E-020

医学・生物論文内の頻度情報を利用したタンパク質相互作用の自動抽出と可視化
Automatic Extraction and Visualization of Protein Interaction from Medical/Biological Papers using Frequency Information in the Documents

藤川 裕充 † **土井 晃一** ‡ **大森 晃** †
Hiromitsu Fujikawa Kouichi Doi Akira Ohmori

1. はじめに

近年，タンパク質の相互作用が実験によって次々と明らかにされてきている．実験の結果は論文に発表され，生物・医学文献データベース Medline[1] に蓄積される．しかし，実験結果は多量で，しかも，次々に実験結果が発表されるので，人手によって整理することは困難である．そこで，タンパク質の相互作用に関する情報を論文から自動的に抽出しようとする研究が行われている．例えば，Ono ら [2] は正規表現を用いることにより酵母菌に関するタンパク質の相互作用情報を論文から抽出した．しかし，現実の論文では，文の範囲，関係詞句の処理などが困難なため，相互作用情報の正確な抽出は困難である．また，生物学の専門家には，タンパク質の相互作用の全貌を知りたいという要求がある．そこで，我々は，

- 論文中でのタンパク質の出現頻度
- タンパク質の相互作用を意味する文字列が論文中に現れるかどうか

を基にして，タンパク質の相互作用を推定した．さらに，相互作用の全貌を見るため，その一部分を可視化した．

2. 準備

MeshTerm[3] の protein のエントリーから，木の葉の方向を探索することにより，タンパク質の名前を抽出した．抽出したタンパク質の一般名を走査するプログラム (字句スキャナ)，および，Ono らの論文を参考にして，表 1 に示した文字列（前後にスペースを含む）を走査するプログラムを作成した．これらのプログラムと，flex を用いてオートマトンを生成した．また，Medline から abstract 部分のみを抽出した．

3. 頻度情報の抽出

作成したデータ，及び字句スキャナを利用して，タンパク質の一般名を 2 つ以上含み，かつ表 1 に示した文字列をどこかに含む abstract を抽出した．抽出した abstract と，それに含まれるタンパク質の一般名のデータを利用して，同一 abstract 内に出現するタンパク質の一般名の組合せが，Medline の全 abstract 内で出現する回数を集計し，200 回以上出現した組合せを抽出した (表 2)．

4. 可視化

タンパク質の一般名の組合せで 200 回以上出現したものを，Graphviz[4] を用いて可視化した．その全体像を図 1 に，その一部を拡大したものを図 2 に示す．可視化する際に，組合せの出現回数により重み付けを行っているため，出現回数が多い組合せほど近くに表示されている．

†東京理科大学, 工学部第二部経営工学科
‡奈良先端科学技術大学院大学,NAIST

	抽出文字列
associate	associate with,associated with
association	association between,association of, association with each other
bind	bind,bind between,bind of,bind to, bound,bound between,bound of, bound to
complex	complex,complex with,complexed, complexed with
interact	interact with,interacted with
interaction	interaction among,interaction of, interaction between

表 1: 抽出に利用した文字列

件数	組合せ
4637	igg - igm
2900	iga - igg
2124	iga - igm
1799	igg - complement
1441	fibrinogen - thrombin
1351	collagen - laminin
1250	fibrin - fibrinogen
1187	igg - immunoglobulins
1049	antithrombin iii - thrombin
1007	fibrin - plasminogen

表 2: 抽出した組合せの一部 (上位 10 組)

可視化の結果を生物学の専門家に見てもらった結果，以下のような所見を得た．

- ほとんど (約 8 割) の組合せが正しい
- タンパク質の相互作用の全貌がわかる
- 分野 (血液，筋肉，目など) ごとに島が出来ている
- 関係が密になっているところが一目でわかるので，重要なタンパク質，あるいは，よく研究されているタンパク質がすぐにわかり，また，そのタンパク質と関係の深いタンパク質が一目でわかる
- 関係が疎になっているところがわかるので，まだ研究の進んでいない，あるいは，相互作用があまりないタンパク質がすぐにわかる
- 機能が分かっていないタンパク質の機能を推定出来そうである

