4H2-2

遺伝的プログラミングによる 複合的なブロック保存型外平面的グラフパターンの獲得

Acquisition of Multiple Block Preserving Outerplanar Graph Patterns by Genetic Programming

徳原 史也 *1 Fumiya Tokuhara　宮原 哲浩 *1 Tetsuhiro Miyahara　久保山 哲二 *2 Tetsuji Kuboyama　鈴木 祐介 *1 Yusuke Suzuki　内田 智之 *1 Tomoyuki Uchida

*1広島市立大学情報科学研究科 Graduate School of Information Sciences, Hiroshima City University

*2学習院大学計算機センター Computer Centre, Gakushuin University

Machine learning and data mining from graph structured data are studied intensively. Many chemical compounds can be expressed by outerplanar graphs. We use block preserving outerplanar graph patterns as representations of graph structures. We report a method for acquiring multiple block preserving outerplanar graph patterns from positive and negative outerplanar graph data by Genetic Programming

1. はじめに

近年グラフ構造を持つデータが増加しており，グラフ構造を持つデータからの機械学習やデータマイニングが注目されている．また，多くの化合物は外平面的グラフの構造を持つ．そこで現在までに機能がわかっている化合物において，機能がある化合物にマッチし，機能が無い化合物にはマッチしないような構造的特徴を獲得することができれば，それを用いる分野において大きな意義があるといえる．本研究では，外平面的グラフの構造的特徴を表現するために，佐々木ら [Sasaki 08, Yamasaki 09] によるブロック保存型外平面的グラフパターン（Block Preserving Outerplanar graph pattern，BPO グラフパターン）を用いる．また，学習手法として，遺伝的プログラミング (Genetic Programming, GP)[Koza 92] を木パターンに適用させた手法 [Nagamine 07] を用いる．

佐々木ら [Sasaki 08, Yamasaki 09] は，BPO グラフパターンの照合アルゴリズムと極小一般化 BPO グラフパターンの発見アルゴリズムを提案し，BPO グラフパターン言語は正事例から多項式時間帰納推論可能であることを示している．さらに，正事例からの頻出 BPO グラフパターン列挙アルゴリズムを提案している．大内山ら [Ouchiyama 15] は，正事例と負事例の外平面的グラフから特徴的な外平面的グラフパターンを獲得する進化的手法を実現し，人工データに適用している．中居ら [Nakai 14] は特徴的な VLDC 木パターン獲得手法を用いる二段階構造の進化的計算で，特徴的な VLDC 木パターン集合を獲得する手法を提案している．

我々は，大内山ら [Ouchiyama 15] の手法を発展させて，正事例からラベルの情報を抽出し，抽出したラベルの情報を利用した遺伝的プログラミングを用いて特徴的な BPO グラフパターンを獲得する進化的手法を実現し，化合物データに適用している [Tokuhara 16a, Tokuhara 16b, 徳原 16]．しかし一つの BPO グラフパターンで特徴をとらえる事が難しいような外平面的グラフの構造を持つデータも存在する．そこで，本研究では中居ら [Nakai 14, 山縣 17] の二段階構造の進化的獲得手法を採用し，遺伝的プログラミングによる複合的な BPO グラフパターンを獲得する手法を提案する．ここで，複合的 BPO グラフパターンとは，BPO グラフパターン集合のことをいう．

連絡先: 徳原史也，広島市立大学情報科学研究科，〒 731-3194，広島市安佐南区大塚東 3-4-1，mb67015@e.hiroshima-cu.ac.jp

2. 準備

本研究で扱う，BPO グラフパターンとブロック木パターンの説明をする [Sasaki 08，Yamasaki 09，Tokuhara 16a, Tokuhara 16b, 徳原 16]．

G をグラフ，Λ と Δ をアルファベットとする．ラベル付きグラフとは，頂点集合 $V(G)$ と辺集合 $E(G)$ の各要素が，それぞれ Λ と Δ の要素によってラベル付けされたグラフをいう．ラベル付きグラフのすべての頂点が外平面に接するように平面埋め込みが可能であるとき，そのグラフを外平面的グラフとよぶ．連結グラフにおいて，削除することでグラフを非連結にすることができる頂点をカット点という．外平面的グラフのブロックとは，頂点数 3 以上のカット点をもたない極大な二重連結成分をいう．ブロックに属さない辺をブリッジとよぶ．BPO グラフパターンとは，ブリッジ変数および末端変数とよばれる 2 種類の構造的変数を持つ連結な外平面的グラフである．本稿では，連結な外平面的グラフのみを扱う．以後，連結な外平面的グラフを単に外平面的グラフということにする．外平面的グラフ G と BPO グラフパターン p に対し，p の全ての変数を適当な外平面的グラフで置き換えることによって G が得られるとき p と G はマッチするという．外平面的グラフと BPO グラフパターンの例を図 1 に示す．図 1 において，BPO グラフパターン p のブリッジ変数 X，Y と末端変数 Z をそれぞれ外平面的グラフ g_1，g_2，g_3 に置き換えることで外平面的グラフ G を得ることができるので，p と G はマッチする．

