Detecting Retroviruses in Genomic Sequences and Applying Signal Processing Techniques to Genomics: Literature Review

Qualifying Paper

Wendy Ashlock
Supervisor: Suprakash Datta

Sept. 24, 2010
Abstract

Identifying retroviruses in genomic sequences is a biological problem that benefits from a computational approach. Retroviruses integrate their genetic material into the genome of their host. They impact the welfare of their host in many ways: they cause disease; they affect the expression of genes; they protect against infection by other viruses. In particular, endogenous retroviruses (ERVs), retroviruses that have permanently integrated into their host genomes and are inherited, have influenced the evolution of the structure of the genome and act as living fossils giving us information about the relationships between species. ERVs are difficult to identify because they have a high mutation rate and often are defective (partly or wholly non-functional). This paper reviews four computational methods developed to annotate ERVs in the human genome and discusses approaches used to detect the broader category of transposons in the genomes of other organisms. It also discusses three signal processing techniques which have been effective for solving other problems in genomics and which are potentially useful for addressing the problem of detecting retroviruses.
## Contents

1 Introduction ........................................... 1

2 Biological Background ................................. 2
   2.1 Genomes and the Genetic Code .................... 2
   2.2 Viruses ........................................... 4
   2.3 Retroviruses ...................................... 7
   2.4 Retroviral Genomic Structure ................... 9

3 Existing Methods for Retroviral Detection .......... 11
   3.1 RepeatMasker .................................... 11
   3.2 HERVd ........................................... 14
   3.3 RetroSearch ..................................... 15
   3.4 RetroTector ...................................... 17
   3.5 ERVs in non-human species ...................... 19
   3.6 Difficulties Created by Sequence Assembly Methods . 21

4 Digital Signal Processing Techniques for Analyzing DNA sequences 23
   4.1 Turning DNA sequences into numerical sequences . 24
   4.2 Discrete Fourier Transform ....................... 25
   4.3 Autoregressive Models ............................ 30
   4.4 Entropy .......................................... 32

5 Conclusion ............................................. 33
List of Figures

1  Retrovirus Structure ......................................................... 11
2  Decoding using different reading frames. The DNA strand is broken into
codons on the left, and the symbols on the right represent amino acids. . . . 12
3  Histogram of phase values computed with a sliding window on a sequence
from a coding region from human chromosome 5. .......................... 27
4  Histogram of phase values computed with a sliding window on a sequence
from a non-coding region from human chromosome 3. ........................ 27
5  Histogram of phase values computed with a sliding window on the sequence
from Figure 3 with one base deleted creating two reading frames. .......... 29
6  Histogram of phase values computed with a sliding window on the sequence
from Figure 3 with two bases deleted creating three reading frames. ....... 29

List of Tables

1  Standard Genetic Code ....................................................... 5
2  Software tools for TE discovery. ......................................... 22
1 Introduction

When the Human Genome Project was completed in 2003, it was believed that understanding the genome amounted to finding all the genes and determining their functions. At that time, biologists viewed the genome using a “beads on a string” analogy, with the beads being genes, i.e. sequences that code for proteins. To their surprise, only a small percentage (about 2%) of the genome consisted of these sorts of sequences. The gene finding problem was much harder than anticipated. Computational tools, many using signal processing techniques, were developed to solve it.

Initially, the non-protein coding part of the genome was referred to as “junk DNA.” Once gene finding was well underway, many biologists turned their attention to the junk DNA and revised their belief that it was, indeed, junk. The junk was divided into different categories: pseudogenes, introns, telomeres, repetitive sequences, transposons, etc. The completion of the Chimpanzee Genome Project led to the discovery that the protein coding sequences of humans and chimps are virtually identical, but that there are significant differences in their non-protein coding sequences. Understanding the junk thus becomes important to understanding differences between closely related species.

This paper reviews approaches to the problem of detecting and annotating one particular category of the non-protein coding genome, the transposons, and, in particular, their subclass, the retroviruses. These elements have the property that they are mobile, i.e. they can make copies of themselves and relocate to other parts of the genome. This means that they have played a key role in the evolution of the structure of the genome. In addition, some are associated with disease and some affect the regulation of genes. Some have been
“domesticated” and now are genes themselves. Retroviruses can serve as genetic fossils, helping biologist build phylogenetic trees.

This literature review is intended as preparation for a project to detect retroviruses in genomes using signal processing techniques and machine learning. It consists of two parts. The first part includes the biological background needed to understand the problem, its significance, and the work that has already been done. The second part includes a review of signal processing techniques that could be useful in addressing the problem, along with discussion of how these techniques have been used to address other problems in genomics.

2 Biological Background

2.1 Genomes and the Genetic Code

From the computer science point of view, genomes are long alphabetic sequences using the alphabet \{A,C,G,T\}. From the biological point of view, they are sequences of nucleotides consisting of a sugar/phospate backbone and four types of bases, adenine (A), cytosine (C), guanine (G), and thymine (T). A record of the sequence of base for many organisms, including human beings, is stored in a public database called NCBI (National Center for Biotechnology Information) \texttt{http://www.ncbi.nlm.nih.gov/}. Even though many of these organisms are diploid (two copies of each chromosome), most of the genomes in NCBI include just one copy of each chromosome for either a single individual representative or some amalgamation of data from multiple individuals into one copy. Genomes are assembled from shorter sequence reads (see Section 3.6). The sequence reads are also publicly available as
are many other short sequences from various research projects. The quality of annotations for these genomes varies widely with the highest quality annotations existing for those organisms which were sequenced earliest and which have been studied the most. Among the first organisms sequenced were: *Saccharomyces cerevisiae* (baker’s yeast), *Caenorhabditis elegans* (roundworm), *Drosophila melanogaster* (fruit fly), *Homo sapiens* (human beings), and *Arabidopsis thaliana* (thale cress, a small flowering plant). The annotations for these organisms can be used as starting points for annotating other sequenced genomes. For example, one can search an unannotated genome for sequences similar to those annotated as retroviruses in another genome.

In addition to the reference genome in NCBI, databases are available with information about individual differences in the human genome. The International HapMap Project [http://hapmap.ncbi.nlm.nih.gov/](http://hapmap.ncbi.nlm.nih.gov/) is a good source of these. This project includes DNA samples from 270 individuals from Nigeria, Japan, China, and the US (of European ancestry). They are not complete genomes, but selected pieces aligned to the reference genome so they can be compared. Most of these individual differences involve inserted, deleted or inverted sequences, some of which could be retroviruses.

