A Hybrid, Recursive Algorithm for Clustering Expressed Sequence Tags in *Chlamydomonas reinhardtii*

Monica Jain[1] Hilary Holz[2] Jeff Shrager[1] Olivier Vallon[3] Charles Hauser[4] Arthur Grossman[1]

Abstract

We present an efficient, fully automated algorithm to assemble ESTs into full-length cDNA sequences that represent the complete coding regions of a gene. Our EST clustering algorithm is neither hierarchical nor incremental, but recursive, processing each EST once. The algorithm exploits a variety of syntactic and statistical features of the ESTs. The resulting assembly shows significant improvement in computational efficiency and information extraction over a previous assembly of C. reinhardtii ESTs. The algorithm was developed using iterative and participatory design on C. reinhardtii; however, it can be used for any organism with a draft genomic sequence.

1. Introduction

Expressed sequence tags (ESTs) are short sequences derived from cloned DNA molecules that represent the transcribed regions of a genome. Several hundred thousand ESTs have been sequenced by EST study projects in the past two decades. EST study projects are now working on constructing *gene indices*. A gene index is a set of transcripts representing the same gene. Assembling ESTs into gene indices via exhaustive search is computationally prohibitive, however, clustering can render the problem tractable. EST sequence assembly algorithms must be specific enough to ensure that only sequences of a unique gene are clustered together.

Early EST clustering procedures (*e.g.*, [1]) used tiered clustering, rapidly developing rough clusters then reprocessing the data to refine the clusters. As these algorithms refine clusters, however, they exploit progressively less definitive information. Accurately clustering ESTs depends on exploiting verifiable criteria to assign cluster membership.

Manufacturing a consensus sequence from each cluster is preferable, although more difficult. The majority of research projects that involve EST clustering employ readily available 'shotgun' assemblers (*e.g.*, [2]) for this step. These tools work very well on small-scale data (<< 100,000 ESTs.)

While many algorithms have been proposed for EST clustering, only a few have been incorporated into a complete system to construct gene indices. d2_cluster [3] begins with each sequence in a separate cluster and merges the clusters using transitive closure. Shrager, *et. al.,* [4] used an iterative process to assemble *Chlamydomonas reinhardtii (C. reinhardtii)* ESTs into unique genes.

We have developed an efficient, fully automated algorithm to assemble *C. reinhardtii* ESTs into full-length sequences that represent the complete coding regions of a gene. Our EST clustering algorithm is neither hierarchical nor incremental but recursive, processing each EST once, exploiting a variety of syntactic and statistical EST features. We developed the algorithm using iterative and participatory design on *C. reinhardtii;* however, it can be used for any organism with a draft genomic sequence.

2. Background

2.1 Expressed Sequence Tags

DNA sequencing finds the precise order of the four bases (A, C, G and T) that comprise a DNA molecule. Through sequencing, huge volumes of sequence information are now being stored as character strings. Sequencing requires the isolation of millions of copies of a population of cDNAs. (cDNA, or complementary DNA, is synthesized in the lab from a messenger RNA template using an enzyme named reverse transcriptase. The term *genomic DNA* refers to the entire gene with

[1] {mojain, jshrager, arthurg}@stanford.edu. Dept. of Plant Biology, Carnegie Inst. of Washington, 260 Panama Street, Stanford, CA 94305
[2] hilary.holz@csueastbay.edu. Dept. of Math & Computer Science, CSU, East Bay, 25800 Carlos Bee Blvd, Hayward, CA, USA 94542
[3] ovallon@ibpc.fr. 7141, CNRS/Université Paris 6, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris FRANCE
[4] chauser@duke.edu. Biology Department, Duke University, Durham NC 27708

coding and non-coding regions, while the term cDNA refers only to the coding regions of the gene.)

Copies can be generated for cDNA sequences by a process called *cloning*. In cloning, cDNA molecules are inserted into cloning vectors called plasmids, generating specific recombinant clones. The plasmid can undergo thousands of rounds of replication in a bacterial host, thus amplifying the clone and generating millions of copies of the inserted cDNA sequences.

