RC23530 (W0502-103) February 17, 2005
Other

IBM Research Report

rna22: A Unified Computational Framework for Discovering miRNA Precursors, Localizing Mature miRNAs, Identifying 3' UTR Target-islands, and Determining the Targets of Mature-miRNAs

Isidore Rigoutsos, Kevin Miranda, Tien Huynh
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

rna22: A Unified Computational Framework for Discovering miRNA Precursors, Localizing Mature miRNAs, Identifying 3' UTR Target-islands, and Determining the Targets of Mature-miRNAs

Isidore Rigoutsos[†,‡] **Kevin Miranda**[†] **Tien Huynh**

Bioinformatics and Pattern Discovery Group
IBM Thomas J. Watson Research Center,
PO Box 218, Yorktown Heights, NY 10598

Keywords: RNA interference, pattern discovery, miRNA precursor, mature miRNA, 3' UTR target-islands

[†] These authors contributed equally to this work.

[‡] Corresponding author: rigoutso@us.ibm.com

ABSTRACT

Mature microRNAs are short (21-22nt long) RNAs that are enzymatically excised from endogenously-encoded, longer precursors and have been shown to hybridize to mRNAs transcripts causing the transcripts' downregulation through the RNA interference mechanism. Because of the importance of RNAi in the regulation of cell processes, attempts to answer the questions of cardinality and location of the miRNA precursors which are encoded by a given genome have been at the center of scientific research for several years already. The arguably open question of cardinality not withstanding, many miRNA precursors have already been reported in the literature for several genomes, and additional are continuously sought. A component that is equally important for shedding more light on the details of the RNAi process is that of determining the cardinality and location of the targets of these mature miRNAs as well as the identity of the mature miRNA that will hybridize to a given target. Generally assumed to be located in the 3' UTRs of mRNA transcripts, these miRNA targets have proven to be much more elusive, and, despite great amounts of work by many scientists around the world, very few of them have been validated experimentally to date. Due to the high cost (in materials and time) of the experimental approach, computational methods are becoming increasingly important as they can help focus the experimentalist's attention and effort while maximizing the rate of experimental success. All of the computational methods that are available in the literature have generally treated the problems of miRNA precursor discovery, mature miRNA localization, miRNA-target-island determination and mature-miRNA/miRNA-target identification as separate tasks with varying degrees of reported success. In this paper, we present a method that simultaneously tackles the four problems of miRNA precursor discovery, mature miRNA localization, miRNA-target-island identification and mature-miRNA/miRNA-target determination, in a single, uniform framework. To the best of our knowledge this is the first method of its kind that addresses all these questions in a unified way. In contrast to some of the previously reported techniques that were developed and focused on specific genomes, our method, *rna22*,[§] is genome-independent and applies equally-well to genomes spanning the spectrum from viruses to mammals. Key to our method is the use of a greatly redundant scheme for representing *locally conserved* signatures that are identified by processing the sequences of known precursors and mature miRNAs using an exhaustive pattern discovery technique. The use of local signatures liberates us from the limitations associated with seeking precursor-wide conservation across the genomes of related species while potentially permitting the identification of precursors and 3' UTR target-islands that are potentially mosaic-like structures composed of known elemental blocks. Using a very extensive computational analysis, we examine the capabilities of our method and demonstrate that it a) identifies essentially all currently known miRNA precursors, b) very accurately locates the mature miRNAs in all known precursors, c) correctly predicts most of the 3' UTR regions that have been shown to be targeted by known mature miRNAs, and, d) correctly predicts a large percentage of the miRNA/mRNA-target pairs that have appeared in the literature. Additionally, our method has the very desirable characteristic of simultaneously exhibiting substantially high sensitivity and specificity values. We have used our method to analyze several genomes and to obtain revised estimates for the number of endogenously coded miRNA precursors as well as for the number of 3' UTR islands that will act as targets of one or more mature miRNAs: summarily, our analysis suggests that both of these numbers are likely to be substantially higher than initially believed. Taken together, our analysis suggests that there exist a very extensive combinatorial mechanism for carrying out post-transcriptional gene regulation within the cell and that the RNA interference-based regulation of cellular processes is a very pronounced and wide-ranging mechanism.

[§] Our method's name, *rna22*, is both a word-play and a tribute to the late Douglas Adams, the brilliant writer who gave us the famous Hitchhiker's Guide to the Galaxy. Among other things he gave us "42," the answer to "Life, Universe and Everything." Just as "42" was the quotable number of the late 20th century, "22" is slowly shaping into the magic number of the early 21st century. And, of course, in keeping with Douglas Adams' sense of humor, we would like to point out that the number 22 happens to also be the string representing the number 8 in a base 3 number system: notably, both 8 and 22 are numbers which have become very meaningful in the context of RNA interference.

INTRODUCTION

Following a serendipitous discovery in the early 1990's, scientists realized before long that a wide spectrum of organisms had the ability to exploit short RNA sequences for the purpose of degrading mRNA or disrupting translation of mRNA into amino acids. The phenomenon has become known as RNA interference (RNAi for short) or "post-transcriptional genomic silencing" (PTGS).

In recent years, RNAi was catapulted to the forefront of attention and research activity. The reason for this attention lies on the potential of harnessing RNAi for a number of purposes including speedy elucidation of gene function, targeted gene silencing for therapeutic uses, etc.

The broad process of RNAi was first uncovered in experimentation on the gene responsible for the purple color of petunia flowers: chalcone synthase. Jorgensen and colleagues wanted to increase the purple color of petunia flowers; they hypothesized that injection of extra copies of the chalcone synthase gene would have the desired effect. Surprisingly they found that injection of the transgene led to a large percentage of entirely white and/or patterned flowers. Experimentation revealed transgenic white flowers had a 50-fold decrease in chalcone syntase mRNA levels compared to the wild-type. They termed this phenomenon “co-suppression” because both the endogenous and introduced chalcone synthase genes were suppressed (Napoli *et al.* 1990).

A similar reduction of gene expression was seen in the fungus *N. crassa*. Romano and Macino were investigating the albino-3 and albino-1 genes that produce enzymes involved in carotenoid biosynthesis. They found that Neurospora transformed with portions of these genes exhibited an albino phenotype. The authors referred to this process as quelling due to the sudden and forceful suppression of carotenoid biosynthesis (Cogoni *et al.* 1994).

