IBM research

Tokyo Research Laboratory

# 生体ネットワーク予測の機械学習

タンパク質
相互作用

教師つき学習

鹿島久嗣
IBM東京基礎研究所

## 生体（とくに、タンパク質）ネットワーク予測への機械学習的アプローチを概観します

- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ

## 生体（とくに、タンパク質）ネットワーク予測への機械学習的アプローチを概観します

- **ネットワーク予測問題**
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ



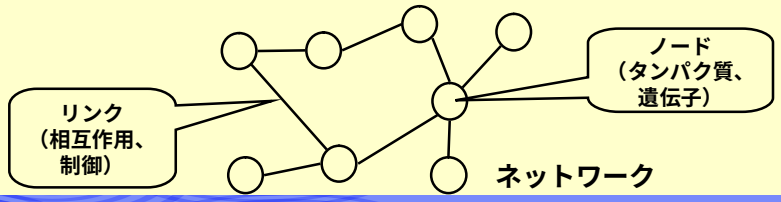
## 生体（とくに、タンパク質）ネットワーク予測への機械学習的アプローチを概観します

- ネットワーク予測問題
  - **ネットワーク予測問題の定義**
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ

## 生物学における「ネットワーク」の予測問題を考えます

- ネットワークとは、
  - タンパク質や遺伝子などの主体を「ノード」として
  - 相互作用や制御など、それらの間の関係を「リンク」として
    （向きのある場合、ない場合がある）

  グラフ表現したもの
- ネットワークの構造を予測することによって、
  - リンクの存在を確かめるための実験（通常、高コスト）を実際にするまえに、候補の絞込みができる
  - 実験的に示されたリンクの信頼度を測ることができる
  - 予測されたネットワークに基づいた、仮説を考えることができる
- 生物以外でのネットワーク： WWW、社会ネットワーク、…

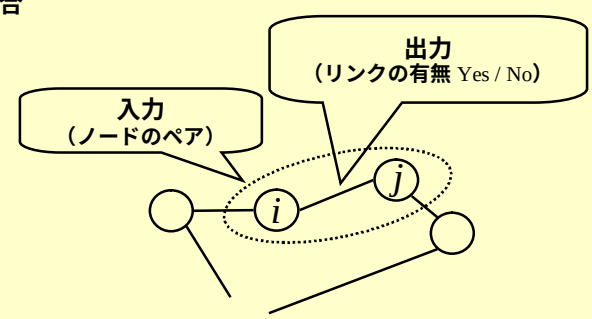





## ネットワーク予測の問題は、任意の2つのノードの間に、リンクがあるかどうかを予測する問題として捉えることができます

- 各ノード間のリンクの有無を予測したい
  - 入力は、2つのノード $i$ と $j$
  - 出力は、2つのノードの間にリンクがあるか否か（Yes / No）
- ただし、ノードの集合はわかっている
- リンクの存在は、
  - 予め、いくつか分かっている場合
  - まったく分かっていない場合

  と、がある

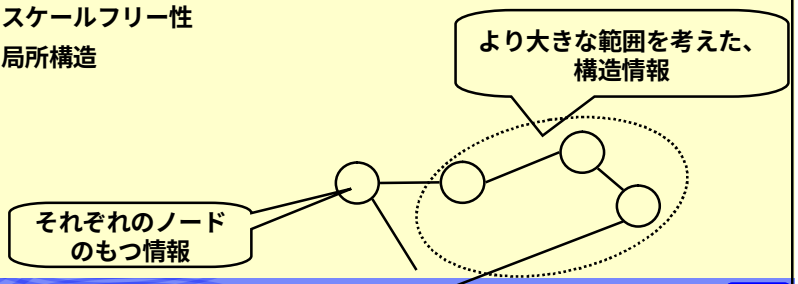


- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ



## 何の情報をもとにネットワークを予測するか？ 大きく分けて、ノードの持つ情報 と 構造の持つ情報の2つがあります

- それぞれのノード（タンパク質）は、いくつかの情報をもっている
  - 配列情報： 文字列
  - 発現情報： 実数値、時系列
  - 立体構造情報： 文字列、座標（の列）
  - その他のアノテーション： 局在部位、機能、文献、…
- 一方、ネットワークを、より大きな視点で見た、構造的な情報がある
  - ネットワークのスケールフリー性
  - ネットワークの局所構造



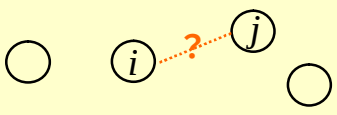
- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ



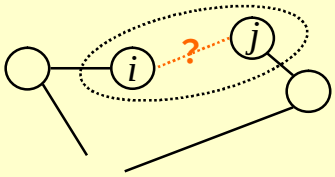
## 予測においてとるべきアプローチは、リンク情報があるかどうかによって大きく２つに分かれます

