Designing oligo libraries taking alternative splicing into account

Avi Shoshan[a], Vladimir Grebinskiy[a], Avner Magen[b], Ariel Scolnicov[b],
Eyal Fink[b], David Lehavi[b] and Alon Wasserman[a]

[a]Compugen Inc., 7 Centre Drive, Jamesburg, NJ 08831, USA
[b]Compugen Ltd., 72 Pinchas Rosen St., Tel-Aviv 69512, Israel

ABSTRACT

We have designed sequences for DNA microarrays and oligo libraries, taking alternative splicing into account. Alternative splicing is a common phenomenon, occurring in more than 35% of the human genes. In many cases, different splice variants have different functions, are expressed in different tissues or may indicate different stages of disease. When designing sequences for DNA microarrays or oligo libraries, it is very important to take into account the sequence information of all the mRNA transcripts. Therefore, when a gene has more than one transcript (as a result of alternative splicing, alternative promoter sites or alternative poly-adenylation sites), it is very important to take all of them into account in the design. We have used the LEADS transcriptome prediction system to cluster and assemble the human sequences in GenBank and design optimal oligonucleotides for all the human genes with a known mRNA sequence based on the LEADS predictions.

Keywords: DNA microarrays, probe design, alternative splicing, EST clustering

1. INTRODUCTION

DNA microarrays have become a widely used technology for monitoring the expression of genes.[1–6] On these microarrays, cDNA clones or synthesized oligonucleotides are deposited in high density on a solid substrate. Oligonucleotide arrays used to be much more expensive than cDNA microarrays, but the diminishing costs of synthesizing oligonucleotides have changed the balance, and now more scientists are starting to experiment with oligonucleotides deposited on slides.

The problem of the choice of sequences for oligonucleotide arrays has not yet been extensively addressed. In order to reduce the redundancy in EST and mRNA databases, where multiple sequences correspond to the same gene, EST and mRNA clustering algorithms and databases are used. Many designs are currently based on UniGene[7–9] (`http://www.ncbi.nlm.nih.gov/UniGene/`). UniGene groups the EST and mRNA sequences into clusters according to their gene association and chooses a representative sequence from each cluster. The sequences for the oligonucleotides are selected, for each UniGene cluster, from the representative sequence.

While generally a good strategy, a lot can be gained by further analyzing the sequences in each cluster. Important phenomena, such as alternative splicing[10–12] and alternative poly-adenylation sites,[13,14] can only be detected by looking at the full set of gene sequences. It is important to take alternative splicing into account even when not trying to differentiate between the splice variants, because then a probe must be chosen from the segments which are common to all the variants. Problems affecting the clustering, such as chimeric sequences and paralogous genes, can be found and corrected using this deeper analysis. All these benefits result in better sequence choices for the oligonucleotides.

We have used the LEADS software platform for predicting the different transcripts for all genes. On that basis, we have developed tools for designing oligonucleotide probes. These tools use the predicted transcripts, well-known probe selection criteria[4] and additional criteria (described in section 3) for choosing optimal probes targeted either at differentiating between splice variants or at the detection of the overall expression levels of genes. These tools were used for selecting oligonucleotides probes for virtually all human genes that have a known mRNA sequence in GenBank. In section 2, we describe the LEADS transcriptome prediction software platform and its various stages. Section 3 details the way we use the LEADS-created transcript database in the design of sequences for oligonucleotide probes. Section 4 shows results and statistics related to a probe design for all human genes that have a known mRNA sequence in GenBank. We compare our designs to those obtainable from UniGene.

Corresponding authors: A.S.: E-mail: avi@cgen.com A.W. : E-mail: alon@cgen.com

2. THE LEADS TRANSCRIPTOME PREDICTION PLATFORM

For optimal oligo sequence selection, one needs to have the most accurate and complete information on the transcriptome of the organism in question. For that, we have used the LEADS platform. This software platform analyzes all the sequence information for a specific organism and produces a prediction of all the different mRNA transcripts for the genes which are represented by the data. LEADS models various biological and experimental phenomena, including alternative splicing, alternative poly-adenylation sites, repeats, vector contamination, sequencing errors, polymorphisms, chimeric sequences, genes with anti-sense transcripts and more.

2.1. The LEADS Process

The main stages of the LEADS process are the clustering and assembly stages. In the clustering stage, we take the sequences and use overlaps between them in order to distribute them according to their gene association. The goal of this stage is to come up with sequence clusters, such that all the sequences from one gene are in the same cluster and all the sequences in a specific cluster originate from the same gene. The assembly stage takes these clusters and analyzes the overlap patterns between the sequences in each cluster. The goal here is to calculate all the different mRNA transcripts of the gene. This is acheived by calculating the multiple alignment of all ESTs and mRNA sequences in each cluster. A consensus sequence is then created, and all the input sequences are aligned (allowing long gaps for alternatively spliced exons) to that sequence. A list of mRNA transcripts is created, each containing all or just a part of the exons. Although these two stages are the central stages, a naïve implementation would result in various problems and artifacts, causing wrong predictions. Therefore, it has become necessary to accompany these stages with preparatory and post-processing stages that deal with these problems.