RNAi was then found to extend into the worm *C. elegans.* Guo and Kemphus were characterizing the par-1 gene which encodes a serin-threonine kinase responsible for asymmetric cleavage of the developing *C. elegans* embryo. Deletion of the gene was lethal to embryos and the authors hypothesized that injection of anti-sense par-1 into the gonad of wildtype worms would produce a similar phenotype in the progeny, which it did. However, the negative control, injected sense par-1, also exhibited this phenotype (Guo and Kemphues 1995). This counterintuitive result was explained by subsequent work by Fire and colleagues (Fire *et al.*, 1998) who varied both the structure and delivery of the RNA. They found that double stranded RNA, compared with either sense or antisense RNA was 10 times more potent at eliciting RNAi. Since then, RNAi has been observed in a variety of organisms including zebrafish, hydra, fungi, drosophila, mammalian systems, and one virus.

Insight into the molecular mechanisms underlying the RNAi process was first revealed in (Hamilton and Baulcombe 1999) who noted that 25 bp species of dsRNA was found in plants under co-suppression. Additionally, they found that these small RNAs had a sequence similar to the gene under suppression. Elbashir and colleagues demonstrated, first in drosophila then mammals, that RNAi was mediated by these small interfering RNAs (Elbashir *et al.* 2001).

RNAi is effected by two classes of short RNAs that are currently considered to be distinct: siRNAs and miRNAs. The two classes differ in the mechanism that triggers them but share a common mechanism in the manner that the class members act on their targets. Briefly, siRNAs, or small interfering RNAs, are activated by exogenously-supplied double stranded RNA which acts as a silencing trigger. With the help of the DICER enzyme, the dsRNA is processed into short fragments, 21 to 23 nucleotides in length (mature siRNA), which in turn act on their target genes by degrading these genes’ mRNA. siRNAs typically exhibit perfect or near-perfect complementarity to the sequence of their target gene. Unlike siRNAs, miRNAs, a.k.a. micro-RNAs or small temporal RNAs, begin their journey by first being transcribed from the host genome as pri-miRNAs. These are in turn cleaved by the RNase III endonuclease *Drosha* into 60-70 nucleotide-long segments that have been termed precursor miRNAs (Lee *et al.* 2002). A

precursor miRNA forms a very characteristic hairpin-like double stranded RNA structure by folding back on itself. The newly transcribed precursor miRNA is then transported across the nuclear membrane into the cytoplasm through a process that depends on *exportin 5* (Lund *et al.*, 2003, ; Yi *et al.* 2003). Once in the cytoplasm, a precursor miRNA is cleaved by the *Dicer* protein into 22mers that are subsequently unwound by a helicase into single-stranded species (Khvorova *et al.* 2003; Schwarz *et al.* 2003; Tomari *et al.* 2004). One of the ssRNA molecules is degraded while the other is incorporated into RISC (=RNA-induced silencing complex), a multi-subunit ribonuclear particle complex (Hutva´gner and Zamore, 2002a; Martinez *et al.* 2002; Tomari *et al.* 2004). Once complexed with RISC, the mature miRNA directs either the translational inhibition or the degradation/cleavage of target messages (Doench and Sharp 2004).

A lot of progress has been made in recent years towards elucidating the specific details of this mechanism. Nonetheless, several questions which naturally arise here still remain open, in our opinion: how many miRNA precursors are encoded by a given genome? how many mature miRNAs will a given precursors give rise to and where are they located? how many and which locations in a given gene's 3' UTR region will become the targets of a RISC-complexed mature miRNA? and, probably the most important question of all, which mature miRNA(s) will target a given gene's 3' UTR and where?

Given a set of candidate sequences, the above questions can be answered in a more or less straightforward manner through wet-lab experimentation. For example, a typical avenue for validating the presence of a given miRNA is the use of Northern blots. However, lack of validation through Northern blotting is not proof positive of the absence of a miRNA. Indeed, there is the possibility that the sought miRNA is expressed in quantities too small to be detectable by mRNA hybridization. Alternatively, the miRNA may be expressed in large quantities but in a very limited set of cells or during specific developmental or cell-life stages (Johnston and Hobert 2003). In the case where the sought RNA exists in very small quantities, the use of PCR could conceivably provide an alternate solution. But the use of PCR necessitates the availability of the appropriately designed, non-universal primers, something that to this date has been hindered by the inability of all previously-reported computational methods to localize a mature miRNA with reasonable confidence.

Computational approaches have an advantage over the traditional, labor-intensive experimental techniques in that they can help focus the latter and avoid time consuming trial-and-error schemes. Consequently, more and more techniques are being proposed in the literature which address various aspects of the RNAi process. One immediate observation is that the methods proposed to date have attempted to answer each of the above questions independently of one another and using methodologies attuned to the specific task. With regard to the discovery of miRNA precursors, this has largely focused on hairpin assessment aided by comparison to known miRNAs and constrained by the requirement for cross-species conservation (Kiriakidou *et al.* 2004; Lim *et al.* 2003; Lim *et al.* 2003; Grad *et al.* 2003; Lai *et al.* 2003; Lee *et al.* 2001; Ambros 2003). Alternatively, one could search for cis-regulatory signals that presumably affect the transcription of miRNA precursors and which signals help locate the latter. Judging from what is currently available in the open literature, one can conclude that the question of how many miRNA precursors are coded by a given genome has been effectively answered. As will become apparent during this discussion, we believe that this question has all but been addressed. In fact, our results suggest that the number of precursors and mature miRNAs that are encoded by genomes such as those of *C. elegans, D. melanogaster, H. sapiens, M. musculus, etc.* is likely to have been underestimated by as much as one order of magnitude in some of the genomes.

Another open question that is amenable to computational treatment pertains to the determination of the number and the location of the sites that are the targets of mature miRNAs. Equally importantly, one would like to determine for each such site the identity of the miRNA that will hybridize to it. Generally assumed to be located in the 3' UTRs of mRNA transcripts, these miRNA target sites have proven to be elusive, and, despite great amounts of work by many

scientists around the world, very few of them have been validated experimentally and reported to date. For reasons that will become apparent shortly, we introduce the term "*3' UTR target-islands*" to refer to the sites that will be targeted by miRNAs – this term is meant to reflect the passive nature of these sites as well as what we believe is a clear separation of these locations or groups of these locations from the surrounding regions.

