Genome analysis

Gene sequence signatures revealed by mining the UniGene affiliation network

Jiexin Zhang, Li Zhang and Kevin R. Coombes*

Department of Biostatistics and Applied Mathematics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Box 447, Houston, TX 77030-4009, USA

Received on April 29, 2005; revised on November 18, 2005; accepted on November 19, 2005
Advance Access publication December 8, 2005
Associate Editor: Martin Bishop

ABSTRACT
Background: In the post-genomic era, developing tools to decode biological information from genomic sequences is important. Inspired by affiliation network theory, we investigated gene sequences of two kinds of UniGene clusters (UCs); narrowly expressed transcripts (NETs), whose expression is confined to a few tissues; and prevalently expressed transcripts (PETs) that are expressed in many tissues.

Results: We explored the human and the mouse UniGene databases to compare NETs and PETs from different perspectives. We found that NETs were associated with smaller cluster size, shorter sequence length, a lower likelihood of having LocusLink annotations, and lower and more sporadic levels of expression. Significantly, the dinucleotide frequencies of NETs are similar to those of intergenic sequences in the genome, and they differ from those of PETs. We used these differences in dinucleotide frequencies to develop a discriminant analysis model to distinguish PETs from intergenic sequences.

Conclusions: Our results show that most NETs resemble intergenic sequences, casting doubts on the quality of such UniGene clusters. However, we also noted that a fraction of NETs resemble PETs in terms of dinucleotide frequencies and other features. Such NETs may have fewer quality problems. This work may be helpful in the studies of non-coding RNAs and in the validation of gene sequence databases.

Availability: http://bioinformatics.mdanderson.org/SequenceQualityCheck/
Contact: kcoombes@mdanderson.org
Supplementary information: http://bioinformatics.mdanderson.org/Supplements/AffiliationNetwork/SupplementaryMaterial.pdf

INTRODUCTION

UniGene is a database of gene sequences widely used in biological research (Pontius et al., 2003). Its content is derived from GenBank, a large collection of cDNA and Expressed Sequence Tags (ESTs) representing the results of decades of worldwide effort. UniGene was created to circumvent the redundancy in GenBank and to weed out contamination. Records in UniGene were generated by partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. To ensure the quality of the UniGene database, care was taken to remove contamination resulting from untrimmed vectors, linkers, ribosomal, mitochondrial, low-complexity sequences, repeats and other external contaminants. Alignments between transcript sequences and genomic sequences were also used to verify the genomic origin of the clusters.

However, because of the enormous size of the database and the complex biological information it contains, not all entries in the database are equally reliable. The current human UniGene database (build 181) contains 5 080 380 GenBank sequences representing 52 924 UniGene clusters (UCs). This number of UCs far exceeds the estimated number of protein-coding genes in the human genome (Fields et al., 1994; Antiquera and Bird, 1993; Lander et al., 2001; Das et al., 2001; Venter et al., 2001). The latest estimate finds 20 000 to 25 000 protein-coding genes in the human genome (IHGSC, 2004). There are multiple potential causes of the excess. First, some UCs may represent unmerged short sequences. However, UniGene takes several steps to minimize the number of such UCs: (1) UniGene clusters must be anchored at the 3' end of a transcription unit; (2) non-overlapping 5' and 3' ESTs are joined into the same cluster using evidence from clone-based studies; (3) singleton clusters (clusters with one EST or sequence) and non-anchored sequences are compared with all anchored clusters at reduced stringency to decrease the number of singleton clusters and non-anchored sequences and (4) a more stringent test of 3' anchoring has been applied with the availability of genome sequence (Pontius et al., 2003; Yuan et al., 2001).

Second, the excessive UCs may come from non-coding RNAs, which were not included in the estimates of functional genes. Some non-coding RNAs may have biological functions, although only a small number of them have been characterized so far. Some non-coding RNAs may result from ‘leakiness’ in the transcriptional machinery in cells (Cases and de Lorenzo, 2001) and can be considered part of the background noise of transcription. It is possible that most of these RNAs have no biological function at all. But not least, some UCs might result from errors, such as incorrect merging of ESTs, contamination of pre-mRNAs or foreign sources, or simple sequencing errors. Therefore, the UniGene database includes many classes of RNA sequences of varying quality. It would be desirable to develop methods to recognize these classes and to investigate them further.

