Identifying DNA Strands Using a Kernel of Classified Sequences

Guillermo Tonsmann[a], David D. Pollock, Wanjun Gu and Todd A. Castoe[b]

Abstract— **Automated DNA sequencing produces a large amount of raw DNA sequence data that then needs to be classified, organized, and annotation. One major application is the comparison of new DNA sequences with previously known classified sequences. In this paper we present a new approach to perform these comparisons. From a kernel of previously classified DNA sequences, we identify distinctive oligomers, or short DNA sequences, that are infrequent and thus highly unique within the kernel. We then search for the presence of these distinctive oligomers in the new unclassified DNA sequences. Their presence indicates a possible relation between a new DNA sequence and every previously classified DNA sequence that shares the distinctive oligomer. Ultimately, unclassified sequences are related to classified sequences with which they share the highest number of distinctive oligomers. We explain the details of our technique and show some experimental results in a kernel of immunoglobulin DNA sequences.**

I. BACKGROUND

THE genetic instructions for the formation and functioning of all known living creatures are encoded in DNA molecules. DNA molecules are made of alternating sequences of 4 monomeric components: Adenine, Cytosine, Guanine and Thymine. These 4 monomers, known as nucleotides, are usually abbreviated by their respective initials: A, C, G, and T. In the typical double-stranded DNA polymer, Adenine molecules are complementary to Thymine molecules, and vice versa, because they bond with each other across the axis of the DNA double helix. This complementarity is also true of the other two nucleotides, Cytosine and Guanine. The well-known double helix configuration of a DNA molecule is produced when a DNA strand is paired side by side with another strand with the complementary nucleotides of the first. Because of this molecular complementarity, knowing the composition of one of the DNA strands is enough to describe both strands of a DNA molecule.

Various techniques are currently in use for DNA sequence determination[4]. Recent developments on these techniques aim to minimize sequencing time and cost, while maximizing sequenced DNA throughput. Among these techniques, our research used parallel pyrosequencing[5] implemented on the 454-FLX sequencing instrument from 454 Life Sciences[6]. These machines may yield millions of sequence reads, and about one billion total nucleotide sequences per day. To harness such sequencing power, new methods for classifying and comparing newly sequenced DNA to sequences previously sequenced and annotated is necessary.

Many DNA sequences are being found, studied, named, classified, and annotated every day. Various public databases constantly collect and organize this new information, like the database maintained by the National Center for Biotechnology Information (NCBI)[3]. When new DNA material is sequenced, it is usually compared against database records. For example, the NCBI web site provides access to BLAST, a well-known sequencing and alignment tool[2].This and similar tools provide a good starting point when studying DNA material of which very little else is known. However, when studying DNA sequences that are known to belong to a certain category, a more focused approach could provide much more rapid analytical results of comparisons than could BLAST algorithms. For example, our research is focused in the study on human immunoglobulins. These are proteins used by the immune system to detect and destroy foreign molecules that invade the body. Immunoglobulins may attach to the foreign molecules (be they viruses, foreign proteins, bacteria) and tag them for destruction by other factions of the immune system. By the attachment itself, they may also neutralize the molecule’s or pathogen’s dangerous capabilities. Like most eukaryotic genes, immunoglobulin genes are comprised of multiple exon regions that are eventually spliced together to form the immunoglobulin protein sequences A typical immunoglobulin DNA sequence is usually divided in regions, each one with a specific function in the lifecycle of the immunoglobulin. The region in charge of its adaptability to various types of attackers is encoded in three identifiable consecutive sections, or exons, V-exon, D-exon, and the J-exon. In our study, immunoglobulin DNA strands were matched against a database of previously identified V-exons and J-exons.

In this paper, we present a methodology to classify newly sequenced raw DNA sequences as related to previously classified DNA sequence or sequences. The set of classified DNA sequences will be known as the kernel. From a kernel we identify distinctive oligomers that characterize the classified sequences[1]. We define an oligomer of size *n* to be

[a]Dr. Guillermo Tonsmann, Associate Professor of Computer Science at Park University – Austin Campus; e-mail: tonsmann@park.edu.

