成分分析を行う．

- 物理化学的性質と GES の疎水性指標の 511 次元を使用した場合 (④)
- 物理化学的性質の 510 次元のみを使用した場合 (⑤)

⑤の場合は主成分分析をした後に，GES の疎水性指標から得られる特徴量を結合する．

### 4.4. T 検定

T 検定では，2 つの母集団 A，B からランダムサンプリングによって標本を抽出し，2 つの標本を比べることによって母集団に差異があるかを調べる．本実験における 2 つの母集団とは，特徴量として pH-pI の差分を用いる場合の実験結果と用いない場合の実験結果である．配列長が 30 以上のタンパク質を使用した実験では提案手法，比較手法で各 400 通り行った実験からそれぞれ 60 通りずつ，配列長が 46 以上 200 以下のタン

パク質を使用した実験では提案手法，比較手法で各120通り行った実験からそれぞれ30通りずつの実験結果をランダムサンプリングによって抽出し，T 検定に使用する．母集団 A, B におけるそれぞれの標本数を$n_A$, $n_B$，自由度を$\emptyset_A$，$\emptyset_B$，標本の平均を$\mu_A$，$\mu_B$，標本の標準偏差を$S_A$，$S_B$とすると，2 つの標本を合わせたときの標準偏差 S，母集団に差異があるかを調べる際に用いる t は以下の式で求めることができる．ここで，自由度とは標本数から 1 を引いた値である．

$$S = \sqrt{\frac{\emptyset_A {S_A}^2 + \emptyset_B {S_B}^2}{\emptyset_A + \emptyset_B}} \quad (1)$$

$$t = \frac{\mu_A - \mu_B}{S \times \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \quad (2)$$

自由度によって定められる t の値によって，2 つの母集団の差異が有用性を調べる．本実験では提案手法と比較手法からそれぞれ 30, 60 通りの標本をサンプリングするため，全体の標本数は配列長が 30 以上のタンパク質を使用する場合は 120，配列長が 46 以上 200 以下のタンパク質を使用する場合は 60 となる．よって自由度はそれぞれ 118, 58 となる．これは自由度が 120, 60 の場合と考えてよい．よって，自由度が 60，120 の場合における有意水準 1%，5%，10%の場合の t の値を表 6 に示す[17].

**表 6 自由度が 60 と 120 の場合の t の値**

| 自由度 | 有意水準 1% | 有意水準 5% | 有意水準 10% |
|---|---|---|---|
| 60 | 2.660 | 2.000 | 1.671 |
| 120 | 2.617 | 1.980 | 1.658 |

### 4.5. 実験結果

提案手法の評価はテストデータを用いた実験において，(3)から得られる *Accuracy* によって行う．ここで，正しく予測できたポジティブデータ数を TP，間違った予測をしたポジティブデータ数を FP とする．同様の基準で，ネガティブデータの場合はそれぞれ TN, FN とする．

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

#### 4.5.1. 最高 Accuracy の比較結果

配列長が 30 以上のタンパク質を使用する場合，配列長が 46 以上 200 以下のタンパク質を使用する場合のそれぞれにおいて，①~⑤の実験結果は表 7，表 8 のようになった．表には *Accuracy* が最も高い場合のガウスカーネルのパラメータ *g*，ソフトマージンのパラメータ *C*，SVM での学習の際にデータセットの正規化の有無を記入する．表 7，表 8 より，主成分分析において最適な方法を選択すれば，どちらの実験においても pH-pI の差分を使用した場合の方が使用しない場合よりも *Accuracy* は高いことがわかる (表 7 では②と④，表 8 では①と⑤を比較した)．表 7，表 8 において，pH-pI の差分を使う場合と使わない場合の最大精度結果を図 3 に示す．これより，pH-pI の差分は特徴量として有用であると言える．

**表 7 配列長 30 以上の場合**

| 主成分分析 | pH | g | C | 正規化 | Accuracy(%) |
|---|---|---|---|---|---|
| ① | 有 | 50 | 10 | 無 | 82.0 |
| ② | 有 | 1 | 500 | 無 | 83.1 |
| ③ | 有 | 0.05 | 0.05 | 有 | 83.0 |
| ④ | 無 | 0.05 | 0.1 | 有 | 82.7 |
| ⑤ | 無 | 0.05 | 0.1 | 有 | 82.0 |

**表 8 配列長 46 以上 200 以下の場合**

| 主成分分析 | pH | g | C | 正規化 | Accuracy(%) |
|---|---|---|---|---|---|
| ① | 有 | 10 | 500 | 無 | 78.08 |
| ② | 有 | 1 | 500 | 無 | 76.33 |
| ③ | 有 | 0.5 | 500 | 無 | 76.42 |
| ④ | 無 | 10 | 500 | 無 | 77.25 |
| ⑤ | 無 | 0.05 | 10 | 有 | 77.33 |

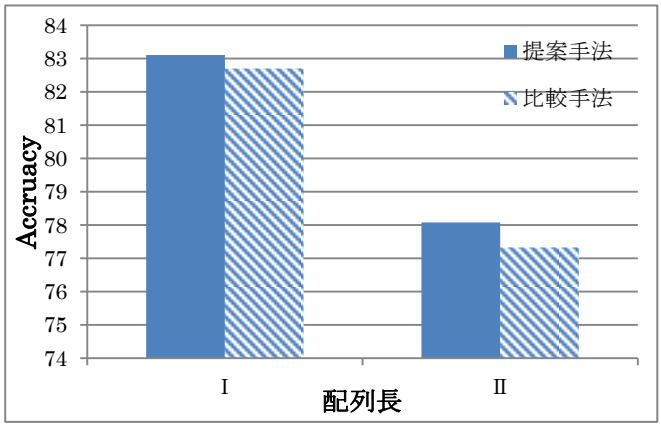


**図 3 提案手法と比較手法の最高 Accuarcy**

ここで，図 3 において，I は配列長 30 以上のみのタンパク質を使用した場合，II は配列長 46 以上 200 以下のタンパク質を使用した場合の実験結果である．