Finding new repeats and vectors

The first stage of the LEADS process is the search for unknown repeats and vectors. Unknown vectors and repeats may cause the clustering of unrelated sequences, due to the occurrence of common vectors and repeats in sequences from different genes. Though we use the most up-to-date repeat and vector databases, these databases are incomplete, and therefore, it is worthwhile to search for undocumented repeats and vectors. This is done by examining the whole set of input sequences and looking for sequences that appear many times. Special care is applied to distinguish between repeats, vectors and abundantly expressed genes which also tend to appear many times in the sequence databases.

Cleaning

Using public, commercial and in-house created databases of repeats and vectors, we look for occurrences of these sequences in the data. Vectors are clipped because they are not part of the gene sequence. Repeats and low-complexity regions (periodic and aperiodic) are masked, so that they do not cause false positive associations in the clustering stage. Low-quality ends of sequences are trimmed and very short sequences are discarded.

Clustering

The goal of the clustering stage is to group the sequences according to the genes from which they originate. In this stage, all sequences are compared to each other and sequences with a significant overlap are clustered together. This is done using fast in-house developed algorithms. At the end of this stage, sequences from the same gene are grouped in the same cluster.

Sequences which form "weak links" within clusters are marked as suspect chimeras. These sequences are removed from consideration and the sequences from the different genes are separated into different clusters.

Assembly

The heart of the process is the assembly stage. At this stage, the sequences in each cluster are analyzed. A multiple alignment of all the sequences in each cluster is calculated, allowing long gaps. This results in a prediction of all the mRNA transcript sequences, based on the input sequences. When the gene has several splice variants, these produce several predicted transcripts. In sequence regions which are covered by more than one sequence, sequencing errors are detected and corrected. Ambiguities which result from internal repeats within the gene are resolved. Sequences which exhibit chimeric characteristics are detected and removed. Low-quality sequence ends which disagree with other sequences are trimmed. The outcome of this stage is a consensus sequence, which ideally consists of the concatenation of all the exons (when the gene is only partially covered by the sequences, this consensus is partial). In addition, the different exon combinations that constitute the different splice variants are reported. Figure 1 contains an example for a LEADS cluster, showing the outcome of the clustering and assembly process.

