NMR測定データのマイニングによるタンパク質不安定性予測

Protein instability prediction from NMR specta data

○荒井 ひろみ[1]，木川 隆則[1,2]，　山村 雅幸[1]

○ Hiromi Arai[1], Takanori Kigawa[2], Masayuki Yamamura[1]

1 東京工業大学 大学院総合理工学研究科 2 理化学研究所ゲノム科学総合研究センター

1 Interdisciplinary Graduate School of Science and Eng.,Tokyo Institute of Technology

2 RIKEN Genomic Science Center

Abstract

In this paper, we explored possibility of protein instability prediction and estimated correlation of protein disorder and instability. We set the problem as binary discrimination problem and tried to predict one protein is stable or not using Support Vector Machine(SVM). We also used program DISOPRED2 and compared the result of disorder prediction and our data set. We extracted protein stability information from Nuclear Magnetic Resonance(NMR) spectra data and made non-homologous (under 25% identity) sequence data set with stability index "fold" or "unfold". Another data set we used is sequence list of disorder part and order part of proteins extracted from Protein Data Bank(PDB). As a results, we achived 72% accuracy with stability prediction and over 90 % accuracy with disorder prediction using linier kernel SVM. In addition, disorder prediction infers a certain level correlation between protein partial disorder and unstability.

1　はじめに

数多くの生物種のゲノムが解読されつつある現在, ポストゲノム研究としてタンパク質の網羅的研究が盛んに行われている. タンパク質とはアミノ酸 20 種が様々な組み合わせで連なった 1 次元の鎖であり, 生理的環境下で折れたたまって (フォールドして) 配列固有の安定な立体構造を取ることにより機能を発揮する. タンパク質は生物の体を構成している主な成分であり, また生体というシステムにおいて重要な機能を果たすため, その機能を知ることは生物学的にも, また製薬等の応用面においても極めて重要である. タンパク質の機能はその分子の立体配置によって決まるために, 機能解明のためにはタンパク質の構造解析が重要な役割を果たす. 現在, 実験的なタンパク質の立体構造決定は一筋縄ではいかず, 大量解析のボトルネックの 1 つに測定用の試料の作成がある. タンパク質がフォールドする条件は極めて敏感であり, タンパク質の配列の長さの僅かな変更や環境の微妙な変化により影響を受けフォールドしなくなってしまう (アンフォールド) 場合がある. また元々フォールドしないタンパク質も存在している. しかし現在タンパク質の安定な立体構造を取る条件を事前に知ることは出来ず, 実験においてトライアンドエラーを繰り返し成功例を探し出す以外に方法はない. 本研究においては構造を取らないタンパク質をそのアミノ酸配列の情報を用いて予測し, タンパク質の試料作成において候補アミノ酸配列を事前にスクリーニングすることを目標としている.

2　問題設定

本研究の目的は安定構造をとらないタンパク質をそのアミノ酸配列の情報を用いて予測することである. 問題として, 実験的に安定性の評価されたタンパク質の配列データから安定なタンパク質と不安定なタンパク質それぞれのデータセットを作成し, この 2 つを判別することを試みた.

データは理化学研究所ゲノム科学総合研究センター (理研 GSC) のタンパク質 NMR 測定のハイスループット実験において得られた網羅的なタンパク質アミノ酸配列及びそのフォールド判定のデータを使用した. ここでフォールド判定とはタンパク質フォールドの時間的安定性の度合いを NMR スペクトルから評価したもので, 本研究ではフォールド判定の 7 段階の評価のうち最も安定なタンパク質をフォールド, 最も不安定なタンパク質をアンフォールドと分類して用いる. これを GSC データと呼ぶことにする. 両

極にあるデータを 2 値判別することで微妙な特徴を持つと考えられるフォールドに影響する因子を発見できると期待できる. 具体的には, フォールドのタンパク質とアンフォールドのタンパク質の Support Vector Machine(SVM) による判別を試みた.

一方, フォールドするタンパク質は安定にフォールドしている order 領域と不安定に揺らいでいる disorder 領域を持つ場合がある. 本研究でのもう 1 つの試みはこの disorder 領域とアンフォールドの関係を調べることである. そのためにタンパク質立体構造データベース Protein Data Bank(PDB) のタンパク質立体構造の座標データを用いた. 実験的にタンパク質の構造を決定する際に, 不安定に揺らいでいるタンパク質領域は測定にかからないために座標が欠損している. このことを利用して X 線タンパク質結晶構造解析,NMR 測定それぞれにおいて得られたタンパク質立体構造データを基にしてタンパク質の order 領域と disorder 領域のデータセットを作成し, 同様に SVM による予測を試みた. このデータセットを PDB データとする. また,disorder 領域を予測する先行研究 DISOPRED2 を利用し GSC データのフォールド判定と disorder 領域の関係を評価する.

