G-002

相互カーネル行列補完によるタンパク質機能予測
Protein Function Prediction by Mutual Kernel Matrix Completion

リベロ　レイシェル † 下山 愛祐美 † 加藤 毅 † § ¶
Rachelle Rivero Ayumi Shimoyama Tsuyoshi Kato

1. まえがき

2000 年以降測定技術の急速な発達により，複数の種類のゲノムワイドのデータが得られるようになった．個々のデータタイプは細胞内の機構を一つの側面から見た観測値となっている．これらを統合して，細胞内でタンパク質の相互作用や細胞内の生物学的な機構の解明に用いられるようになった [1].

カーネル行列を介した異種データの統合がしばしば用いられている [1]．データをカーネル行列で表現しておくと，カーネル法と呼ばれる一連の方法論が利用可能になる．データの統合は単純で，各々のデータタイプから得られたカーネル行列を単純に平均をとるだけで統合することができる．この統合法の正当性は，カーネルの半正定値性を仮定してカーネル法が構成されており，半正定値カーネルは凸錘上にあることからいえる．

分子生物学において複数のデータを統合するときに一部のタンパク質に対して，一部のデータタイプが欠けているような状況がある．このような状況においてカーネルの線形結合によるデータ統合を行うには，2 つのアプローチがある．その一つは，そのタンパク質を解析対象から除外するか，そのデータタイプを除外する方法である．すると，解析対象が少なくなる，もしくは，有用なデータタイプを一部のデータの欠落だけのために放棄することになってしまう．もう一つのアプローチは，欠落しているカーネルの値を推定する方法である [2, 3].

従来のカーネル行列補完の方法は，1 個のデータタイプだけ不完全で，そのほかのデータタイプはすべてのタンパク質に対して数値が観測できていることを仮定していた [2, 3] (図 1(a),(b) 参照)．しかし，現実的には，複数のデータタイプが不完全である場合 (図 1(c) 参照) がよくあり，従来法の想定とはギャップがある．本研究では，図 1(c) のように所与のカーネル行列各々が不完全でも相互に補完できる新しい方法論を開発した．

2. 問題設定

本研究で議論する**相互カーネル行列補完** *(MKMC)* というタスクを述べる．K 個の不完全カーネル行列 $\boldsymbol{Q}^{(k)}$ $(k = 1, \ldots, K)$ が 図 1(c) のように与えられていたとする．$\boldsymbol{Q}^{(k)}$ の行列のサイズはいずれも $\ell \times \ell$ である．K 個の情報源のいずれかの対象に関する情報が欠損しているとする．すると，それぞれの情報源のカーネル行列 $\boldsymbol{Q}^{(k)}$ における対応する行と列は欠損していることになる．$\boldsymbol{Q}^{(k)}_{vh,vh}$ は，$\boldsymbol{Q}^{(k)}$ の行と列の順番を次のように並び替えたものとする．最初の $n_k < \ell$ 個の対象に関

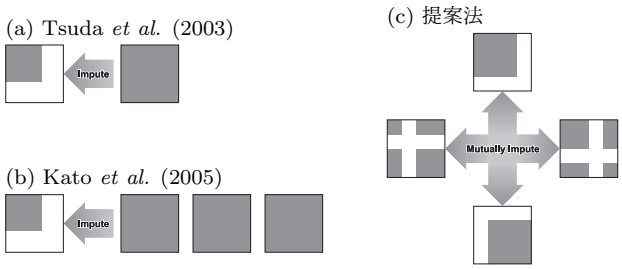

図 1: 問題設定．

する情報が得られているとし，残りの $(\ell - n_k)$ 個の対象に関する情報が欠損しているとする．このようにして $\boldsymbol{Q}^{(k)}$ の行と列の順番を並び替えて得られる $\boldsymbol{Q}^{(k)}_{vh,vh}$ を次のように分割する：

$$\boldsymbol{Q}^{(k)}_{vh,vh} = \begin{bmatrix} \boldsymbol{Q}^{(k)}_{v,v} & \boldsymbol{Q}^{(k)}_{v,h} \\ \boldsymbol{Q}^{(k)}_{h,v} & \boldsymbol{Q}^{(k)}_{h,h} \end{bmatrix} \tag{1}$$

