Vol.2012-BIO-28 No.17
2012/3/29

パス頻度の上下限制約を満たす木状化合物の二段階列挙法

鈴木政喜[†1]　永持仁[†1]　阿久津達也[†1]

分子構造の部分情報が与えられたとき, それに基づく化合物の推定問題は生物情報学において重要な課題の一つであり, 薬剤設計などの多くの分野に応用することが期待される. 本研究では, 部分的な分子構造として, 長さが与えられた整数 $K \geq 0$ 以下である部分パスを取扱い, このような部分パスに関する特徴ベクトルの上限と下限が与えられたときに, この上下限制約の範囲内の特徴ベクトルに一致する木状の化合物を全て列挙する問題を考える. ここで特徴ベクトルは, 化合物における原子のパスの出現頻度を表す特徴量により定める. これまで一本の特徴ベクトルに一致する木状化合物を列挙する問題に対しては, すでに石田ら (2008), (2010) によって高速な分枝限定法アルゴリズムとして一段階法, 二段階法が提案されている. さらに, 清水ら (2010) は, このうち一段階法アルゴリズムを基に, 上下限制約を持つ木状化合物列挙問題を解く分枝限定法を用いた厳密アルゴリズムを実現した. 本研究では, 上下限制約を持つ木状化合物列挙問題に対する二段階法アルゴリズムを提案する. この提案するアルゴリズムは, 分枝操作として中野と宇野 (2003) によるラベル付き木の列挙アルゴリズムを用いる. 限定操作としては, 特徴ベクトルに上限と下限が与えられているため, 石田ら (2010) による二段階法の限定操作をそのまま用いることはできないので, 上下限制約にも対応できる新たなアルゴリズムを提案する.

A 2-Phase Algorithm for Enumerating Tree-like Chemical Graphs Satisfying Given Upper and Lower Bounds

Masaki Suzuki,[†1] Hiroshi Nagamochi[†1]
and Tatsuya Akutsu[†1]

Enumeration of chemical graphs satisfying given constraints is one of the fundamental problems in chemoinfomatics since they lead to a variety of useful applications including drug design. In this extended abstract, we consider the problem of enumerating all tree-like chemical graphs from a given set of feature vectors, which is specified by a pair of upper and lower feature vectors, where a feature vector represents the frequency of prescribed paths in a chemical conpound to be constructed. To solve the problem for a single feature vectors, Ishida et al. proposed 1-Phase Algorithm and 2-Phase Algorithm. The problem for a given set can be solved by applying the algorithms proposed by Ishida et al. to each single feature vector in the given set, but this method may take a large amount of computation time because in general there are many feature vectors in a given set. Therefore Shimizu et al. proposed a new exact branch-and-bound algorithm based on 1-Phase Algorithm for the problem so that all the feature vectors in a given set are handled directly. We propose a new exact branch-and bound algorithm based on 2-Phase Algorithm for the problem. In our algorithm, the branching operation is based on Nakano and Uno's enumeration algorithm on labeled trees. Since we cannot use the bounding operation proposed Ishida et al. due to the new upper and lower constraints, we introduce new bounding operations based on upper and lower feature vectors, a bond constraint, and a detachment condition.

1. 序　論

1.1 背　景

現在，生体生命情報学にとって重要な目的の一つとなっている薬剤設計における，その重要なステップの一つとして，望ましい性質を持つ化合物を見分ける作業が挙げられる．近年，サポートベクターマシン (SVM) や，カーネル法に基づいた化合物の分類は広く研究されている[4),6),11),12)]．それらに通じる考えは，化合物を特徴ベクトルに写像し，サポートベクターマシンを用いてその特徴ベクトル空間からの予測を行うことである[5)]．特徴ベクトルによる化合物の表現方法はいくつかあるが，ラベル付きのパス[11),12)] や部分構造[4),6)] の出現頻度に基づいたものが広く使われている．

現在，カーネル法を用いて，入力空間上で化合物を設計・最適化する手法が提案されている[2),3)]．この手法では，化合物は適当な関数によって特徴ベクトル空間上での一つの点として表され，その点からの元の入力空間への写像 (原像) を求める．グラフに対する原像問題は，目的関数をうまく選ぶことができれば，所望の性質を持つような化合物を見つけることが期待され，薬剤の設計における候補化合物のスクリーニングへの応用が考えられる．

原像問題に関しては，様々な研究がなされてきた[2),3)]．しかし，これらはヒューリスティックスや確率的な手法に基づいており，厳密アルゴリズムが取り上げられるようになったのは最近のことである．阿久津と深川[1)] は原像問題を，長さ K 以下のラベル付きパスの出現頻

