1N-04

タンパク質データベースにおける高速な相関問合せ手法の提案

直井 悠馬 † 真次 彰平 †† 塩川 浩昭 ‡

† 筑波大学情報科学類 †† 筑波大学理工情報生命学術院 ‡ 筑波大学計算科学研究センター

1 はじめに

相関問合せとは，タンパク質データベースの中からクエリと共起して出現するタンパク質を検出する処理であり，医療や生命科学分野で広く利用されている．大量のタンパク質から構成される大規模なデータベースに対する相関問合せは，タンパク質構造の複雑さによって計算コストが大きくなるという問題がある．

Ke [1][2] らは相関値を厳密に計算せずに候補グラフを枝刈りすることによって探索空間を縮小できることに着目し，効率的な相関グラフ検出手法を提案した．しかしグラフデータベースを構成するグラフがより大規模になると相関判定のコストは大きくなるため依然として問題がある．

そこで本研究ではタンパク質データベースにおける高速な相関問合せ手法を提案する．提案手法はタンパク質構造を要約したグラフを構築することで相関問合せの判定コストの削減を図るとともに，乱択アルゴリズムに基づくランダムな要約グラフを複数利用することで高精度な Top-k 相関グラフ検出を実現する．本研究では実データを用いた提案手法の評価実験を行い，提案手法が Ke らの手法と比較して最大 6.7 倍高速であること，精度が 100 %になることを確認した．

2 前提知識

本研究ではひとつのタンパク質をラベル付無向連結グラフ $g = (V, E, l)$ として考える．V は頂点集合，E は辺集合，l は各頂点と辺にラベル付けする関数である．N 個のグラフから構成されるタンパク質データベースを $\mathcal{D} = \{g_1, g_2, ..., g_N\}$ とする．2 つのグラフ g, g' について，g が g' の部分グラフならば $g \subseteq g'$ と表記し，g' は g のスーパーグラフであると呼ぶ．データベース $\mathcal{D}$ とあるグラフ g について，g のスーパーグラフとなる $\mathcal{D}$ のグラフ集合を $\mathcal{D}_g$ と表記する．g の出現頻度の指標として支持度 $supp(g; \mathcal{D}) = \frac{|\mathcal{D}_g|}{|\mathcal{D}|}$ を定める．文脈上明らかであるときは $supp(g; \mathcal{D})$ を $supp(g)$ と表記する．また 2 つのグラフ g_1, g_2 が同時に出現する割合である結合支持度は $supp(g_1, g_2) = \frac{|\mathcal{D}_{g_1} \bigcap \mathcal{D}_{g_2}|}{|\mathcal{D}|}$ である．

相関グラフ検出とは $\mathcal{D}$ とクエリグラフ q が与えられたとき，クエリと共起して出現する部分グラフを $\mathcal{D}$ から検出することである．2 つのグラフ g_1, g_2 が与えられた時，それらの相関値 $\phi(g_1, g_2)$ は以下のピアソン相関係数によって定義される．

$$\phi(g_1, g_2) = \frac{supp(g_1, g_2) - supp(g_1)\, supp(g_2)}{\sqrt{supp(g_1)\, supp(g_2)\, (1 - supp(g_1))\, (1 - supp(g_2))}}$$

$\phi(g_1, g_2)$ は $[-1, 1]$ の値をとり，正の値が大きいほど g_1, g_2 が $\mathcal{D}$ 内で共起して出現することを表す．

本研究で扱う問題：本研究では Top-k 相関グラフ検出について扱う．Top-k 相関グラフ検出とは，データベース $\mathcal{D}$ の中からクエリグラフ q との相関値が最も高い k 個の部分グラフを見つける問題である．一般的に Top-k 相関グラフ検出は探索すべき候補グラフ数が膨大にあることや支持度計算はコストが高いことが原因となり膨大な計算時間を必要とする．

3 提案手法

本研究ではグラフ要約と乱択アルゴリズムを利用した高速な Top-k 相関グラフ検出手法を提案する．データベース $\mathcal{D}$ 内の各グラフに対して，その構造を要約したグラフを事前に構築する．クエリ q が与えられた際には，要約したグラフを用いて Top-k 相関グラフ検出を行うことによって相関グラフ判定のコストを小さくし高速化を図る．グラフ要約は Top-k 相関グラフ検出に対して偽陽性・偽陰性となる結果をもたらす場合がある．そこで，提案手法は乱択アルゴリズムを利用することにより確率的に偽陽性・偽陰性となる結果を除外し，高精度な Top-k 相関グラフ検出を実現する．提案手法は (1) グラフ要約，(2) 探索，(3) 検証 の段階で構成される．

(1) グラフ要約: $\mathcal{D}$ 内の各グラフを要約したグラフデータベース $\mathcal{D}'$ を構築する．$\mathcal{D}$ が与えられたとき，要約データベース $\mathcal{D}'$ は以下の手順で生成される．

(i) $\mathcal{D}$ からラベル対集合 R をサンプリングする．このとき，R は次の性質を満たすラベル対の集合である：

Fast Correlation Query for Protein Databases

Yuma Naoi†, Shohei Matsugu†† and Hiroaki Shiokawa‡

†College of Information Science, University of Tsukuba ††Graduate School of Science and Technology Degree Programs, University of Tsukuba ‡Center for Computational Sciences, University of Tsukuba

（性質 1）R 内のラベル対は少なくとも 1 つ以上，$\mathcal{D}$ 内のいずれかのグラフで隣接している．（性質 2）ラベル対の両端点は R 内の他のラベル対と重複しない．

