社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－MPS－62 (33)
2006－BIO－7 (33)
2006／12／22

SVMを用いた金属イオン結合部位予測システムの開発

中澤昌美+ 高田雅美* 横田恭宣† 野口保†
関嶋政和† 城和貴*

+ 奈良女子大学理学部情報科学科
* 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科学研究センター

概要

金属イオンの結合は，タンパク質の構造や機能の変化に大きな影響を与える．ゲノム配列がシーケンサーにより網羅的に読解されていくのに対して，タンパク質の立体構造は NMR や X 線結晶解析によって解かれることから，あるゲノム配列が得られてから立体構造や金属イオン結合情報が明らかになるまでには相当の時間が要される．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データベースから抽出し，SVM による機械学習から，タンパク質と金属イオンとの結合予測を行うシステム開発をしている．

Development of a Prediction System for Metal Binding Sites in Protein by Using SVM

Masami Nakazawa+ Masami Takata* Kiyonobu Yokota†
Tamotsu Noguti† Masakazu Sekijima† Kazuki Joe *

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

Abstract

Metal ion binding is important for structure and function of proteins. One third of known proteins require metal binding to play their function. And some proteins are caused structure transition by metal binding. Genome sequences are decided by high through-put method. On the other hand, conformations of protein are determined by NMR and X-ray crystallographic analysis. These determinations are took large cost. In this paper, we will describe development of a prediction system for metal binding site in protein using by SVM.

1. はじめに

生体内には，金属タンパク質と呼ばれる金属イオンと結合したタンパク質が存在している．金属タンパク質に含まれる金属イオンは，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，金属タンパク質は，エネルギー代謝，物質代謝，シグナル伝達など様々な生理機能の発現に深く関与していることも知られている．既知のタンパク質の 1/3 がその機能を果たすために金属イオンが補助因子として必要であることが知られている [2,3]．また，Zn イオンが酵素反応に直接関係しているのに対して，Mg イオンは構造変化へ関係していることが多い[4]．

NMR（Nuclear Magnetic Resonance）や X 線結晶構造解析，様々な分光学などの手法により金属タンパク質のメカニズムの解明が進められている．しかし，これらの手法は，多額な実験装置や実験時間が必要となる．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ている．

近年，コンピュータ上で，金属イオンとタンパク質の複合体形成のメカニズムを解明する手法が考えられている．しかし，生体内では結合可能であるにもかかわらず，ポテンシャル関数が全ての金属イオンに揃っている訳ではない為に，コンピュータシミュレーションでは解析できない金属タンパク質がある．また，分子動力学シミュレーションを始めとするコンピュータシミュレーションによるアプローチは，タンパク質の構造が決定している場合に適用可能であり，ゲノム配列，アミノ酸配列情報を得ることに対して求められる立体構造を得るためのコスト圧縮に貢献しない．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エアが必要となる．

本研究では，タンパク質と金属イオンの結合可能性を予測するシステムを開発する．このシステムでは，あるタンパク質の一次構造（アミノ酸配列）に対し，ある金属が結合可能かどうかの予測を，機械学習の一種である SVM (Support Vector Machine) を適用する．SVM は，線形分類不可能な場合，カーネル関数を用いることによって高次元空間に写像し，線形分離可能な状態に移行させることができる．

以下，2 章では，タンパク質と金属イオンの結合についてカルモジュリンの例を示しながら説明する．3 章では，学習に用いるデータセットを抽出する際の条件を述べる．4 章において，金属イオン結合部位予測システムの開発の概要を説明する．5 章では，実験とその結果について考察する．6 章において，まとめと今後の課題について述べる．

2. タンパク質と金属イオンの結合

生物は主にペプチド，タンパク質などの生体分子と水から構成されている．タンパク質は，20 種類のアミノ酸の結合体である．そのため，同じ数のアミノ酸で構成されるタンパク質であっても，その数は非常に多い．

