EST Assembly for the Creation of Oligonucleotide Probe Targets

Charlie Nelson
Agilent Technologies, Inc.

1. ABSTRACT

For several years now, Expressed Sequence Tags (ESTs) have provided researchers with a low cost, yet extensive survey of the genomes they were derived from. ESTs have proven to be invaluable in the research domains of structure prediction, gene discovery, and genomic mapping for many species. While most of the key mammalian model organisms have had their transcriptomes defined, ESTs continue to provide value in detecting splice variants, and in the exploration of poorly characterized transcriptomes. Researchers that are interested in utilizing array-based gene expression technologies need to be able to confidently predict RNA transcripts from ESTs. Here we discuss methodologies to generate EST assemblies with the purpose of generating transcript representatives that can be used as sources of high quality sequence for oligonucleotide probe design.

2. INTRODUCTION

2.1 EST Problem Overview

Expressed Sequence Tags (ESTs) are 200-500 bp sequences that are obtained as part of a 3′ or 5′ single pass read of individual clones (1). These clones are derived from cDNA libraries that may be specific to a tissue and/or development state of an organism, which can be used to identify expressed genes that are considered "rare" when exploring the overall expression of an organism's transcriptome. These rare transcripts may have splice variants and/or alternative polyadenylation (2). ESTs are popular primarily because they can be generated in high volume, using a high-throughput data production method at low cost. Despite serving as a rich source of sequence information as a survey of a transcriptome, ESTs must be used with an understanding of their limitations. The principle problems are that EST sequences contain errors, represent only a portion of a gene product, and are found in vast, highly redundant datasets.

ESTs contain sequencing errors (sequence compression and frame-shift errors) at high rates (3%) due to the single pass read process they were generated from (3). These errors do not follow a normal distribution along the length of the sequence, but rather are biased toward the start and end of the sequence (4), leaving EST base pair positions 100 to 300 to be the most accurate part of the EST (5).

Each EST in itself represents only a portion of a gene product. Being 200-500 bp means that there are potentially several thousand base pairs of the underlying transcript that are unrepresented in the EST sequence. The attenuation of the sequencing reaction used in generating the EST leads to a length that is shorter than the cDNA clone it was derived from. EST single pass reads can be generated from the 5′ or 3′ end. ESTs can also be generated using random primers, which may result in ESTs with ambiguous orientation, from different parts of the same, non-overlapping RNA (6).

Databases that contain EST data also have two primary limitations. 1) ESTs, as a whole, are poorly annotated, both in terms of source and sequence quality, which makes it difficult to determine what gene product a given EST represents. In addition, 2) EST databases contain a huge number of sequences, at a high rate of redundancy, which makes it difficult for a researcher to negotiate and derive concise value from.

2.2 *Why Cluster and Assemble?*

The value of ESTs with respect to microarray target creation is greatly enhanced by utilizing a process of clustering and assembling. The primary goals of sequence clustering and assembling are to 1) reduce redundancy and mass of the EST dataset; 2) increase the sequence quality of the reduced data; and 3) create a contiguous sequence from ESTs that approximate the cDNA clone from which the ESTs were derived.

In principle, *clustering* is a grouping of like entities on one or more dimensions. These groupings allow for the creation of partitions that provide a binary definition of ‘self’ and ‘non-self’ for any given entity. With respect to nucleic acid sequence, clustering is often done on the dimension of sequence *similarity*, were sequences that have a defined degree of similarity are considered part of the same cluster. There are as many ways to define similarity, both in terms of variables and approach, as there are clustering applications that employ them. We will not discuss the specific definitions of sequence similarity in this paper, however.

The clustering of gene sequence is often referred to as *gene indexing*, where all data concerning a single gene or gene isoform is partitioned into a single *index class*, and each index class contains the information for only one gene (7). Gene indexes can be used to define *exemplar* transcripts (transcripts that provide the best available representative for a given cluster of transcripts), as well as provide a basis for an assembly, which would create *consensus* transcripts. The UniGene dataset provided by NCBI is a gene index that does not contain sequence assemblies, but rather represents the exemplar sequence of a given cluster in its *Unigene Unique* set (8). The TIGR Gene Indices, on the other hand, represents assemblies called tentative consensus (TC) sequences, which are meant to approximate the underlying mRNA transcripts (9).

Clustering and assembling methodologies can be used to define sequence strings that reliably represent transcripts, which can, in turn, be used to define probe sequences for an Oligonucleotide array.

2.3 Clustering Types

Sequence clustering methodologies can be categorized into three primary groups; *Unsupervised, Semi-supervised, and Supervised*.

Unsupervised Clustering (also called non-seeded clustering) is a computationally intensive process where a N x N comparison is made between all ESTs, and all ESTs with a defined degree of similarity to each other are grouped into the same partition. With this methodology, there is no pre-conceived notion as to the number of clusters that will be formed, and there is a higher chance of clustering error because the data content of the sequences are generally poor. Because this methodology has the greatest chance of error, only researchers that only have EST data for their transcriptome would principally use this methodology.

