様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　５　月　１０　日現在

機関番号：　３２６５７
研究種目：　基盤研究(C)（一般）
研究期間：　2012～2016
課題番号：　２４５０００２３

研究課題名（和文）計算困難な問題への科学と工学の両面からのアプローチ

研究課題名（英文）Scientific and Practical Approaches to Computationally Hard Problems

研究代表者
陳　致中（Chen, Zhi-Zhong）

東京電機大学・理工学部・教授

研究者番号：００２４２９３３

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：主にバイオインフォマティクス分野の重要な組み合わせ最適化問題（生物系統樹のrSPR距離問題，網状ネットワークの構築問題，中心列問題など）について研究を行った．どの問題も計算困難で今まで盛んに研究され，たくさんのアルゴリズムが提案されていた．本研究で，これらの問題を解く新しいアルゴリズムを設計してその性能を理論的に解析するだけでなく，アルゴリズムを実装して実際のデータに対する性能を計測して先行研究のアルゴリズムと比較検証した．その結果，先行研究で提案されたアルゴリズムより理論的性能がよいだけでなく，実際のデータに対する性能もよいアルゴリズムを設計できたことが分かった．

研究成果の概要（英文）：We mainly studied several important and fundamental combinatorial optimization problems in bioinformatics such as the rSPR distance problem of two phylogenetic trees, the reticulate network problem of multiple phylogenetic trees, and the closest string problem. All the studied problems are computationally hard and various algorithms had been designed for them before. In this project, we designed new algorithms for the problems and not only rigorously analyzed the theoretical performance of our new algorithms but also implemented them and compared their practical performance against the previous bests. Our results show that the new algorithms are the bests not only in theory but also in practice.

研究分野：　Algorithm Science and Engineering

キーワード：　近似アルゴリズム　固定パラメータアルゴリズム　バイオインフォマティクス　組み合わせ最適化　ハプロタイプアセンブリー　網状ネットワーク　生物系統樹のrSPR距離　中心列問題

１．研究開始当初の背景

計算困難な問題に対して今まで行ってきた研究をアルゴリズム科学とアルゴリズム工学に大別できる．前者では，アルゴリズムを設計してその理論的性能を解析して理論的保障を証明するが，実際のデータに対する性能を検証しない．一方，後者では，アルゴリズムを設計して実装した後，結果のプログラムに実際のデータを与えてその性能を計測するが，アルゴリズムの性能を理論的に解析しない．両者にそれぞれ短所と長所があり，両者を融合することによって理論的保障付きの実用的なアルゴリズムを開発することが望ましい．しかし，このような研究が今まであまり行われてこなかった．

２．研究の目的

本研究の目的は様々な計算困難な問題を解くための理論的保障付き実用的なアルゴリズムを設計・解析・実装することである．具体的には、主にバイオインフォマティクスにおける重要で計算困難な問題を選定して、それらを解くアルゴリズムを設計してから理論的にその性能を解析するだけでなく、アルゴリズムを実装して模擬データだけでなく生物学者によって得られた実データに対してその性能を計測して評価する。目的として、理論的にも実際のデータに対しても高性能が保証されているようなソフトウエアを生物学者に提供して、満足に使ってもらえるようにしたい。

３．研究の方法

本研究者が以前の研究で近似・乱択・並列化の３つのアプローチを融合して計算困難な問題の計算限界を打破するという混成アプローチを提唱し、いくつもの重要で計算困難な問題に対するアルゴリズムを設計した。本研究で、今までの研究で採用してきた「近似・乱択・並列化を融合した混成アプローチ」をさらに発展させるために、「固定パラメータ化」というアプローチも取り入れる。一般的に、ある最適化問題が計算困難であれば、それを厳密に解くアルゴリズムの時間量は最適解の値に指数的に依存する。その依存度が低く、かつ、実際の応用に現れる実例の最適解の値が小さければ、その指数時間アルゴリズムはその最適化問題にまだ有効である。これは「固定パラメータ化」というアプローチの考え方である。このアプローチはバイオインフォマティクスに現れる一部の最適化問題に有効であることが知られている。本研究では、近似・乱択・並列化を融合した混成アプローチに固定パラメータ化を取り入れて、バイオインフォマティクスに現れるより多くの最適化問題の実用的なアルゴリズムを設計することを目指した。

また，一部の組み合わせ最適化問題の整数線形計画モデルを設計して優秀なソルバー（CPLEX や GUROBI など)を使って解くようにした．さらに，現代のコンピュータが複数のコアを持っているのでアルゴリズムを並列化して計算速度を上げるようにした．