- いくつかのリンクについて、その有無が分かっているかどうかによってアプローチの仕方が異なってきます
  - タンパク質相互作用や遺伝子の制御関係の場合、実験的に確認された相互作用がある
    - タンパク質なら Y2H (Yeast Two-Hybrid)、 TAP (Tandem Affinity Purification) などの実験手法
- 無い（使わない）場合には、生物学的知見をもとにした推測（次頁）
- ある場合には、そのデータをもとに、リンクを予測するルールを発見 （あとで）

過去に観測されたリンク情報がない場合




過去に観測されたリンク情報がある場合



- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ



## リンク情報がないときの予測は、生物学的な知見にもとづくルールを用いて、予測を行います

- 相互作用するタンパク質がもつであろう特徴を利用して、予測する
- 遺伝子フュージョン法
  - 祖先において、ひとつの遺伝子だった場合、相互作用する可能性が高い
- 系統発生プロファイル法
  - 種を超えて、共起のしかたが一致している場合、相互作用する可能性が高い
- 遺伝子近傍法
  - 種を超えて、順序が保存されている／近くに固まっている ものは相互作用する可能性が高い
- 発現データを使う方法
  - 発現データが似ているものは、相互作用する可能性が高い
  - より賢く、グラフィカルモデル（今回は深入りません）

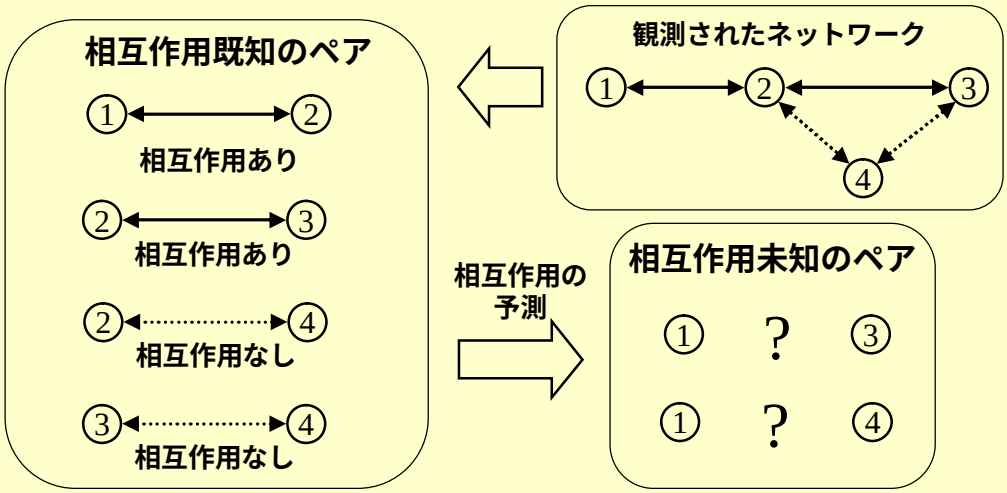
- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ



- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ

## リンク情報があるときの予測は、教師つき学習の問題として捉えることができます

- せっかく実験的に得られた相互作用があるので、これを用いて予測を行いたい
- 機械学習においては、教師つき学習の文脈で定式化される

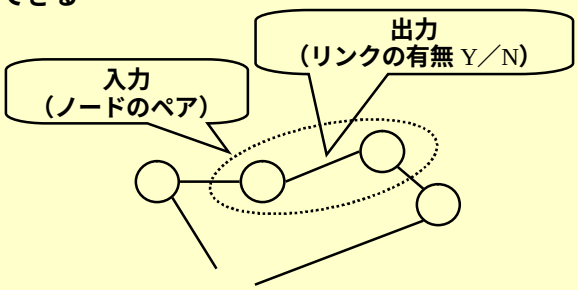





## 教師つき学習は、リンク情報があるときに、これをもとにリンクを予測することのできる枠組みです

- 教師つき学習とは、（ネットワーク予測の文脈では）
  - 入力：2つのノード
  - 出力：2つのノードの間にリンクがあるか否か（Yes／No）

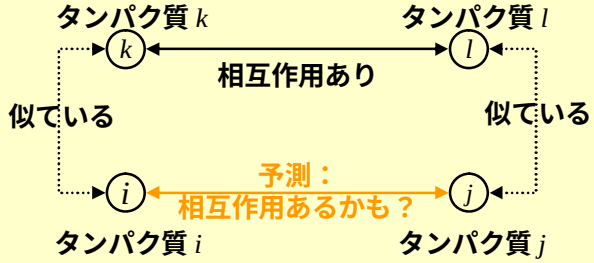
  の関係（入力から出力を予測する関数）を、予め与えられた入出力の事例（リンクの有無が分かっているノードペア）から推定する問題
- 一旦、入出力の関係がわかれば、リンクが未知のノードペアに対して、リンクの有無を「予言」することができる