†1 京都大学
Kyoto University

度からグラフを推定する問題に定式化し，この問題が次数制約を持たせた平面的グラフにおいても NP 困難であることを示した．永持は[13] は最大パス長 $K = 1$ のグラフ推定問題は，連結したデタッチメントを見つける問題として定式化することにより，多項式時間で解を得ることができることを証明した．藤原ら[7] は木状の化学構造の推定問題に対して，分枝限定法に基づくアルゴリズムを提案した．彼はこの中で，中野と宇野[14] によって考案された左荷重（left heavy）である木を列挙し，同型なものの重複列挙を防いでいる．また，石田ら[9] が従来のアルゴリズムに，永持[13] によって考案されたデタッチメントによる限定操作を新たに導入し，アルゴリズムの高速化を図った．さらに石田ら[8] が新たに提案した二段階法アルゴリズムでは，アルゴリズムの第一段階で空の木に節点を一つずつ追加し全体の木構造が決定された単純木を構築した後に，第二段階でその単純木の各枝に多重木を割り当てることで解となる多重木を得る．そのため，従来の一段階法アルゴリズムとは異なり，将来化学構造的な制約を追加するときに，全体の木構造が決定された段階で，新たな化学的な限定操作を導入することで容易にアルゴリズムを修正できる (図 1)． 従来の研究では，与えられた特徴ベクトルに完全に一致する化合物を列挙していたが，本研究では，与えられた二つの特徴ベクトル，すなわち上限と下限の特徴ベクトルの「間」にある化合物を列挙する問題を考える．これにより，柔軟な入力を与えることができる．上限と下限の特徴ベクトルの間に存在する全ての特徴ベクトル一つ一つに対して，石田ら[8),9)] によるアルゴリズムを用いれば，この問題は解くことができる．しかし，一般に上限と下限の特徴ベクトルの間には多くの特徴ベクトルが存在するため，時間がかかることが分かっている[15]．そこで，より高速に列挙を行うための分枝限定法に基づく一段階法アルゴリズムを清水ら[15] が提案した．本研究は，上下限制約を持つ問題に対して，さらなる列挙の高速化と新たな化学的な限定操作を導入する余地を得るために，石田ら[8] が提案した上下限制約を持たない木状化合物列挙問題を解く二段階法アルゴリズムをもとに上下限制約を持つ問題にも対応した新たな二段階法アルゴリズムを提案する．このアルゴリズムは，ウエブサーバーとして公開されている（http://sunflower.kuicr.kyoto-u.ac.jp/tools/enumol2/）．

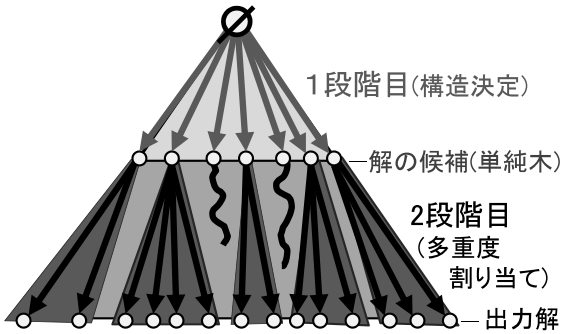

図 1 二段階法アルゴリズムの探索空間図．解の候補となる単純木を作った時点で構造が決定するので，新たな限定操作を導入する余地を作ることができる．
Fig. 1 A branching search space for 2-Phase Algorithm.

1.2 化合物列挙問題

本研究では，水素原子を考慮しない化学グラフをモデルとして取り扱うが，このモデルに対して二つの定式化が提案されている[7),9)]．本節では，用語の定義と石田ら[9] の提案した上下限制約を持たない化合物列挙問題の定式化，そして本研究で扱う清水ら[15] が提案した上下限制約を持つ化合物列挙問題の定式化を解説する．

同じ組み合わせの節点を結ぶ 2 本以上の枝を多重枝という．そうでない枝を単純枝という．多重枝の存在を許すグラフのことを多重グラフと呼び，そうでないグラフのことを単純グラフと呼ぶ．特に，閉路や自己ループを持たない連結な多重グラフのことを多重木と呼び，そうでないグラフは単純木と呼ぶ．ラベルの集合を Σ とし，d 以下の正整数の集合を $B = \{1, 2, \ldots, d\}$ とする．ここで $\Sigma^{k,B}$ は列を要素とする集合であるとし，列の奇数番目の項は Σ の要素であり，偶数番目の項は B の要素となるような長さが $2k-1$ の全ての列を含む．$\Sigma^{\leq k,B} = \cup_{j=1}^{k} \Sigma^{j,B}$, $\mathcal{F}_k(\Sigma, B) = \{g : \Sigma^{\leq k+1,B} \to \mathbb{Z}_+\}$ と定義する．ここで，$\mathbb{Z}_+$ は非負整数の集合とする．

多重グラフ G の各節点 v に対し，$l(v)$ $(\in \Sigma)$ で表されるラベルが与えられているとき，G は Σ-ラベル付きグラフという．さらに，各節点に $val(l(v)) \in \mathbb{N}$ で表される価数が与えられているとき，G は (Σ, val)-ラベル付きグラフという．このとき，木状の化合物は閉路のない連結な (Σ, val)-ラベル付き多重グラフとして表現することができ，ラベル付けされたそれぞれの節点は原子を，節点間の枝の多重度はそれぞれ対応する原子間の結合の多重度を，節点の次数は対応する原子の価数を表す．節点と枝によるパス P を $P = (v_0, m_1, v_1, m_2, \ldots, m_s, v_s)$ としたとき，P のパス長は s であり，そのラベル列を $l(P) = (l(v_0), m_1, l(v_1), m_2, \ldots, m_s, l(v_s))$ と定義する．ここで，v_i $(i = 0, 1, 2, \ldots, s)$ をグラフの節点とし，m_i $(i = 1, 2, \ldots, s)$ を v_{i-1} と v_i を結ぶ枝の本数とする．ラベル列 t に対し，$occ(t, G)$ は G に含まれるラベル列が t のパスの数とする．また，G の最大パス長が K の特徴ベクトルを，$f_K(G) = (occ(t, G))_{t \in \Sigma^{\leq K+1,B}}$

