Vol.2012-BIO-30 No.2
2012/8/9

タンパク質複合体サイズ分布を用いたマルコフ連鎖モンテカルロ法に基づく複合体予測手法の研究

田附 大典[1,a)]　丸山 修[2,b)]

概要：本研究では，タンパク質間相互作用情報からタンパク質複合体を予測するサンプリング手法を提案する．既存手法の多くはタンパク質間相互作用ネットワークの部分グラフの密度に基づき複合体を予測するので，小さな複合体の正確な予測は相対的に困難である．ところが，酵母の代表的なタンパク質複合体データベースである CYC2008 を調べると，複合体のサイズ分布はスケール・フリーであり，42%の複合体は最小サイズ 2 であることが分かる．そこで，本研究では，複合体のサイズ分布情報を活用したメトロポリス-ヘイスティングス法に基づく予測手法 PPSampler (Proteins' Partition Sampler) を提案する．この PPSampler が，既存手法と比べて高い精度を実現することを計算機実験により確認した．

キーワード：マルコフ連鎖モンテカルロ法，MCMC, サンプリング，タンパク質複合体，タンパク質間相互作用，スケール・フリー，PPSampler

MCMC Strategy for Protein Complex Prediction Using Cluster Size Frequency

Tatsuke Daisuke[1,a)]　Maruyama Osamu[2,b)]

Abstract: In this paper we propose a Markov chain Monte Carlo sampling method for predicting protein complexes from protein-protein interactions (PPIs). Many of existing tools are directly or indirectly designed based on a density measure of a subgraph of the PPI network. This kind of measure is less effective for smaller complexes. However, we can see that the distribution of complex sizes of CYC2008, a database of curated protein complexes of yeast, is scale-free, and that 42% of the complexes are heterodimeric, i.e., of size 2. In this work, we propose PPSampler, which is a protein complex prediction algorithm designed based on the Metropolis-Hastings algorithm using a parameter representing a target value of the relative frequency of predicted protein complexes of a particular size. In performance comparison, we found that PPSampler outperforms other existing algorithms. Furthermore, about 65% of the predicted clusters that are not matched with any known complexes in CYC2008 have more than 90% coverages by cellular component terms. Some of them are expected to be true complexes.

Keywords: Markov chain Monte Carlo, MCMC, sampling, protein complex, protein-protein interaction, scale-free, PPSampler

1. はじめに

タンパク質間相互作用 (protein-protein interaction) からのタンパク質複合体の予測は計算生物学分野の重要な問題である [1]．この問題の解決のために，MCL [2]，MCODE [3], RNSC [4], CFinder [5], DPClus [6], COACH [7], RRW [8]，NWE [9] などの様々な予測手法が提案されている．この予測問題の既存手法の多くは直接的もしくは間接的にタンパク質間相互作用ネットワークにおける密な部分グラフに基づき複合体を予測する．しかしながら，部分グラフが小さくなればなる程，密度は粗い指標となり，複合体判定

[1] 九州大学大学院数理学府
[2] 九州大学マス・フォア・インダストリ研究所
[a)] ma211030@math.kyushu-u.ac.jp
[b)] om@imi.kyushu-u.ac.jp

は理論的に困難となる．

一方，既知のタンパク質複合体データベースにはサイズの小さな複合体が数多く存在する．例えば，*S. cerevisiae* のタンパク質複合体データベースである CYC2008 [10] は 408 個の複合体を有するが，そのうちの 42%の 172 個が二量体（サイズ 2）である．実際，最頻出のサイズは 2 となっている．従って，サイズ 2 の複合体の予測に重きをおいた予測手法は，予測精度の向上が期待できる．タンパク質二量体の予測手法に関しては，丸山 [11] による教師付き学習手法による予測手法があるが，この手法の対象は二量体のみに限定されている．

