2009－MPS－73 (54)
2009／3／6

SVMを用いた生体分子への金属結合部位予測手法の提案

中澤昌美† 高田雅美† 横田恭宣* 野口保* 関嶋政和* 城和貴†

†奈良女子大学大学院人間文化研究科
＊産業技術総合研究所 生命情報工学研究センター

タンパク質は，生体内に存在する金属イオンと結合することで立体構造が変化し，それに伴い生体内での機能が起こる．金属タンパク質はさまざまな生理機能の発現に深く関与し，生命活動において特徴のある非常に重要な役割を果たしている．現在タンパク質の立体構造はX線結晶構造解析やNMRによって解析が進められているが，これらの実験手法は金属の結合情報を得るまでに相当な時間を要する．そこで本研究では，既知の金属イオンとタンパク質の結合情報をタンパク質の立体構造データベースであるPDB (Protein Data Bank)から抽出し，SVMによる機械学習を用いて，金属イオンの結合性を予測する手法と結合部位を予測する手法を提案する．

Prediction of metal binding sites from protein sequence

Masami Nakazawa†, Masami Takata†, Kiyonobu Yokota*,
Tamotsu Noguchi*, Masakazu Sekijima* and Kazuki Joe†

† Graduate School of Humanities and Sciences, Nara Women's University
*Computational Biology Research Center (CBRC),
National Institute of Advanced Industrial Science and Technology (AIST)

Metal ion binding has a profound effect on the structure and the function of proteins. Metal proteins play very important roles in life activity. One third of known proteins require metal binding to play their functions. Although conformation of proteins is investigated by using X-ray crystallographic analysis and NMR, these engineered methods require a long time to get information about the binding of metal ion. In this paper, we propose a method for extracting known binding information of metal ion and protein from the PDB (Protein Data Bank), and prediction metal ion binding sites in proteins by using SVM.

1．はじめに

生物の生体内には金属が存在する．金属は体内に多量に存在すると毒となるが，一方で，全くないと生命活動に支障をきたす．生物が生きていくためには，体内で金属量が一定に保たれることが必要不可欠である．

金属イオンと結合したタンパク質は，金属タンパク質と呼ばれる．金属タンパク質中の金属イオンは，タンパク質と相互作用することで，タンパク質の主鎖の構造を含めてフォールディングに強い影響を与えることが知られている[1]．また，既知のタンパク質の約3分の1は，各タンパク質が機能を果たすために，補助因子として金属イオンが必要であることが知られている[2][3]．このように，金属タンパク質はエネルギー代謝，物質代謝，シグナル伝達などの様々な生理機能の発現に深く関与している．

金属タンパク質のメカニズムの解析は，現在，NMR (Nuclear Magnetic Resonance)やX線結晶構造解析，その他様々な分光法などの手法により進められている．しかし，これらの解析手法には，高価な実験装置や多くの実験時間が必要である．そのため，実験設定を再現しやすく，ターンアラウンドタイムの短縮が望めるコンピュータによる解析が期待され，コンピュータ上でタンパク質複合体形成のメカニズムを解明するための研究が，活発に行われている．分子動力学シミュレーションをはじめとするコンピュータ・シミュレーションによる解析は，タンパク質の構造が決定している場合にのみ適用可能であるが，生体内では結合可能であるにも関わらず，すべての金属イオンにポテンシャル関数が揃っているわけではない．そのため，コンピュータ・シミュレーションでは解析できないタンパク質がある．この問題を解決するためには，別のアプローチのソフトウェアが必要となる．

本稿では，SVMを用いてタンパク質と金属イオンの結合可能性を予測する手法，さらに，結合部位の予測手法を提案する．

以下2章では金属結合の予測手法について述べる．3章では，提案手法による実験を行い，結果について考察する．4章で本稿のまとめを行う．

2．金属結合予測手法の提案

本稿では，タンパク質と金属イオンの結合予測をタンパク質の構成要素であるアミノ酸と金属イオンの結合情報を機械学習で学習することにより行う．

2.1節では学習と予測に用いるデータベースである

Protein Data Bank (PDB)について説明する．2.2節ではデータセットの抽出方法とその条件について述べ，2.3節では抽出した PDB ファイルから特徴ベクトルを取得し，学習データファイルを作成する手法を示す．2.4節において，予測データの作成手法を説明する．2.5節では，学習と予測に用いる LIBSVM と学習の際の設定方法，予測手法を示す．