タンパク質は糖，脂質，金属イオンなどさまざまな物質と結合することによって，特定の機能を示す．特に，生体内で金属イオンが結合した金属タンパク質は，構造や機能が大

図１．カルモジュリン（左）カルシウムイオンが結合していない構造(PDBID: 1DMO)，（右）カルシウムイオンが結合した構造(PDBID:3CLN)

きく変化し，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．

金属イオンが結合するタンパク質の例として，カルモジュリンがある．図１は，カルモジュリンの立体構造である．カルモジュリンは細胞内でカルシウムイオンの濃度変化に応じた調節機能を行うタンパク質である．通常，細胞内のカルシウムイオン濃度が低く，カルシウムイオンを含まない構造をとっている(PDBID: 1DMO) [5]．この構造のカルモジュリンは，不活性酵素と結合することが不可能である．しかし，細胞外からの刺激に応答し細胞内のカルシウムイオン濃度が上昇すると，カルモジュリンは構造を変化させながらカルシウムイオンと結合する(PDBID:3CLN) [6]．カルモジュリン-カルシウムイオン複合体は，不活性酵素との結合が可能となる．また，不活性酵素はカルモジュリンと複合体をつくることにより活性型となる．

金属イオンが体内に多量にある場合，人体の正常な活動を阻害することがある．しかし，金属イオンが全くない状態では，既知の生体分子の 1/3 は生理機能が正常に働かない可能性がある．ゆえに，金属イオンの種類と量を適切に保たれることが重要である．

タンパク質の構造を調べるための方法として，NMR や X 線結晶構造解析，様々な分光学などの手法がある．これらの工学的手法は，多額な実験設備や実験時間を必要とする．また，実験結果を再現するためには，非常に細かい設定が必要となる．そのため，多くのタンパク質を調べるためには，膨大な時間が必要となる．

このコストを回避するための手段として，コンピュータを使った解析が求められている．ポテンシャル関数が全ての金属イオンに無いため，また実行に膨大な時間がかかる為に，分子動力学法などのコンピュータシミュレーションで解析することは困難である．

本研究では，タンパク質の一次構造（アミノ酸配列）と金属イオンの結合情報を基にして，学習機械の一種である SVM を適用することで，任意のアミノ酸配列に対する金属イオンの結合可能性を予測するシステムを開発する．

３．データセットの抽出条件

タンパク質と金属イオンが結合したデータセットとして，PDB（Protein Data Bank）[7]がある．本研究において，SVM の学習データとして，この PDB （Version Jul 04, 2006, 37556 Structures, 89849 chain）エントリーからデータを抽出する．その抽出条件は以下のとおりである．

１）モノマーで構造解析されたもの
PDB の１エントリー中に chain が１つだけのもの．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．

２）X線で構造解析されたもの
構造解析手法において X-ray と記述されているもの．NMR で構造解析された

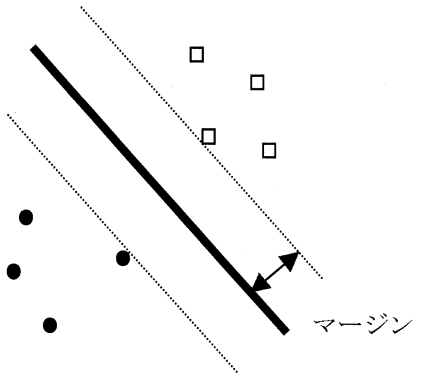

図2 データと識別面の関係

データは複数の構造を含んでいる可能性があるため，今回は定量性を持たせるためにX線で構造解析されたものだけを用いる．

3）部位特異的変異が行われていないタンパク質

PDBエントリーファイルにmutationの記述があるものは削除．部位特異的置換が入っていると構造が野生型とは異なることがあるため．

4）天然アミノ酸のみが含まれているもの天然アミノ酸以外の配列を含んでいるものは削除．タンパク質中のアミノ酸が化学修飾や翻訳後修飾されて構造解析をされている場合は非野生型と判断しデータセットからは除く．

5）金属イオンを含んで構造解析されているもの

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もの．

溶媒中に水と金属イオン以外の分子が入っている場合には，その分子が構造に影響を与える可能性があるためにデータセットからは除く．

以上のような条件から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するデータセットを用いる．本研究においては，以上の条件を満たす82個のPDBファイルを用いる．