ただし，$\boldsymbol{Q}^{(k)}_{v,v} \in \mathbb{S}^{n_k}_{++}$, $\boldsymbol{Q}^{(k)}_{v,h} = \left(\boldsymbol{Q}^{(k)}_{h,v}\right)^{\top} \in \mathbb{R}^{n_k \times m_k}$, および $\boldsymbol{Q}^{(k)}_{h,h} \in \mathbb{S}^{m_k}_{++}$ である．開発した算法では欠損した部分行列 $\boldsymbol{Q}^{(k)}_{v,h}$, および $\boldsymbol{Q}^{(k)}_{h,h}$ の値を推定する．$\mathcal{H} := \left\{\boldsymbol{Q}^{(k)}_{v,h}, \boldsymbol{Q}^{(k)}_{h,h}\right\}_{k=1}^{K}$ とおく．

3. 提案法：相互カーネル行列補完法

提案法では，モデル行列 $\boldsymbol{M} \in \mathbb{S}^{\ell}_{++}$ を導入し，各 $\boldsymbol{Q}^{(k)}$ が $\boldsymbol{M}$ との距離が最小になる $\mathcal{H}$ を求めるものである．すなわち，目的関数を各 $\boldsymbol{Q}^{(k)}$ と $\boldsymbol{M}$ との**カルバックライブラ (KL) 距離**の和

$$J(\mathcal{H}, \boldsymbol{M}) := \sum_{k=1}^{K} \mathrm{KL}(\boldsymbol{Q}^{(k)}, \boldsymbol{M}) \tag{2}$$

で定義し，この目的関数を最小にする $\mathcal{H}$ と $\boldsymbol{M}$ の組み合わせを見つける．KL 距離は，本来，確率分布間の擬距離であるが，本研究では，文献 [2, 3] に倣って，カーネル行列と正規分布とを関連付け，正規分布間の KL 距離

$$\mathrm{KL}(\boldsymbol{Q}^{(k)}, \boldsymbol{M}) \;:=\; \mathrm{KL}(\mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{(k)}) || \mathcal{N}(\boldsymbol{0}, \boldsymbol{M})) \tag{3}$$

をカーネル行列間の KL 距離とした．この定義を用いると，最小解が正定値になることが保証されるので，カーネル行列たるために必要な正定値性に関する追加的制約が不要になる．しかし，$(\mathcal{H}, \boldsymbol{M})$ を同時最小化する解は閉形式では与えられないので，適当な正定値行列

†群馬大学大学院理工学府
§群馬大学次世代モビリティ社会実装センター (CRANTS)
¶早稲田大学規範科学総合研究所 (IIRS)

$\boldsymbol{M} \in \mathbb{S}^{\ell}_{++}$ を初期値とし，E-step と M-step を繰り返す次の算法を用いることにする．

1: **repeat**
2: {E-step} $\mathcal{H} := \operatorname{argmin}_{\mathcal{H}} J(\mathcal{H}, \boldsymbol{M})$;
3: {M-step} $\boldsymbol{M} := \operatorname{argmin}_{\boldsymbol{M}} J(\mathcal{H}, \boldsymbol{M})$;
4: **until** convergence.

