アミノ酸配列によるタンパク質disorder領域の予測

東京大学 大学院農学生命科学研究科　石田 貴士[1], 中村 周吾, 清水 謙多郎

概 要

近年、一部のタンパク質でdisorder領域と呼ばれる一定の立体構造を持たない領域がその機能と深く関係している例が発見され、現在ではこのdisorder領域についての研究が盛んになっている。タンパク質のdisorder領域は特徴的なアミノ酸配列を持つことが知られており、この特徴を利用することで、disorder領域はアミノ酸配列から予測可能であると考えられる。本論文では局所配列情報からの予測と配列アライメントを利用した大域的な配列情報からの予測を組み合わせて行うことで予測精度の向上を図った新たなdisorder領域予測手法について報告する。

1 はじめに

1.1 タンパク質のdisorder領域

フィッシャーの『鍵と鍵穴説』に代表されるようにタンパク質の立体構造がその機能を決定するという『構造—機能』パラダイムは構造生物学における中心的な考え方である。そのため、タンパク質の立体構造の決定はこの分野の中心的なテーマの一つとなっており、現在構造ゲノミクスの名の下に進行している大規模なプロジェクトにより、数多くの新規フォールドを持つタンパク質の立体構造が解明されつつある。しかし、多くのタンパク質の立体構造が明らかになる一方で、構造の決定が不可能なタンパク質、つまり機能を持っているにも関わらず天然状態で一定のフォールドを形成しなかったり、数十残基にわたる一定の構造をとらなかったりする領域を持つようなタンパク質が数多く存在することが明らかになってきた。幾つかのタンパク質がそのアミノ酸配列の特徴から天然状態でフォールドしないことは既に20年以上前に示唆されていた。しかし、近年の研究により、これらのタンパク質のフォールドしない領域の中にそのタンパク質の機能にとって重要な領域が含まれていることが明らかになり、現在ではそのようなタンパク質及び領域は“intrinsically unstructured”、もしくは“intrinsically disordered”という呼称が与えられ、それらに関する研究が盛んに行われている。本論文ではこうした領域を単にdisorder領域と呼ぶことにする。このタンパク質disoder領域の明確で定量的な定義は存在しないが、disorder領域はNMR分光法やX線結晶構造解析、プロテアーゼ消化など幾つかの実験的手法によって決定することが可能である。このようなdisorder領域は古細菌や原核生物のタンパク質では全体の10%以上、真核生物のタンパク質ではその30%以上において存在するとも見積もられており[1, 2]、大きな注目を集めている。タンパク質disorder領域と機能の関係についてはDNA結合やその他の分

[1]E-mail: tak@bi.a.u-tokyo.ac.jp

子認識に関与していることが実験的に示されている。Disorder 領域上の結合サイトは通常の安定な構造を持つ結合サイトとは異なり、その構造の柔軟さから複数のターゲットとの結合が可能であり、また高い特異性と同時に低いアフィニティで分子を認識することが可能であると考えられている [2]。

1.2 タンパク質 disorder 領域の予測

タンパク質の立体構造はそのアミノ酸配列によって決定されるが、同様に一定の立体構造をとらないという disorder の状態もそのアミノ酸配列から決定されているのだろうか？既に disorder 領域のアミノ酸構成には単純な配列の繰り返しや、A, R, G, Q, S, P, E, K といった特定のアミノ酸が order 領域に比べ多く見られるといった特徴的なパターンが存在していることが過去の研究から明らかとなっており [2, 3]、2 次構造予測等と同様にニューラルネットワークなどの機械学習を用いた予測手法が開発されている。しかしながら、それらの予測精度は 70〜80%程度にとどまっており [4]、更なる改良が望まれている。そこで、本論文では従来行われてきた局所配列情報からの予測と共に配列類似のタンパク質とのアミノ酸配列のアライメントを利用した大域的な配列情報からの予測を併用することで予測精度の向上を図った新たな disorder 領域予測手法を提案する。

2 ホモログにおいて disorder 領域は保存される

ホモログ、すなわち配列の似たタンパク質は似たような立体構造を持つことはよく知られた事実である。また、僅かな配列相同性しか存在しないタンパク質間であっても立体構造がよく保存されている例は数多く見受けられ、そのことは比較モデリングによる立体構造予測などに利用されている。ある残基が disorder しているのか否かということは立体構造の情報の一部であると考えることができる。それではアミノ酸配列が大きく変異した配列相同性の低いタンパク質同士でも disorder 領域とその位置は立体構造と同様に保存されているのであろうか？