本研究では，生命活動に重要であり，生体内に二番目に多い遷移金属であるZnと結合した金属タンパク質を抽出した．

4．SVMを用いたタンパク質と金属イオンの結合システムの開発

タンパク質はアミノ酸で構成されているため，タンパク質と金属イオンの結合の可能性を調べるのではなく，アミノ酸と金属イオンの結合の可能性を判明させればよい．アミノ酸配列はシーケンサーにより網羅的に解析することができる．一方，タンパク質の立体構造はNMRやX線解析などにより実験的に決定する必要がある．ゆえに，アミノ酸配列と金属イオンの組み合わせを1組見つけるためだけでも非常に長い時間を必要とする．そこで，機能解析までのターンアラウンドタイムを短縮するために機械学習を用いて，金属イオン結合部位を予測するシステムを開発する．

本研究では金属イオンとタンパク質の組み合わせを調べるための学習機械としてSVMを用いる．SVMは2クラスの分類を行う学習機械の一種で，現在知られている多くの手法の中で最も認識性能が優れた学習モデルである．SVMでは，トレーニングデータの中でサポートベクトルとよばれる．図2は，クラス境界付近に位置するトレーニングデータと識

表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		

別面との距離であるマージンをあらわす．SVMでは，マージンが最大となるように分離超平面を構築し，クラス分類を行う．

線形分類が困難な場合，カーネルトリックにより入力空間をより高次の特徴空間に写像し線形分類を行う．このことによって，非線形の問題に対応できる．従来パターン認識の分野で使用されていた多層パーセプトロンに比べ高い汎化性能をもち，ラグランジュ乗数法により2次の凸計画問題として定式化されるため最適解を得ることができるという特徴をもつ．

本研究では，金属イオンがあるアミノ酸に結合可能かどうか判断する際，高い汎化性能が最も必要であると考え，学習にSVMを採用する．また，多数存在するSVMのライブラリの中から今回は台湾国立大学のLINらによって開発されたLIBSVM (Version2.82, April 2006) [8]を使用する．

まず抽出したPDBファイルからZnが結合するタンパク質のアミノ酸残基を抜き出す．その際アミノ酸残基のCαとZnイオンとの距離が4.7Å以内の残基を結合すると定義し，結合する残基に加えその前後4残基も取り出す．LIBSVMのトレーニングデータは数値に変換する必要があるため，本研究では，表1のように，20種のアミノ酸残基を1～20の整数に対応させてトレーニングデータファイルを作成する．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）はそれぞれ21，22を割り当てる．

抽出したデータをLIBSVMのデータフォーマットに変換し，トレーニングファイルをスケーリングする．スケーリングの最大値と最小値の値は，デフォルトでは[-1,+1]である．この値は，オプションで指定することによって変更することができる．スケーリングを行う主なメリットは，より大きな数値域の特徴値を避け，小さな数値域の特徴値は反映されることである．また，特徴ベクトルの内積はカーネル値に依存するため，スケーリングを行うことで複雑な計算を回避できるという特長もある．ゆえに，SVMを適用する前にスケーリングを行うことは非常に重要であるといえる．

次にカーネルモデルの選択を行う．カーネルモデルの代表的なものに，以下のものがある．

linear：

$$K(x_i, x_j) = x_i^T x_j.$$

polynomial:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0.$$

radial basis function(RBF)