さらに，CYC2008 のサイズ分布を調べると，その分布はスケール・フリー性を有することが分かる．これは，サイズ k の複合体の相対頻度が $k^{-\gamma}$(γ は定数) に比例することを意味する [12]．そこで，本研究ではこの事実を事前知識として活用するメトロポリス-ヘイスティングス法に基づく予測手法 PPSampler (Proteins' Partition Sampler) を提案する．PPSampler は, 与えられた確率分布に従ってタンパク質のクラスター集合をサンプルとして生成する．その確率分布は，タンパク質のクラスター集合 C に対する３つの異なる評価関数から構成される．これらの評価関数は，それぞれ，C 内のタンパク質間相互作用の重みに基づくもの，C に属する予測されるクラスターのサイズ分布に基づくもの，そして C に含まれるタンパク質の総数に基づくものである．既存手法の中で評価の高い MCL [2] などと予測精度の比較を行った結果，PPSampler はより高い予測精度を有することが分かった．

2. PPSampler

本節では，我々の提案手法である PPSampler について説明する．まず，PPSampler の骨格であるメトロポリス-ヘイスティングス (Metropolis-Hastings; M-H) アルゴリズム [13] をどのように具体化するかを述べる．

2.1 M-H アルゴリズム

M-H アルゴリズムはある確率分布からランダムにサンプルを生成するためのマルコフ連鎖モンテカルロ (Markov chain Monte Carlo; MCMC) 法 [14] の一種である．M-H アルゴリズムを**図 1** に示している．M-H アルゴリズムは，次の３つの構成要素を決めることにより具体化される:

(i) 状態の集合 D

(ii) 状態 $C \in D$ から状態 $C' \in D$ の提案分布 $Q(C'|C)$

(iii) サンプルを生成する確率分布 $P(C)$

次に PPSampler で用いる M-H アルゴリズムの以上の 3 要素を定式化していく．

2.2 タンパク質間相互作用データ

タンパク質間相互作用データは，タンパク質複合体予測

Input:
温度パラメータ T;
反復回数 K;
初期状態 C_0;
提案分布 $Q(C'|C)$;
評価関数 $f(C)$;

Output:
サンプルされた状態 K 個の列;

Procedure:
$C = C_0$; /*初期状態の設定*/
for $k = 1$ to K:
$Q(C'|C)$ より候補状態 C' を提案;
$P(C) \propto \exp\left(-\frac{f(C)}{T}\right)$;
$r = \dfrac{P(C')\,Q(C|C')}{P(C)\,Q(C'|C)}$;
区間 $[0, 1]$ 上の一様乱数 R の生成;
if $r > R$ **then** $C = C'$;

図 1 Metropolis-Hastings アルゴリズム．
Fig. 1 Metropolis-Hastings algorithm.

において重要な入力データである．本稿では，このデータを次のように定式化する．V をある生物種のタンパク質の集合とし，タンパク質間相互作用の集合を

$$E \subseteq V \times V$$

で表す．各 $e \in E$ の重みを $w(e) \in \mathbb{R}_+$ で表す．ただし，$e = \{u, v\} \notin E$ に対しては，$w(e) = 0$ と仮定する．

2.3 状態

次に M-H アルゴリズムの状態について述べる．V の分割 (partition) を C とする．つまり，C は次のように書ける：

$$C = \left\{ c_1, \ldots, c_n \subseteq V \;\middle|\; \begin{array}{l} \forall i, c_i \neq \emptyset, \\ \cup_{1 \leq i \leq n} c_i = V, \\ \forall i, j(\neq i), c_i \cap c_j = \phi \end{array} \right\}.$$

C の要素をクラスターとも呼ぶ．以後，分割はすべて V の分割を意味することとする．個々の分割 C は M-H アルゴリズムにおける 1 つの状態に対応する．

2.4 提案分布

次に，分割 C から分割 C' を提案する提案確率 $Q(C'|C)$ を定義する．C' は，次の二通りの方法により C から派生する．まず，どちらの場合であっても，クラスター間を移動させるタンパク質として，V の中から一様分布に従いランダムに一つのタンパク質 u を選択する．つまり，特定のタンパク質 u が選択される確率は $\frac{1}{|V|}$ となる．次に，C' の二通りの作り方のそれぞれに対する確率 $Q(C'|C)$ を定める．ここで，次の (i) の u のみからなる新しい分割の要素を生成する場合を選択する確率を β とする．