## 教師つき学習の基本的な考え方は「入力が似ていれば、出力も似ている」です

- 教師つき学習は、過去の事例との類似によって、未来の事例についての予測を行う
  - 過去の事例： 過去に、入力 X に対する出力が Y であることがあった
  - 未来の事例と過去の事例との類似： 出力未知の入力 X' は X に似ている
  - 未来の事例に対する予測： 従って、入力 X' に対する出力は、恐らく Y である
- タンパク質の相互作用予測の文脈で考えると...

過去の事例： 相互作用の有無がわかっているタンパク質ペア



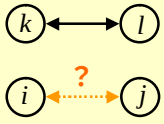

未来の事例： 相互作用の有無が未知のタンパク質ペア

## カーネル法は「入力が似ていれば、出力が似ている」を賢く実現するための、ひとつの方法です

- カーネル法の一般的な形は、

$$y(i,j) = \sum_{(k,\ell)} \alpha(k,\ell)\ K^{pair}((i,j),(k,\ell))$$

過去の事例（重み：$\alpha(k, l)$）




未来の事例

たんぱく質ペア$(i,j)$についての予測（正ならはリンクあり、負ならばリンク無し）

タンパク質ペア$(k,l)$の重み（正負の値をとる）

タンパク質ペア$(i, j)$とタンパク質ペア$(k, l)$の類似度

- 過去の事例がそれぞれ重要度 αをもっていて、それぞれが投票を行うことで、多数決による予測を行う
  - 過去の各事例（タンパク質ペア $k\&l$ ）が 重みをもっている
    - リンクがあるほうに投票するなら正、無いほうなら負の重み
  - 投票の量は、事例の重要度 $\alpha(k,l)$ と、事例と予測したいペアの類似度 $K^{pair}((i,j), (k,l))$ をかけたもの

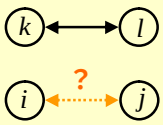


## カーネル法のパラメータの決め方 １：カーネル法における各事例の重みαは、過去の事例をもとに学習します

- 重みαは、過去の事例をつかった訓練によって予め調整（学習）される
  - リンクのあるペアについては、結果の値が大きくなるように
  - リンクのないペアについては、結果の値が小さくなるように
- α の決め方は、単純には、
  - $(k,l)$ 間に相互作用があるなら $\alpha(k,l) = +1$
  - $(k,l)$ 間に相互作用がないなら $\alpha(k,l) = -1$

とすればよい

過去の事例（重み：$\alpha(k, l)$）




未来の事例

$$y(i,j) = \sum_{(k,\ell)} \alpha(k,\ell)\ K^{pair}((i,j),(k,\ell))$$

たんぱく質ペア$(i,j)$についての予測（正ならはリンクあり、負ならばリンク無し）

タンパク質ペア$(k,l)$の重み（正負の値をとる）

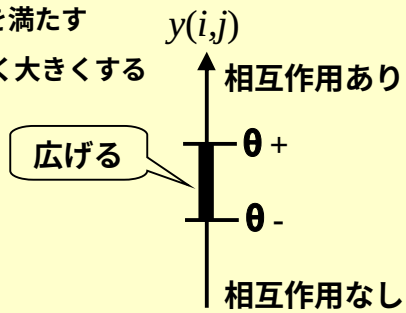
タンパク質ペア$(i, j)$とタンパク質ペア$(k, l)$の類似度

## カーネル法のパラメータの決め方 2：もう少し賢く決めるなら、サポートベクトルマシン

- $(i,j)$ 間に相互作用があるなら $y(i,j) > \theta^+$ を満たす
- $(i,j)$ 間に相互作用がないなら $y(i,j) < \theta^-$ を満たす
- 分離が良くなるように、 $\theta^+ - \theta^-$ をなるべく大きくする

となるように、$\alpha$ を調整する

  - 問題としては「2次計画問題」となる




$$y(i,j) = \sum_{(k,\ell)} \alpha(k,\ell)\ K^{pair}((i,j),(k,\ell))$$

たんぱく質ペア$(i,j)$についての予測（正ならはリンクあり、負ならばリンク無し）

タンパク質ペア$(k,l)$の重み（正負の値をとる）

タンパク質ペア$(i,j)$とタンパク質ペア$(k,l)$の類似度



## タンパク質同士の類似度は、タンパク質の特徴を表現した「特徴ベクトル」の積として定義できます

- タンパク質ペア同士の類似度を定義するまえに、まず、タンパク質同士の類似度を定義する
- ひとつのタンパク質は、その特徴を列挙した「特徴ベクトル」によって表現する

