BIOINFORMATICS

Vol. 19 Suppl. 2 2003, pages ii73–ii80
DOI: 10.1093/bioinformatics/btg1063

A rapid method for detection of putative RNAi target genes in genomic data

Yair Horesh[1], Amihood Amir[1], Shulamit Michaeli[2] and Ron Unger[2,*]

[1]Department of Computer Science and [2]Faculty of Life Sciences Bar-Ilan University, Ramat-Gan, 52900, Israel

Received on March 17, 2003; accepted on June 9, 2003

ABSTRACT

RNAi, inhibition of gene expression by double stranded RNA molecules, has rapidly become a powerful laboratory technique to study gene function. The effectiveness of the procedure raised the question of whether this laboratory technique may actually mimic a natural cellular control mechanism that works on similar principles. Indeed recent evidence is accumulating to suggest that RNAi is a natural control mechanism that might even serve as a primitive immune response against RNA viruses and retroposons. Three different interference scenarios seem to be utilized by various RNAi mechanisms. One of the mechanisms involves degradation of mRNA molecules. Here we suggest a method to systematically scan entire genomes simultaneously for RNAi elements and the presence of cellular genes that are degraded by these RNAi elements via exact short base-pair matching. The method is based on scanning the genomes using a suffix tree data structure that was specifically modified to identify sets of combinations of repeated and inverted repeated sequences of 20 bp or more. Initial scan suggest that a large number, about 7% of *C.elegans* and 3% of *C.briggsae* genes, have the potential to be subject to natural RNAi control. Two methods are proposed to further analyze these genes to select the cases that are more likely to be actual cases of RNAi control. One method involves looking for ESTs that can provide direct evidence that RNAi control element are indeed expressed. The other method looks for synteny between *C.elegans* and *C.briggsae* assuming that genes that might be under RNAi control in both organisms are more likely to be biological significant. Taken together, supportive evidence was found for about 70 genes to be under RNAi control. Among these genes are: transposase, hormone receptors, homeobox proteins, defensin, actins, and several types of collagens. While our method is not capable of detecting all cases of natural RNAi control, it points to a large number of potential cases that can be further verified by experimental work.

Key words: RNAi, Suffix tree, *C.elegans*, *C.briggsae*, control mechanism

Contact: ron@biocom1.ls.biu.ac.il

INTRODUCTION

A lot of attention was drawn recently to RNA molecules that have the ability to control expression of genes. Recently, the journal Science (Couzin, 2002) awarded these molecules as the ‘Molecule of the Year’ for the year 2002. The phenomenon, known as RNAi, (RNA interference) was first discovered (Fire *et al.*, 1998) in *C.elegans*, where it was shown that introducing double stranded RNA (dsRNA) molecules can interfere with specific mRNA (the messenger RNA molecules that carry the template for protein production) that contain homologous sequence and thus block the expression of the corresponding gene. Here we suggest a computational method to detect such potential RNAi control elements and the genes controlled by them.

The ability to inhibit the expression of any particular gene, simply by introducing a short RNAi molecule, greatly facilitates studies of gene function in adult tissues. Indeed admisitration of synthetic RNAi to mammalian cells caused the specific degradation of the specific mRNAs (Elbashir *et al.*, 2001). Unlike knock-out experiments, where a missing gene is taken out of the genome and can not function in critical developmental stages of the organism, a gene under RNAi control can be shut off at will, especially if the system to synthesize dsRNA is under inducible promoter. In many organisms, RNAi became the method of choice for massive genome analysis. These experiments are especially convenient in *C.elegans* where the worms can be fed with bacteria engineered to silence the gene of interest (Maeda *et al.*, 2001). RNAi was also shown to be useful tool in studying genes in plants (Wesley *et al.*, 2001) and even in the protozoan trypanosomes, a parasite that causes sleeping sickness (Shi *et al.*, 2000).

As a laboratory technique to shut down gene expression,

[*]To whom correspondence should be addressed.

the dsRNA can be introduced directly to cells as a synthetic molecule or the dsRNA can be synthesized in vivo to form a stem-loop RNA or hairpin loop (hpRNA). In both cases, the dsRNA domain is cleaved to shorter fragments of about 20–25 nts (nucleotides) known as siRNA by an RNaseIII-type enzyme known as Dicer. After cleavage by Dicer, siRNAs join a multicomponent nuclase complex, termed RISC (RNA Induced Silencing Complex) that is competent for triggering degradation of the target mRNA. Recent studies suggest that the initial dsRNA information is amplified by the action of RNA-dependent RNA polymerase. The availability of this complex cellular machinery to handle dsRNA as a control molecule made it clear that the laboratory technique of RNAi is using a natural cellular mechanism.