In this paper, we have examined the properties of UCs via an affiliation network. Affiliation networks contain two kinds of nodes, with a restriction that each edge must connect different kinds of nodes. Affiliation networks have been applied in various fields to elucidate global properties that might not be obvious.
from individual elements (Watts, 2003; Tsonis and Tsonis, 2004; Ding et al., 2004). We believed that understanding the gene affiliation network would provide useful biological insights.

For this study, we constructed an affiliation network between UCs and the tissues in which they were expressed. The tissues for a UC include all tissue libraries from which the ESTs in the cluster were derived. We grouped UCs according to the number of tissues in which they were expressed. Depending on the prevalence of expression, UCs can be put into two general categories: narrowly expressed transcripts (NETs) if expression is confined to a small number of tissues, and prevalently expressed transcripts (PETs) otherwise.

Our aim is to identify the properties of UCs that are associated with their expression prevalence. In general, NETs are more likely to represent tissue-specific genes with special functions, while PETs are more likely to represent genes that perform common functions needed for the normal operation of many cell types. We investigated several properties of UCs, including their tissue distribution, cluster size, sequence length, genome location information, dinucleotide frequencies and expression levels extracted from a wide range of biological conditions.

To investigate the relationship between PETs and NETs, we also collected a set of intergenic sequences by randomly excising genomic sequences. We compared the dinucleotide frequencies of PETs, NETs and intergenic sequences. We then trained a quadratic discriminant analysis (QDA) model (Krzanowski, 1988) to distinguish PETs from intergenic sequences. Support vector machines (SVM) were used as an alternative classification scheme to corroborate the QDA results. We further examined the subclasses of PETs and NETs as classified by the QDA model.

We performed the same analysis on both the human and the mouse UniGene databases. In this article, we shall focus on the human UniGene database since the results from the mouse database are similar to those from the human database.

MATERIALS AND METHODS
Source of UniGene data
The annotations and representative sequences for UCs from the human UniGene build 181 and from the mouse UniGene build 145 were downloaded from http://www.ncbi.nlm.nih.gov/UniGene. The human version contained 5 080 380 sequences, of which 96.1% are ESTs. These sequences corresponded to 52 924 human UCs, of which 22 795 contained LocusLink annotations. The mouse database contained 3 760 414 sequences corresponding to 45 717 UCs, of which 25 366 contained LocusLink annotations. The dataset of representative sequences contained one sequence for each UC. This unique sequence was selected from the cluster because it contained the longest region of high-quality sequence data (Pontius et al., 2003).

Classification of UniGene clusters by their degrees in the affiliation network
In the UniGene database, the ‘EXPRESS’ information for a UC, if available, records the tissue name(s) from which the sequences in that UC were derived. In total, the human UniGene database contains 279 different tissues; the mouse database, 90. We constructed an affiliation network by drawing an edge between a UC and a tissue whenever the UC is expressed in that tissue. It should be noted that ‘tissue’, as used here, is defined so that ‘spleen’ and ‘spleen tumor’ represent two different tissues, even though both come from spleen. A UC contains no tissue expression information only when no tissue library is known for any corresponding sequences. We excluded from our study ~4.0% of UCs in the human database (3.4% in mouse) that contain no tissue expression information. The final numbers of UCs included in our study were 50 800 for human and 44 177 for mouse.

The degree of a UC is defined as the number of tissues to which the UC is connected in the affiliation network. To investigate the differences between NETs and PETs, we collected UCs according to their degrees and partitioned the UCs into four groups by manually setting breakpoints on the distribution curves (Fig. 1). G1 contains UCs expressing in one tissue. The separation points for groups G2–G4 are 20 and 33 for human, 15 and 22 for mouse. Thus, human group G2 contains UCs expressing in 2–20 tissues; human G3 contains UCs expressing in 21–33 tissues and human G4 contains UCs expressing in more than 33 tissues.