Computational methods are needed for dealing with the vast amount of data generated by these projects. Many of these methods rely on statistical properties inherent in particular regions of the genome. For example, when DNA codes for proteins, it uses a genetic code (see Table 1) consisting of groups of three bases (codons). Each codon codes for an amino acid or a start or stop signal. (In the table, amino acids are represented by their one-letter abbreviation.) Strings of amino acids make up proteins. Most of life is thought to use this genetic code, but a few variant codes are also used. This code is degenerate – multiple codons
code for the same amino acid. This degeneracy leads to detectable statistical properties in the genome or regions of the genome. For example, in regions that use this genetic code (protein coding regions), the third position of the codon is more random than the other two positions. In addition, sometimes a particular choice of codon for a given amino acid is preferred. This “codon usage” bias can be detected with statistical techniques.

Regions that do not use the genetic code also often have statistical features that can be detected. Some regions have biases for particular bases. This is characterized in terms of AT-richness (percentage of bases which are A or T). A C base directly followed by a G base is written CpG. These pairs are rare in most of the genome because their chemistry encourages mutation. However, there are regions of the genome where they are common. These regions are called CpG islands and can be detected statistically. It is also common to have short sequences repeated many times. These are called tandem repeats and are statistically detectable.

2.2 Viruses

Viruses are genetic parasites that can only replicate using the cellular systems of a host. They are similar to living organisms in many ways: they can die; they evolve by natural selection; virus species can become extinct. The most studied viruses are associated with disease (examples include the H1N1 influenza virus, the herpes virus, and the HIV virus), but the majority of known viruses do not cause disease, and some are even beneficial to their hosts. Viruses are ancient. They have been part of life and part of evolution for hundreds of millions of years. Virus particles are simple. They have only two or three parts: genes made
of RNA or DNA, a protein coat protecting the genes, and, sometimes, an envelope made of lipids surrounding the entire particle.

Viruses are particular about what type of cell they infect. In order for a virus to infect a cell there must be the proper type of receptor on the cell’s surface. For example, the viruses that cause the common cold (rhinoviruses) bind to ICAM-1 receptors, receptors which are also used by white blood cells. These receptors are found on skin cells in the respiratory tract. HIV, the virus that causes AIDS, infects only cells in the immune system and central nervous system. Herpes viruses infect skin cells and nerve cells. Viruses infect cells from all types of life. There are viruses that infect animals, plants, bacteria, fungi, algae, even other viruses. Some viruses are species specific; they only infect cells from a particular species. Many others invade cells from a broad range of species. This can result in horizontal gene transfer, the transfer of genes from one adult organism to another adult organism (possibly

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>TTT F</td>
<td>TCT S</td>
<td>TAT Y</td>
<td>TGT C</td>
</tr>
<tr>
<td></td>
<td>TTC F</td>
<td>TCC S</td>
<td>TAC Y</td>
<td>TGC C</td>
</tr>
<tr>
<td></td>
<td>TTA L</td>
<td>TCA S</td>
<td>TAA STOP</td>
<td>TGA STOP</td>
</tr>
<tr>
<td></td>
<td>TTG L</td>
<td>TCG S</td>
<td>TAG STOP</td>
<td>TGG W</td>
</tr>
<tr>
<td>C</td>
<td>CTT L</td>
<td>CCT P</td>
<td>CAT H</td>
<td>CGT R</td>
</tr>
<tr>
<td></td>
<td>CTC L</td>
<td>CCC P</td>
<td>CAC H</td>
<td>CGC R</td>
</tr>
<tr>
<td></td>
<td>CTA L</td>
<td>CCA P</td>
<td>CAA Q</td>
<td>CGA R</td>
</tr>
<tr>
<td></td>
<td>CTG L</td>
<td>CCG P</td>
<td>CAG Q</td>
<td>CGG R</td>
</tr>
<tr>
<td>A</td>
<td>ATT I</td>
<td>ACT T</td>
<td>AAT N</td>
<td>AGT S</td>
</tr>
<tr>
<td></td>
<td>ATC I</td>
<td>ACC T</td>
<td>AAC N</td>
<td>AGC S</td>
</tr>
<tr>
<td></td>
<td>ATA I</td>
<td>ACA T</td>
<td>AAA K</td>
<td>AGA R</td>
</tr>
<tr>
<td></td>
<td>ATG START</td>
<td>ACG T</td>
<td>AAG K</td>
<td>AGG R</td>
</tr>
<tr>
<td>G</td>
<td>GTT V</td>
<td>GCT A</td>
<td>GAT D</td>
<td>GGT G</td>
</tr>
<tr>
<td></td>
<td>GTC V</td>
<td>GCC A</td>
<td>GAC D</td>
<td>GGC G</td>
</tr>
<tr>
<td></td>
<td>GTA V</td>
<td>GCA A</td>
<td>GAA E</td>
<td>GGA G</td>
</tr>
<tr>
<td></td>
<td>GTG V</td>
<td>GCG A</td>
<td>GAG E</td>
<td>GGG G</td>
</tr>
</tbody>
</table>

Table 1: Standard Genetic Code
of another species). This is in contrast to the more usual vertical gene transfer in which genes are transferred from parent to child.

Viruses have two possible life strategies: some are acute, and some are persistent. Acute viruses kill their host cells. Persistent viruses integrate into their host cells, becoming a permanent part of them. Most epidemic disease viruses are acute. Examples include smallpox, measles, and flu. Persistent viruses can either be latent or chronic. Latent (or cryptic) viruses become a part of their host cell, but do not reproduce. Chronic viruses maintain a more or less constant level of reproduction. Examples include herpes, papillomavirus, hantaviruses, and retroviruses. Some viruses are defective, requiring not only the cellular systems of their host cells for reproduction, but also the help of other viruses. This phenomenon has been experimentally verified for retroviral families with many copies in a given genome. Their expression level is consistent with the number of copies, even though only some are intact.