で定義する．図 2 は，化合物を現す (Σ, val)-ラベル付き多重木と，最大パス長が 1 のその多重木の特徴ベクトルの例を示している． 石田ら[9] が提案した与えられた一つの特徴ベクトルから，木状の化合物をすべて列挙する問題は，以下のように定式化される[7]．

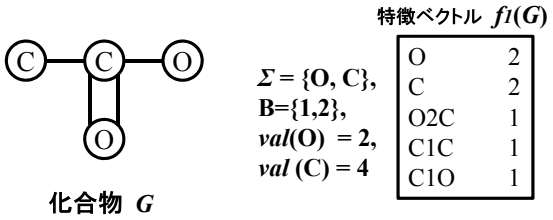

図 2 化合物 G とその特徴ベクトル $f_1(G)$.
Fig. 2 Compound G and the Feature vector $f_1(G)$.

与えられたパス頻度に基づく木状化合物列挙（Enumeration of Tree-like chemical graphs with given Path Frequency, ETPF） 与えられたラベル集合 Σ，正整数集合 B, 自然数 K，特徴ベクトル $g \in \mathcal{F}_K(\Sigma, B)$ および価数 $val : \Sigma \to \mathbb{N}$ から $f_K(T) = g$ かつ，T に含まれる全ての節点 v に対し $deg(v) = val(l(v))$ を満たすような全ての (Σ, val)-ラベル付き多重木 T を出力せよ．もしそのような T が存在しなければ「解なし」と出力せよ．

本研究では，与えられた上限と下限に対応する二つの特徴ベクトルから，条件をみたす木状の化合物を全て列挙することを目的とする．それは問題 ETPF を元にして，以下のように定式化できる[15]．なお，上限を表す特徴ベクトルを g_U，下限を表す特徴ベクトルを g_L と書き，二つの特徴ベクトル $g_1, g_2 \in \mathcal{F}_K(\Sigma)$ において，$g_1 \leq g_2$ であるとは g_1 の全ての成分の値が g_2 以下であると定義する．

与えられた上下限付きパス頻度に基づく木状化合物列挙（Enumeration of Tree-like chemical graphs with given Upper and Lower bounds on path Frequencies, ETULF） 与えられたラベル集合 Σ，正整数集合 B，自然数 K，特徴ベクトル $g_U, g_L \in \mathcal{F}_K(\Sigma, B)$ および価数 $val : \Sigma \to \mathbb{N}$ から $g_L \leq f_K(T) \leq g_U$ かつ，T に含まれる全ての節点 v に対し $deg(v) = val(l(v))$ を満たすような全ての (Σ, val)-ラベル付き多重木 T を出力せよ．もしそのような T が存在しなければ「解なし」と出力せよ．

ここで本研究における上限，下限の特徴ベクトルに設ける条件について述べる．まず，パス $l(v), v \in \Sigma$ については上限と下限が等しいものとする．すなわち，各原子の数は固定される．それ以外のパスについては，上限の特徴ベクトルの成分が下限の特徴ベクトルの成分以上であるとする．すなわち，上限と下限が等しいとする部分については，$g_L = g_U$ であるという条件を課し，それ以外については，$g_L \leq g_U$ という条件を課す．

2. 二段階法アルゴリズムの概要

本研究では，価数と特徴ベクトルの上限と下限の条件を満たす多重木を列挙することが目的である．二段階法アルゴリズムは二段階構成になっている．

アルゴリズムに入る前にまず準備として，入力で与えられた特徴ベクトル g_L, g_U を「単純化」した新たな特徴ベクトル g'_L, g'_U を用意する必要がある.「単純化」については 3 章で詳しく述べる．第一段階では，単純化された特徴ベクトル g'_L, g'_U の条件と価数の条件を満たすような単純木を清水ら[15] の一段階法アルゴリズムに沿って列挙し，解の候補となる木を全て列挙する．清水ら[15] の一段階法アルゴリズムは g'_L, g'_U のような単純化された入力を相手にするとき，高速に処理できることが実験から分かっている．第二段階では、第一段階で得られた解の候補となる単純木の各枝に，入力で与えられた特徴ベクトル g_U, g_L の条件と価数の条件を満たすように，多重度を割り当てることで解となる多重木を列挙する．第一段階で用いる清水ら[15] の一段階法アルゴリズムは分枝限定法アルゴリズムである．分枝操作では同型な単純木を重複して列挙するのを防ぐために，左荷重な根付き木をグラフの標準形とする[14]．空の木 $T = \emptyset$ に根となる節点を追加し，T の節点数が入力で与えられた節点数と等しくなるまで T への追加を繰り返していく．このとき，単純木 T には左荷重の標準形を保ったまま節点を追加していく．また，節点が一つ追加される度に価数カット，特徴ベクトルカットの二つの限定操作をかけて，それぞれの限定操作の条件に反していないかを調べ，もしも一つでも条件に反しているものがあればそこで単純木 T の探索を打ち切る．単純木 T の節点数が入力で与えられた節点数に等しくなったとき，単純木 T が特徴ベクトル g'_L, g'_U と価数の条件を満たしていれば，T を第一段階の出力とし，第二段階の入力とする．第二段階における単純木の各枝への多重度の割り当ては，分枝限定法アルゴリズムを用いる．同型の木を重複して列挙するのを防ぐために，多重木における標準形を導入する．分枝操作は，多重木 T が標準形を保つように単純木の各枝に一本ずつ多重度を割り当てること