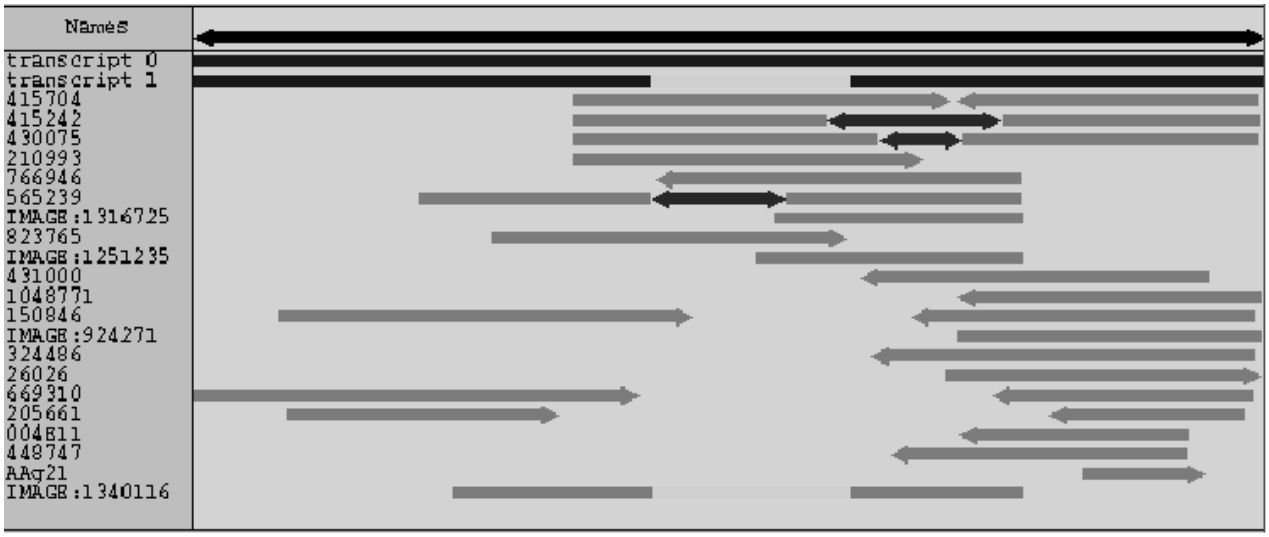

Figure 1. This is an example of a LEADS cluster with two predicted transcripts. The top dark line represents the consensus of the cluster, that is the concatenation of all the predicted exons. The next two lines represent the transcripts predicted for this cluster. The first transcript contains all the 3 predicted exons, and the second contains only the first and the third. The following lines represent the sequences in the cluster, as they align to the consensus, the transcripts and within themselves. ESTs from the same clone appear in the same line. Note that the last EST is the indication for the existance of the second transcript, as it contains the end of the first and the beginning of the third exon, but lacks the middle exon.

Genomic assembly

The consensus sequences of all the clusters are then compared with all the genomic sequences in order to map the gene sequences on the genome. Genomic sequences may span several genes and therefore may be associated with several clusters. All the mapped clusters are reassembled using the genomic sequence as a reference.

Genomic curation

Mapping sequences on the genome is one of the most useful tools for finding problems. The alignments of the expressed sequences on the genome are used to fix various artifacts of the previous stages. Chimeric sequences are identified and removed. Sequences from paralogous genes are separated. New repeats are found. Low-quality ends of sequences are trimmed. Clusters whose overlap is too small for clustering together but map on the same genomic region are joined. All affected clusters are reassembled to obtain transcript sequences of higher quality.

Genes with anti-sense transcripts

There is a significant number of human genes in which a certain genomic region is transcribed in both directions. This may be due to a regulatory mechanism, the existence of overlapping genes in the two strands,[15] the existence of an unknown inverted repeat or some other phenomenon. Such cases are recognized by the existence of a significant number of contradictions between documented strands. In such cases, the sequences of the different strands are assembled separately.

3. MICROARRAY DESIGN USING LEADS TRANSCRIPTS

The most common application of DNA microarrays is the measurement of expression levels of mRNA transcripts. This is achieved by hybridizing labeled cDNA with pre-chosen complementary sequences. In oligonucleotide microarrays, these sequences are short (25–80 bps) synthesized sequences taken from the gene sequence. Many factors affect the utility of these oligonucleotide "probes". The different considerations will be discussed in this section. We have developed a software system on top of the LEADS-predicted transcriptome which chooses sequences for the oligonucleotide probes.

3.1. The Relevance of Splicing Patterns

Each oligonucleotide probe acts as a detector for the presence of its complementary sequence. When a gene has several splice variants, each probe can only detect the expression of those mRNA transcripts that contain its complementary sequence. Therefore, the different splicing patterns of the gene should be considered when choosing a representative probe. This is done according to the design objectives, as described below.

3.1.1. Measuring the overall expression level

In many cases, one is first and foremost interested in the list of genes which are expressed in a specific condition or differentially expressed between two conditions. Knowing which of the splice variants is the one that is expressed is the next step to be pursued at a separate, more focused experiment. If one wishes to measure the overall expression level of the gene without distinguishing between the different splice variants, it makes sense to choose probes from the segments of the gene which are common to as many splice variants as possible.

3.1.2. Distinguishing between splice variants

It is becoming more and more accepted that alternative splicing is an important phenomenon.[11,12] In many cases, the different splice variants have a different function, are expressed in different tissues or developmental stages or may even indicate a disease stage. Thus it is not only important to know which genes are expressed in each condition, but also to know which of the splice variants is the one that is expressed. Often it is possible to achieve this using oligonucleotide microarrays.

Each segment of the gene takes part in a specific subset of the splice variants. This may be used in the choice of probes. For example, if a specific exon is present in only one splice variant, any probe from this exon can be used in order to measure the expression level of that particular variant. This results in the choice of multiple probes from a single gene, each from a different exon, chosen so that once the experiments are done, the results will enable one to distinguish between the expression levels of the different variants.

For example, assume a gene has three exons and two splice variants. The first includes all three exons and the second includes only exons 1 and 3. If we choose probes from exons 2 and 3, the probes from exon 2 will detect the specific expression of the first variant, while the probes from exon 3 will detect both variants.

3.2. Specificity

Every probe serves as a detector for a specific short sequence. If that sequence appears in several genes, the probe will not be specific and will light up in case any of these genes is expressed. It will then be impossible to know which of these genes is expressed. This is also true when there is a high similarity between the sequence of the probe and the sequence of another gene, because hybridization can occur even with a partial match.