The oldest classification system for viruses was created in 1971 by David Baltimore and is called the Baltimore classification. It divides viruses into seven classes based on type of nucleic acid, number of strands, strand direction, and method of replication. There are double-stranded DNA viruses (Class I), single-stranded DNA viruses (Class II), double-stranded RNA viruses (Class III), sense-strand RNA viruses (Class IV), and antisense-strand RNA viruses (Class V). In addition, there are viruses that contain a gene for reverse transcriptase (RT), an enzyme that converts RNA into DNA. There are two types of reverse transcriptase viruses: double-stranded DNA RT viruses (Class VII) and single-stranded RNA RT viruses or retroviruses (Class VI). Retroviruses are the focus of this paper. RT viruses are of historical interest because they contradict the Central Dogma of Molecular Biology. The Central Dogma says that DNA is transcribed into RNA which is translated into protein. RT viruses
operate the other way around: RNA is transcribed into DNA.

Beyond the basic Baltimore classification, there is a desire to subdivide the viruses into smaller groups. Many criteria have been proposed including those used by the Baltimore classification and others such as which species they infect, which diseases they cause, and sequence similarity. A taxonomy of viruses which is considered somewhat of a standard is maintained by the International Committee on Taxonomy of Viruses (ICTV) [1], but there are many viruses that are not included. Researchers who work on retroviruses tend to classify them in families based on sequence similarity of the gene that codes for reverse transcriptase. For more information about viruses and their role in evolution, see [67].

2.3 Retroviruses

Retroviruses are single-stranded RNA RT viruses. When they enter a cell, their RNA is converted to DNA using reverse transcriptase and inserted into the DNA of their host. When the host cell is a germ cell (egg or sperm), the retroviral DNA becomes heritable. This DNA is referred to as an endogenous retrovirus (ERV), a provirus, or an LTR retrotransposon. Human ERVs are referred to as HERVs, cow ERVs as BERVs (bovine ERVs), sheep ERVs as OERVs (ovine ERVs), etc. To distinguish them from ERVs, retroviruses that have not integrated into a cell’s DNA are sometimes referred to as exogenous retroviruses or wild retroviruses. The ERV will, of course, only become a permanent feature in the species’s population if it is not harmful to the host, or, at least, not too harmful. If it is beneficial, it will undergo positive selection. Some endogenous retroviruses become defective or non-functional due to mutation. It is estimated that about 8% of the human genome is made up
of endogenous retroviruses [72]. Some can be found in similar locations on the chromosomes of related species. Some are found in distantly related species, implying they have been part of the genome for a very long time. A retrovirus named Phoenix was estimated to have been part of the human genome for five million years [15]. (Modern humans have only existed for 50,000 - 100,000 years.) Phoenix was shown to be still capable of producing infectious particles. Endogenous retroviruses like Phoenix are useful to evolutionary biologists as living molecular fossils which help them determine the relationships between species.

Endogenous retroviruses are part of a class of genomic elements called transposons, transposable elements (TEs), or mobile genetic elements (MGEs) which were first discovered by Barbara McClintock [18]. About half of the human genome is composed of transposable elements. Transposons that use reverse transcriptase to transcribe (like ERVs) are called retrotransposons. Other important retrotransposons include LINEs (long interspersed elements) and SINEs (short interspersed elements). LINEs range from 900-6000 base pairs (bps); SINES from 200-400 bps. SINEs and LINEs are derived from RNA polymerase transcripts. RNA polymerase is the enzyme that catalyzes RNA synthesis from DNA. LINEs are derived from RNA polymerase II; SINEs are derived from RNA polymerase III. LINEs make up about 21% of the human genome; SINEs about 11% [49]. LINEs and SINEs differ from retroviruses in that they do not have long terminal repeats (LTRs), identical sequences at their beginnings and ends. ERVs are part of a group called LTR retrotransposons. Transposons that do not transcribe using reverse transcriptase are called DNA transposons. They are common in bacteria; less common but still present in eukaryotic organisms including humans.

In addition to the original insertion, copies of transposons are made and inserted else-
where in the genome (retrotransposition). This is why they comprise so much of the genome. The most common transposon in humans is the SINE family Alu. Our genome has about 300,000 copies of Alu, one for every 6K of DNA. Some retroviruses are repeated only a few times; others are repeated hundreds of times [39]. Sometimes these copies carry part of the cellular DNA with them, rearranging the genome. Sometimes they affect the function of neighbouring genes just through their presence. For example, they can interfere with regulatory sequences. Retroviruses can cause cancer in this way [35]. Feline leukemia virus is an example of a retrovirus that causes cancer by activating an inactive cancer-causing gene (oncogene) in the cell. Some cancer-causing retroviruses carry oncogenes. The Rous sarcoma virus is an example of this.

Some retroviruses perform useful functions like producing the protein which causes immunosuppression in the human placenta enabling extended human gestation [42]. They may protect us from infection by exogenous retroviruses. Our immune system appears to have arisen from a retrovirus, although the viral ancestor of adaptive immunity remains to be found [67]. Understanding retroviruses is also important because of the potential for the use of retroviruses in gene therapy. In gene therapy, a gene is inserted in a harmless retrovirus which can then insert the gene into the patient’s cells. For more information about retroviruses, see [10, 39, 65, 22, 19].

2.4 Retroviral Genomic Structure

Retroviruses are difficult to detect using sequence homology, i.e., by looking for sequences similar to known retroviral sequences, because they are diverse, having only small portions
of their genomes in common. For example, it is estimated that there are $10^{60}$ variants of HIV [67]. An analysis of HIV-1 copies within a single nucleus showed 28% variation in the $env$ gene [70], more than the average protein variation between birds and humans. One explanation for this variability is that when RNA is converted to DNA using reverse transcriptase, there is no error correction such as there is when DNA is copied. Also, due to a lack of selective pressure, many endogenous retroviruses are heavily mutated to the point of being defective or even completely non-functional. These defective retroviruses are still of interest, however, because of their past influence on the genome, their value as molecular fossils, and because they can function with the help of other retroviruses.

