

遺伝子データベースからの
テキストマイニングによる
Pathwayにおけるタンパク質の可視化

1815070 武藤 克弥

富山県立大学 電子・情報工学科

November 24, 2021

背景

遺伝子データベースにテキストマイニングを適用し，生命科学分野の新たな知見を得る試みは，依然として需要が高い．その中でもタンパク質間の相互作用を示した Pathway を分析して，タンパク質の機能予測や，化合物の反応経路予測を行うことが盛んである．

目的

- ① Pathway 内のタンパク質に対して共起分析を行い，共起ネットワークを 3D グラフに描画する
- ② 得られたネットワークに対して新たな分析をユーザに促すきっかけを作る

Pathway とは

- 代謝で起こる酵素反応, シグナル (神経) 伝達を描いたマップ
→ 遺伝子・タンパク質の相互作用が見られる
- 各分野で得られた実験結果から専門家が手動で作成している
- 四角: タンパク質 (遺伝子), 丸: 化合物 (酵素), 矢印+p: リン酸化, 矢印+m: メチル化 などを表す

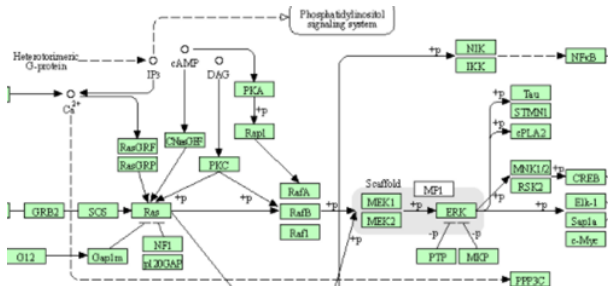


図 1: Pathway の例

- Web サイト上にある文章中のある単語に対して、付随して現れる単語を分析
- 共起のネットワーク (3D グラフ) で表現
- 単語どうしのつながりからアイデアの発送支援を行う

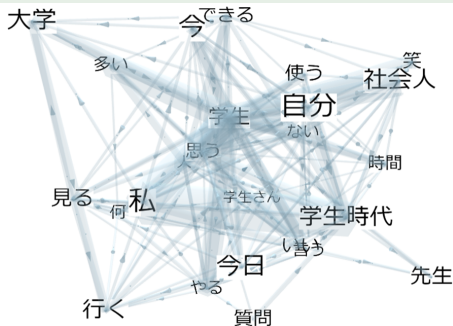


図 2: 3D グラフの共起ネットワークの例

文献内共起

- PubMed の Mesh ターム等を用いて、各文献内に出現する遺伝子・タンパク質の類似度・共起性を分析

ゲノム配列内共起

- ゲノム配列内に出現する対象の遺伝子から、一定の範囲内に出現する他遺伝子の出現頻度、最小距離などを分析

未知経路の予測

既存の Pathway データを用いて未知の反応経路、遺伝子相互作用を予測する

- 予測精度を上げるための特徴量の検討
- オリジナルの特徴量を組み込んだ分類モデルの評価

重要ノードの特定

Pathway もノード（頂点）とエッジ（辺）を持つグラフ構造とみなせる

- R 言語の KEGGgraph パッケージを用いて、Pathway(xml 形式)を解析し、すい臓がん Pathway とつながる 8つの Pathway をすべて統合
- グラフ理論の媒介中心性（どれくらい他ノードに情報伝達しているか）を用いて、すい臓がんパスウェイの中で重要な上位 3つのノードを特定

Pathway 分析 (2)

7/13

背景
研究概要
調べたこと
提案手法
課題
今後の予定

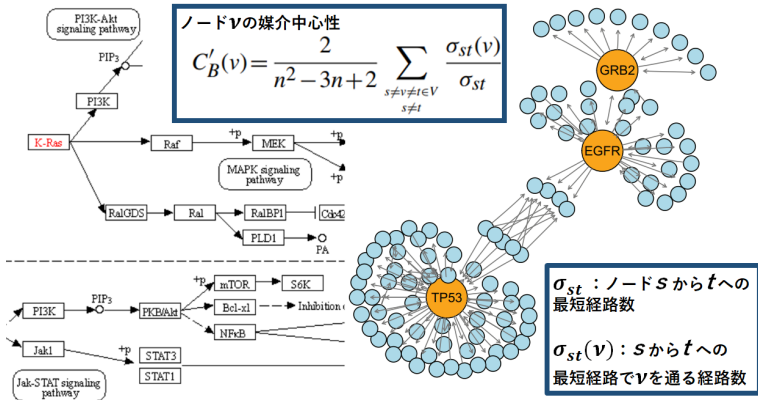


図 3: 統合された Pathway から特定された 3 つの重要ノード

Pathway 上での共起分析

- ① ヒトの Pathway と遺伝子情報を KEGG から取得
- ② 同じ Pathway 内に同時出現するタンパク質 = 互いに共起し合っていると仮定し、共起頻度をカウント
- ③ 共起頻度に応じてエッジの色分け→ 3D グラフ出力

	hsa:5594	hsa:5595	hsa:5291	hsa:5293	hsa:5290
hsa:5594	0	117	84	84	84
hsa:5595	117	0	84	84	84
hsa:5291	84	84	0	106	106
hsa:5293	84	84	106	0	106
hsa:5290	84	84	106	106	0
hsa:8503	83	83	103	103	103
hsa:5296	83	83	103	103	103
hsa:5295	83	83	103	103	103
hsa:207	81	81	90	90	90
hsa:10000	81	81	90	90	90
hsa:208	81	81	90	90	90
hsa:5604	89	89	71	71	71