(i) u のみからなる新しい分割の要素を生成する場合．
このときの提案確率は

$$Q(C'|C) = \frac{\beta}{|V|}$$

となる．

(ii) C からランダムに選択したクラスター c に u を移す場合．
u 以外の全タンパク質 $v \in V$ を $w(\{u, v\})$ に従い降順に並び替え，第 i 番目のタンパク質を v_i と記す．つまり，

$$w(\{u, v_1\}) \geq w(\{u, v_2\}) \geq \ldots w(\{u, v_{|V|-1}\})$$

となる．分割 C から c が選ばれる確率は

$$\sum_{v_i \in c} \frac{1}{i}$$

に比例すると定める．従って，提案分布 $Q(C'|C)$ は

$$Q(C'|C) \propto \frac{1-\beta}{|V|} \sum_{v_i \in c} \frac{1}{i}$$

となる．

確率パラメータ β の値は，本稿を通して $\beta = 1/100$ に固定している．

2.5 評価関数

次に M-H アルゴリズムで使用する評価関数 f を構成する C の評価関数 f_1, f_2, f_3 を定義する．これらは全て最大化関数である．

まず C に含まれるタンパク質間相互作用の重みに基づく評価関数 $f_1(C)$ を定義する．そのために，まず一つの要素 $c \in C$ に対する評価関数 $f_1(c)$ を次のように定義する：

$$f_1(c) = \begin{cases} 0 & \text{if } |c| = 1, \\ -\infty & \text{else if } |c| > N \text{ または} \\ & \quad \exists u \in c, \forall v(\neq u) \in c, \\ & \quad w(\{u, v\}) = 0, \\ \displaystyle\sum_{u, v(\neq u) \in c} w(u, v) & \text{otherwise.} \end{cases}$$

ただし N はクラスター c のサイズの上限値を与えるパラメータである．上記の $f_1(c)$ の定義における最後の場合は，クラスター c 内の全てのタンパク質ペアの相互作用の重みの総和を表している．次に $f_1(C)$ を次のように定義する：

$$f_1(C) = \sum_{c \in C} f_1(c).$$

次に分割 C のクラスターのサイズ分布に基づく評価関数 $f_2(C)$ を定義する．C に対して $|c| = i\,(= 2, 3, \cdots, N)$ となる $c \in C$ の数の全体に対する割合を $\psi_C(i)$ で表すことにする．各サイズ i のクラスター数の相対頻度の目標値をパラメータ $\psi(i)$ で表す．$\psi(i)$ の値と $\psi_C(i)$ の値の二乗誤差とサイズ i に対する誤差ペナルティ i^2 との積の逆数の積を $f_2(C)$ と定義する．つまり

$$f_2(C) = \prod_{i=2}^{N} \frac{1}{1 + i^2 \cdot (\psi(i) - \psi_C(i))^2}$$

となる．ただし，分母が 0 になることを避けるため分母に 1 を足している．

分割 C のサイズ 2 以上のクラスター c 内のタンパク質の総数を $s(C)$ で表す．つまり，$s(C) = \bigcup_{c \in C \text{ s.t. } |c| \geq 2} c$ と書ける．$s(C)$ をその目標値を表すパラメータ λ の値に近づけるため，第 3 の評価関数 $f_3(C)$ を

$$f_3(C) = \frac{1}{1 + \dfrac{(s(C) - \lambda)^2}{10^3}}$$

と定義する．f_2 と同様に，分母が 0 になることを避けるため 1 を足している．

以上の関数 f_1, f_2, f_3 を組み合わせて最終的な評価関数 f を

$$f(C) = -f_1(C) \cdot f_2(C) \cdot f_3(C)$$

と定義する．

2.6 初期状態

次に，図 1 の M-H アルゴリズムが用いる初期状態 C_0 を定める．初期状態 C_0 は，次の 2 種類のクラスターすべてから構成する：

- タンパク質間相互作用の重み $w(u,v)$ が最大である二つのタンパク質 u と v のみから成るクラスター．
- 残りの各タンパク質 $w \in V \setminus \{u,v\}$ のみからなるサイズ 1 のクラスター．