On the miRNA-target-discovery front, the proposed methods have been more varied. Here as well, almost all methods place a strict requirement for a potential site to be conserved across several species. Summarily, the methods for the discovery of miRNA targets have been based on one of the following categories: a) dynamic programming (Enright *et al.* 2003, Kiriakidou 2004); b) signature-based – here it is typical to use as a signature 6 consecutive nucleotides taken from the first 8 nucleotides in the 5' region of the miRNA at hand and this 'signature' is used either explicitly (Lewis *et al.* 2003, Lewis *et al.* 2005) or implicitly (Rajewsky and Socci, 2004); c) hidden Markov models (Stark *et al.* 2003); d) semi-exhaustive techniques – for example, running the miRNA along a candidate 3' UTR, calculating interactions at every site and subselecting those that are significant according to a specific statistical measure (Rehmsmeier *et al.* 2004).

In what follows, we present a method that addresses all of the above questions: cardinality of precursors and mature miRNA, cardinality of miRNA islands in 3' UTRs, miRNA/3'UTR pairs, in a single, unified framework. Key to our method is the use of a greatly redundant set of signatures, that are derived through a pattern discovery process, and which capture locally conserved signatures. This is continuing along a line of research that we begun several years ago on the discovery of locally conserved motifs in amino acid sequences and their exploitation in the context of protein annotation, gene discovery, etc. (Rigoutsos and Floratos 1998a; Rigoutsos and Floratos 1998b; Rigoutsos *et al.* 1999; Shibuya and Rigoutsos 2002; Rigoutsos *et al.* 2002; Rigoutsos *et al.* 2003; Huynh and Rigoutsos 2004). Interestingly, the possibility of miRNA precursors having a modular nature has received very little attention, presumably due to the fact that there is not enough available data at this point that would allow us to draw such conclusions. Of course, one 'module' that is mentioned in the literature is the 8-nt block which has been identified in the 5' region of the currently known miRNAs (Lewis *et al.* 2003, Lewis *et al.* 2005).

The use of local signatures gives us the ability to locate miRNA precursors without the limitations imposed by the requirement for precursor-wide conservation across the genomes of related species. In parallel, the use of local signatures permits, at least in principle, the identification of precursors and 3' UTR target-islands that are potentially mosaic-like structures composed of currently unknown elemental blocks. Summarily, our method allows us to identify essentially all currently known miRNA precursors and mature miRNAs in these precursors. The method also allows us to correctly predict most of the 3' UTR regions which have been shown experimentally to be targeted by known mature miRNAs. Finally, the method correctly predicts a large percentage of mature miRNA/miRNA-target pairs that have been published in the literature. These results are coupled with simultaneously high values for sensitivity and specificity.

We have applied our method to the analysis of several genomes and report on estimates for the number of endogenously coded miRNA precursors. Additionally, we have obtained first and report on estimates for the number of 3' UTR target-islands which we believe will form complexes with one or more mature miRNAs. Finally, we give examples of predicted RNA/RNA complexes that have been derived using our method. The results of our analysis suggest that the number of precursors in at least some of the publicly available genomes may have been underestimated, in some cases by a very significant amount. Also, the estimated number of predicted 3' UTR target-islands suggests the existence of a very extensive combinatorial mechanism for carrying out post-transcriptional gene regulation within the cell and that the regulation of cellular processes through RNAi is a very pronounced and wide-ranging mechanism.

In the Methods section, we describe our method, *rna22*, and explain how it is applied to solve the various problems that we mentioned above. In the Results section, we describe preliminary experimental results and briefly examine their ramification. And finally, we conclude with a brief discussion and some closing comments.

METHODS

Method: Background

The discovery and exploitation of patterns in computational biology has a long history that goes back at least two decades (see Rigoutsos *et al.* 2000 for a related review). For the most part, the use of pattern discovery in the biological context has been confined at the amino acid level whereas most of the very interesting applications and results have appeared in the literature only in the last 10 years. The use of pattern discovery on DNA inputs has for the most part been limited to the discovery of tandem repeats or of cis-regulatory signals in the 5' UTR of genes. To the best of our knowledge, this is the first time that a pattern discovery method is proposed as a mechanism for addressing questions that arise in the context of studying RNAi.

Method: The Key Idea

Let us consider the situation where we are presented with a database of sequences and a query and are asked to determine whether the query can be edited into a sequence that is already present in the database, under the constraint that the number of edit operations be bounded from above. Traditionally, this version of the string editing problem has been solved by variations of a basic process during which the query is compared with every one of the sequences in the database in turn – this comparison can either be exhaustive, e.g. using dynamic programming techniques (Smith and Waterman 1981), or rely on heuristics that are meant to speed up the process e.g. FASTA (Pearson 1996), BLAST (Altschul *et al.* 1997).

Several years ago, we proposed an approach which is the logical inversion of this basic process (Rigoutsos *et al.* 1999a; Rigoutsos *et al.* 1999b, Rigoutsos *et al.* 2000): beginning with the sequence database, we use pattern discovery (Rigoutsos and Floratos 1998a) and combinatorially generate a collection *C* of regular expressions (i.e. patterns) that is a redundant representation of the original database and 'covers' it as completely as possible. Given the collection *C* of patterns, we can answer the above problem by searching the query sequence for one or more instances of the patterns from *C*. Unlike the previous approach where one searches the database with a query, we search the query for instances of an alternative representation of the database's contents.

One can think of the sequence database in question as a repository of information that describes a specific context, a training set; for example, the database could comprise the sequences of all currently known miRNA precursors. Patterns derived from such a treatment of the various sequences are bound to capture intra- and inter-family signatures. Clearly, in a context such as the one we are discussing, i.e. RNAi, the concept of a sequence "family" is not well-defined yet (except perhaps for definitions aligned with species boundaries). Intra-family signatures are typically global in terms of their span/extent whereas the inter-family ones are local.