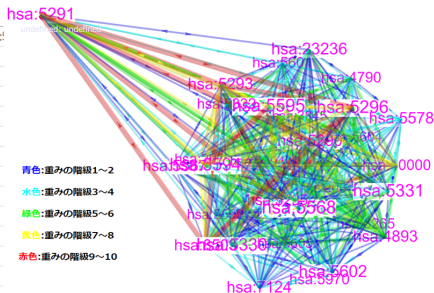


図 4: Pathway 共起分析

3D グラフから分かったこと

- 共起頻度が高い赤色のネットワークに着目
- ホスファチジルイノシトール 3-キナーゼ (PI3K) というシグナル伝達タンパク質のネットワーク, セリン/スレオニンキナーゼ (Akt) の 3 種類のアイソフォームからなるネットワークが得られた.
- 2つのネットワークともがんに関連している→なぜこの2つが1番強く現れたのかを分析することにつながる?

共起分析の改善

- 現状は同じ Pathway 内にあれば共起
→離れすぎていたら共起といえないのではないか
- 同 Pathway 内に存在 かつ 矢印でつながっているとき共起頻度をカウントするように改善
- KEGGgraph パッケージを使って試作中

提案した背景

- Pathway= 世界中様々な分野の実験結果等から作成
→ Pathway 内のノード同士のつながりから遺伝子の共起性を調べることで、未知の関係が分かるのでは？

課題

- 何を持って共起しているとみなせるのか分かっていない
- 3D グラフで得られる情報が一般的な知識で、さらなる分析にならないのではないか
- 現在：ヒトのパスウェイ & 遺伝子全体で分析（範囲が広く、何が重要なのか不明瞭）
→ 特定の疾患・パスウェイなどに焦点を絞った方がよいのでは？

逆合成経路の推定

- ① 欲しい物質（図5左）を作る方法が複数あるとしたとき
- ② 生成するコスト・手数がかからない最適な合成パターン（経路）を推定する

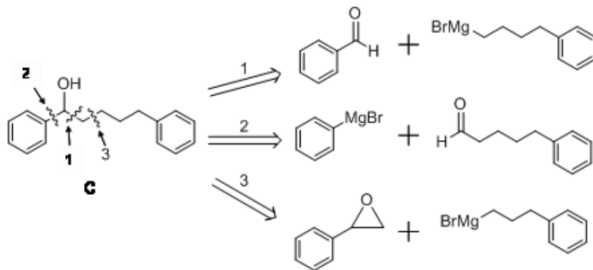


図 5: 3 通りの合成経路がある場合の例

KEGG REACTION データベースから反応式の抽出

- 反応パスウェイとそのパスウェイ内に登録されている反応式を記した対応表を作成
- 反応式の R 番号から反応式を抽出

map_id	description	
m00010	Glycolysis / Gluconeogenesis	R00014; R00199; R00947; R00959; R02073; R02187; R
m00020	Citrate cycle (TCA cycle)	R00014; R00268; R
m00030	Pentose phosphate pathway	R00305; R01049; R01544; R01621; R02740; R02749; R
	Pentose and	R00264; R00286; R01529; R01540; R01903; R01904; R

ENTRY	NAME	DEFINITION	EQUATION
R00001	polyphosphate polyphosphohydrolase	Polyphosphate + n H2O <=> (n+1) Oligophosphate	C00404 + n
			C00001
			<=> (n+1)
			C02174
R00002	Reduced ferredoxin:dinitrogen oxidoreductase (ATP-hydrolysing)	16 ATP + 16 H2O + 8 Reduced ferredoxin <=> 8 e- + 16 Orthophosphate + 16 ADP + 8 Oxidized ferredoxin	16 C00002
			+ 16
			C00001 + 8
			C00138
			<=> 8
			C05359 +
			16 C00009
			+ 16
R00004	diphosphate phosphohydrolase;;pyrophosphate phosphohydrolase	Diphosphate + H2O <=> 2 Orthophosphate	C00013 +
			C00001
			<=> 2
			C00009
			C01010 +

図 6: 対応表と反応式の表

逆合成経路推定の場合

経路置き換えに対応するシステムに変更する

- 目的の物質・中間生成物等をキーワードにして，Pathway 内でその物質を生成している経路を検索
- 抽出して置き換える

3D グラフ共起分析の場合

- 意味のある課題を見つけ出し，それに向けたシステムを作成する．