（タンパク質 $i$ の特徴ベクトル）＝（特徴 $1$, 特徴 $2$, 特徴 $3$, ...., 特徴 $N$）

  - 特徴の例：
    - 配列情報： あるアミノ酸配列が、部分文字列として含まれているかどうか
    - 発現情報： ある時点での発現量
- ２つのタンパク質の類似度 $K(i,j)$ は、特徴ベクトルの内積として定義される

$K(i,j) = <$(タンパク質 $i$ の特徴ベクトル), (タンパク質 $j$ の特徴ベクトル)$>$

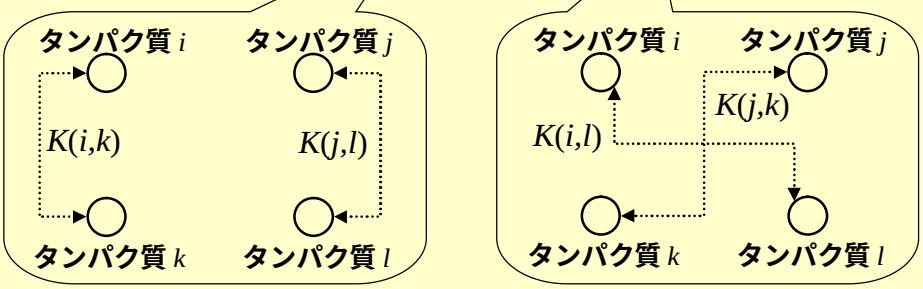
- 特徴ベクトルが定義しにくい場合には、直接類似度を定義してもよい
  - 配列アラインメントのスコア
  - 時系列間の相関係数

## タンパク質ペア同士の類似度関数（カーネル関数） $K^{pair}((i,j),(k,l))$ は、タンパク質同士の類似度関数から構成されます

- タンパク質ペア同士の類似度 $K^{pair}$ を、タンパク質同士の類似度 $K$ の積として表す

$$K^{pair}((i,j),(k,\ell)) = K(i,k)K(j,\ell)+K(i,\ell)K(j,k)$$

  - 掛け算はAND、足し算はOR イメージ




- リンクに向きがある場合には、足さずに片方だけつかう

Ben-Hur & Noble: *Kernel methods for predicting protein–protein interactions*, Bioinformatics, Vol. 21 Suppl. 1, 2005



## カーネル法のひとつであるサポートベクトルマシン(SVM)を用いて、予測精度が改善された例があります

- $K^{pair}$内のタンパク質同士の類似度$K$： モチーフ、ドメイン、文字列などを用いて構成された特徴ベクトルによって定義する
- さらに、相互作用するタンパクペアの関係自身を使った別の類似度関数も考えられる

$$K^{pair'}((i,j),(k,\ell)) = K'(i,j)K'(k,\ell)$$

  - $K'(i,j)$は、タンパクペア $(i,j)$ が
    - 同じ場所に局在する
    - 同じ機能に携わる
    - 別の種で相互作用がある
    - 共通の隣接ノードをもつ

    などによって定義する
- 全ての類似度を組み合わせたものを使って、SVM を適用する

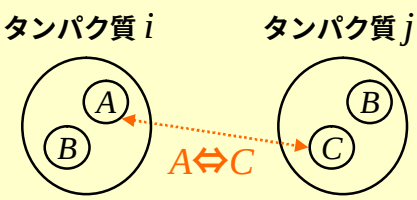
Ben-Hur & Noble: *Kernel methods for predicting protein–protein interactions*, Bioinformatics, Vol. 21 Suppl. 1, 2005

- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ



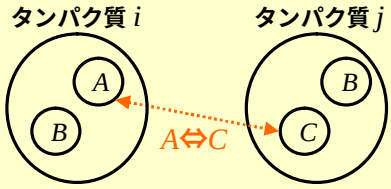
より、タンパク質事情に特化した手法として、
「ドメイン情報を活用する」という手があります

- 実際に相互作用しているのは、「ドメイン」と呼ばれる進化的に保存された領域
  - タンパク質が相互作用するということは、ドメインが相互作用するということ
- 漠然と配列全体を使うのではなく、ドメイン単位で考えたほうがいいかもしれない
  - タンパク質を、ドメインの集合としてとらえる
  - 各ドメインがどのような配列をしているか、という情報はデータベース(pfamなど)に蓄積されている