This pattern discovery approach is particularly suitable in situations where the nature and the cardinality of the characteristics shared by a given group of sequences cannot be easily identified or described. To the extent that the input sequences in such a database represent a large and diverse sampling of the underlying sequence space, the patterns which would be derived from this process would also represent an exhaustive collection of intra- and inter-family signals that have been discovered in an unsupervised manner.

One can distinguish two stages in the *rna22* method: "off-line" and "on-line." The off-line stage is also the training stage during which we generate descriptors from the available data. These descriptors are then used during the on-line stage to address each one of the questions of interest.

Figure 1 below outlines the generic steps that we follow during the off-line stage. We begin with an appropriately selected training set (miRNA precursor sequences and mature miRNA sequences respectively), then use the Teiresias pattern discovery algorithm [Rigoutsos and Floratos 1998a; Rigoutsos and Floratos 1998b] to discover patterns that are contained in it. We use statistical significance criteria to subselect from among these patterns and conclude with the generation of a set of patterns that can be used as predicates for membership in the collection that the training set represents.

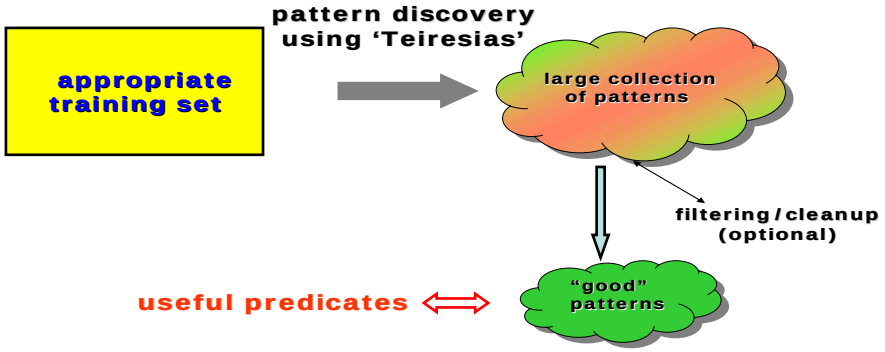

Figure 1. The generic process that is used in the off-line stage of *rna22*.

If the training set comprises the sequences of all miRNA precursors that are currently known, then, by definition, the resulting patterns will capture various characteristics that are conserved locally and possibly globally and are shared by 2 or more of the input sequences. If, on the other hand, the training set comprises the sequences of all known mature miRNAs, then the pattern set will correspond to local (or global) characteristics shared by two or more mature miRNA sequences. What is important and should be stressed here is that pattern discovery as a process obviates the requirement that there be a global alignment between any two or more sequences in the training set. As such, *rna22* can accommodate the possibility that sequences such as our miRNA precursors are actually composed of as-yet-unidentified modules in a manner analogous to what has been encountered and described in amino acid sequences (as well as in the rest of the hierarchy that extends from nucleic acid to genome, for that matter).

Method: Discovering Precursors

We begin with the sequences of known miRNA precursors that are contained in RFAM (Griffiths-Jones 2004) and following the process of Figure 1 we generate a set of precursor-specific patterns P_{pre} that appear in two or more of the RFAM precursor sequences and have been filtered for statistical significance. It should be stressed here that, unlike previously proposed methods, we do not draw lines across organismal boundaries but rather use the union of all known miRNAs from all organisms to generate P_{pre}. Subsequently, we use P_{pre} to process intergenic/intronic regions from the genome of interest. If a sequence of nucleotides corresponds to a miRNA precursor, we expect that it will also contain numerous instances of many patterns

from P_{pre} in clear contrast with its surrounding region – each candidate precursor island is subsequently filtered to ensure a minimum length (typically 60 nucleotides) and that a minimum number of patterns from P_{pre} have instances across the candidate precursor's span. The candidate sequences are then folded using RNAfold (Hofacker *et al.* 1994) to ensure that they form a hairpin. Those candidates that do not form a hairpin or contain internal self –hybridizations are discarded at this stage. Candidates can also be discarded based on the Gibbs free energy of the formed hairpin structures (typical threshold -28 Kcal/mol). Since we are also interested in the mature miRNAs that will be derived, we will discard candidate precursors if a mature miRNA cannot be localized in them.

Method: Localizing a Mature miRNA

We begin with the sequences of known matures miRNAs that are contained in RFAM and following the process of Figure 1 we generate a set of mature-miRNA-specific patterns P_{mat} that appear in two or more of the RFAM mature sequences and have been filtered for statistical significance. As in the case of miRNA precursors, we process all known miRNAs from all organisms at once to generate P_{mat}. With the collection P_{mat} at hand, we process the sequences of the candidate precursors from the previous step looking for concentrations of instances of mature miRNA patterns. The sought regions will be clearly separated from their surrounding area as they will be 'hit' by many more patterns. Candidate mature miRNA regions are checked to ensure that they do not overlap with the loop region, have a minimum length (typically 18 nts), and that a minimum number of patterns from P_{mat} have instances across their span. If more than one region within a precursor exceed threshold, *rna22* will report each and every one of them.

Method: Identifying 3' UTR target-islands

This is arguably one feature of *rna22* that, to the best of our knowledge, really distinguishes it from all previously-proposed methods. In particular, *rna22* can identify automatically those segments in a gene's 3' UTR that will be the targets of *some* miRNA and can do so *in the absence of any knowledge* about the targeting miRNA. This is revisited below in the context of the LSY-6/COG-1 interaction.

As we mentioned above, the patterns which are generated using the generic process shown in Figure 1 capture generic properties of the training set and abstract them in the form of regular expressions. This will also hold true for the patterns contained in P_{mat}. Recall now that, by definition, the target site of a given miRNA is expected to look like the reverse complement of the targeting miRNA's sequence. This is precisely the property we use to identify 3' UTR target-islands. Starting with the set P_{mat} we first generate its reverse complement ${}^{revcompl}P_{mat}$ I.e. if pattern [AT][CG].TTTTT[CG]G..[AT][AT][AT]G[CG].CTT is contained in P_{mat} then the pattern AAG.[CG]C[AT][AT][AT]..C[CG]AAAAA[CG][AT] will be contained in ${}^{revcompl}P_{mat}$. We then use the set ${}^{revcompl}P_{mat}$ to process the 3' UTR of each gene seeking concentrations of instances of the reverse complements of mature miRNA patterns. As before, the sought regions will be clearly separated from their surroundings as they will be 'hit' by many more patterns. Candidate 3' UTR target-islands are filtered to ensure a minimum length (typically 16 nts), and that a minimum number of patterns from ${}^{revcompl}P_{mat}$ have instances across their span. Clearly, if there is reason to believe that miRNAs can target other regions in addition to a gene's 3' UTR, then the process we just described can be used to examine these additional regions as well.

