Concurrent Bioinformatics Software for Discovering Genome-wide Patterns and Word-based Genomic Signatures

Jens Lichtenberg[1], Kyle Kurz[1], Xiaoyu Liang[1], Rami Al-ouran[1], Lev Neiman[1], Lee Nau[1], Joshua Welch[1], Edwin Jacox[2], Thomas Bitterman[3], Klaus Ecker[1], Laura Elnitski[2], Frank Drews[1], Stephen Lee[4], Lonnie Welch[1,5,6,§]

1 - Bioinformatics Laboratory, School of EECS, Ohio University, Athens, Ohio, USA
2 - Genomic Functional Analysis Section, National Human Genome Research Institute, NIH, Rockville, MD, USA
3 - Department of Statistics, University of Idaho, Moscow, Idaho, USA
4 - Cyberinfrastructure Group, Ohio Supercomputer Center, Columbus, OH, USA
5 - Biomedical Engineering Program, Ohio University, Athens, Ohio, USA
6 - Molecular and Cellular Biology Program, Ohio University, Athens, Ohio, USA

§ - Presenting author: welch@ohio.edu

URL for the overall project web site: www.word-seeker.org
URL for accessing the code: http://word-seeker.googlecode.com/svn/trunk
Open Source License being used: GNU General Public License v3

The importance of discovering the patterns and features in genomic sequences is motivated by a number of problems in biology. The Encyclopedia of DNA Elements project, ENCODE, seeks 'to identify all functional elements in the human genome sequence'. The study of co-regulated genes involves the analysis of the promoter sequences, introns, and UTRs of genes that were determined, by microarray experiments, to be co-regulated. Similarly, transcription factor binding regions identified by ChIP-chip and ChIP-seq experiments are examined to identify genomic patterns. Genome-wide pattern discovery studies seek to identify vocabularies of genomes. The search for genomic signatures seeks unique elements that characterize specific organisms, tissues, pathways, and functions. A number of algorithms and software tools have been developed to address some of these problems. However, very few provide the scalability needed to process large (genome-scale) data sets. Furthermore, none provides the comprehensive, integrated set of capabilities required by the complete set of biological problems mentioned above.

This manuscript presents WordSeeker, a general purpose, concurrent software suite that addresses these shortcomings. The Open Word Enumeration Framework (OWEF) class performs a central role in WordSeeker. When it was presented at BOSC 2009, OWEF had been used only to support a single data structure for word enumeration. Since that time, it has been employed successfully with three different data structures (radix tree, suffix tree, and suffix array). Additionally, it has been deployed on multi-core and distributed computational platforms. The concurrent implementation of WordSeeker divides the data space among nodes by using nucleotide prefixes. A controller task coordinates the activities of the worker nodes, each of which enumerates a subset of the DNA *word space*. To build a distributed Markov chain model for the computation of word scores, the nodes communicate with each other to obtain word occurrence information. WordSeeker is deployed on the Ohio Supercomputer Center's *Glenn* cluster, which is an IBM e1350 system with more than 4200 Opteron processor cores that are connected by 20 Gbps Infiniband.

To measure the performance enhancement due to concurrency, WordSeeker's performance was evaluated on a 5 node distributed system and on a single node. Execution time was measured for the 27,167 core promoters of *A. thaliana*, and for the entire *E. coli* genome. Two different algorithms, radix tree and suffix tree, were compared. Table 1 shows the results for DNA words of length 20.

Table 1. Performance measurements for a distributed implementation of WordSeeker.

		5 Nodes						1 Node					
		Radix Tree			Suffix Tree			Radix Tree			Suffix Tree		
Data Set	**Word Length**	**Complete (h:m:s)**	**Enum. (secs)**	**Scoring (secs)**	**Complete (h:m:s)**	**Enum. (secs)**	**Scoring (secs)**	**Complete (h:m:s)**	**Enum. (secs)**	**Scoring (secs)**	**Complete (h:m:s)**	**Enum. (secs)**	**Scoring (secs)**
A. thaliana Core Prom.	20	0:05:10	3.74	303.51	0:05:39	20.74	315.43	0:45:56	29.69	6247.6	0:50:27	126.63	6218.3
E.coli Genome	20	0:01:18	5.49	68.89	0:01:27	13.03	70.15	0:03:48	60.4	165	0:03:23	78.65	134.57