An alternative to detection using sequence homology is detection based on structure. Retrovirus genomes have a consistent structure. See Figure 1 for an illustration. They range in size from 5000 to 20,000 nucleotides. An intact retrovirus begins and ends with a short (18-250 nucleotides) sequence called a direct repeat or long terminal repeat (LTR). Another short non-coding sequence follows. There is then an 18 nucleotide primer binding site (PBS). A somewhat longer (90-500 nucleotides) sequence, called a leader, follows. Then, come the genes. All retroviruses have three genes that always occur in the same order ($gag$, $pol$, and $env$). Some retroviruses have additional genes. For example, lentiviruses (like HIV) have six additional genes. After the genes, there is a very short (about 10 nucleotides) sequence called the polypurine tract (PPT). Then there is a short non-coding sequence, followed by the same direct repeat sequence there was at the beginning. In addition, the DNA flanking the endogenous retrovirus often includes a target site duplication (TSD) consisting of the 4-6 bases at the insertion point. Copies of these bases can be found on either side of the inserted retrovirus.
Retroviral genes have an unusual structural feature – they use overlapping reading frames. Reading frames in DNA arise from the fact that the genetic code maps nucleotides onto proteins in groups of three. This means that the code translates differently depending on whether decoding begins at position 0, position 1, or position 2. For an illustration of this see Figure 2. Starting at position 3 will have the same result as starting at position 0 (excluding the first protein), so the codes starting at position 0 and position 3 are said to be in the same reading frame. On any segment of DNA, there are six possible reading frames, three in each direction (sense and antisense). The genes in the cells that host the viruses mostly use a single reading frame. Viral genes, however, use all three reading frames in a given direction with the end of one gene often overlapping the beginning of another.

3 Existing Methods for Retroviral Detection

3.1 RepeatMasker

RepeatMasker [57] is a program designed to screen DNA sequences for repeated elements (including ERVs) and for low complexity sequences. About half the human genome falls into these categories. Users of RepeatMasker have to identify the species of the input sequence. Best results are obtained for human and mouse sequences since these have the best collections
Reading Frame 1: TCA GGT GCC AAC GTG GA? → SGANV . . .
Reading Frame 2: CAG GTG CCA ACG TGG A?? → QVPTW . . .
Reading Frame 3: AGG TGC CAA CGT GGA ?? → RCQRG . . .
Reading Frame 4: TCC ACG TTG GCA CCT GA? → STLAP . . .
Reading Frame 5: CCA CGT TGG CAC CTG A?? → PRWHL . . .
Reading Frame 6: CAC GTT GGC ACC TGA ?? → HVGT{STOP} . . .

Figure 2: Decoding using different reading frames. The DNA strand is broken into codons on the left, and the symbols on the right represent amino acids.
of reference consensus sequences. Repeatmasker’s primary purpose is to mask parts of a DNA sequence whose presence could result in false positive matches in another search. Since it categorizes the type of repeats it finds, it is also useful for those interested in a particular type of repeat, like ERVs.

RepeatMasker starts with a database of consensus sequences for each repeat type taken from RepBase [28, 27, 26]. RepBase is a database of consensus sequences for repeated elements in genomes. Consensus sequences are sequences derived from a group of similar sequences. The sequences are aligned (matched, possibly with gaps), and the most common value at each position is assigned to the consensus sequence. RepBase has 475 consensus sequences for HERVs. A dynamic programming algorithm, called the Smith-Waterman-Gotoh algorithm [58, 20] is used to align the consensus sequences to the input sequence. Sequence alignment is challenging, because mutations can cause substitutions, insertions, and deletions in sequences. So, to get the best match, it is necessary to decide where to put gaps and how big to make them. Also, if the sequences are different lengths, there can be many choices of where to start the alignment. The Smith-Waterman-Gotoh algorithm finds the optimal alignment by constructing a matrix with scores for possible alignments. Scores for alignments are based on a weighting system designed so that matches improve the score, mismatches detract from it, and there is a penalty for gaps. The score at position \((i,j)\) in the matrix represents the best possible score for the first \(i\) bases in sequence 1 and the first \(j\) bases in sequence 2. The optimal alignment can be constructed by backtracking from the highest score in the matrix.
3.2 HERVd

HERVd [46, 45] is a database of retroviruses in the human genome. It is available at http://herv.img.cas.cz/ and was last updated in 2003. It is based on the build of the human genome current at that time. The user can search the database for a specific HERV (using its ID number), for HERVs with a specified range of lengths, for HERVs in a particular family, for HERVs on a particular chromosome or in a particular section of a chromosome, for HERVs with a given orientation (sense or antisense), or for HERVs with a specified GC content (proportion of G and C bases), or for any combination of these. The original database included 39 HERV families and identified about a third of HERVs. They were able to increase the number of HERVs identified when more consensus sequences were identified, doubling the number of HERV families they could include.

The database was assembled starting with the consensus sequences for various families of HERVs in RepBase [28, 27, 26]. RepeatMasker [57] was used to screen for non-retroviral transposons that would confuse the search and to search for matches to the consensus sequences. To cope with identification problems caused by mutations resulting in insertions and deletions (including retroviruses inserted into other retroviruses), HERVd employs a defragmentation algorithm. Complete details of this algorithm are not given in [46, 45]. All that is said is that it uses profile hidden Markov models [17] to define HERV families and to assign HERV sequences to these families. The defragmentation algorithm pieces together the sequence fragments identified by RepeatMasker. In addition, the flanking DNA is examined in an attempt to identify TSDs. The presence or absence of TSDs is used to distinguish HERVs from other sorts of transposons with similar sequences.
Advantages of this method include having the identified retroviruses classified by family and being able to find retroviruses that are heavily mutated and fragmented. The major disadvantage is that the process must be repeated every time a new build of the genome is done and that it must be custom designed for a particular species (in this case, human beings). Another disadvantage is that it is only able to find retroviruses that belong to known families.

The HERVd database has been used in various ways by researchers. Some of the authors of the database, together with some other researchers, used it to study the HERV-W family [47]. The HERV-W family is important, because of its possible role in multiple sclerosis, rheumatoid arthritis, and schizophrenia. Their studies suggested that scientists studying these diseases should focus on a particular subset of HERV-Ws. In [11], the HERVd database was used to study the sequence CCTGTT, a sequence that occurs unusually often in the human genome. The authors used HERVd to discover that this sequence occurs even more often in HERVs. [7] used HERVd to study mutations in the \textit{env} gene of various HERV families. Since the \textit{env} gene is only needed if the retrovirus leaves the cell, the type of mutations found in it are a clue to whether copies of the retrovirus were created by infection from exogenous viruses or from retrotransposition.

