早稲田大学大学院理工学研究科

博 士 論 文 概 要

論 文 題 目

生物配列から重要部位を予測する文字列処理に関する研究

String processing for predicting important regions from biological sequences

申 請 者

氏 名	清水	佳奈
	Kana	Shimizu

専攻・研究指導
(課程内のみ)　情報ネットワーク専攻情報構造研究

2006 年 7 月

医療技術や創薬技術を向上させるためには、生体内で重要な役割を果たすタンパク質の機能を明らかにする必要性がある。そのため、DNA からタンパク質の設計図となっている部位を発見する研究や、タンパク質の機能を司る立体構造を明らかにする研究は、タンパク質の機能解析に役立つ重要な研究である。その中でもタンパク質のコーディング領域をより高い確率で保持している cDNA を扱った研究や、機能に関係している可能性の高いディスオーダーと呼ばれる構造を扱った研究が注目を集めている。

本論文では cDNA のコーディング領域を予測する手法、及び、タンパク質のディスオーダーを予測する手法を提案し、提案手法がタンパク質の機能解析に有効であることを明らかにした。

cDNA のコーディング領域予測（2, 3 章）

本論文の 3 章で提案した手法の新規性は、タンパク質の立体構造情報が利用可能な点である。cDNA からタンパク質のコーディング領域を同定するためには、コーディング領域内に頻出する配列のパターンなどを用いる必要がある。従来研究では DNA 配列の塩基 6 つ分（ダイコドン）のパターンをモデル化して予測を行っていた。しかし、本論文の 3 章で示すように、コーディング領域が持つ情報量は、従来の手法で表現されてきたダイコドンの情報よりも豊富で、離れたアミノ酸同士にも強いパターンが現れている。これは、DNA がタンパク質に翻訳された際に形成する立体構造と関係があると考えられ、これらのパターンを用いることでより精度の高い予測を行うことが期待できる。しかし、従来研究で用いられてきた隠れマルコフモデルによって離れたアミノ酸同士の相関をモデル化しようとすると、アミノ酸同士の距離が広がるほど状態数が増加して、最適なパラメータを推定するには莫大な学習データが必要となってしまう問題点があった。また、隠れマルコフモデルなどの確率モデルでは確率的に依存する情報を同時にモデル化できないため、生物学的には有効であってもモデルに組み込むことができない情報もあった。これに対し、3 章ではブースティングとマルコフモデルのハイブリッドなアルゴリズムを提案し、離れたアミノ酸同士の情報を用いることに成功した。具体的には固定長のスライディングウィンドウで入力配列を区切り、分類器を用いて各領域についてのコーディング領域らしさを判定し、マルコフモデルによって最適なパスを探索した。

提案手法の評価にはシークエンシングエラーの含有率が 0.1%～0.5%のデータセットを 4 種類用いた。実験結果より、従来手法と比較して Matthews Correlation Coefficient(MCC)の値で 0.43%～4.6%の予測精度の向上が確認できた。また、入力配列の長さが 1000 塩基以下の結果では従来手法よりも MCC の値で 9.26%高いという大幅な精度の向上を実現できた。一般的に入力配列が短いほど配列から得ら

れる情報量が少なく、予測が難しくなる。提案手法は短い配列に対しても予測精度が高いため、cDNA の解析にとって有益である。さらに、提案手法は多数の特徴量を用いて、短い領域の情報のみからコーディング領域かどうかを判定するため、長さに依存しにくい予測を行うことができた。具体的には、1000 塩基以下と 1000 塩基以上 3000 塩基未満の配列に対する予測精度の差を従来手法と比較した場合、提案手法の方が平均で 4.1 倍小さかった。また、提案手法は入力配列の塩基一文字に対してコーディング領域らしさを示すスコアを与えるため、cDNA を人の手でアノテーションする作業を行う際に利便性が高い。このように、提案手法は予測精度、及びユーザビリティの両方で cDNA の解析に貢献ができた。

タンパク質のディスオーダー予測　（4,5,6 章）

ディスオーダーはタンパク質の構造上の特徴であるが、生物学的な目的の違いによって、ゲノム全体や大量の cDNA からタンパク質単位での発見が望まれる場合と、アミノ酸一文字単位での発見が望まれる場合がある。このため、下記の二つのように別々に目標を定め、それぞれについて予測手法を提案した。