[b] Drs. David D. Pollock, Wanjun Gu and Todd A. Castoe, currently at the Consortium of Comparative Genomics, Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine.

any DNA subsequence with exactly *n* nucleotides in a sequence. A single DNA sequence of size *m* (where *m* is bigger than or equal to *n*), may contain up to *m-n+1* different oligomers of size *n*. Oligomer repetition may reduce this number. A similar consideration is to be made in a kernel with *k* DNA sequences. If all the sequences in the kernel are of size *m*, then the maximum number of different oligomers in the kernel, *k·(m-n+1)*, may be reduced by repetitions. This will be particularly true, if *n* is small, because repetitions will be very likely with only 4 possible nucleotides for every position on the oligomer.

A distinctive oligomer is an oligomer with a low number of repetitions in a kernel. The cut-off value, *c*, is the maximum number of repetitions allowed in a kernel for an oligomer to be considered distinctive. The set of distinctive oligomers from a kernel is directly proportional to the cut-off value. Also, the longer an oligomer is, the larger the likelihood that it will be distinctive. These are not necessarily the only factors on determining oligomer distinctiveness. Additions, replacements, deletions and mutations of nucleotides in a DNA sequence may also contribute to the emergence of distinctive oligomers. The set of all distinctive oligomers of size *n* contained in a DNA sequence that belongs to a kernel constitutes the *n*-signature of the DNA sequence in the kernel under a given cut-off value. Ideal n-signatures of DNA sequences would discriminate among sequences of the same kernel, but in practice, kernel's sequences may share oligomers to a certain degree, especially if they are also related among themselves.

Our method consists of two basic steps: first, identifying distinctive oligomers in a kernel and determining the frequencies for the *n*-signatures of every DNA sequence in the kernel. Secondly, we classify the raw DNA strands as related to the DNA sequence or sequences which share the most distinctive oligomers in common. Details on these two steps will be shown in the following sections II and III, respectively. Section IV will talk about clustering DNA sequences. The paper will end with some final remarks in section V.

II. IDENTIFYING OLIGOMERS

Given a kernel of classified DNA sequences we find all distinctive oligomers of size *n* that appear in the population with a frequency that is less than or equal to a cut-off value of *c* occurrences. The choices of oligomer size and cut-off value will determine the number of distinctive oligomers found. Oligomer size cannot be too small, because shorter oligomers will have too many repetitions in the kernel. On the other hand, large oligomer sizes would be too specific, and although they may characterize a classified DNA sequence completely, they may not be found in the raw DNA strands, even when the strands are related to the classified DNA sequences. Evolutionary changes in the raw DNA strands and the classified DNA sequences will be the main reason for this discrepancy. In our studies we explore oligomer sizes in between 5 and 12.

Also, smaller cut-off values will mainly identify rare oligomers. For example a cut-off value of one will find unique oligomers in the kernel. However, more than one classified DNA sequence may share oligomers with others, especially if they are related; therefore considering some repetition is advisable to capture these relations. In our research we tested cut-off values from 1 to 12.

To see the effect of these two parameters, we analyzed two different kernels, one containing DNA sequences classified as V-exons and another one classified as J-exons. The number of oligomers we found under different oligomer sizes and cut-off values increased proportionally with both parameters. Figures 1 and 2 show this trend. The curves join together experimental points that share the same cut-off value. Notice that given an oligomer size, the percentage of distinct oligomers found increases when the cut-off value increases, but at higher cut-off values this effect is less pronounced.

Figure 1: Oligomers from a V Exon Kernel

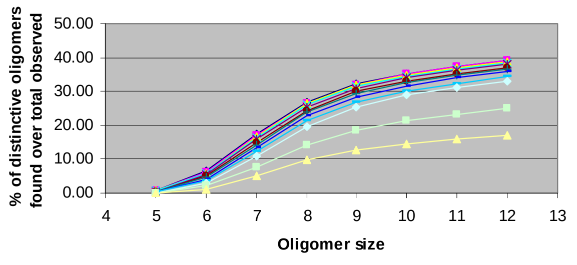
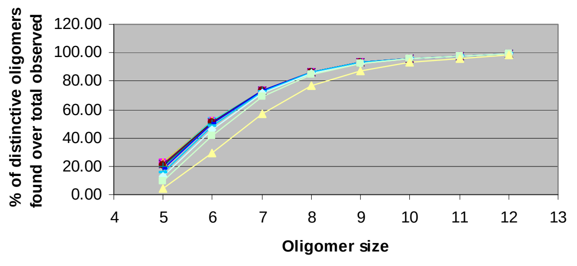