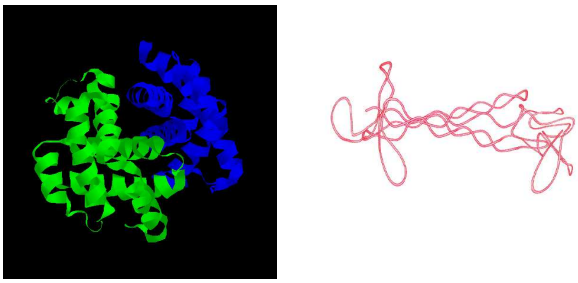

Fig. 2.1　フォールドタンパク質とアンフォールドタンパク質

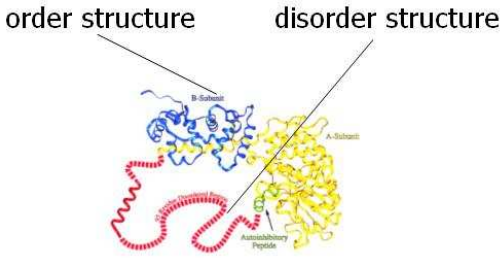

Fig. 2.2　タンパク質 disorder 領域

3　実験

3.1　データセットの作成

ある 2 つのタンパク質はそのアミノ酸配列をアラインメントした際に 25%以上一致すると構造が一致する可能性がある. よって本研究ではまずデータセットを配列相同性が 25%以下になるようにクレンジングした.

GSC データについてはアミノ酸配列のアラインメントに BLAST[5] を用い, 配列一致度が 25%以上のものでクラスタリングし, フォールド判定が最も良かったものを代表データとした. アミノ酸配列数 12496 のうち代表データは 1670 件, そのうちフォールドデータは 94 件, アンフォールドデータは 1118 件となった.

GSC データの配列長は図 3.1 に示すように分布している. 最短配列でも 40 残基以上あったため, 今回は配列のアミノ酸組成比率を特徴量として用いた.

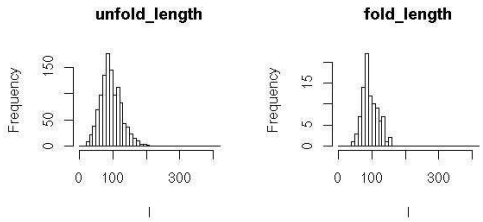

Fig. 3.1　GSC データ配列長の分布

PDB データについては産業技術総合研究所の PAPIA サービス [2] を用い, 膜タンパク質を除いて配列相同性が 25%以下のクラスタリングを行い代表データを作成した. 結果,X 線結晶構造解析によるデータが 2512 件 (うち order 配列が 3411 件,disorder 配列が 3232 件),NMR によるデータが 532 件 (うち order 配列が 538 件,disorder 配列が 124 件) となった. ここで, タンパク質の立体構造解析によく用いられる手法として X 線結晶構造解析及び NMR 測定の 2 つがあるが, この 2 つは X 線の実験がタンパク質の結晶を,NMR 測定がタンパク質の高濃度溶液をそれぞれ試料として用いることから, 研究対象となるタンパク質の特性や測定にかからない disorder 領域の揺らぎの程度がことなると考察されるためにデータを分けて実験を行った.

PDB データの disorder 領域は図 3.2,3.3 に示すように分布している. グラフは全配列,N 末 15 残基を含むデータ,

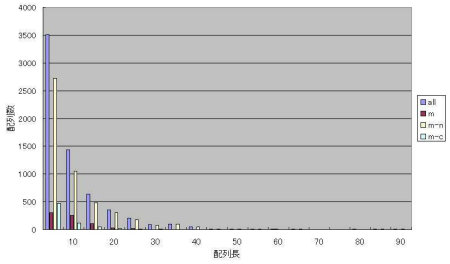

Fig. 3.2　PDB X 線データ disorder 領域配列長の分布

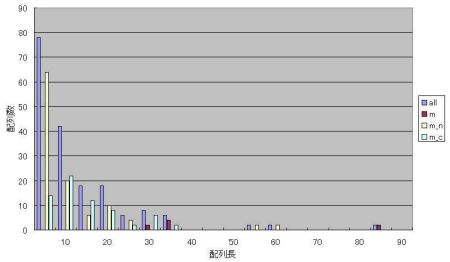

Fig. 3.3　PDB NMR データ disorder 領域配列長の分布

両端 15 残基を除いたデータ,C 末 15 残基を除いたデータそれぞれについて取った. 結果,N,C 末端には短い disorder 領域が多く含まれていることが分かる. これは経験的に分かっており本研究の関心の対象ではない. また,NMR データから,NMR による測定において両端を除いた領域に 20 残基以下の短い disorder 領域は観測されなかったことが分かる. よって本研究では配列長が 20 以上のものを扱う.