## ドメイン情報にもとづく予測手法は、ドメインの相互作用強度を見積もり、これをもとにタンパク質相互作用を予測します

- ドメイン同士が相互作用する強さがわかれば、タンパク質が相互作用するかどうかも予測できるはず
  - if $A \Leftrightarrow C$ :ドメイン$A$とドメイン$C$は強く相互作用
  - then タンパク質 $i$ とタンパク質 $j$ は相互作用



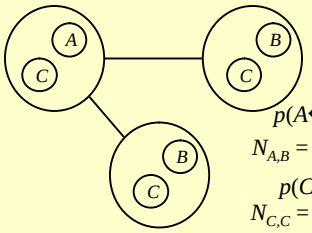

- 問題点： どのドメインとどのドメインが相互作用するかは、あまりわかっていない
- まず、ドメイン間が相互作用する確率（あるいは強さ）を、何らかの方法で推定
  - ドメイン$A$とドメイン$B$の相互作用確率を $p(A, B)$ とする
- それを使って、タンパク質間相互作用を予測する。
  - たとえば、$p(i \Leftrightarrow j) = \max_{A \in D_i} \max_{B \in D_j} p(A \Leftrightarrow B)$

$D_i$：タンパク質 $i$ 上のドメイン集合
$D_j$：タンパク質 $j$ 上のドメイン集合



## ドメイン相互作用確率推定法：もっとも単純な「アソシエーション法」なら数えるだけでOK

- ドメイン間相互作用の確率を推定する、もっともシンプルな予測方法
- 考え方： ドメイン$A$とドメイン$B$が、相互作用するタンパク質ペアに頻繁に出現するならば、ドメイン$A$とドメイン$B$は相互作用するだろう
- ドメイン$A$とドメイン$B$が相互作用する確率を以下によって推定する
  - ドメイン$A$とドメイン$B$ が現れたときに、これが相互作用する確率の推定値




$p(A \Leftrightarrow B) = 1$
$N_{A,B} = 2 \quad I_{A,B} = 2$
$p(C \Leftrightarrow C) = 1$
$N_{C,C} = 6 \quad I_{A,B} = 4$

ドメイン$A$とドメイン$B$が相互作用するタンパク質ペア中に出現する数

$$p(A \Leftrightarrow B) = \frac{I_{A,B}}{N_{A,B}}$$

ドメイン$A$とドメイン$B$が、全タンパク質ペア中に出現する数

- 本当は、分子が「ドメイン$A$とドメイン$B$が、相互作用するタンパク質ペア中で実際に相互作用している数」であれば正しい値が推定できるが、わからないので、これで代用する

Sprinzak et al.: *Correlated sequence-signatures as markers of protein-protein interactions*, J. Mol. Biol, 311,2001

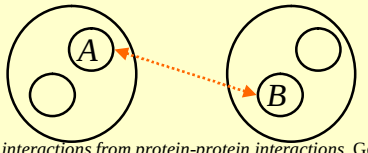
## ドメイン相互作用の特徴を考えた、より精緻な推定法 1：「ドメインがどれかひとつでも相互作用すれば、タンパク質相互作用する」

- 相互作用をしているタンパク質ペアのなかで、実際にドメインペア(A,B) が、相互作用A⇔Bしている回数がわかれば、正しい確率が推定
- 仮説： 2つのタンパク質のドメインペアのうち、どれかひとつでも作用すれば、タンパク質が相互作用するはず
- これを確率で書くと、タンパク質ペア $(i,j)$ の相互作用 $i \Leftrightarrow j$ が起こる確率は

$$p(i \Leftrightarrow j) = 1 - \prod_{A \in D_i} \prod_{B \in D_j} (1 - p(A \Leftrightarrow B))$$

ドメインAとBが相互作用しない確率

たんぱく質 $i$ と $j$ の含むドメインがひとつも相互作用しない確率




Deng et al.: *Inferring domain-domain interactions from protein-protein interactions*, Genome Research, 12



## ドメイン相互作用の特徴を考えた、より精緻な推定法 1：確率はEMアルゴリズムによって推定することができます

- EMアルゴリズムによって、ドメイン相互作用の確率 $p(A \Leftrightarrow B)$を逐次改善する

1. 【初期化】$p(A \Leftrightarrow B)$ の初期推定値を適当に求める（アソシエーション法などで）
2. 【ドメイン相互作用の推定】相互作用をしているタンパク質ペアのなかで、実際にドメインペア $(A,B)$ が、相互作用$A \Leftrightarrow B$している回数を推定する
   - ドメインペア $(A,B)$ を含む、実際に相互作用しているタンパクペア $(i, j)$ それぞれについて条件付確率 $p(A \Leftrightarrow B \mid i \Leftrightarrow j)$ を求める
     - ベイズの定理 $p(A \Leftrightarrow B \mid i \Leftrightarrow j) = p(i \Leftrightarrow j \mid A \Leftrightarrow B)\, p(A \Leftrightarrow B) / p(i \Leftrightarrow j)$
       $= p(A \Leftrightarrow B) / p(i \Leftrightarrow j)$ ← 現在の推定値を代入
3. 【ドメイン相互作用確率の推定】AとBを含むペア$(i, j)$について、
   $p(A \Leftrightarrow B) = \sum_{(i,j)} p(A \Leftrightarrow B \mid i \Leftrightarrow j) / N_{A,B}$
   によってあたらしい $p(A \Leftrightarrow B)$ の推定値を求める
- 2と3を繰り返す