(1) タンパク質単位でディスオーダーを予測する。（5 章）

本論文の 5 章で提案した手法の新規性は、タンパク質全体の構造空間を考慮して予測を行った点である。従来研究では、立体構造が実験的に確かめられているタンパク質から、ディスオーダータンパク質固有の配列パターンを用いて予測を行っていた。数多くの生物種で多数のディスオーダーの存在が予測されているにもかかわらず、実験的に確かめられているディスオーダータンパク質は非常に少数である。そのため、既知のディスオーダータンパク質の分布は実際のタンパク質の構造空間に対して偏っている可能性が高い。一方で、立体構造が明らかになっていないタンパク質の配列情報は多数存在し、この中にはディスオーダータンパク質が自然界のタンパク質の分布に近い状態で存在していると考えられる。そのため､構造が未知の配列を利用することでタンパク質全体の構造空間を考慮した、偏りの少ない予測を行うことが期待できる｡5 章では Transductive learning の一種である Spectral Graph Transducer を用いることで、構造が分かっていない多数の配列からタンパク質全体の分布を考慮して予測する手法を提案した。

従来手法との比較実験を行ったところ、MCC の値で 20.2%～22.1%の予測精度の向上を確認できた。従来研究との比較実験の他にも、学習器として性能の高い Support Vector Machines(SVM)を用いて、同じ特徴量を用いた予測を行い、精度を比較した。その結果、提案手法の方が MCC の値で 7.05%高い精度を得ることができた。SVM との比較では、異なる学習データを用いた場合の予測結果の相関についても調べたが、提案手法の方が平均で相関係数が 0.14 高かった。このことから提案手法の予測結果は学習データに依存しにくく、学習データが増えにくい現在

の環境であってもよりロバストな予測を行えることが分かった。

(2) アミノ酸一文字単位でディスオーダーを予測する。（6 章）
本論文の 6 章で提案した手法の新規性は、ディスオーダー領域に適した配列の分割と特徴選択を行った点である。ディスオーダーには出現しやすいアミノ酸とそうでないアミノ酸があるため、従来研究ではそれらを特徴として学習・分類が行われてきた。6 章では、統計情報を用いて配列の位置によってディスオーダーに見られる特徴が異なっていることを示し、より精度の高い予測を行うためには学習データの正確な分割・及び分割されたデータごとの特徴選択が必要であることを考察した。提案手法では、カイ二乗値を用いてアミノ酸の組成が類似している箇所を一つの領域として定義した。その結果 7 つの新しい領域が定義された。また、7 つの領域それぞれでディスオーダーの傾向を示すアミノ酸の物理化学的性質の組成が異なっていることを示し、領域ごとに異なる要素を持つ Position Specific Scoring Matrix を分類器の特徴量とする手法を提案した。分類には Support Vector Machines を用いた。
領域の分割方法と特徴選択の方法に対して 2 種類のデータセットを用いて従来研究との比較実験を行った。CASP6 のデータでは、中央部と末端部の間の領域において、提案手法は従来手法よりも MCC の値で 11%～18%良い予測精度を得ることができた。従来手法は固定長の 3 分割を用いているが、中央部分の学習データに対して他の領域の学習データは数が少ないため、従来の分割法では中央部分のディスオーダーが持つ特徴に学習器がオーバーフィッティングしてしまったと考えられる。一方で、端と中央部が比較的類似した特長を持つデータにおける予測精度は、従来手法と提案手法ではそれほど差がなかった。また、提案手法は領域ごとに特徴選択を行っているため、全ての領域で同じ特徴量を選択した従来手法と比較した場合、ほとんどの結果において提案手法が良い精度を上げることができた。提案手法は、従来では予測することが難しかった場所でディスオーダーを発見することが可能であるため、広域的なディスオーダーの予測、局所的なディスオーダーの予測の両方に貢献することができた。

本研究ではタンパク質の機能解析にとって重要な二つの課題に取り組み、従来手法よりも良い成果を上げることができた。cDNA を用いた研究は今後もより多くの生物種で行われていくため、本研究が提案した手法はその解析に貢献するだろう。また、近年注目をされはじめたディスオーダーは様々な生物種の機能を推定する上でより重要性を増していくと考えられる。これに対して、本研究が提案した手法は、ディスオーダーの広域な発見および局所的な発見の両方に貢献するだろう。