2.1 Protein Data Bank (PDB)

金属イオンが結合するタンパク質のデータセットとして，Protein Data Bank(PDB)[4] がある．PDB は，タンパク質と核酸の三次元構造の構造座標を蓄積した，国際的な公共のデータベースである．PDB に保存されている構造データは，主に NMR 法や X 線結晶構造解析などにより実験的に構造が決定されている．

PDB ファイル中の SEQRES 行には，アミノ酸や核酸の配列，HET 行には，アミノ酸や核酸以外の残基の情報，ATOM 行には，原子のシリアルナンバーや原子名，アミノ酸の名前やその残基番号，原子の座標，HETATM 行には，アミノ酸や核酸以外の残基の原子のシリアルナンバーや原子名，原子の座標などがそれぞれ記載されている．

2.2 PDB データの抽出

本稿では，SVM の学習データとして，この PDB エントリからデータを抽出する．その際，以下の 5 つの抽出条件を設ける．

- モノマーで構造解析されたもの．
- X 線で構造解析されたもの．
- 部位特異的変異が行われていないもの．
- 天然アミノ酸のみが含まれているもの．
- 金属イオンを含み，構造解析されたもの．

モノマーで解析されたタンパク質を抽出するため，PDB の 1 エントリ中に chain が 1 つだけのものを抽出する．核酸と結合しているタンパク質は 2chain 扱いとし，データセットからは除く．

NMR 法により構造解析されたデータは複数の構造を含む可能性がある．解析手法に定量性をもたせるため，PDB ファイルの解析手法が X-ray と記述されているタンパク質を抽出し，X 線結晶構造解析により同定されたタンパク質は，野生型と構造が異なる場合があるため，PDB エントリファイルに mutation と記述があるものはデータから削除し，部位特異的変異が行われていないデータを抽出する．

タンパク質中のアミノ酸が化学変化や翻訳後修飾されて構造が解析されている場合，そのタンパク質は非野生型と判断する．また，天然アミノ酸以外の配列を含んでいるものはデータから削除し，天然のアミノ酸のみが含まれるタンパク質を用いる．

金属イオンを含み構造解析されたデータを用いるため，PDB ファイルの HET 行に金属イオンの記述があり，かつタンパク質以外の分子が入っていないデータを抽出する．ただし，溶媒中に水と金属イオン以外の分子が入っている場合は，その分子が構造に影響を与える可能性があるため，データセットから除く．

以上の条件を設定することにより，金属イオンが野生型の立体構造に与える影響のみを考慮したデータセットを抽出することができる．本稿では，以上の条件を満たす PDB ファイルを用いて学習を行う．

2.3 学習データの作成

2.2 節の条件に従って抽出した PDB ファイルから，金属タンパク質に関する必要なデータを取り出し，学習データを作成する．本稿では，「結合距離の条件を満たすデータを用いる方法」と「結合情報データを用いる方法」の 2 つの手法を用いた学習データの作成方法を提案する．

結合距離の条件を満たすデータを用いて学習データファイルを作成する．各金属に対して，2.2 節の抽出条件に従って取り出した PDB ファイルに対し以下の操作を行う．手順は次の通りである．

1. 金属イオンの座標を取得.
2. α炭素原子の座標を取得.
3. 金属イオンとα炭素原子との原子間距離を計算.
4. 原子間距離の結合条件を設定.
5. 4)の条件を満たすアミノ酸とその前後数残基のアミノ酸を抽出.
6. 5)で取り出したアミノ酸の個数をアミノ酸の種類ごとにカウント.
7. LIBSVM の学習データファイルのフォーマットに従い，特徴ベクトルを作成.
8. できた特徴ベクトルを学習データファイルに出力.
9. 1)〜7)の操作を 2.2 節で抽出した PDB ファイルに対して実行.