$$p(i \Leftrightarrow j) = 1 - \prod_{A \in D_i} \prod_{B \in D_j} (1 - p(A \Leftrightarrow B))$$

Deng et al.: *Inferring domain-domain interactions from protein-protein interactions*, Genome Research, 12

## ちなみに、前述の手法は、線形計画法によって書くこともできます

- EMアルゴリズムでやったことを、制約として書き直す
  - 相互作用のあるタンパク質ペアには、

$$p(i \Leftrightarrow j) = 1 - \prod_{A \in D_i} \prod_{B \in D_j} (1 - p(A \Leftrightarrow B)) \geq \theta$$

  - ないペアには、

$$p(i \Leftrightarrow j) = 1 - \prod_{A \in D_i} \prod_{B \in D_j} (1 - p(A \Leftrightarrow B)) < \theta$$

ある確率θより大きければ相互作用あり

- この制約をなるべく満たす（制約を破る量をなるべく減らす）ように、θを決める
- 制約は、線形制約として書き直せるので、線形計画問題になる
  - 制約の対数をとって、変数をおきなおすと、

$$\sum_{A \in D_i} \sum_{B \in D_j} \log(1 - p(A \Leftrightarrow B)) \leq \log(1 - \theta)$$

$$\Longrightarrow \sum_{A \in D_i} \sum_{B \in D_j} \lambda(A \Leftrightarrow B)) \leq \beta \quad (\leftarrow 線形)$$

Hayashida et al.: *Inferring strengths of protein-protein interactions from experimental data using linear programming, Bioinformatics, 19, 2003.*



## ドメイン相互作用自体も、重要な生物学的知識です：ドメイン相互作用の重要度を「なかったらどのくらい困るか」で測ります

- タンパク質相互作用よりも、ドメイン相互作用のほうを知りたいことがある
  - ドメイン相互作用は完全に知られているわけではないので、コレ自体も重要な生物学的知識となる
- 各ドメイン相互作用の重要度を検証する方法を考えたい
  - ドメイン*A*と*B*が相互作用する確率 $p(A,B)=0.8$ だったとして、これの相互作用はどのくらい重要だろうか？
- いま検証したいドメインペアの作用が存在しないとしたら「どのくらい困るか」の度合いをもって、ドメインペアの作用の信頼度とする
  - 前頁で紹介したEMアルゴリズムを適用し、相互作用のあるタンパク質ペアについてその尤度 $p(i \Leftrightarrow j)$ の積 $\prod p(i \Leftrightarrow j)$ を計算する
  - 同じことを、 $p(A,B)=0$ （AとBは絶対に作用しないとして）として計算しなおす
  - $\prod p(i \Leftrightarrow j)$ の変化を観て、
    - 大きく減るようなら、$A \Leftrightarrow B$ の作用は重要と判断する
    - あまり減らないなら、あまり重要でない

Riley et al.: *Inferring protein domain interactions from databases of interacting proteins*, Genome Biology, 6(R89), 2005

## ドメイン相互作用の特徴を考えた、より精緻な推定法 2：「節約」仮説 ～ 生物はなるべく少ないドメインでネットワークを作っている

- 仮説： 生物は、なるべく少ない数のドメイン間相互作用で、ネットワークをつくっているはず
- 「節約しているはず！」の心を最適化問題に表すと…
  「$\sum_A \sum_B p(A,B)$ を最小化せよ」
- ただし、タンパク質の相互作用がちゃんと説明できなくてはならないので
  「相互作用のあるタンパク質ペア $(i, j)$ について
  $\sum_{A \in D_i} \sum_{B \in D_j} p(A,B) \geq \gamma$ をみたす」という制約が入る
  - $\gamma$：適当な0より大きい定数
  - $D_i$：タンパク質 $i$ にあるドメイン集合
  - $D_j$：タンパク質 $j$ にあるドメイン集合
- この問題は線形計画問題になり、高速に解ける
  - ちなみに前頁の線形計画との関連から $1-\exp(-p(A, B))$ は確率としても解釈できる