４．研究成果

下記の通り，7 つの計算困難な問題について研究を行い，予想以上の成果をあげることができた．

(1) 中心列問題について．

まず，「3-string アプローチ」という斬新な手法に基づく固定パラメータアルゴリズムを新たに設計して解析した．先行研究で提案された最良のアルゴリズムが入力文字列からハミング距離が最も大きい 2 本を選ぶことによって計算を始めるのに対し，新しいアルゴリズムがその 2 本の入力文字列以外にもう 1 本の入力文字列をうまく選ぶことによって計算を始める．このように変更を行うと，時間量の解析でよりタイトな上界を証明することが可能になる．結果として，新しいアルゴリズムは入力文字列のアルファベットが小さいときに先行研究で提案された最速のアルゴリズムより（少なくとも）理論上かなり速い．たとえば，入力文字列が 2 進列の場合，先行研究の最良アルゴリズムの時間量は $O(nL+nd8^d)$であるのに対し，新しいアルゴリズムの時間量は $O(nL+nd^3 6.731^d)$である．また，入力文字列が DNA 配列の場合，先行研究の最良アルゴリズムの時間量は $O(nL+nd\cdot 13.921^d)$であるのに対し，新しいアルゴリズムの時間量は $O(nL+nd\cdot 13.183^d)$である．

次に，乱拓を利用して中心列問題をより速く解くことができるかについて研究を行った．具体的には，中心列問題を解くための乱拓固定パラメータアルゴリズムをいくつか設計して解析した．1 つ目の乱拓アルゴリズムの期待時間量は $O^*(2.5^d\sigma^d)$である．ここで，σ は入力文字列のアルファベットのサイズである．σ=2 のとき，$O^*(2.5^d\sigma^d)$が $O^*(5^d)$となるので，先行研究で提案された最良のアルゴリズムの時間量 $O^*(6.731^d)$より遥かによい．また，σ=4（すなわち，DNA 配列）のとき，$O^*(2.5^d\sigma^d)$が $O^*(10^d)$となるので，先行研究で提案された最良のアルゴリズムの時間量 $O^*(13.18^d)$より遥かによい．2 つ目の乱拓アルゴリズムは 1 つ目に対して，アルファベットのサイズが大きいときに威力を発揮する．その期待時間量が $O^*((2\sigma+4)^d)$である．σ=20（すなわち，蛋白質）のとき，$O^*((2\sigma+4)^d)$が $O^*(44^d)$となるので，先行研究で提案された最良のアルゴリズムの時間量 $O^*(47.21^d)$より遥かによい．3 つ目の乱拓アルゴリズムは 1 つ目と 2 つ目を融合したもので，両方の優れた点を持っている．その期待時間量は，σ=4 のとき $O^*(9.81^d)$，σ=20 のとき $O^*(40.09^d)$である．したがって，新しく設計した乱拓アルゴリズムは 2 進列，DNA 配列，蛋白質の 3 つの重要な場合において，先行研究で提案した固定パラメータアルゴリズムより（少なくと

も）理論上遥かに速いことが分かった．

中心列問題の既知のアルゴリズムの中に理論上遅いが実際のデータに対して高速なアルゴリズムも開発されている．このようなアルゴリズムを発見的アルゴリズムと呼ぶことにする．本研究で設計した「3-string アルゴリズム」と乱拓アルゴリズムを発見的アルゴリズムと比較するため，「3-string アルゴリズム」と乱拓アルゴリズムを実装した．実際のデータに対する性能を比べた結果，「3-string アルゴリズム」は最も速いことが分かった．ということで，「3-string アルゴリズム」は理論上だけでなく，実際の応用上最も速いので，理論的保障付きの実用的アルゴリズムであることが言える．

(2) 共有中心列問題について．

この問題を解くため固定パラメータアルゴリズムを２つ設計した．１つ目の時間量が $O(m^3L(n-k)+ m^2Lk + kd(6\sqrt{3})^d m^{\log(d+1)+2})$で，２つ目の時間量が $O(m^3L(n-k)+ m^2Lk + k^2d8^d m^{\log(d+1)+2})$である．さらに，有名な ETH 予想が成り立つ限り，その時間量にある指数 $\log(d+1)$を $o(\log d)$に改善できないことを証明した．ということで，２つのアルゴリズムとも漸近的に最適であると言える．