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel), \gamma > 0.$$

sigmoid：

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r).$$

ここで，γ, r, d はそれぞれ，カーネルパラメータである．

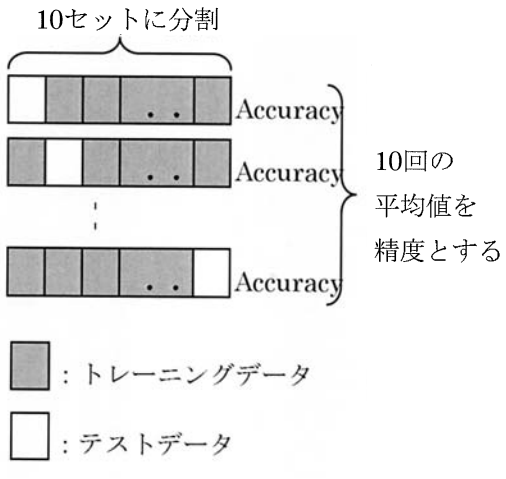

図3　n-fold cross-validation

RBF カーネルは，非線形のものでも扱うことができる．本研究は，タンパク質と金属イオンの結合に SVM をはじめて適用するため，RBF カーネルを用いる．ただし，今後，研究を進めるにつれ，大まかなデータのプロット状況が判明する場合，それにあわせたカーネルに変更する．

次に svm-train でトレーニングデータセットから予測のためのモデルを生成する．今回のトレーニングデータは Zn が結合するデータのみとし，結合しないデータは含まない．SVM タイプは one-class SVM を選択する．トレーニングを行ってできたモデルの汎化性能を調べるため，cross-validation を行う．そのために，図 3 が示すような n-fold cross-validation を使用する．この検定方法は，データを n セットに分割し，(n-1)セットでトレーニングを行い，残りの 1 セットでテストをし，その精度を調べる方法である．この作業を n 回行い，n 回の精度の平均を求めることで汎化性能を求めることができる．

表 2　残基を変化させた場合の実験結果

Data file	Accuracy
Zn1_3	47.619%
Zn1_4	50.000%
Zn1_5	45.6311%

5．実験

5.1 節では，アミノ酸の残基数を変換した場合の SVM の結果を示す．5.2 節では，金属イオンである Zn とアミノ酸のイオンの距離を変更した場合の実験結果を示す．5.3 節では，LIBSVM のパラメータ選択を行う．5.4 節において，5.1 および 5.2 節の実験に対する考察を行う．

5.1　残基数がシステムに及ぼす影響

本節では，アミノ酸から取り出す残基数を変化させた場合の影響について実験を行う．Zn とアミノ酸 Cα との距離が 4.7Å以内で結合すると定義する．実験手順は以下のとおりである．

はじめに，Zn が結合するアミノ酸残基とその前後 n 残基を取り出す．次にアミノ酸配列を表 1 のアミノ酸対応表に従って数値に変換し，トレーニングデータファイルを作成する．データファイルをスケーリングし，LIBSVM でトレーニングを行う．その際，10-fold cross-validation を行う．

表 2 は，取り出す前後のアミノ酸の個数を 3 残基，4 残基，5 残基と変化させた場合の実験結果を示す．ここで，Zn1_3, Zn1_4, Zn1_5 は，それぞれ，前後 3 残基，前後 4 残基，前後 5 残基を示す．

表 3　アミノ酸と金属イオンの距離を変更した場合の実験結果

	Zn_3(%)	Zn_4(%)	Zn_5(%)
4.0Å	47.8261	30.4348	52.1739
4.5Å	44.7368	44.7368	46.0526
5.0Å	49.6599	48.6301	48.9655
6.0Å	49.1803	48.6842	49.1749

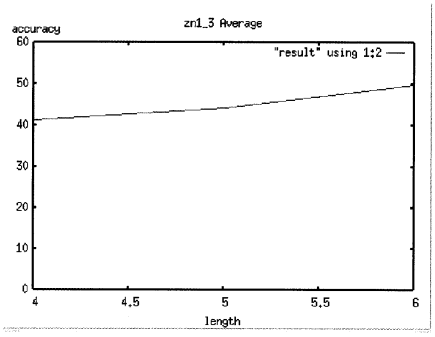

図4　距離の影響

5.2 結合する原子距離が及ぼす影響

5.1 の実験では結合の定義を Zn イオンと Cα 原子の距離が 4.7Å以内とした．本節では，この距離を変化させた場合の実験を行う．

結合の定義を，「Zn イオンと Cα 原子の距離が 4.0Å以内」と「Zn イオンと Cα 原子の距離が 5.0Å以内」の 2 種類として前節同様の実験を行う．表 3 は，この実験結果である．また，このデータの内，Zn_3 に関して，10 回の cross-validation の平均をとり，その結果を図 4 のグラフに表す．