Genomes vary in size from millions (e.g., bacteria) to billions (e.g., humans) of nucleotides. Current sequencing methods cannot determine the order of more than 1000 bases at a time. Therefore, the cDNA inserts are sequenced from both ends (called 5' and 3'), yielding a pair of specific sequencing reads, or *Expressed Sequence Tags* (ESTs), for each clone. Sequencing the 5' and 3' ends of multiple clones representing the same transcript allows for the assembly of the sequence information into full-length or near full-length cDNAs that can ultimately be used to establish a set of unique coding sequences present on the genome.

2.2 EST Clustering

The EST reads are assembled into long consensus sequences called *contigs.* The contigs are then used to analyze the structure of the genome and the function of each of the genes present on the genome.

Assembling ESTs requires accurate clustering of high volume sequence information. An EST database consists of sequences associated with a collection of ESTs derived from multiple cDNA libraries. This sequence collection is often highly redundant. The ESTs should be partitioned into clusters such that ESTs from each gene are put together in a distinct cluster.

In EST assembly, overlapping ESTs within a cluster are aligned to form a consensus sequence much larger than an individual EST. EST assembly generates long stretches of the coding regions of genes derived from several different, but often overlapping, sequences.

Identifying related ESTs is very difficult, as all ESTs are composed of just four characters and have approximately the same length. Historically, the results of the cluster analyses have required manual curration by biologists due to potential repeat sequences in the genome and the subtle nature of the information conveyed by the sequence data.

Factors in the computational costs of EST assembly include: repetitive sequences (many naturally-occurring DNA sequences are present on the genome as repeats of varying sizes); experimental errors (e.g., base substitutions, insertions, and deletions); sequencing errors (e.g., insertions/deletions of DNA bases, errors due to repetitive elements within DNA sequences and statistical errors in base positioning); and gene families (separating the ESTs for two or more closely related genes is problematic.)

2.3 Scaffolding: a draft genome sequence

The Joint Genome Institute has generated a draft genomic sequence for *C. reinhardtii* [5]. The draft sequence covers over 90% of the genome and is distributed over a number of "scaffolds" of various lengths. Version 2.0 consists of 3211 scaffolds, of which 72 make up 50% of the genome. The largest scaffold covers 2,159,859 nucleotides. Many scaffolds contain sequence gaps due to incomplete coverage or sequence assembly artifacts. Therefore, the gene models generated from the sequence information do not precisely represent the complete set of *C. reinhardtii* genes. A more refined genome sequence with only a few regions of known length that have not been sequenced can only be achieved through dedicated efforts involving expensive physical mapping and gap closure procedures. Unless technological breakthroughs simplify the arduous task of gap closure, more and more eukaryotic genomes are likely to linger at an advanced draft stage.

3. Methodology

Development of the algorithm required manual intervention because definitions for optimal assembly are varied and imprecise, and because optimal solutions may not have minimal error rates. Once we finalized the parameters of the algorithm, the process was completely automated. We present an overview of our algorithm here; see [6] for complete details.