Semi-supervised Clustering includes well-characterized reference objects in the EST data set. These reference objects are often well-characterized sequence (e.g., full-length mRNA), which serve as an implicit scaffold against which an EST can cluster. The primary role of such a methodology is to leverage the length of known sequence in serving as a bridge for 5′ and 3′ EST sequence when creating a representative transcript. Semi-supervised clustering helps in determining which ESTs already are represented by a full-length transcript, and which ESTs are derived from novel, or unknown transcripts. Despite the fact that the full-length transcripts have higher quality information content than the ESTs, they are not given any special weighting in the Semi-supervised clustering process, simply relying on the inherent value as long, well characterized sequence.

Supervised Clustering (or seeded clustering) utilizes reference objects in much the same way that Semi-supervised clustering does, however these reference objects are given explicitly defined value within the sequence data set. These reference objects define the boundary for clusters. Reference objects in Supervised Clustering can either be well-characterized sequence (e.g., full-length mRNA) that act as seeds during the clustering process, or can be partial sequences that have clearly defined transcript features (polyadenylation sites, poly A tail, etc.) that define the edge of a cluster, and ultimately the assembled, consensus transcript. Edge definition helps to prevent the merging of sequences that are actually represent different transcripts, as well as prevent the creation of chimerical assemblies. Defining explicit reference objects can also help to differentiate ESTs that belong to different splice variants.

3. A GENERAL CLUSTERING AND ASSEMBLY METHODOLOGY FOR ESTs

3.1 Sequence Preparation

With any EST set, it is important to prepare the data for similarity searching. The goal of preparation is to 1) eliminate all sequence substrings that that have no or low probability of existing in the biological sample (contaminants), and 2) eliminate all sequence substrings that may lead to spurious results in similarity searching.

Contaminants may be cloning vectors (e.g. plasmid, BACs, YACs); adapters and linkers; transposons; and other impurities. During an EST cluster and assembly, present contaminants can lead both false positives and negatives during the similarity search process. These contaminants can be identified and masked using NCBI's Univec database (10). The low quality reads of the 3' and 5' end of ESTs are candidates for masking as well.

Repetitive elements and low complexity regions provide a problem for sequence alignments used in clustering. Repetitive elements (interspersed repeats) are transcript sequence substrings found to be similar among phylogenetically related organisms, and decrease in similarity as species diverge. Within repeat databases, repeats are often represented at the taxonomic Class or Order level. Low complexity regions and simple repeats are the primary cause of false positives in a similarity search. Low complexity sequence is a more general term for stretches of DNA with or without detectable repetitive structure.

After conducting vector and repeat masking, sequences that contain a contiguous base pair stretch that falls short of a defined minimum, should be removed from the data set before sequence assembly. These resulting sequences may contain low information content, and lead to spurious results during the clustering and/or alignment procedures. For example, as part of the Unigene build procedure, NCBI determines that a sequence must contain at least 100 informative bp to be a candidate for entry into Unigene (11).

3.2 First-Pass Clustering

The goal of *first-pass clustering* is to partition a large input sequence set into reasonably smaller sequence subsets that can be processed by the assembly program without running out of memory (12). In addition to creating manageable data sets, this initial clustering reduces redundancy; while at the same time increases base position confidence for each sequence assembly that is generated from the cluster. The first-pass cluster and assembly process should generate a reduced and reliable dataset that has a low chance of containing type I errors.

First Pass Clustering should be highly stringent, using an application (such as MegaBLAST) that finds similarity between sequences that have only slight differences. Slight differences observed between sequences have a greater chance of resulting from sequencing errors, not evolutionary divergence (13). In a pair-wise alignment, high stringency is observed when the two sequences being compared have 95% identity and 90% overlap for both sequences. If sequence similarity is being defined with BLAST, *E* value scores of 10^{-4} could also be considered a stringent cutoff value (14).

3.3 *First-Pass Alignment and Assembly*

The goal of first-pass alignment and assembly is to align sequences found within each of the clusters, and to create a consensus sequence for each cluster. This process can be conducted with, or without the use of Sequence Quality Scores, which are generated during the conversion of the raw, continuous chromatogram data into discrete EST sequences (15). Most public domain EST sequence information is not provided with quality scores, which makes the use of highly redundant EST sets useful in eliminating errors by determining a consensus for each base position, after an alignment has been performed. EST Assemblies can be conducted using an assembly application such as CAP3 (16), or PHRAP (17), which both can use quality score information during the assembly process.

3.4 *Second-Pass Clustering*

The goal of *second-pass clustering* is to partition the first-pass assemblies into groups of sequence that potentially belong to the same transcript. Second pass assemblies should utilize a high identity threshold (95%) but should allow for less overlap (40-70%), particularly for ESTs that are generated using a random primer methodology. If the ESTs contain potential splice variants, it may be best to use an iterative second pass clustering process that uses an ever-lessening overlap threshold.

