IN SILICO RETRIEVAL AND CATALOGING OF GENBANK DNA SEQUENCES ADJACENT TO INTERSPERSED REPETITIVE ELEMENTS

Author:
Steven Steinway

Faculty Sponsor:
Sudhir Nayak
Department of Biology

ABSTRACT

Repetitive elements are often undrepresented among drafts of genomic sequencing projects, especially during the early stages of such projects. The assembly of clones containing repetitive DNAs poses some technical problems and the sequencing of clones containing these DNAs is generally considered low priority despite the fact that they constitute 50-80% of the DNA in many eukaryotes. A comprehensive understanding of the architecture of genomes requires a detailed knowledge of the organization of these forsaken DNAs. A good example of this paucity of repetitive DNA representation is the soybean genome, where one family of long terminal repeat (LTR) retrovirus-like retroelements, the *SIRE*1 family, is virtually absent from Genbank, despite the fact that its members alone constitute as much as 1% of the soybean genome. We have found, however, that the *SIRE*1 family is well represented among BAC-end sequences deposited in the Genbank (GSS) database, with over 1800 independent entries. The repetitive nature of elements like *SIRE*1 makes characterizing their insertions amenable to *in silico* search strategies. Using *SIRE*1 as our model, we have developed a computational method that searches all appropriate databases to collect and evaluate DNAs that are adjacent to individual members of transposable element families. This allows us to characterize the immediate neighborhoods into which these repetitive elements are found and by extension, a major portion of eukaryotic genomes that are in various stages of sequence assembly. In our study, 998 unique entries were determined to flank SIRE1 in *Glyicine Max,* and 179 unique flanking sequences were given preliminary annotation. This iterative search strategy can be applied to virtually any moderately repetitive transposable element family whose members are fairly well conserved and can be used to search databases containing incomplete and complete drafts of eukaryotic genomes.

INTRODUCTION

Repetitive DNAs comprise a considerable part of most eukaryotic chromosomes. They were first discovered in the early 1990s, and currently there are several known classes of these DNAs. Nearly 50% of the human genome and the majority of many plant genomes are repetitive DNAs (Bromham, 2002). Repetitive DNAs can be subdivided into two classes: tandem repeat and interspersed repeat DNAs, also known as transposable elements (TEs). It has been proposed that TEs are major contributors to the evolution of virtually all species, driving major evolutionary changes (Bowen and Jordan, 2002). TEs can be subdivided into Class I (retrotransposons) and Class II (DNA transposons) elements. Class I elements can further be subdivided based on autonomy of replication. The distribution of TEs in eukaryotic genomes is quite varied; some have shown apparent random distributions while others have been shown to be clustered at specific sites (Miyao, 2003). Despite their significant presence in eukaryotic genomes, TEs are virtually absent from genome projects, even in those species with extensive whole genome coverage (Bromham, 2002). However, unpublished databases like the Genbank Genome Survey Sequence (GSS) Database have been shown to be rich sources of these elements. Preliminary work has shown that a database search strategy can be employed to identify and annotate these elements; this analysis has also shown preferences in target site selection. However, previous studies were tedious, riddled with human error, and extremely slow (Laten, unpublished).

Tandemly repeated DNAs include satellite DNAs, which consist of consecutively repeated short DNA sequences. Satellite classification can be broken down into more specific classifications of micro- and mini-satellites based on lengths of repeating unit. These elements are generally found in the heterochromatic and telomeric regions, and are used extensively in forensic analysis, genotyping, and in the creation of genetic maps. Although their purpose has yet to be definitively determined, satellite DNAs have been identified in regulatory roles, and variations in repeat length have been linked to genetic repeat diseases such as Huntington's disease (Legendre and Verstrepen, 2007).

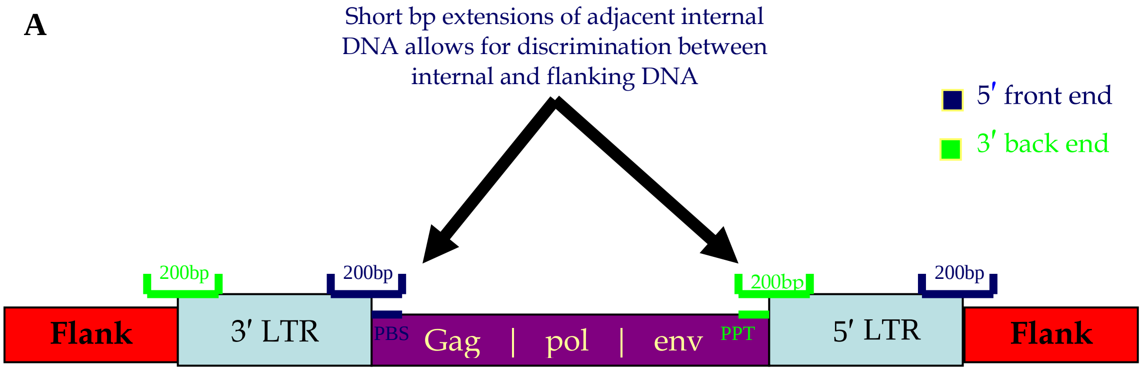
Interspersed repetitive elements are usually mobile transposable elements dispersed throughout the genome. They can be subdivided into two families distinguished by their respective mechanisms of transposition. Class I elements, collectively known as retroelements, transpose through a reverse transcribed RNA intermediate. These include autonomous elements: LTR retrotransposons, endogenous retroviruses, and long interspersed nucleotide elements (LINEs). Individually, retroelements are composed of a very small number of genes that sponsor their own proliferation within the cells of their hosts using just one to three structural proteins and enzymes, including a reverse transcriptase. Collectively they make up much of eukaryotic chromosomes (Kumar and Bennetzen, 1999).