(ii) 次の手順を R の要素がなくなるまで繰り返し，要約データベース $\mathcal{D}'$ を構築する．まず，R からラベル対 r を取り出す．その後，$\mathcal{D}$ 内の各グラフ g_i に対して，r を含む場合，それらのノードを 1 ノードに集約する．

乱択アルゴリズムの導入: この要約処理はランダムに辺を集約することになるため，集約の順序や R の内容に依存して Top-k 相関グラフ検出において偽陽性や偽陰性を生じさせる場合がある．そこで提案手法では乱択アルゴリズムを導入することで確率的に偽陽性・偽陰性の排除を図る．具体的には，複数のラベル対集合 $R_1, R_2, \ldots, R_i$ を構築し，それぞれの集合をもとに要約データベース $\mathcal{D}'_1, \mathcal{D}'_2, \ldots, \mathcal{D}'_i$ を構築する．以降の処理で，複数の要約データベースに対して Top-k 相関グラフ検出を実行することで，偽陽性・偽陰性となる部分グラフを除外する．

(2) 探索: 要約データベース集合 $\mathcal{D}'_1, \mathcal{D}'_2, \ldots, \mathcal{D}'_i$ から，クエリグラフ q に対する Top-k 相関グラフ $T_1, T_2, \ldots, T_i$ をそれぞれ検出する．提案手法ではこの処理には TopCor [2] を利用する．このとき，クエリグラフ q はそれぞれの要約データベースに対応するラベル対集合 $R_1, R_2, \ldots, R_i$ によって要約する．要約したクエリグラフ $q'_1, q'_2, \ldots, q'_i$ と $\mathcal{D}'_1, \mathcal{D}'_2, \ldots, \mathcal{D}'_i$ を用いることで各 $\mathcal{D}'$ における Top-k 相関部分グラフを検出する．

(3) 検証: 要約データベースに対して Top-k 相関グラフ検出を行った $T_1, T_2, \ldots, T_i$ に対して正確な Top-k 相関グラフを求めるために検証を行う．$T_1, T_2, \ldots, T_i$ に含まれるグラフは要約グラフであるため，$\mathcal{D}$ の隣接ノード情報と要約ルール集合 R に基づき要約を解除して復元する．その後，復元によって獲得したグラフ g に対して $g \subset g_i$ となる g_i を D から選択し，再度 TopCor を実行する．この手続きにより得られたグラフの中から相関値の高い k 個のグラフが Top-k 相関グラフとなる．

4 評価実験

提案手法の計算時間と処理精度の評価を行うために，実データを用いて提案手法と先行研究 TopCor [2] の比較を行った．実験に用いたデータセットは実際のタンパク質のグラフデータベースである [2]．データセット nci10k, nci20k, nci40k, nci60k はそれぞれ 10K , 20K , 40K , 60K 個のグラフから構成され，それぞれのグラフの平均頂点数は 34，平均辺数は 35 である．クエリグラフは [0.001,0.005] , (0.005,0.01] , (0.01,0.03] , (0.03,1) の範囲の支持度を持つクエリグラフを利用し，それぞれ Q1 , Q2 , Q3 , Q4 とする．精度の評価指標には適合率を用いる．また，実験には Intel Xeon 3.50GHz , 128GB RAM の Linux サーバを用い，実装は C++を用いた．

提案手法は主に事前計算部分である (1) グラフ要約と問合せ処理である (2) 探索以降の処理の 2 つに分けられる．事前に構築する要約データベースの数を 5 つとしたときの事前計算時間を表 1 に示す．

表 1: 事前処理の実行時間

データセット	グラフ数	実行時間 [s]
nci10k	10,000	3.844
nci20k	20,000	7.633
nci40k	40,000	15.537
nci60k	60,000	23.305

また，提案手法における問合せ処理の時間と TopCor の計算時間を比較した結果を図 1 に，要約データベースの数を 1 から 5 まで変化させた際の適合率の推移を図 2 に示す．スペースの都合により，図 1 , 2 では nci10k を使用した結果のみを示す．表 1 より事前計算にはグラフ数に比例した実行時間がかかることがわかるが，図 1 に示した問合せ処理時間と比較すると極めて小さい時間で事前計算を完了していることがわかる．また，図 1 に示したように提案手法は TopCor よりも問合せ処理が高速であり，支持度が大きいほど提案手法が高速になることがわかる．これはグラフ要約によって TopCor の探索空間が小さくなることや相関判定のコストが小さくなることが高速化の要因だと考えられる．

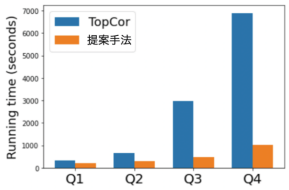

図 1: 問合せ処理時間比較

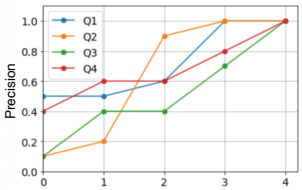

図 2: 適合率

5 まとめ

本研究では，事前に要約したデータベースを構築することで相関判定コストを下げ，乱択アルゴリズムによる高速な Top-k 相関グラフ検出を行う手法を提案した．実データを用いた評価実験により，提案手法は既存手法よりも高速に計算できることが示された．

参考文献

[1] Yiping Ke, James Cheng, and Wilfred Ng. Correlation search in graph databases. In Proc. KDD, pages 390-399, New York, NY, USA, 2007.

[2] Yiping Ke, James Cheng, and Jeffrey Xu Yu. Top-k Correlative Graph Mining. In Proc. SDM, pp.1038-1049, 2009.