3.3 RetroSearch

Another database of HERVs is called RetroSearch [68, 2, 30], \url{www.retrosearch.dk}. RetroSearch lets users search for HERVs in the 2003 Human Genome Assembly by ID number, HERV family, location (a range can be specified), minimum length, minimum ORF length,
The RetroSearch database was built starting with a query database of 237 endogenous and exogenous retroviruses from a variety of host organisms. These sequences were edited so that they contained only the part of the retrovirus that codes for genes. This was to avoid finding solo LTRs. (Sometimes when a retrovirus is copied, it is deleted from the host genome leaving behind one of the non-coding repeated regions at the beginning or end of its genome. This is called a solo LTR.) A BLAST search [5] of the human genome from the query database was done. The BLAST algorithm is similar to the Smith-Waterman-Gotoh algorithm, except that it uses a heuristic instead of an exhaustive search. This allows it to run about 50 times faster at the cost of some accuracy. It identifies matches (hits) between the query and the input sequence and assigns scores. Overlapping hits were clustered together and assigned a region score based on the BLAST scores of the sequences. If the score was high enough, the flanking DNA on either side was examined for LTRs. The results were called “putative HERVs”. The next step was to find ORFs in the putative HERVs. This was done by searching for stop codons within the putative HERVs. Regions between two stop codons that were long enough (> 62 nucleotides) were compared to a database of over 6000 retroviral and non-retroviral proteins using a FASTA search [37, 48]. A FASTA search is similar to a BLAST search, except that it is especially tuned for aligning proteins. Regions with retroviral ORFs were identified as HERVs. The original database using this method
identified 1.1% of the human genome as containing HERVs (about 14% of HERVs). This database has been updated and expanded since. Data is displayed online together with data for the same regions from RepeatMasker. Often, there are noticeable differences.

Advantages of this method include having the HERV ORFs identified and searchable. One can, for example, search for all HERVs that have an env ORF that is more than 200 nucleotides long. (It finds 493 of these.) This makes it possible to assemble custom databases for studying particular retroviral genes. This database is pickier about its choices than HERVd, identifying fewer, but doing more to verify that the retroviruses in it are actually retroviruses and not just sequences that resemble retroviruses. The disadvantages are similar to those of HERVd: the process must be repeated for every new build of the genome sequence, it only includes human ERVs, and only retroviruses similar to known retroviruses are found.

RetroSearch has been used primarily for studying intact retroviral genes. In particular, it was used to study a HERV envelope gene that is expressed in the placentas of monkeys and apes [30].

### 3.4 RetroTector

The RetroTector algorithm [59, 60] takes as input a genomic sequence of length 5,000 to 10,000,000 bps. It was originally designed only for human genomic sequences, but has been extended to accommodate primate, chicken, cow, dog, elephant, horse, lizard, mouse, and zebrafish genomes. However, the species must be given as input to the algorithm. RetroTector scans the input sequence, finds ERVs, identifies the LTRs and retroviral protein genes, and
outputs them in an annotated format such as is found in textbooks. It is available online at http://www.kvir.uu.se/RetroTector/RetroTectorProject.html.

The algorithm has five parts:

1. First, it “sweeps” the input sequence, masking out all Alu and L1 fragments, since these could be confused with ERVs. Alu is the most common SINE, and L1 is the most common LINE.

2. Next, it searches for LTR pairs and attempts to find solo LTRs. Solo LTRs are usually remnants of a cut-and-paste transposition in which the “cut” is incomplete.

3. Then, it searches for motifs (using a library of 275). Motif is defined loosely. From a programmer’s point of view, a motif is a procedure for detecting a conserved ERV trait, taking into account the possibility of mutation. Most of the motifs are procedures that detect close matches to specified amino acid sequences, but they also use trained neural nets and some other procedures. The program is designed so that it is easy to add new motif modules. Each motif has constraints on its relationship to other motif hits and the LTRs (distance from and relative position). Also, each motif is assigned to a particular retrovirus genera. Using this information, the algorithm puts together chains of motifs and LTRs, creating a putative ERV. This is the part of the algorithm that is species specific. New motifs and constraints have to be devised for each species.

4. Given the putative ERV, the algorithm tries to identify the genes for gag, pro, pol, and env proteins. These are proteins common to all retroviruses.

5. The last step is to look for other possible genes.
Retrotector has been used to build a retroviral taxonomy [23], to study HERV gene expression in cancer [3], and to study the impact of retroviral defense mechanisms, known to disable exogenous retroviruses, against endogenous retroviruses [24].

3.5 ERVs in non-human species

Researchers focusing on organisms other than human beings prefer different terminology. Instead of “endogenous retroviruses” they say “LTR retrotransposons” (a somewhat broader category) and the focus is more often directed towards discovering how they affect the operation and evolution of the genome rather than on how they impact disease. The researchers are interested not in specifically finding ERVs, but in finding all repeated elements, all transposable elements (abbreviated TE or MGE for mobile genetic element), or all retrotransposons (SINEs, LINEs in addition to ERVs). Another distinction they make is between Class I and Class II transposable elements. Class 1 TEs do not use RNA intermediates when they replicate. Class II TEs do use RNA intermediates and are also called “retrotransposons” or “retroelements”. Other common distinctions are “replicative” (copy-and-paste transposition) and “nonreplicative” (cut-and-paste transposition) and “autonomous” (encode genes for replication) and “nonautonomous” (use genes from other TEs to transpose). See a good genetics textbook, such as [34] or [49], for more information.

There are many software tools designed to perform these tasks. For a partial list, see Table 2. For a review, see [8]. There are four different approaches: repeat finding methods, homology-based methods, structure-based methods, and comparative genomic methods. Table 2 lists some of the tools along with which methods they use. Repeat finding methods look
for repeated sequences in the genome. They use computational strategies such as suffix trees and hashing. Homology-based methods take advantage of prior knowledge by comparing sequence fragments to a database of classified sequences. Some use a direct comparison with the sequences in the database; others compare the sequences using profile hidden Markov models. Structure-based methods use known characteristics of the structure of the elements to find them. They look for characteristic features of LTRs, TSDs, PBSs, PPTs, and sequences that code for proteins found in the elements and identify subsequences that have these features spaced appropriately. Comparative genomic methods compare closely related genomes, either from the same species or closely related species. Regions that exist in one genome but not in others are likely transposons. After a new transposon family has been discovered, a consensus sequence is created and put in Repbase [28] to be used in future searches using homology-based methods.

It seems odd that so many tools have been created, especially since they are freely shared. Why not just develop one good tool and use it? The reason is that each tool has strengths and weaknesses, and the best results are obtained by using a combination of them. [55] compared six de novo repeat finding algorithms using the same data from the rice genome and found that their results were profoundly different. [25, 21, 52] are examples of studies in which a combination of tools were used with good results in the rice, chicken, and fruit fly genomes.