2.7 出力

PPSampler は，図 1 が生成する全てのサンプル C の中から確率 $P(C)$ が最大となる C を予測複合体の集合として出力する．ただし，C に含まれるサイズ 1 のクラスターは予測複合体に含めない．また，確率最大のサンプルを選ぶために実際にサンプル C の確率 $P(C)$ を計算する必要はなく，$P(C)$ の比例値である $\exp\left(-\frac{f(C)}{T}\right)$ を用いて個々の $P(C)$ の大小関係を判定すればよい．

2.8 手法の評価

予測されたクラスター集合の評価を 適合率 (precision), 再現率 (recall), F 値 (F-measure) の 3 つの尺度で行う．これらを定義するため，まず二つのクラスターの重複度 (overlap ratio) を定義する．

クラスター s と t の重複度 $ov(s,t)$ を，$|s|$ と $|t|$ の幾何平均に対する s と t の共有タンパク質数の割合を用いて次のように定義する：

$$ov(s,t) = \begin{cases} \frac{|s \cap t|}{\sqrt{|s| \cdot |t|}} & \text{if } |s \cap t| > 1, \\ 0 & \text{その他の場合.} \end{cases}$$

この重複度 $ov(s,t)$ は，もしサイズ 2 以上のクラスター s と t が完全に一致しているなら最大値の 1 となる．また，s と t により共有されているタンパク質が 1 個以下ならば 0 となる．共有されているタンパク質が 1 個の場合も重複度を 0 とする理由は以下のとおりである．文献における重複度の典型的な閾値の値は $0.4472(=\sqrt{0.2})$ である（例えば [3]）．しかしながら，この閾値の値 0.4472 ではサイズ 2 の s と t に対しては共通するタンパク質が 1 つしかない場合でも $\frac{|s \cap t|}{\sqrt{|s| \cdot |t|}} = 0.5 > 0.4472$ となりマッチしたと判定される．これは偶然に起こりうることであり重複しているとは認めがたい．従って，この不適切な状況を避けるために重複度 $ov(s,t)$ を以上のように定義した．クラスター s のクラスター集合 T に対する重複度 $ov(s,T)$ を，s の $t \in T$ に対する $ov(s,t)$ の最大値と定義する．すなわち

$$ov(s,T) = \max_{t \in T} ov(s,t)$$

と書ける．

C を予測されたタンパク質複合体の集合とし，K を既知のタンパク質複合体の集合とする．また，予め与えられた重複度の閾値を t とする．このとき，C の K に対する適合率を

$$precision_K(C) = \frac{|\{c \in C | ov(c,K) \geq t\}|}{|C|}$$

と定義し，再現率を

$$recall_K(C) = \frac{|\{k \in K | ov(k,C) \geq t\}|}{|K|}$$

と定義する．最後に，F 値を適合率と再現率の調和平均と定義する．つまり，

$$F_K(C) = 2 \cdot \frac{precision_K(C) \cdot recall_K(C)}{precision_K(C) + recall_K(C)}$$

となる．

3. 結果

本節は，様々な観点からの提案手法 PPSampler の性能評価について述べる．

3.1 予測精度比較

まず，PPSampler と既存手法の予測精度の比較を行う．予測精度を比較するアルゴリズムは，文献 [1], [15] 等の予測精度の比較実験において高い評価を得ているクラスタリング・アルゴリズム MCL [2] と，再スタート・ランダム・ウォーク (random walk with restarts) 手法に基づく二つの予測アルゴリズム RRW [8] と NWE [9]，そして PPI ネットワーク上のタンパク質の連結性に基づく手法である MCODE [3] である．以上のアルゴリズムに与えるタンパク質間相互作用データは，WI-PHI [16] の全ての相互作用とする．また，既知のタンパク質複合体として 408 個の CYC2008 [10] の全ての複合体を用いる．