3.2　特徴量の分布

SVM を用いた学習の際に, 本研究では配列中のアミノ酸組成比を特徴量として用いる. 学習の前に,GSC データ,PDB X 線データ,PDB NMR データそれぞれに対してアミノ酸組成比の統計を取った. 結果を図 3.4-3.6 に示す. ここでエラーバーは四分位偏差を示す.

結果, アンフォールドタンパク質には疎水性のアミノ酸が少ない傾向が,disorder 領域には疎水性のアミノ酸が少なく親水性のアミノ酸が多い傾向にあることが分かる. しかし, 組成比の分布は重なっており, アミノ酸組成のみの情報からフォールドや disorder 領域の判定は難しいことが分かる.

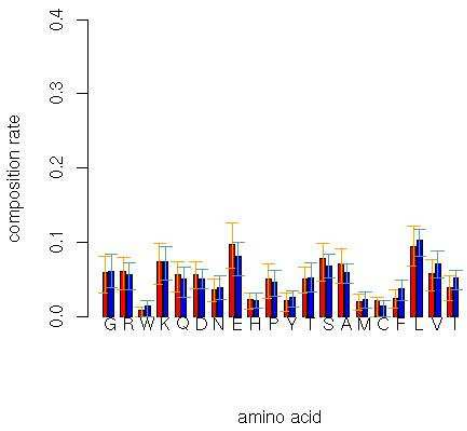

Fig. 3.4　GSC データ残基割合の統計 (赤:アンフォールド, 青:フォールド)

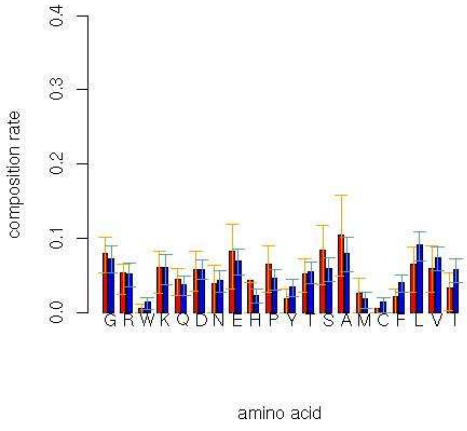

Fig. 3.5　PDB X 線データ アミノ酸組成比の統計 (赤:disorder, 青:order)

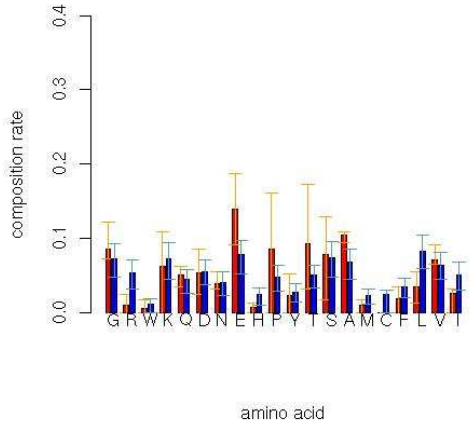

Fig. 3.6　PDB NMR データ　アミノ酸組成比の統計 (赤:disorder, 青:order)

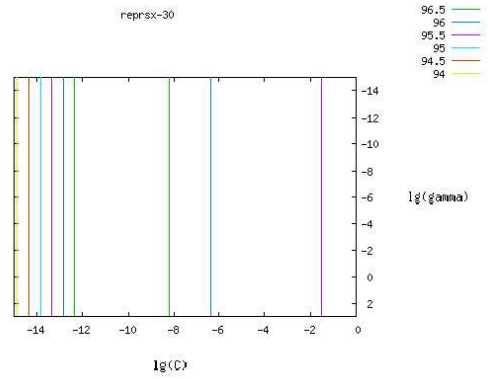

Fig. 3.7　線形カーネルを用いた PDB X 線データの予測精度

3.3　Support Vector Machine による判別学習

SVM のプログラムは LIBSVM[4] を用いた. 本研究では線形カーネルを用いた. コストパラメータ C は主問題における以下の量である.

$$\text{minimize}_{\xi,\mathbf{w},b} \quad k(\mathbf{w}\cdot\mathbf{w}) + C\sum_{i=1}^{l}\xi_i^2 \tag{3.1}$$

GSC データの学習において, コストパラメータ $\log_2 C$ を-5 から 15 まで 2 刻みに動かしたところ, パラメータに関係なく予測精度は 71.2%であった.