[52] describes the process by which the transposons in the Drosophila genome were annotated. Since Drosophila was the first large genome sequenced, it is of particularly high quality and its assembly has been well verified, making it a good choice for developing a model process. Quesneville et. al. developed a pipeline for annotating the transposons in
Release 4 of the Drosophila genome using the manually curated set of annotations from the Release 3 genome as a benchmark to test their technique. They compiled results using four homology methods and four *de novo* methods and gave these results to five human curators each working on a separate chromosome arm. A single human curator did a second pass to improve consistency. The fact that to get a high quality annotation they needed to use eight methods plus manual curators demonstrates the importance of developing new methods using novel approaches.

### 3.6 Difficulties Created by Sequence Assembly Methods

The retroviral detection methods described in the previous sections operate on assembled sequences. A limitation of their effectiveness results from the way sequence assembly is done. With current technology, the longest DNA strand whose sequence can be directly determined is 1000 bps long. The shortest human chromosome is 50,000,000 bps. In order to sequence a chromosome, a process called shotgun sequencing is done. In shotgun sequencing, the DNA is randomly shattered into pieces using ultrasound, the pieces are inserted into cloning vectors (known sequences of DNA), the cloning vectors are inserted into a bacteria, multiplied until there are enough to sequence, sequenced, and then the pieces are put them together like a jigsaw puzzle. The putting together step is called sequence assembly. Typically, the genome is oversampled by 20-30 times so that little is missed. Sequence assembly is not a trivial task and the primary reason is the repeated regions, like SINEs, LINEs, and retroviruses. It is like putting together a jigsaw puzzle in which large numbers of the pieces are identical.
<table>
<thead>
<tr>
<th>name</th>
<th>method</th>
<th>website</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSAHA [44]</td>
<td>repeat finding</td>
<td><a href="http://www.sanger.ac.uk/resources/software/ssaha/">http://www.sanger.ac.uk/resources/software/ssaha/</a></td>
</tr>
<tr>
<td>REPuter [33, 32]</td>
<td>repeat finding</td>
<td><a href="http://bibiserv.techfak.uni-bielefeld.de/reputer/">http://bibiserv.techfak.uni-bielefeld.de/reputer/</a></td>
</tr>
<tr>
<td>RepeatScout [51]</td>
<td>repeat finding</td>
<td><a href="http://bix.ucsd.edu/repeatscout/">http://bix.ucsd.edu/repeatscout/</a></td>
</tr>
<tr>
<td>HMMER [16]</td>
<td>homology</td>
<td><a href="http://hmmer.janelia.org/">http://hmmer.janelia.org/</a></td>
</tr>
<tr>
<td>MGEscan-LTR [53]</td>
<td>structure</td>
<td><a href="http://darwin.informatics.indiana.edu/cgi-bin/evolution/ltr.pl">http://darwin.informatics.indiana.edu/cgi-bin/evolution/ltr.pl</a></td>
</tr>
<tr>
<td>LTR_STRUC [41]</td>
<td>structure</td>
<td><a href="http://www.mcdonaldlab.biology.gatech.edu/finalLTR.htm">http://www.mcdonaldlab.biology.gatech.edu/finalLTR.htm</a></td>
</tr>
<tr>
<td>LTR_par [29]</td>
<td>structure</td>
<td><a href="http://www.eecs.wsu.edu/~ananth/software.htm">http://www.eecs.wsu.edu/~ananth/software.htm</a></td>
</tr>
<tr>
<td>LTR_FINDER [73]</td>
<td>structure</td>
<td><a href="http://tlife.fudan.edu.cn/ltr_finder/">http://tlife.fudan.edu.cn/ltr_finder/</a></td>
</tr>
<tr>
<td>LTRdigest [61]</td>
<td>structure</td>
<td><a href="http://genometools.org">http://genometools.org</a></td>
</tr>
</tbody>
</table>

Table 2: Software tools for TE discovery.
The assembled sequence also includes gaps, due to the fact that the sampling isn’t really random; the cloning vectors prefer some pieces over others. The assembly process is described in [43, 50, 56]. The consequences of using assembled sequences for detecting retroviruses are that the ERVs are likely to be included in the unsequenced gaps (for example, if they are toxic to the bacteria in which they are grown), are likely to be put in the wrong place, and some copies are likely to be left out entirely.

4 Digital Signal Processing Techniques for Analyzing DNA sequences

Digital signal processing techniques have been applied to various problems in genomics. One of the most important is gene finding. Before the Human Genome Project was completed in 2003, it was believed that chromosomes were strings of genes. In fact, it turns out that only a small percentage of the human genome (about 2%) consists of genes. In addition, a gene is not a straightforward sequence beginning with a start code and ending with a stop code. Instead, it is a complex mixture of regions that code for protein (exons) and regions containing regulatory and other elements (introns). Thus, the problem of identifying genes and the exons within them is a non-trivial one. Overviews of how signal processing has been used in gene finding and in other genomics problems can be found in [38, 66, 6, 4].
4.1 Turning DNA sequences into numerical sequences

In order to apply signal processing techniques to DNA sequences, it is necessary to turn them into numeric sequences. There are many ways to do this. One common way is to use binary indicator sequences: four sequences, one for each nucleotide. The A-sequence, for example, would have a 1 everywhere the sequence had an A and a zero elsewhere. This method was first used by Voss [69] and is referred to as the Voss representation. It has since been used by [62, 14, 31] and many others.

There is some concern that results based on these sequences could be an artifact of the representation. In [54], Rushdi and Tuqan compare the Voss representation to four other numeric representations and show that they all yield the same DNA Fourier spectrum (explained in Section 4.2). The four representations they examine are: tetrahedral mappings, quaternions, simplex mappings, and Z-curve mappings.

Tetrahedral mappings map \( \{ A, C, G, T \} \rightarrow \{ 1 + j, -1 + j, -1 - j, 1 - j \} \) where \( j = \sqrt{-1} \). These values represent the corners of a tetrahedron projected onto the complex plane. Quaternions, a generalization of the tetrahedral mapping, map \( \{ A, C, G, T \} \rightarrow \{ i + j + k, i - j - k, -i - j + k, -i + j - k \} \) where these are hypercomplex numbers such that \( i^2 = j^2 = k^2 = ijk = -1 \). Simplex mappings are a transformation of tetrahedral mappings from four sequences to three sequences.