で行う．また，節点が一つ追加される度に価数カット，特徴ベクトルカットの二つの限定操作をかけて，それぞれの限定操作の条件に反していないか調べ，もしも一つでも条件に反しているものがあればそこで多重木 T の探索を打ち切る．多重木 T の全ての枝に多重度が割り当てられたとき，特徴ベクトル g_L, g_U の条件と価数の条件を満たすならば多重木 T を解として出力し，そうでなければ多重木 T を破棄する．3 節では，入力で与える特徴ベクトル g_L, g_U から第一段階で用いる特徴ベクトル g'_L, g'_U をどのように用意するかを述べる．5 節で，二段階法アルゴリズムの第二段階について述べる．

3. 第一段階で用いる特徴ベクトルの生成

この節では，第一段階で用いる特徴ベクトルの生成について述べる．この生成を，特徴ベクトルの「単純化」と呼ぶ．入力で与えられた特徴ベクトル g_L, g_U の全ての成分を 2 以上の正整数を全て 1 にした成分に修正したものを新たに g'_L, g'_U として生成する．ただし，それぞれの特徴ベクトルにおいて，修正した成分がすでに存在していた場合，既に存在した成分の値に，修正によって得られたその成分の値を加える (図 3)． 単純化によって得ること

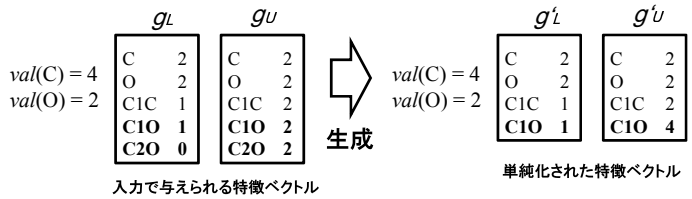

図 3 特徴ベクトルの単純化の例．成分 {C2C} は {C1C} に修正されて，成分の値が加算される．
Fig. 3 An example for simplifying Feature Vector.

ができる特徴ベクトル g'_L, g'_U の条件を満たす木は全て多重枝を持たず，単純木である．ここで，特徴ベクトルの条件を列挙したいグラフが持つべき部分構造の条件として見れば，単純化は多重枝を含む部分構造の全ての多重枝を単純枝に修正する作業にあたる．アルゴリズムの第二段階において，単純枝に修正された枝に多重度を割り当てることで元の多重枝に戻すことができれば，入力で与えた特徴ベクトルの条件，すなわち入力で与えた部分構造の条件を満たす多重木が最終的に得られることは明らかである．

4. アルゴリズムの第一段階

第一段階では，清水ら[15] の一段階法アルゴリズムに「単純化」された特徴ベクトルを入力として与えることで，解の候補となる単純木を全て高速に列挙することができる．分枝操作，限定操作ともに一段階法アルゴリズムを用いるが，多重度カットとデタッチメントカット[15] は単純化した入力に対しては効果が薄いことから導入していない．次の章からは，解の候補となる単純木が一段階法アルゴリズムをもとに得られたとして第二段階の説明に入る．

5. アルゴリズムの第二段階

ここでは，二段階法アルゴリズムの第二段階について述べる．5.1 節では，第二段階で多重木が満たすべき標準形について述べる．5.2 節では，第二段階における分枝操作について述べる．5.3 節では，第二段階における限定操作について述べる．

5.1 第二段階の標準形

この節では以下より，石田ら[8] が導入した多重木の標準形の導入について述べる．n を多重木の節点数とする．$v_0, \ldots, v_{n-1}$ を与えられた木の根から深さ優先探索順で左から右へと順序付けした節点の番号とし，e_i $(1 \leq i \leq n-1)$ は v_i と同様に順序付けして与えられた木の各枝とする．また，v_i の親を $P(v_i)$ とする．木のラベルシーケンス L を以下のように導入する．

$$L := (l(v_0), l(v_1), \ldots, l(v_{n-1})$$

木の多重度シーケンス M を以下のように定義する．

$$M := (m_1, m_2, \ldots, m_{n-1})$$

ここで，m_i $(1 \leq i \leq n-1)$ は枝 e_i の多重度を指す．多重木は単純木 T と多重度シーケンス M から与えられ，それらを合わせて (T, M) と表す．多重木 (T_1, M_1) と (T_2, M_2) は，多重度を考慮しない単純木として同じ型である，すなわち $L(T_1) = L(T_2)$ ならば同型構造といい，同型構造かつ多重度シーケンスは異なってもグラフとして同じ型 (isomorphic) であるとき，同型という．同型な多重木の中で，多重度シーケンスが辞書式順序で最大のものを多重木の標準形とする (図 4)． 次に，単純木から与えられる全ての多重木における親子関係を定義する．$m_i = 0$ は e_i の多重度がまだ決定されていない状態だとする．多重木 (T, M) の親である $(T, P(M))$ は，$M = (m_1, m_2, \ldots, m_i, 0, \ldots, 0)$ である (T, M) に対して，$P(M) = (m_1, m_2, \ldots, m_{i-1}, 0, \ldots, 0)$ となる．明らかに (T, M) が標準形ならば $(T, P(M))$ も標準形となる (図 5)．このようにして多重木の家族木を定義できる．それゆ