PPSampler のパラメータに関しては，温度パラメータを $T = 5$ そして反復回数を $K = 10^8$ としている．最大クラスターサイズ N は，CYC2008 の最大複合体のサイズが 81 なので，近似的に $N = 100$ と設定している．タンパク質総数パラメータ λ はデフォルト値 $\lambda = 2000$ とし，後でその他の値の場合の予測精度を検証している．

パラメータ $\psi(i)$ は各サイズ $i(= 2, 3, \ldots, N)$ のクラスター数の相対頻度の目標値を表すパラメータである．CYC2008 に含まれる複合体のサイズ分布を調べると，その分布はスケール・フリー性を有していることが分かる．そこで，$2 \leq i \leq 100$ の範囲で相対頻度の二乗誤差の最小化によりべき乗に回帰させて正規化を行うと $1.62 \times i^{-2.02}$ となる．そこで本研究では，近似的に，$\psi(i)$ を i^{-2} に比例する形に設定する．つまり，

$$\psi(i) = \frac{i^{-2}}{\sum_{j=2}^{N} j^{-2}}$$

となる．現在の PPSampler では，$\psi(i) \propto i^{-\gamma}$ の形でパラメータ $\psi(i)$ を指定可能となっている．後で，γ の値は PPSampler の予測精度にさほど影響を与えないことを確認する．

RRW と NWE の最小クラスター・サイズ・パラメータを 2 に設定している．さらに NWE の overlap ratio のデフォルト値は 0.3 であるが，これを RRW と同じ 0.2 にしている．この二つのアルゴリズムのその他のパラメータ値は全てデフォルト値であり，さらに他のアルゴリズムのパラメータ値も全てデフォルト値である．

各アルゴリズムの予測結果を**表 1** に与えている．表 1 において「タンパク質数」の行は，サイズ 2 以上のクラスターに属するタンパク質の総数を示し，「クラスター数」の行は，サイズ 2 以上のクラスターの総数を表している．また，適合率，再現率，F 値の 3 つの尺度においてはそれぞれの最高値を太字で表している．

PPSampler のタンパク質数は目標値 $\lambda = 2000$ とほぼ同じ 2001 であることから，評価関数 f_3 がよく効いていることが分かる．第 2 行のクラスター数に関しては，アルゴリズムごとに様々な値を取っており，PPSampler のクラスター数は比較的少なめの 350 個である．適合率に関しては，PPSampler の 0.54 が他を凌駕しており，2 番目に高い NWE の 0.28 より約 93%優れている．再現率では，MCL の 0.60 が最高値であるが，PPSampler の 0.53, NWE の 0.52, RRW の 0.50 と，MCODE の 0.08 を除き，MCL の最高値と遜色ない値を得ている．特に，MCL の 0.60 と次に良い PPSampler の 0.53 の値の差は，適合率における PPSampler の 0.54 と NWE の 0.28 の値の差に比べると非常に小さい．再現率と適合率から計算される F 値においては，PPSampler が最も高い 0.54 を得ており，その次に良いのは NWE の 0.37 である．よって PPSampler のスコアは NWE よりも約 46%優れていることが分かる．以上より，PPSampler は適合率と再現率の双方においてバランスよく高い値を得ており，その結果，予測精度の総合的評価基準である F 値においても優れた値を得ている．

3.2 Gene Ontology による評価

Gene Ontology プロジェクト (GO) は，あらゆる生物種の遺伝子と遺伝子産物の属性を表す共通語彙を策定するプロジェクトである [17]．予測されたクラスター c 内の多くのタンパク質によって共有された GO term は c を特徴付ける有用な情報と考えられる．そこで，予測された各クラスターが GO term によりどれ程うまく特徴付けられているかを知るために次の被覆率を定義する．ここで用いる GO term は，その全体集合の中から代表的なものを集めた部分集合である GO-slim [18], [19] を用いる．タンパク質のクラスター c の GO-slim term t に対する被覆率 (coverage) を，$|c|$ に対する t により注釈付けられた c 内のタンパク質の個数の割合と定義する．さらに，c の各 GO-slim term に対する被覆率の最大値を c の GO-slim 全体に対する被覆率と定義する．

