公募研究：2002年度

ゲノム配列の高次圧縮・索引構築と高次幾何構造解析による知識発見

●定兼 邦彦[1)]　◆徳山 豪[1)]　◆今井 浩[2)]　◆稲葉 真理[3)]

1) 東北大学大学院情報科学研究科　2) 東京大学大学院情報理工学系研究科　3) 東京大学大学院理学系研究科

〈研究の目的と進め方〉

ゲノム計画では様々な生物のDNA配列の解読が進んでおり，さらに類似配列の検索，DNA配列中のタンパク質のコーディング領域の推定や，RNAの二次構造，タンパク質の構造，機能の予測，進化系統樹の推定などの知識発見が行われている．これらの解析はゲノム配列のデータベースを用いて行うが，データベースのサイズは増大する一方であるため，既存のアルゴリズム・データベースの構造では限界がある．本研究では大量のデータに対しても効率のよいアルゴリズムおよび，データの効率的な格納法の研究を行う．

〈研究開始時の研究計画〉

ゲノム配列に対する索引生成とそれを用いた高速な配列検索，および配列の幾何構造解析による知識発見を行なう．大量のゲノム配列からの検索を可能にするために圧縮した索引を構築する．そしてその索引を用いて高速に配列検索を行なうアルゴリズムを開発する．

〈研究期間の成果〉

文字列を圧縮したまま高速に検索するデータ構造の研究を行った．ヒトゲノム28億塩基に対する従来の索引である接尾辞配列のサイズは約 12G バイトであり、それを約 2G バイトに圧縮可能であることは既に示しているが，その圧縮アルゴリズムは多くのメモリが必要としていた．そこで圧縮索引を作成する省スペースなアルゴリズムを考案した．これにより，ヒトゲノムに対する圧縮索引を 3G バイトのメモリで構築できるようになった．

ヒトとマウスのゲノムのアラインメントの計算など，長い配列に対するアラインメントアルゴリズムとしてMUMmerが有名であるが，接尾辞木というサイズの大きい索引が用いられており，大規模配列に対しては適応できなかった．本研究では圧縮接尾辞木を提案し，少ないメモリでも全ゲノムアラインメントの計算が行なえるようにした．

ある問合せパタンを含む遺伝子を重複なく列挙する演算は従来の索引では効率よく解けなかったが，それを可能にする省スペースな索引を考案した．そのサイズはゲノム配列のサイズの約 3 倍であり，従来手法の1/5以下のメモリで実行できる．

文字列検索を高速にするための情報であるlcp (最長共通接頭辞) の長さを高速に求めるための省スペースなデータ構造を提案した．

大量のゲノムデータを共有するための高速通信技術を開発し，大規模データ共有システムの基盤を作った．

ゲノム情報を表す関係データベースにおいて，与えられた目的関数を最適にするデータの分類を求めるアルゴリズムを開発した．また，数値属性を持つデータベースにおいて，データの数値誤差を平滑化し，予測精度を上げる手法を開発した．さらに，それを 2 次元データに拡張し，2 次元実数配列をディスクレパンシーの小さい0,1配列に丸める手法を開発した．

〈達成できなかったこと、予想外の困難、その理由〉

省スペースな索引の構築はできたが，それを用いた高速な類似配列検索アルゴリズムの開発はできなかった．これは問題自体の難しさに加え，我々の開発した索引の特徴を生かした専用アルゴリズムの開発に時間を要したからである．

〈今後の課題〉

高速な類似配列検索アルゴリズムの開発が必要である．従来手法では最悪ケースの性能を保証していないものがほとんどである．よって今後の課題はどのような配列データベースおよび問合せ配列に対してもデータベースサイズの対数多項式時間で検索が終了するアルゴリズムと，索引サイズが配列長の線型になるものの開発が挙げられる．そしてそのアルゴリズムと索引を用いてゲノム配列からの知識発見を行なうことも課題である．

〈研究期間の全成果公表リスト〉

1. 202251429. Sadakane, K.: Succinct Representations of lcp Information and Improvements in the Compressed Suffix Arrays. In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp. 225–232 (2002).
2. Hon, W.-K., Lam, T.-W., Sadakane, K, Sung, W.-K., and Yiu, S.M.: A Space and Time Efficient Algorithm for Constructing Compressed Suffix Arrays, Algorithmica, accepted.
3. Hon, W.-K., and Sadakane, K.: Space-Economical Algorithms for Finding Maximal Unique Matches. Proc. CPM 2002: 144-152
4. Sadakane, K.: Succinct Data Structures for Flexible Text Retrieval Systems, Journal of Discrete Algorithms, accepted.
5. Sadakane, K.: Compressed Suffix Trees with Full Functionality, Theory of Computing Systems, accepted.
6. Morimoto, Y., Fukuda, T., and Tokuyama, T.: Algorithms for Finding Attribute Value Group for Binary Segmentation of Categorical Databases. IEEE Trans. Knowl. Data Eng. 14(6): 1269-1279 (2002)
7. Katoh, N., and Tokuyama, T.: K-Levels of Concave Surfaces. Discrete & Computational Geometry 27(4): 567-584 (2002)
8. Asano, T., Katoh, N., Obokata, K., and Tokuyama, T.: Matrix rounding under the Lp-discrepancy measure and its application to digital halftoning. Proc. SODA 2002: 896-904
9. Chun, J., Sadakane, K., and Tokuyama, T.: Linear time algorithm for approximating a curve by a single-peaked curve, Algorithmica, 44(2):103–115, 2005.
10. Hiraki, K., Inaba, M., Tamatsukuri, J., Kurusu, R., Ikuta, Y., Koga, H., and Jinzaki, A.:Data Reservoir: Utilization of Multi-Gigabit Backbone Network for Data-Intensive Research, Proc. Super Computing 2002, High Performance Networking and Computing, (SC2002) CD-ROM , Nov., 2002.