Figure 2: Oligomers from a J Exon Kernel



The final choice of parameters to use may vary with the nature of the kernel and the application. For example, kernels of DNA sequences that are tightly related may require higher cut-off values; also, applications required to maximize the number of classifications of raw DNA strands may benefit with longer oligomer sizes.

Once the oligomer size and the cut-off value are selected, and all distinctive oligomers in the kernel are found, for every DNA sequence in the kernel we count the number of occurrences of each distinctive oligomer they contain. This process generates a matrix of frequencies *P[d][k]*, where *d* is the number of distinctive oligomers, and *k* is the number of classified DNA sequences in the kernel. An entry *P[i][j]* will contain the number occurrences of the distinctive oligomer *i* in the classified DNA sequence *j*, and the j^{th} column of this matrix is the signature of the classified DNA sequence under this circumstances. The i^{th} row in the matrix contains information about how the copies of the distinctive oligomer *i* are distributed in the kernel. An oligomer *i* that is unique to a classified DNA sequence *j* will have a single entry in the matrix at *P[i][j]*, while an oligomer that appears in more than one classified DNA sequence will have multiple entries in the i^{th} row. However, the matrix of frequencies is a sparse matrix. Most of the times, the entries will contain zeros and ones, but they may contain any value that is lesser than or equal to the cut-off value. A matrix of frequencies is generated for every kernel. This matrix is independent of the raw DNA strands and once it is computed, it can be used for any classification of DNA strands. Figure 3 shows a section of a matrix of frequencies with a selected sample of distinctive oligomers and classified DNA sequences. The oligomer size was 10 nucleotides and the cut-off value was also 10.

Figure 3: Matrix of Frequencies (sample)

	Clasified DNA sequences (Exons)																			
Distinctive Oligomers	TRAV3-1-Exon2-1	TRAV13D-1-Exon2	TRAV13D-2-Exon2	TRAV3D-3-Exon2	TRAV13D-3-Exon2	TRAV13D-4-Exon2	TRAV13N-1-Exon2	TRAV13N-2-Exon2	TRAV3N-3-Exon2	TRAV13N-3-Exon2	TRAV13N-4-Exon2	TRAV13-1-Exon2	TRAV13-2-Exon2	TRAV3-3-Exon2-1	TRAV13-3-Exon2	TRAV13-4/DV7-Exon	TRAV3-4-Exon2-1	TRAV13-5-Exon2	TRAV21/DV12-Exon	TOTAL
CTGAACCTCA	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	5
CGGTCTACAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
GGCACTTATT	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	3
GGCACTTATC	0	1	1	0	1	0	1	1	0	1	1	1	0	0	1	1	0	0	0	10

III. CLASSIFYING DNA STRANDS

A DNA strand may be classified as related to one of more already classified DNA sequences in a kernel. In order to do that, we will use a vector of frequencies *p[k]*, with *k* elements, one for each of the classified DNA sequences in the kernel. The entries on this vector are all initialized to zero. We then proceed to search for all distinctive oligomers contained in the DNA strand. Every time we find an oligomer, we add its corresponding row from the matrix of frequencies *P[d][k]* onto the vector of frequencies *p[k].* For example, if a DNA strand contains only the first two distinct oligomers shown in Figure 3 (CTGAACCTCA and CGGTCTACAA), then its vector of frequencies will be the addition, column by column, of the first two rows in the matrix of frequencies. When this process ends, the entries in the vector of frequencies will indicate the relative strength in which every classified DNA sequence is related to the raw DNA strand. The DNA sequence or sequences with the highest value are the most likely to be related to the raw DNA strand, because they indicate how many oligomers they have in common.