#### 4.5.2. T 検定の結果

配列長が 30 以上のタンパク質を使用した実験では提案手法，比較手法で各 400 通り行った実験からそれ

ぞれ 60 通りずつ，配列長が 46 以上 200 以下のタンパク質を使用した実験では提案手法，比較手法で各 120 通り行った実験からそれぞれ 30 通りずつの実験結果をランラムサンプリングによって抽出し，T 検定に用いる各実験結果の標本とした．その結果，各実験における標本の平均を$\mu_A$，$\mu_B$，標本の標準偏差を$S_A$，$S_B$は表 9 のようになった．

**表 9 各配列長における平均値と標準偏差**

| 配列長 | $\mu_A$ | $\mu_B$ | $S_A$ | $S_B$ |
|---|---|---|---|---|
| 30 以上 | 80.335 | 80.015 | 0.998 | 0.871 |
| 46 以上 200 以下 | 79.775 | 79.025 | 0.928 | 0.927 |

表 9 の値を式(1)，(2) に代入することによって，各配列長パターンにおける S, t は表 10 のようになった．

**表 10 各配列長における S 値と t 値**

| 配列長 | S | t |
|---|---|---|
| 30 以上 | 0.937 | 1.872 |
| 46 以上 200 以下 | 0.927 | 3.133 |

表 6，表 10 より，配列長が 30 以上のタンパク質を使用する場合は有意水準 10%，配列長が 46 以上 200 以下のタンパク質を使用する場合は有意水準 1%を満たしている．よって，最高 *Accuracy* の比較によって得られた提案手法と比較手法の差異は，統計学的に有用であることがわかる．

## 5. 今後の課題

本研究では特徴量ベクトルの次元削減方法として主成分分析を用いたが，LSI や SVD などの方法を用いた次元削減によって結果が変わると思われる．また，次元削減を行わず，512 次元の特徴量をすべて用いた場合も実験結果は変わると思われる．よって，新たな次元削減手法を用いた実験が今後の課題として考えられる．

また，本稿で行った実験では提案手法が有用であると判断できたが，既存手法との比較を行い，その結果を考察することによってより有用な検討ができるため，他の手法との比較を今後の課題として挙げる．

## 6. まとめ

本稿ではタンパク質の結晶化予測を行う際，特徴量として pH を使用する手法を提案した．実験の結果，特徴量の次元削減の方法によっては pH を使用しない場合よりも良い予測精度が得られた．また，この予測精度の差異は T 検定によって統計学的にも有用であることが示された．よって，pH が結晶化予測に有用な要素であることが確認できた．

## 参 考 文 献


[1] NCBI. http://www.ncbi.nlm.nih.gov/

[2] EBI. http://www.ebi.ac.uk/

[3] TargetDB. http://targetdb.pdb.org/

[4] PDB. http://www.pdb.org/pdb/home/home.do

[5] K. Chen, L. Kurgan and M. Rahbari, “Prediction of protein crystallization using collocation of amino acid pairs”, Biochemical and Biophysical Research Communications, vol. 355, pp. 764-769, 2007.

[6] L. Kurgan, A. A. Razib, S. Aghakhani, S. Dick, M. Mizianty and S. Jahandideh, “CRYSTALP2: sequence-based protein crystallization propensity prediction”, BMC Structural Biology, vol. 9, pp. 50-63, 2009.

[7] P. Smialowski, T. Schmidt, J. Cox, A. Kirschner and D. Frishman, “Will My Protein Crystallize? A Sequence-Based Predictor”, PROTEINS: Structure, Function, and Bioinformatics, vol. 62, pp. 343-355, 2006.

[8] D. M. Engelman, T. A. Steitz and A. Goldman, “Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins”, Annual Review of Biophysics and Biophysical Chemistry, vol. 15, pp. 321-353, 1986.

[9] 坂部知平，相原茂夫，“タンパク質の結晶化-回折構造生物学のために”，京都大学学術出版会，2005.

[10] K. A. Kantardjieff, M. Jamshidian and B. Rupp, “Distributions of pI vs pH provide prior information for the design of crystallization screening experiments”, Bioinformatics, vol. 20, pp.2171-2174, 2004.

[11] M. J. Zvelebil, G. J. Barton, W. R. Taylor and M. J. E. Sternberg, “Prediction of protein secondary structure and active sites using the alignment of homologous sequences”, Journal of Molecular Biology, vol. 195, pp. 957-961, 1987.

[12] M. A. Hall, “Correlation-based feature selection for machine learning”, University of Waikato, Department of Computer Science, 1999.

[13] W. Li, L. Jaroszewski and A. Godzik, “Clustering of highly homologous sequences to reduce the size of large protein databases”, Bioinformatics, vol. 17, pp.282-283, 2001.

[14] W. Li, L. Jaroszewski and A. Godzik, “Tolerating some redundancy significantly speeds up clustering of large protein databases”, Bioinformatics, vol. 18, pp.77-82, 2002.

[15] M. Carson, D. H. Johnson, H. McDonald, C. Brouillette and L. J. DeLucas, “His-tag impact on structure”, Acta Crystallographica Section D Biological Crystallography, D63, pp.295-301, 2007.

[16] ExPASy Proteomics Server. http://au.expasy.org/

[17] 松原望，縄田和満，中井検裕，“統計学入門”，東京大学出版会, 1991.