BPO グラフパターン p のブロック部分を，ブロックの辺ラベルの情報を保持したブロック頂点へと置き換える事で得られる，根なし無順序木の構造を持つグラフパターンを p のブロック木パターンとよび，$t(p)$ で表す．ブロック木パターンの例を図 2 に示す．

3. 遺伝的プログラミングによる複合的な BPO グラフパターンの獲得

本研究では，遺伝的プログラミングにより BPO グラフパターン集合を獲得する．BPO グラフパターン集合の獲得では，特徴的な BPO グラフパターンを獲得する GP[Tokuhara 16a, Tokuhara 16b, 徳原 16] を手続きとして使う．

3.1 特徴的な BPO グラフパターンの獲得

我々は，特徴的な BPO グラフパターンを獲得するための進化的手法 [Tokuhara 16a, Tokuhara 16b, 徳原 16] を提案して

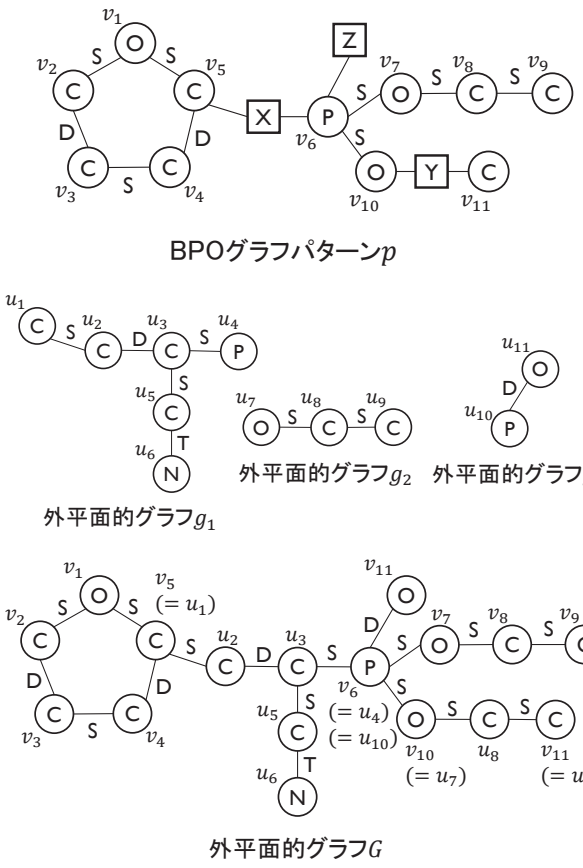

図 1: BPO グラフパターンと外平面的グラフの例．頂点内の文字列は頂点ラベル，辺のそばの文字列は辺ラベル，正方形は変数を表すものとする．

いる．

特徴的な BPO グラフパターン獲得問題

入力: 正事例及び負事例からなる外平面的グラフの有限集合 D.

問題: D に関する適合度の高い BPO グラフパターンを獲得する．

BPO グラフパターン p の D に関する適合度 $fitness_D(p)$ は，$fitness_D(p) = (p$ が D の正事例にマッチする割合 $+p$ が D の負事例にマッチしない割合 $)/2$ で定義する．よって，D に関する適合度の高い BPO グラフパターンとは，D の多くの正事例にマッチし，D の負事例にあまりマッチしないような特徴的な BPO グラフパターンであるといえる．GP の遺伝操作は BPO グラフパターン p のブロック木パターン $t(p)$ に適用する．ブロック木パターンに対する交叉の適用例を図 3 に示す．

3.2 特徴的な BPO グラフパターン集合の獲得

VLDC 木パターンに対する二段階構造の進化的手法 [Nakai 14]，TTSP グラフパターンに対する二段階構造の進化的手法 [山縣 17] を基に，特徴的な BPO グラフパターン集合を獲得するための進化的手法を以下に示す．

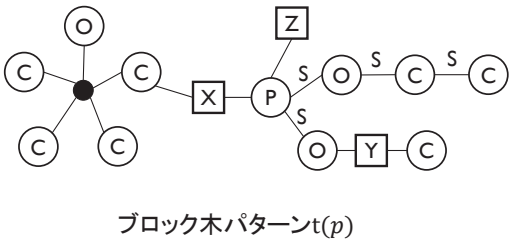

図 2: 図 1 の BPO グラフパターン p のブロック部分を変換して得られた p のブロック木パターン $t(p)$.