Construction of training set for prediction by QDA and SVM
We constructed a training set containing two groups of sequences: intergenic sequences and PET gene sequences. Intergenic sequences were randomly excised from all 24 chromosomes in the human genome build 35.1, with lengths ranging from 250 to 4000 bp. Repetitive regions were removed before sequence extraction. Because only ~5% of the genome corresponds to functional products, most of the sequence segments excised from the genome came from intergenic regions. The gene sequences in the training set were excised from PETs. More precisely, we first randomly selected representative sequences from groups G3 and G4 and concatenated them into a single long sequence. Next, we randomly excised sequence segments from the concatenated sequence, with lengths ranging from 250 to 4000 bp. Care was taken to avoid overlaps in the excised sequences. The complete training set included 12 000 intergenic fragments and 12 000 PET fragments. The distributions of sequence lengths in the two groups were the same.

Quadratic discriminant analysis
Discriminant analysis is a technique for classifying a set of observations into predefined classes based on a set of predictor variables (Krzanowski, 1988),
We computed the dinucleotide frequencies for each sequence in the training set. Repetitive regions and linker segments were removed before dinucleotide frequency calculation. Nine features, representing the dinucleotide frequencies with large variance (AA, AT, CC, CG, GA, GC, GG, TA and TT as shown in Fig. 2), were employed as predictors in the QDA classifier. QDA was computed using the ‘classify’ method in MATLAB 7.1 (The MathWorks, Natick, MA). To evaluate the performance of the QDA method, we used non-redundant 5-fold cross-validation. For each round of validation, one-fifth of the training set was used for prediction and the remaining four-fifths was used for testing. The procedure was repeated five times. The five sets of dinucleotide frequencies used for training were mutually exclusive. The average posterior probability, from the five predictions, of a UC being intergenic-sequence-like is reported. A Perl script was written to calculate dinucleotide frequency of DNA sequences and to make predictions using the QDA parameters generated in MATLAB; these predictions can be accessed at http://bioinformatics.mdanderson.org/SequenceQualityCheck/

**Source of microarray data**

Microarray data were obtained from 571 microarray experiments using the Affymetrix human genome HG-U133A GeneChip®. The data were collected in the microarray core facility of The University of Texas M.D. Anderson Cancer Center from cancer research projects of many different laboratories. In order to reduce the effect of variations in the probe binding affinities, expression values were computed using the position-dependent nearest-neighbor (PDNN) model (Zhang et al., 2003).

Mapping information between probe sets on the HG-133A chip and UniGene clusters was downloaded from the Affymetrix web site (http://www.affymetrix.com/support/technical/libraryfilesmain.affx). Of the 22,283 probe sets on the HG-U133A chip, 19,751 were annotated into the human UniGene build 181, corresponding to 12,753 distinct UCs. Groups G1–G4 contain 253, 5124, 7050 and 7324 probe sets, respectively.

**RESULTS**

We constructed an affiliation network between UCs and tissues as described in the Materials and methods section. Important properties of a network can be derived from the degree distribution (Newman et al., 2002; Strogatz, 2001). The degree distribution of UCs, which represents the expression prevalence of the UCs, differs from a simple power law distribution (Fig. 1). Such a distribution suggests that there may be complex biological factors affecting the expression prevalence of the UCs.

To characterize the UCs, we partitioned them into four main groups (Fig. 1). Group G1 contains the most nearly expressed transcripts, because each UC in this group is expressed in a single tissue. The number of UCs in G1 is 15,945, which accounts for >30% of all UCs in the dataset. G1 contains 8631 UCs that are singleton clusters containing only one EST or sequence. UCs in G2 are expressed in 2–20 tissues; UCs in G3 are expressed in 21–33 tissues; UCs in G4 are expressed in at least 34 tissues. (In mouse, UCs in G2 are expressed in 2–15 tissues; UCs in G3, in 16–22; UCs in G4, in 23 or more.) We viewed UCs in G1 as NETs and UCs in G3 and G4 as PETs. We considered UCs in G2 to be a mixture of NETs and UCs. The median number of ESTs associated with each UC in human group G4 was 432, with an inter-quartile range (IQR) equal to 351. In G3, the median was 163 with IQR = 113; in G2, median = 8 and IQR = 22; in G1, median = 1 and IQR = 1.

