公募研究：2000～2001年度

ゲノム配列からの高次圧縮・クラスタリングによる知識発見

●稲葉　真理[1)]　◆今井　浩[1)]　◆定兼邦彦[2)]

1) 東京大学大学院情報理工学系研究科　　2) 九州大学大学院システム情報科学研究科

〈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DNA配列中のたんぱく質のコーディング領域の推定やたんぱく質の機能予測などが行われている。また、異なる生物のDNAやアミノ酸配列の間の類似度を計算し、それによって構造や機能の予測や、進化系統樹の解析が行われている。配列の類似度は複数の配列にギャップを挿入し一致させるためのコストで定義するアラインメントによる方法が一般的には用いられているが、これは進化の過程で起こる文字列反転などを適切に扱えないという問題がある。本研究では、ある文字列の圧縮情報を用いて別の文字列の圧縮を行った場合、似ている文字列ほど圧縮率が高くなるという性質に着目し、ある配列の圧縮情報を利用して別の配列の圧縮を行った時の圧縮率を配列間の類似度として用いることで、従来のアライメントのギャップコストとは異なり文字列反転も類似と認識できる枠組みを提供し、その枠組みを利用して、クラスタリング・学習を応用することで、コーディング領域の推定や、構造・機能予測、進化系統樹の解析を行うことを目標とする。

〈研究開始時の研究計画〉

研究開始時、定兼が文字列圧縮を、稲葉・今井がクラスタリングと学習をそれぞれ担当する研究計画をたてた。具体的には、定兼が DNAやアミノ酸配列のような大規模データの圧縮を高速に行うためのアルゴリズムを研究する。稲葉・今井は、情報幾何の応用としてダイバージェンスを類似度とするクラスタリング問題について、幾何的性質を利用した高速アルゴリズムの研究開発を行い、その応用として最尤推定を高速に効率よく行うことを研究計画としていた。

〈研究期間の成果〉

文字列圧縮　DNAやアミノ酸配列のような大規模データの解析を行うために、文字列を圧縮したまま高速に検索を行うアルゴリズムを示し実装を行った。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。たとえば、ヒトゲノム２７億塩基に対して、任意のパタンを検索可能な索引である接尾辞配列のサイズは約１１Gバイトであるが、それを約２Gバイトに圧縮した。これにより、コンピュータのメモリ内に格納することが可能になり、高速検索が行えるようになった。任意のパタンの高速検索が可能な接尾辞配列の圧縮に関する研究は、当時は、新しい研究で、まだ理論的な結果のみしか示されておらず、実際に索引を構成している例は、いずれもデータ量が小さかった。本研究では、大量のデータに対して実際に索引を作成した。圧縮された索引を用いて配列を検索する場合、圧縮されていない索引を用いた場合と同じアルゴリズムが使えるが、アルゴリズムによっては速度が低下する場合がある。また、索引生成時に必要なメモリ量が大きいため、索引の生成にはスーパーコンピュータを用いたが、少ないメモリによる索引生成は今後の課題である。配列中のあいまい一致の検索は、アルゴリズムの工夫により一致する箇所の候補を減らした。

クラスタリング　ダイバージェンスを用いた幾何クラスタリング行った。クラスタリングの基準にはKolmogorov complexity の近似であるエントロピーに関係する量を用い、確率分布の間のダイバージェンスにより距離を定義した場合、ユークリッド距離の場合とほぼ同じアルゴリズムを適応することができる。したがって、情報幾何におけるDivergence Voronoi Diagram を利用した相互情報量によるクラスタリングアルゴリズムを提案した。

大規模データ共有システム　当初の研究計画にはなかった項目であるが、遠距離ネットワークを利用した大規模データ共有システムの研究を行った。ゲノム計画では、扱うデータサイズが膨大であり、データが採取される生物学的実験施設とデータ解析を行うためのコンピュータが整備されている情報科学的実験施設は、距離的に離れていることが多く、また採取された生物学的データは複数の研究グループで共有されることも多い。従ってデータ解析については、いかに共同研究者と大量データを共同研究者達と、タイムラグなしに共有できるかが一つの重要なポイントとなる。ここでは遠距離通信と近距離通信を分離し、アプリケーショントランスペアレントであり、ネットワークバンド幅とディスク容量に対してスケーラブルであるデータ利用基盤の提案を行った。

