Parallel Implementation of a Bioinformatics Pipeline for the Design of Pathogen Diagnostic Assays

Ravi Vijaya Satya, Kamal Kumar, Nela Zavaljevski, and Jaques Reifman
US Army Medical Research and Materiel Command (MRMC), Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, Ft. Detrick, MD
{rvijaya, kamal, nelaz, jreifman}@bioanalysis.org

Abstract

The genomes of hundreds of pathogens and their near neighbors are now available and many more are being sequenced. With the availability of this genome information, sequence-based pathogen identification has become an increasingly important tool for clinical diagnostics and environmental monitoring of biological threat agents. Chief among sequence-based identification tools are DNA microarrays, which have the ability to test for thousands of pathogens in a single diagnostic test. The design of microarray diagnostic assays involves the identification of short DNA sequences unique to a pathogen or groups of pathogens, where these unique sequences, or "fingerprints" (also referred to as probes) are used to identify the pathogens. To design pathogen fingerprints, we developed TOFI (Tool for Oligonucleotide Fingerprint Identification), a high performance computing software pipeline that designs microarray probes for multiple related pathogens in a single run.

The TOFI pipeline is extremely efficient in designing microarray fingerprints for multiple pathogens. Parallel implementation of computationally expensive specificity analysis of the designed fingerprints drastically reduces the overall execution time of the software. Comprehensive performance analysis shows that TOFI achieves super-linear speedup for up to 74 processors. A Web-based user interface, developed using the User Interface Toolkit, provides easy access to the pipeline. Using 74 processors, TOFI took approximately nine hours to design 5,015 in-silico probes for eight Burkholderia genomes with a combined size of more than 50 million base pairs. Experimental validation of these probes with various Burkholderia genomes showed that nearly 80% of the designed fingerprints identify the intended targets.

1. Introduction

Diagnostic assays provide the means for the identification of pathogens, including biological threat agents, in clinical and environmental samples. In particular, due to advances in genome sequencing technology that have led to the availability of many pathogen genome sequences, sequence-based diagnostic assays have become an attractive alternative. The availability of these genomic sequences has further opened opportunities for the development of whole-genome-based diagnostic assays, such as DNA microarrays and polymerase chain reaction assays, which offer more flexibility than traditional methods based on a single gene or selected regions of a target genome[1]. Microarray-based pathogen diagnostic assays have the ability to test for hundreds, or even thousands, of pathogens in a single diagnostic test[2], and due to this capability they are being widely used for various diagnostic applications.

A microarray-based diagnostic assay consists of thousands of oligonucleotide sequences attached to a glass plate. These oligonucleotide sequences (also referred to as probes) are used as "fingerprints" for identifying pathogens, and hence, should be unique to the pathogen (or target) genome with respect to all other non-target genomes. As a result, the design of microarray-based pathogen diagnostic assays entails the computationally expensive comparison of target genomes with all available non-target sequences. The use of high performance computing (HPC) bioinformatics tools is essential for completing these comparisons in a reasonable amount of time. Although many different methods have been developed to guide the design of pathogen diagnostic assays[1,3–9], none of them have the ability to make use of HPC resources to design oligonucleotide probes suitable for microarray-based diagnostic assays. In this paper, we describe the development of Tool for Oligonucleotide Fingerprint Identification (TOFI), an integrated, scalable, HPC

978-0-7695-3946-1/10 $26.00 © 2010 IEEE
DOI 10.1109/HPCMP-UGC.2009.36

IEEE computer society

pipeline for the design of highly specific microarray probes for pathogen identification. The TOFI pipeline designs microarray probes for multiple, related, bacterial and viral pathogens by identifying probes from the input target sequences that are unique with respect to all available non-target sequences. TOFI performs these computations efficiently by: 1) pre-processing the input sequences and identifying a small set of non-redundant target sequences, 2) running in parallel various steps in the probe design process, and 3) using the parallel BLAST[10] implementation mpiBLAST[11] for performing the specificity analysis. The pipeline scales well with increase in the number of target genomes, and can potentially design fingerprints for hundreds of related target genomes in a single run.