Nonautonomous small interspersed nucleotide elements (SINEs) use the products of other TEs for their copying and integration. The copy numbers of autonomous element families can range from a few to several hundred thousand (Boeke and Stoye, 1997). The class II elements transpose by a cut and paste mechanism mediated by an enzyme called transposase. These elements are generally present at lower copy numbers, but some can reach copy numbers equal to LTR retroelements (Feschotte and Jiang, 2002).

Distributions of TEs throughout the genomes of eukaryotes are as varied as the elements themselves. The locations of their insertions can have major impacts on the genomes TEs inhabit. Many are clustered around centromeres in regions where there are few if any genes, while others are found scattered along the lengths of most chromosomes, interspersed with coding regions (Miyao, 2003). Despite the apparent randomness in the distribution of individual members of noncentromeric families, unrelated elements are often clustered, as reported in maize and wheat (SanMiguel, 2002). On a larger scale, the genomes of Arabidopsis and rice have been surveyed for TE distribution and both centromeric and noncentromeric insertions are present, with large clusters near the former. However, in rice, new insertions of the Tos17 retrotransposon are spatially isolated and adjacent to functional genes (Miyao, 2003). It has been suggested that evolutionary forces have led to the targeting of high copy number elements to already silenced repetitive DNA regions, while low copy number families have escaped this scrutiny (Bennetzen, 2000). Thus, an analysis of retroelement insertions can provide valuable information about the extent to which these TEs clustered and the neighborhoods in which they exist in the genome. Knowledge about the nature of target site similarities and/or differences provides important insights into the mechanism of target site selection and the extent to which particular retroelement families may influence economically important traits and contribute to far reaching evolutionary changes.

Retroelements have also been major contributors to the evolution of virtually all species. These elements promote DNA rearrangements and changes in gene regulation that are proposed to have sponsored major evolutionary leaps throughout the history of life (Bowen and Jordan, 2002). In plants, these elements are the source of major expansions of genome size on time scales that can be measured in units of generations as well as in millions of years (Bennetzen, 2002). With the exception of rice, the genome sequencing projects of important agricultural crop plants have lagged far behind those of model organisms. For example, the soybean genome sequencing project is largely incomplete, while the model legume genome projects that are considerably farther along are distantly related to major bean crops (Swaminathan and Hudson, 2007).

Even for species with extensive whole genome coverage (e.g., humans) the regions that are dense in TEs, including retroelements, are often absent from advanced drafts (Bromham, 2002). The scarcity of genomic DNA sequences from sequencing projects makes Soybean Interspersed Repetitive Element 1 (SIRE1) sequences in published DNA databases such as the Genbank Nonredundant Nucleotide Database nearly nonexistent (Benson and Wheeler, 2007), despite the fact that they constitute 1% of the soybean genome (Laten and Gaucher, 1998). On the other hand, the Genbank Genome Survey Sequence (GSS) database is a rich source of these entries. The GSS database contains short (< 1kb), unpublished single pass read sequences from ends of genomic clones consisting of large DNA sequences.


A
Short bp extensions of adjacent internal DNA allows for discrimination between internal and flanking DNA
■ 5′ front end
🟩 3′ back end
200bp
200bp
200bp
200bp
Flank
3′ LTR
PBS
Gag | pol | env
PPT
5′ LTR
Flank

B

QS$_{5'}$: *5′-3′*

ccaaaggggagat | tgttagtgcttagcactactgagtttaaaaaggttggctaagattttgttaaaacataagcacttagacaatgaa ggaaagctggagttgctgcacatgatgtccaacgttatgtcaaggaataagatcgggctgcataatgcacaaggcaagataaagtgtca agtgatgaattgaagttgaagg