この算法の性質から，目的関数は単調非増加することは明らかである．加えて，次のような望ましい性質を持つ：

Theorem 1. 上述の算法における E-step および M-step は閉形式で表される．

各ステップを，それぞれ，E-step および M-step と表記した所以は，それぞれ統計学でいう**期待値最大化** (EM) 法の E-step および M-step で説明できるからである．その理由を次節で述べる．また，紙面の制約から具体的な更新式は次節にのみ記述した．

4.EM 法との関係

EM 法再訪． EM 法は，観測値 $\boldsymbol{V}$ および潜在変数 $\boldsymbol{H}$ の同時確率のモデル $p(\boldsymbol{V}, \boldsymbol{H} \,|\, \boldsymbol{\Theta})$ のパラメータ $\boldsymbol{\Theta}$ を最尤推定するための枠組みである．すなわち，**経験分布** $q(\boldsymbol{V})$ に対して，EM 法は周辺分布 $p(\boldsymbol{V} \,|\, \boldsymbol{\Theta})$ の対数尤度関数

$$L(\boldsymbol{\Theta}) := \mathbb{E}_{q(\boldsymbol{V})} \left[\log p(\boldsymbol{V} \,|\, \boldsymbol{\Theta}) \right] \tag{4}$$

を最大化する $\boldsymbol{\Theta}$ を見つけようとする．いま，同時確率モデルの対数密度が

$$\log p(\boldsymbol{V}, \boldsymbol{H} \,|\, \boldsymbol{\Theta}) = \langle \boldsymbol{S}(\boldsymbol{V}, \boldsymbol{H}), \boldsymbol{G}(\boldsymbol{\Theta}) \rangle - A(\boldsymbol{\Theta}) \tag{5}$$

のように指数分布族で表されているとする．ただし，$\boldsymbol{S}(\boldsymbol{V}, \boldsymbol{H})$, $\boldsymbol{G}(\boldsymbol{\Theta})$，および $A(\boldsymbol{\Theta})$ は**十分統計量**，**自然パラメータ**，**キュムラント関数**である．すると，EM 法における Q 関数は

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{\mathrm{old}}) := \langle \mathbb{E}_{\boldsymbol{\Theta}^{\mathrm{old}}} [\boldsymbol{S}(\boldsymbol{V}, \boldsymbol{H})], \boldsymbol{G}(\boldsymbol{\Theta}) \rangle - A(\boldsymbol{\Theta}) \tag{6}$$

と表される．E-step で期待値 $\mathbb{E}_{\boldsymbol{\Theta}^{\mathrm{old}}}[\boldsymbol{S}(\boldsymbol{V}, \boldsymbol{H})]$ を計算し，M-step では期待値を固定して，Q 関数を最大化するように $\boldsymbol{\Theta}$ を更新する．

提案法は EM 法． 次のようにモデル分布と経験分布を定めると，提案法は EM 法であることがわかる．今，完全データを $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K \in \mathbb{R}^{\ell}$ とする．$\boldsymbol{x}_k$ の各要素は $\boldsymbol{Q}^{(k)}$ のある行と列に関連付ける．$\boldsymbol{v}_k \in \mathbb{R}^{n_k}$ および $\boldsymbol{h}_k \in \mathbb{R}^{\ell - n_k}$ を $\boldsymbol{x}_k$ の部分ベクトルとし，$\boldsymbol{v}_k$ は $\boldsymbol{Q}^{(k)}$ の可視データに $\boldsymbol{h}_k$ は $\boldsymbol{Q}^{(k)}$ の欠損データに対応する．EM 法における観測データを $\boldsymbol{V} := (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)$ で構成し，潜在データを $\boldsymbol{H} := (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K)$ で構成する．経験分布 $q(\boldsymbol{V})$ は $\mathbb{E}_{q(\boldsymbol{V})}[\boldsymbol{v}_k \boldsymbol{v}_k^{\top}] = \boldsymbol{Q}^{(k)}_{v,v}$ を満たすものとし，モデル分布は $\boldsymbol{\Theta} = \boldsymbol{M}$ とし，完全データの密度は $p(\boldsymbol{V}, \boldsymbol{H} \,|\, \boldsymbol{\Theta}) := \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_k \,|\, \boldsymbol{0}, \boldsymbol{M})$ で与えるとする．このモデルは，

$$\boldsymbol{S}(\boldsymbol{V}, \boldsymbol{H}) := \sum_{k=1}^{K} \boldsymbol{x}_k \boldsymbol{x}_k^{\top}, \text{ および } \boldsymbol{G}(\boldsymbol{M}) = -\frac{1}{2} \boldsymbol{M}^{-1}$$