Method: Determining Complexes between miRNA and 3' UTR Target-islands

At this point, and for a given organism we have described methods that allow us to determine candidate miRNA precursors and their corresponding miRNAs, and candidate target-islands that are present in the 3' UTRs of all of the genes of the organism (or elsewhere – see also above). Let us assume that we have a collection C_{miRNA} of mature miRNAs and a collection of C_{genes} of interest. The task at hand is one of determining the answers to the following questions: (a) given

a miRNA *m* from C_{miRNA}, determine its target genes and report the locations in each gene's 3' UTR that *m* will target; (b) given a gene *g* from C_{genes} determine the miRNAs from C_{miRNA} that will target *g*. Given the previous discussion, the process we outline here should be straightforward. In order to answer (a) above, we do the following: for each gene *g* in $C_{islands}$ determine g's 3' UTR target-islands ${}^{g}C_{islands}$ then proceed by generating and ranking the interactions between *m* and $\{{}^{1}C_{islands}, {}^{2}C_{islands}, {}^{3}C_{islands}, \ldots\}$. In order to answer (b) above, we do the following: we compute gene *g*'s 3' UTR target-islands ${}^{g}C_{islands}$ then proceed by generating and ranking the interactions between each member of C_{miRNA}, and each island (if more than one) in ${}^{g}C_{islands}$.

Clearly, this approach assumes that for a given gene *g* the set ${}^{g}C_{islands}$ is non-empty, that it contains the 'correct' target sites (which are of course unknown), and, that what it contains is but a small subset of the *g*'s original 3' UTR (so as to also enjoy performance gains). Whether these constraints are satisfied is also dependent on the employed thresholds. As we will see in the Results section, and to the extent that we can evaluate our performance by analyzing all currently known RNA/RNA complexes, *rna22*'s does indeed perform well.

The last remaining item is that of specifying how these interactions are computed. At least two possibilities exist at this point. First, one can use only the subset of ${}^{revcompl}P_{mat}$ that is formed from only those patterns of P_{mat} that are also present in the mature miRNA *m* under consideration. This choice will generally lead to increased specificity in determining the correct target but is likely to adversely affect sensitivity. As a second possibility, one can simply form putative complexes of *m* that start at each location of every island in ${}^{g}C_{islands}$. From a performance standpoint, this is actually an acceptable process because ${}^{g}C_{islands}$ will contain only a small fraction of *g*'s original 3' UTR. Several other possibilities exist, each with its own sensitivity and specificity characteristics but their discussion escapes the score of this document.

RESULTS

This section is organized as follows. We begin by presenting results on publicly available data that highlight the various aspects of our method. We then proceed with a performance evaluation of our method's various modules again using publicly available data. This is followed by the reporting of new findings and the mention of several interesting observations which showcase the power of our method. Finally, we conclude by presenting estimates on the numbers of miRNA precursors, mature miRNAs and target sites for several genomes of interest.

But first, we would like to point out that our training set intentionally comprises a set of publicly available data that is more than 1 year old. In particular, we worked with Release 3.0 of the RFAM database from January 2004. This release contained 719 entries from human, mouse, worm, fly and plants. Since the available datasets are still relatively small, we wanted to use an older dataset for training as it would create an ideal setting for evaluating *rna22* using more recent data and determining how well it can extrapolate.

Discovering Precursors / Locating Mature miRNAs:

We begin by showing an example of the kinds of pattern coverage that will be received by a nucleotide region that encodes for a miRNA. Figure 2 below shows such a plot for hsa-miR-224 which was part of the training set. The x-axis shows nucleotide position. The y-axis of the left plot shows the coverage in terms of P_{pre} patterns, whereas the y-axis of the right plot shows the coverage in terms of P_{mat} patterns. As can be seen from this Figure, the support that the region obtains is non-zero for the length of the precursor. Moreover, the P_{mat} pattern support clearly delineates the location of the mature miRNA for this precursor between positions 8 and 29 inclusive.

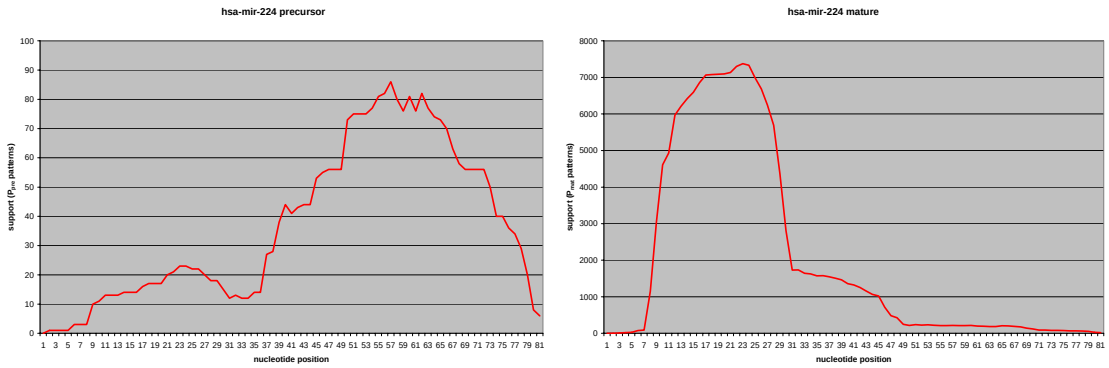

Figure 2. Example of pattern coverage for the region that encodes the has-mir-224 precursor. Left: coverage in terms of P_{pre} patterns as a function of position. Right: coverage in terms of P_{mat} patterns as a function of position.