PDB X 線データ,NMR データそれぞれについて同様にコストパラメータ ((3.1) 式参照)$\log_2 C$ を-5 から 15 まで 2 刻みに動かした結果を図 3.7,3.8 に示す. 最高予測率はそれぞれ 96.7%,97.7%であった.

ここで, 学習に用いたデータ数を表 3.1 に示す. なお,PDB データについては正例, 負例はそれぞれ disorder 領域,order 領域とし,GSC データについてはアンフォールドタンパク質, フォールドタンパク質とした.SVM を用いた学習の際には正例と負例が同数になるように荷重パラメータを設定した.

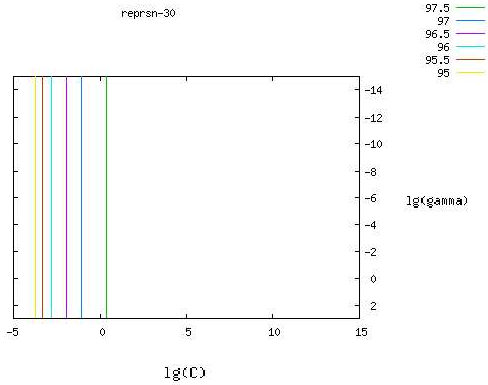

Fig. 3.8　線形カーネルを用いた PDB NMR データの予測精度

データ	正例数	負例数
PDB X 線データ	418	6326
PDB NMR データ	26	598
GSC データ	1118	94

Table 3.1　学習に用いたデータ数

3.4　学習結果を用いたフォールド判定

3.2 における実験において予測精度が最も良かったパラメータを用いた PDB データの学習結果を用いて GSC

PDB X 線データ	
予測対象	予測率 (正例/負例)
PDB X 線データ	97.3% (60.3 % /99.7 %)
PDB NMR データ	94.5 %(38.5 %/ 96.3 %)
GSC データ	23.6 %(17.2% / 100 %)
PDB NMR データ	
予測対象	予測率 (正例/負例)
PDB X 線データ	92.4% (30.1 % /96.6 %)
PDB NMR データ	97.7 %(100 %/ 97.7 %)
GSC データ	14.9 %(7.9 %/97.9 %)
GSC データ	
予測対象	予測率 (正例/負例)
PDB X 線データ	81.9% (76.5 % /82.3 %)
PDB NMR データ	59.0 %(84.6 %/ 58.4 %)
GSC データ	71.7 %(70.8 %/ 83.0%)

Table 3.2　SVM 学習結果による予測

データの予測を行った. 結果を表 3.2 に示す.

3.5　disorder 領域予測を用いたフォールド判定の予測

タンパク質のアミノ酸配列が与えられたときに disorder 領域を予測する先行研究が幾つかなされている. 本研究では, このうち DISOPRED2 を用いて GSC データの disorder 領域を予測し, フォールド判定との比較を行った.DISOPRED2 では本研究の PDB X 線データ同様,PDB X 線データの disorder 領域を元に学習を行い, 類似パターンを検出するものである.

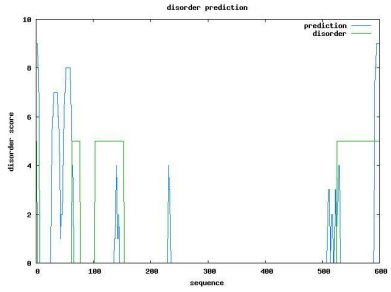

Fig. 3.9　DISOPRED2 予測精度

DISOPRED2 における disorder 予測は図 3.9 のようにスコアで評価される. これを 2 値の評価にするためにスレッショルドを動かして ROC 曲線を作成した.DISOPRED2 によるデータ PDB X 線データおよび NMR データの予測精度は図 3.10 のようになる.