Therefore, it is important to choose probes from the parts of the gene sequence which are as specific to that gene as possible. In order to measure that, it is important to compare each candidate probe with every possible continuous segment of any gene. For that, one has to consider all splice variants, because such segments can appear as part of exons which are not common to all the splice variants or at specific splice junctions. We note that because of that, it is important to consider the whole transcriptome even when designing probes for a small set of genes. Apart from having high overall specificity, it is also important for a probe not to have a significant continuous stretch which is non-specific. Another important thing is to be careful not to choose probes which have a high match with the sequence of a repeat.

We have incorporated extensive homology searches between all candidate probes and all the mRNA transcripts. Additional searches were conducted with a non-redundant assembly of the genomic sequences (the Oct. 7 2000 release at UCSC[16]), in order to detect probes which appear several times in the genome.

3.3. Physio-chemical Considerations

The chemical properties of the probes have a critical effect on its hybridization affinity. The best microarray is the one in which all the probes have the same affinity coefficient. This cannot be achieved or accurately measured by current computational models, but can be approximated by choosing probes from a narrow range of melting temperatures. Another important consideration is the probe's secondary structure. In order to avoid self-hybridization (hairpin loops), it is important to choose probes without palindromes. Additional local criteria[4] have been applied.

3.4. Advantages of Using LEADS for Microarray Design

It has become a tradition to use UniGene[7–9] for choosing probes for DNA microarrays. According to that methodology, a probe is chosen from the representative UniGene sequence (usually the longest 3' sequence in each cluster). Since DNA microarrays are used for monitoring the expression of mRNA transcripts, it is clearly essential to have the most accurate and complete information about the transcriptome when choosing the sequences for the different probes.

Alternative splicing complicates the picture since different probes from the gene sequence participate in different collections of transcripts. Knowing the splicing patterns, namely the different splice variants, is essential for choosing a probe that will detect all (or, when not possible, the maximal number of) splice variants.

Using an assembled consensus sequence rather than a representative sequence also improves the sequence quality, owing to the high coverage. Even when only two sequences cover a position, the sequence quality is significantly improved, because one can avoid choosing probes from positions where the two sequences disagree. Restricting to those positions where the two sequences agree results in a much higher quality, due to the generally random behavior of sequencing errors. Cases where there are more than two sequences which cover a certain position have an even higher sequence quality.

Under-clustering may cause redundancy and over-clustering may cause genes to be missing from the representative set. Therefore, it is essential to have clustering which is as accurate as possible. From our experience, assembling the sequences in a cluster is an excellent way to find problems in the clustering, such as chimeric sequences or the clustering together of paralogous genes. Mapping the sequences to the genomic sequences also helps significantly in finding and virtually eliminating such problems.

Another benefit of having a more comprehensive transcript database is ensuring the specificity of probes. In order to ensure that a probe is unique, it is very important to compare it with all the possible subsequences from all the different splice variants and with the longest possible gene sequences.

4. RESULTS

We report results obtained on a run which was restricted to all the human genes for which an mRNA sequence is known. We have chosen to represent each gene with a single probe, selected to represent as many splice variants as possible (as explained in section 3.1.1). The designed oligonucleotides were synthesized and aliquotted in 96- and 384-well plates. Each well contains 10–20μl of a 60–65-mer at the concentration of 50μM, modified with a 5'-C6 amino group to enable covalent attachment to different substrates.[17,18]

4.1. General Statistics

General clustering statistics

We have started with 53644 mRNA sequences and 2513461 ESTs (all the human mRNAs and ESTs from GenBank release 120). By mRNAs, we mean all human sequences documented as "RNA" in the primate section of GenBank. 13951 of them contain the complete coding sequence. 20308 of the mRNAs were discarded because they consisted mostly of repeats or vectors or belonged to the highly variant and abundant family of immunoglobulins or T cell receptors or they were too short or with a bad sequence quality. The remaining 33336 mRNAs (12065 of the complete coding sequence mRNAs remained) were distributed to 17432 clusters. 7174 of these clusters contained complete coding sequence mRNAs. 887099 of the ESTs belonged to these clusters. Clustering allows one to overcome the redundancy which is inherent in the random sampling process of EST library construction. Highly expressed or extensively researched genes usually have many ESTs and, in some cases, several mRNA sequences (corresponding in some cases to several splice variants or several alleles). In fact, only 2582 mRNAs are singletons, whereas the average number of sequences in a cluster with an mRNA is 52.8. The average number of mRNAs in such clusters is 1.9. Figure 2 shows the distribution of cluster sizes for clusters with mRNA. clusters of a specific size for each cluster size.