2. Methods and Implementation

Given a set of target genomes, TOFI finds microarray fingerprints that are unique to any subset of the target genomes with respect to all sequenced non-target genomes. In the following, we briefly describe the various components of the TOFI pipeline, which consists of the three main stages illustrated in Figure 1. The stages are designed so that large portions of the target genomes are eliminated in the less expensive two initial stages, and the computationally more expensive searches for specific fingerprints are performed over smaller regions of the target genome in the final stage.

Before submitting the target genomes to the first stage of the pipeline, they are first pre-processed to build a set of non-redundant target sequences. In general, there is significant sequence similarity among genomes of related organisms. Hence, to take advantage of this sequence similarity and reduce the effective size of the input target genomes, TOFI uses the suffix-tree-based MUMmer program[12] to compare the target genomes with each other and eliminates any repeated occurrences of identical segments. This pre-processing step reduces the input target genomes to a set of non-redundant target sequences. From our experience with bacterial genomes, this step can reduce the combined length of input target sequences by as much as 80% of their original size.

The first stage of TOFI also uses the MUMmer program, but here it is used to perform pairwise comparisons of the non-redundant target sequences with each non-target genome and eliminate regions in the target sequences that have exact matches with any of the non-target genomes. TOFI uses MUMmer to find these maximal exact matches and eliminate regions in the target sequences that are covered by them. This procedure ensures that every segment of the target sequences that satisfies the restrictions on probe length and specificity parameters is part of the surviving regions of the target genomes. These surviving regions, referred to as candidate sequences, are then passed on to the second stage of the pipeline.

In the second stage, TOFI identifies oligonucleotides of desired length from the candidate sequences that satisfy DNA microarray experimental conditions, such as melting temperature (T_m) and GC content. TOFI uses the open source UNAFold[13] software to identify these oligonucleotides. UNAFold uses the nearest-neighbor hybridization model[14] to calculate T_m and to estimate whether a probe forms any secondary structures that may prevent it from identifying the intended targets.

In the third stage of the pipeline, TOFI performs a BLAST[10] search for each probe against a comprehensive sequence database, such as the *nt* database provided by the National Center for Biotechnology Information. The BLAST comparisons are performed in parallel on multiple processors using the blastn program of mpiBLAST[11]. TOFI computes the specificity of each probe based on multiple specificity criteria selected by the user. Probes with significant alignments to non-target genomes are eliminated and the surviving probes become the *in-silico* DNA fingerprints for the target genomes.

The TOFI pipeline also incorporates a further post-processing step in which each probe is aligned with all the input target genomes using the pairwise BLAST program, bl2seq. Based on these alignments, TOFI computes the fingerprints that are unique to each individual genome as well as fingerprints that are common to subsets of input genomes. These probes are then subjected to experimental sensitivity and specificity validation tests.

2.1. Parallel Implementation of the TOFI Pipeline

All the three stages of TOFI pipeline are implemented in parallel. In Stage 1, the target sequences are compared against a very large database of non-target sequences using the MUMmer program. As the size of the non-target sequence database is very large [>24 billion base pairs (bp)], these comparisons take 8–10 hours using a single processor. To perform these comparisons in parallel, the non-target sequence database is split into n fragments, where n is the number of processors specified by the user. The input sequences are compared against each fragment in parallel and the segments matching the non-target sequences are eliminated. The surviving candidate sequences from each processor are pooled together and only the candidate sequences reported by all processors are provided as the output candidate sequences from Stage 1.

In Stage 2 of the pipeline, the candidate sequences are equally distributed among the available processors. At each processor, a separate instance of the UNAFold program is run on the candidate sequences assigned to that processor. The microarray probes designed by each individual processor are combined to form the complete set of microarray probes for the candidate sequences.

Stage 3 is the most computationally expensive part of the pipeline. On average, more than 90% of TOFI's execution time is spent in Stage 3. For Stage 3, TOFI uses mpiBLAST, a parallel implementation of the BLAST program. The mpiBLAST program provides two levels of parallel execution—database fragmentation and query segmentation. With the database fragmentation option, the user splits the non-target database into smaller fragments offline. At run time, the same query sequence is aligned against different fragments of the database at individual processors and the results are assembled together by a master process. For best performance, the size of the individual fragments should be small enough to fit in the memory available for each process. However, if the fragments are too small, the communication overhead and inter-process dependencies will cause significant delays, thereby increasing the overall execution time.