Indeed, all eukaryotes carry gene-silencing mechanisms that have dual cellular functions. RNAi serves as a defense systems against invasion either by nucleic acids such as aberrant transcript derived from mobile genetic elements or viruses. For example, most plant viruses are degraded by this mechanism known also as post-transcriptional gene silencing (PTGS) and many plant viruses even developed proteins that are able to interfere and inactivate this silencing. In plants these silencing mechanism can also operate during overproduction of a transgene that may result in accumulation of aberrant RNA, a phenomenon known for years as co-suppression. In mammalian cells the introduction of dsRNA longer than 30 nts causes apoptotic response, the so called interferon response that is initiated by the dsRNA-dependent kinase (PKR). This is why, in order to circumvent this general response elicited by long dsRNA, the use of RNAi as mean to silence specific genes in mammalian cells was based on introduction of only short synthetic siRNA molecules or expression of a very short hairpin RNA precursor structure.

At least three independent pathways are related to RNAi in eukaryotes (for recent reviews see (Eddy, 2001; Hannon, 2002)). One that elicits specific *degradation* of mRNA as discussed above, a second one that *inhibits* the translation of mRNAs, the third involved chromatin silencing directed by siRNAs. Dicer is known to be involved in all these different pathways. Indeed, mutations in Dicer caused pleotropic developmental abnormalities in Arabidopsis (Hannon, 2002). These effects were later implicated to be the result of the inability to process small RNAs known as miRNA from their precursor RNA (usually a tiny hairpin RNA). The miRNAs act via binding to the 3′ untransalted regions of mRNA, and by yet unknown mechanism inhibits translation. While the miRNA-target mRNA interaction can tolerate mismatches, the degradation by siRNA requires perfect (or almost perfect) base-pair matching.

Interestingly, whereas we hear more and more on the widespread presence of miRNA in human, plant, drosophila and nematodes and their inhibition effect on translation to regulate gene expression (Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee and Ambros, 2001). little is known about the presence of natural regulatory siRNA that can elicit degradation of a particular mRNA. Interestingly, there is growing evidence for the expression of anti-sense RNA that can potentially form dsRNA in vivo and elicit the degradation of mRNA in a regulated manner.

To date there is no single candidate of a control RNA that leads to mRNA degradation. However, evidence suggests that such control RNA should exist. For instance, *C.elegans* mutants defective in the mRNA degradation induced by RNAi prevent mobilization of the endogenous transposons, suggesting that one of the RNAi function is in transposon silencing (Tabara *et al.*, 1999). Indeed, studies in trypansomes suggest that in trypanosomes, the housekeeping function of RNAi involves the silencing of transposition (Djikeng *et al.*, 2001). Thus, regulation by RNA control elements is much more common than was previously believed and there are apparently many more families of such elements to be discovered in genomic data. Identification of additional RNA control elements is a major bioinformatic challenge.

Based on what is known about the mechanism of siRNA to elicit the degradation of target mRNA both from in vivo and in vitro systems, a general scenario emerges: For a gene that is under RNAi degradation control, in addition to the coding information, a control element is expected to appear elsewhere in the genome. This control element may fold into a stem-loop-stem structure that will be cleaved by Dicer to yield siRNA molecules that can potentially degrade the target mRNA. A schematic view of the process is presented in Figure 1.

Our method is based on the assumption that for an RNAi control mechanism, a triple repeat is necessary: Two occurrences form the stem, and must appear relatively close to each other and as an inverted repeat in head to head orientation. An additional occurrence, that can be anywhere in the genome, is part of the target gene. As mentioned in (Lau *et al.*, 2001) for miRNA, the complementary fragment of the RNA can come from both arms (5′ and 3′) of the control element (with preference to the 3′), thus we allow each arm in the inverted repeat to complement the coding gene. Detection of such ‘triplets’ is expected to be a good first screen for genes that may be regulated by an RNAi-like process. Note that unlike the situation in the inhibition pathway, where the dsRNA and its target are not fully matched, RNAi activity by degradation requires 100% (or close to 100%) match between the dsRNA of size 20–25 derived from the stem and its target gene.

A straight forward search for such triple repeats would take time proportional to $(NL + NS)$ where N is the