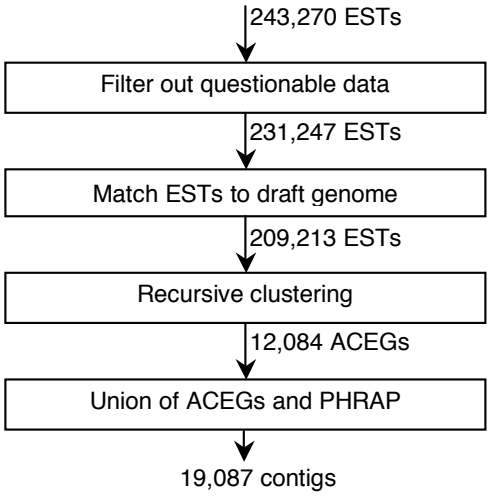

Figure 1: Overall algorithm

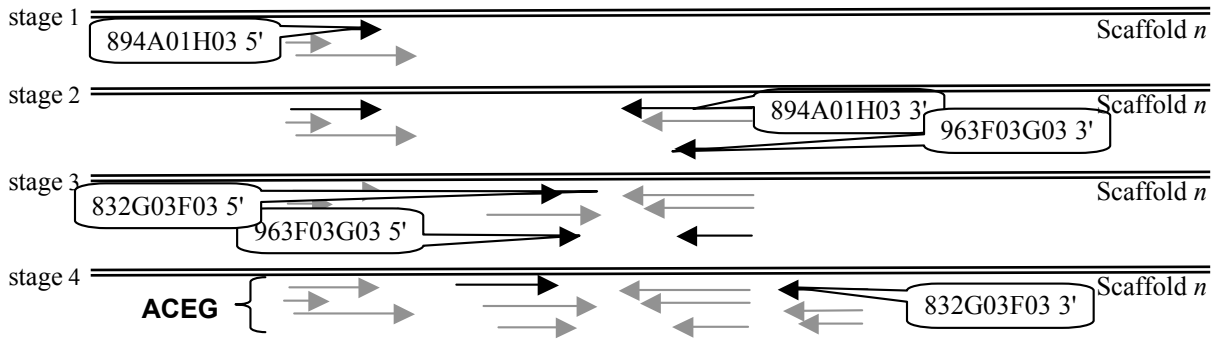

Figure 2. Recursive EST clustering algorithm

We collected 243,270 ESTs from a variety of sources. The heart of the algorithm is the generation of EST clusters using *recursive* clustering. Clusters are formed such that each cluster contains all the EST sequences from a single gene. Before forming clusters, we built an automated tool to convert the ESTs to a single format and to filter out reads that are likely to contain errors: poor quality ESTs; short read ESTs; vector sequences and ambiguous bases. The filter dropped approximately 5% of the sequences.

3.1 Match Against Genome

The ESTs are matched against the draft genome using BLAT [7]. BLAT matches are scored as follows:

$$100 \times \left(matches - mismatches - \frac{repeats}{length(EST)} \right)$$

where repeats are bases covered by another match in the same or another scaffold.

An EST was dropped if the EST mapped to different scaffolds with scores of ±10% or the EST mapped to the same scaffold with a score of ±10%. An EST sequence may match to more than one scaffold if it represents a family of closely related genes or contains a repeat element. Each EST is then paired with its matching genomic sequence and the combined sequence information is designated a HSP (Highest Segment Pair). A single EST sequence will generate more than one HSP if a non-coding region interrupts the corresponding genomic sequence. The EST is dropped if the sum of the HSP lengths is less than 75% of the EST length. All HSPs derived from a single, continuous EST sequence are fused to generate the corresponding genomic sequence. Matching against the genome eliminated approximately 14% of the EST reads. The sizable loss results primarily from large gaps and regions of low quality genomic sequence information.

Matching the ESTs to the genomic sequence increased the statistical power of the data by doubling the sequence information. (Note that errors are also present in the draft genomic sequence, thus the two sequences are best viewed as two samples, rather than one sample and truth).

3.2 Clustering

Each EST is an object to be clustered in the *feature space*. We combine the *genomic position* of the ESTs and the EST *clone IDs* to generate a novel measure of EST similarity. The genomic position identifies overlapping ESTs. The clone ID identifies 3' and 5' ESTs (reads) that belong to the same cDNA clone.

The clustering algorithm starts by randomly selecting a 5' read and assigning it to a cluster. All of its overlapping 5' reads are also added to the cluster (stage 1, Figure 2), as measured using genomic position. The minimum overlap between two ESTs to be placed in the same cluster was 20 bases. Next, the 3' reads corresponding to the 5' reads are added to the cluster unless they are more than 20,000 bases away from their 5' read (stage 2, Figure 2.) As most genes have fewer than 10,000 bases, separation by 20,000 bases suggests an error.

The clustering algorithm is recursive: for each new 3' read, the overlapping 3' reads and corresponding 5' reads are added to the cluster, effectively adding 5' reads that belong to the same clone but were omitted during prior steps because they did not overlap (for example 963F03G03 5' in Figure 2 was not included in stage 1 because of the lack of the overlap). We continue to call the algorithm recursively on each added 3' read and 5' read, until no more additional reads are placed into the cluster (stages 3 and 4, Figure 2). We call the resulting cluster an ACEG (Assembly of Contiguous ESTs verified on Genome.) Once an ACEG is complete, all the corresponding reads are removed from the dataset. The process is then repeated for another randomly selected 5' read. The process continues until all the reads are assigned to some ACEG. Thus, each ACEG consists of ESTs that are

likely to belong to the same gene. A total of 12,084 ACEGs were generated from 209,213 EST sequences.