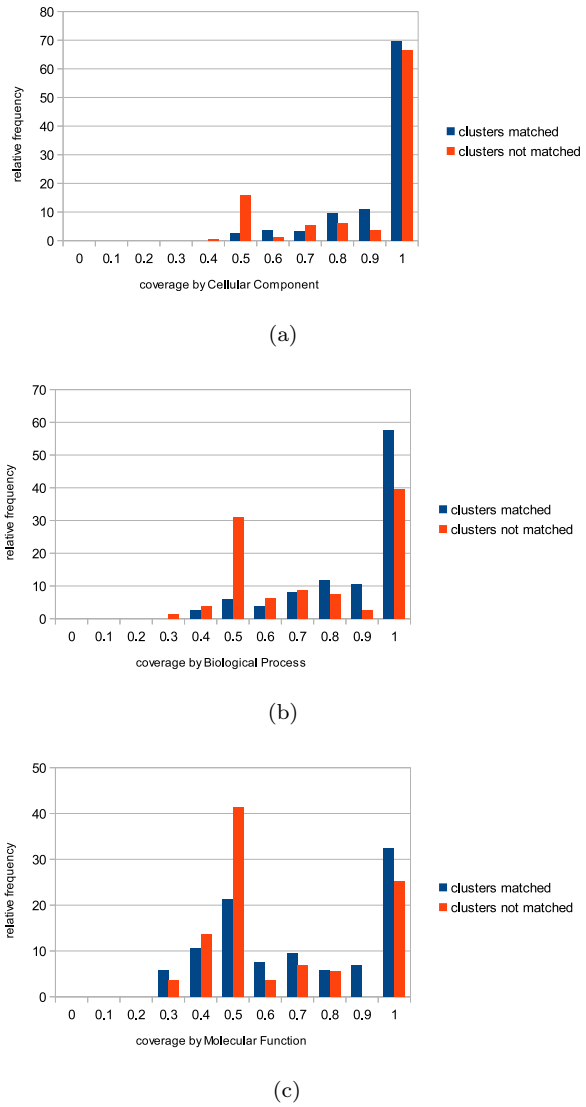

図 2 PPSampler により予測された各クラスターの GO-slim term による被覆率．

Fig. 2 The relative frequency distributions of coverages by GO-slim term.

各オントロジーごとの被覆率の相対頻度を**図 2** に示している．まず，cellular component オントロジーに関する被覆率が図 2 (a) に示されている．0.1 刻みの各ビンごとに，既知の複合体とマッチした予測クラスタに関する被覆率の相対頻度（左側の青色）とどの既知の複合体ともマッチしてない予測クラスタに関する被覆率の相対頻度（右側の橙色）を表している．

既知の複合体とマッチした予測クラスタの場合，被覆率の区間 (0:9; 1:0] のみでピークを持っている．このビンは

表 1 予測精度の比較.
Table 1 Performance comparison.

	MCL	MCODE	RRW	NWE	PPSampler
タンパク質数	5869	2432	4240	1626	2001
クラスター数	880	156	1984	720	350
適合率	0.23	0.17	0.10	0.28	**0.54**
再現率	**0.60**	0.08	0.50	0.52	0.53
F 値	0.34	0.11	0.16	0.37	**0.54**

マッチした予測クラスタの 65%を有している．区間を (0:8; 1:0] に広げると，この区間が有するクラスターは 79%にも増加する．故に，既知の複合体とマッチした予測クラスタの多くは，cellular component で注釈付られていることが分かる．

一方，どの既知の複合体ともマッチしてない予測クラスタに関しては，二つの区間 (0:4; 0:5] と (0:9; 1:0] にピークが存在することが分かる．そのうち区間 (0:9; 1:0] の場合は，既知の複合体とマッチした予測クラスタの場合とほぼ同じ相対頻度である．これらクラスターは既知の複合体とマッチしてないが被覆率が 90%であるため，これらクラスターが真の複合体もしくはそれらと大きく重複する可能性が強く示唆される．例えば，PPSampler が予測した Sgt2/Yor007c と Mdy2/Yol111c からなるクラスターは，どの CYC2008 の既知複合体ともマッチしてないが，cellular component オントロジーの term “cytoplasm” による被覆率は 100%であり，さらに，biological process オントロジーの term “protein targeting” による被覆率も 100% である．そして，このクラスターは，3 つのタンパク質からなる Get4-Get5-Sgt2 複合体の二つのタンパク質に一致している [20].

