編集距離と遺伝的プログラミングを利用した特徴的木パターンの獲得
Acquisition of Characteristic Tree Patterns Using Edit Distance and Genetic Programming

中居　翔平　宮原　哲浩
Shohei Nakai　Tetsuhiro Miyahara
広島市立大学情報科学研究科
Email: nakai@ml.info.hiroshima-cu.ac.jp

Abstract— **We propose a learning method for acquiring characteristic tree patterns from positive and negative tree structured data, by using edit distance and Genetic Programming. We report some experimental results on applying our method to glycan data.**

I.　はじめに

遺伝的プログラミング(GP) [1]とは，遺伝的アルゴリズム(GA)の遺伝子型を拡張し，構造的表現（木構造）を扱えるようにしたものである．木構造を用いることにより一般的な GA では扱うことが難しかった数式やプログラムなどの構造表現を扱うことができる．

木構造データの例として糖鎖がある．糖鎖は核酸(DNA)とタンパク質に続く 3 番目に重要な生体分子である．その構造の複雑さから糖鎖の機能や構造の解析は核酸やタンパク質に比べて進んでいない．本研究では，木パターンと木データの編集距離と遺伝的プログラミングを利用して，正事例と負事例の木データから特徴的な VLDC 付き木パターン[4]を獲得する手法を提案する．VLDC(variable-length don't care)は木データの一部を代入できる構造的変数である．

関連研究[2]では遺伝的プログラミングによる特徴的なタグ木パターン抽出手法が提案されている．

II.　準備

A.　木パターン

本研究では，木構造データの構造的特徴を表現するため，VLDC 付き木パターンと呼ぶ木構造パターンを用いる．以降 VLDC 付き木パターンを木パターンという．木パターンは，ノードラベルでデータを表現し，木データの一部を代入できる VLDC 変数を持つ．この VLDC 変数には Path-VLDC と Umbrella-VLDC の二種類がある．それぞれの定義を以下で述べる．

木データの根から葉までの経路の一部が代入可能な VLDC 変数を Path-VLDC という．表記では“|”で表される．

木データの根から葉までの経路の一部と，その経路の一部上のノードから出ているすべての部分木も代入可能な VLDC 変数を Umbrella-VLDC という．ただし経路で一番下のノードから出ている部分木は含まなくてもよい．表記では“Λ”で表される．

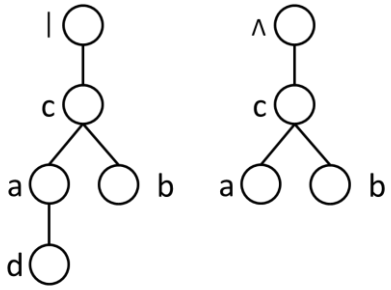

図 2.1：木パターンの例

B.　編集距離

木 T_1 と木 T_2 の編集距離 $\mathrm{treedist}(T_1,T_2)$は，$T_1$ を T_2 に変換するためにノード削除，ノード挿入，ノードラベル置換の 3 種類の編集操作を用いて編集する際にかかるコストの総和の最小値として定義される[3]．ノード削除，ノード挿入，ノードラベル置換の編集コストは全て 1 としている．

S を木パターン P における可能な VLDC 変数への代入の全ての集合とする．P の VLDC 変数に代入 $s \in S$ を適用して得た木を P(s)とする．P(s)と木データ T との編集距離が最小になるときの P(s)と T の編集距離を，P と T の編集距離 $\mathrm{treedist}(P,T)$ と定義する．すなわち $\mathrm{treedist}(P,T) = \min_{s \in S} \mathrm{treedist}(P(s),T)$ と定義する[4]．