Vol.2012-BIO-28 No.17
2012/3/29

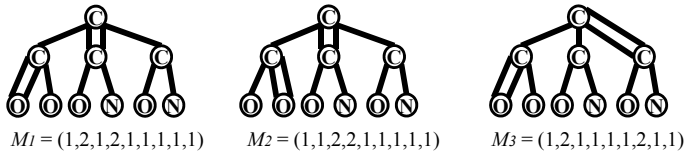

図 4 互いに同型である 3 つの多重木 (T, M_1),(T, M_2),(T, M_3) の例と多重度シーケンス.(T, M_1) が標準形となる.

Fig. 4 (T, M_1) is an canonical represantation.

え，家族木の節点を列挙するだけでよい．これは単純木の e_1 から e_{n-1} まで各辺に多重度を繰り返し割り当てることで実行できる．

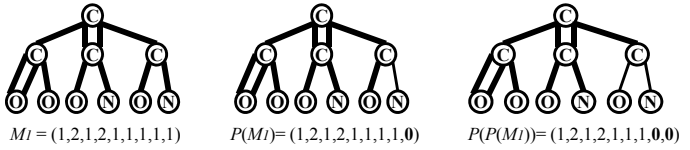

図 5 多重木の親子の関係性を示す図．細い枝はまだ多重度が決定されていないことを表す.
Fig. 5 A relation between a parent and a child for multi-tree.

5.2 第二段階の分枝操作

この節では，石田ら[8] が導入した第二段階における多重木の標準形から成る家族木をなぞる分枝操作について述べる．列挙される多重木の標準形を保つためにコピーフラグを導入する．以下ではコピーフラグと，そのコピーフラグを用いた第二段階における分枝操作について詳しく説明する．

多重木 (T, M) において $T(v)$ と $M(v)$ をそれぞれ節点 v を根とした単純木 T の部分木とその多重度シーケンスとする．同様に，節点 v とその親 $P(v)$ をつなぐ枝 e を用いて，$T(e)$ と $M(e)$ はそれぞれ，$T(v)$ に枝 e を加えた部分木と，その部分木の多重度シーケンスを表す．多重木の節点 v において，$left(v)$ は v と同じ親を持つ節点の中で v の左隣の兄弟の節点を表し，e_v は v と $P(v)$ を繋ぐ枝を表す．$M_1 = (m_1, m_2, \ldots, m_i, 0, \ldots, 0)$, $M_2 = (m'_1, m'_2, \ldots, m'_j, 0, \ldots, 0)$ $(i \geq j)$ とした場合の多重木 (T, M_1) と (T, M_2) で，$1 \leq k \leq i$ において $m_k \geq 1$ かつ，$1 \leq k \leq j$ において $m'_k \geq 1$ となり，$m_k = m'_k$ $(0 \leq \forall k \leq j)$ ならば $M_1 \supseteq M_2$ とする．以上を踏まえてコピーフラグの関数 $copy[v; i]$ を定義する．多重度シーケンス $M = (m_1, m_2, \ldots, m_{n-1})$ が，$m_j \geq 1$ $(1 \leq j \leq i)$，$m_j = 0$ $(i < j \leq n-1)$ となる i に対して，各節点 v について，$copy[v; i]$ は以下のように定義される．

$$copy[v; i] = \begin{cases} 1 & L(T(left(v))) = L(T(v)) \text{ and } M(e_{left(v)}) \subseteq M(e_v) \\ 0 & otherwise. \end{cases}$$

ここで，枝 e_i まで多重度が決定している多重木 (T, M) に，まだ多重度が決定していない全ての枝に多重度を割り当てることで多重木 $(T, \hat{M})$ を得ることを考える．このとき $copy[v; i] = 0$ とは，多重木 (T, M) の残りの枝に今後どのように多重度を割り当てても，部分木 $(T(e_v), \hat{M}(e_v))$ と左兄弟の部分木 $(T(e_{left(v)}), \hat{M}(e_{left(v)}))$ が同型になるような $\hat{M}$ が存在しないことを表す．つまり，e_i まで多重度が決定した時点で，e_v を根とする部分木がその左兄弟の部分木と同型になる可能性がないとき，$copy[v; i] = 0$ となる．ここで，多重木 (T, M) の節点集合 V を扱う関数 $\mu^- : V^2 \to V$ を導入する．μ^- は $j = i - |T(u)|$ の関係性を持つ v_i，v_j，u を用いて，$\mu^-(v_i, u) = v_j$ と定める．この関数は，u を根とした v_i を含む部分木 T_i とその左隣の部分木 T_j が互いに同型構造であるとき，T_i から T_j への同型写像で v_i が写される節点を得ることができ，それが v_j となる (図 6)． 次に，ある節点 v_i

