A Multi-Granularity Parallel Approach to the Analysis of Evolutionary History Using Genomic Resources

Jesse D. Walters,[2,4] Thomas L. Casavant,[1,2,3,4,6] John P. Robinson, [2,4] Thomas B. Bair,[2] Terry A. Braun,[1,2,3,5] and Todd E. Scheetz[1,2,3,5]

[1]Center for Bioinformatics and Computational Biology,
http://genome.uiowa.edu
[2]Coordinated Laboratory for Computational Genomics,
{jwalters, tomc, tbair, tabraun, tscheetz}@eng.uiowa.edu
john-robinson@uiowa.edu
http://genome.uiowa.edu/clcg.html
[3]Department of Biomedical Engineering
http://www.bme.engineering.uiowa.edu
[4]Department of Electrical and Computer Engineering,
http://www.ece.engineering.uiowa.edu
[5]Department of Ophthalmology and Visual Sciences
http://www.ece.engineering.uiowa.edu
Iowa City, IA
[6]Stockholm Bioinformatics Center
http://www.sbc.su.se
Stockholm, Sweden

Abstract. This extended abstract describes XenoCluster, a parallel approach to analyzing large genomic sequence datasets in an evolutionary context. The general biological aim of this work is understanding the functions of the genes of all species. This work was originally motivated by the desire to identify rare evolutionary events by detecting 'unusual' phylogenetic gene trees. With the availability of large grid environments, as well as genomic sequence from 1,000s of species, the possibility exists to perform gene clustering using phylogenetic tree "similarity" as a metric. A direct assault on this problem would require years of uni-processor CPU time. While some phases are readily parallelized, making only modest IPC demands, at least one phase of the solution is best-suited to fine-grained parallelism, so provides an excellent vehicle for exploring the domain of multi-grained parallel scientific computations in the area of genomics and bioinformatics.

1. Introduction

The mass availability of trillions of nucleotides of genomic sequence from more than 2,000 species containing as many as 35,000 genes each, makes it possible to pose biological and biomedical questions that just a few years ago would have been inconceivable. However, without large-scale parallel computational power, it would still be infeasible to practically address and answer these same questions. In multiple phases of computation, *XenoCluster* identifies genes in a species which are either highly similar to other genes, or which appear anomalous – *from an evolutionary perspective*. To simplify the presentation, the case of identifying anomalous genes (aka xenologs), possibly resulting from a process known as Horizontal Gene Transfer (HGT), is discussed. By modifying threshold parameters of this suite of software, it is

trivial to direct the results toward the identification of functionally-related genes as implied by a highly similar (rather than divergent) pattern of evolutionary behavior.

2. Approach and Methods

The core of the XenoCluster algorithm is divided into 3 major phases:

1. Identification of a maximal set of orthologous genes across species.
2. Generation of phylogenetic trees resulting from the orthologous groups.
3. Clustering of these trees into groups corresponding to genes which show consistent evolutionary behavior.

In phase 1, it is necessary to identify potential homologous genes for every gene in the union of a complex set of 1000s of species. This is accomplished by BLASTing each gene against the set of all known genes in all species, and then performing a reciprocal BLAST operation to verify that the best hit for each gene hits the original gene with the highest rank score. This becomes the base set of orthologous gene groups to be used in phase 2 among which xenologs may be identified. The second phase involves sequence trimming and multiple alignment of all members of each of the orthologous gene groups, followed by the automated generation of a phylogenetic tree for each aligned group. Phase 3 performs an all-pairs distance analysis of phylogenetic trees for all gene groups, and then uses a clustering technique to identify maximal sets of trees, which represent sets of genes which share a common evolutionary history. Details of this three-phase clustering may be found in [1, 2]. The tree comparison metric is based on work by Wang [3].

Table 1. Benchmark timings on 20,364 RefSeq genes for the component phases of XenoCluster run with 1 dual CPU node (cluster size, N=1).

Phase/component	**Time (Seconds)**	**# of Iterations**	**Total (Seconds)**
Intra Cluster IPC	124	1	124
Inter Cluster IPC	311	1	311
Initial BLAST	301	20364	6129564
Reciprocal BLAST	12	794196	9530352
Sequence Alignment	33	20364	672012
PHYLIP tree generation	2518	20364	51276552
Tree Clustering	1036800	1	1036800
Total	**1,040,099**	**855,291**	**68,645,715**
Days to completion			**794.5 days**

Each of the phases described in detail above were implemented in a LINUX environment (2.2GHz dual Athlon with 2GB RAM running either Fedora or Redhat 9.0), and benchmark executions were performed using the largest set of human genes known at the time of publication (RefSeq release 12, July 2005). Table 1 summarizes the detailed serial wall-clock execution times of the five computational and two communication phases of XenoCluster. The intra-cluster communication overhead was empirically based on a 100 Mbit connection, while the inter-cluster cost was based on a 1 Gbit connection. The tree clustering phase was done using the *pthread* multi-threading package for Linux 2.4 based kernels.

3. Results and Discussion

Grid Results on Human Genome Data

For reporting our results, we have employed a simple grid model based on two parameters – **K**: the number of clusters, and **N**: the number of nodes in each cluster. We examined both the effect of increasing overall system size while maintaining the number of clusters constant, and while holding the cluster size constant. Figure 1 shows a composite of these many analyses, examining a range of combinations of K and N. As shown, the optimal configuration corresponds to K=16, and N=128. Such a 2,048 node grid solution would yield an execution time of 12.8 days. Note that this execution time is roughly equal to the non-parallelized final tree clustering phase.