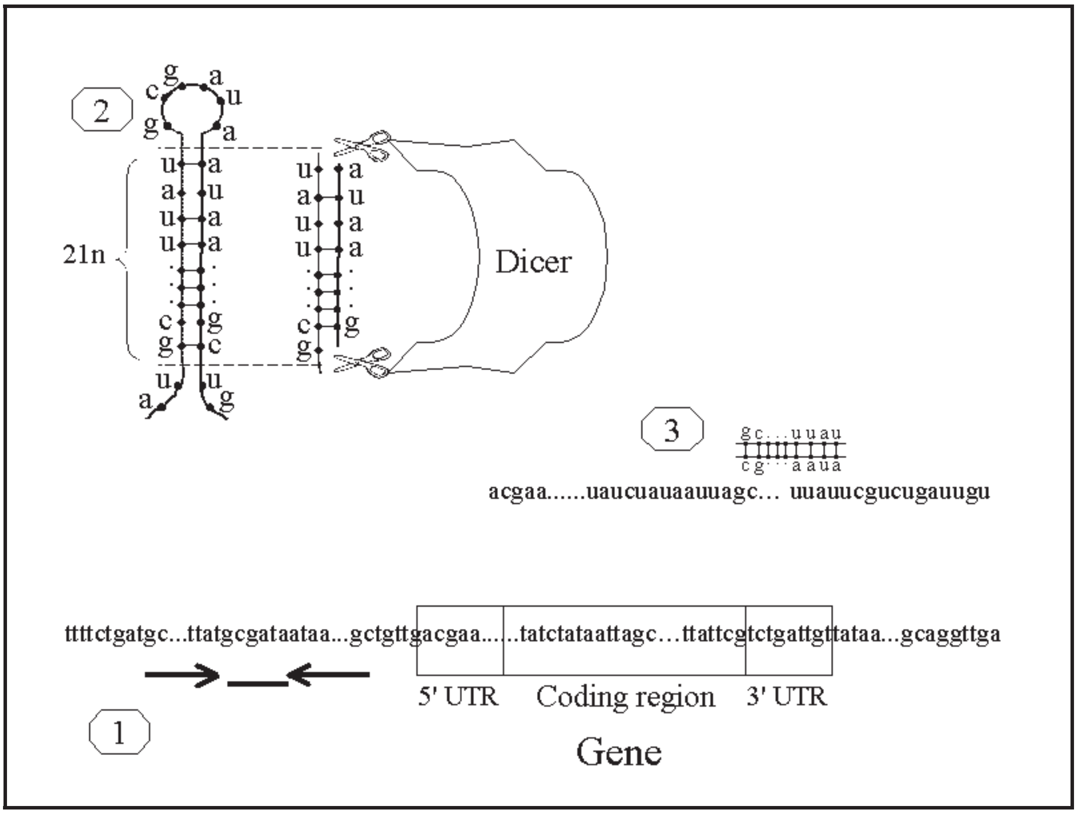
2
21n
Dicer
3
gc...uuau
cg···aaua
acgaa......uaucuauaauuagc... uuauucgucugauugu
ttttctgatgc...ttatgcgataataa...gctgttgacgaa......tatctataattagc…ttattcgtctgattgttataa...gcaggttga
5' UTR
Coding region
3' UTR
1
Gene

Fig. 1. A possible mechanism of RNAi action: (1) According to this scenario, the genome should contain three occurrences of the same sequence of size at least 20 nucleotides: Two occurrences (left) should form an inverted repeat of palindromic DNA with a gap between them, which can form a SLS (Stem-Loop-Stem) structure. The third occurrence of the sequence should be inside a gene, either in its coding region, or in its 5′ or 3′ UTR control regions. (2) The SLS structure would be transcribed and form an RNA structure. The stem of the RNA structure is cut (for example by Dicer, a ribonuclease III protein) to form a double strand RNA species. (3) The double strand RNA homes in on its target, the mRNA of the target gene that includes the third occurrence of the repeat sequence, and degrade it before its translation to a protein.

size of the genome (in nucleotides), L is the maximal size of the loop allowed between the two fragments of the palindromic repeat and S is the approximate number of inverted repeats with a gap shorter than L. This time analysis is based on finding palindromic repeats which takes time proportional to NL, using every point on the genome as a possible center and searching outward for half L nucleotides to form a palindromic match. Then, each match must be scanned against the genome to find the third repeat. For *C.elegans*, N is 10^8, L is set to 1000nt, and S is about 10^5, an overall time of about 10^{13} steps. Such a search would take a long time (on the order of few weeks of CPU time). Clearly, extending the search to larger genomes like human or mouse is not feasible.

To significantly speed up the process we propose to use a data structure of suffix tree (Weiner, 1973; McCreight, 1976). A suffix tree is a data structure that enables various string searches over text in linear time. In this structure all the suffixes of the text are stored in the tree in an overlapping manner that enables efficient search operations. Originally, the data structure was designed to answer questions of finding an occurrence of a short string in a large text. But actually many more string matching questions can be efficiently handled by using suffix trees. In particular, it can enable the identification, in linear time, of the repeat structure that we are looking for, i.e. all the fragments of size longer than 20 nt that appear at least three times. However, various modifications are needed in order to use a suffix tree for the problem at hand. A description of the suffix tree data structure, and the modifications included in our implementation, is given in the Methods section.