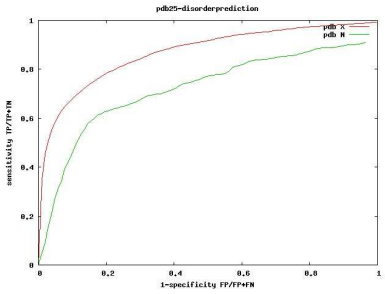

Fig. 3.10　DISOPRED2 予測精度 (赤:X 線, 緑:NMR)

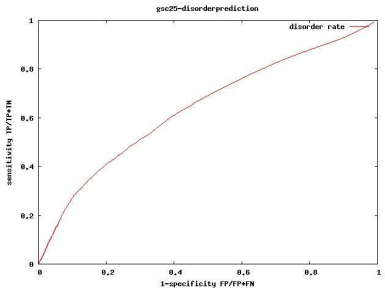

Fig. 3.11　DISOPRED2 を用いた予測:総残基割合

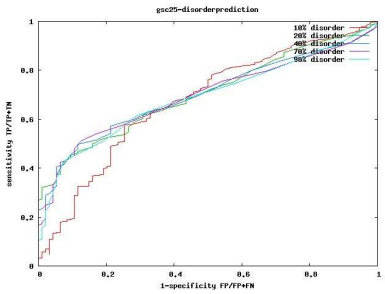

Fig. 3.12　DISOPRED2 を用いた予測:配列毎の残基割合による分類

さらに,DISOPRED2 を GSC データに対して適用した. ここでフォールドとアンフォールドと disorder との関係は,disorder が残基ごとに予測をするのに対し, フォールドの指標は配列ごとについているために基準を設けて判断する必要がある. 図 3.11 はフォールドの配列全体を disorder, アンフォールドの配列全体を order として判断したもの, 図 3.12 は配列中の disorder と予測された残基の割合があ

る割合 (10%,20%,40%,70%,90%) 以上であればアンフォールドと判断し, 配列ごとの正答率をプロットしたものである. 結果, 配列中の disorder 予測領域が 20%以上の配列はアンフォールドする傾向があることが分かった.

4 おわりに

SVM を用いた実験の結果, 表 3.2 によりフォールドタンパク質とアンフォールドタンパク質のアミノ酸組成比による識別がある程度可能であることが分かる. また,3.2 及び図 3.10 により,PDB の X 線データ,NMR データの予測精度が高いことからこれらの disorder 領域はそのアミノ酸組成によってよく特徴付けられることが分かる. さらにこれらの相互の予測が良い精度をあげていることから X 線データ,NMR データの disorder 領域は類似の特徴を持つと推測される.

一方,NMR 測定によるタンパク質全体のフォールド判定と disorder 領域の関係は, 図**??**よりある程度以上 (実験では 20%) の disorder 領域を持つことがフォールド判定に影響することが考察される. よって, 表 3.2 の GSC データの学習結果は disorder 領域が order 領域と分けられていないことが精度の低さと関係すると考察される. また,disorder 領域以外にもフォールド判定に影響する因子があることが推測される.

以上により, タンパク質の disorder 領域は NMR 測定によるフォールド判定に影響することがわかった. 今後の課題としてこの影響の評価をすすめ, フォールド判定から disorder 領域以外のタンパク質アンフォールド条件を考察していくことで, スクリーニングの精度を上げることが考えられる.

参考文献

[1] Burkhard Rost(1999). Twilight zone of protein sequence alignments Protein Engineering. **12**,85-94

[2] Noguchi,T. and Akiyama,Y.(2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003.Nucleic Acids Research,**31**,492-493.

[3] S.O.Garbuzynskiy,M.Y.Lobanov and O.V.Galzitskaya(2004). To be folded or to be unfolded? Protein Science. **13**,2871-2877

[4] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using the second order information for training SVM (2005) Journal of Machine Learning Research 6, 1889-1918 (http://www.csie.ntu.edu.tw/ cjlin/libsvm/)

[5] E.Michael Gertz(2005) BLAST Scoring Parameters (NCBI documentation http://ncbi.nih.gov/BLAST/developer.shtml)

[6] J.J.Ward, J.S.Sodhi, L.J.McGuffin, B.F.Buxton and D.T.Jones(2004). Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. Journal of Molecular Biology. **337**,635-645

[7] Pedro Romero, Zoran Obradovic, Niaohong Li, E.C.Garner, C.J.Brown , A.K.Dunker(2001). Proteins:Structure,Function,and Genetics **42**, 38-48

[8] Rune Linding, R.B.Russell, V. Neduva and T.J.Gibson(2003). GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Research**31**,3701-3708

[9] J.Kyte and R.F.Doolittle(1982) A Simple Method for Displaying the Hydropathic Character of a Protein Journal of Molecular Biology **157** 105-132.