QS$_{3'}$: *3′-5′ (reverse complement)*

gctctgataccaat | tgaaattctgataccaggggacagatgtcgtacaggatgtcacgacatcacgcttcagaacatgcagtttatgtgt gtccgtatgaacagattaaacaagtaaataacacaagagaattgtttacccagttcggtgcaacctcacctacatctgggggctaccaag ccagggaggaaatccactct

Figure 1. (A) *SIRE-1* LTR Retrotransposon has genes coding for an envelope, reverse transcriptase, and viral coat. The exact same DNA sequence at 5′ and 3′ LTR requires our program to check to ensure that DNA flanking the outsides of the LTRs and not the ends of the LTRs adjacent to the coding regions are selected. (B) Sequences used to query GSS database for possible SIRE1 members. " | " denotes the break between the primer-binding site (PBS) and LTR sequence in QS5 and between polypurine tract (PPT) and LTR sequence in QS3. These sequences allowed for distinction between external and internal regions of the LTR.

This study employed a database search strategy to collect, annotate, and evaluate DNA sequences in Genbank databases of published and unpublished DNA sequences that are adjacent to the hundreds of copies of retroelements in soybean and other legume genomes. Since legume genome projects, especially the soybean initiative, are in the early phase of development, only a handful of retroelements' sequences have been identified in these projects. Preliminary work with the SIRE1 retroelement in soybean, a relatively young TE that was active as recent as 30,000 years ago (Laten and Morris, 1993), showed that 90% of SIRE1 members inserted themselves directly into other repetitive DNAs (Laten, unpublished). Thus, an analysis of retroelement insertions can provide valuable information about the extent to which these TEs are clustered and the neighborhoods in which they exist in the genome. However, the methods used to collect and identify these neighbors require a very labor-intensive and iterative approach.

As in Laten's previous study, the DNA sequences at the ends of SIRE1 have been used as a model for a streamlined computational method (Figure 1A). This strategy can be applied to virtually any moderately repetitive TE family whose members are somewhat conserved and can be used to search databases containing incomplete and complete drafts of eukaryotic genomes. This allowed for characterization of immediate neighborhoods in which TEs are found and by extension, major portions of eukaryotic genomes that are in various stages of sequence assembly.

This analysis provides valuable information about the nature of retroelement insertion sites. Knowledge about the nature of target site similarities and/or differences provides important insights into the mechanism of target site selection and the extent to which particular retroelement families may influence economically important traits and contribute to far reaching evolutionary changes. Analyses of retroelement insertions can provide valuable information about the extent to which these TEs are clustered and the neighborhoods in which they exist in the genome. Results showed the accuracy and efficiency of this approach to be far superior to previous "manual" search strategies.

MATERIALS AND METHODS

A four-step search strategy was implemented to identify and annotate DNAs. Step one involved a local retrieval of GSS database entries from NCBI that contain identifiable flanking sequences. Next, the flanking sequences were obtained from Genbank, and then duplicate sequences were eliminated. The final step was annotation of flanking sequences.

The National Center for Biological Information (NCBI) maintains a group of databases collectively known as Genbank for retrieval of information in their numerous online databases consisting of biological data as well as other scientific resources. These databases store published and unpublished sequence entries for DNA, RNA, and protein, of a vast number of organisms. Genbank has implemented a search algorithm called Basic Local Alignment Search Tool (BLAST) which allows for search and retrieval of sequences. A BLAST search requires a query sequence and returns similar sequences (hits) that share sequence similarity above a threshold defined by parameters.

An interfaced implementation of BLAST is accessible via the internet; however, each search can take anywhere from two to three minutes based on query and site traffic. BLAST also has been packaged with Genbank databases as *blastall* for download and use as a command-line program. The latter version was implemented for this study.

As a model, the *SIRE1* LTR retrotransposon was used as a query, and DNA submissions (from the GSS database) for the Soy Bean, *Glycine max,* were retrieved. The presented software was written in Java, an object-oriented programming language. *Blastall* was implemented through the java.lang's Runtime, which allowed for access of the NCBI's BLAST and FASTA algorithms. A flow-chart of work-flow is shown in Figure 2.

Step 1 – Local retrieval of all GSS database entries from NCBI that contain identifiable flanking sequences

BLASTn searches using the 5′ and 3′ ends of the Long Terminal Repeat (LTR) region of the SIRE1 DNA sequence were run via the locally downloaded GSS database, and the results were written to an XML file. For each hit, BLAST output was parsed for alignment and flanking sequence information. It was next determined whether the LTR returned was internal or flanking DNA (Figures 1 and 3). Genbank Identification (GI) objects, were created to hold the Genbank ID number of a hit, orientation of query-hit alignment, and all calculated values specific to a hit sequence For each hit meeting specified criteria, a GI object was created and stored in a list and text file. 1565 sequences were collected for SIRE1.