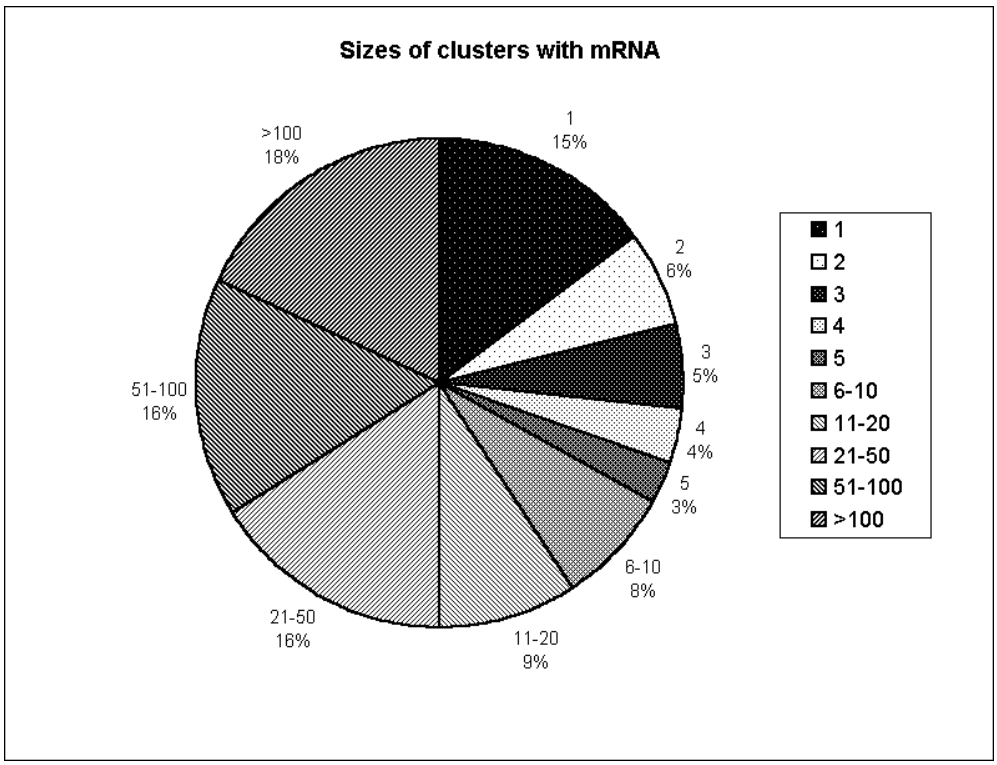

Figure 2. This chart shows the distribution of cluster sizes for clusters with mRNA. There were 17432 such clusters. 15% of them were singletons, 50% contained more than 50 sequences.

Alternative splicing

Alternative splicing is a very common phenomenon (it is claimed that it occurs in at least 35% of the human genes[10]). The higher the number of sequences in a cluster, the more likely it becomes to detect the different splice variants, if they exist. Of the 17432 clusters with an mRNA sequence, alternative splicing was detected in 9453 clusters, that is in 54% of the clusters. 2868 of them have two transcripts, the rest 6585 have three or more transcripts. In 73.6% (5282 out of 7174) of the clusters with an mRNA with complete coding sequence alternative splicing was detected. Figure 3 shows the disribution of the number of different transcripts in clusters with mRNA and detected alternative splicing.

Sequence coverage

Because of the high redundancy in the sequence databases, in most cases there are multiple sequences covering most of the gene sequence positions. This can be used in order to detect and correct sequencing errors and avoid places with suspected SNPs. In our dataset, only 33.1% of the positions are covered by only one sequence. 14.1% are covered by two sequences and the rest 52.8% are covered by three or more sequences.

Sequence elongation

Another benefit of the assembly process is that one gets elongated gene sequences. In many cases, none of the sequences in the public databases covers the full gene sequence, so the assembled consensus sequence is longer than the longest sequence in the database from that gene. To measure the effect of this, we calculated the difference between the assembled consensus length and the length of the longest database sequence for each gene. The average elongation is 524bp and in 3001 genes (17.2%), the elongation was higher than 1000bp.

4.2. Comparison with UniGene

It has become customary to use UniGene[7–9] as a clustering tool for designing oligonucleotide microarrays. In many cases designs are made by choosing a probe from the suggested UniGene representative for each UniGene cluster.