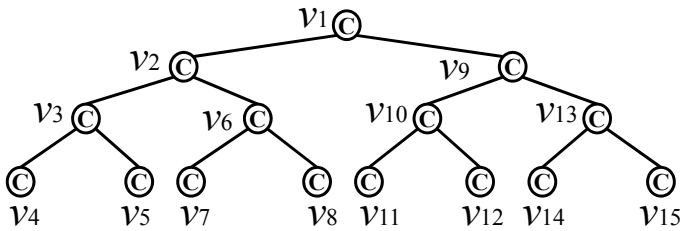

図 6 関数 μ^- の例として，$\mu^-(v_{15}, v_9) = v_{12}$, $\mu^-(v_{15}, v_1) = v_8$ が挙げられる.
Fig. 6 An example for μ^-. $\mu^-(v_{15}, v_9) = v_{12}$, $\mu^-(v_{15}, v_1) = v_8$.

から多重木の根までのパスに含まれる全ての節点を v_i の祖先と定義する．$copy[u_h; i] = 1$ となるような v_i のある祖先 u_h が存在するとき，$v_j = \mu^-(v_i, u_h)$ とその親 $P(v_j)$ との枝 e_j の多重度 m_j は，標準形を崩さず e_i に割り当てることができる多重度 m_i の最大値であり，$m_i \leq m_j$ を満たす．もしも v_i が $copy[u_h; i] = 1$ となるような二つ以上の祖先 u_h を持つとき，以下の補題[8] にもしたがう．

補題 1. v_i の二つの祖先 u_0，u_1（$copy[u_0; i] = 1$，$copy[u_1; i] = 1$）が存在すると仮定す

る．ただし u_0 は u_1 の祖先であるとする．m_j と m_k をそれぞれ e_j と e_k の辺の多重度を表し，e_j と e_k を v_j と v_k とそれぞれの親との枝であり，それぞれ $v_j = \mu^-(v_i, u_0)$ と $v_k = \mu^-(v_i, u_1)$ であるとする．このとき，以下が成立する．

$$m_j \geq m_i \text{ ならば } m_k \geq m_i,$$
$$m_j > m_i \text{ ならば } m_k > m_i.$$

節点の番号付けが深さ優先探索順であることから，上記補題において $j < k < i$ であり，m_j と m_k は一意に多重度が決定している．このコピーフラグを用いた条件と補題は，多重度を割り当てようとしている枝を e_i とすれば，標準形を崩さずに割り当てることができる多重度 m_i の最大値がその枝の祖先にあたる節点から定められることを表している．

5.3 第二段階の限定操作

この章では，第二段階で用いる限定操作について述べる．石田ら[8] は，価数の条件を調べる価数カットのみ導入していたが，本研究では新たに上下限制約を持つ特徴ベクトルの条件を調べる特徴ベクトルカットを提案し，導入する．一つの枝に多重度が割り当てられる度に上記二つの限定操作をかけ，一つでも限定操作の条件を破るものがあればそこで探索を打ち切る．

5.3.1 価数カット

多重木 (T, M) において節点 v の持つ次数を $deg(v; (T, M))$ とする．ここで次数 $deg(v; (T, M))$ は，m 重枝を 1 本持っていたとすれば，その 1 本の枝に対して，次数を m 数えるものとする．$m_j \geq 1$ $(1 \leq j \leq i)$，$m_j = 0$ $(i < j \leq n-1)$ とする．つまり，枝 e_i まで多重度が決定しているとする．

$$val(l(P(v_i))) \geq deg(P(v_i); (T, M)), val(l(v_i)) \geq deg(v_i; (T, M))$$

の条件を満たしているかどうか多重度が枝に割り当てられる度に調べる．もしも条件を満たしていなければ，そこで多重木 (T, M) の探索を打ち切る．

5.3.2 特徴ベクトルカット

$m_j \geq 1$ $(1 \leq j \leq i)$，$m_j = 0$ $(i < j \leq n-1)$ であるとする．つまり，枝 e_i まで多重度が決定しているとする．このとき，多重度シーケンス $M = (m_1, m_2, \ldots, m_{n-1})$ は，$m_j \geq 1$ $(1 \leq j \leq i)$，$m_j = 0$ $(i < j \leq n-1)$ となる．ここで，多重木 (T, M) の部分木 (T', M') を考える．(T', M') は，$e_j (1 \leq j \leq i)$ の枝とそれらの枝の両端の節点で構成される部分木であるとする (図 7)． 多重度が枝に割り当てられる度に以下の条件を満たしているかどうか調べ，満たしていなければそこで多重木 (T, M) の探索を打ち切る．

$$f_K((T', M')) \leq g_U$$

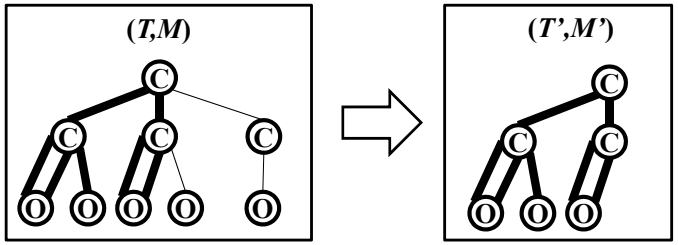

図 7 多重木 (T, M) とその部分木 (T', M') の例. 細い枝はまだ多重度が決定していないものを表す.
Fig. 7 An example for (T, M) and (T', M').