Repeats and inverted repeats are common in genomic data (see for example Heringa, 1998). Thus, it is clear that not in every case where this specific type of repeat occur it points to an actual case of RNAi control. In order to focus on data that have a higher probability to lead us to cases of biological significance, another source of information is needed to support the existence of RNA control elements for some of these cases. We explored using two types of data, one is using ESTs and the other is using synteny between *C.elegans* and *C.briggsae*.

ESTs (Expressed Sequence Tags) are small fragments of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an *expressed* gene. For our purposes, a database of ESTs can be considered as a collection of fragments of genes. The collection is redundant, since one gene can be the source of more than one EST, and is not complete, since not every gene is expressed when the collection is made. EST collections are designed to contain mRNA sequences that lead to proteins, since they are produced by reverse transcription of mRNA which contain poly-adenine tract that are mainly found in 3′ UTR of coding genes. Still, EST collections are known to include additional types of sequences like precursor mRNA that contain introns, and also occasionally tRNA and other types of RNA. Thus, for our application, EST collections are useful as they may include at least some of the RNA control elements that we wish to find.

We also look for supportive evidence by looking for ‘synteny’ between *C.elegans* and *C.briggsae* genes that might be under RNAi control. Since the two nematodes are closely related, it makes sense to suggest that there is a significant overlap between the sets of genes that might be under RNAi control in both organisms. Note that we are not looking for conservation of the RNAi elements themselves. Rather, we suspect that in many cases, if a gene is under RNAi control in one organism, its ortholoug gene will be under RNAi control in the other. Thus, we searched for genes in *C.briggsae* that have the potential to be under RNAi control, and compared them with the *C.elegans* results.

The current study is based on the genome of *C.elegans*, a small worm (about 2mm long) whose genome was one of the first to be sequenced (Blaxter, 1998). The genetics of this organism is known to outstanding detail as a result of a collaborative international effort that was rewarded by the Nobel prize for 2002 to Brenner, Horvitz and Sulston. The genome of *C.elegans* has about 10^8 nucleotides (100 MB) divided into 6 chromosomes, and about 20 000 genes. We have chosen *C.elegans* because of the wealth of the genomic data that will help us to determine gene location, exon/intron locations etc. An additional important reason for choosing *C.elegans* is the fact that in laboratory experiments, RNAi is extremely effective. Almost every gene of the organism has been shut down in a large scale RNAi experiment (Maeda *et al.*, 2001). This might suggest that natural RNAi-like processes are common in *C.elegans*. As mentioned above, the ability to compare results between *C.elegans* and *C.briggsae*, another worm whose genome has been recently sequenced (Mullikin and Ning, 2003) is also a useful feature.

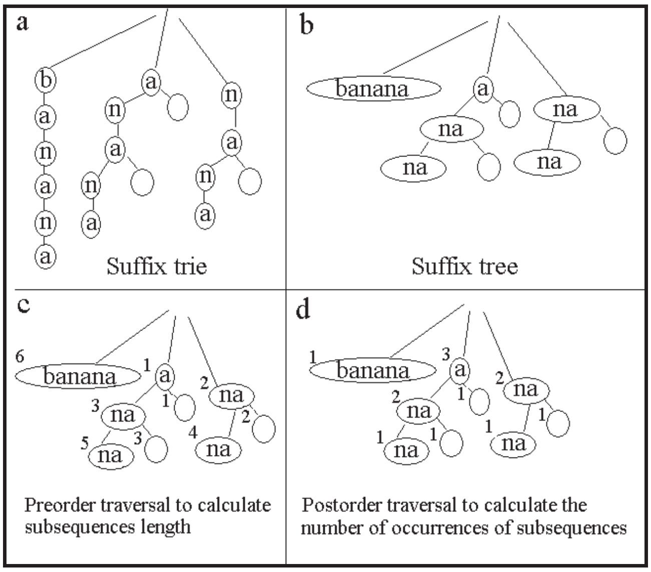
a
b
a
n
a
n
a
a
n
a
n
a
n
a
n
a
Suffix trie
b
banana
a
na
na
na
na
Suffix tree
c
6
banana
1
a
2
na
3
na
1
2
5
na
3
4
na
Preorder traversal to calculate subsequences length
d
1
banana
3
a
2
na
2
na
1
1
na
1
1
na
1
Postorder traversal to calculate the number of occurrences of subsequences

Fig. 2. An example of a suffix trie (a) and suffix tree (b) that stores the text ‘banana’ and all of the suffixes of that text. The path from the root to a node, i.e. the concatenation of the strings stores in the nodes along the path, contains the subsequence this node represents. The number of leaves below a node is equal to the number of occurrences of this subsequence. (Note that the sequence stored in each node, represented by a pair of start and end indices, is not an independent data item, it is the entire path from the root that counts.) A post order traversal of the tree (c) is used to calculate the length of subsequence represented in each node. A pre-order traversal is used (d) to calculate the number of occurrences of each subsequence.