Local search parameters were used that matched those of the previous manual search strategy: -v (number of hits to return) = 5000, -a (number of processors used) = 8, -r (reward for a match) = +2, -F (use of filters, false) = F, -e (expected value) = 10^{-5} –m (specifies output format, XML) = 7. A client-side (optional) specification of organism was implemented using "*Glycine max*," which limited the returned sequences to a specific organism (soy bean).

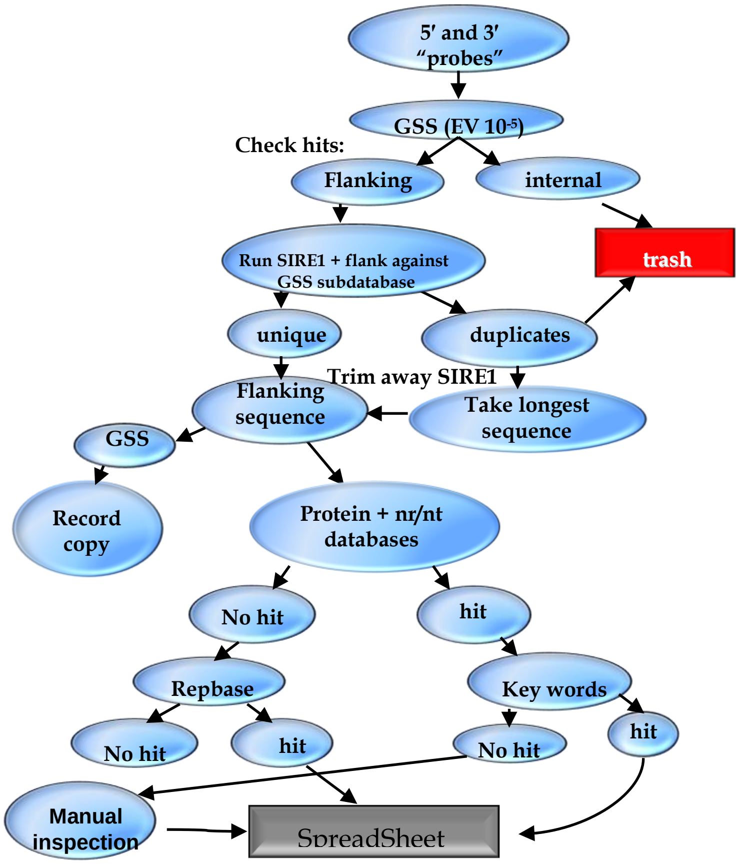


Figure 2. Flow of Program. 200 bp sequences of 5′ and 3′ Long Terminal Repeat (LTR) regions of SIRE1 were used to query the Genome Survey Sequence (GSS) Database. Returned database entries were those that met search-defined parameters. Since Long Terminal Repeat regions are identical (at time of insertion) at 5′ and 3′ ends of SIRE1, hit sequences were checked for external (flanking) or internal position. External sequences were kept and duplicate sequences were eliminated. SIRE1 sequence adjacent to flanking DNA was removed and flanking sequences were queried against Genbank databases. Hits from these searches were run in a keyword search. If a match occurred, that sequence was entered into a spreadsheet. Copy number in the GSS database was also recorded. Sequences that returned no hits were flagged for manual inspection.

For identification of whether the aligned sequence was flanking or internal, <Hsp_query-from> (QF) and <Hsp_query-to> (QT) tags were parsed for each sequence returned. These tags allowed designation of where the alignment began and ended. One of these values had to be between 11 and 17. That is, sequences returned had to start between the 11th to 17th nucleotide of this sequence (Figures 3 and 4). To be sure of the orientation of the query sequence (QS), the QF value was subtracted from the QT value, to determine the direction of the sequence. Since a query sequence will never be only one nucleotide long, QT ≠ QF. A positive value was identified as "+" orientation and a minus value as a "-" orientation.

Within the returned hits, the orientation of each hit strand was determined by subtracting the values in the <Hsp_hit-to> (HT) and <Hsp_hit-from> (HF) tags. If the HT was greater than HF, then a "+" value was assigned to the hit, and if HT was less than HF), then "-" orientation was assigned. Sequence flanking the alignment of each hit was checked to be longer than 20bp. Flanking sequence length was identified for each hit that met the minimum length specification. A query/hit alignment could be identified by one of four orientations: +/+, +/-, -/+, and -/-. The method for calculating a sequence's Flanking Length (FL) differed for each of the four query/hit orientation cases identified. Flanking length will be used later in the program in Step 2.

Each sequence submitted in a database is given a unique Genbank Identification (GI) Number. GI numbers were retrieved from the <Hit_id> tag of each hit. Each GI number was stored in a .txt file for use in step two.

