# タンパク質間相互作用データからの相関ルール発見

○尾山 卓也[1], 北野 景彦[1], 佐藤 賢二[2], 伊藤 隆司[3]
[1]インテック W&G, [2]北陸先端科学技術大学院大学 知識科学研究科, [3]金沢大学がん研究所

## 1 はじめに

近年、two-hybrid 法を用いた網羅的なタンパク質間相互作用検出実験が頻繁になされるようになり、その結果大量のタンパク質間相互作用事例が蓄積されつつある[1]。しかしながら、個々のタンパク質間相互作用に関しては多く蓄積されているが、それらの多くに共通する一般的な知識（例えば、ある特徴を持つタンパク質は、別のある特徴を持ったタンパク質と相互作用しやすいなど）の発見といった意味ではまだ多くはない。一方、膨大なデータの中からそこに隠れた有用な情報を発見する「データマイニング」が近年注目され、バイオインフォマティクス分野への適用研究もいくつか見られる[2]。

このような背景のもと、筆者らはデータマイニングの代表的な手法である「相関ルール発見」を用いて、大量に蓄えられたタンパク質間相互作用データから相互作用に関する知識を発見する手法およびシステムの研究開発に取り組んでいる。

## 2 相互作用からの知識発見の概要

図１は相互作用からの知識発見の概要を表している。データマイニングに用いるデータは two-hybrid 実験等で得られた数千の酵母タンパク質相互作用データと、相互作用する各タンパク質の特徴から生成される。相互作用は伊藤ら[1]の two-hybrid 実験や YPD*、MIPS*等のウェブサイトから得ている。また各タンパク質の特徴は、ウェブ上で公開されている様々なゲノムデータベースを用いて、機能的、配列的あるいは構造的な様々な観点で定義した。以下に本研究で定義した６つのタイプの特徴を示す。

◆YPD において、タンパク質は「生化学的機能」、「細胞内での役割」等の観点に基づき数十のカテゴリに分類されている。第１のタイプの特徴は、タンパク質が分類されるカテゴリによって定義される。すなわち、あるタンパク質がカテゴリ C に分類される場合、そのタンパク質はカテゴリ C が意味する特徴を有するとみなす。

◆酵素はその機能に基づき、EC 番号でラベル付けされた多くのカテゴリに分類される。第２のタイプの特徴は、各タンパク質がどの EC 番号に対応づけられているかによって定義される。

◆酵母タンパク質の多くは SWISS-PROT*や PIR*に対応するエントリが存在する。それぞれのエントリには多くの場合、機能的、配列的あるいは構造的特徴を表すキーワードが記述されている。第３のタイプの特徴は、どのようなキーワードが割り当てられているかによって定義される。

◆第４タイプの特徴は、PROSITE*に登録されたモチーフが当該タンパク質のアミノ酸配列上に存在するか否かによって定義される。

◆第５タイプの特徴は、タンパク質アミノ酸配列上のアミノ酸の偏り、即ち配列上のある短い区間に特定の種類のアミノ酸が多数出現しているか否かによって定義される。

◆約 6,000 個の酵母タンパク質間の総当りのホモロジー検索により、全タンパク質上のアミノ酸部分配列を、互いに類似するグループに分類し、各タンパク質がどのグループの部分配列を持つかについて調べた。その結果を用いて第６のタイプの特徴、即ち「各タンパク質の部分配列が分類される部分配列グループ」という特徴を定義した。

酵母の全タンパク質について特徴を定義した結果、数千種類の特徴を得た。それらの各タンパク質の特徴及び同じく数千の相互作用事例からデータマイニングを適用するデータを作成した。相関ルール発見を用いたデータマイニングでは、一つのタンパク質（あるいは一つの遺伝子）を各トランザクションに対応させる場合が多い。しかしながら本研究は相互作用するタンパク質間の関係に注目しているため、タンパク質よりも寧ろ一つの相互作用を各トランザクションに対応させる必要があった。すなわち、各トランザクションは一つの相互作用を表し、かつ相互作用する二つのタンパク質（図１で示される「タンパク質 A」および「タンパク質 B」）それぞれの特徴をも持たせる必要があった。そしてタンパク質 A の特徴が条件部、タンパク質 B の特徴が結論部に存在する相関ルールのみを抽出する。

*Web 上で公開されているゲノム情報データベース

## 3 マイニング結果

データマイニングの実行の結果、数多くの相関ルールが得られた。その中には、「SH3 ドメインはプロリンリッチな部位に結合する」という既に生物学的に認知されている知見も含まれていた。その他に未知のルールも多く含まれているが、それらのいくつかは生物学的実験によって検証していきたい。

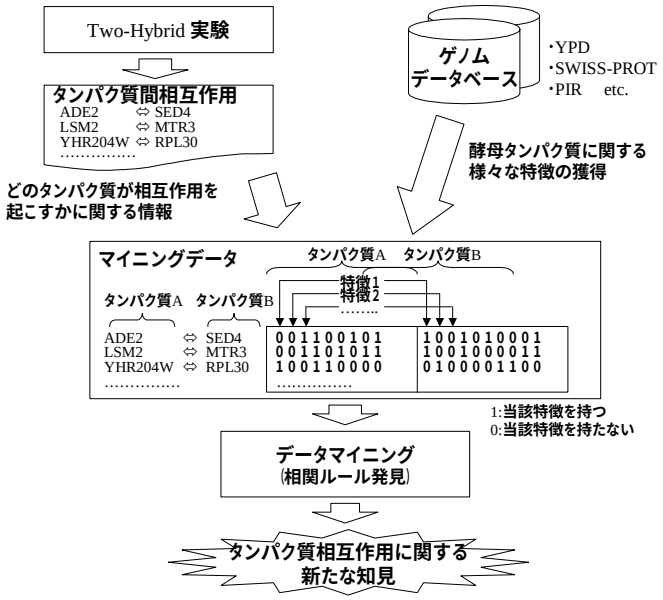


図１ 本手法の概要

## 参考文献

[1] T. Ito, et al., Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins, Proc. of the National Academy of Science of USA, 97(3):1143-1147, 2000.

[2] K. Satou, et al., Finding association rules on heterogeneous genome data, Proc. Pacific Symposium on Biocomputing ’97, 397-408, 1997.