Extrapolating From Old-Known to New-Unknown Cases: Precursors and Matures

A question that typically arises in data mining applications has to do with overfitting a given training set: the mining method at hand has learned to fit the training set so well that it has lost its ability to generalize. Here we report on how many of the entries of Release 5.1 from December 2004 can be correctly predicted by *rna22*. It should be noted that Release 5.1 is almost twice the size of our training set and contains 1420 entries. Table 1 below shows how many new entries have been added to RFAM since version 3.0 and what percentage of these entries is correctly predicted by *rna22*. Numbers are given separately for each genome. As can be seen, despite the fact that the original training set is rather small, *rna22* can extrapolate and can discover *de novo* almost 90% of all miRNAs that have been added to RFAM since release 3.0 with a notable weakness in the plant genomes.

	Precursors in training set	**RFAM 5.1 Dec 2004**	**Difference**	**discovered by *rna22***	**% of new items discovered by *rna22***
ATH	43	112	69	18	25.0%
CBR	48	79	31	10	30.3%
CEL	106	116	10	0	0.0%
DME	78	78	0	0	N/A
DPS	0	73	73	67	91.8%
DRE	0	30	30	27	90.0%
EBV	0	5	5	3	60.0%
GGA	0	121	121	109	90.1%
HAS	176	222	46	71	65.7%
MMU	202	224	22	100	76.3%
OSA	28	134	106	35	32.4%
RNO	38	186	148	148	93.7%
ZMA	0	40	40	37	78.7%
Total=	**719**	**1420**	**701**	**625**	**89.2%**

Table 1. This table shows for each genome separately, the percentage of the *new* entries which have been added to RFAM since release 3.0 and which can be correctly discovered *de novo* by *rna22*.

Identifying Target-islands and the Ability to Extrapolate

As we discussed above, *rna22* has the ability to identify miRNA target sites in a gene's 3' UTR without any need to know the identity of the miRNA that will bind to it. We have selected the gene COG-1 as an example which will demonstrate this capability.

Since the release of RFAM 3.0 that we used for training, many examples of miRNA/mRNA complexes have appeared in the literature. One such notable case is the LSY-6/COG-1 complex (Johnston and Hobert, 2003). LSY-6 is the first miRNA that has been identified as having a role in neuronal patterning and is responsible for controlling left/right asymmetry in the taste chemoreceptors of *C. elegans*. LSY-6 has been shown to repress COG-1, Nkx-type homeobox gene, by binding to its 3' UTR. Moreover, LSY-6 has two very interesting characteristics: it is only expressed in a very small fraction of the nematode's cells, and, the precursor is not conserved in the closely related nematode *C. briggsae* (but the mature miRNA itself is conserved).

LSY-6 was not contained in RFAM 3.0 and thus was not part of our training set. Moreover, and this is evidenced by a similarity search, LSY-6 shares no discernible similarity with any other mature miRNAs contained in RFAM 3.0, whether from nematodes or other genomes. Consequently determining the site in COG-1's 3' UTR where LSY-6 will bind is a non-trivial discovery event. Figure 3 below shows the coverage in terms of ${}^{revcompl}P_{mat}$ patterns as a function of position in the 3' UTR of COG-1. As can be seen, *rna22* is able to correctly determine the binding site even though it has never seen LSY-6, the targeting miRNA. The yellow region indicates the unusually-long, 29 nucleotide binding region which was reported in (Johnston and Hobert 2003). Its agreement with the result reported by *rna22* is exceptional. Equally importantly, several more clearly-delineated target-islands, all of them being 21-22 nucleotides long in width are present in COG-1 3' UTR suggesting the possibility that more miRNAs than just LSY-6 target COG-1.

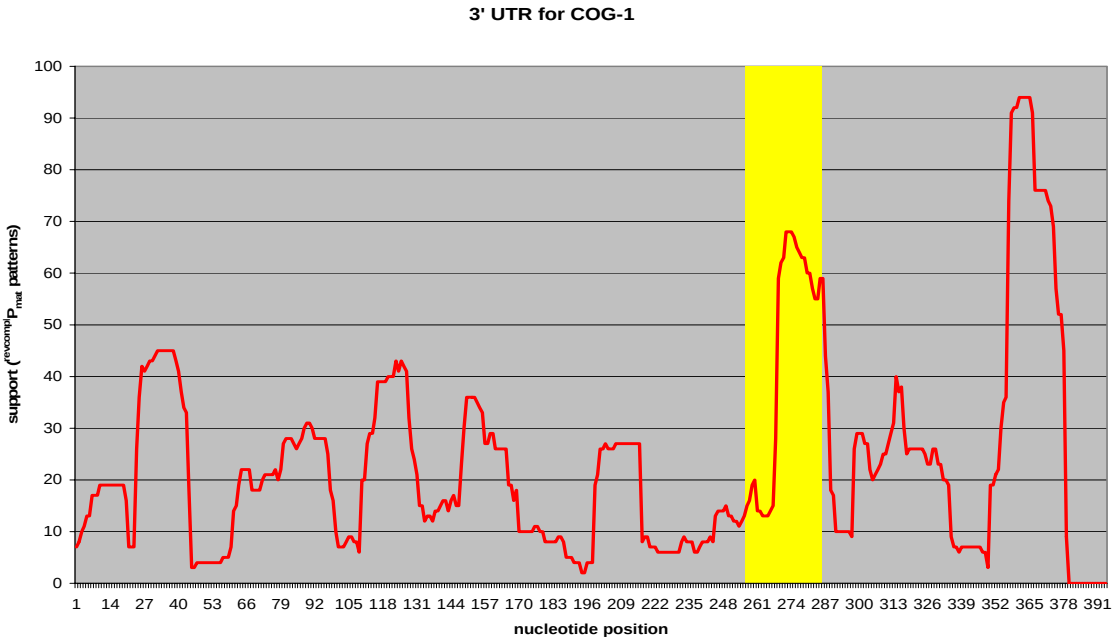

Figure 3. Plot showing the identification of distinct target-islands in the 3' UTR of COG-1 where LSY-6 has been shown to bind. The yellow region indicates the unusually long, 29 nt region which was reported in the literature as the target site for LSY-6 and which is in near-perfect agreement with *rna22*'s estimate for a 3' UTR target-island. See also text.

In Table 2 below we show which of the 3' UTR target sites that have been reported in the literature as being the targets of interfering miRNAs are picked up by *rna22* for various threshold values (=number of ${}^{revcompl}P_{mat}$ patterns that cover the reported region).