METHODS

A suffix tree is a data structure that allows for efficient storage and retrieval of substrings of a text that is presented as one long string. The definition of a suffix tree starts with the definition of a trie. A trie is a tree-like data structure for storing strings in which there is one node for every common prefix. The strings are stored in extra leaf nodes. A suffix tree is a compact representation of a trie corresponding to the suffixes of a given string where all nodes with one child are merged with their parents. See Figure 2a and 2b for an example of a suffix trie and suffix tree representation of a string.

The strength of this ingenious data structure comes from the surprising fact the tree can be built in linear time (Weiner, 1973; McCreight, 1976). While linear time suffix tree algorithms have been known for a long time, they have not been used frequently in biological applications. Although the awareness to the effectiveness of suffix tree data structures for applications in computational biology is growing (see many examples in the comprehensive book of Gusfield, 1997), a search in PubMed (June 2003) revealed only 11 related publications. (Clearly PubMed is not the only source for published papers in related areas). The reasons for that anomaly are probably threefold: Suffix trees are not easy to implement especially if the data one wants to extract from these trees are not standard; they require keeping large data-structures in memory; and more fundamentally, suffix trees, in a straightforward application, are useful only for exact matching. Fortunately, our application requires an exact, or almost exact match of the sequence of the control element to that of its target. Thus, the use of a suffix tree, with the required modifications, was very suitable for our purpose.

One preliminary problem that we had to solve is that regular suffix trees can handle regular repeats while we wish to identify a pair of the occurrences that appear as a palindromic repeat plus an additional occurrence. This problem was solved using a simple modification of the suffix tree. The entire genome was reversed and complemented and then concatenated to the original sequence. In the 'combined' genome, an inverted repeat will appear as a regular repeat, provided that one occurrence of the repeat is from the first half of the sequence, and the other repeat from in the second half. In addition, we have to check that the original indices of these repeats are close to each other such that they can form the SLS structure.

In order to use the tree to find all of the triple occurrences of sequences above a certain length, we had to preprocess the information stored in the tree in a specific order. A 'pre-order' (all parents are visited before their sons) scan was needed to calculate the length of each fragment (Fig. 2c). Then, a 'post-order' scan (all sons are visited before their parents) was used to calculate the number of repeats of each subsequence (Fig. 2d). Combining the information from these two scans it was possible to calculate the indices of each subsequence stored in the tree. With this information, it was then possible to calculate the indices of each repeat that fits the minimal size occurrence and appear at least three times where two of the occurrences appear as an inverted repeat with a distance between them of less than a threshold value. Note that all of these scans are done in linear time.

Note that in some cases, more than one repeat or more than one target share the same sequence. Thus, we split the set of repeats, if needed, such that each set will contain a unique target.

It is important to note that by definition a repeat of length N includes two overlapping repeats of length $N - 1$, three overlapping repeats of length $N - 2$, etc. And indeed the suffix tree will report on all of these redundant cases. Thus, we need to get rid of all of this redundant information. This was done by sorting the repeats by length, and then using them to 'cover' a Boolean array of the size of the genome. A repeat was considered only if it covered a range of the array that was not covered before. This method clearly eliminated the redundant sub-copies of each repeat.

The next point to consider is that some repeats may participate in more than one control element-target combination. Thus, we had to decide in which order to resolve these conflicts. We have chosen the simple greedy criterion that prefers the combination that covers the greatest part of the genome area. Thus, the algorithm will prefer, for example, a combination of 4 repeats of length 20 bp each over a combination that has three repeats of length 25 bp each. Again the Boolean array that shadows the genome was used to implement this preference.

After generating a genomic set of candidate sets by the suffix tree, the next step was to examine individually each triplet to determine the exact boundary of the SLS structure. This step is needed since the suffix tree match is based on exact matches, but it is clear that a small number of mismatches (either as a result of sequencing errors or because stems and target sequence do not require 100% matches) could be tolerated as part of a natural RNAi. Thus, we performed an 'edit distance' alignment to extend the stems and determine their exact boundaries.

The next step was to scan the triplets against a database of EST entries. The purpose here was to look for supporting evidence that the SLS elements are truly expressed in the genome. Clearly, ESTs that are fully mapped to coding regions, i.e. exons, should be screened out since these will show up in the EST collection by virtue of the mRNA derived from the coding regions and can not provide support to the existence of transcribed control element. Since EST sequences are often short, they may not contain the entire SLS sequence but only part of their sequence. It was therefore important to decide which part of the SLS structure to search for in the EST collection. It is clear that the stem part is not useful because these sequences appear in the EST collection as part of the target sequence which is, by definition, an expressed coding gene. Similarly, taking a fragment only from the loop part is not relevant because loops are not part of the repeat and thus may be not related at all to the target gene. Thus, we searched for fragments that cover the boundary regions between the stem and the loop regions. We chose 60 nucleotides, 30 from the stem side and 30 from the loop. If the loop was not too short, then we had two fragments to consider: the left and right stem-loop junctions. The search was done by mega-Blast (http://www.ncbi.nlm.nih.gov/blast/megablast.html) to rapidly determine which triplets that appear in the genome might also appear in the EST data.

