Clustering and Assembly of a Large Number of EST and cDNA Sequences Using Hyseq's Proprietary Software

John Tillinghast **Ankura Sinku** **Y. Tom Tang**
john@sbh.com ankura@sbh.com tom@sbh.com
Bioinformatics Division, Hyseq, Inc. 670 Almanor Ave., Sunnyvale, CA 94086, USA

Keywords: clustering, assembly, EST, distributed, PHRAP

1 Introduction

Many organizations sequence cDNA clones from diverse tissue sources. Clustering and assembly of these sequences is a critical step for further genomic studies. Output from the clustering and assembly provides the expressed coding regions of genes with no introns, and contains important information about gene polymorphisms (SNPs hereafter) and splice-variants [1].

Existing bioinformatic tools for the clustering and assembly of cDNA segments (primarily EST sequences) include PHRAP [2], TIGR-ASSEMBLER [4], the NCBI UniGene approach (clustering without assembly), and STACK-PACK (clustering by word counts, keeping track of tissue types) [3]. Each of the above packages has its own limitations. For example, both PHRAP and TIGR-ASSEMBLER were initially developed to assemble genomic sequences. When used for EST/cDNA assembly, PHRAP is limited to assemble small numbers of sequences. For anything more, it requires supercomputers, large amounts of system memory, and a lot of CPU time. It does not address the specific needs of cDNA/EST assembly, such as SNP and splice-variant detection. The TIGR-ASSEMBLER, on the other hand, does not use the sequence quality scores reported by PHRED, and does not perform SNP or splice-variant detection either.

We have developed a proprietary package of programs to solve these problems. The software introduces a practical, yet efficient approach to the assembly of very large EST/cDNA databases. We also provide SNP and splice-variant detection.

2 Method and Results

Hyseq's software uses a third-party intermediate assembly engine. The package takes a cDNA/EST database as its input. The output contains assembled genes (or gene segments when a complete gene assembly is not achievable), and potential SNP and splice variants. Parameters related to sequence comparison, assembly, and SNP and splice variant detection, can be specified by the user. Those parameters determine the quality and accuracy of the output; reasonable parameters for most situations are suggested in the documentation. Hyseq's software package also provides ways around some of the limitations usually associated with the third-party software involved, such as high system memory requirements and difficulties in assembling genes with high sequence coverage.

In the simplest scenario, where overlapping fragments of a full-length gene with no SNP or splice variant are contained in the database, our program starts with a single cDNA/EST seed segment. The seed is extended by its significant homologous hits in the database, and is then assembled into an intermediate contig. This routine is called until every fragment belonging to the gene is assembled. The output is updated at each such point so that the program can be continued after an interruption.

The Hyseq system offers a much fuller set of convenient features than any other package now available. It features include:

1. Significant speedup of the clustering/assembly process. Its computational complexity (the total number of alignment/assembly cycles to run) is less than $O(n^2)$, where n is the total number of entries in the database. Other methods have $O(n^2)$ complexity.

2. Ability to distribute processing power over any UNIX network, especially an economical LINUX-cluster configuration. Many other packages cannot be distributed or require expensive, dedicated programming efforts to distribute them across a network.

3. Handling any size set of input cDNA/EST sequences. Essentially there is no upper limit: one can easily assemble over 10 million sequences.

4. It can terminate at any point of the assembly, and continue to run when a machine becomes available. There is no need to panic and redo everything from the beginning when you accidentally trip over the plug.

5. Detecting SNPs and splice variants as a by-product at the same time as it assembles.

6. Screening against any family of genes the user wants to exclude during the assembly process, for example the immunoglobulin and T-cell receptor families, or a collection of housekeeping genes that are of no interest to the user.

Hyseq's proprietary software includes a suite of flat-file format database utility tools, such as indexing, retrieving, insertion, deletion, and updating. Those tools can be used on FASTA formatted sequence and quality files, as well as on any other formatted files, such as table-formatted files, and concatenated sequence alignments.

References

[1] Boguski, M.S. and Schuler, G.D., Establishing a Human Transcript Map, *Nature Genetics*, 10:369–371, 1995.

[2] Green, Phil., PHRAP Documentation, Available at *http://www.phrap.org/phrap/docs/phrap.html*, 1996.

[3] Hide, W., Burke, J., Christoffels, A., and Miller, R., A novel approach towards a comprehensive consensus representation of the expressed human genome, *Genome Informatics*, 8:187–196, 1997.

[4] Sutton, G., White, O., Adams, M.D. and Kerlavage, A.R., TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects, *Genome Science and Technology*, 1:9–18, 1995.