科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２５　日現在

機関番号：１４３０１
研究種目：若手研究(B)
研究期間：2011～2013
課題番号：２３７００００１１

研究課題名（和文）Ｃｏｍｐａｒｉｎｇ　ａｎｄ　Ｃｏｍｂｉｎｉｎｇ　Ｔｒｅｅｓ

研究課題名（英文）Comparing and Combining Trees

研究代表者
ジャンソン　ジェスパー（Jansson, Jesper）

京都大学・白眉センター・特定准教授

研究者番号：６０５３６１００

交付決定額（研究期間全体）：（直接経費）　　2,600,000円、（間接経費）　　780,000円

研究成果の概要（和文）：この研究プロジェクトの目的は、巨大なデータ集合に関わる木構造を比較したり統合したりする効率的なアルゴリズムを開発することであり、特に以下の３つのトピックに焦点があてられていた。(i)最小の上位木を構成する。(ii)合意木を構築する。(iii)２つの木様構造の類似度を測定する。
いくつかの新しい理論的なアルゴリズムがデザインかつ分析され、関連する計算複雑さへの結果が得られた。

研究成果の概要（英文）：The purpose of this research project was to develop efficient algorithms for comparing and combining trees involving huge datasets. It focused on three particular topics: (i) Building a minimally resolved supertree; (ii) Constructing a consensus tree; and (iii) Measuring the similarity of two treelike structures.

Some new theoretical algorithms were designed and analyzed, and a number of related computational complexity results were obtained.

研究分野：総合領域

科研費の分科・細目：情報学・情報学基礎

キーワード：アルゴリズム　計算複雑さ　系統樹

１．研究開始当初の背景

木は階層構造の情報を表現するための基本的なデータ構造である。木の基本的なタイプのひとつは順序木であり、各ノードの子の順序は左から右へ固定されている。興味深いことに、一見まったく異なるように見えるテキスト処理、データアーカイブ、電子 XML 文書の更新検出、コンパイラの最適化、ソフトウェア保守、モチーフと呼ばれる RNA 分子の２次構造の共通箇所の発見、などの問題が、２つの順序木の類似度を測る動的計画法に基づくアルゴリズムで解くことができる。順序木以外の特殊なタイプの木構造としては系統樹が知られており、進化の歴史を記述するために 1850 年代より科学者達に用いられてきた。その核となるアイディアは、葉を生物種や自然言語などの研究対象に対応させる一方、木の内部頂点が共通祖先に対応するような分岐構造を選択する。多数の葉を含む信頼度の高い系統樹を構築することは、その背景にある最適化問題の計算複雑さが原因で難しい。まとめると、科学や情報処理の、様々な現実的状況において、木構造を扱うための汎用的な方法が求められていた。

２．研究の目的

この研究プロジェクトの目的は、巨大なデータ集合に関わる木構造を比較したり統合したりする効率的なアルゴリズムを開発することであり、特に以下の３つのトピックに焦点があてられていた。

(i)　　最小の上位木を構成する。
(ii)　　合意木を構築する。
(iii)　　２つの木様構造の類似度を測定する。

生物学者は時として、大きな系統樹を構築しなければならないという困難な仕事に直面する。近年、人気が増している上位木によるアプローチでは、分割統治による上位木が用いられ、最初、大きな計算量を伴う手法で小さく正確な木を部分的にオーバーラップする葉集合の部分集合に対し構築し、組み合わせアルゴリズムを用いて小さな木を統合して上位木と呼ばれる１つの大きな木を得る。この方法に対するよくある批判は、上位木が示す進化関係は、どの与えられた木からも直接的にはサポートされておらず、「偽の新規生物群」を意味する誤ったグループを作りだしている、という内容である。それゆえ、そのような誤りを必要最小限に抑えたうえで合意木を構成するのは自然な発想である。この理由により、トピック(i)では最小の合意木、すなわち内部頂点数を可能な限り少なくし、かつ入力の木と矛盾が無いようにする。このトピックは、これまでのアルゴリズム研究において無視されていた重要な問題に対処することになると考えられる。

関連する問題として、葉のラベル集合は同じだが分岐構造の異なる系統樹の集合が与えられた時に、どのようにしてそれらを統合して１つの木にするのが最善であろうか？そのような木を合意木と呼ぶ。合意木は、異なるデータ集合や異なる木構造の推定方法が、葉に対する同じラベルを持つものの少し異なる構造を生成した時に、それらを１つの木で表現するべき時に使われる。異なる分岐情報は様々に解釈することができ、合意木に対する多くの異なる定義が提案され、何年にもわたり文献上で分析されてきた。トピック(ii)の目標は、*R**合意木のような、あるタイプの合意木を構成する高速アルゴリズムを開発することであった。