図 1 は PISCES サーバ [5] によって作成された、相互の配列相同性が 40%以下の高解像度の X 線結晶構造のみを含むタンパク質立体構造セットの中で 20 残基以上の disorder 領域を含む 492 本のチェインについて、そのホモログとの間の全体配列相同性と disorder 領域の保存の割合を示したものである。Disorder 領域が保存されているかどうかは、disorder 領域を含むタンパク質全体の配列をクエリとして PDB に対する BLAST を実行し、ヒットしたホモログでクエリの disorder 領域にアラインされた領域に disorder の残基が含まれているかどうかを調べ、もし含まれていれば disorder 領域は保存されているとした。プロットから明らかなように、高い配列相同性を示したホモログにおいては disorder 領域はかなり高い割合で保存されている。

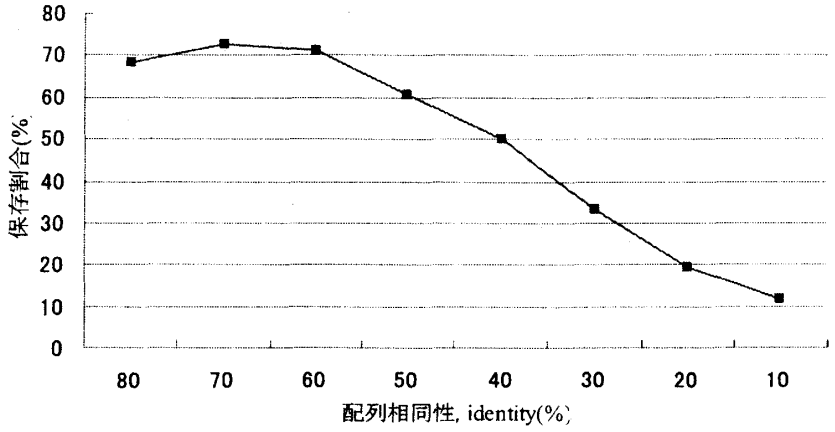

図 1: 配列相同性と disorder 領域の保存割合

3 予測手法

3.1 大域的な配列アライメントからの予測

上記の知見から開発されたのが、配列アライメントを利用した予測手法である。予測の対象となるタンパク質のアミノ酸配列をクエリとして PDB に対して BLAST による検索を行い、そこで得られたアライメントから各残基について予測値 P を求める。

$$P = \sum_{i=1}^{N} \alpha_i \mathrm{I}_i$$

N は BLAST で E-value が 0.01 以下であったヒットの件数であり、I_i は i 番目のヒットの配列相同性 (identity) である。α_i は i 番目のヒットでアラインされた残基の状態が disorder であれば 1、そうでなければ 0 となる。この予測値 P が閾値以上の場合に、その残基を disorder と判定する。

3.2 局所配列情報からの予測

アライメントによる予測とは別に局所配列情報からも予測を行う。まず、アミノ酸の配列情報を PSI-BLAST[6] により位置特異スコア行列 (PSSM, Position Specific Score Matrix) へと変換する。その後、予測対象となる残基を中心としたそれぞれ 9, 15, 33 残基幅の window における位置特異スコア行列を入力として学習機械の一種である Support Vector Machine(SVM) により予測を行う。SVM は 2 クラスの分類を行う機械学習のアルゴリズムで、与えられた訓練点のなかでサポートベクトルと呼ばれるクラス境界近傍に位置する訓練点と識別面との距離であるマージンを最大化するように分離超平面を構築しクラス分類を行う線形識別機である。また、線形では分類が難しい場合には、カーネルトリックによって入力空間をより高次の特徴空間に写像し、そこで線形分離を行うため、非線形の問題に対しても適用が可能である。SVM の学習には LibSVM[7] を用い、訓練セットとして PISCES サーバによる配列相同性 25%以下の高解像度の X 線結晶構造の

みを含む493チェインのデータセットを利用した。SVMは予測の結果として識別面との距離を表すdecision valueを与える。decision valueはdisorderと強く予測されれば大きな正の値に、逆にorderであると強く予測されれば大きな負の値となる。最後に、上記の3つのSVMから得られたdecision valueを重み付きで平均し、その値が正となればその残基はdisorderであると判定する。

3.3 予測の組み合わせ

最終的な予測は上記の2つの予測結果を組み合わせることで得る。2つの予測結果を組み合わせるため、配列アライメントによる予測から得られた予測値Pと局所配列情報からの予測の際に得られたdecision valueを重み付きで足し合わせる。これより得られた値が閾値より大きければ最終的な予測をdisorderとし、閾値より小さければorderとする。

