科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ３ 日現在

機関番号：１３３０１
研究種目：研究活動スタート支援
研究期間：2012～2013
課題番号：２４８１００１０

研究課題名（和文）レクチン・複合糖鎖結合予測のための統計的方法論の開発

研究課題名（英文）Statistical inference of lectin-carbohydrate binding

研究代表者
広瀬　修（Hirose, Osamu）

金沢大学・電子情報学系・助教

研究者番号：３０５４９６７１

交付決定額（研究期間全体）：（直接経費）　　2,400,000円、（間接経費）　　720,000円

研究成果の概要（和文）：糖鎖は単糖類が重合した物質で，特に複合糖鎖と呼ばれる糖鎖は細胞表面などに存在する糖タンパク質に結合し，発生・分化・免疫・癌化・感染といった生体内のあらゆる局面で重要な役割を果たす．糖鎖に対する構造解析法の近年の急激な進展によって，糖鎖や糖鎖認識タンパク質であるレクチンに関する情報が急速に蓄積しつつあり，糖鎖生物学の飛躍的な進展が期待されている．本研究課題期間に特にレクチンに関して以下の様な進捗があった．(1)タンパク質データベースからのレクチンの配列情報の取得およびキュレーション．(2)局所フィッシャー判別分析による配列情報からのレクチン予測．(3)レクチンを特徴付けるアミノ酸配列の推定．

研究成果の概要（英文）：Among carbohydrates which are polymers of monosaccharaides, glycoconjugates bind to glycoproteins and play important and various roles in the process of development, differentiation, immunity, canceration, and infection. Owing to the recent development of mass spectrometry and liquid chromatography, information about glycoconjugates and lectins has rapidly accumulated. The accumulation is expected to activate researches about glycoinformatics. In our study, we obtained the following results: (1) Collection and curation of amino acid sequences for lectins, (2) Prediction of lectins from amino acid sequences using local Fisher discriminant analysis, and (3) Inference of amino acid subsequences that characterizes whether proteins are lectins or not.

研究分野：複合新領域

科研費の分科・細目：ゲノム科学・ゲノム生物学

キーワード：レクチン　複合糖鎖　判別分析　機械学習

１．研究開始当初の背景

糖鎖は単糖類が多数重合した物質であるが，そのなかでも複合糖鎖と呼ばれる糖鎖は細胞表面などに存在する糖タンパク質に結合し，発生・分化・免疫・癌化・感染といった生体内のあらゆる局面で重要な役割を果たす．複合糖鎖の構成要素はグルコースやガラクトース，マンノースといった主に 10 種類ほどの単糖類であり，それらが木構造状に結合することによっていろいろな機能や情報を担う．その代表的なものとして，例えば，複合糖鎖がタンパク質運搬のための標識となることやウィルス侵入 のセンサーとなることが挙げられる．このように，複合糖鎖はタンパク質や脂質に結合しその働きを制御することから,DNA･アミノ酸配列に次ぐ情報を担う第三のバイオポリマーとして近年非常に注目を浴びている．

そのような状況から盛んに研究されているにもかかわらず，複合糖鎖の機能の多くが未解明である，その理由は，一つ目に，比較的最近まで糖鎖の一次構造を迅速に決定する実験技術が存在しなかったことが挙げられる．また，二つ目の理由として，糖鎖認識タンパク質であるレクチンの糖鎖認識メカニズムが構造生物学的によく研究されている少数のものを除き，未だ研究の途上であることが挙げられる．一方で，質量分析法や液体クロマトグラフィー法などのような糖鎖に対する構造解析法の進展によって，複合糖鎖やレクチンに関する情報が急速に蓄積されつつある．

２．研究の目的

現在 2 値判別を行うために非常に多くの方法が提案されており，素朴な方法として，さまざまなデータに対して高い予測精度を達成する SVM の利用が考えられる．しかしながら，SVM は結果の解釈可能性という点においては必ずしも優れた手法ではない．SVM に代表される空間の分割による 2 値判別手法では，データの局所的な分布を少なくとも明示的には考慮しないため，データの背後にあるグループ構造を捉えることは困難である.例えば，アミノ酸配列からそのタンパク質がレクチンであるか否かを予測する場合，レクチンであるか否かを予測することはできても，それと同時にそのレクチンがＣ型レクチンなのかガレクチンなのかについて答えることは難しい

３．研究の方法

本研究は，結合することが既知である複合糖鎖とレクチンの組の集合をもとに，結合することが未知の組に対し，糖鎖・レクチンの分子構造あるいは機能上の意味で類似するグループがあれば，その結合未知の組も互いに結合する可能性が高いという仮定に基づき，統計モデルを構成するものである．レクチンや複合糖鎖はその機能上，あるいは，分子構造の特徴に基いていくつかのグループに分類される．このグループの特徴を自動的に抽出することで結合予測の精度向上を目指した．