⇐ Value of Pattern Threshold ⇒

Target GENE	miRNA	20	25	30	35	40	45	50	55	60	65	70	StartPos
lin41-cel-2	cel-let-7												738
hid-dme-1	dme-bantam												874
hid-dme-2	dme-bantam												1711
hoxb8-has	hsa-mir-196a												411
lin28-hsa	hsa-let-7b												890
mapk14-hsa.	hsa-mir-24												651
fbxwib-hsa	hsa-mir-103												2392
brn3b-hsa-3	hsa-mir-23a												463
enx1-hsa-2	hsa-mir-101												114
cog-1-cel	cel-lsy-6												257
myotrophin-	has-mir-375												3126
bdnf-hsa-1	hsa-mir-1b												220
lin28-cel	cel-lin-4												328
g6pd-hsa-2	hsa-mir-1b												433
nmyc-hsa.	hsa-mir-101												494
laminin-hsa	hsa-mir-199b												209
lin41-cel-1	cel-let-7												689
dmtf1-hsa	hsa-mir-15a												123
smad-hsa-2	hsa-mir-26a												103
clock-hsa	hsa-mir-141												214
cgi38-hsa	hsa-mir-16												294
bdnf-hsa-3	hsa-mir-1b												1322
brn3b-hsa-1	hsa-mir-23a												102
smad-hsa-1	hsa-mir-26a												46
smc1l1-hsa	hsa-let-7e												73
enx1-hsa-1	hsa-mir-101												59
g6pd-hsa-1	hsa-mir-1b												97
pten-hsa-1	hsa-mir-19a												411
pten-hsa-2	hsa-mir-19a												N/A

Table 2. Color coded reporting of the number and validated targets sites are discovered by *rna22* as a function of the used pattern threshold.

From this Table, and for a rather conservative threshold of 35 ${}^{revcompl}P_{mat}$ patterns, we can generate a rough estimate of *rna22*'s sensitivity in identifying a true 3' UTR target-island to be equal to 21/29=72.4% - the true number is likely to be higher. In terms of specificity when identifying a 3' UTR target-island, *rna22*'s performance can be estimated with the help of the negative examples mentioned in (Kiriakidou *et al.* 2004). We again use the same threshold of 35 patterns. Of the 9 reported sites that were predicted to be targets of several known miRNAs but could not be validated experimentally, 3 are reported by rna22 as corresponding to 3' UTR target-islands. Since the experiments in (Kiriakidou *et al.* 2004) only investigated the possibility of complexes with *specific* miRNAs, the possibility exists that the 3 sites which are reported by *rna22* do indeed function as miRNA targets but for a miRNA other than the one that was investigated by the authors. Thus, using this admittedly limited set of negative examples, we estimate the specificity of *rna22* to be *at least* 66%.

Next, let us examine the binding energies for the mRNA/miRNA complexes that have been reported in the literature. These energies are shown in Table 3 below. This is of interest because

it permits us to estimate an appropriate threshold for the binding energy of miRNA/mRNA complexes that will be reported by *rna22*: only predicted complexes with energies *lower* than the value of the selected threshold should be reported. Being conservative, we decided to select a threshold of -28 Kcal/mol for *rna22* and did so with the full understanding that we would be forced to ignore a substantial number of true interactions.

⇐ Value of Gibbs Free Energy (in –Kcal/mol) ⇒

mRNA	miRNA	-15.0	-17.5	-20.0	-22.5	-25.0	-27.5	-30.0	-32.5	-35.0	-37.5	-40.0	Energy of complex
hoxb8-has	hsa-mir-196a												-45.7
delta-hsa-1	hsa-mir-34												-38.3
lin28-hsa	hsa-let-7b												-38.2
lin41-cel-1	cel-let-7												-34
mapk14-hsa.	hsa-mir-24												-32.3
delta-hsa-3	hsa-mir-34												-31.4
laminin-hsa	hsa-mir-199b												-31.3
lin41-cel-2	cel-let-7												-31.1
hid-dme-2	dme-bantam												-30.9
lin28-cel	cel-lin-4												-30.6
fbxwib-hsa	hsa-mir-103-												-30
dmtf1-hsa	hsa-mir-15a												-29.6
nmyc-hsa.	hsa-mir-101												-28.1
g6pd-hsa-2	hsa-mir-1b												-26
hid-dme-1	dme-bantam												-25.7
enx1-hsa-1	hsa-mir-101												-25.6
smad-hsa-1	hsa-mir-26a												-25.5
pten-hsa-2	hsa-mir-19a												-24.6
smc1l1-hsa	hsa-let-7e												-24.2
clock-hsa	hsa-mir-141												-24
g6pd-hsa-1	hsa-mir-1b												-24
myotrophin-hsa	has-mir-375												-23.4
enx1-hsa-2	hsa-mir-101												-23.4
cgi38-hsa	hsa-mir-16												-22.7
bdnf-hsa-1	hsa-mir-1b												-21.9
bdnf-hsa-3	hsa-mir-1b												-21.7
brn3b-hsa-1	hsa-mir-23a												-21.1
pten-hsa-1	hsa-mir-19a												-21.1
brn3b-hsa-3	hsa-mir-23a												-20.2
smad-hsa-2	hsa-mir-26a												-19
cog-1-cel	cel-lsy-6												-17.8

Table 3. Color coded reporting of the experimentally validated interactions that would be reported by *rna22* at a given choice of threshold for the binding energy of the RNA/RNA complex.

Naturally, the question now arises of how many 3' UTR target-islands can *rna22* identify in the 3' UTRs of the genes for which experimentally verified bindings are available. Table 4 below shows precisely this as a function of threshold (=number of $^{revcompl}P_{mat}$ patterns that cover each region).