Query segmentation options of mpiBLAST enable the processing of multiple query sequences simultaneously. Using query segmentation, two or more query sequences are processed simultaneously, with a set of processors working on each individual query. Depending on the number of processors available, using database fragmentation and query segmentation together will result in the best performance. To achieve optimal performance, TOFI uses both database fragmentation and query segmentation options of mpiBLAST.

2.2. Graphical User Interface

The TOFI pipeline is available to users of the DoD Supercomputing Resource Center (DSRC) via a Web-based graphical user interface (GUI), accessible at https://applications.bioanalysis.org/tofi/. The Web-based GUI allows users to access HPC clusters and run TOFI jobs from any Web browser. It uses the User Interface Toolkit (UIT) for communicating with HPC clusters at the Maui High Performance Computing Center (MHPCC), where the TOFI pipeline is currently deployed. UIT is a Web-service application programming interface (API) that provides secure access to HPC resources. TOFI users are authenticated using their Kerberos and SecurID credentials via the UIT Web-service.

Without requiring any special plug-ins, the TOFI GUI provides rich desktop-like interface within a Web browser with capabilities such as remote file browser for easy access to HPC files and drag-and-drop components. Figure 2 shows a screenshot of the TOFI GUI. The Web-based GUI was developed as a Web-application using Java, J2EE, JavaServer Faces (JSF) (http://java.sun.com/javaee/javaserverfaces/), ICEfaces (http://www.icefaces.org/), and asynchronous JavaScript and XML (AJAX)[15]. The main Web application consists of server-side Java codes that use JSF- and AJAX-based APIs from ICEfaces. ICEfaces provides a rich set of user interface components, such as buttons, drag-and-drop lists etc., and generates updates of Web pages on the fly with all of the application logic hidden on the server side. The Web application is deployed on an Apache Tomcat (http://tomcat.apache.org/) server, using a secure hypertext transfer protocol over a secure socket layer connection for encrypting all of the data flowing to and from the user's Web browser.

3. Results

In this section, we present some performance results from TOFI, when it was run with a set of eight *Burkholderia* genomes, which include four strains of *Burkholderia mallei* and four strains of *Burkholderia pseudomallei*. The combined size of these eight genomes is more than 51 million bp.

In all, TOFI identified 5015 *in silico* fingerprints for the eight *Burkholderia* genomes. A detailed breakdown of these fingerprints is presented in Figure 1. There were nearly 1,000 fingerprints common to all eight input genomes, 32 fingerprints common to the four *B. mallei* genomes, and nearly 500 fingerprints common to the four *B. pseudomallei* genomes. While there were no fingerprints unique to each individual *B. mallei* genome, there were many fingerprints unique to each individual *B. pseudomallei* genome.

3.1. Performance of the TOFI Pipeline

The pre-processing step was very effective in reducing the effective size of the input sequences. Figure 2 shows the effectiveness of the pre-processing step as more and more genomes are added. Initially, when the first genome is added, the combined size of the non-redundant sequences is the same as that of the input genome. However, the size of non-redundant sequences increases very slowly as more and more genomes are added. This is because TOFI eliminates the redundant portions of the new genome and adds only the non-redundant segments. The combined size of the non-redundant candidate sequences for the eight genomes is only 12 million bp, which is less than 25% of the original size of the input genomes. That is, for the current test case the pre-processing step results in a fourfold reduction of the overall execution time.

TOFI obtains nearly linear speedup in Stage 1 and in Stage 2, as expected. However, runtime is dominated by Stage 3. For the eight *Burkholderia* genomes described above, the serially-performed pre-processing step of TOFI, takes less than one minute. Using 74 processors, Stage 1 takes approximately 30 minutes, Stage 2 takes 10 minutes, and Stage 3 takes nearly 7 hours. The post-processing step is completed in less than two minutes. The total execution time is less than 9 hours using 74 processors. In this case, nearly 90% of the total execution time is spent in Stage 3.