本研究では，そのための準備として，アミノ酸配列からのレクチン予測を行った．解釈可能性を失うことなく高精度の予測を行うために，局所フィッシャー線形判別分析を利用した．この方法は，手法の線形性を失うことなく，ラベル情報だけでなくデータの背後にあるグループ構造を考慮しているため，解釈可能性と予測性能を両立する方法として，非常に有効である．以下この手法の概要を示す．x_i, x_j をデータ点とするとき，群間散布行列 B および群内散布行列が以下のように定義される．

$$B = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}^{(b)}(x_i - x_j)(x_i - x_j)^T$$

$$W = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}^{(w)}(x_i - x_j)(x_i - x_j)^T$$

この２つの行列に対し，その Rayleigh 比

$$\frac{a^T B a}{a^T W a}$$

を最大にするような線形写像 a を見つける方法である．すなわち，グループ間のばらつきを小さくすると同時に，グループ間のばらつきを大きくするような線形写像を見つける方法である．この手法では，重み行列 $W^{(b)}$, $W^{(w)}$にラベル情報とともに，データの局所性を考慮するため，グループ内に隠されたグループ構造が存在する場合にも有効な手法となる．

この手法は，比較的高次元の分類問題に非常に有効であるが，本研究でのレクチンデータのように，入力ベクトルが 8000 次元に達するという超高次元データの分類に対しては決して有効な方法ではない．一般的に標本数がデータの次元よりも小さい場合群間散布行列は常にランクが欠損する．比較的低次元のデータであれば，正則化と呼ばれる手法によって，このランク欠損の問題を回避可能である．しかしながら，本研究でのデータセットのように極端な超高次元のデータの場合，群間散布行列の零空間は膨大であるため，正則化を行ったとしても，得られる分類器の予測性能は不安定なものとなりうる．この問題を回避するために，本研究では，正則化ではなく，ランク制約によるランク欠損の回避するような局所フィッシャー判別分析手法を開発した．

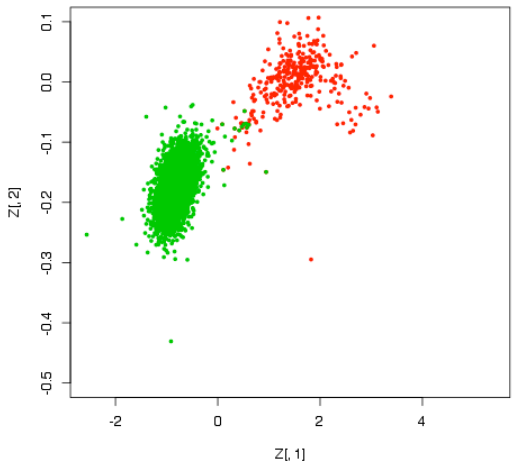
４．研究成果

本研究課題期間に特に糖鎖認識タンパク質であるレクチンに関して次のような進捗があった．

(1) タンパク質データベースからのレクチンの配列情報の取得，およびキュレーション．
(2) 局所フィッシャー判別分析を利用した配列情報からのレクチン予測．
(3) レクチンを特徴付けるアミノ酸配列の推定．

特に(2)では，線形判別による解釈可能性を失うことなく SVM と同程度の予測性能を達成した．

また，この課題に取り組む課程で局所フィッシャー判別分析という分析手法そのものに関する重要な知見が得られ，判別分析手法自体の拡張に繋がったことも成果の１つとなった．今後もさらに研究を続け，配列情報だけでなくレクチンの立体構造を考慮に入れた結合予測手法の開発などを目指していきたい．



(図１) レクチン（赤）とレクチンではないタンパク質（緑）の局所フィッシャー判別分析による２次元プロット．明確な違いが見てとれる．x 軸，y 軸がそれぞれ同手法による第１および第２固有ベクトルを表す．

５．主な発表論文等

〔雑誌論文〕（計 3 件）

① Tu Kien T. Le, Osamu Hirose, Vu Anh Tran, Thammakorn Saethang, Lan Anh T. Nguyen, Xuan Tho Dang, Duc Luu Ngo, Mamoru Kubo, Yoichi Yamada, Kenji Satou, Predicting residue contacts for protein-protein interactions by integration of multiple information, *Journal of Biomedical Science and Engineering*, **7**, pp.28-37,査読有．

DOI:10.4236/jbise.2014.71005

② Naoki Nariai, Osamu Hirose, Kaname Kojima, Masao Nagasaki, TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference, *Bioinformatics*, **29**(18), pp.2292-2299, 査読有．

DOI:10.1093/bioinformatics/btt381

③ Xuan Tho Dang, Osamu Hirose, Thammakorn Saethang, Vu Anh Tran, Lan Anh T. Nguyen, Tu Kien T. Le, Mamoru Kubo, Yoichi Yamada, Kenji Satou, *Journal of Biomedical Science and Engineering*, A novel over-sampling method and its application to miRNA prediction, **6**, pp.236-248

DOI:10.4236/jbise.2013.62A029

〔学会発表〕（計　0 件）

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
広瀬 修（HIROSE Osamu）
金沢大学・電子情報学系・助教
研究者番号：30549671

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：