Step 2 – Retrieve the flanking sequences from Genbank

Local FASTA searches using each GI in our list from Step 1 were run to retrieve corresponding Genbank sequences using NCBI's *fastacmd* (Pearson and Lipman, 1988). The Flanking Sequence (FS) was extracted for each GI number in the list. Since FSs end at the 14th nucleotide of a hit's corresponding QS, but hits may start anywhere between the 11th and 17th nucleotides, the values were offset accordingly to retrieve only the FS by their query/hit orientation.

The sequences were extracted in the range of each FS. A substring was made containing a 70 bp sequence from the flanking region for identification and elimination of duplicate sequences in Step 3.

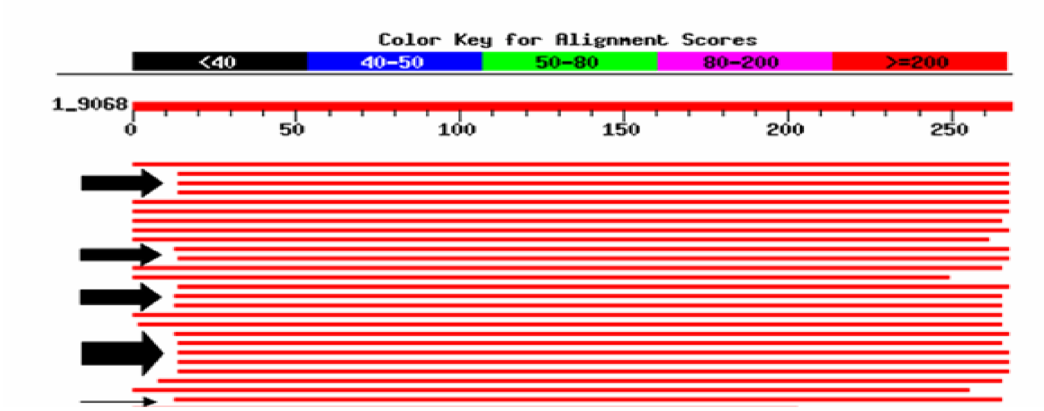


Figure 3. Graphical output from BLASTn search with SIRE1 LTR (5') query. Arrows indicate sequences of interest with flanking DNA. These sequences start between 11th and 17th nucleotides.

Step 3 – Eliminate duplicates of collected GSS sequences

Because it was possible that the same DNA copy existed on multiple clones in the GSS database, each retrieved sequence (~1500) was compared to every other sequence by creating a sub-database using the 1500 sequences and doing a local BLASTn search of the database using a 70 nt sequence of each member of the database as a query. The sub-database was created using *blastall.* Since not many duplicates were anticipated, this search was limited to 20 hits (if a sequence produced 20 legitimate duplicates, the limit was increased to 50).

The query contained 20 nt of the SIRE LTR and 50 nt of the flanking sequence, and up to three mismatches (95% identity) were allowed in a query-hit alignment to mark a sequence as a duplicate. For each query made against the created sub-database, a list was made containing the query and all its hits, thus producing a list of duplicates for a single query. A sequence similarity as low as 95% was allowed because of the possibility of a sequencing or data interpretation error by the submitting party. Lists containing at least one sequence shared between any two lists were merged (Figure 4); thus, the definition of a duplicate became *any* sequence that had 95% identity with any other. After merging these lists, the duplicate in each list that was longest was added to a unique list of flanking sequences.

