ESTmapper: Efficiently Clustering EST Sequences Using Genome Maps

Xue Wu, Woei-Jyh (Adam) Lee, Damayanti Gupta, Chau-Wen Tseng [1]

Keywords: Bioinformatics, EST clustering, genome map, suffix tree, high performance computing

1 Introduction

Expressed sequence tags (ESTs) are nucleotide sequences of transcribed genes that provide important information about gene discovery and gene expression in many organisms. Because individual ESTs are incomplete, many ESTs must be *clustered* to discover the full sequence for each gene. Many earlier EST clustering algorithms (TGICL [1], d2_cluster [2], GeneNest [3], CLOBB [4]) rely on performing pairwise comparisons between ESTs, with closely matching pairs put into a single cluster. But obviously these algirithms do not scale well because of the $O(n^2)$ number of pairwise comparisons needed. More recent approachs (PaCE [5], Xsact [6]) build a *suffix tree* of all ESTs to identify pairs of ESTs with long common substrings. But these algorithms are still pairwise comparison based, and the time complexity are $O(n \log n)$. In addition, since ESTs represent large numbers of incomplete, error-prone, and redundant fragments of genes, the above EST to EST comparison based algorithms cannot guaranteen the accuracy of the clustering results. Biologists' help are needed to correct and assure the precision of the final results. With the advent of high-throughput sequencing of the entire genomes of many species, an alternative approach becomes possible. In this paper, we describe ESTmapper, a new tool for clustering EST sequences based on efficiently mapping ESTs to the genome.

2 ESTmapper

The algorithm used by ESTmapper consists of two steps: preprocessing genome and clustering ESTs. It first builds a suffix tree for the genome, then searches for long common substrings between each EST and the genome. We use them to build gapped matching regions to account for sequencing errors and splicing, and use the longest overall matching region to map the EST to locations in the genome. ESTs mapped to overlapping locations are then placed in a cluster. Preliminary experiments show that ESTmapper is not only very efficient for its linear processing time, but also precise with respect to the input EST data when compared to other popular EST clustering tools.

3 Performance and Statistics

We compared the performance of ESTmapper, TGICL and PACE on a AMD Athalon PC cluster. The dataset are a selection of ESTs from Arabidopsis thaliana and its second chromosome. TGICL and ESTmapper are executed on a single PC node, but PaCE is executed on 8 nodes. Resutls are presented in the following table. We also evaluated the

[1] Computer Science Department, University of Maryland at College Park. E-mail: {wu,adamlee,dami,tseng}@cs.umd.edu

scalability of ESTmapper. Figure 1 shows its running time when mappig from 250K to 1.5 million human ESTs against human chromosome 21.

# ESTs	Execution time (seconds)			Memory use (MB)		
	TGICL	PaCE	ESTmapper	TGICL	PaCE	ESTmapper
2,980	14	131	72	77	103	661
5,960	34	282	85	79	106	661
11,921	73	686	111	81	112	661
23,842	142	872	174	85	181	661
47,685	313	1702	278	88	336	662
95,370	675	4285	460	102	603	662
190,740	1987	-	861	129	1192	663

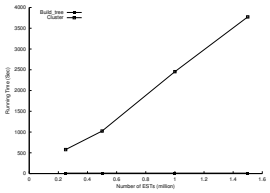

Figure 1: Scalability of ESTmapper

In addition, ESTmapper also provides various statistics about the mapping and clustering results. Table 1 shows several parameters that affect ESTmapper's clustering precision. The dataset are 2224 Arabidopsis ESTs and its second chromosome. At the same time the results also gives one example of statistics information provided by ESTmapper.

Min Common Substring length	Number Unmapped ESTs	Average Mapping Region length	mapping region/EST length ratio	average gap length	average gap/mapping length ratio	average number gaps	number cluster	number singleton
10	0	408	97.80%	1	0.04%	0.1	711	310
20	0	411	98.50%	1	0.07%	0.2	711	310
40	0	407	97.60%	3	0.04%	0.1	713	311
80	15	407	97.30%	1	0	0.02	708	309
160	91	416	97.20%	0	0	0.005	689	302
320	505	452	97.60%	0	0	0	568	263
640	2166	688	97.90%	0	0	0	50	53

Table 1: Impact of minimum common substring length

4 Conclusions

The measured performance results show that ESTmapper has constant memory usage and linear execution time with regard to the number of processed ESTs. Since the size of genome is relatively constant, our EST clustering is more scalable than other clustering algorithms. Besides, our EST clustering tool can also provide useful statistic information about the clustered ESTs, which we believe can help biologists with their research.

References

[1] G. Pertea, X. Huang, F. Liang and V. Antonescu et al, TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets *Bioinformatics* Vol 19(5), pp. 651 - 652; March, 2003

[2] J. Burke, D. Davison and W. Hide, d2_cluster: A Validated Method for Clustering EST and Full-Length cDNA Sequences *Genome Res.*, 9(11): 1135 - 1142; November, 1999

[3] S. Haas, T. Beissbarth, E. Rivals, A. Krause and M. Vingronu, GeneNest: automated generation and visualization of gene indices *Trends Genet.*, 16(11), 521-523; 2000

[4] J. Parkinson, D. Guiliano and M. Blaxteur, Making sense of EST sequences by CLOBBing them *BMC Bioinformatics*, Vol 3(1):31; 2002

[5] A. Kalyanaraman, S. Aluru, S. Kthari and V. Brendeul, Efficient clustering of large EST data sets on parallel computers *Nucleic Acids Research*, Vol 31(11), pp. 2963-2974; 2003

[6] K. Malde, E. Coward and I. Jonassen, Fast sequence clustering using a suffix array algorithm *Bioinformatics*, Vol. 19, pp. 1221-1226; 2003