Z-curve mappings create three sequences based on pairings of bases: AG/CT, AC/GT, and AT/GC. These pairings are used because they have biological meaning: AG/CT distinguish purines and pyrimidines; AC/GT distinguish amino and keto bases; AT/GC distinguish bases with weak and strong hydrogen bonds. The sequence has a 1 if the base is one of the
pair, -1 otherwise. Their name is derived from Z-curves [75], a method used to create a graphical representation of DNA.

Wang and Schonfeld further develop the theory needed for comparing representations in [71]. They use their theory to compare the Voss representation to the representation that uses two sequences such that in one sequence there is a -1 for A, a 1 for T, and zeros for C and G, and in the other sequence there is a -1 for C, a 1 for G, and zeros for A and T. They show that these different representations do not produce equivalent results. They also compare the Voss representation to a mapping that creates four sequences: one sequence maps A to \( \frac{1}{\sqrt{2}} \) and G to \(-\frac{1}{\sqrt{2}}\) and C and T to 0; one maps T to \( \frac{1}{\sqrt{2}} \) and C to \(-\frac{1}{\sqrt{2}}\) and A and G to 0; one maps A and G to \( \frac{1}{\sqrt{2}} \) and C and T to zero; the last maps C and T to \( \frac{1}{\sqrt{2}} \) and A and G to zero. They show that this representation produces the same Fourier spectral results as the Voss representation. They show that rotation is the unique equivalent transformation from one mapping to another that leads to consistent results, and that, when there is inconsistency, it increases as the window size of the analyzed DNA sequences increases.

Real number mappings are used as well, particularly with AR models (see Section 4.3). A common one is \( A = -1.5; T = 1.5; C = 0.5; G = -0.5 \). This is used, for example, in [13]. This mapping has the nice property that it is easy to calculate the sequence on the opposite strand: just multiply by -1 and reverse the sequence.

### 4.2 Discrete Fourier Transform

The discrete Fourier transform (DFT) of a sequence \( X[n] \) of length \( N \) is defined as:

\[
\hat{X}[k] = \sum_{n=0}^{N-1} X[n] e^{-j\frac{2\pi kn}{N}} \tag{1}
\]
It is known to be useful in finding periodicities, so this was its first application in genomics. In [63], Trifonov found periodicities of 3, 10.5, 200, and 400 bases. He explained the 10.5 periodicity based on the need for the DNA to deform and fold in the nucleus and based on the coiled structure of some of the proteins coded for by the DNA. The 200 and 400 base periodicities were explained by the segmented organization of the genome. The 3 periodicity was found only in protein coding regions (exons), which led to the use of the DFT in gene finding.

The reason for the 3 periodicity in protein coding regions is that the identity of the third base of a codon matters less than the identities of the other two bases, due to the way the genetic code is constructed. This generates a 3-periodicity in the DFT, making the value of the DFT at $\frac{N}{3}$ particularly useful for analyzing DNA sequences. The DFT produces a complex number that has a magnitude and a phase ($r$ and $\theta$ in polar coordinates). Both the magnitude and phase have been used to distinguish between protein coding and non-protein coding regions of the genome [14, 31, 64]. High magnitudes at $\frac{N}{3}$ signify coding regions.

The phase value is used in [31] by Kotlar as part of his Spectral Rotation Measure. This measure relies on the fact that histograms of phase values computed for a sliding window on a region of a genome look considerably different depending on whether the region is protein coding or not. Figure 3 shows an example histogram for a coding region of the human genome; Figure 4 shows a histogram for a non-coding region. These phase values were calculated using a 240 bp sliding window on the DNA sequence, sliding 3 bp between calculations.
Figure 3: Histogram of phase values computed with a sliding window on a sequence from a coding region from human chromosome 5.

Figure 4: Histogram of phase values computed with a sliding window on a sequence from a non-coding region from human chromosome 3.
Kotlar’s spectral rotation measure is given by:

\[ |V|^2 = \left| \frac{e^{-i\mu_A}}{\sigma_A} A(s) + \frac{e^{-i\mu_T}}{\sigma_T} T(s) + \frac{e^{-i\mu_C}}{\sigma_C} C(s) + \frac{e^{-i\mu_G}}{\sigma_G} G(s) \right|^2 \]  

(2)

where \( A(s), T(s), C(s), \) and \( G(s) \) are complex numbers representing the values of the DFT at frequency one-third for the Voss representation of DNA sequence \( s; \mu_A, \mu_T, \mu_C, \) and \( \mu_G \) are the approximated average phase values for coding regions, and \( \sigma_A, \sigma_T, \sigma_C, \) and \( \sigma_G \) are the standard deviations of the phases for coding regions. The \( \mu \) and \( \sigma \) values are species specific. This measure has higher value for coding regions than for non-coding regions, because it selects out the parts of \( A(s), T(s), C(s), \) and \( G(s) \) pointing in the direction of the peak value of the histogram similar to that shown in Figure 3. Kotlar also defines a G Rotation Measure based only on the binary sequence defined by G bases:

\[ |V_G|^2 = |e^{-i\tilde{\mu}} G(s) + |G(s)||^2 \]  

(3)

where \( \tilde{\mu} \) is the value of \( \{\mu, \mu + \frac{2\pi}{3}, \mu - \frac{2\pi}{3}\} \) which is maximal (an adjustment for reading frame). Kotlar finds that both the Spectral Rotation Measure and the G Rotation Measure are effective for finding coding regions in yeast, and, in fact, perform similarly.

In addition to being useful for detecting coding regions, the phase histogram gives information about which reading frame is being used (which is why Kotlar needs to make the \( \tilde{\mu} \) adjustment). An insertion or deletion (i.e., a shift in reading frame) in exons in coding regions shifts the reading frame by \(-\frac{2\pi}{3}\) and \(\frac{2\pi}{3}\) respectively. Figure 5 shows the impact of deleting one base from the middle of the sequence generating the histogram in Figure 3. The two reading frames in the sequence are represented as two groups in the histogram shifted \(\frac{2\pi}{3}\) from each other. Similarly, Figure 6 shows the three reading frames created when a second base is deleted.
Figure 5: Histogram of phase values computed with a sliding window on the sequence from Figure 3 with one base deleted creating two reading frames.