Figure 5 shows the overall speedup of the TOFI pipeline, which is nearly identical to the speedup of Stage 3. TOFI obtains super-linear speedup for up to 94 processors. However, the speedup starts declining beyond 74 processors. When TOFI is run with four processors (which is the base case in Figure 5), the database fragments are too large to fit in the memory available at each individual processor. As the number of processors is increased, the database fragments get smaller and smaller until they are small enough to fit in the memory available for each processor. As a result, initially we obtain super-linear speedup; however, as the number of processors is increased further, communication overhead becomes increasingly significant, gradually reducing the initial speedup.

3.2. Experimental Validation

The fingerprints designed by TOFI for the *Burkholderia* genomes have been experimentally validated by life scientists at the US Army Medical Research Institute of Infectious Diseases (USAMRIID) at Ft. Detrick, MD. More than 80% of these fingerprints were found to identify the intended targets with high sensitivity and specificity. In addition, in a one-way blinded test, fingerprints designed to identify common signatures of multiple bacterial strains of the *B. pseudomallei* species were successful in identifying a different, unsequenced strain of the same species. Detailed analysis of these results was previously reported by Vijaya Satya et al.[16].

4. Conclusions

The TOFI pipeline described in this paper is a highly scalable software system with the ability to design microarray fingerprints for multiple bacterial and viral genomes. It is currently being used by USAMRIID life scientists to design diagnostic assays for various viral and bacterial pathogens. In addition, it has been used by plant pathologists at the US Department of Agriculture to design fingerprints for various plant pathogens. The pipeline is highly efficient and can potentially design fingerprints for hundreds of related pathogen genomes in a single run.

Acknowledgments

This work was sponsored by the US Department of Defense High Performance Computing Modernization Program, under the High Performance Computing Software Applications Institutes initiative.

Disclaimer

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense. This paper has been approved for public release with unlimited distribution.

References

1. Slezak, T., T. Kuczmarski, L. Ott, C. Torres, D. Medeiros, J. Smith, B. Truitt, N. Mulakken, M. Lam, E. Vitalis, A. Zemla, C.E. Zhou, and S. Gardner, "Comparative genomics tools applied to bioterrorism defence." *Brief Bioinform*, vol. 4, pp. 133–49, Jun. 2003.

2. Loy, A. and L. Bodrossy, "Highly parallel microbial diagnostics using oligonucleotide microarrays." *Clin Chim Acta*, vol. 363, pp. 106–19, Jan. 2006.

3. Fitch, J.P., S.N. Gardner, T.A. Kuczmarski, S. Kurtz, R. Myers, L.L. Ott, T.R. Slezak, E.A. Vitalis, A.T. Zemla, and P.M. McCready, "Rapid Development of Nucleic Acid Diagnostics." *Proceedings of the IEEE*, vol. 90, pp. 1708–20, 2002.

4. Kaderali, L. and A. Schliep, "Selecting signature oligonucleotides to identify organisms using DNA arrays." *Bioinformatics*, vol. 18, pp. 1340–9, Oct. 2002.

5. Phillippy, A.M., J.A. Mason, K. Ayanbule, D.D. Sommer, E. Taviani, A. Huq, R.R. Colwell, I.T. Knight, and S.L. Salzberg, "Comprehensive DNA signature discovery and validation." *PLoS Comput Biol*, vol. 3, p. E98, May 18, 2007.

6. Rimour, S., D. Hill, C. Militon, and P. Peyret, "GoArrays: highly dynamic and efficient microarray probe design." *Bioinformatics*, vol. 21, pp. 1094–103, Apr. 1, 2005.

7. Rouillard, J.M., M. Zuker, and E. Gulari, "OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach." *Nucleic Acids Res*, vol. 31, pp. 3057–62, Jun. 15, 2003.

8. Tembe, W., N. Zavaljevski, E. Bode, C. Chase, J. Geyer, L. Wasieloski, G. Benson, and J. Reifman, "Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays." *Bioinformatics*, vol. 23, pp. 5–13, Jan. 1, 2007.