Most tissues are expected to express tissue-specific genes that are required to carry out tissue-specific functions. Thus, we expected G1 UCs to be widely distributed across most tissues. Surprisingly, however, 208 out of 279 tissues do not contain any G1 UCs. In fact, the top ten most UniGene-rich tissues (i.e. tissues with the largest number of associated UCs, which include brain, lung, testis, kidney, eye, uterus, placenta and colon along with the less informative...
many UCs in groups G1 and G2 are predicted to be intergenic-like. In
the training set. We found that most UCs in G3 and G4
resemble genomic sequences more closely than UCs in other groups
(Supplementary Figure S1 and Supplementary Table S1).

Since the dinucleotide frequencies of NETs are distinctively dif-
ferent from those in PETs, we hypothesized that we could use
dinucleotide frequencies for class prediction. We randomly sampled
dinucleotide frequencies of segments excised from genomic
sequences and from PETs to train a QDA model to discriminate
the two classes. [We used SVM as an alternative classification
scheme to corroborate the QDA results. QDA and SVM produced consistent predictions in most cases (Supplementary Table S2).] We
repeated the QDA classification five times using mutually exclusive
training sets for cross-validation. The QDA classifier reported the
average posterior probability that a UC is intergenic-sequence-like.
The prediction for a sequence in G3 or G4 is based on the four QDA
models in which the sequence was used in the test set but not in the
training set. The predictions for a sequence in G1 or G2 are based on
all five QDA models, since genes in G1 and G2 were never included in
the training set. We found that most UCs in G3 and G4
were predicted to be gene sequences; i.e. they have a low probability
of being intergenic sequences (Fig. 3). By contrast, many UCs in
groups G1 and G2 were predicted to be intergenic-like. In
human UniGene, 67% of G1 and 40% of G2 sequences had >0.5
probability of being intergenic-like. Further, 49% of G1 and 24%
of G2 sequences had >0.8 probability of being intergenic-like.
About 35% of these intergenic-like sequences in G1 and G2
were singleton UCs.

We suspected that our QDA classifier might have implicitly used
some sequence properties of translated regions of protein-coding
sequences, especially those of the housekeeping genes. To test this
hypothesis, we examined if our QDA classifier was good at recog-
nizing functional non-coding RNA sequences. We downloaded
sequences from a non-coding RNA database (http://biobases.
ibch.poznan.pl/ncRNA/) that contained 49 human functional non-
coding RNAs. Our QDA classifier identified 37 out of 49 non-
coding RNAs as PET-like with a posterior probability larger
than 0.5 (Supplementary Table S3). By contrast, less than 20%
of randomly selected genomic sequences are predicted as
PET-like with a posterior probability larger than 0.5. This result
suggested that it was unlikely that our QDA classifier was limited to
picking up protein-coding genes.

In addition to the dinucleotide frequencies, we examined other
properties of UCs that might be associated with their expression
prevalence. Figure 4 shows the relationship between the expression
prevalence of a UC and the length of its representative sequence. In
G4, the shorter the sequence, the more prevalently the UC is
expressed. This finding is consistent with the idea that housekeeping
genes are shorter than other genes (Eisenberg and Levanon, 2003).
For UCs in G1 and G2, the trend goes the opposite way: the shorter
the sequence, the more rarely the UC is expressed. Most UCs in G1
are extremely short, with an average length of 703 nt. One possi-
bility is that G1 contains clusters with non-overlapping 3′ and 3′
ESTs. The representative sequences for UCs in G1 group might
become longer when more ESTs are sequenced.