上述のように、与えれた２つの木の間の類似度の測定は、多くの場面で直面する問題である。バイオインフォマティクスにおいて、種々の木構成方法による構造の正確さを査定したり、合意木を評価したり、実験で得られた木に信頼性に関する情報を付加したりすること等は必要である。類似度を測定するための簡単で直観的な方法は、根付き三つ組み距離を用いて、ちょうど３つの同じ葉集合を持ちながら、その構造が２つの木の間で異なる系統樹の数を数えることである。その他のよく使われる尺度である *Robinson-Foulds* 距離では、異なるクラスタ数を数える。トピック(iii)では、そのような距離を効率的に計算する方法を研究する。

３．研究の方法

我々の結果を得るために、再帰、ボトムアップ的な木のなぞり、２項探索、２色塗り分け、行列の掛け算、基数ソート、組み合わせ問題間の尺度保存帰着のような、よく知られているアルゴリズム的技術を参考にした。一方で、それらに比べるとあまり知られていない、セントロイド経路分解や、木の分離、直交範囲を数える方法、Day のアルゴリズム、レベル祖先、Apresjan クラスタ、理想系統ハプロタイプなどの技術や概念も用いた。

研究トピック(i)に対しては、1981 年に Aho, Sagiv, Szymanski, Ullman により発明された BUILD という名前の古典的なアルゴリズムがあり、与えられた矛盾しない根付き系統樹を統合することができる。その簡便性と効率性により、矛盾する木を統合する上位木を用いる既存の多くの方法で使われている。出発点として、我々は BUILD により構成された木の分岐構造を詳細に研究した。

(ii)において、R*合意木を計算する高速アルゴリズムを得るために、、葉のペアの最も低い共通祖先への葉からの距離に基づく定式化を用いて、木に埋め込まれた根付き３つ組の集合をコンパクトに記号化した。

また、それぞれの木において同じ葉が複数回現れることができるマルチラベル木に対し

ても、合意木の研究を行った。この研究の主な難しい点は、対応するクラスタのコレクションがもはや集合ではなく、マルチ集合であり、ある主な問題はマルチ集合に対し NP 困難になることである。それゆえ我々は最初に、情報に富むある種の合意木をマルチラベル木に対し特定することにより効率的なアルゴリズムを構築した。

(iii)において我々は、根付三つ組距離を計算する問題を、無向枝色付きグラフにおける単色かほぼ単色の三角形を数える問題に変換した。グラフにおける異なる種類の色付き三角形を効率よく数えるために、行列積に関する既存のアルゴリズムを拡張した。

３．研究成果

トピック(i): n を葉の数とした時に、BUILD アルゴリズムは、必要よりも n/2 倍も多くの内部頂点を含む系統樹を出力するかもしれないという驚くべき事実を発見した。この結果、BUILD に基づくいかなる上位木による方法も膨大な数の間違ったグループ分けを生じさせるかもしれないという、悲観的な観測が成り立つ。我々は P=NP でなければ、根付き三つ組集合を入力とする最小の合意木を推定する問題は、いかなる 0 < c <= 1 に対しても n^{1-c}以内で多項式時間アルゴリズムによる近似ができないことを証明した。しかも、処理の間に葉のあるブロックを統合することを助ける、最適グラフ塗り分けアルゴリズムを伴う増加する BUILD は、最悪ケースでは出力の解ををそれほど改良しないことを証明する反例を見つけた。楽観的な面としては、いくつかの特殊な場合に問題が素早く解けることを示した。

トピック(ii):
２つの入力の系統樹に対する R*合意木を３乗より速く計算するアルゴリズムを開発した。これは既存のアルゴリズムに対する大幅な改良である。なぜなら３乗の数の要素を含むかもしれない集合に関して R*合意木は定義されるので、この集合を数えるのを明示的に避けることにより我々の方法は速くなるのである。また、マルチラベル木に対する合意木を構築する初めての多項式時間アルゴリズムをデザインした。

トピック(iii): 入力が２つのいわゆる galled 木で構成される時、根付き三つ組距離はナイーブなアルゴリズムでは O(n^3)時間で計算できる。我々はこれを o(n^2.688)時間に改良した。副産物として、ソーシャルネットワークやその他の応用にとって興味があるかもしれない、三角形を数える問題に対するいくつかのアルゴリム的結果を得た。例えば、o(n^1.686)の枝をもつ疎な入力グラフにおいて、三角形の数を数える問題に対する現時点での世界記録を我々の方法は保持している。より正確には、その計算時間は m を枝の数とすると、o(m^1.488)である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 6 件）