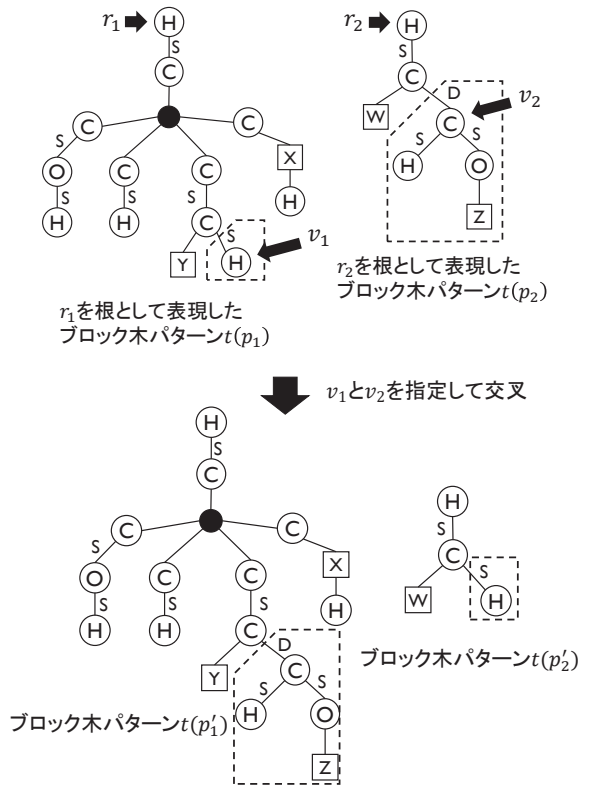

図 3: ブロック木パターンに対する交叉の適用例.

特徴的な BPO グラフパターン集合獲得問題

入力: 正事例集合 P 及び負事例集合 N からなる外平面的グラフの有限集合 D，正整数 $c\,(1 \leq c < |P|)$.

問題: D に関する適合度の高い BPO グラフパターン集合 $\Pi\,(1 \leq |\Pi| \leq c)$ を獲得する．

ここで集合 A の要素数を $|A|$ と表す．BPO グラフパターン集合 Π に含まれる BPO グラフパターンの少なくとも 1 つと外平面的グラフ G がマッチするとき，Π と G はマッチするという．BPO グラフパターン集合 Π の D に関する適合度 $fitness_D(\Pi)$ は，$fitness_D(\Pi) = (\Pi$ が D の正事例にマッチする割合 $+\Pi$ が D の負事例にマッチしない割合 $)/2$ で定義する．

BPO グラフパターン集合を獲得する進化的手法 (メインルーチン) は，特徴的な BPO グラフパターンを獲得する GP をサブルーチンとして使う二段階の構造をしている．世代 s のサブルーチン終了時の BPO グラフパターン π の適合度を π の基本適合度といい，メインルーチンで獲得した BPO グラフパ

ターン集合での π の出現回数に比例した値を π の加算適合度という．π の基本適合度に π の加算適合度を加えた値を π の適合度として，世代 $s+1$ の GP を開始する．各 GP で得られた π_i からなる BPO グラフパターン集合 $\{\pi_1, \pi_2, \ldots, \pi_c\}$ を，BPO グラフパターン列 $[\pi_1, \pi_2, \ldots, \pi_c]$ として扱う．