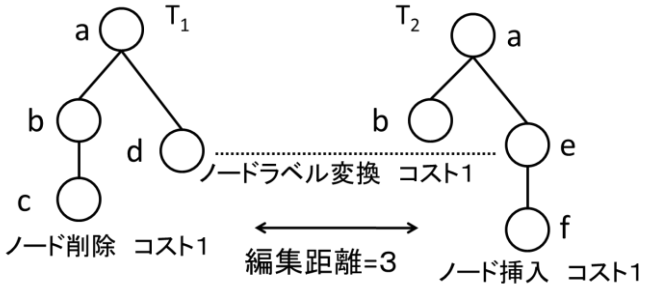

図 2.2：木 T_1 と木 T_2 の編集距離

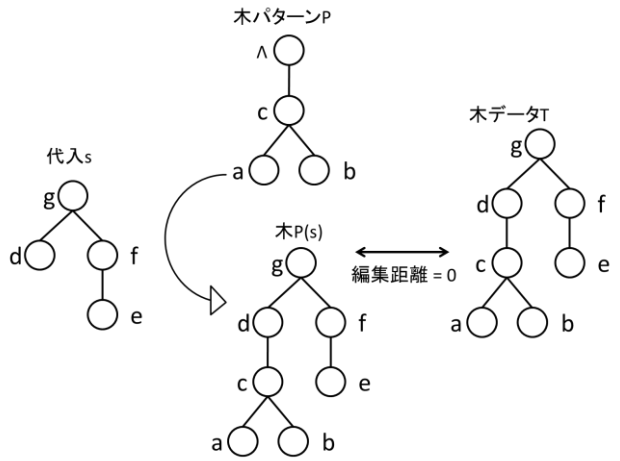

図 2.2：木パターン P とマッチする木データ T

III. 遺伝的プログラミングを用いた VLDC 付き木パターンの発見

木パターン P と木データ T の編集距離が 0 となるとき P と T がマッチするといい，これが 0 より大きいときマッチしないという．

VLDC 付き木パターン発見問題とそれに対するアルゴリズムを以下に示す．

入力：正事例，および負事例からなる木構造データの有限集合 D

問題：GP により，適合度の高い特徴的な木パターン P を発見する．

本研究では VLDC 付き木パターンを GP の個体として用いる．木パターン P の適合度 = (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 と定義する．Pos を D の正事例すべての集合とする．P の正事例距離和を P と Pos の木データとの編集距離の和，すなわち $\sum_{T \in Pos} \text{treedist}(P,T)$ と定義する．

アルゴリズム

1.正事例木データ集合から使用されているノードラベル，ラベルの上下関係，木のサイズの最大値，子の数の最大値を求める．

2.で求めた値を基にランダムに初期木パターン集合を生成する．

3.木パターンの適合度を求める．

4.場合によって以下の 4.a，4.b のいずれかの手順を行う．

4.a 適合度が大きい順に個体を並べ替える．

4.b 適合度がほぼ同じとき正事例距離和の小さい方が上位になるように並べ替える．

5.交叉，突然変異，逆位，複製の遺伝的操作により，次世代の集団を生成する．図 3.1 に交叉の例を示す．

6.終了条件である世代数まで達していれば終了．そうでなければ 5 で生成された次世代の集団を現世代の集団として 3 へ戻る．

A. 適合度計算の工夫

木データ頂点数を木データのサイズ，木パターンの VLDC 変数を除いた頂点数を木パターンのサイズと呼ぶ．木パターン P サイズが，負事例の最もサイズの大きい木データ T_M のサイズよりも大きい場合，その P はすべての負事例データにマッチしないことがわかる．これを利用して適合度の計算にかかる時間を減らすことができる．これは木パターン P のサイズが木データ T のサイズよりも大きい場合，P から T へ変換する時少なくとも 1 回は編集操作を必要とする．この時の編集距離は必ず 0 より大きくなるため，P と T はマッチしない．そのため，T_M のサイズよりも P のサイズが大きい場合，P はどの負事例のサイズよりも大きい．先ほどの理由からすべての負事例にマッチしないといえる．また，負事例とはマッチするかどうかのみを判断しているため適合度の計算にも支障はない．

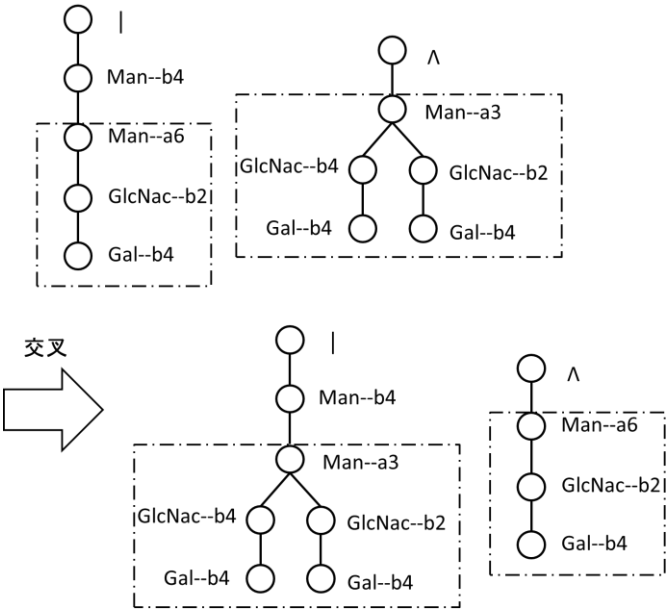

図 3.1：木パターンの交叉例

IV. 実験

実験データとして KEGG GLYCAN データベースに登録されている白血病に関する正事例データ 176 個，負事例データ 304 個を用いる．GP のパラメータは次の通りである．