Guimaraes et al.: *Predicting domain-domain interactions using a parsimony approach*, Genome Biology, 7(R104), 2006

## ネットワークの構造情報を活用することで、これまでとは異なる大局的な情報に基づいた予測が行えるかもしれません

- これまでは、個々のタンパク質の情報（いわゆるローカルな情報）に基づいた予測を行ってきた
  - 2つのノードペアについて、その間にリンクがあるかどうかのみ考慮してきた
  - 少し引いて眺めたときの、ネットワークとしての特徴や整合性などは考慮していない
    - 全体的に疎で、ところどころに密な領域がある、など
- 大きな視点で、ネットワークの構造情報を用いた予測も考えられるはず
  - ネットワークのスケールフリー性を利用する
  - ネットワークの局所的な構造を利用する



- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ

## ネットワークのスケールフリー性を利用して、べき乗則に従うようなネットワークを予測しようとするアプローチがあります

- 現実世界のネットワーク構造は、しばしばスケールフリーネットワークになっている
  - ＝ 次数 $k$ の分布が、裾野の厚い、ベキ分布に従う

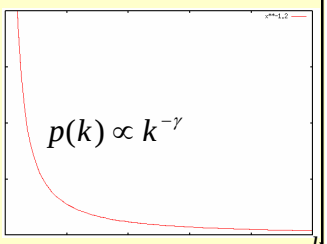
$$p(k) \propto k^{-\gamma}$$ （$\gamma$ は定数；タンパク質の場合2.5くらいらしい）

- ネットワーク構造に「スケールフリーネットワークになるような力」を入れて予測を行うことで、より現実のネットワーク構造に近いものが構成できるはず
- その気持ちを表現したネットワーク構造全体の確率↓を最大にするネットワークをみつける

ドメインなどに基づく確率

$$\prod_{(i,j)} p(i,j) \times \frac{N}{m_0! m_1! \cdot m_N!} \prod_k p(k)^{m_k}$$

べき乗則を目指す確率




- ドメイン情報などに基づくタンパク質間相互作用の確率 $\prod p(i \Leftrightarrow j)$ の部分と、
- $m_k$を次数$k$をもつノードの個数とすると、ネットワークのトポロジーの実現確率は $\prod_k p(k)^{m_k}$ に場合の数を掛けたもの
  - $p(k)$自体は、予め推定しておいたものをつかう

Gomez et al.: *Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks*, Genetics 159, 2001



- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ

## ネットワークの局所的な構造に基づいて、ネットワーク構造を予測する方法もあります

- 社会ネットワーク研究やスケールフリーネットワーク研究においていくつかの指標が提案されている
- いくつかの指標が、リンクの予測能力があることが知られている
  - 共通の隣接ノードが多いほど、リンクが張られやすいとするモデル 「友達の友達は友達」

$$\text{common neighbors} := |\Gamma(i) \cap \Gamma(j)|$$

Γ(i) はノード $i$ の隣接ノード集合

  - common neighbors の重み付きバージョン 「友達が少ない人ほど付き合いは深い」

$$\text{Adamic/Adar} := \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(k)|} \qquad \text{Jaccard's coefficient} := \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}$$

（こちらは情報検索で使われる指標）

  - 遠距離の影響もとりいれたcommon neighbors

$$\text{Katz}_\beta := \sum_{l=1}^{\infty} \beta^l |\text{paths}_{i,j}^{\langle l \rangle}|$$

$\text{path}_{i,j}^{\langle l \rangle}$はノード $i$ から $j$ への長さ $l$ のパスの集合

  - preferential attachment モデル （友人が多いほど、より多くの友人を得る）

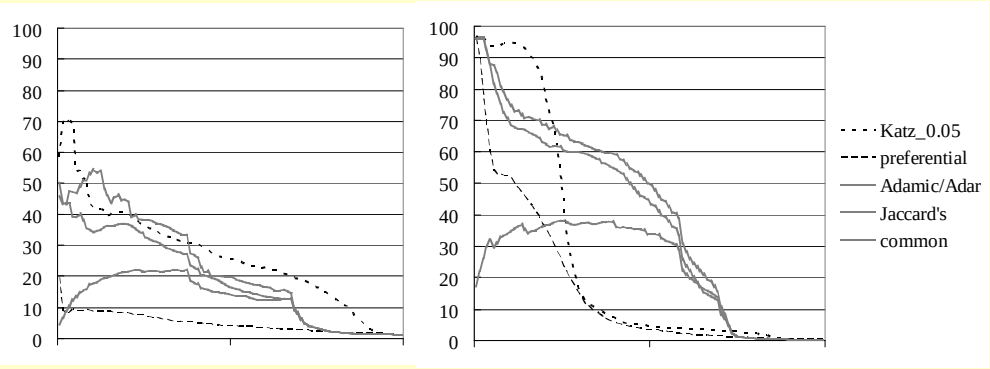
$$\text{preferential attachment} := |\Gamma(i)| \cdot |\Gamma(j)|$$