9. Wang, D., A. Urisman, Y.T. Liu, M. Springer, T.G. Ksiazek, D.D. Erdman, E.R. Mardis, M. Hickenbotham, V. Magrini, J. Eldred, J.P. Latreille, R.K. Wilson, D. Ganem, and J.L. DeRisi,

"Viral discovery and sequence recovery using DNA microarrays." *PLoS Biol*, vol. 1, p. E2, Nov. 2003.

10. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic local alignment search tool." *J Mol Biol,* vol. 215, pp. 403–10, Oct. 5, 1990.

11. Darling, A., L. Carey, and W. Feng, "The Design, Implementation, and Evaluation of mpiBLAST "*4th International Conference on Linux Clusters: The HPC Revolution 2003 in conjunction with the ClusterWorld Conference & Expo*, San Jose, CA, 2003.

12. Kurtz, S., A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg, "Versatile and open software for comparing large genomes." *Genome Biol*, vol. 5, p. R12, 2004.

13. Markham, N.R. and M. Zuker, "UNAFold: software for nucleic acid folding and hybridization." *Methods Mol Biol*, vol. 453, pp. 3–31, 2008.

14. Santa Lucia, Jr., J. and D. Hicks, "The thermodynamics of DNA structural motifs." *Annu Rev Biophys Biomol Struct*, vol. 33, pp. 415–40, 2004.

15. Paulson, L.D., "Building rich web applications with Ajax." *Computer*, vol. 38, pp. 14–17, 2005.

16. Vijaya Satya, R., N. Zavaljevski, K. Kumar, E. Bode, S. Padilla, L. Wasieloski, J. Geyer, and J. Reifman, "In silico microarray probe design for diagnosis of multiple pathogens." *BMC Genomics*, vol. 9, p. 496, 2008.

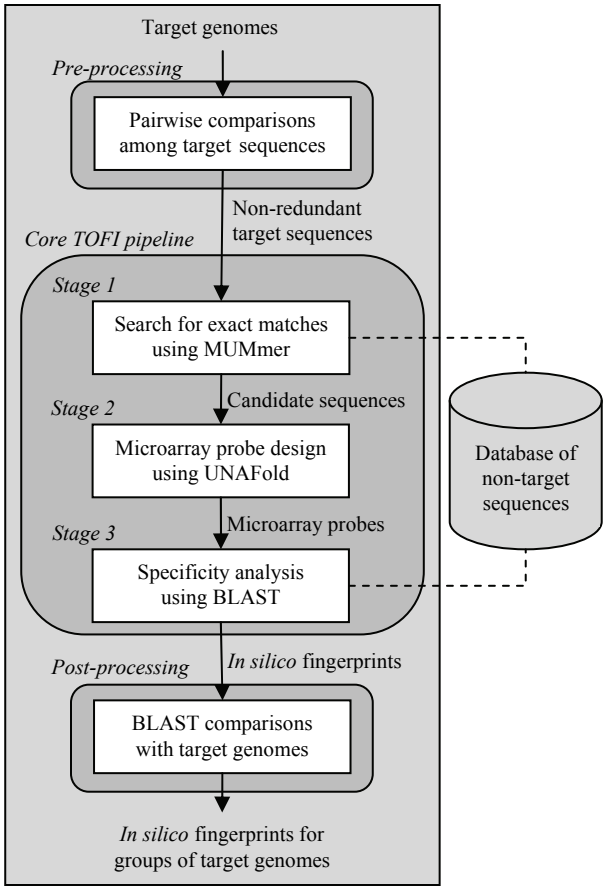

Figure 1. Overview of TOFI pipeline. The pre-processing stage of TOFI eliminates redundant sequences from the target genomes. The actual fingerprint design process, including comparison with non-target genomes, happens in the three stages of the core TOFI pipeline. The post-processing module identifies fingerprints that are common to groups of target genomes.