とおくことにより，指数分布族であることが示される．この設定を EM 法にあてはめる．$\boldsymbol{M}^{(k)}_{h|v} := \boldsymbol{M}^{(k)}_{h,h} - \boldsymbol{M}^{(k)}_{h,v} (\boldsymbol{M}^{(k)}_{v,v})^{-1} \boldsymbol{M}^{(k)}_{v,h}$ とおくと，

$$\mathbb{E}\left[\boldsymbol{v}_k \boldsymbol{h}_k^{\top}\right] = \boldsymbol{Q}^{(k)}_{v,v} (\boldsymbol{M}^{(k)}_{v,v})^{-1} \boldsymbol{M}^{(k)}_{v,h}, \text{ および}$$

$$\mathbb{E}\left[\boldsymbol{h}_k \boldsymbol{h}_k^{\top}\right] = \boldsymbol{M}^{(k)}_{h|v} + \boldsymbol{M}^{(k)}_{h,v} (\boldsymbol{M}^{(k)}_{v,v})^{-1} \boldsymbol{Q}^{(k)}_{v,v} (\boldsymbol{M}^{(k)}_{v,v})^{-1} \boldsymbol{M}^{(k)}_{v,h},$$

が成立し，それぞれ $\boldsymbol{M}$ を固定して $\mathcal{H}$ を最適化したときの $\boldsymbol{Q}^{(k)}_{v,h}$ および $\boldsymbol{Q}^{(k)}_{h,h}$ に等しい．また，Q 関数を最大化する $\boldsymbol{M}$ は $\boldsymbol{Q}^{(k)}$ の平均で与えられるので，$\operatorname{argmin}_{\boldsymbol{M}} J(\mathcal{H}, \boldsymbol{M})$ に等しくなる．以上の導出により，次の理論的結果を得た：

Theorem 2. 前節で述べた算法は EM 法の一つである．

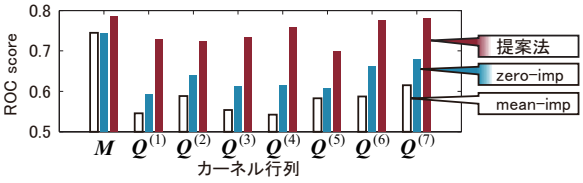

図 2: タンパク質機能予測の性能．

5. 実験結果

文献 [1] で用いられていた 7 個のカーネル行列を使って性能の評価を行った．0 で埋める方法 (zero-imp) および平均で埋める方法 (mean-imp) と提案法を比較した．各データタイプで 50% を欠損させた．無作為に選んだ 200 個を SVM の訓練に用い，評価用タンパク質が膜タンパクか予測した．各カーネル行列における予測性能を ROC スコアで評価したところ，図 2 を得た．いずれのカーネル行列でも，0 で埋めたり平均で埋めるような単純な補完法より提案法で補完したほうが顕著に高い汎化性能を得ることができた．

謝辞:本研究は JSPS 科研費 26249075, 40401236 の助成を受けたものである．

参考文献

[1] G. R. Lanckriet et al. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–35, Nov 2004.

[2] Koji Tsuda et al. The em algorithm for kernel matrix completion with auxiliary data. *JMLR*, 4:67–81, 2003.

[3] Tsuyoshi Kato, Koji Tsuda, and Kiyoshi Asai. Selective integration of multiple biological data for supervised network inference. *Bioinformatics*, 21:2488–2495, 2005.