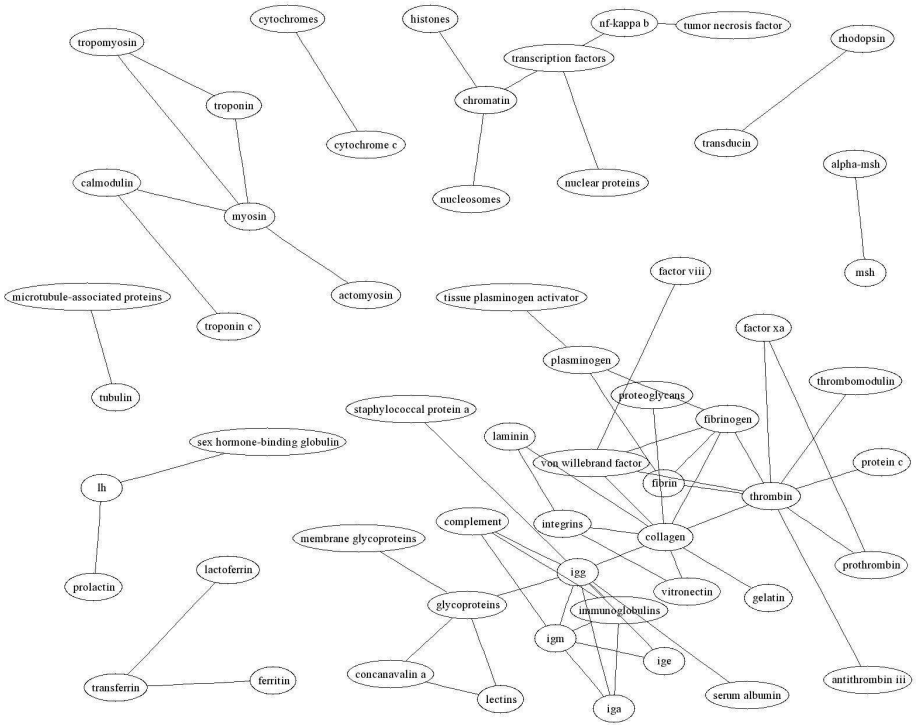

図 1: タンパク質の一般名の組合せの可視化

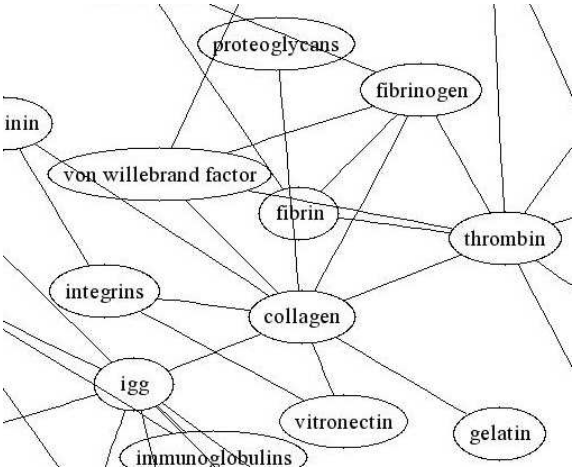

図 2: タンパク質の一般名の組合せの可視化 (一部)

- タンパク質名として一般的なものが並んでいる．もっと特殊なもの (タンパク質の個有名) も欲しい
- 特定の生物種に関してできないか? そのほうが評価しやすい
- タンパク質名ではなく，対応する遺伝子名で書かれているものはどうするのか

5. おわりに

本研究は，以下の 2 点で有用と考える．

1. Medline をベースにして，タンパク質の相互作用の全体像を見えるようにした
2. それによって，たんぱく質の機能を予測できる可能性を開いた

謝辞

本研究を進めるにあたって，ご助言と評価に協力して下さった，奈良先端大の土居先生，小笠原先生，森先生，三森氏，Sevrani 氏に感謝致します．

参考文献

[1] Medline http://www.ncbi.nlm.nih.gov/

[2] Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami, Toshihisa Takagi: Automated extraction of information on protein-protein interactions from the biological literature. BIOINFOMATICS 17(2): 155-161 (2001)

[3] MeshTerm http://www.nlm.nih.gov/mesh/MBrowser.html

[4] graphviz http://www.research.att.com/sw/tools/graphviz/