〈国内外での成果の位置づけ〉

国内では定兼が FIT2002 において、「柔軟な文書検索のためのコンパクトなデータ構造」で船井ベストペーパー賞を受賞した。国外では稲葉らが、Supercomputing 2002(Baltimore), 2003(Phoenix), 2004(Pittsburgh)の Bandwidth Challenge において、それぞれ、High Performance Bandwidth Challenge　Most Efficient Use of Available Bandwidth Award (最高効率賞), Distance x Bandwidth Product & Network Technology Award

(最高バンド幅・距離積　&　ネットワークテクノロジー賞), Single Stream, Longest Path, Standard MTU TCP Throughput Award （シングルストリーム最長パススタンダードMTU　TCP　スループット賞）を３年連続受賞した。従って、本研究で得られた成果は国内外で高い評価を受けたと言って差し支えないと思われる。

〈達成できなかったこと、予想外の困難、その理由〉

実現できなかった点は，圧縮率による類似度の定義、およびクラスタリングアルゴリズムを利用した学習方式による具体的な生物知識の発見である．

〈今後の課題〉

圧縮および検索のさらなる高速化をめざしFPGA 等ハードウェアアルゴリズムの研究を行う。具体的に状態遷移表のサイズを押さえオンチップメモリで処理が可能で複数文字を並列処理するためのアルゴリズムの提案・実

装を行いたい。

〈研究期間の全成果公表リスト〉

[1] Kunihiko Sadakane, Takumi Okazaki, and Hiroshi Imai, "Implementing the context tree weighting method for text compression", In Proceedings of the IEEE Data Compression Conference(DCC 2000), pages 123-132. IEEE Computer Society Press, March 2000.

[2] Mary Inaba and Hiroshi Imai. "Geometric Clustering for Multiplicative Mixtures of Distributions in Exponential Families." Proceedings of the 12th Annual Canadian Conference on Computational Geometry, pp.195-196, 2000.

[3] Mary Inaba Hiroshi Imai, and Kunihiko Sadakane, "Voronoi diagrams and clustering in information geometry - their computational and combinatorial complexity". In ISM Reports on Statistical Computing, volume 132, pages 74-87. The Institute of Statistical Mathematics, 2000

[4] Hiroshi Imai and Mary Inaba. "Geometric Clustering by Divergence and Its Underlying Discrete Proximity Structures." IEICE Transactions Information and Systems, Vol.E83-D, No.1, pp.27-35, 2000.

[5] Mary Inaba, Naoki Katoh and Hiroshi Imai. "Variance-Based k-Clustering Algorithms by Voronoi Diagrams and Randomization". IEICE Transactions on Information and Systems, Vol.E83-D, No.6, pp.1199-1206, 2000.

[6] T. Matsumoto, K. Sadakane, H. Imai and T. Okazaki "Can General-Purpose Compression Schemes Really Compress DNA Sequences?" In "Currents in Computational Molecular Biology" (S. Miyano, R. Shamir, and T. Takagi, eds.), Universal Academy Press, 2000, pp.76–77.

[7] K. Doi and H. Imai: "Complexity Properties of the Primer Selection Problem for PCR Experiments". Proceedings of the 5th Japan-Korea Joint Workshop on Algorithms and Computation, 2000, pp.152–159.

[8]T. Matsumoto, K. Sadakane and H. Imai: "Biological Sequence Compression Algorithms". In "Genome Informatics 2000" (A. K. Dunker, A. Konagaya, S. Miyano and T. Takagi, eds.), Universal Academy Press, 2000, pp.43–52.

[9] K. Doi and H. Imai: Sequencing by Hybridization in the Presence of Hybridization Errors. In "Genome Informatics 2000" (A. K. Dunker, A. Konagaya, S. Miyano and T. Takagi, eds.), Universal Academy Press, 2000, pp.53–62.

[10] Hiraki, K. Inaba, M., Tamatsukuri, J., Kurusu R., Ikuta, Y., Hisashi, K. and Jinzaki, A., "Data Reservoir: A 4Gbps Long Distance File Sharing Facility for Science Data Processing" Poster, SC2001, Nov. 2001.