Figure 6: Histogram of phase values computed with a sliding window on the sequence from Figure 3 with two bases deleted creating three reading frames.
Fourier magnitude and phase values at frequency one-third are quick and easy to calculate. In [40], the following formula is derived from the position count functions, $C_i$, where $i \in 1, 2, 3$. The value of $C_i$ is the number of ones in the $i$th position of each group of three scanning across the sequence. The Fourier phase value at frequency $1/3$ is:

$$\arctan\left(\frac{\sqrt{3}(C_2 - C_1)}{2C_0 - C_1 - C_2}\right).$$  \hspace{1cm} (4)

The Fourier magnitude value is:

$$\frac{1}{2} \left[(C_0 - C_1)^2 + (C_1 - C_2)^2 + (C_2 - C_0)^2\right].$$ \hspace{1cm} (5)

### 4.3 Autoregressive Models

Autoregressive models are also used for analysis of genomic sequences [13]. They have the advantage over the DFT that they work with smaller window sizes and, thus, shorter sequences. The idea of a forward predictive autoregressive model is that, given a number $p$ of previous values in a sequence $x$, the value of $x(n)$ can be predicted using:

$$x(n) = \sum_{k=1}^{p} a_k x(n - k) - e(n)$$ \hspace{1cm} (6)

where $a_1, a_2, \ldots, a_p$ are prediction coefficients and $e(n)$ is the prediction error. Likewise, a backward predictive AR model predicts the value of $x(n)$ based on following values in the sequence. The prediction coefficients are calculated by minimizing the mean squared forward prediction error by solving the Yule-Walker equations or by using the Burg Method.

The Yule-Walker equations are:

$$\sum_{k=1}^{p} a_k r_{xx}(i - k) = r_{xx}(i), \ i = 1, 2, \ldots, p$$ \hspace{1cm} (7)
where $r_{xx}$ is the autocorrelation function. Since the sequences are not infinite, an estimator must be used for the autocorrelation. [13] uses:

$$\hat{r}_{xx}(i) = \frac{1}{N} \sum_{k=0}^{N-|i|-1} x(k + |i|)x(k)$$  \hspace{1cm} (8)

where $N$ is the length of the window.

The Burg Method is based on the Levinson-Durbin recursion algorithm for solving the Yule-Walker equations. In order to get a more stable solution, Burg’s Method minimizes not just the forward prediction error, but the sum of the forward and backward prediction errors.

Once the prediction coefficients have been calculated, they can be used in various ways. One way is for comparing sequences. A model for one sequence (for example, a gene) can be computed and then used to calculate the error for another sequence (one being tested to see if it’s a gene) using that model. This gives a measure of “goodness of fit.” In [13] it was found that the “goodness of fit” test did not work well for gene prediction and that it was highly specific for particular genes, especially as the model order increased. An alternative is to use the prediction coefficients as features of the sequence. This is more useful. In [13] it produced good results for gene finding, looking for repeated sequences, and identifying sequences with similar chemical structures.

Autoregressive models were also used for finding tandem repeats (short sequences repeated many times in a row) in [76] and for classifying HIV-1 subtypes in [74]. In [76] peaks in the power spectral density function $P(\omega)$ calculated from the prediction coefficients for the model using:

$$P(\omega) = \frac{\sigma^2}{|1 + \sum_{k=1}^{p} a_k \exp(-j\omega k)|^2}$$  \hspace{1cm} (9)
indicated period $m$ repeats, where the peak value $\omega = \frac{2\pi}{m}$. In [74] a neural net trained on prediction coefficients was used for classification.

### 4.4 Entropy

A method from information theory that has been useful in gene finding is Shannon entropy. This method utilizes the fact that in protein coding regions not all codons (groups of three nucleotides coding for an amino acid) are used uniformly, while in non-protein coding regions they are. In [9] Bernaola-Galvan et al. compute entropy of a sequence using a 12-symbol alphabet $\{A_0, A_1, A_2, C_0, C_1, C_2, G_0, G_1, G_2, T_0, T_1, T_2\}$. The letters A, C, G, and T represent the four possible bases; the subscripts represent their position in the sequence mod 3, i.e. their position in their codon. Other alphabets are possible: the 4-symbol alphabet of bases, the 16-symbol alphabet of dinucleotides (pairs of bases occurring in a row), the 64-symbol alphabet of trinucleotides. For each sequence, they computed the frequency vector $F = \{f_1, \ldots, f_12\}$ for the 12 symbols. Shannon entropy $H(F)$ was calculated using the formula:

$$H(F) = -\sum_j f_j \log_2 f_j$$

To compare two sequences, they calculate the Jensen-Shannon divergence $C(F_1, F_2)$ using the frequency vectors ($F_1$ for the first sequence, $F_2$ for the second sequence, and $F$ for the concatenated sequence), lengths ($n_1$ for the first sequence, $n_2$ for the second sequence and $N$ for the concatenated sequence), and the formula:

$$C(F_1, F_2) = 2 \ln 2[NH(F) - n_1H(F_1) - n_2H(F_2)].$$

Two sequences are considered different if their Jensen-Shannon divergence is greater than that of two random sequences.
This method was used to find the boundaries between protein coding and non-protein coding regions in the following way. A sliding pointer was moved through the genome and the Jensen-Shannon divergence was calculated for the sequences on either side of it. The point that had the maximum divergence was calculated. If that divergence was greater than that for random sequences, a cut was made. Then, the procedure was repeated on the subsequences created. This resulted in a segmentation of the genome that was a good match for coding/non-coding boundaries.

5 Conclusion

The problem of detecting endogenous retroviruses in genomes is both non-trivial and important. ERVs contribute to mutation, development, and disease, and they play a crucial role in the evolution of genome structure. They serve as a source of new genes and as molecular fossils which researchers can use to better understand the connections between species and the process of speciation. Some work has already been done to detect and annotate ERVs in humans and other species, but much remains to be done. Some of the early genomes to be sequenced, including Drosophila (fruit fly), S. cerevisiae (yeast), and Arabidopsis (the first plant to be sequenced), have good annotations. These can be used as training examples for machine learning tools, such as support vector machines, random forest, and naive Bayes classifiers, which can then be used to detect new retroviruses in these genomes or to annotate retroviruses in genomes of other organisms.

Signal processing techniques have been useful in solving other problems in genomics, such as gene finding. They have not yet been applied to the problem of ERV detection.
Unique properties of ERVs, such as their overlapping reading frames and other structural features and their choice of which codons to use to code for particular amino acids, could be extracted with the signal processing techniques discussed in this paper and used as input for the machine learning tools to classify sequences as retroviral or not and to scan for retroviruses in genomes.

References