[1] J. Jansson and A. Lingas: Computing the Rooted Triplet Distance between Galled Trees by Counting Triangles, *Journal of Discrete Algorithms*, Vol. 25, pp. 66-78, 2014.
doi:10.1016/j.jda.2013.10.002

[2] J. Jansson and W.-K. Sung: Constructing the R* Consensus Tree of Two Trees in Subcubic Time, *Algorithmica*, Vol. 66, Number 2, pp. 329-345, 2013.
doi:10.1007/s00453-012-9639-1

[3] T. Asano, J. Jansson, K. Sadakane, R. Uehara, and G. Valiente: Faster computation of the Robinson-Foulds distance between phylogenetic networks, *Information Sciences*, Vol. 197, pp. 77-90, 2012.
doi:10.1016/j.ins.2012.01.038

[4] Y. Cui, J. Jansson, and W.-K. Sung: Polynomial-Time Algorithms for Building a Consensus MUL-Tree, *Journal of Computational Biology*, Vol. 19, Number 9, pp. 1073-1088, 2012.
doi:10.1089/cmb.2012.0008

[5] J. Jansson, K. Sadakane, and W.-K. Sung: Ultra-Succinct Representation of Ordered Trees with Applications, *Journal of Computer and System Sciences*, Vol. 78, Number 2, pp. 619-631, 2012.
doi:10.1016/j.jcss.2011.09.002

[6] J. Jansson, R. S. Lemence, and A. Lingas: The Complexity of Inferring a Minimally Resolved Phylogenetic Supertree, *SIAM Journal on Computing*, Vol. 41, Number 1, pp. 272-291, 2012.
doi:10.1137/100811489

〔学会発表〕（計 6 件）

[1] J. Jansson, C. Shen, and W.-K. Sung: Algorithms for the Majority Rule (+) Consensus Tree and the Frequency Difference Consensus Tree, in Proceedings of the Thirteenth International Workshop on Algorithms in Bioinformatics (WABI 2013), *Lecture Notes in Computer Science*, Vol. 8126, pp. 141-155, Springer-Verlag,

2013.
Nice, France; 2013-09-02 -- 2013-09-04.

[2] J. Jansson, C. Shen, and W.-K. Sung: An Optimal Algorithm for Building the Majority Rule Consensus Tree, in Proceedings of the Seventeenth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2013), *Lecture Notes in Computer Science*, Vol. 7821, pp. 88-99, Springer-Verlag, 2013.
Beijing, China; 2013-04-07 -- 2013-04-10.

[3] J. Jansson, C. Shen, and W.-K. Sung: Improved Algorithms for Constructing Consensus Trees, in Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2013), pp. 1800-1813, Society for Industrial and Applied Mathematics (SIAM), 2013.
New Orleans, U.S.A.; 2013-01-06 -- 2013-01-08.

[4] J. Jansson and A. Lingas: Computing the Rooted Triplet Distance between Galled Trees by Counting Triangles, in Proceedings of the Twenty-Third Annual Symposium on Combinatorial Pattern Matching (CPM 2012), *Lecture Notes in Computer Science*, Vol. 7354, pp. 385-398, Springer-Verlag, 2012.
Helsinki, Finland; 2012-07-03 -- 2012-07-05.

[5] Y. Asahiro, J. Jansson, E. Miyano, and H. Ono: Upper and Lower Degree Bounde d Graph Orientation with Minimum Penalty, in Proceedings of Computing: the Eighteenth Australasian Theory Symposium (CATS 2012), Australian Computer Science Communications, Vol. 34, Number 8, pp. 139-146, Australian Computer Society Inc., 2012.
Melbourne, Australia; 2012-01-30 -- 2012-02-02.

[6] Y. Cui, J. Jansson, and W.-K. Sung: Algorithms for Building Consensus MUL-trees, in Proceedings of the Twenty-Second International Symposium on Algorithms and Computation (ISAAC 2011), *Lecture Notes in Computer Science*, Vol. 7074, pp. 744-753, Springer-Verlag, 2011.
Yokohama, Japan; 2011-12-05 -- 2011-12-05.

〔その他〕
ホームページ等

http://sunflower.kuicr.kyoto-u.ac.jp/~jj/

６．研究組織
(1)研究代表者
ジャンソン　ジェスパー（Jansson，Jesper）
京都大学・白眉センター・特定准教授
研究者番号：60536100