3.3 Union of Clusters

ACEG length is calculated from the starting genomic position of the leftmost EST to the ending genomic position of the rightmost EST. ACEGs that overlap by more than 50% of the length of the shorter ACEG are fused into a single ACEG. The 50% limit is set in order to avoid fusing distinct but overlapping genes (genes often do overlap to some extent at their 3' ends). After fusion, there were a total of 11,387 ACEGs. Generating and fusing the ACEGs took ~15 minutes on a Pentium 4 with 1GB of RAM.

3.4 Consensus Finding

ACEGs were assembled into contigs using PHRAP. PHRAP combines user-supplied and internal data quality information to improve assembly accuracy in the presence of repeats. PHRAP assembles contigs as a mosaic using the highest quality parts of the reads (rather than a consensus sequence). PHRAP found 19,087 contigs in ~1 day on a Linux server with 3GB of RAM.

4. Results

We started with a total of 243,270 EST sequences, of which 209,213 were used in the final clustering; 5% of the EST sequences were dropped during quality filtration and 14% were dropped during generation of genomic sequences corresponding to ESTs.

Out of the 209,213 EST sequences, we generated 11,387 clusters (ACEGs). Each cluster was analyzed by the PHRAP assembly program, yielding 19,087 contigs. We have submitted the ACEGs to the ChlamyDB [8] and JGI [5].

An ACEG can have one contig, in which case it potentially represents a full-length cDNA (if it contains both 3' and 5' reads), or more than one contig, in which case it represents parts of the gene. In the current assembly, 56% of the ACEGs are composed of a single contig and the remaining 44% of the ACEGs are composed of two, three or four contigs. Of the 6,334 ACEGs that contain a single contig, 2,670 are composed of both 5' and 3' reads, while the remaining are composed of only 5' reads.

5. Discussion

Currently, there are nearly 20,000 gene models for *C. reinhardtii* predicted by JGI using the existing gene prediction programs. We evaluated the quality of ACEGs by comparing them to the existing gene models for *C. reinhardtii* using BLAST [9]. Out of 20,000 gene models, 8,728 were similar to 8,346 ACEGs.

Out of 11,387 ACEGs that we generated, 3,049 ACEGs were not associated with a gene model. This phenomenon can be a consequence of missing genomic information (there are still many gaps in the genomic sequence), low quality genomic sequence information, the inability of the gene model predictors to identify certain genes, or errors in ACEG generation because of limitations in the PHRAP assembly program. Those ACEGs of high quality that are not represented by gene models can be used to identify and predict new genes on the *C. reinhardtii* genome.

The number of ACEGs that are associated with gene models is less than the number of gene models associated with ACEGs because a single ACEG could map to two similar gene models that are on separate scaffolds. A number of the smaller scaffolds are really contained within larger scaffolds but they were not put together because of the low quality sequence data.

The current assembly results show significant improvement in computational efficiency and information extraction over the previous assembly of *C. reinhardtii* ESTs [4]. In addition, the current, recursive approach can easily be optimized for parallel processing because each cluster can be processed independently of the others.

10. References

[1] L.D. Hillier, *et. al.,* "Generation and analysis of 280,000 human expressed sequence tags." *Gen Res 6*: pp. 807-828.

[2] PHRAP, http://www.phrap.org/phredphrap/phrap.html

[3] J. Burke, D. Davison, and W. Hide, "d2_cluster: a validated method for clustering EST and full-length cDNA sequences," *Gen Res 9* (11), pp. 1135-1142.

[4] J. Shrager, et. al., "Chlamydomonas reinhardtii Genome Project: A guide to the generation and use of the cDNA information.," *Plant Physiol 131*: pp. 1-8.

[5] http://genome.jgi-psf.org/chlre2/chlre2.home.html

[6] M. Jain, "Computational Algorithms for the EST Clustering Problem to Identify Unique Genes," MS Thesis, CSU, East Bay, 2005.

[7] W.J. Kent, "BLAT -- The BLAST-Like Alignment Tool," *Gen Res* 4: pp. 656-664.

[8] ChlamyDB, http://www.chlamy.org/chlamydb.html

[9] A. Altschul, et. al., "Basic Local Alignment Search Tool," *Journal of Molecular Biology 215*: pp. 403-410.