Liben-Nowell & Kleinberg: *The Link Prediction Problem for Social Networks*, CIKM 2004



## リンク指標にはリンクの予測能力があることが示されています

- Katz指標は比較的良好な予測性能をもつ
- このまま使ってもよいし、ノードペアの特徴ベクトルに加えてもよい




**代謝ネットワーク**
#nodes：700 (KEGG/S. Cerevisiae)

**タンパク質相互作用ネットワーク**
#nodes：3000（von Mering)

## 構造情報の利用方法のひとつとして、リンク指標をターゲットとして学習するアプローチがあります

- 予測したいタンパク質ペアの周りに、これまで観測された相互作用がない場合がある
  - 実験していないタンパク質に対して、予測を行いたいことが多いだろう
- このような場合、学習時には構造情報があるが、予測時にはノード情報（タンパク質自身の情報）しかない
  - が、せめて、学習時に構造情報を有効活用して、ノード情報を補いたい
- 解決法： 予測する値を、枝の有無ではなく、リンク指標の値にすることで、ノード情報から構造情報を考慮した予測をするように学習する

$$y(i,j) = \sum_{(k,\ell)} \alpha(k,\ell)\ K((i,j),(k,\ell))$$

リンク指標を予測するように学習

ノード情報に基づくカーネル予測器

Katz指標（厳密にはdiffusion kernel）を予測するように学習

Yamanishi et al.: *Kernel Matrix Regression*, Technical Report HAL-00133355, 2007
Yamanishi et al.: *Protein Network Inference from Multiple Genomic Data: A Supervised Approach*, ISMB 2004



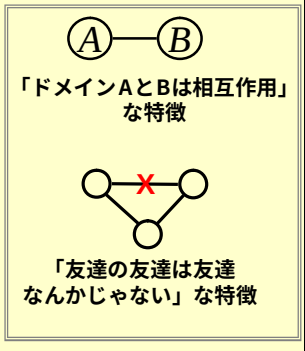
## 関係マルコフネットワークは、局所構造のパターンを限定しない、一般的なモデルですが、計算は大変です

- ネットワーク構造 $G$ の特徴ベクトル $\Phi(G)$ を定義する
  - G = {X, Y} は与えられている部分 $X$ と、予測したい部分 $Y$ からなる
    - $X$：各たんぱく質の特徴ベクトルと、わかっている相互作用
    - $Y$: わかっていない相互作用
  - 各特徴（$\Phi(G)$の各要素）は、テンプレート（構成部品）の出現回数として定義される
    - テンプレートの集合は予め定義しておく
- 関係マルコフネットワークは、テンプレートを使って、ネットワーク全体の出現確率を定義するモデル

$$P(Y|X) = \frac{\exp\left(\langle \mathbf{w}, \Phi(G)\rangle\right)}{Z(X)}$$

$$Z(X) = \sum_{Y} \exp\left(\langle \mathbf{w}, \Phi(G)\rangle\right)$$

w: 各テンプレの重要度を表すパラメータベクトル




テンプレート（構成部品）の例

- 計算はちょっと大変、サンプリングなどによる近似が必要

Jaimovich et al.: *Towards an Integrated Protein-protein Interaction Network*, J. Comp. Bio, 13(2), 2006

- ネットワーク予測問題
  - ネットワーク予測問題の定義
  - ネットワーク予測に用いる情報
- アプローチ大別
  - リンク情報が無い場合のアプローチ
  - リンク情報がある場合のアプローチ
    - 教師つき学習の考え方
    - カーネル法
    - ドメイン情報に基づく方法
  - 大局的な情報の利用
    - ネットワークのスケールフリー性を利用する方法
    - 局所構造情報を用いる方法
- まとめ



## まとめ

- 生物におけるネットワーク（特にタンパク質の相互作用ネットワーク）予測の問題は、配列データ（配列そのものや、ドメイン）や発現データ、立体構造情報そのほかのアノテーション情報などをもとに、2つのノードの間のリンクを予測する問題として捉えることができます
- そのアプローチは、既知のリンク情報があるかどうかによって大きく2つ分けることができます
- リンク情報が利用できる場合には、教師付き学習の問題になり、ノードペア同士の類似度をいかに設計するかというところに帰着されます
- 特に、ドメインに基づいて予測する手法では、ドメインに関するいくつかの仮説をもとに、より特化した方法が提案されています
- さらに、ネットワークの大局的な性質を利用して、予測の精度を高めようという試みもあります