FPGAを用いた遺伝子情報処理のパターンマッチングシステムの設計

1ZB−05　High Speed Pattern Matching System for Amino Acid with FPGA

宮島洋介[†1]　山口佳樹[†2,†3]
丸山勉[†4]　小長谷明彦[†5,†6]

1. はじめに

ゲノムの機能解析における有効な手法に専用の計算機と優秀な解析ソフトの使用が挙げられる[5)6)7)8)]。何故なら膨大にあるデータの中からある条件と関連するデータを高速に探索する作業は計算機が得意とする作業の一つだからである。

しかしいざ専用計算機を開発するとなると、膨大な費用と開発時間が必要となるため実現することは非常に難しい。また解析ソフトの開発もそのソフト単体で飛躍的な速度向上を望むことはできない。その理由の一つには既存のマイクロプロセッサの性能限界 (動作周波数、入出力) などが挙げられる。

そこで、本稿では書き換え可能ハードウェアを利用した高速解析システムの構築を提案する。このシステムは，問題に応じたパターンマッチングのアルゴリズムを書き換え可能ハードウェアに実装することで飛躍的な高速化と低コスト化を実現するものである。

2. システム概要

本稿で提案するシステム構成は、市販の計算機とPCIボードを組み合わせたものである。PCIボードには書き換え可能ハードウェアであるFPGA(Field Programmable Gate Array) が組み込まれている。これらは小規模で比較的安価 (市販の計算機数台分) でありながらも解析時間を飛躍的に短縮することができる。

本稿のシステムを使用するユーザは特別な知識は必要なく、遺伝子の一次構造解析において必要とするパラメータ (データベース配列、遺伝子長, 類似性スコアなど) を指定するだけで簡単に使用することができる (図1)。

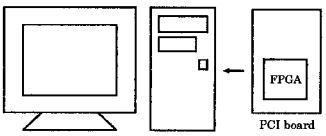

図1　システム構成

また、新しいFPGAデバイスについては、PCIボードを差し換え、パラメータを変更するだけで比較的容易にシステムに組み込むことが可能であり、常に最新のデバイスを利用することができる。

3. 書き換え可能ハードウェア

FPGA(Field Programmable Gate Array) は、その構成を問題にあわせて実行時に再構成可能なデバイスである。現在では約300万ゲート相当のものまで市販化されており、一世代前のGate Arrayの規模を大きく上回っている。

本稿ではFPGA(XILINX社製VirtexEシリーズXCV2000E[10)]) が搭載されているPCIボード (Celoxica社製RC1000-PP[9)]) を使用している。このPCIボードの概略図を図2に示す。

図2に示されている通り、RC1000-PPはボード上に2MByteの容量のSRAMを4バンク持つ。これらはPC側、FPGA側のいずれからも各バンク個別にアクセスすることができる。

4. 遺伝子の一次構造解析

遺伝子の一次構造の解析手法に、Smith-Waterman アルゴリズム (以下SW法)[1)] を使用した。

SW法は、置換、挿入、欠損という変異を含む二つ

†1 筑波大学第3学群工学システム学類
College of Engineering Systems, University of Tsukuba
†2 筑波大学大学院工学研究科
Doctoral Program in Engineering, University of Tsukuba
†3 日本学術振興会特別研究員
Research Fellow of the Japan Society for the Promotion of Science
†4 筑波大学機能工学系
Institute of Engineering Mechanics and Systems, University of Tsukuba
†5 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology
†6 ゲノム科学総合研究センター
Japan Riken Genomic Sciences Center

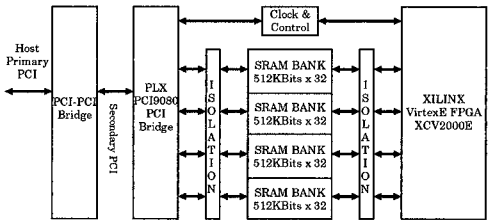

図 2 RC1000-PP 概略図

の配列に対して、一致する要素の数が最大になるような配列を新しく生成する。SW 法はダイナミックプログラミングを採用しているためホモロジを検索するのに適しているが計算量が多いため実際に比較を行うことは難しい。

そこでゲノムネットでは、より簡便な方法である BLAST[2)4)]、FASTA[3)] が利用されている。しかし、これらの手法は処理を高速にするためにギャップ検索を行わない（もしくはギャップ検索を行うが荒い検索にする）ことで検索時間を短縮し実時間で実行できるようにしているので、配列のホモロジーを見落とす可能性がある。

そこで、本稿ではホモロジー検索手法に SW 法を採用し，これを専用ハードウェアを FPGA 上に実装することで、検索時間の短縮と検索の信頼性の向上の両立を図った。

5. 遺伝子情報処理システム

本稿で構築したシステムの基本的な流れは以下のようになっている。

(1) 検索を行いたい問い合わせ配列とデータベース配列を RC1000-PP 上の SRAM に DMA 転送する。
(2) SRAM に転送されたデータを元に FPGA 上でホモロジー検索をを行う。
(3) 検索結果を RC1000-PP 上の SRAM に書き込み、PC は DMA により結果を取り込む。
(4) 処理されたデータを元に検索結果をディスプレイ上に表示する。

PC と FPGA ボードとの通信には Celoxica 社が提供する専用の関数群を利用する。これらは C 言語のプログラム内で簡単に呼び出すことができる。

問い合わせ配列として与えられる塩基配列は非常に長い文字列である。また、テンプレートとするデータベース配列もハードウェアの大きさに対して十分な長さを持っている。そこで問い合わせ塩基配列およびデータベース配列を分割しマルチスレッドで処理を行うことが可能である。以上のようにシステムの速度向上率は FPGA の回路量および FPGA ボードの個数に比例して得ることができる。

6. おわりに

本稿で提案した遺伝子の一次構造のホモロジサーチシステムは、FPGA ボード 1 枚で 200MByte のデータベース配列の検索を約 57 秒で終えることができた。今後は遺伝子と遺伝子の全体全のマッチングができるようなシステムを構築していく。また、ユーザが利用しやすいようユーザインタフェスの整備、および類似性スコアや探索手法などのライブラリを充実させていく。

謝辞 本稿の一部は、文部科学省科学研究費特定領域研究 C「ゲノム」ゲノム情報科学、および、文部科学省科学研究費補助金（特別研究員奨励費）の援助を受けて行った。

参考文献

1) Smith T. F. and Waterman M. S.: *Identification of Common Molecular Subsequences*, Journal of Molecular Biology 147, 195-197, (1981).
2) Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J.: *Basic Local Alignment Search Tool*, J. Mol. Biol. 215, 403-410, (1990).
3) Pearson,W.R. and Lipman,D.J.: *FASTA: Improved tools for biological sequence comparison*, Proc. Natl. Acad. Sci. USA 85, 2444-2448, (1988).
4) Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J.: *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucl. Aci. Res. 25(17), 3389-3402, (1997).
5) Yoshiki YAMAGUCHI, Tsutomu MARUYAMA and Akihiko KONAGAYA: *High Speed Homology Search with FPGAs*, Pacific Symposium on Biocomputing 2002, pp.271-282, (2002).
6) http://www.compugen.com
7) http://www.paracel.com/index.html
8) http://www.timelogic.com
9) http://www.celoxica.com
10) http://www.xilinx.com