特徴的な BPO グラフパターン集合獲得手法

1. BPO グラフパターンの上位数 k，進化的手法の集団サイズ b，進化的手法のエリートサイズ e，GP の集団サイズ b'，GP のエリートサイズ e'，最大世代数 n，BPO グラフパターンの加算適合度の最大値 C_{add} を設定する．
2. 初期世代として $s = 1$ とする．
3. P を c 個に分類したクラスタが与えられたとき，$D_j (1 \leq j \leq c)$ を，P の j 番目のクラスタである pos_j を正事例集合，N を負事例集合とするデータの与え方とする．
4. $D_1, D_2, \ldots, D_c$ それぞれに対して，正事例からラベルの情報を抽出し，それを利用して初期 BPO グラフパターンを生成し，特徴的な BPO グラフパターンを獲得する GP による学習過程を始める．それぞれの学習過程を $GPL_1, GPL_2, GPL_3, \ldots, GPL_c$ とする．
5. 現世代 s の $GPL_1, GPL_2, GPL_3, \ldots, GPL_c$ を実行する．
6. $GPL_1, GPL_2, GPL_3, \ldots, GPL_c$ の個体の BPO グラフパターンの適合度を評価して基本適合度とする．
7. PAT_{seq}^{prv} は前世代 $(s-1)$ の適合度上位 e 個の BPO グラフパターン列の集合とする．PAT_{seq} を現世代 s の各 GPL_j の基本適合度が上位 k 個の BPO グラフパターン $\pi_j (1 \leq j \leq c)$ からなる BPO グラフパターン列 $[\pi_1, \pi_2, \ldots, \pi_c]$ のすべてからなる集合とする．PAT_{seq_best} を $PAT_{seq} \cup PAT_{seq}^{prv}$ の適合度上位 b 個の BPO グラフパターン列の集合とする．PAT_{seq_best} を現世代 s の BPO グラフパターン列の集団とする．
8. 終了世代 $s = n$ に達していれば終了とする．
9. GP の学習過程 GPL_j の各 BPO グラフパターン π_j に対し，PAT_{seq_best} の BPO グラフパターン列の j 番目の要素としての BPO グラフパターンの出現回数を $n(\pi_j)$ とする．GPL_j の中の BPO グラフパターンの基本適合度に π_j の加算適合度 $\left(C_{add} * \frac{n(\pi_j)}{b}\right)$ を加えた値を π_j の適合度とする．
10. GP の処理過程 $GPL_1, GPL_2, GPL_3, \ldots, GPL_c$ を継続し，現世代 s の集団を次世代 $s+1$ の集団とし，$s = s+1$ として 5. へ戻る．

4. 実験

設定した BPO グラフパターン集合 $\Pi = \{\pi_1, \pi_2, \pi_3\}$ について，人工的に生成した外平面的グラフのうち Π にマッチするものを正事例，マッチしないものを負事例として，500 個の正事例と 500 個の負事例からなる集合 D を作成した．3.2 章の特徴的な BPO グラフパターン集合獲得手法を Intel Xeon CPU E5-2630 v2 2.60GHz のプロセッサ，実装メモリ 32.0GB の Windows10 Pro 64bit OS 上に Java 言語で実装し，人工データ D を用いて，正整数 c を 3 とし 10 試行の実験を行った．また，比較のため 3.1 章の特徴的な BPO グラフパターン獲得手法を同様の環境で実装し，D を用いて 10 試行の実験を行った．設定した BPO グラフパターン集合を図 4 に示す．

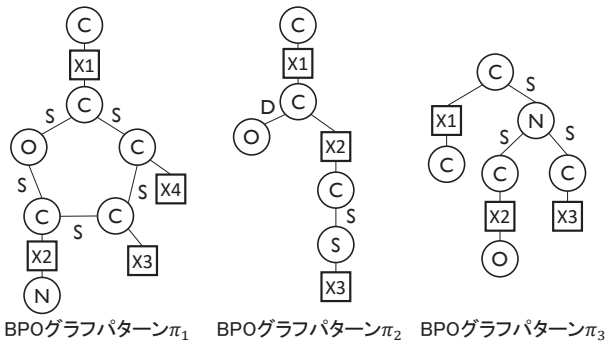

図 4: 人工データ作成のために設定した BPO グラフパターン集合 $\Pi = \{\pi_1, \pi_2, \pi_3\}$.

特徴的な BPO グラフパターン集合獲得手法 (メインルーチン) のパラメータを次のように設定した．BPO グラフパターンの上位数 (k):5，進化的手法の集団サイズ (b):30，進化的手法のエリートサイズ (e):5，最大世代数 (n):200，加算適合度の最大値 (C_{add}):0.1.

特徴的な BPO グラフパターン獲得手法 (サブルーチン及び比較実験) のパラメータを次のように設定した．集団サイズ (b'):50，エリートサイズ (e'):3，最大世代数:200，トーナメントサイズ:2，複製確率:0.05，交叉確率:0.50，突然変異確率:0.45.

特徴的な BPO グラフパターン獲得手法，特徴的な BPO グラフパターン集合獲得手法の全 10 試行の実行時間の平均はそれぞれ，16315 秒，37975 秒であった．また特徴的な BPO グラフパターン獲得手法，特徴的な BPO グラフパターン集合獲得手法の各試行における最終世代の最良個体の適合度を表 1 に，各世代における最良個体の適合度の 10 試行の平均を図 5 に示す．最終世代の最良個体である BPO グラフパターンと BPO グラフパターン集合をそれぞれ図 6，図 7 に示す．