また，２つのアルゴリズムとも実装して，模擬データだけでなく国際 HapMap プロジェクトで得られた実データを用いてその性能を計測した．その結果，両方とも以前の近似アルゴリズムよりも遺伝子の突然変異領域の探知に有効であることを確認できた．そのついでに，１つ目のアルゴリズムは２つ目よりかなり高速であることも確認できた．

(3) rSPR 距離の計算問題について．

２本の生物系統樹の rSPR 距離を求める新しい近似アルゴリズムと固定パラメータアルゴリズムを設計・解析・実装した．先行研究で提案された最速の固定パラメータアルゴリズムの時間量 $O(2.415^d n)$を改善して，本研究で時間量 $O(2.344^d n)$の固定パラメータアルゴリズムを設計した．先行研究で Schalekamp らによって提案された最良の近似アルゴリズムが近似率２を達成するが時間量が厳密に解析されていなく，単に多項式時間だと述べられている．一方，本研究で新たに提案された近似アルゴリズムが近似率２を達成するだけでなく，時間量が三次である．また，Schalekamp らの近似率の解析に線形計画問題の双対性理論が使われ，直感的に分かりづらい．それに対して，本研究のアルゴリズムの近似率の解析は純粋に組み合わせ的であるので，直感的に理解することが可能である．さらに，両方のアルゴリズムとも rSPR 距離の上界と下界を出力するが，模擬データに対して計測した結果，本研究のアルゴリズムの方がより小さい上界とより大きい下界を出力することを確認した．

新しい近似アルゴリズムを枝刈りに利用して rSPR 距離を厳密に求める新しい固定パラメータアルゴリズムを実装した．新しい近似アルゴリズムがよい下界を出力するので枝刈りにかなり有効で，Whidden らによる既知の最速アルゴリズムよりも高速に rSPR 距離を求めることができることを確認した．ということで，本研究の $O(2.344^d n)$時間アルゴリズムは理論的保障付きの実用的アルゴリズムである．

(4) 網状ネットワークの構築問題について．

この問題を解く新しいアルゴリズムを設計して実装した．そのアルゴリズムが分枝限定法に基づいている．その主なアイディアは，rSPR 距離問題の近似アルゴリズムまたは厳密アルゴリズムを探索木の枝刈りに利用したところにある．模擬データと実データに対して本研究で得たプログラムの性能を計測した結果，その速度が以前のソフトウエア（HybridNet や Dendroscope3）よりはるかに速いことが分かった．

(5)ハプロタイプ組み立て問題について．

この問題を解くための厳密アルゴリズムを設計して実装した．先行研究で提案された最良の厳密アルゴリズムが入力行列をブロックに分けてから，各ブロックに関する問題を Max-Cut 問題かまたは Max-SAT 問題に帰着して解いた．本研究では，この問題の新しい性質を証明して利用し，ブロックをさらにサブブロックに分割できることを証明した．その結果，元の大きな問題をより小さい部分問題に変換でき，より高速に解けることが分かった．また，各サブブロックに関する問題を整数線形計画問題（ILP）に帰着して，優秀な ILP ソルバー（CPLEX や GUROBI など）を用いて解いた．結果として，以前のアルゴリズムは最適な解を出力できないかまたは PC クラス上でも長い時間を必要としたのに対し，本研究で得たアルゴリズムが１台の PC 上でも高速に最適解を出力できることが分かった．

最適な解の計算に長時間がかかってしまう一部の難しいサブブロックに関する問題を近似的に解くための heuristics も複数設計してその実用性を模擬データと実データの両方で検証した．さらに，どの heuristic も改善できる手法を３つ提案した．提案手法では，まず heuristic を用いて近似解（ハプロタイプ h と入力行列の行の２分割 P）を求めておく．その後，１つ目の手法では h のビットの中から信頼度の高いビットを固定して残りのビットを ILP ソルバーで計算しなおす．一方，２つ目の手法では，重なりが多くて重なりの中に共通部分が多い２つの行が P の同じグループに入っているならばそうなるように固定し，重なりが多くて重なりの中に共通部分が少ない２つの行が P の異なるグループに入っているならばそうなるように固定してから，入力行列の行を ILP ソルバー