Figure 4, in the next page, shows the results of the classification of a DNA strand with a kernel of immunoglobulin sequences. The signatures for the kernel were produced with an oligomer size of 10 and cut-off value of 10. Two V-exons and one J-exon produce the highest scores from their respective vector of frequencies. As indicated before, the scores indicate the relative support we have to consider these exons as related to the DNA strand. The numbers can be compared with elements of the same vector of frequencies, but not against other vectors, they are relative weights. The big difference between the weight for the first J-exon (weight of 46) and the other J-exons (weight of 2) clearly indicates that the first J-exon has a greater relationship with the DNA strand. The partial alignment of the three J-exons with the end of the DNA strand highlights the common distinctive oligomers. These distinctive oligomers are represented as capital letters. They may overlap, being this the reason why we see stretches of capital letters longer than the oligomer size of 10. Because the J-exon sequences are short we can appreciate the relation between the scores in the classification and the number of oligomers in the DNA sequences. The winner J-exon contains 46 distinctive oligomers overlapping each other, and all of them match the DNA strand. The other J-exons only have 2 distinctive oligomers each, also overlapping and matching the DNA strand. In this case the numbers 46 and 2 are the exact weights we obtained in the classification, indicating the number of distinctive oligomers in common. This will always be the case when the matrix of frequencies contains entries that are only zeroes and ones. The same happens in the classification of the DNA strand with the V-exons, but we cannot appreciate it in Figure 4 because it only contains a partial view of the alignment at the beginning of the DNA strand. What we can observe instead is that both V-exons provide a pretty good match to the DNA strand. This raises the possibility that both V-exons may be related and in fact could be close enough to form a cluster of similar DNA sequences. How we dealt with this situation is the topic of the following section.

IV. FINDING CLUSTERS OF RELATED DNA SEQUENCES

Whenever a DNA strand is closely related to more than one classified DNA sequence in a kernel, it bears the question if the classified DNA sequences are so related than they should form a cluster, rather than stand on their own. To discover if such clustering exists, we computed a square matrix of similarities *S[k][k]* among all the *k* sequences in the kernel. The element *S[i][j]* of this matrix contains a measurement of the similarity between the i^{th} and the j^{th} DNA sequences. We obtained these measurements by applying the classification algorithm described in the previous section to all classified DNA sequences in the kernel, a sort of bootstrapping process. As a consequence of this process, the matrix of similarities is made of vector of frequencies *p[k]* for all DNA sequences in the kernel.

Figure 4: Classification of DNA strand

```
Sample : S_572 was matched with the following V-exons:
   VExon 19            Score (relative weight): 73
   VExon 29            Score (relative weight): 72
```

Partial alignment of the beginning of sample with matched V-exons

```
-------------------------------------------------------------------------------------------------------
       | 0         1         2         3         4         5         6         7         8         9
       | 0123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
-------------------------------------------------------------------------------------------------------
S_572  |        gGGGGCTGCAGCTGCTCCTCAAGTACTATTCTGGAGACCCAGTGGTTCAAGGAGTGAACGGCTTCGAGGCTGAGTTCAGCAAGAGC
Vexon19| catggccgtggcctccagtttCTCCTCAAGTACTATTCgggaaACCCAGTGGTTCAAGGAGTGAACGGCTTCGAGGCTGAGTTCAGCAAGAGt
Vexon29| ccgcggcaGGGGCTGCAGCTGCTCCTCAAGTACTATTCaGGAGACCCAGTGGTTCAAGGAGTGAAtGGCTTCGAGGCTGAGTTCAGCAAGAGt
-------------------------------------------------------------------------------------------------------
```

```
Sample : S_572 was matched with the following J-exons:
   JExon 31            Score (relative weight): 46
   JExon 05            Score (relative weight):  2
   JExon 02            Score (relative weight):  2
```

Partial alignment of sample’s end with matched J-exons

```
-------------------------------------------------------------------------------------------------------
       |  0         1         2         3         4         5         6         7         8         9
       |  0123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
-------------------------------------------------------------------------------------------------------
S_572  |  gtgtacTTCTGTGCTATGAATTCTGGGACTTACCAGAGGTTTGGAACTGGGACAAAACTCCAAGTCGTTCCAAacatccagaacccagaac
Jexon31|                  caAATTCTGGGACTTACCAGAGGTTTGGAACTGGGACAAAACTCCAAGTCGTTCCAA
Jexon05|            tagcatcctcctccttcagcaagctggtgtttgggcaggggacatccttatcAGTCGTTCCAA
Jexon02|                ctcctgggacacccgacagatgtttTTTGGAACTGGcatagagctctttgtggagcccc
-------------------------------------------------------------------------------------------------------
```