3.5 *Chimerical Sequence Detection and Removal*

Chimeric transcripts are sequences that are hybrids between heterologous mRNAs, which can be generated by a variety of molecular mechanisms. Analysis of NCBI's Unigene database has revealed that ~1% of all transcripts within the dataset contain chimeric sequence (18). Chimeric sequence is extremely problematic for EST assembly procedures in that they may cause truly unique transcript clusters to join together. Because of this, Identifying and removing these sequences is an important part of the EST clustering and assembly process.

Because chimeric sequences are anomalous, there should be few or no such sequences within any given cluster. Identifying and removing putative chimeras may be a process of simply extracting all potential chimeras from the EST dataset, applying a rule set to refute or acknowledge their chimeric nature, and remove them from the EST clusters. Potential Chimeras could be found by analyzing the *chimerism point* of each cluster; that is the weakest subsequence of a cluster relatively to alignment depth (19). Once the chimerism point has been identified, it can be tested for the presence of internal poly A stretches and polyadenylation sites, and even endonuclease restriction sites. Chimeras can also be identified by genome comparison. These sequences would then be removed from the cluster, and a reclustering step would be performed.

3.6 Final Assembly

Once clustering has been completed with all putative chimeric sequences removed, a final assembly can be constructed that contains transcript targets against which oligonucleotide probes can be designed. The alignment and assembly process can be similar to what was conducted for first-pass alignment and assembly, however sequence quality data would no longer be available for use, unless you choose to average base positional scores for the first alignment and assembly that was performed.

3.7 Target Dataset Construction

After all ESTs have been assembled into high quality, non-redundant putative partial or full-length transcripts, an additional step is necessary to construct the target data set for probe design. If 5′ derived EST were included in the dataset, and their assemblies include no 3′ EST sequences, the 5′ EST assembly should be removed from probe design consideration if a 3′ biased linear amplification sample preparation methodology is to be employed. Sequences assemblies that were derived from small clusters should also be potentially removed from probe design consideration, as they may not represent an actual transcript.

If the assembled ESTs belong to a well characterized genome, conducting a similarity search against genomic sequence, or even protein sequence can help to validate the presence of the putative transcripts within the transcriptome of interest.

4. CONCLUSIONS

Designing probes to assembled ESTs, as opposed to raw EST sequence, can significantly improve the performance of the microarrays containing them. There are many methodologies that can be used in constructing EST assemblies, however most are variations on the cluster and assembly paradigm. Incorporating full-length, well-characterized transcript sequence into the EST data set increases overall sequence quality, which can have a positive impact on the creation of putative transcripts, against which oligonucleotide probes can be designed.

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science. Jun 21; 252(5013):1651-6

2. Gauntheret, D., Poirot, O., Lopez, F., Audic, S. and Claverie, J.-M. (1998) Alternative Polyadenylation in Human mRNAs: A Large-Scale Analysis by EST Clustering. Genome Res. 8:524-530

3. Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993) dbEST – database for "expressed sequence tags." Nat. Genet. 4:332-333.

4. Ewing, B., and Green, P. (1998) Genome Res. 8: 186-194.

5. Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., More, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Marra, M., and et al., (1996) Generation and analysis of 280,000 human expressed sequence tags. Genome Res. 6:807-828.

6. Kapros, T., Robertson, A.J., and Waterborg, J.H. (1994) A simple method to make better probes from short DNA fragments. Mol. Biotechnol. 2(1):95-8.

7. Burke, J., Davison, D., and Hide, W. (1999) d2_cluster: A validated method for clustering EST and full-length cDNA Sequences. Genome Res. 9:1135-1142.

8. Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information; 2003.

9. Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000) The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. Nucl. Acid Res. 28:141-145.

10. http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html.

11. Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information; 2003.

12. Liang, F., Holt, I., Pertea, Ge., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. (2000) An optimized protocol for analysis of EST sequences. Nuc. Acid Res. 28(18):3657-3665.

13. Zhang, Z., Schwartz, S., Wagner, L. Miller, W. A greedy algorithm for aligning DNA sequences." J. Computational Biology (2000) 7:203-214.

14. Schmitt, A.O., Specht, T., Bechmann, G., Dahl, E., Pilarski, C.P., Hinzmann, B., and Rosenthal, A. (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. Nuc. Acid Res. 27(21):4251-4260.

15. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998) Base-calling of automated sequencer traces using Phred I: Accuracy assessment. *Genome Res.* **8**: 175-185

16. Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868-877

17. Green, P., 1994, Phrap, unpublished. http://www.genome.washington.edu/

18. Romani, A., Guerra, E., Trerotola, M., and Alberti, S. (2003) Detection and analysis of spliced chimeric mRNAs in sequence databanks. Nuc. Acid Res. 31(4):e17.

19. Just, J., Barillot, E., Legeai, F., Duclert, A. (unpublished) Bellerophon: Graph analysis for improved EST clustering.

Printed in the U.S.A.

Agilent Technologies
Bioresearch Solutions Unit
3500 Deer Creek Road
Palo Alto, CA 94304
E-mail: dna_microarrays@agilent.com

Agilent Gene Expression Microarrays Website:
www.agilent.com/chem/dna

Part Number: 5989-0750EN June 1, 2004

Agilent Technologies