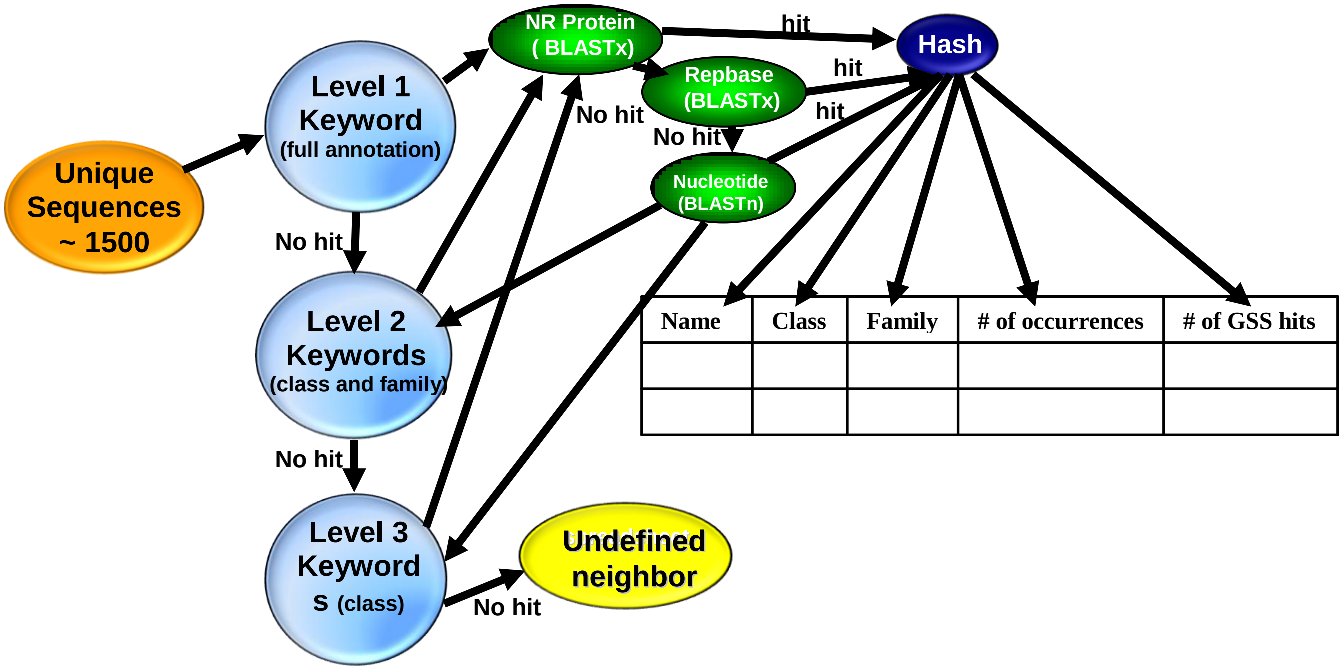


Figure 4. Program flow for flanking sequence annotation. Each unique sequence is first queried using BLASTx of the Protein Database and hit annotation is compared against Level 1 key words. If a match is found, that sequence is "annotated," no more searches will be done, and it is stored in the Hashtable. If a Level 2 or Level 3 Keyword is matched, those values are stored temporarily. Subsequent searches are done using BLASTn queries and hits. If no Level 1 match occurs, the highest annotation level from the first searched database will be stored. If no annotation is found, the sequence is defined as an "Undefined Neighbor." Repbase was not implemented in this search strategy.

Step 4 – Flanking sequence annotation

Sequence entries in the GSS database have essentially no annotation; therefore, to determine the identity of flanking sequences of SIRE1, unique flanking sequences were queried against two local databases. A flow diagram of the search strategy is in Figure 2. BLASTx searches of the nonredundant Protein Database and BLASTn searches of the nonredundant nucleotide database were queried locally. <Hit_def> tags were parsed and database annotation was run against a keyword list. When a keyword was found, it triggered storage of specific values (Figure 4).

When a keyword was found, it was stored in a Hashtable, a data structure that uses a key (the specific keyword) to store and access data. There are 4 classes: I, II, Satellite, Repeats. In Class I, there are 3 families: Gypsy, Copia, and other. In Class II, there is 1 family: Transposon. In class Satellite, there is one family: Satellite. In class Repeats, there is 1 family: Repeats. Each family within a class has a list of names associated with it, and each name has a list of GI numbers associated with it. If a specific name matched a keyword, then Level 1 or full annotation was designated. If only a family was matched with a keyword, then Level 2 annotation was given. If only a Class matched a keyword, then that sequence was given Level 3 annotation. If a sequence is given Level 1 annotation it is removed from the list, and no more searches are done with it. The number of GIs per name was recorded, and each GI was run through the GSS database. The number of hits it returned was recorded (a hit is a sequence with error <Hsp_evalue> values less than 10^{-5}). Program flow of keyword annotation is shown in Figure 4.

RESULTS

Local GSS retrieval

Program output was compared to "manual" output at various instances throughout development. In the initial query of the local GSS database with 2 queries, 1,572 hits were returned. In the manual search, 1,467 hits were returned. Comparison between this automated approach and manual approach showed that 77 manual hits were not included in the automated approach, and that 14 hits from the automated query were not included in the results of the manual approach.

Flanking sequence retrieval

After SIRE1 DNA adjacent to flanking sequences was removed, online BLAST searches were run with 10 random GIs to check that calculations for isolating the flanking sequence were implemented properly and by query/hit orientation. It was confirmed that flanking sequences were isolated from SIRE1 properly. No cases of +/+ or +/- orientation were seen throughout development.

Elimination of duplicate sequences from the GSS Database

Because the GSS database contains unpublished sequences, the same sequence could have been submitted many times into the database. After elimination of duplicates, 998 unique sequences were determined to flank SIRE1 in *Glycine max*. Manual search found 812 unique junctions. Comparison of these lists and subsequent spot searches by comparison with online BLAST searches revealed that erroneous and inconsistent duplicate designation did occur in the manual approach.

Table 1. Annotated sequences flanking SIRE1 in *Glycine max*. Keyword represents the keyword in the Keyword List that matched the Hit_Def tag in the XML output of the BLASTx search of the nr-protein database.