α炭素原子とは，アミノ酸の中心となる炭素原子のことである．

次に，PDB ファイルの結合情報データを用いる学習データファイルの作成方法について述べる．金属の種類別に，2.2 節の抽出条件に従って取り出した PDB ファイルに対して，以下の操作を行う．

1. 金属イオンのシリアルナンバーを取得.
2. 金属と結合している原子(結合原子)のシリアルナンバーを取得.
3. 結合原子のアミノ酸残基番号を取得.
4. 3)で取り出したアミノ酸残基とその前後数残基を抽出.
5. 4)で取り出したアミノ酸の個数をアミノ酸の種類ごとにカウント.
6. LIBSVM の学習データファイルのフォーマットに従い，特徴ベクトルを作成.
7. できた特徴ベクトルを学習データファイルに出力.
8. 1)〜7)の操作を 2.2 節で抽出した PDB ファイルに対して実行.

残基を取り出す際，取り出すアミノ酸の残基数が学習に影響を与えるか調べるため，抽出するアミノ酸残基数を 7〜9 に変化させる．学習データファイルのデータは，<label> と<index>:<value>で表す．<label>はクラスラベル，<index>:<value>は特徴ベクトルである．<label>は整数でクラスラベルを表し，<index>は整数で 1 から昇順に，<value>は実数で表す．

表 2.1. アミノ酸の対応表

ALA	1	GLU	7	MET	13	TYR	19
ARG	2	GLY	8	HPE	14	VAL	20
ASN	3	HIS	9	PRO	15	ASX	21
ASP	4	ILE	10	SER	16	GLX	22
CYS	5	LEU	11	THR	17		
GLN	6	LYS	12	TRP	18		

本稿では，表 2.1 のように 20 種類のアミノ酸残基を 1～20 の整数に対応させる．アミノ酸残基が決定していない ASX(アスパラギンもしくはアスパラギン酸)と GLX(グルタミンもしくはグルタミン酸) は<index> をそれぞれ 21，22 として割り当てる．学習データファイルや予測データファイルを作成する際には，アミノ酸を 1～22 の index 値で表す．抽出した全 PDB ファイルに対してこのデータを出力する．

2.4 予測データの作成

新規に発見されたタンパク質に金属イオンが結合するか判定するために，予測データファイルを作成する．予測したいタンパク質のアミノ酸の並びを PDB の SEQRES 行より調べて前から 1～7 残基，2～8 残基，・・・，(n-6)～n 残基を取り出し，2.2 節と同様にして予測データファイルを作成する．

2.5 LIBSVM

本稿では，金属イオンがあるアミノ酸に結合可能かどうか判定する際，未知のデータに対する予測の精度が高い識別器が有効であると考え，学習と予測に SVM を採用する．多数開発されている SVM のライブラリの中から，Tiwan National University の Chih-Jen Lin らによって開発された LIBSVM(A Library for Support Vector Machine: Version2.88 released on October 30,2008) を使用する[5].

LIBSVM を用いて学習を行う際に必要な設定を行う．

- Gaussian カーネル
- Grid search により決定したパラメータを使用
- C-SVC (結合性判定)
- ε -SVR (結合部位予測)

Gaussian カーネルは他のカーネルに比べ計算が行いやすい．また，Gaussian カーネルの値は，原点付近に写像されるため扱いやすい[6]．これらの理由から Gaussian カーネルを用いる．

SVM の 2 つのパラメータ C と α を決定する際に Grid search を用いる．またオーバーフィッティングを回避するため，Cross Validation を行い，その確率が最も高い C と α のペアを最適値とする．

結合性の判定をする際，SVM のタイプに C-SVC (C-Support Vector Classification)を用いる．予測結果は，結合の場合 1，非結合の場合-1 を出力する．

結合部位の予測をする際，SVM のタイプに ε -SVR (ε -Support Vector Regression)を用いる．予測結果は，クラスラベルと同時に数値が出力される．この予測値によりどちらのクラスにどの程度近いかを知ることができる．予測値がある値以上であると，結合すると仮定する．このある値を結合基準値と呼ぶことにする．