表 4.1：GP のパラメータ

個体数	50
交叉確率	0.7
突然変異確率	0.1
逆位確率	0.1
複製確率	0.1
次世代の選択方法	ルーレット選択，トーナメント方式(サイズ 4)，エリート保存(サイズ 10)
最大世代数	200

GP のパラメータは同じにして，GP の選択において Ⅲ 章で記したアルゴリズム 4.1 を用いた場合と 4.2 を用いた場合で、適合度の高い特徴的な木パターンを発見する実験を行った．

実験 A：4.1 を用いて、GP の選択で適合度のみで個体を並べ替えた場合
実験 B：4.2 を用いて、GP の選択で適合度と正事例距離和を用いて個体を並べ替えた場合

それぞれの実験での，各世代で最も適合度の高い個体の適合度の 30 試行の平均値の推移を図 4.1，4.2 に示す．各試行の最終世代での最も適合度の高い個体の適合度の 30 試行の平均値と 1 試行の平均実行時間を表 4.2 に示す．また，それぞれの実験の 30 試行での，世代ごとのエリート保存された個体を除いた個体の，サイズ(VLDC 変数を除いた頂点数)の 30 試行の平均値の推移を図 4.3 に示す．

実験 B において 30 試行の最終世代の最も適合度の高い個体（最良個体）を図 4.4 に示す．また，その最良個体の適合度を表 4.3 に示す．最良個体にマッチするような木データを図 4.5 に示す．図 4.1，4.2 のグラフと表 4.2 から適合度に関しては実験 A，B ともにほぼ同じ結果が得られた．

サイズは図 4.3 からわかるように正事例距離和を考慮した実験 B の方が考慮しない実験 A より小さく，わずかだが実行時間も少なくなった．これらの結果からどちらの実験でも同じように最終世代の個体は適合度が高いことがわかる．よって，正事例に多くマッチし，負事例にほとんどマッチしない木パターンを獲得することができたといえる．

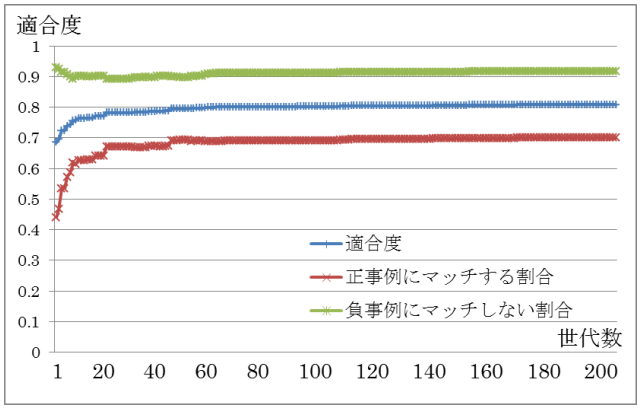

図 4.1：実験 A の適合度の推移

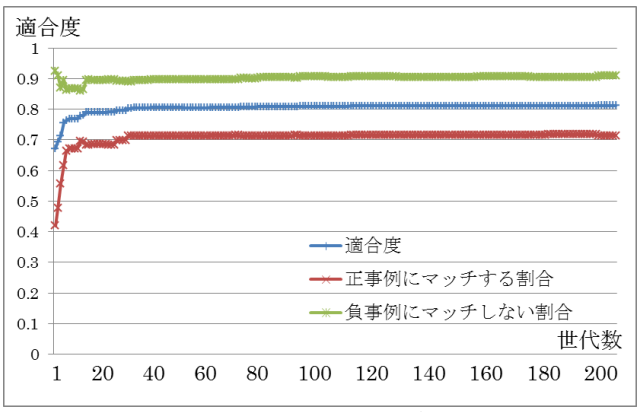

図 4.2：実験 B の適合度の推移

表 4.2：実験 A，実験 B の適合度と実行時間

	実験 A	実験 B
正事例にマッチする割合	0.700	0.714
負事例にマッチしない割合	0.916	0.909
適合度	0.808	0.811
実行時間	8,835 (秒)	8,089 (秒)