5.3 パラメータγの設定

LIBSVM では，パラメータが大きな影響を与える．LIBSVM には最適なパラメータを求める grid.py が付属している．これを用いて最適なパラメータγを求め，オプションでγの値を指定してトレーニングを行う．表 4 は，この結果である．

表 4　パラメータγが及ぼす影響

	Zn_3(%)	Zn_4(%)	Zn_5(%)
4.0Å	39.1304	39.1304	43.4783
4.5Å	44.7368	43.4211	46.0526
5.0Å	49.6599	50.0000	48.9655
6.0Å	49.5082	49.3421	49.505

表 5　結合距離とデータ数の関係

距離(Å)	4.0	4.5	5.0	6.0
データ数	23	76	147	305

5.4 結果と考察

図 4 より，結合する距離の定義が厳しければ厳しいほど，精度が悪いという傾向がみられる．これはトレーニングデータの数が少ないためであると考えられる．

表 5 は，結合する距離とデータ数の関係を示す．この表と図 4 から，データ数が多いほど精度が高くなる傾向があるといえる．しかし，5.0Å付近で精度はほぼ最大になり，それ以上距離を広げてもほとんど変化がない．

実験の結果，全体的に精度は約 50%であった．本研究では Zn イオンとタンパク質の Cα 原子との距離で結合の判定をしているが，側鎖の配向と距離を考慮することで精度が向上すると考えられる．また，金属イオンは酸素，窒素，水素などの親水性の原子群に囲まれ，さらにその周りを炭素などの疎水性の原子群が取り囲んでいるという特徴があり[10]，このような知見を組み込むことでも精度を向上することが可能であると考えられる．

また今回の結果からはパラメータγの値を変化させることによる精度の向上は認められなかった．しかし，実験を何度も繰り返す場

合，γの最適値を使用した方が，実験結果の分散が小さくなる可能性があるため，パラメータγは適切に設定すべきであると考える．

本実験では，カーネルはデフォルトのRBFを用いたが，精度を上げるために，最適なカーネルを選ぶ必要があるものと考える．

6. まとめ

本研究において，タンパク質の一次構造（アミノ酸配列）と金属イオンの結合性を予測するためのシステムを開発した．今後は，より高精度なシステムにするために，トレーニングデータの数の増加，物理化学的特徴の考慮，SVM のカーネルを変更などといった工夫をしていく必要がある．

また，本研究では生体の活動に重要であり，生体内に２番目に多い遷移金属である Zn イオンに焦点を絞って結合可能性の予測を行ったが，今後は Fe, Mn, Ni, Cu, Co などのソフトな金属イオンへの適用を進めて行く．

参考文献

[1] M. Babor, H.M. Greenblatt, M. Edelman, V. Sobolev, 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**. pp.223-235 (1980).

[3] J.A. Tainer, V.A. Roberts, E.D. Getzoff, Protein metal-binding sites. *Curr Opin Biotechnol*, **3**, pp.378-387 (1992).

[4] A.K. Katz, J.P. Glusker, S.A. Beebe, C.W. Bock, Calcium ion coordination: a comparison with that of beryllium, magnesium, and zinc. *J Am Chem Soc*, **118**, pp.5752-5763 (1996).

[5] K.L. Yap, J.B. Ames, M.B. Swindells, M. Ikura. Diversity of conformational states and changes within the EF-hand protein superfamily. *Proteins*, **37**, pp.499-507 (1999).

[6] G. Barbato, M. Ikura, L.E. Key, R.W. Pastor, A. Bax, Backbone dynamics of calmodulin studied by 15N relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible, *Biochemistry*, **31**, pp.5269-78 (1992).

[7] 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).

[8] 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

[9] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press (2000).

[10] M.M. Yamashita, L. Wesson, G. Eisenman D. Eisenberg. Where metal-ions bind in proteins. *Proc Natl Acad Sci USA*, **87**, pp.5648-5652 (1990).