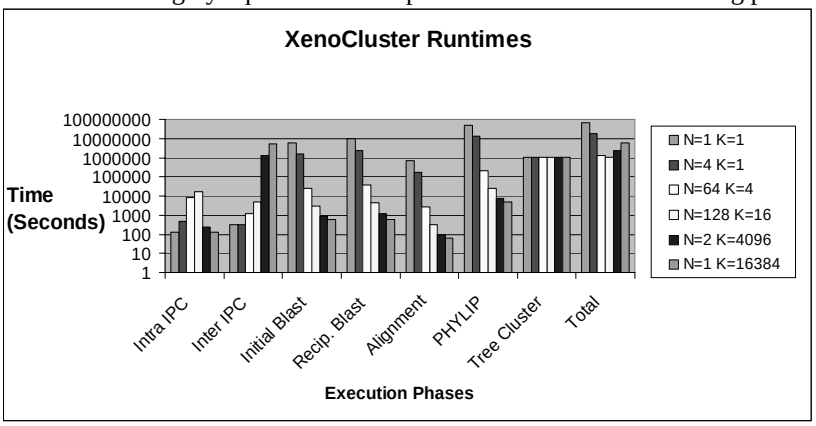

Figure 1. XenoCluster execution times for varying number of clusters (K) and cluster sizes (N). Total number of compute nodes varies from 1 to 16384.

Multi-threading Results on Yeast Genome Data

Both for purposes of providing a more manageable benchmark dataset, and for allowing us to compare our biological results to known instances of HGT, we also performed an analysis of yeast genome data. The multi-threaded execution results based on this yeast data are illustrated in Table 2.

Biological Validation

In order to evaluate the potential for XenoCluster to produce useful predictions of evolutionary anomalies, we evaluated our predictions versus a set of 'known' horizontally transferred genes in yeast [4]. A set of 10 experimentally validated genes were used, whose GenBank yeast identifiers are shown in column 2 of Table 3. 4 of these 10 genes lacked sufficient ortholog data to perform reliable predictions. However, 6 genes we not only present in the data, but also ranked within the top 6% of all possible yeast genes.

A critical parameter in the use of this software is the use of a similarity threshold. Columns 4 and 5 of Table 2 show the threshold values above which the gene either remained as a singleton cluster (column 4), or the threshold at or below which the gene remained in the "consensus cluster" (column 5).

Table 2. XenoCluster results on 10 known yeast genes [4]. 6 of 10 candidates with sufficient data available ranked in the top 6% of all 4,234 yeast genes.

HGT candidate	Yeast ID	Best Identity	Forms Singleton %	Joins large cluster %	# of Inter-tree links
1	YFR055W	Hypoth Prot			
2	YMR090W	Hypoth Prot			
3	YOL164W	BDS1			
4	YJL217W	Hypoth Prot			
5	YDR540C	Hypoth Prot	85%		
6	YJL218W	Hypoth Prot	90%		
7	YPL245W	Hypoth Prot		97%	25
8	YKL216W	URA1		97%	12
9	YNR057C	BIO4		97%	7
10	YNR058W	BIO3		97%	1

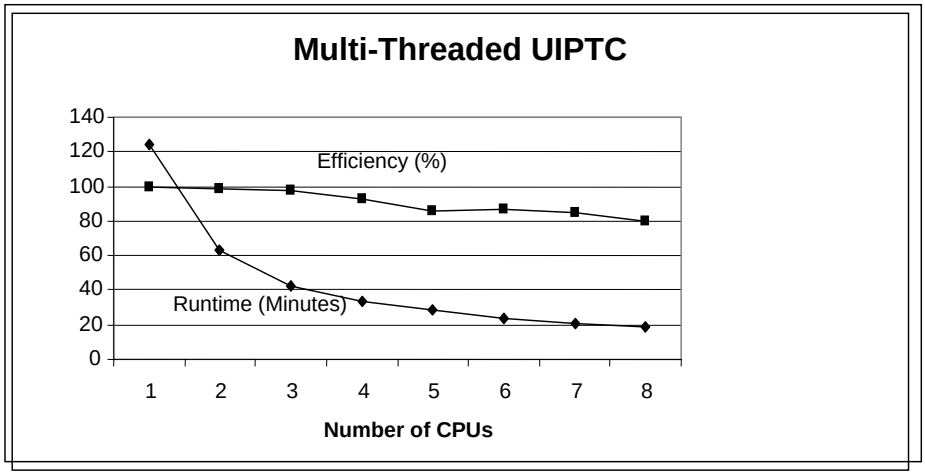

Figure 2. The benefits of multi-threading in phase 3 – UI Parallel Tree Clustering. Execution time is reduced from just over 2 hrs to under 20 minutes.

References

[1] Walters, J. "Xenocluster: A Parallel Computational Method to Identify Horizontal Gene Transfer Events", M.S. Thesis, University of Iowa, December 2005.

[2] Walters, J., T. Casavant, J. Robinson, T. Bair, T. Braun and, T. Scheetz, "XenoCluster: A Grid Computing Approach to Finding Ancient Evolutionary Genetic Anomalies," LNCS 3606, 9/2005, pp 355 – 366.

[3] Wang JTL, Shan H, Shasha D and Piel WH., "TreeRank: A Similarity Measure for Nearest Neighbor Searching in Phylogenetic Databases", Proceedings of the 15th International Conference on Scientific and Statistical Database Management (SSDBM 2003), Cambridge, Massachusetts, pp. 171-180.

[4] Hall, C. Brachat, S. Dietrich, F., Contribution of Horizontal Gene Transfer to the Evolution of Saccharomyces cerevisiae Eukaryotic Cell 2005, 4(6), 1102-1115.