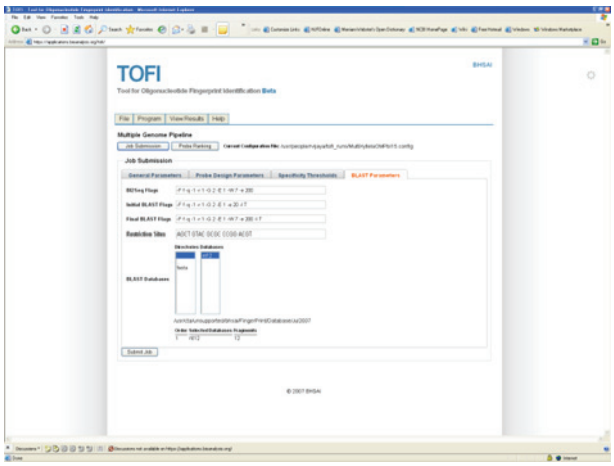

Figure 2. A screenshot of the TOFI Graphical User Interface (GUI). The TOFI GUI allows users to run the pipeline for single and multiple genome targets. The GUI runs entirely from a Web browser, and does not require the installation of any software at the user's end.

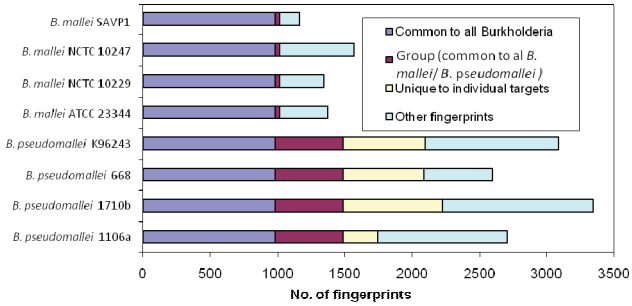

Figure 1. Properties of the 5015 *in silico* fingerprints designed for the eight *Burkholderia* genomes. Each horizontal bar indicates the total number of in silico fingerprints that identify the corresponding *Burkholderia* genome. There are nearly 1,000 fingerprints that are common to all eight input genomes. Thirty two fingerprints are common to the four *B. mallei* genomes, and nearly 500 fingerprints are common to the four *B. pseudomallei* genomes. In addition, there are a significant number of fingerprints that are unique to each *B. pseudomallei* genome.

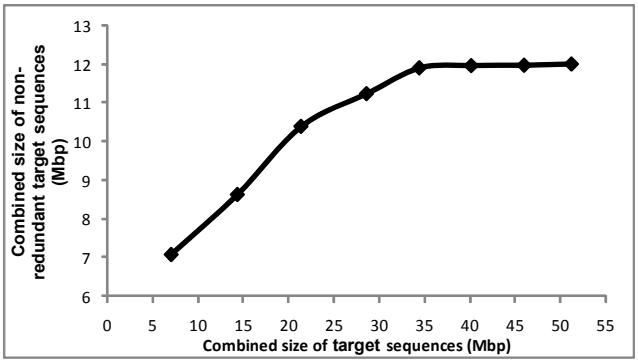

Figure 2. Effectiveness of the pre-processing step. The pre-processing step effectively reduces the size of the input target genomes. As the size of the input sequences increases from 7.2 million base pairs (bp) for one genome to 51 million bp for eight genomes, the size of the non-redundant target sequences only increases from 7.2 to 12 million bp.

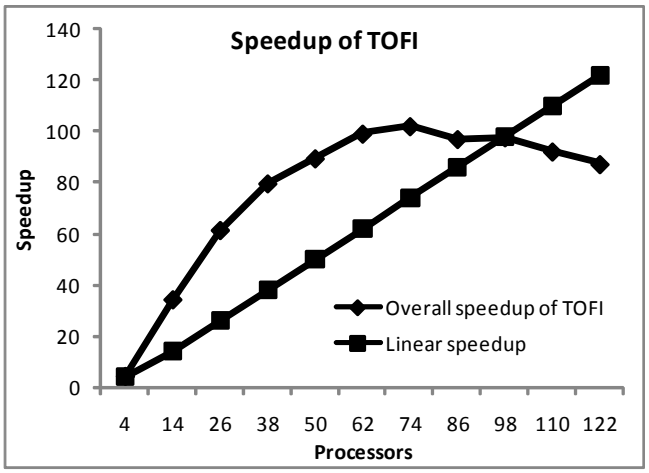

Figure 3. Overall speedup of TOFI. The overall TOFI speedup is dominated by the speedup of Stage 3, as nearly 90% of the total execution time is spent in the last stage of the pipeline. TOFI obtains the best speedup with 74 processors, with the database split into 12 fragments.