4 性能評価

Disorderの残基とorderの残基を1対1で含み、SVMの訓練セットと25%以上の配列相同性を持たない高解像度のX線結晶構造124チェインを含むテストセットを利用して性能評価を行った。その結果を表1に示す。組み合わせ予測は局所配列情報からの予測と配列アライメントからの予測を組み合わせたもので、再現率の点では局所配列情報からの予測に若干劣り、より多くの疑陽性を生じるが、非常に高い適合率を示しており、その結果としてQ2のスコアが最も高いものとなっている。

また、上記のSVMの学習に利用した訓練セットを用いて、5-foldのクロスバリデーションを行い、予測手法の性能評価を行った。図2は予測のROC(Receiver Operating Characteristic)カーブである。訓練セットのdisorder残基の存在比率はPDB中のdisorder残基の存在比率と同様全体の4%程度であるため予測精度の向上に対してグラフの変化が多少わかりにくいが、局所配列のみを用いた予測に比べ、配列アライメントによる予測を組み合わせた場合の方が全般的に良い結果を示しており、特に疑陽性の存在比率の低い領域ではより高い適合率を示している。

	Q2(%)	適合率 (recall)(%)	再現率 (precision)(%)
配列アライメントからの予測	68.21	53.21	73.21
局所配列情報からの予測	77.86	59.70	96.02
組み合わせ予測	81.14	67.39	94.69

表1: テストセットに対する性能評価

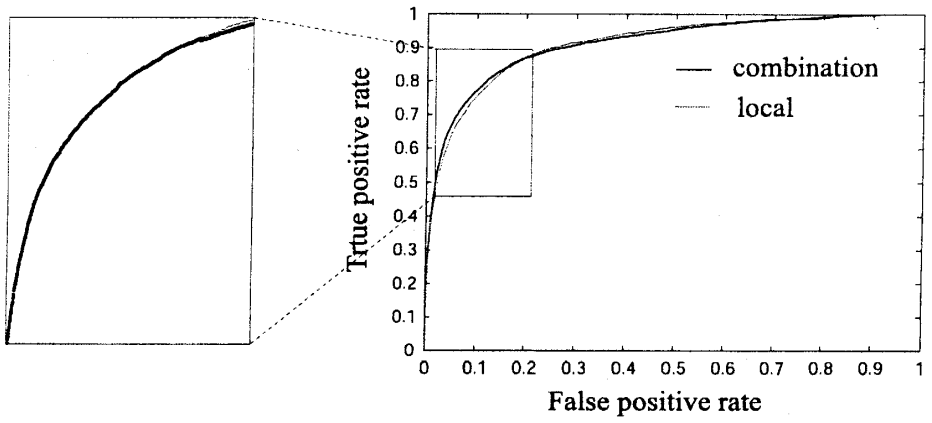

図 2: クロスバリデーションによる ROC カーブ。太線は組み合わせ予測による ROC カーブを、細線は局所配列情報のみからの予測の ROC カーブを表す。

5 まとめ

局所配列の情報のみを入力として利用する予測手法に対して、大域的な情報を予測に加えることで予測精度の向上が得られた。これは配列相同なタンパク質の大域的な情報が明示的に加わったことで、フォールドや残基周辺の環境の情報が取り込めたことの結果であると考えられる。この手法の問題点は、データベース中に配列相同なタンパク質の立体構造情報が存在する場合にのみ有効な点である。また、逆に新規のフォールドのタンパク質に関しては疑陽性を増加させる可能性があり、今後は新規フォールドや既知構造との配列相同性の低いタンパク質に対しても有効となるよう、立体構造予測の結果を取り込むといった改良が必要であると考えられる。また、単なる disorder 領域の予測にとどまらず、同時にその disorder 領域に関連する機能を予測するというのも重要な課題である。

参考文献

[1] J. J. Ward, et al., Journal of Molecular Biology, **337** (2004), 635.

[2] K. Dunker, et al., Journal of Molecular Graphics and Modeling, **19** (2001), 26.

[3] V. Uversky, Protein Science, **11** (2002), 739.

[4] E. Melamud, J. Moult, Proteins, **53** (2003), 561.

[5] G. Wang, R. Dunbrack, Bioinformatics, **19** (2003), 1589.

[6] SF. Altschul, et al., Nucleic Acids Res., **25** (1997), 3389.

[7] E. Agirre, et al., Computers and the Humanities, **34** (2000), 103.