![Density Estimation](image)

**Fig. 3.** Density estimation of the posterior probability of being intergenic
sequence predicted by QDA using dinucleotide frequencies as predictors.
QDA was repeated five times with mutually exclusive training sets for cross-
validation. The averages of posterior probabilities from five predictions were
reported here. (a) Human UniGene prediction results. (b) mouse UniGene
prediction results.
We explored the relationship between expression levels and expression prevalence. Microarray expression data used for this study are described in the Materials and methods section. Affymetrix probe sets are mapped to UniGene clusters via their sequence accession numbers. Of the 22,283 probe sets on the HG-U133A chip, 19,751 were annotated into human UniGene build 181, corresponding to 12,753 distinct UCs. Groups G1–G4 contain 253, 5124, 7050 and 7324 probe sets, respectively. We plotted the distributions of expression levels for probe sets in each group in a density plot (Fig. 5). We found that expression levels (whether measured by the mode, median or mean) increase as the prevalence of the UCs increases. It should be noted that the definition of expression prevalence does not take the expression level into account. Tissue-specific genes might be expressed at a high level in a particular tissue because it is central to that tissue’s functionality. If this were the case, we should have seen a population of NETs with high expression levels. However as shown in Figure 5, probe sets corresponding to UCs in G1 and G2 express at low level in general.

Another property of UCs that we examined is whether a UC possesses LocusLink annotation. The non-existence of such an annotation implies that it is difficult to map the UC to the genome (Pruitt et al., 2000). Such difficulties are often caused by chimeric sequences resulting from artifacts of cDNA cloning and other contaminated sequences. Hence, a UC sequence without a LocusLink is more likely to represent an error. As expected, much smaller percentages of UCs in G1 and G2 have LocusLink annotations than in G3 and G4 (Fig. 6).

We believed that PETs are more reliable than NETs in general because PETs are easily found in many tissues and hence are often extensively studied. It is not surprising that PETs are more likely to have LocusLink annotations than NETs. Interestingly, we found that 32.7% of G1 are classified as PETs (with probability >0.5) by our QDA classifier (Fig. 3). These PET-like NETs also have higher probability of possessing LocusLink annotations (Table 1). The P-values (<10^-16 except group G4) calculated by χ²-tests for each group suggest that the association is statistically significant. It is important to note that this association is significant even for the UCs in groups G1 and G2. These PET-like NETs may have better quality than other NETs.

**DISCUSSION**

In this study, we grouped UCs into NETs and PETs according to the degree of each UC in the gene affiliation network and examined their properties from many perspectives. We found that NETs were associated with smaller cluster size, shorter sequence length, a lower likelihood of having LocusLink annotations, and lower and more sporadic levels of expression. Most importantly, we found that the dinucleotide frequencies of NETs are similar to those of intergenic sequences, and they differ from those of PETs.
Table 1. Contingency table for the association between QDA predictions and the LocusLink annotations

<table>
<thead>
<tr>
<th>Human UC groups</th>
<th>Classification result</th>
<th>LocusLink number</th>
<th>$\chi^2$-test</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1 Intergenic seq</td>
<td>10410</td>
<td>324</td>
<td>(&lt;)E−16</td>
<td></td>
</tr>
<tr>
<td>G1 Gene seq</td>
<td>4329</td>
<td>882</td>
<td>(&lt;)E−16</td>
<td></td>
</tr>
<tr>
<td>G2 Intergenic seq</td>
<td>7696</td>
<td>1774</td>
<td>(&lt;)E−16</td>
<td></td>
</tr>
<tr>
<td>G2 Gene seq</td>
<td>5712</td>
<td>8380</td>
<td>(&lt;)E−16</td>
<td></td>
</tr>
<tr>
<td>G3 Intergenic seq</td>
<td>143</td>
<td>1093</td>
<td>(&lt;)E−16</td>
<td></td>
</tr>
<tr>
<td>G3 Gene seq</td>
<td>166</td>
<td>5397</td>
<td>(&lt;)E−16</td>
<td></td>
</tr>
<tr>
<td>G4 Intergenic seq</td>
<td>33</td>
<td>3704</td>
<td>2.87E−02</td>
<td></td>
</tr>
<tr>
<td>G4 Gene seq</td>
<td>14</td>
<td>743</td>
<td>(&lt;)E−16</td>
<td></td>
</tr>
<tr>
<td>Total Gene seq</td>
<td>18263</td>
<td>3934</td>
<td>(&lt;)E−16</td>
<td></td>
</tr>
</tbody>
</table>

A UC was classified as intergenic sequence when the average posterior probability, from 5-fold predictions, of it being intergenic sequence is larger than 0.5; and a UC was classified as gene sequence otherwise.