研　究　業　績

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
論文 ○	題目　ANGLE: A sequencing errors resistant program for predicting protein coding regions in unfinished cDNA 発表箇所　Journal of bioinformatics and computational biology, vol.4, No.3, p649-664 発表年月　2006年6月 著者　Kana Shimizu, Jun Adachi, Yoichi Muraoka
○	題目　Feature Selection Based on Physicochemical Properties of Redefined N-term Region and C-term Regions for Predicting Disorder 発表箇所　Proceedings of 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, p262-267 発表年月　2005年11月 著者　Kana Shimizu, Yoichi Muraoka, Shuichi Hirose, Tamotsu Noguchi
○	題目　Using Boosting to predict the protein coding region in cDNA with frame-shift errors 発表箇所　Proceedings of The 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Science 発表年月　2002年6月 著者　Kana Shimizu, Jun Adachi, Yoshihide Hayashizaki, Yoichi Muraoka
講演	題目　構造が未知な配列データを使ったディスオーダータンパク質の予測 発表箇所　第六回日本蛋白質科学会年会 発表年月　2006年5月 著者　清水佳奈　廣瀬修一　村岡洋一　野口保
	題目　2ステップSVMによる長いdisorder領域予測システムの開発 発表箇所　第六回日本蛋白質科学会年会 発表年月　2006年5月 著者　廣瀬修一　清水佳奈　金井理　野口保
	題目　Prediction of Long Disorder Region Using Two-Step SVM 発表箇所　Proceedings of the 15th International Conference on Genome Informatics 発表年月　2005年12月 著者　Shuichi Hirose, Kana Shimizu, Satoru Kanai, Tamotsu Noguchi
	題目　Feature selection based on physicochemical properties of redefined N-term regions and C-term regions for predicting disorder 発表箇所　International Symposium on Computational Biology & Bioinformatics 2005 発表年月　2005年10月 著者　Kana Shimizu, Yoichi Muraoka, Shuichi Hirose, Tamotsu Noguchi

研 究 業 績

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
講演	題目　Predicting the protein disordered region using modified position specific scoring matrix 発表箇所　Proceedings of the 15th International Conference on Genome Informatics, P150 発表年月　2004年12月 著者　Kana Shimizu, Shuichi Hirose, Tamotsu Noguchi, Yoichi Muraoka 題目　Prediction of disordered coil regions in proteins by threading and secondary structure prediction 発表箇所　Proceedings of the 6th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, P34-35 発表年月　2004年12月 著者　Tamotsu Noguchi, Shuichi Hirose, Kana Shimizu, Kentaro Tomii 題目　Predicting protein disordered regions using SVMs 発表箇所　Proceedings of the 6th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, P35-36 発表年月　2004年12月 著者　Kana Shimizu, Shuichi Hirose, Tamotsu Noguchi 題目　Predicting protein coding region in cDNA with frame shift errors using boosting based classifier 発表箇所　Proceedings of the12th International Conference on Intelligent Systems for Molecular Biology, P165 発表年月　2004年7月 著者　Kana Shimizu, Yoichi Muraoka 題目　ブースティング法を用いたマウスcDNAのコーディング領域判定 発表箇所　第25回日本分子生物学会年会 発表年月　2002年12月 著者　清水佳奈　足立淳　村岡洋一　林崎良英 題目　ブースティング法を応用したcDNA配列のコーディング領域予測 発表箇所　第63回情報処理学会全国大会 発表年月　2001年9月 著者　清水佳奈　足立淳　村岡洋一
その他	題目　N末領域・C末領域の再定義によるディスオーダー予測 発表箇所　第４回産総研　生命情報科学人材養成コースシンポジウム 発表年月　2005年9月 著者　清水佳奈　廣瀬修一　村岡洋一　野口保 題目　タンパク質のdisorder regionの予測 発表箇所　第３回産総研　生命情報科学人材養成コースシンポジウム 発表年月　2004年10月 著者　清水佳奈　廣瀬修一　村岡洋一　野口保

研　究　業　績

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
	題 目　フレームシフトを考慮した cDNA 配列のコーディング領域予測 発 表 箇 所　平成 15 年度ライフサイエンス分野融合会議・生命工学部会バオテクノロジー研究会合同研究発表会 発 表 年 月　2004 年 2 月 著 者　清水佳奈　村岡洋一 題 目　THE DESIGN METHOD OF MELODY RETRIEVAL SYSTEM ON PARALLEL-ZIED COMPUTERS 発 表 箇 所　Proceedings of Web Delivering of Music 2002 発 表 年 月　2002 年 12 月 著 者　Tomonari Sonoda, Toshiya Ikenaga, Kana Shimizu and Yoichi Muraoka 題 目　A Melody Retrieval System on Parallel-ized Computers 発 表 箇 所　Proceedings of International Workshop on Entertainment Computing 2002 発 表 年 月　2002 年 5 月 著 者　Tomonari Sonoda, Toshiya Ikenaga, Kana Shimizu and Yoichi Muraoka