The *i*th row value in the matrix of similarities is the vector of frequencies *p[k]* associated with the *i*th DNA sequence. The elements *S[i][i]* in its diagonal contain the total number of distinctive oligomers found in the *i*th DNA sequences. These values also indicate the maximum numbers for similarities with other DNA sequences, because no other one can share more distinctive oligomers with any DNA sequence than itself. This fact gives a mechanism to perform clustering of DNA sequences.

Whenever a DNA strand is classified as related to a set of DNA sequences, using the process outlined in the previous section, we consider the matrix of similarities and evaluate if the entries for the selected DNA sequences are close enough to consider them as a cluster. A threshold value for clustering, *T*, will be used to determine what close enough means. This will be a parameter for the clustering process. If the difference between the maximum number of similarities found in the diagonal *S[i][i]* of the matrix of references and the value for similarity with another DNA sequence *S[i][j]* is smaller than or equal to the threshold *T*, we can declare the DNA sequences to possibly be part of the same cluster, and the DNA strand related to the cluster.

Figure 5 shows the classification of another DNA strand. This time the strand is associated with a cluster of two DNA sequences. The cluster is created because the difference between the entries in the similarity matrix was 2, smaller than 20, the threshold value used for clustering. As the partial alignment of sequences show, both DNA sequences were very similar to each other, even in sectors unmatched by the raw DNA strand.

Figure 5: Classification of DNA strand to a cluster

```
Sample:S_393 was matched with the following V-exons:
  VExon-39           Scores (relative weight): 104
  VExon-49           Scores (relative weight): 103
```

Clustering analysis

```
      The following V-exons may be a cluster:
  VExon-39                   Similarity count: 245
  VExon-49                   Similarity count: 243
```

Partial alignment of the beginning of sample with matched V-exons

```
--------------------------------------------------------------------------------------------------------
     |  0         1         2         3         4         5         6         7         8         9
    |  012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
--------------------------------------------------------------------------------------------------------
S_393  |                                gGGGGCTGCAGCTGCTCCTCAAGTACTATCCAGGAGACCCAGTGGTTCAAGGAGTGAATGGCT
Vexon39|  atctgttctggtatgtccagtacccgcggcaGGGGCTGCAGCTGCTCCTCAAGTACTATCCAGGAGACCCAGTGGTTCAAGGAGTGAATGGCT
Vexon49|  atctgttctggtatgtccagtacccgcggcaGGGGCTGCAGCTGCTCCTCAAGTACTATCCAGGAGACCCAGTGGTTCAAGGAGTGAATGGCT
--------------------------------------------------------------------------------------------------------
```

V.FINAL REMARKS

We presented our technique to classify DNA strands using a kernel of previously known DNA sequences. We are currently using this technique to classify immunoglobulin samples from various species. This classification is a fast screening process to determine relations among our research subjects that will lead us to understand them better. We are currently working in a user-friendly interface to our algorithm for use in day-to-day analysis.

REFERENCES

[1] Gu, Wanjun., Castoe, Todd A., Hedges, Dale J., Batzer, Mark A. and Pollock, David D. (2008). Identification of repeat structure in large genomes using repeat probability clouds, *Analytical Biochemistry*, 2008.

[2] National Center for Biotechnology Information, Basic Local Alignment Search Tool (BLAST), http://blast.ncbi.nlm.nih.gov/Blast.cgi.

[3] National Center for Biotechnology Information, www.ncbi.nlm.nih.gov.

[4] Neil Hall (2007). Advanced sequencing technologies and their wider impact in microbiology. *The Journal of Experimental Biology,* 209:1518-1525.

[5] Ronaghi M, Uhlén M, Nyrén P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 1998 Jul 17; 281(5375):363, 365.

[6] 454 Life Sciences, a Roche Company, www.454.com.