で再分割する．3 つ目の手法は単に 1 つ目と 2 つ目を融合したものである．検証の結果，1 つ目の手法が最もよい近似解を出力することを確認できた．

(6)Scaffold の構築問題について．
ゲノム組み立ての際，paired-end リードから組み立てた contig をより長い scaffold にマージする必要がある．この問題を一種の巡回セールスマン問題として定式化できる．この問題が NP 困難であるので，近似アルゴリズムが先行研究で設計されていた．その中で最良の近似アルゴリズムが近似率 0.5 しか達成しなかった．本研究でよりよい近似率(5-5ε)/(9-8ε)を達成する近似アルゴリズムを設計できた．ここで，ε は 0 より大きく 1 より小さい任意の定数である．
Chateau らがこの問題を拡張してから，ある条件の下で近似率 1/3 を達成する近似アルゴリズムを設計した．その条件を外しても近似率 1/3 を達成できるかは未解決問題であった．本研究でその条件を外せることを示して，未解決問題を解くことに成功した．さらに，Chateau らの条件よりも緩い条件の下で，1/3 よりよい近似率 0.5 を達成する近似アルゴリズムを設計した．条件をもう少し厳しくすれば，近似率(5-4ε)/9 や(7-6ε)/13 などを達成する近似アルゴリズも設計できることを証明した．
新しく設計したアルゴリズムを実装して実際のデータに対する性能を検証したところ，先行研究で提案されたアルゴリズムよりよい近似解を出力することを確認した．

(7)最多内部頂点スパニング木問題について．
これは与えられた無向グラフから最多の内部頂点を持つスパニング木を求める問題である．この問題は古典の巡回セールスマン問題の拡張だけでなく，水道システムの構築への応用も知られているが，NP 困難であるため今まで近似アルゴリズムと固定パラメータアルゴリズムがたくさん設計されてきた．本研究以前最良の近似アルゴリズムが近似率 3/4 を達成していた．本研究で，そのアルゴリズムの本質を見出して大幅に簡単化した後，新しいアイディアを注入することによってよりよい近似率 13/17 を達成する新しいアルゴリズムを設計して解析した．今後このアルゴリズムを実装して，実際の性能を検証したい．

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 15 件）

① Zhi-Zhong Chen, Y. Harada, F. Guo, and L. Wang. Approximation algorithms for the scaffolding problem and its generalizations. Theoretical Computer Science, to appear.
② Zhi-Zhong Chen, Q. Feng, C. Shen, J. Wang, and L. Wang. Algorithms for pedigree comparison. IEEE/ACM Transactions on Computational Biology and Bioinformatics, to appear.
③ Zhi-Zhong Chen, F. Deng, C. Shen, Y. Wang, and L. Wang. Better ILP-based approach to haplotype assembly. Journal of Computational Biology, Vol. 23, No. 7, pp. 537-552, 2016. DOI: 10.1089/cmb.2015.0035
④ Zhi-Zhong Chen, B. Ma, and L. Wang. Randomized fixed-parameter algorithms for the closest string problem. Algorithmica, Vol. 74, No. 1, pp. 466-484, 2016. DOI: 10.1007/s00453-014-9952-y
⑤ Zhi-Zhong Chen, Y. Fan, and L. Wang. Faster exact computation of rSPR distance. Journal of Combinatorial Optimization, Vol. 29, No. 3, pp. 605-635, 2015. DOI: 10.1007/s10878-013-9695-8
⑥ Zhi-Zhong Chen, Y. Fan, and L. Wang. Parameterized and approximation algorithms for finding two disjoint matchings. Theoretical Computer Science, Vol. 556, pp. 85-93, 2014. DOI: 10.1016/j.tcs.2014.03.030
⑦ Zhi-Zhong Chen, W. Ma, and L. Wang. The parameterized complexity of the shared center problem. Algorithmica, Vol. 69, No. 2, pp. 269-293, 2014. DOI: 10.1007/s00453-012-9730-7
⑧ Zhi-Zhong Chen, F. Deng, and L. Wang. Exact algorithms for haplotype assembly from whole-genome sequence data. Bioinformatics, Vol. 29, No. 16, pp. 1938-1945, 2013. DOI: 10.1093/bioinformatics/btt349
⑨ Zhi-Zhong Chen, and L. Wang. An ultrafast tool for minimum reticulate networks. Journal of Computational Biology, Vol. 20, No. 1, pp. 38041, 2013. DOI: 10.1089/cmb.2012.0240
⑩ Zhi-Zhong Chen, L. Wang, and S. Yamanaka. A fast tool for minimum hybridization networks. BMC Bioinformatics, 13:155, 2012. DOI: 10.1186/1471-2105-13-155
⑪ Zhi-Zhong Chen, F. Deng, and L. Wang. Simultaneous identification of duplications, losses, and lateral gene transfers. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 9, No. 5, pp. 1515-1528, 2012. DOI: 0.1109/TCBB.2012.79
⑫ Zhi-Zhong Chen and L. Wang.