[11] Kunihiko Sadakane and Tetsuo Shibuya, "Indexing Huge Genome Sequences for Solving Various Problems." Genome Informatics 2001, Universal Academy Press, 2002, 175-183

[12] Kunihiko Sadakane and Hiroshi Imai, "Fast Algorithms for k-Word Proximity Search." IEICE Trans. Fundamentals. E-84-A, 9 2002 2311-2318

[13]Kunihiko Sadakane, "Succinct Representation of 1cp Information and Improvements in the Compressed Suffix Arrays." Proceedings of ACM-SIAM Symposium on Discrete Algorithms. 2002, 225-232

[11] Mary Inaba, Ryutaro Kurusu, Junji Tamatsukuri, Hisashi Koga, H, Akira Jinzaki, and Kei Hiraki, "Data Reservoir: A very high-speed long distance file sharing facility for scientific data processing," Proc. High-Performance Computing Systems, IPSJ, pp.81-88, Jan 2002.

[15] Kurusu, R., Sakamoto, M., Ikuta, Y., Hiraki, K., Inaba, M., Tamatsukuri, J., Koga, H., and Jinzaki, A.,"Data Reservoir: Multi-Gigabit Data Transfer Facility, Its Design and Implementation", Proceedings of the third International Conference on Parallel and Distributed Computing(PDCAT), 2002, pp. 100-108, Sep., 2002.

[16] Kei Hiraki, Mary Inaba, Junji Tamatsukuri, Ryutaro Kurus, Yukichi Ikuta, Hisashi Koga, and Akira Jinzaki,"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.

[17] Kei Hiraki, Mary Inaba, Junji Tamatsukuri, Ryutaro Kurusu, Yukichi Ikuta, Hisashi Koga, and Akira Jinzaki. "Data Reservoir: A New Approach to Data-Intensive Scientific Computation," Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks, ISPAN 2002, pp.269-274, May 2002.

[18] Makoto Nakamura, Mary Inaba, and Kei Hiraki,"Fast Ethernet is sometimes faster than Gigabit Ethernet on LFN–Observation of congestion control of TCP streams" Proc. Int. Conf. on Parallel and Distributed Computing And Systems(PDCS2003) Nov. 2003 pp.854-859

[19] Tsuyoshi Ito, Mary Inaba,"Theoretical Analysis of Performances of TCP/IP Congestion Control Algorithm with Different Distances", Networking2004 May 2004 pp. 962-973

[20] Hiroyuki Kamezawa, Makoto Nakamura, Mary Inaba, and Kei Hiraki"Coordination between parallel TCP streams on Long Fat Pipe Network", 1st International Workshop on Data Processing and Storage Networking: towards Grid Computing(DPSN04), pp. 41-48

[21] Yutaka Sugawara, Mary Inaba, and Kei Hiraki"Over 10Gbps String Matching Mechanism for Mluti-Stream Packet Scanning System", LNCS 3203 Field-Programmable Logic and Applications, 14th International Conference, FPL2004, pp. 484-493

[22] Hiroyuki Kamezawa, Makoto Nakamura, Junji tamatsukuri, Nao Aoshima, Mary Inaba, and Kei Hiraki"Inter-layer coordination for parallel TCP streams on Long Fat pipe Networks", Super Computing 2004, High Performance Networking and Computing SC2004, CD-ROM

[受賞]

Supercomputing 2002, High Performance Bandwidth Challenge, "Most Efficient Use of Available Bandwidth Award(最高効率賞)"

Supercomputing 2003, High Performance Bandwidth Challenge

"Distance x Bandwidth Product & Network Technology

Award (最高バンド幅・距離積　&　ネットワークテクノロジー賞)"
Supercomputing 2004, High Performance Bandwidth Challenge,
"Single Stream, Longest Path, Standard MTU TCP Throughput Award（シングルストリーム最長パススタンダードMTU　TCP　スループット賞）"
Internet2 2005 Spring, Land Speed Record
"Single and Multiple Stream, Longest Path, Standard MTU TCP Throughput"