この限定操作は，(T, M) からの探索によって得られる多重木は必ず (T', M') を含むことから，上記の条件を満たさない時点でそれ以上探索をしても得られる多重木が上限の特徴ベクトル g_U を必ず越えることが約束されるため，探索を打ち切ることができる．

5.4 解の出力条件

多重木の全ての枝の多重度が決定したとき，その多重木が解としてふさわしいか調べる解の出力条件について述べる．石田ら[8] が提案した上下限制約を持たない化合物列挙問題に対する解の出力条件は以下のようになる．

$$g = f_K((T, M))$$

の条件を全ての枝の多重度が決定した時点で満たしていれば T を解の一つとして出力し，満たしていない場合は破棄する．

しかし本研究においては，与えられる特徴ベクトルが上下限制約を持つため，上記の出力条件をそのまま用いることはできないので，新たな解の出力条件を以下のように提案する．

$$g_L \leq f_K((T, M)) \leq g_U$$

の条件を全ての枝の多重度が決定した時点で満たしていれば T を解の一つとして出力し，満たしていない場合は破棄する．

6. 実験結果と考察

計算実験は PC (Intel Core i 5 @3.20GHz 4GiB) 上で行った．問題例は，KEGG LIGAND データベース[10]（http://www.genome.jp/ligand/）から抽出したものである．ここで，ベンゼン環は価数が 6 の新たな原子であるとみなしている．本研究では，抽出した問題例を特徴ベクトル t の形で表現し，上限の特徴ベクトルと下限の特徴ベクトルのパス頻

度の幅を $w \in \mathbb{Z}_+$ とする．つまり，ベクトル t の各成分値 $a > 0$ に対し，その値を $a + w$ としたものを上限の特徴ベクトル，逆に $a - w$（$a - w < 0$ なら $a := 0$）としたものを下限の特徴ベクトルとする．特に，$w = 0$ とすれば ETULF の問題例は ETPF の問題例に帰着される．本研究における計算実験では，幅 w を 1 として，KEGG LIGAND データベースから抽出した問題例から ETULF の問題例を生成し，清水ら[15] の一段階法アルゴリズム 1-Phase と本研究の二段階法アルゴリズム 2-Phase に対し，計算時間と実行可能解の数を出力し比較を行った．**表 1** は，ETULF に対する二つのアルゴリズム（1-Phase, 2-Phase）の比較実験の結果である．ここで，入力例は KEGG LIGAND データベースにおいて用いられているエントリー番号であり，n は各問題における原子数，K は特徴ベクトルの最大パス長，解は制限時間内に計算された実行可能解の数を表している．制限時間は 1800 秒とし，各問題に対する計算時間，制限時間内に計算された実行可能解の数を記した．計算時間における”T.O.”とは，制限時間以内に条件の満たす全ての多重木を列挙できなかったことを表す．

まず，K の値が小さいほど，1-Phase と 2-Phase 共に計算時間と解の数が大きいことが分かる．これは，K が小さいことで，特徴ベクトルの条件が少なく，特徴ベクトルカットの限定操作が効きにくいため探索空間が大きくなってしまうことが，計算時間が大きい理由である．解の数が大きい理由は同様に，多重木が満たすべき特徴ベクトルの条件が少ないため，出力条件を満たす多重木が多くなる．

K の値が小さい入力例に対しては，全ての問題例で 2-Phase が 1-Phase より早く問題を解くことが分かる．制限時間内に終わらなかった入力例でも，制限時間内に出力した解の数は 2-Phase の方が 1-Phase より大きいため，2-Phase が 1-Phase より高速に列挙できていることが分かる．ただし，C03630 の K が大きな入力例においては，1-Phase の方が高速に解くことができている．これより，2-Phase は多重木が満たすべき条件が少ないような入力に対して 1-Phase より高速に解くことができ，1-Phase は満たすべき条件が多いような入力に対して，稀に 2-Phase より高速に解くことができることが分かる．

よって，二段階法アルゴリズムは，将来の化学的な構造に基づく限定操作を導入する余地を与えるとともに，一段階法アルゴリズムよりも計算を高速にすることができたといえる．

7. まとめと今後の課題

本研究では，与えられた上下限を表す二つの特徴ベクトルから，条件を満たす木状化合物を全て列挙する問題に対し，石田ら[8] の二段階法アルゴリズムを基に，分枝限定法に基づ