Note that control elements might exist and function even if they do not show up in the EST collection, since the EST collection is produced in a manner designed to find specifically mRNA sequences (i.e. those leading to proteins) and not other types of RNA sequences. Thus, as an alternative procedure to find supportive data for genes that might be under RNAi control, we made a parallel search for such genes in *C.briggsae.* The genome of *C.briggsae* became available recently (Mullikin and Ning, 2003). The genome contains 102M bp in 142 pieces. The genome contains about 14 000 genes. Assuming that the pieces are large enough such that most likely SLS structures reside are not broken into two pieces, we concatenated the pieces to form a ‘complete’ genome and used the same procedure as in the *C.elegans*. Note that since the SLS structure is local, and the target gene can be anywhere in the genome, we can ignore the ‘correct’ order of the genome which is still not available and get away with arbitrary ordering.

The final step involved manual inspection of promising examples, as determined by the annotation of the target gene, to confirm that it is a plausible case of RNAi control.

RESULTS

A suffix tree was built to represent the sequence of the *C.elegans* genome. In order to search for the palindromic repeats, the sequence was duplicated by adding an inverted and complementary copy of the genome. We then ran the suffix tree algorithm on this sequence database. First, we noticed that the genome contains more pairwise repeats than pairwise inverted palindromic repeats, for example there are about three times more repeats of size 50nt than inverted repeat. In length 100 the ratio grows to about 5. We then ran the suffix tree algorithm to detect the triplets of SLS plus the third occurrence in a target gene. The length of the repeats was set to at least 20 nts, and the size of the maximal gap that forms the loop in the SLS was set to 1000nt. We noticed an unusual distribution of the targets in the *C.elegans* genome. Table 1 compares the proportions of the genome devoted to exons, introns, and UTR (UnTranslated Regions) with the distributions of RNAi targets in these regions. It is clear that targets are over-represented in introns and in the UTR of genes. While the over-representation in the UTR regions is expected in light of recent studies (e.g. Lee and Ambros, 2001) that show a preponderance of RNA control elements (miRNA) to target these regions, the over—representation in the intron regions is intriguing and so far not explained. Anyhow, since we are looking for RNAi activity by mRNA degradation, we focus here on the cases where the targets were found within exons.

The first scan using the suffix tree revealed 10 350 possible candidate triplets. These triplets were then analyzed by a dynamic programming algorithm to determine the exact boundaries of the stem and the loop. In most cases multiple SLS structures target the same gene, either because the same SLS structure appear multiple times or because different regions of the same gene are targeted. When eliminating this multiple counting we found that 1453 of *C.elegans* genes are potentially targeted by SLS structures. This is about 7% of the 20 000 of the total number of genes in *C.elegans*.

Table 1.

	Exon	Intron	3′ UTR	5′ UTR
Area of genome	25%	28%	2.2%	1.5%
Num. Of Hits	20%	48%	8%	5%
Ratio	**0.8**	**1.7**	**3.6**	**3.3**

The SLS sequences where compared with the sequence of known miRNA in *C.elegans* and no significant hits were found. It is of interest to note however, that similar to the situation in miRNA (Lau *et al.*, 2001), in a significant majority (71%) of the cases, the part that matches the target come from the 3′ arm of the SLS structure.

C.elegans has about 190 000 EST sequences. Out of these sequences, about 160 000 sequences were clearly mapped using mega-Blast to coding exons. Using the remaining EST sequences we found that for about 100 target genes there are ESTs that might have originated in the corresponding SLS structures. These hundred cases were then analyzed manually, using visual tools provided in the wormbase site (http://www.wormbase.org) to make sure that the entire scenario is correct, i.e. that the cases are not redundant, that indeed the SLS matches both the target and the EST, that the relevant EST is not related to any coding region, etc. At the end we were left with about 20 cases for which RNAi seems to be a relevant control mechanism. Out of the 20 cases, only one is a gene with a known function, transposes. All the other cases are of genes with unknown function. Figure 3 shows two examples of SLS structures found in the search.

The other approach to support the existence of RNAi controlled genes was by comparing the results with a similar scan of *C.briggsae*. Out of the 14 000 genes tentatively identified in *C.briggsae*, we have found that 376 (2.6%) can be potentially targeted by SLS structures. Intersecting the two lists of genes, we have found about 50 pairs of orthologous genes that may be under RNAi control in both worms. Among the genes that were identified in this way we found hormone receptors, homeobox proteins, defensin, actins, and several types of collagens.