We found that 67.3% of UCs in G1 and 40.2% of UCs in G2 resemble intergenic sequences (with probability >0.5). Because the false positive rate in G3 and G4 is ~17%, it is likely that some of the intergenic-like sequences in G1 and G2 represent misclassified genes. Nevertheless, there are still many UCs in G1 and G2 with poor quality. It has previously been noted that there are multiple sources of potential contamination in the UniGene database. Because most UCs were based upon assemblies of ESTs, which were ‘single pass’ cDNA sequences with error rates as high as 3%, various kinds of contamination could occur (Schuler, 1997; Hillier et al., 1996). Sorek and Safer (2003) found that some EST libraries may be particularly prone to be contaminated by human genomic DNA, pre-mRNA or non-canonical introns. The contaminated libraries were characterized by an unusually high percentage of unsliced singleton ESTs or ESTs overlapping with introns. Our result is consistent with their study. We found that the top 10 most UniGene-rich tissues also contain the highest percentage of singleton clusters and 76% of singleton clusters are expressed in those top 10 tissues. Brain and lung are the most popular tissues in the human UniGene database. Sorek and Safer (2003) also found that the brain library was contaminated by non-canonical introns and the lung by pre-mRNAs. However, 65.2% of highly intergenic-sequence-like UCs (with probability >0.8) contain multiple ESTs, suggesting the existence of questionable UCs besides singleton clusters.

NETs contained in the UniGene databases, excluding those caused by sequence errors, may represent a class of largely unknown non-coding RNAs. There is now accumulating evidence that the number of transcribed RNAs in a cell is much larger than previously thought (Bertone et al., 2004; Kampa et al., 2004; Kapranov et al., 2002). Some non-coding RNAs seem to play regulatory roles (Cawley et al., 2004). Others may merely reflect reproducible transcriptional noise and have no specific biological functions. It is not yet clear what percentage of transcribed RNAs have biological functions (IHGSC, 2004). Our current method probably does not have any power to differentiate sequence errors from non-coding RNAs; further investigation is needed in this direction.

Previous studies have established that dinucleotide frequency pattern is an important feature of gene/genome sequences (Karlin and Burge, 1995; Karlin, 1998). For example, it has been noticed that most housekeeping genes have a CpG island in the 5’ promoter region (Larsen et al., 1992; Gardiner-Garden and Frommer, 1987). Thus, the housekeeping genes are associated with high GC content. Our dinucleotide frequency pattern (Fig. 2) shows that there is more at play than merely the GC content. For example, frequencies of GA and TC are similar in genomic sequences, as well as in groups G1 and G2, which is expected because GA and TC are complementary. But the frequencies diverge in groups G3 and G4. Such information would be lost if merely the GC content was used.

Our study uncovered a distinct difference between PETs and intergenic sequences in terms of dinucleotide frequencies. This finding may be of general use for quality control purposes in the development of gene sequence databases. There are many known patterns of protein-coding gene sequences that differ from intergenic sequences. Such patterns are the basis of computer models for gene discovery (Ashurst and Collins, 2003; Burge and Karlin, 1997; Zhang 1997). However, our QDA classifier is not limited to protein-coding sequences, as we have shown that it can also recognize functional non-coding RNAs.

It should be pointed out that we only expect a limited sensitivity and specificity from our QDA classifier for detecting UCs with quality problems. The method can often tell PETs from intergenic sequences, but some UCs may represent functional narrowly expressed transcripts. Thus, considering properties besides the dinucleotide frequencies, such as sequence length, LocusLink annotation and expression levels, may enhance the detection power of the method.

Conflict of Interest: none declared.

REFERENCES


390