区間 (0:4; 0:5] に存在するもう一つのピークは，どの既知の複合体ともマッチしてない予測クラスタの約 18%を含んでいる．このビンに含まれるクラスターは相対的に一番低い被覆率 40%から 50%のクラスターとなっている．故に，これらは間違って複合体と予測された可能性が高いと言えそうである．

Biological process と molecular function のオントロジーに関する被覆率の相対頻度分布は，図 2 (b) と (c) に示されている．これらは，(a) の cellular component オントロジーと類似したトレンドを有している．例えば，どの既知の複合体ともマッチしてない予測クラスタの分布は，同じ区間 (0:4; 0:5] と (0:9; 1:0] にピークをもつ．しかしながら，これらの分布は，(a) の場合と比較して形状が緩やかであるので，これらのオントロジーによる予測クラスターの特徴付けは，cellular component より幾分弱いものとなっている．

表 2 クラスター・サイズ相対頻度の目標値パラメータ $\psi(i) \propto i^{-\gamma}$ と予測精度の関係.
Table 2 Relationship between parameter $\psi(i) \propto i^{-\gamma}$ and performance.

γ	1.5	2	3
タンパク質数	2001	2001	2001
クラスター数	260	350	418
適合率	**0.54**	0.54	0.47
再現率	0.40	0.53	**0.55**
F 値	0.46	**0.54**	0.51

表 3 タンパク質総数の目標値パラメータ λ と予測精度の関係.
Table 3 Relationship between parameter λ and performance.

λ	1000	2000	3000	4000	5000
タンパク質数	1002	2001	3000	4000	5000
クラスター数	186	350	501	793	1158
適合率	**0.67**	0.54	0.38	0.24	0.18
再現率	0.35	0.53	0.60	0.60	**0.65**
F 値	0.46	**0.54**	0.47	0.34	0.28

3.3 クラスター・サイズ相対頻度の目標値パラメータ $\psi(i) \propto i^{-\gamma}$ と予測精度の関係

各クラスター・サイズ $i\ (= 2, 3, \cdots, N)$ のクラスター数の相対頻度の目標値を与えるパラメータ $\psi(i) \propto i^{-\gamma}$ の γ の値と予測精度の関係を**表 2** に示しており，γ のデフォルト値 $\gamma = 2$ の結果と $\gamma = 1.5$ と 3 の場合を比較している．この比較実験において，その他のパラメータ値は前節と同じである．

$\gamma = 2$ の場合の F 値 0.54 に比べて，$\gamma = 1.5$ の場合は 0.46 そして $\gamma = 3$ の場合は 0.51 となっている．これらの値は，3.1 節において 2 番目に高かった NWE の 0.37 よりも高いので，異なる γ の値に対して相対的に高い F 値を維持していることが分かる．故に，PPSampler の F 値は γ への依存度は高くないと言える．

3.4 タンパク質総数の目標値パラメータ λ と予測精度の関係

分割 C のサイズ 2 以上のクラスター内のタンパク質の総数 $s(C)$ の目標値パラメータ λ の値と予測精度の関係を**表 3** に示している．λ の値は，1000 から 5000 まで 1000 刻みで増やしている．

この表から分かることは以下のとおりである．まず，適合率は，λ の増加に従い，単調に減少している．これは顕著な傾向である．一方，再現率は，$\lambda = 3000$ のときまで増加しているが，それ以降は飽和していることが分かる．その結果，F 値は，$\lambda = 1000$ から 3000 までが比較的高い値となっている．故に，PPSampler の予測精度は λ と相関があると言える．また，今回の実験においては λ の値は 1000 から 3000 ぐらいが適当と言えるが，新規のデータに対しては，λ の値の選定は重要であることが強く示唆される．