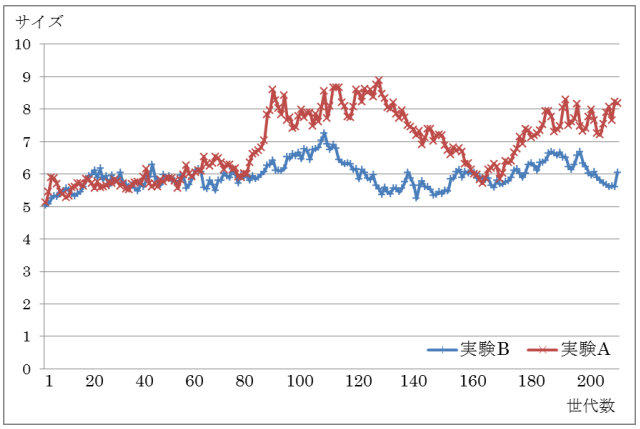

図 4.3：実験 A と実験 B の個体のサイズの推移

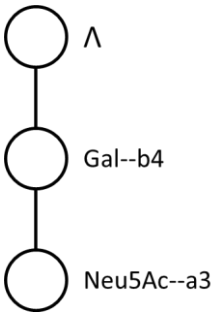

図 4.4：実験 B における最良個体

表 4.3：実験 B における最良個体の適合度

	実験 B の最良個体
正事例に マッチする割合	0.727
負事例に マッチしない割合	0.925
適合度	0.826

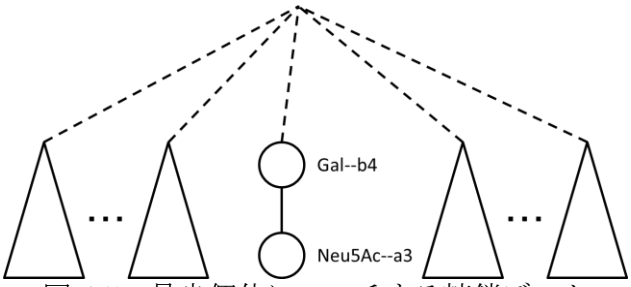

図 4.5：最良個体にマッチする糖鎖データ

V. 考察

実験 B において実行時間が減少する理由は次のように考えられる．この GP の実行時間の大部分は適合度の計算時間で占められている．その適合度の計算時間は，木のサイズと木の深さの最小値によって変化する．

GP の選択で正事例距離和を考慮することで木データ集合内の木データと形の似ていない木パターンが次世代に残りにくくなる．特に，サイズが大きい木パターンがそうなる．その理由として，サイズの小さい木パターンは VLDC 変数への代入を行うと木データと形が似たものとなる可能性がある．この時正事例距離和は小さくなる．しかし，サイズが大きい木パターンは VLDC 変数への代入を行っても木データ全体に対して形が似ない可能性が高い．この時の正事例距離和は大きくなる．これらの理由から，適合度の計算に時間のかかる形の似ていないサイズの大きな木パターンが減っていくため，全体の計算時間が減ると考えられる．

VI. 終わりに

本研究では編集距離と GP を利用して正事例データと負事例データから適合度の高い特徴的な VLDC 付き木パターンを獲得する手法を提案した．実験では糖鎖データを対象に GP の選択において適合度のみで個体を並べ替える実験 A と，適合度のみではなく正事例距離和も考慮して個体を並べ替える実験 B で適合度の高い特徴的な VLDC 付き木パターンを獲得する実験を行った．その結果，最終世代の適合度はどちらの実験でもほぼ同じくらい高い値が得られたが、実験 B の方が各世代での個体のサイズの平均値が実験 A に比べて抑えられるため，わずかだが実行時間が短くなるという結果が得られた．

参考文献

[1] 川上浩司　編,進化技術ハンドブック〈第 1 巻〉基礎編,第 4 章,近代科学社,2010

[2] M. Nagamine et al., "A Genetic Programming Approach to Extraction of Glycan Motifs Using Tree Structured Patterns," Proc. AI 2007, Lecture Notes in Artificial Intelligence, Springer-Verlag vol.4830, pp.150-159, 2007

[3] K.Zhang and D. Shasha “Simple Fast Algorithms for the Editing Distance between Trees and Related Problems,” SIAM J.Computing, Vol.18, No.6, pp.1245-1262, 1989

[4] K.Zhang and D. Shasha and J. Wang “Approximate Tree Matching in the Presence of Variable Length Don’t Cares,” Journal of Algorithms Vol.16, No.1, pp.33-66, 1994

問い合わせ先
〒731-3194
広島市安佐南区大塚東 3 丁目 4 番 1 号
広島市立大学情報科学研究科知能工学専攻
中居　翔平