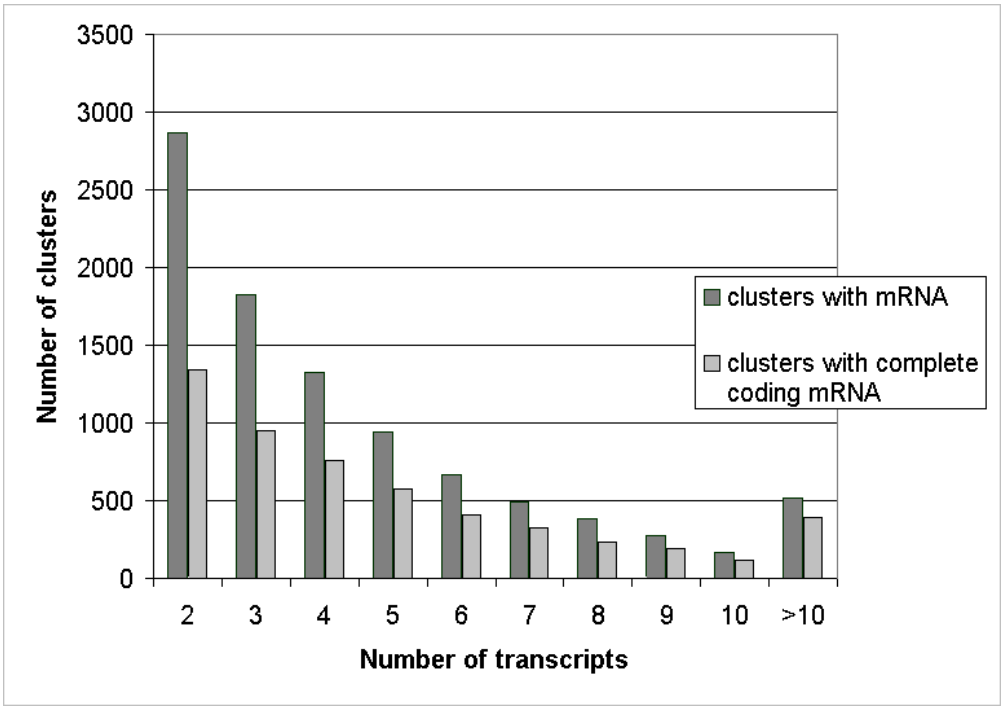

Figure 3. This chart shows the number of clusters with mRNA and 2 or more predicted transcripts. There were 17432 clusters with mRNA, 7979 of which with 1 transcript, and all the rest 9453 (54%) with 2 or more. There were 7174 clusters with complete coding sequence mRNA 1892 of which with 1 transcript and all the rest 5282 (73.6%) with 2 or more.

We therefore include a comparison between the clustering results of LEADS (based on GenBank version 120) and those of UniGene (based on build 127) and their relevance to the quality of the chosen probes.

Clustering comparison

Since the complete structure and sequence of all human genes is unknown, we compare the clustering characteristics to see which of the methods tends to cluster more sequences together. Since the collections of input sequences in UniGene and in our LEADS run were not identical and in order to filter out differences which are caused by this, we chose to limit our comparisons to the set of common sequences, so that if clusters in UniGene and in LEADS differ only by sequences which are unique to one of the sets, this difference is ignored. In general the LEADS clustering used more sequences than the UniGene clustering. Out of the 920435 sequences that ended up in the LEADS clusters that contain mRNAs, 97921 sequences did not appear in UniGene. Out of the 17432 LEADS clusters with mRNA, 2339 contain only sequences that do not appear in Unigene, and 10200 are identical to UniGene clusters. There are 3277 LEADS clusters that clustered together 6937 UniGene clusters, such that each LEADS cluster corresponds to several UniGene clusters. There are 1498 LEADS clusters that were clustered to 917 UniGene clusters, such that each UniGene cluster corresponds to several LEADS clusters. In 848 LEADS clusters that corresponded to 643 UniGene clusters, the correspondence between the two clustering methods was more complicated.

To further inquire the clustering differences, we checked the clustering differences as they appear on the set of UniGene representatives. In our set of 17432 LEADS clusters with mRNA, 16542 UniGene representatives appeared. Only in 10788 cases (65.2% of the cases) a single representative appeared in a LEADS cluster. In all other cases where a UniGene representative appeared in a LEADS cluster, it appeared with other representatives. In particular, there were 1972 LEADS clusters that contained 2 different UniGene representatives, and 520 LEADS clusters that contained 3 or more different UniGene representatives.

We conclude that LEADS tends to cluster more than UniGene. One explanation could be the relatively high number of sequences that do not appear in UniGene. If indeed there is a tendency for under clustering in UniGene, this might have a bad effect on probe designs since in many cases probes for the same genes will be chosen more than

once without knowing that causing unnecessary redundancy in the design. Even more serious than that is the effect on the cross homology score for each probe. Separating the same gene into different clusters might make one think that an area that is unique to this gene is not such and therefore have less choices for probes, ending with a choice of probes with lower quality. However, there is also a small but significant number of cases where LEADS splits UniGene clusters, and we believe that this happens in cases where chimeric sequences were identified by LEADS, and in cases where using the genome as a reference suggested a split.

