Using the Human Genome as a Framework for Sequence Clustering and Microarray Design

Barbara Lin[1], Timothy Burcham[1]

Keywords: human genome, BLAT, sequence, alignment, microarray design

1 Introduction.

At diaDexus, Inc., we utilize genomic and bioinformatic tools to identify and validate cancer-associated molecular targets for diagnostic and therapeutic applications. These gene discovery efforts generate large amounts of sequence data on a daily basis, which poses significant challenges to organize and eliminate redundancy between sequences. One approach commonly used to group similar sequences is to BLAST the sequences against one another. However, BLAST results can be hard to interpret and difficult to reproduce, and may be inconclusive with large sequence sets. With the completion of the human genome and its sequence information publicly available, a reliable, stable framework now exists and can be utilized to classify and structure large sequence sets. We have used the genome as a means of organizing sequence data, and have developed an algorithm to quickly isolate non-overlapping clusters of sequences. These clusters allow us to quickly find the most prevalent forms of sequences that are of interest to disease target discovery, and also enable a systematic approach for oligonucleotide microarray design.

2 Software and Method.

Using the human genome (Build 30-34) as a template, we developed a process to identify a unique set of sequences from a highly redundant set of sequences. This process required an ability to identify the highly similar sequences, but not identical, sequences in an efficient and reproducible manner. We used the resulting data to identify a high quality, non-redundant set of oligonucleotides to print on microarrays for gene expression profiling.

First, approximately 263,000 sequences of specific interest to target discovery, each having different quality and length, and generated from different sources and methods, were aligned to the human genome using BLAT[2]. Out of the 263,000 sequences, about 16,000 sequences were from the Human Refseq mRNA database, 23,000 were Ensembl genes, 129,000 were from UniGene, and 95,000 were either generated as a result of internal sequencing efforts or from other proprietary sources. Sequences were allowed to map to more than one location on the genome and the coordinates of each mapping were stored in a relational database for downstream processing.

After the BLAT alignment, we developed and implemented an algorithm to cluster sequences based on their exact coordinates and locations on the genome. In the first step in the algorithm, "chromosome walking", we iterate through every chromosome and walk each chromosome from one end to another, in both orientations, defining and separating stretches of overlapping sequences into individual "super-clusters". Sequences are grouped together if their beginning and ending chromosomal coordinates overlap one another. After chromosome walking, we were able to very rapidly reduce the complexity of the set of the 263,000 sequences, and further group the sequences into approximately 80,000 bins or super-clusters.

[1] diaDexus, Inc, 343 Oyster Point Blvd, South San Francisco, CA 94080. E-mail: tburcham@diadexus.com

After the first step, many super-clusters contained sub-clusters which had no overlap with the parent clusters. For example, a small sub-cluster of sequences would be intronic to a super-cluster whose chromosomal coordinates encompassed the sub-cluster. The second step of the algorithm, "exon walking", was used to identify all sets of truly overlapping sequences from the results generated by chromosome walking, based on the coordinates of the exons within each super-cluster.

Using exon walking, sequences with non-overlapping exons are separated out of the chromosome clusters and given their own bin or cluster. This is accomplished by taking all the sequences in a super-cluster, and for each sequence, creating a virtual sequence in which every nucleotide is characterized as a bit. The nucleotides that are within an exon are represented as 1's, and those that fall outside exons are represented as 0's. By imposing a logical bitwise AND relationship between the virtual bit sequences, we were able to quickly and efficiently identify the sequences with overlapping regions. Approximately 90,000 bins or clusters resulted after exon walking, a number significantly less than the 263,000 we started out with.

We were then able to use the resulting clusters to identify a non-redundant, high quality set of oligonucleotides to print on microarrays. About 38,000 pre-designed oligonucleotides[1] from proprietary sequences were melded into the clusters obtained from the method described above. We then prioritized the clusters them based on a combination of factors. The criteria that we looked for in characterizing these clusters were:

1. Clusters with a public annotated gene, transcript, or EST from RefSeq, Ensembl, or UniGene.
2. Clusters with multiple exons.
3. Clusters with multiple sequences (non-singleton clusters).
4. Clusters containing oligos with measured, disease-specific, expression in our in-house expression profiling studies.

Using the criteria above, we identified 19,000 "high priority" cluster bins. After the high priority clusters were selected, we selected the best, most representative, oligo from each cluster for array design. This algorithm will be discussed in more detail in the poster, but generally we printed the most 3′ oligo that most generally represented the sequences in the cluster, and for which we had good previous expression results. These oligos were then printed onto several custom oligonucleotide microarrays and are currently in use for cancer-target discovery and validation.

In conclusion, we were able to use the genome a powerful framework to organize and efficiently cluster a large sequence set into individual sequence clusters. Using these clusters, we were able to design high quality microarrays with minimally redundant oligonucleotide sequences.

3 References.

[1] Hughes, T.R., *et al.*, 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 19: 342-347.

[2] Kent, W. J. 2002. BLAT —The BLAST-Like Alignment Tool. *Genome Res.* 12: 656-664.