4. まとめ

本稿では，PPI データからタンパク質複合体を予測する問題に対して，複合体のサイズの相対分布を事前知識として用いる M-H アルゴリズムに基づくサンプリング予測手法 PPSampler を提案した．予測精度の比較実験において，PPSampler が既存手法より優れていることを確認した．とくに，遺伝子オントロジーによる評価では，既知の複合体とマッチしてない予測クラスターの多くが共通の GO term を共有していることが分かった．これらは真の複合体であることが期待できる．

参考文献

[1] Brohée S, van Helden J: **Evaluation of Clustering Algorithms for Protein-Protein Interaction Networks**. *BMC Bioinformatics* 2006, **7**:488.

[2] Enright A, Dongen SV, Ouzounis C: **An Efficient Algorithm for Large-Scale Detection of Protein Families**. *Nucleic Acids Research* 2002, **30**:1575–1584.

[3] Bader GD, Hogue CW: **An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks**. *BMC Bioinformatics* 2003, **4**:2.

[4] King A, Prŭlj N, Jurisica I: **Protein Complex Prediction via Cost-Based Clustering**. *Bioinformatics* 2004, **20**:3013–3020.

[5] Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T: **CFinder: Locating Cliques and Overlapping Modules in Biological Networks**. *Bioinformatics* 2006, **22**:1021–1023.

[6] Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S: **Development and Implementation of an Algorithm for Detection of Protein Complexes in Large Interaction Networks**. *BMC Bioinformatics* 2006, **7**:207.

[7] Wu M, Li X, Kwoh C, Ng S: **A Core-Attachment Based Method to Detect Protein Complexes in PPI Networks**. *BMC Bioinformatics* 2009, **10**:169.

[8] Macropol K, Can T, Singh A: **RRW: Repeated Random Walks on Genome-Scale Protein Networks for Local Cluster Discovery**. *BMC Bioinformatics* 2009, **10**:283.

[9] Maruyama O, Chihara A: **NWE: Node-Weighted Expansion for Protein Complex Prediction Using Random Walk Distances**. *Proteome Science* 2011, **9(Suppl 1)**:S14.

[10] Pu S, Wong J, Turner B, Cho E, Wodak S: **Up-to-date Catalogues of Yeast Protein Complexes.** *Nucleic Acids Res* 2009, **37**:825–831.

[11] Maruyama O: **Heterodimeric Protein Complex Identification**. In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine* 2011:499–501.

[12] Barabási AL, Albert R: **Emergence of Scaling in Random Networks**. *Science* 1999, **286**:509–512.

[13] Hastings W: **Monte Carlo Sampling Methods Using Markov Chains and Their Applications**. *Biometrika* 1970, **57**:97–109.

[14] Liu JS: *Monte Carlo Strategies in Scientific Computing*. Springer 2008.

[15] Vlasblom J, Wodak S: **Markov Clustering Versus Affinity Propagation for the Partitioning of Protein Interaction Graphs**. *BMC Bioinformatics* 2009, **10**:99.

[16] Kiemer L, Costa S, Ueffing M, Cesareni G: **WI-PHI: A Weighted Yeast Interactome Enriched for Direct Physical Interactions**. *Proteomics* 2007, **7**:932–943.

[17] Consortium TGO: **Gene Ontology: Tool for the Unification of Biology**. *Nat. Genet.* 2000, **25**:25–29.

[18] **GO Slim and Subset Guide. http://www.geneontology.org/GO.slims.shtml**.

[19] **SGD project. http://www.yeastgenome.org/download-data/curation/literature/go_slim_mapping.tab**.

[20] Brodsky JL: **The Special Delivery of a Tail-Anchored Protein: Why It Pays to Use a Dedicated Courier**. *Molecular Cell* 2010, **40**:5–7.