Name	**Class**	**Family**	**Keyword**	**Level**	**Copy Number**
athila	I	Ty3/Gypsy	athila	1	14
cacta	II	DNA transposon	cacta	1	11
cinful	I	Ty3/Gypsy	cinful	1	13
hopscotch	I	Ty1/Copia	hopscotch	1	3
mudr	II	DNA transposon	mudr	1	1
Opie	I	Ty1/Copia	opie	1	28
ram	I	Ty3/Gypsy	ram	1	9
retrofit	I	Ty1/Copia	retrofit	1	1
rire	I	Ty3/Gypsy	rire	1	9
SINE-like	I	Non-LTR	SINE	1	13
SIRE	I	Ty1/Copia	SIRE	1	1
tgm	II	DNA transposon	tgm	1	1
Undefined	II	DNA transposon	transposon	2	7
Undefined	I	Ty1/Copia	copia	2	4
Undefined	I	Ty1/Copia	ty1	2	2
Undefined	I	Ty3/Gypsy	ty3	2	21
Undefined		Ty3/Gypsy	gypsy	2	1
Undefined	I	Undefined	gag	3	28
Undefined	I	Undefined	gag-pol	3	3
Undefined	I	Undefined	ltr	3	1
Undefined	I	Undefined	retroelement	3	1
Undefined	I	Undefined	retrotransposon	3	4
Undefined	I	Undefined	transposase	3	2

Flanking sequence annotation

The 998 unique sequences were run through BLASTx search of the NR Protein Database. Keyword search of the output led to annotation of 179 flanking sequences. Of these, 105 received Level 1 annotation, 25 received Level 2 annotation, and 47 received Level 3 annotation. Because of computational difficulties, searches of the Nucleotide and Repbase databases were not completed. A list of annotated sequences flanking SIRE1 retrotransposons in the *Glycine max* genome is shown in Table 1.

DISCUSSION

Eukaryotic genomes consist largely of repetitive DNA, and although mostly non-coding and often considered "junk" DNA, these regions are known to have important implications in whole genome architecture and regulation (Shapiro, 1999). TEs have played a critical role in gene evolution as well. TEs can lead to speciation through creation of new loci caused by insertions and "jumping" out of coding regions. (Brunner and Duncan, 1986). TEs have played direct roles in genome regulation as well. IS elements of *E. Coli* have been shown to transpose in response to environmental stresses. TEs also have important roles in chromatin structure, responsible largely for eukaryotic gene regulation (Hall, 1999). It is apparent from this data that although putative roles have thus far evaded identification, these DNAs play essential roles in genomic events.

Local GSS retrieval

Discrepancies between sequences returned by this approach and the manual approach were deemed insignificant. Parameterization through NCBI's local BLAST program, *blastall,* differed slightly from that of online BLAST protocol. Thus, search requirements may have varied slightly between the two approaches. The local databases are also updated at discrete times, whereas the online databases are updated continuously; thus, some sequences in the online database were not yet part of a local database update.

In terms of processing time, the average local BLAST of the GSS database took nine seconds using a 2.8 Ghz Intel Quad-Core processor. Manual searches took anywhere from two to three minutes. The 1500 initial BLAST searches done programmatically were completed in 3.75 hours. Manually, this would take 85 hours of continuous searching, not including user time spent between searches.

Flanking sequence retrieval

In each sequence there are two important regions, nucleotides 1-14 (known as the Polypurine Tract (PPT) and the Primer-binding site (PBS) region, which give polarity to the external regions of SIRE-1. Nucleotides 15-200 of our query were part of the LTR region. In a SIRE-1 element, the two LTR regions are *exactly* the same at the time of genomic insertion. For this reason, it was important that the PPT/PBS region was included in the query sequences. Sequences that aligned approximately from nucleotides 1 to 200 hit at the whole PPT/PBS-LTR sequence, which was the query. However, sequences that were required hit the QS only from ~15 to 200, excluding the PPT or PBS portion. Excluding the PPT/PBS portion guaranteed that the hit sequence *before* the corresponding ~15th nucleotide of the QS was in the flanking region (Figures 1 and 2). Hits that started anywhere from 11-17 nucleotides were included as a "flanking hit" as well. This was chance, because it was possible to have nucleotides 11-14 match the flanking part of the hit sequence, and mutational mismatches in nucleotides 15-16 that would make the hit start at a later or early QS nucleotide match.

Elimination of duplicate sequences from the GSS Database

Because it was possible that the same DNA copy exists on different clones, each retrieved sequence was compared with every other sequence by creating a sub-database using the 1500 sequences and doing a local BLASTn search of the database using a 70 nt sequence of each member of the database as a query. A duplicate was considered for *any* sequence that has 95%

identity with any other, thus leading to a minimum of 90% sequence similarity between any DNA deemed a duplicate. This was allowed because of the possibility of a sequencing or data interpretation error by the submitting party. Lists containing at least one sequence shared between the two were merged.