表 1 ETULF に対する二つのアルゴリズムによる比較実験
Table 1 Comparison of 1-Phase and 2-Phase

入力例	n	K	1-Phase solution	1-Phase Time	2-Phase solution	2-Phase Time
C00062 C6H12N2O4	12	1	128,184	2.98	128,184	0.13
		2	132	0.04	132	0.03
		3	46	0.02	46	0.01
		4	1	0.01	1	0.01
		5	1	0.01	1	0.01
		6	1	0.01	1	0.01
		7	1	0.01	1	0.01
C03343 C16H22O4	15	1	2,187,665	29.01	2,187,665	4.06
		2	67	0.02	67	0.02
		3	38	0.02	38	0.02
		4	7	0.02	7	0.02
		5	7	0.02	7	0.02
		6	7	0.02	7	0.02
		7	5	0.02	5	0.01
C07178 C21H28N2O5	18	1	6,788,308	640.76	6,788,308	61.17
		2	45	0.15	45	0.17
		3	6	0.09	6	0.02
		4	1	0.03	1	0.01
		5	1	0.03	1	0.01
		6	1	0.03	1	0.01
		7	1	0.02	1	0.01
C03690 C24H38O4	23	1	114,441,124	T.O.	551,188,027	T.O.
		2	29,756	23.44	29,756	12.61
		3	4,505	14.01	4,505	7.46
		4	130	4.02	130	1.35
		5	61	2.59	61	0.91
		6	12	1.31	12	0.45
		7	5	1.30	5	0.34
C04036 C19H39O7P	27	1	5,153,380	T.O.	407,505,280	T.O.
		2	7,833	T.O.	9,167	T.O.
		3	862	1208.42	862	564.53
		4	16	43.53	16	39.94
		5	14	15.99	14	11.31
		6	13	9.63	13	6.15
		7	12	8.77	12	5.23
C03630 C21H39O7P	29	1	0	T.O.	23,006,164	T.O.
		2	39,857	T.O.	228,886	T.O.
		3	8,438	T.O.	2,117	T.O.
		4	204	335.83	204	302.27
		5	86	69.98	86	81.13
		6	45	26.53	45	41.85
		7	25	12.29	25	35.46

いた厳密アルゴリズムを構成した．その結果，本研究で扱う問題を解くことができる清水ら[15] の一段階法アルゴリズムに比べて，制限時間内に計算を終了するのが困難になるような与える条件の少ない入力に対して，特に高速に列挙することができた．また，二段階法アルゴリズムは，そのアルゴリズムの性質上，化学的な構造による限定操作を導入する余地を持つことができるため，上下限制約を満たす木状化合物の列挙問題に対してそのような余地を作ることができた．

今後の課題としては，新たな限定操作の導入によって，探索節点数を減少させることが挙げられる．また，入力で与えられた特徴ベクトルの条件を見直すような，より効果的に特徴ベクトルカットの限定操作が働くようなアルゴリズムの前処理の導入をすることができれば，さらなる探索節点数の減少と計算時間の短縮が期待できる．

参 考 文 献

1) Akutsu, T. and Fukagawa, D.: Inferring a graph from path frequency, *Lecture Notes in Computer Science*, Vol.3537, pp.371–392 (2005).
2) Bakir, G.H., Weston, J. and Schölkopf, B.: Learning to find pre-images, *Advances in Neural Information Processing Systems*, Vol.16, pp.449–456 (2003).
3) Bakir, G.H., Zien, A. and Tsuda, K.: Learning to find graph pre-images, *Lecture Notes in Computer Science*, Vol.3175, pp.253–261 (2004).
4) Byvatov, E., Fechner, U., Sadowski, J. and Schneider, G.: Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, *Journal of Chemical Information and Computer Sciences*, Vol. 43, pp. 1882–1889 (2003).
5) Cristianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press (2000).
6) Deshpande, M., Kuramochi, M., Wale, N. and Karypis, G.: Frequent substructure-based approaches for classifying chemical compounds, *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, pp.1036–1050 (2005).
7) Fujiwara, H., Wang, J., Zhao, L., Nagamochi, H. and Akutsu, T.: Enumerating Tree-like Chemical Structures from Feature Vector, *IPSJ SIG Technical Reports*, Vol.2006, No.135, pp.111–118 (2006).
8) Ishida, Y.: Improved algorithms for enumerating tree-like chemical graphs with given path frequency, *Kyoto-university graduation master's thesis* (2010).
9) Ishida, Y., Zhao, L., Nagamochi, H. and Akutsu, T.: Improved algorithms for enumerating tree-like chemical graphs with given path frequency, *Genome Informatics*, Vol.21, pp.53–64 (2008).
10) Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M.: KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Res*, Vol.38, pp.D355–D360 (2010).
11) Kashima, H., Tsuda, K. and Inokuchi, A.: Marginalized kernels between labeled graphs, *Proceedings of the 20th International Conference Machine Learning*, pp. 321–328 (2003).
12) Mahé, P., Ueda, N., Akutsu, T., Perret, J.L. and Vert, J.P.: Graph kernels for molecular structure-activity relationship analysis with support vector machines, *Journal of Chemical Information and Modeling*, Vol.45, pp.939–951 (2005).
13) Nagamochi, H.: A detachment algorithm for inferring a graph from path frequency, *Algorithmica*, Vol.53, pp.207–224 (2009).
14) Nakano, S. and Uno, T.: Efficient Generation of Rooted Trees, *NII Technical Report* (NII-2003-005E, 2003).
15) Shimizu, M., Nagamochi, H. and Akutsu, T.: Enumerating Tree-like Chemical Graphs with Given Upper and Lower Bounds on Path Frequencies, *IPSJ SIG Technical Reports* (2011).