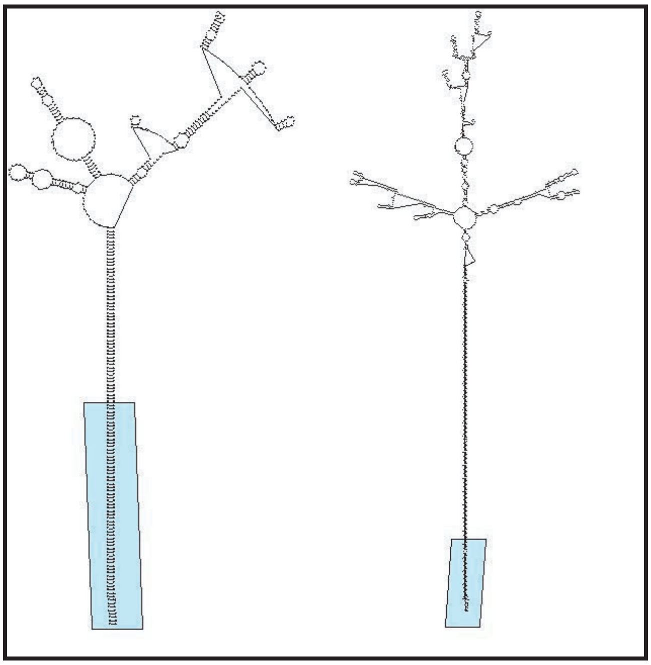

Fig. 3. The secondary structure of two SLS elements found in the search. (Left) The size of the stem region is 118 bp and the loop region is 306nt. Out of the stem, a dsRNA of 65 bp (marked in green) matches the target gene which is Tc5 transposes, a protein for which there are evidences that it is controlled by RNAi. (Right) The size of the stem is 259 bp, the loop size is 975nt and the length of the region that matches the target gene is 34. The function of the target gene, D2024.9, is not known.

DISCUSSION

Our results suggest that RNA control elements that fit the scenario of mRNA degradation exist in the genome of *C.elegans*. We have identified about 70 genes (out of a total of about 20 000 in the *C.elegans* genome) which may be controlled by RNA molecules. Interestingly, one of the genes that emerged from the search is a transposase. Transposase is an enzyme that binds to single-stranded DNA and recognizes the repetitive ends of a transposon and participates in the cleavage of the recipient site into which the new transposon copy inserts. Although it was suggested that this gene is regulated by RNAi (Tabara *et al.*, 1999; Djikeng *et al.*, 2001), there was no evidence for how RNAi control element to this gene is produced in vivo. Our computational method points to a regulatory RNA element in the genome that has the potential for producing SLS to silence this gene. Moreover it suggests that this computational method can unravel novel RNAi targets that their biological role has to be tested experimentally. Several other of the genes that we have found might indeed be controlled by RNAi, notably hormone receptors and homeobox genes. Even collagen, which is usually considered to be only a 'structural' gene, was shown to be controlled during development processes (Johnstone, 2000).

Another interesting observation made in this study is the over-representation of possible target regions in introns, an observation that we are currently attempting to understand in terms of its evolutionary and functional significance. However, it should be noted that RNAi does not operate to silence pre-mRNA (i.e. mRNA that contains introns), as it was not possible to silence genes present in a polycistronic transcript. However, silencing of intronic sequences may take place in the nucleus and may be an efficient system to degrade transcripts that carry fortuitous introns and that would not be spliced and accumulate in the nucleus. Based on this finding one might speculate that additional, yet unknown, mechanism of RNAi control might exist.

We describe here a procedure that enables the scanning of entire genomes for potential RNA control elements. The main element of the procedure is the use of a suffix tree to locate triple repeats, including one set of nearby palindromic repeat. Palindromic repeats were located by duplicating the genome and concatenating it to an inverted and complementary copy of the genome sequence. Even with the duplication, the entire genome of *C.elegans* (of 100 Mb) can be scanned in about 4 h on a single processor. This performance and the fact that the dependence of the size of the genome is linear, suggest that running similar procedures on the larger mammalian genomes, which are one order of magnitude larger, should be possible. On the other hand, memory issues will have to be addressed, since the current application required a very large memory of about 12 GB. We are currently exploring ways to get an implementation that will be more memory efficient, as well as looking into algorithms that consider efficient implementation of suffix trees using external swap space.

Overall, our conclusion is that suffix trees are highly suitable tools for locating RNA control elements especially for case where RNAi works via mRNA degradation, since the level of complementarity between RNAi and their targets is very high (Elbashir *et al.*, 2001), and thus methods based on exact matches like suffix trees are appropriate.