表 3.1 結合距離を用いた亜鉛結合予測精度

	4.0Å	5.0Å	6.0Å
7 残基	37.2 %	31.8 %	35.7 %
9 残基	62.1 %	46.4 %	47.1 %
11 残基	71.1 %	73.5 %	72.0 %

表 3.2 結合距離を用いた亜鉛非結合予測精度

	4.0Å	5.0Å	6.0Å
7 残基	89.0 %	100 %	99.0 %
9 残基	82.0 %	100 %	100 %
11 残基	87.0 %	100 %	100 %

表 3.3 結合情報データを用いた亜鉛結合予測精度

	7 残基	9 残基	11 残基
結合予測精度	96.7 %	100 %	100 %
非結合予測精度	86.0 %	100 %	100 %

表 3.4 結合情報データを用いた鉄(II)結合予測精度

	7 残基	9 残基	11 残基
結合予測精度	41.5 %	22.6 %	42.5 %
非結合予測精度	100 %	100 %	100 %

結合基準値は 0.5～0.95 までの値を 0.05 刻みで設定する．予測する結合部位は，基準値を超えるアミノ酸残基とその前後 3 残基と幅を持たせる．予測結果は，11～16 のように残基番号で出力する．

3．実験

機械学習に SVM を用いて，金属イオンの結合予測を行う．

3.1 節で金属の結合判定を行う実験について，3.2 節で結合部位を予測する実験について述べたあと，3.3 節でその結果についての考察を行う．

3.1 金属の結合判定を行う実験

2.3 節の「結合距離の条件を満たすデータを用いる方法」に従い，予測する金属の学習データを作成する．金属ごとに LIBSVM を用いて学習・予測を行う．取り出す残基数を 7～11 に変化させ，また，結合距離の定義を 4Å～6Å に変化させる．ここでは Leave One Out Cross Validation の結果を予測精度とする．これは，1 つのデータを除いて学習を行い，除いたデータで予測を行うことを，データ数回繰り返す手法である．金属イオンとして亜鉛を用いた結合予測の結果を表 3.1 に，非結合予測の結果を表 3.2 に示す．

「結合情報データを用いる方法」に従い，予測する金属の学習データを作成する．取り出す残基数を 7～11 に変化させて行った実験の結果を表 3.3 に示す．他の金属と比較するため，鉄(II)の結果を表 3.4 に示す．

3.2 金属の結合部位を予測する実験

3.1 節は，タンパク質に金属イオンが結合するか判定する実験を行った．次に，金属がタンパク質中のどの部位に結合するかを予測する．学習データは，これまでの実験で最も精度が良い「結合条件を用いる方法」を用いて作成した学習モデルを採用する．

亜鉛イオンが結合するタンパク質 1A1F への結合部位予測を行った結果を示す.

基準値：予測部位の予測結果
0.95 ：6~13，34~45，48~59，63~71
0.90 ：6~13，25~45，47~71，80~90
0.85 ：6~13，25~72，80~90
0.80 ：6~14，25~72，79~90
0.75 ：5~16，20~72，75~90
0.70 ：4~16，19~72，75~90

ε-SVR の基準値を 0.95 に設定すると，このタンパク質への結合予測部位は，上の 4 部位が得られる．1A1F ファイルの CONECT 行より，亜鉛イオンは，7，25，29，37，40，53，57，65，68，81，85 の 11 か所に結合する．ゆえに，予測した 4 部位はすべて正解である．基準値別に，(正解部位数)/(予測部位数)で表すと，0.95~0.70 まで，4/4，4/4，3/3，3/3，3/3，3/3 となる．