Each sequence should have found at least one duplicate (itself), when queried in the sub-database. Interestingly, two sequences (CZ500563.1 and CZ527290.1) found no duplicates. Analysis of these sequences showed that they were riddled with Ns. Ns are entered by chromatogram software when a nucleotide at a position cannot be detected unambiguously. These sequences were deemed invalid and removed from analysis. Discrepancy between duplicate identification with our protocol and the manual protocol showed that numerous duplicate identification errors occurred through the manual protocol because of human error.

Flanking sequence annotation

Output from the BLASTx keyword search confirmed that the computational method was more efficient and just as accurate as the previous attempt to annotate unknown sequences. Annotated sequences flanking SIRE1 retrotransposons in the *Glycine max* genome are shown in Table 1. Manual annotation attempted prior to this project over the course of a year identified ~600 of ~1500 flanking sequences (Laten, unpublished). This tool did the same work in less than five hours (depending on processor capabilities). This protocol is independent of the human error inherent in this kind of iterated, long term task. Future implementation will include annotation using BLASTn and Repbase data, and parameters will be user-specified to allow for analysis with other TEs. A networked computing cluster with distributed memory will be implemented as well as a distributed databases solution.

Repetitive genetic elements have proven difficult to sequence by conventional techniques which accounts for their absence in genome projects. Thus, these DNAs are greatly under-represented in genome projects (Swinathan and Hudson, 2007). This tool can be utilized for preliminary identification of repetitive DNAs and insertion preferences in a genome. Knowledge about the nature of target site similarities and/or differences provide important insights into the mechanism of target site selection and the extent to which particular TE families may influence economically important traits and contribute to far reaching evolutionary changes. Analyses of transposable element insertions can provide valuable information about the extent to which these TEs are clustered and the neighborhoods in which they exist in the genome. Implementation of this analysis provides preliminary annotation for repetitive DNAs and thus major constituents of eukaryotic genomes at various stages of sequencing projects.

ACKNOWLEDGEMENTS

I would like to thank Professor H. M. Laten and Professor Chandra N. Sekharan of Loyola University Chicago for their mentoring during the summer. I thank Professor Amanda Norvell and Professor Jeffery Erickson for their mentoring in BIO399: Research Internship and Professor Shelley and Professor Sudhir Nayak for mentoring the preparation of this paper. Lastly, I thank Aadeel Akhtar, the co-collaborator on this project.

REFERENCES

Bennetzen, J.L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology* 42:251-269.

Bi, Y.A., Laten, H.M. (2003). Sequence analysis of a cDNA containing the gag and prot regions of the soybean retrovirus-like element, SIRE-1. *Plant Molecular Biology* 20:1222-1230.

Feschotte, C. Ning, J., and Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics* 3:329-341.

Hall, B.G. (1999). Transposable elements as activators of cryptic genes in *E.Coli. Genetica* 107:181-187.

Hull, R. and Covey, S.N. (2005). Retroelements: Propagation and adaptation. *Virus Genes* 11:105-118.

Kumar, A. and Bennetzen, J.L. (1999). Plant Retrotransposons. *Annual Review of Genetics* 33:479-532.

Laten, H.M. and Gaucher, C. (1998). SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein. *Proceedings of the National Academy of the Sciences USA* 95:6897-6902.

Laten, H.M. and Morris, R.O. (1993). SIRE-1, a long interspersed repetitive DNA element from soybean with weak sequence similarity to retrotransposons: initial characterization and partial sequence. *Gene* 134, 153-159.

Miyao, A., Onosato, K., Takeda, S., Kiyomi, A., Yoriko, S., and Hirochika, H. (2003). Target Site Specificity of the *Tos17* Retrotransposon Shows a Preference for Insertion within Genes and against Insertion in Retrotransposon-Rich Regions of the Genome. *Plant Cell* 15:1771-1780.

Pearson, W.R. and Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of the Sciences USA* 85(8): 2444 – 2448.

Rostoks, N., Parks, Y.J., Ramakrishna, W., Ma, J., Druka, A., Shiloff, B.A., and SanMiguel, P.J. (2002). Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. *Functional Integrative Genomics* 2:70-80.

SanMiguel, P.J. and Bennetzen, J.L. (1998) Grass Genomes. *Proceedings of the National Academy of the Sciences USA* 95:5.

Shapiro, J.A. (1999). Transposable elements as a key to a 21st century view of evolution. *Genetica* 107:171-179.

Swaminathan, K., Varala, K., and Hudson, M.E. (2007). Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics* 8:132.

Yano S.T., Panbehi B, Das A., Laten, H.M. (2004) *Diaspora,* A large family of Ty3-gypsy retrotransposons in Glycine max is an envelope-less member of endogenous plant retrovirus lineage. *BMC Evolutionary Biology* 5:30.