Alternative splicing

In our general analysis, we observed that 54% of the clusters in our dataset showed evidence of alternative splicing. When choosing probes from a representative sequence ignoring the other sequences in the cluster, one runs the risk of choosing a probe from a segment of the gene sequence which is not common to all the transcripts. This may result in missing the expression of the gene if other transcripts are expressed. In some cases, the detected alternative splicing occurs outside the region covered by the UniGene representative sequence (it may occur, for instance, in the 5' UTR or in the 3' UTR). In order to estimate the possible effect of alternative splicing on the probe selected from a UniGene representative, we have checked how many of UniGene represetatives in the above 17432 genes have alternative splicing in the region covered by a UniGene representative. We checked all the UniGene representatives that appeared in the group of LEADS clusters that contain an mRNA. There were 16542 UniGene representatives in these clusters, sometimes more than one UniGene representative in one LEADS cluster. It turns that in 11556 of the UniGene representatives checked (69.8% of the total number) alternative splicing occured within the region covered by the representative. We conclude that choosing probes from the UniGene representative might result in many cases of probes that measure only some of the splice variants of a gene and not all of them.

Sequence quality

In order to measure the effect of high coverage on sequence quality, we estimated the sequence error rate in UniGene representatives by looking at positions with a coverage of three sequences or more. We have found that the error rate is 0.43%, which translates to a probability of 23% for an error somewhere inside a 60-mer. Note that we checked only representatives for clusters that contain mRNAs, and therefore most of the representatives are mRNAs with a better sequence quality than a usual EST. In clusters without mRNAs the error rate for UniGene representatives will be higher, but this is beyond the scope of this comparison.

4.3. The Chosen Probes

For each LEADS cluster with an mRNA, we have chosen a 60-mer according to the following criteria:

- Coverage of a maximal number of transcripts.
- Maximal cross-homology of 70%.
- A continuous hit of no more than 17bp to the sequence of another gene.
- Distance of no more than 1500bp and no less than 50bp from the 3' end.
- GC content in the range of 30%-70%, without significant windows of local imbalance.
- No more than 2 palindromes of length 6.
- A hit of no more than 15bp to a repeat, vector or low-complexity region.
- No long stretches of identical nucleotides.

The probe was chosen among the probes satisfying these criteria according to a score which balances the relative importance of the different probe characteristics. The above criteria show the acceptable ranges for each of the parameters. By choosing a probe of minimal score, we usually get probes for which the values of the parameters are much closer to the optimal values.

To calculate the cross-homology we calculate for each probe its best alignment to a different gene (we use here the overall prediction of the transcriptome by LEADS). We count the number of matches in the best alignment to a different gene, add 25% of the bases that do not appear in the alignment, since this is the expected number of

nucleotides that will match. Thus, in our scheme, the lowest level of cross hybridization is 25%, which means no homology was found to any other gene.

Unfortunately, it is not always possible to choose a probe that satisfies all these requirements. For example, in a significant number of cases, it is impossible to choose a probe without any risk of cross-hybridization, because the gene is highly paralogous over its full sequence to another gene. Of the 17432 chosen probes, 15730 satisfy all the above criteria. In the other cases, we have done our best to choose the optimal probe given the limitations, 98.1% of the probes have a GC content in the range 0.35-0.65. 94.1% of the probes have a cross-homology to a different LEADS cluster figure smaller than 70%. 90.1% have a cross-homology smaller than 50%. 98.2% of the probes have a distance smaller than 1500bp to the 3' end. 98.6% of the probes cover more than 50% of the transcripts, and 81.3% of the probes cover all the transcripts. 76.4% of the probes were selected from a region covered by more than 2 sequences.

5. SUMMARY

We described the different parts of the LEADS transcriptome prediction system, that clusters and assembles expressed and genomic data, and its value to DNA oligonucleotide chip design. We discussed the different parameters needed to be taken into account when designing DNA chips, and in particular we discussed the different approaches to take alternative splicing into account. We described a specific design for all the genes that contain at least one mRNA, and a comparison of the LEADS clusters we used to UniGene clusters. We also discussed the effects of choosing probes from a UniGene representative without using an assembly mechanism, and without analayzing the effects of alternative splicing.

Our conclusion is that using LEADS as a basis for DNA chip designs can significantly improve designs based on UniGene. The improvenment rises due to more updated and accurate clustering, running an assembly process for each cluster, using the genome whenever possible, taking alternative splicing into account, and improving sequence quality.

Sequences that are missing in the UniGene clustering can cause under clustering, cause redundant designs and unprecise cross homology searches. Chimeric sequences that can be detected by analysis of an assembly of a cluster, are hard to find in UniGene, and might cause chimeric clusters. Using the genome as a reference in LEADS, in cases where it is possible, significatnly improves the clustering and assembly quality.