表 1: 各試行における最終世代の最良個体の適合度

試行	BPO グラフパターン獲得	BPO グラフパターン集合獲得
1	0.718	0.871
2	0.718	0.878
3	0.718	0.867
4	0.718	0.866
5	0.718	0.877
6	0.718	1.000
7	0.718	0.876
8	0.718	0.880
9	0.718	0.871
10	0.718	0.869
平均	0.718	0.886

5. おわりに

本研究では，特徴的な BPO グラフパターン獲得手法を基に特徴的な BPO グラフパターン集合獲得手法を提案し，人工データに適用した．今後の課題として実データへの適用，計算時間の短縮など挙げられる．

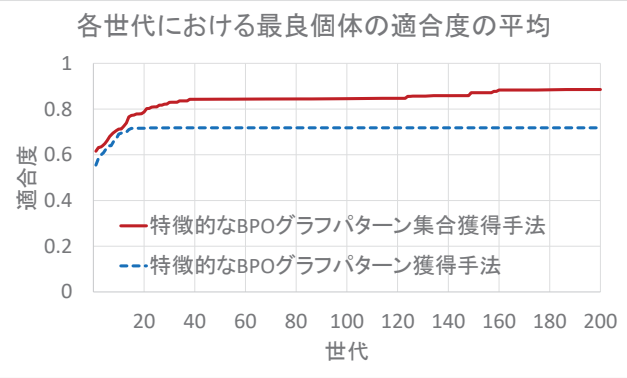

図 5: 各世代における最良個体の適合度の平均.

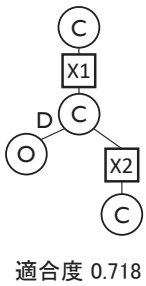

図 6: 特徴的な BPO グラフパターン獲得手法の第 4，5，7，10 試行目の最終世代の最良個体.

参考文献

[Sasaki 08] Y.Sasaki et al., Mining of Frequent Block Preserving Outerplanar Graph Structured Patterns, Proc. ILP 2007, LNAI 4894 Springer, pp.239-253, 2008.

[Yamasaki 09] H.Yamasaki et al., Learning Block-Preserving Graph Patterns and Its Application to Data Mining, Machine Learning, Vol.76 No.1, pp.137-173, 2009.

[Koza 92] J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, 1992.

[Nagamine 07] M.Nagamine et al., A Genetic Programming Approach to Extraction of Glycan Motifs using Tree Structured Patterns, Proc. Al 2007, LNAI 4830, Springer, pp.150-159, 2007.

[Ouchiyama 15] Y.Ouchiyama et al., Acquisition of Characteristic Block Preserving Outerplanar Graph Patterns from Positive and Negative Data using Genetic Programming and Tree Representation of Graph Patterns", Proc. IWCIA 2015, pp.95-101, 2015.

[Nakai 14] S.Nakai et al., Acquisition of Characteristic Sets of Tree Patterns with VLDC's using Genetic Programming and Edit Distance, Proc. IWCIA 2014, pp.113-118, 2014.

[Tokuhara 16a] F.Tokuhara et al., Acquisition of Characteristic Block Preserving Outerplanar Graph Patterns by Genetic Programming using Label Infomation, Proc. IIAI-AAI 2016, pp.203-210, 2016.

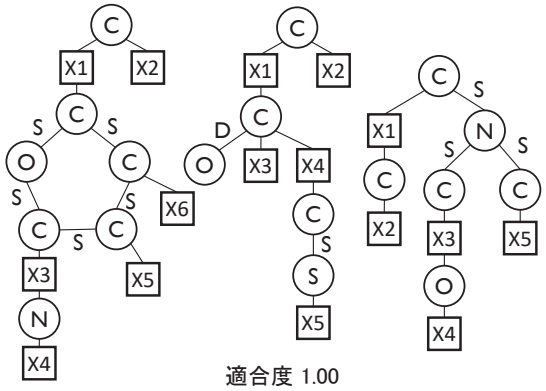

図 7: 特徴的な BPO グラフパターン集合獲得手法の第 6 試行目の最終世代の最良個体.

[Tokuhara 16b] F.Tokuhara et al., Using Canonical Representations of Block Tree Patterns in Acquisition of Characteristic Block Preserving Outerplanar Graph Patterns, Proc. IWCIA 2016, pp.93-99, 2016.

[徳原 16] 徳原史也ほか，" ラベルの情報を利用した遺伝的プログラミングによる特徴的なブロック保存型外平面的グラフパターンの獲得 " 火の国情報シンポジウム 2016 論文集，6C-3，2016.

[山縣 17] 山縣佑貴ほか，" TTSP グラフパタン集合を個体とする進化的手法による複合的グラフ構造パターンの獲得 " 火の国情報シンポジウム 2017 論文集，B5-2，2017.