3.3 結果と考察

「結合距離の条件を満たすデータを用いる方法」により学習データを作成し，金属の結合判定を行う実験では，結合距離の定義と，取り出す残基数を変化させている．亜鉛イオンに関する結合予測精度の結果より，結合距離の定義を変化させたことによる精度の変化は見受けられない．一方，取り出す残基数の変化による精度の変化は，取り出す残基数が多いほど精度が良くなっている．このことから，亜鉛イオンが結合するアミノ酸の種類には特徴があり，その特徴は結合する残基だけに見られるのではなく，前後の 11 残基という広い範囲に及ぶことがわかる．非結合予測精度の結果より，結合距離の定義を緩く設定するほど精度が高く，厳しいほど精度が下がる．これは，結合データ数の影響を受けているためであると考えられる．結合距離の定義を 4Åに設定すると，亜鉛イオンの結合データは 23 しか得られない．本稿では，特徴ベクトル数が 22 である．一般的に LIBSVM において，学習データは特徴ベクトルの数倍必要であるとされている．つまり，この実験には，特徴ベクトルに対して学習データ数が圧倒的に不足している．そのため，十分な学習ができず，予測精度が低くなったと考えられる．取り出す残基数による精度の変化は見られない．

「結合情報データを用いる方法」により学習データを作成し，金属の結合判定を行う実験では，亜鉛イオン結合・非結合の両予測精度で高い精度が得られている．しかし，鉄(II)イオンに対する結合予測精度はかなり低くなっている．この違いはデータ数にあると考えられる．亜鉛イオンは結合データが 216 あるのに対し，鉄(II)イオンは結合データが 11 しかなく，学習不足であると考えられる．他の金属に関しても，結合データが 9 以下のものは特に精度が低い．データ数が 10～20 の金属の精度は良いものと悪いものがある．20 以上ある金属に関しては，すべて予測精度は 100%に近い値が得られた．少ないデータ数でも精度がよい金属は，そのデータに特徴がよく表れているためだと考えられる．

3.2 節では，金属イオンが結合する部位の予測を行った．1A1F への亜鉛イオン結合予測結果から，基準値が 0.95 と厳しい条件を設定した場合でも，11 個の結合候補中 7 個の予測部位を予測することが可能である．他のタンパク質(1A1R と 1A1T)に対する予測において，基準値を 0.95 に設定した場合でも，少なくとも 1 つの正解部位予測が可能である．本稿では，結合候補のすべてを網羅するより，いかに予測部位の誤りを率を低く保ったまま結合部位を正しく予測できるかに重点をおく．そのため，基準値はできるだけ高く設定するほうがよい．

4．まとめ

Vapnik によって提案された Support Vector Machine (SVM)を用い，タンパク質のアミノ酸配列と金属イオンの結合性を予測する手法，更に，金属イオンの結合部位を予測する手法の提案とその評価を示した．

「結合距離」と「結合情報」の 2 つ手法により学習データを作成し，それぞれについて学習と予測を行い，タンパク質に金属イオンが結合するか判定を行った．距離の条件を用いた場合，取り出す残基数により精度が変わった．亜鉛イオンでは，7 残基取り出すと予測精度は 3 割程度であるが，11 残基に増やすと予測精度は 7 割程度まで上がる．

結合情報を用いた場合，学習データ数の違いにより予測精度に差が出た．学習データが十分にある亜鉛イオンは，予測精度が 100%に近く，高い精度を示した．学習データが 20 以下の金属の予測精度はあまりよくなかった．この例として鉄(II)イオンが挙げられ，予測精度は約 4 割となった．

金属が結合する部位の予測手法を提案し，亜鉛イオンを例に挙げて予測した．C-SVC の結果では，結合部位予測を行うことは困難であるが，ε-SVR による回帰を行うことにより，結合確率が高いものを選択すると，結合部位の予測精度を上げることができた．

参考文献

[1] M. Babor, H.M. Greenblatt, M. Edelman, V. Soboler, Flexibility of metal binding sites in proteins on a database scale, *Proteins*, **59**, pp.221 -30 (2005)

[2] J.A. Ibers, R.H. Holm, Modeling coordination sites in metallobiomolecules, *Science*, **209**.

[3] J.A. Tainer, V.A. Roberts, E.D. Getzoff, Protein metal-binding sites. *Curr Opin Biotechnol*, **3**, pp.378-387

[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank. *Nucleic Acids Research*, **28**, pp. 235-242 (2000).

[5] C. C. Chang and C.J. Lin, LIBSVM : a library for support vector machines, (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] Keerthi, S. S. and C.-J. Lin Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation 15*(7), 1667 – 1689 (2003).