We showed that LEADS predicted 54% alternative splicing for clusters with mRNA, and 73.6% of alternative splicing in clusters with complete coding mRNA sequence. Therefore, especially for this interseting set of genes, alternative splicing is a very abundant phenomenon that must be taken into account. We also showed that for this set of clusters, alternative splicing occures within 69.8% of the UniGene representatives. Therefore, choosing a probe from these representatives without analyzing the effects of alternative splicing might end up with choosing probes that cover only a few of the splice variants, and not all or most of them. This means that many transcripts that are expressed in the cell might not be detected and measured.

In general we believe that designing oligo sets for DNA chips should be done carefully using all available data, and taking into account all the possible effects, and in particular alternative splicing. Better designs may lead to better and more reliable results from DNA chips.

ACKNOWLEDGMENTS

We would like to thank our colleagues in Compugen, that developed different parts of the LEADS platform. In particular we wish to thank Avi Rosenberg, Hershel Safer, Guy Kol, Ariel Farkash, Alex Golubev, Galit Fuhrmann, Gil Dogon, Iftah Nachman , Raveh Gill-More, Eran Halperin, Sarah Pollock, Mor Amitai, Amit Gal and Dror Efrati. We wish to thank Simchon Faigler, Liat Mintz and Paul Nisson for their helpfull remarks in the process of developing the design system. We would also like to thank Andrew Olson for his help in running the LEADS on our datasets.

REFERENCES

1. D. J. Lockhart and E. A. Winzeler, "Genomics, gene expression and DNA arrays," *Nature* **405**, pp. 827–836, 2000.

2. E. Southern, S. C. Case-Green, J. K. Elder, M. Johnson, K. U. Mir, L. Wang, and J. C. Williams, "Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids," *Nucleic Acids Res.* **22**, pp. 1368–1373, 1994.
3. S. P. Fodor, R. P. Rava, X. C. Huang, A. C. Pease, C. P. Holmes, and C. L. Adams, "Multiplexed biochemical assays with biological chips," *Nature* **364**, pp. 555–556, 1993.
4. D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nat. Biotechnol.* **14**, pp. 1675–1680, 1996.
5. M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science* **270**, pp. 467–470, 1995.
6. J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* **278**, pp. 680–686, 1997.
7. M. S. Boguski and G. D. Schuler, "ESTablishing a human transcript map," *Nat. Genet.* **10**, pp. 369–371, 1995.
8. G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, and P. Rodriguez-Tomé, "A gene map of the human genome," *Science* **274**, pp. 540–546, 1996.
9. G. D. Schuler, "Pieces of the puzzle: expressed sequence tags and the catalog of human genes," *J. Mol. Med.* **75**, pp. 694–698, 1997.
10. A. A. Mironov, J. W. Fickett, and M. S. Gelfand, "Frequent alternative splicing of human genes," *Genome Res.* **9**, pp. 1288–1293, 1999.
11. G. J. Kilpatrick, F. M. Dautzenberg, G. R. Martin, and R. M. Eglen, "7TM receptors: The splicing on the cake," *Trends Pharmacol. Sci.* **20**, pp. 294–301, 1999.
12. A. J. Lopez, "Alternative splicing of pre-mRNA: Developmental consequences and mechanisms of regulation," *Annu. Rev. Genetic.* **32**, pp. 279–305, 1998.
13. D. Gautheret, O. Poirot, F. Lopez, S. Audic, and J. M. Claverie, "Alternative polyadenylation in human mRNAs: A large-scale analysis by EST clustering," *Genome Res.* **8**, pp. 524–530, 1998.
14. G. Edwalds-Gilbert, K. L. Veraldi, and C. Milcarek, "Alternative poly(A) site selection in complex transcription units: mean to an end?," *Nucleic Acids Res.* **25**, pp. 2547–2561, 1997.
15. J. S. Aaronson, B. Eckman, R. A. Blevins, J. A. Borkowski, J. Myerson, S. Imran, and K. O. Elliston, "Toward the development of a gene index to the human genome: An assesment of the nature of high-throughput EST sequence data," *Genome Res.* **6**, pp. 829–845, 1996.
16. "`http://genome.ucsc.edu/`." Oct. 7 2000 data set.
17. M. Yang, H. L. Chan, W. Lam, and W. F. Fong, "Covalent immobilization of oligonucleotides on modified glass/silicon surfaces for solid-phase DNA hybridization and amplification," *Chemistry Letters* **3**, pp. 257–258, 1998.
18. M. Beier and J. D. Hoheisel, "Versatile derivatisation of solid support media for covalent bonding on DNA-microchips," *Nucleic Acids Res.* **27**, pp. 1970–1977, 1999.