Algorithms for reticulate networks of multiple phylogenetic trees. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 9, No. 2, pp. 372-384, 2012. DOI: 10.1109/TCBB.2011.137

⑬ Zhi-Zhong Chen, B. Ma, and L. Wang. A three-string approach to the closest string problem. Journal of Computer and System Sciences, Vol. 78, No. 1, pp. 164-178, 2012. DOI: 10.1016/j.jcss.2011.01.003

⑭ W. Ma, Y. Yang, Zhi-Zhong Chen, and L. Wang. Mutation region detection for closely related individuals without a known pedigree. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 9, No. 2, pp. 499-510, 2012. DOI: 10.1109/TCBB.2011.134

⑮ Zhi-Zhong Chen, T. Tsukiji, and H. Yamada. Parameterized algorithms for disjoint matchings in weighted graphs with applications. IEICE Trans. Inf & Syst., Vol. 99-A, No. 6, pp. 1050-1058, 2016. DOI: 10.1016/j.jcss.2011.01.003

〔学会発表〕（計 10 件）

① Zhi-Zhong Chen, Y. Harada, and L. Wang. An approximation algorithm for maximum internal spanning tree. Proceedings of 11th International Conference and Workshops on Algorithms and Computation, Lecture Notes in Computer Science, Vol. 10168, pp. 385-396, 2017. Hsinchu, Taiwan, March 29-31, 2017.

② Zhi-Zhong Chen, E. Machida, and L. Wang. An approximation algorithm for rSPR distance. Proceedings of 22nd International Computing and Combinatorics Conference, Lecture Notes in Computer Science, Vol. 9797, pp. 4680479, 2016. Ho Chi Minh city, Vietnam, August 2-4, 2016.

③ Zhi-Zhong Chen, Y. Harada, E. Machida, F. Guo, and L. Wang. Better approximation algorithms for scaffolding problems. Proceedings of 10th International Frontiers of Algorithmics Workshop, Lecture Notes in Computer Science, Vol. 9711, pp. 17-28, 2016. Qingdao, China, June 30-July 2, 2016.

④ Zhi-Zhong Chen, Q. Feng, C. Shen, J. Wang, and L. Wang. Algorithms for pedigree comparison. 14th Asia Pacific Bioinformatics Conference, 2016. San Francisco, U.S.A., January 11-13, 2016.

⑤ Zhi-Zhong Chen, B. Ma, and L. Wang. Randomized and parameterized algorithms for the closest string problem. Proceedings of 25th Annual International Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science, Vol. 8486, pp. 100-109, 2014. Moscow, Russia, June 16-18, 2014.

⑥ Zhi-Zhong Chen, Y. Fan, and L. Wang. Parameterized and approximation algorithms for finding two disjoint matchings. Proceedings of 7th Annual International Conference on Combinatorial Optimization and Applications, Lecture Notes in Computer Science, Vol. 8287, pp. 1-12, 2013. Chengdu, China, December 12-14, 2013.

⑦ Zhi-Zhong Chen and L. Wang. Faster exact computation of rSPR distance. Proceedings of 3rd Joint International Conference on Frontiers in Algorithms and Algorithmic Aspects in Information and Management, Lecture Notes in Computer Science, Vol. 7924, pp. 36-47, 2013. Dalian, China, June 26-28, 2013.

⑧ Zhi-Zhong Chen, F. Deng, and L. Wang. Identifying duplications and lateral gene transfers simultaneously and rapidly. Proceedings of 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 128-135, 2013. Singapore, April 15-19, 2013.

⑨ Zhi-Zhong Chen and L. Wang. An improved approximation algorithm for the bandpass-2 problem. Proceedings of 6th Annual International Conference on Combinatorial Optimization and Applications, Lecture Notes in Computer Science, Vol. 7402, p. 188-199, 2012. Banff, Canada, August 5-9, 2012.

⑩ Zhi-Zhong Chen, L. Wang, and W. Ma. The parameterized complexity of the shared center string problem. Proceedings of 23rd Annual International Symposium on Combinatorial Patter Matching, Lecture Notes in Computer Science, Vol. 7345, pp. 4390452, 2012. Helsinki, Finland, July 3–5, 2012.

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ: http://rnc.r.dendai.ac.jp

６．研究組織
(1)研究代表者
陳　致中　（Zhi-Zhong Chen）
東京電機大学・理工学部・教授
研究者番号：00242933

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）