Nevertheless, it is clear that the fact that suffix trees only consider 100% exact matches is a limitation (for example it is likely that a small number of G–U pairs can be tolerated), and thus we have begun to explore ways to build suffix trees that are able to accommodate a small number of mismatches. We recently presented (Amir *et al.*, 2000) a suffix tree variant that enables identification of repeats despite a single mismatch. The data structure presented in this paper enables to answer only simple inquiries regarding repeats. We are currently attempting to modify the data structure such that it can be used to identify more complicated combinations of repeats then those described here. It is clear to us that even within the scenario of RNAi activity by mRNA degradation, the method presented here can not detect all molecules involved, and that not all of the molecules that we have identified will turn out to be regulated by RNAi control. Yet, we believe that the current work is a significant step forward.

As we gain a greater understanding of the actual molecular mechanisms of RNA regulation operative in the cell, our search criteria may need to be modified. Nevertheless, we believe that the principles presented here will be applicable in a large number of scenarios.

The function of RNA molecules in controlling gene expression and other critical cellular processes is being gradually unveiled. The ultimate proof for RNA regulation in each example we identify depends on experimental work, and indeed, the candidate cases presented here are being studied experimentally. It is clear that a large contribution to this field will come from bioinformatic studies on the genomic level. We hope that the methods presented here will facilitate further work in this direction.

ACKNOWLEDGMENTS

Ron Unger and Amihood Amir are partially supported by the Bikura foundation of the Israeli Academy of Science.

REFERENCES

Amir,A., Landau,G., Keselman,D., Lewenstein,M., Lewenstein,N. and Rodeh,M. (2000) Text indexing and dictionary matching with one error. *J. Algorithms*, **37**, 309–325.

Bernstein,E., Caudy,A.A., Hammond,S.M. and Hannon,G.J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**, 363–366.

Blaxter,M. (1998) *Caenorhabditis elegans* is a nematode. *Science*, **282**, 2041–2046.

Couzin,J. (2002) Breakthrough of the year: small RNAs make big splash. *Science*, **298**, 2296.

Djikeng,A., Shi,H., Tschudi,C. and Ullu,E. (2001) RNA interference in *Trypanosoma brucei*: cloning of small interfering RNAs provides evidence for retroposon-derived 24-26-nucleotide RNAs. *RNA*, **7**, 1522–1530.

Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.

Elbashir,S.M., Harborth,J., Lendeckel,W., Yalcin,A., Weber,K. and Tuschl,T. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.

Fire,A., Xu,S., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.

Gusfield,D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge.

Hannon,G.J. (2002) RNA interference. *Nature*, **418**, 244–251.

Heringa,J. (1998) Detection of internal repeats: how common are they? *Curr. Opin. Struct. Biol.*, **8**, 338–345.

Johnstone,I.L. (2000) Cuticle collagen genes. Expression in *C.elegans*. *Trends Genet.*, **16**, 21–17.

Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

Lee,R.C. and Ambros,V. (2001) An extensive class of small RNAs in *C.elegans*. *Science*, **294**, 862–864.

Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *C.elegans*. *Science*, **294**, 858–862.

Maeda,I., Kohara,Y., Yamamoto,M. and Sugimoto,A. (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol.*, **11**, 171–176.

McCreight,E.M. (1976) A space-economical suffix tree construction algorithm. *J. ACM*, **23**, 262–272.

Mullikin,J.C. and Ning,Z. (2003) The phusion assembler. *Genome Res.*, **13**, 81–90.

Reinhart,B.J., Slack,F.J., Basson,M., Pasquinelli,A.E., Bettinger,J.C., Rougvie,A.E., Horvitz,H.R. and Ruvkun,G. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *C.elegans*. *Nature*, **403**, 901–906.

Shi,H., Djikeng,A., Mark,T., Wirtz,E., Tschudi,C. and Ullu,E. (2000) Genetic interference in *Trypanosoma brucei* by heritable and inducible double-stranded RNA. *RNA*, **6**, 1069–1076.

Tabara,H., Sarkissian,M., Kelly,W.G., Fleenor,J., Grishok,A., Timmons,L., Fire,A. and Mello,C.C. (1999) The rde-1 gene, RNA interference, and transposon silencing in *C.elegans*. *Cell*, **99**, 123–132.

Weiner,P. (1973) Linear pattern matching algorithms. *Proceedings of the 14th IEEE Symp on Switching and Automata* 1–11.

Wesley,S.V., Helliwell,C.A., Smith,N.A., Wang,M.B., Rouse,D.T., Liu,Q., Gooding,P.S., Singh,S.P., Abbott,D. Stoutjesdijk,P.A. *et al.* (2001) Construct design for efficient, effective and high-throughput gene silencing in plants. *Plant J.*, **27**, 581–590.

Zuker,M. and Jacobson,A.B. (1998) Using reliability information to annotate RNA secondary structures. *RNA*, **4**, 669–679.