⇐ Value of Pattern Threshold ⇒

GENE	**miRNA**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**
lin41-cel-2	cel-let-7	12	12	12	13	10	8	8	7	7	7	4
hid-dme-1	dme-bantam	31	31	25	29	23	21	21	19	19	18	17
hid-dme-2	dme-bantam	31	31	25	29	23	21	21	19	19	18	17
hoxb8-has	hsa-mir-196a	9	10	10	10	9	7	7	4	4	4	4
lin28-hsa	hsa-let-7b	31	28	29	25	24	21	19	16	16	12	11
mapk14-hsa.	hsa-mir-24	26	28	26	26	24	22	21	17	17	17	15
fbxwib-hsa	hsa-mir-103	31	33	30	28	22	22	21	17	15	13	12
brn3b-hsa-3	hsa-mir-23a	17	19	19	17	14	11	9	9	8	7	6
enx1-hsa-2	hsa-mir-101	3	3	2	2	2	2	2	2	2	2	2
cog-1-cel	cel-lsy-6	7	6	5	5	5	3	2	2	2	2	1
myotrophin-hsa	has-mir-375	35	41	32	31	27	24	19	16	15	14	13
bdnf-hsa-1	hsa-mir-1b	30	34	34	29	27	24	22	20	18	16	13
lin28-cel	cel-lin-4	39	36	36	31	30	27	24	21	20	15	14
eif3s1-hsa	hsa-let-7e	13	14	12	10	11	10	7	7	7	7	6
g6pd-hsa-2	hsa-mir-1b	6	6	4	4	4	3	1	0	0	0	0
nmyc-hsa.	hsa-mir-101	12	10	11	9	9	9	8	6	6	5	4
laminin-hsa	hsa-mir-199b	16	15	14	15	15	15	11	10	8	6	5
lin41-cel-1	cel-let-7	12	12	12	13	10	8	8	7	7	7	4
dmtf1-hsa	hsa-mir-15a	12	14	14	13	11	10	10	9	7	6	5

smad-hsa-2	hsa-mir-26a	2	2	1	1	1	0	0	0	0	0	0
ste20-hsa	hsa-mir-141	16	17	16	15	15	10	9	8	9	8	7
clock-hsa	hsa-mir-141	34	32	31	32	31	29	27	23	22	20	18
c22orf5-hsa	hsa-mir-15	22	19	19	15	12	10	9	8	7	6	6
cgi38-hsa	hsa-mir-16	4	4	3	2	2	2	2	2	2	2	1
bdnf-hsa-3	hsa-mir-1b	30	34	34	29	27	24	22	20	18	16	13
kiaa0152-hsa	hsa-mir-24	66	64	57	55	53	45	43	42	36	33	28
brn3b-hsa-1	hsa-mir-23a	17	19	19	17	14	11	9	9	8	7	6
vegf-hsa	hsa-mir-16	23	20	19	16	16	14	13	11	11	8	6
smad-hsa-1	hsa-mir-26a	2	2	1	1	1	0	0	0	0	0	0
smc1l1-hsa	hsa-let-7e	62	65	59	63	57	56	51	46	42	36	35
enx1-hsa-1	hsa-mir-101	3	3	2	2	2	2	2	2	2	2	2
g6pd-hsa-1	hsa-mir-1b	6	6	4	4	4	3	1	0	0	0	0
pten-hsa-1	hsa-mir-19a	2	3	3	3	3	3	2	2	2	2	2
pten-hsa-2	hsa-mir-19a	2	3	3	3	3	3	2	2	2	2	2
klf5-hsa	hsa-mir-141	18	14	13	14	15	13	12	10	11	9	9
tmod3-hsa	hsa-mir-145	11	11	10	10	8	7	5	4	3	2	1
ripa-hsa	hsa-let-7b	9	8	9	6	6	6	7	6	5	5	4
gpd1-hsa	hsa-mir-103	18	17	16	17	12	10	9	6	6	6	6

Table 4. Number of target-islands that are discovered by *rna22* in the 3' UTRs of the genes that has been experimentally studied in the literature, as a function of the used pattern threshold.

As can be seen, the 3' UTRs for several of these genes contain numerous predicted target-islands that persist at very high threshold values. And given *rna22*'s results on COG-1 that we presented above we are inclined to accept that many of these islands are indeed valid. We will return to this topic below when discuss the specific case of BDNF.

Validating *rna22* on Old-Known Cases: miRNA / 3' UTR Target-island Interactions

Could we have predicted the mRNA/miRNA complexes that exist in the literature *de novo* using the current version of *rna22*? To answer this question, we used a threshold of 20 for the ${}^{revcompl}P_{mat}$ set of patterns (so as to accommodate the discovery of as many of the known islands as possible – see Table 2) and a binding energy threshold to -17Kcal/mol (so as to accommodate COG-1 – see Table 3), then used *rna22* to process all of the genes for which targets have been validated in the literature. From each of these genes, many target-islands were extracted (Table 4 above shows the exact numbers) and were used to form the set $\{{}^{1}C_{islands}, {}^{2}C_{islands}, {}^{3}C_{islands}, \ldots\}$. We then combined all of the miRNAs from the latest RFAM 5.1 into the set C_{miRNA} of candidate targeting miRNAs. Finally, we used *rna22* to determine which entries from C_{miRNA} should pair up with which island from $\{{}^{1}C_{islands}, {}^{2}C_{islands}, {}^{3}C_{islands}, \ldots\}$ and form a complex with binding energy less than or equal to -17 Kcal/mol. Invariably and for every single one of the processed islands, *rna22* identified the correct miRNA as the top ranking targeting miRNA for the island, in agreement with what had been reported in the literature. Additional interactions between miRNAs in RFAM 5.1 and the genes of Table 4 were also reported by *rna22* – this is precisely what we discuss next. No interactions were reported for enx1-hsa-1, g6pd-hsa-1, pten-hsa-1 and pten-hsa-2 since *rna22* cannot find the corresponding island even at a pattern threshold of 20 (this is shown in Table 2).

Predicting miRNA / 3' UTR Target-island Interactions

We have just described that *rna22* can correctly identify sites in 3' UTR regions which are known to form RNA/RNA complexes with interfering RNAs. We now show an example of how one can go about predicting a previously unreported mRNA/miRNA pair. In what follows, *rna22* used a threshold of 35 ${}^{revcompl}P_{mat}$ and a binding energy threshold of -28 Kcal/mol – these are actually *rna22*'s default settings. Let us assume that BDNF is our gene of interest. Figure 4 below shows the coverage of BDNF's 3' UTR by patterns from the ${}^{revcompl}P_{mat}$ collection. As in

the case of COG-1 above, several clearly delineated target-islands, approximately 22 nucleotides in length, are present.

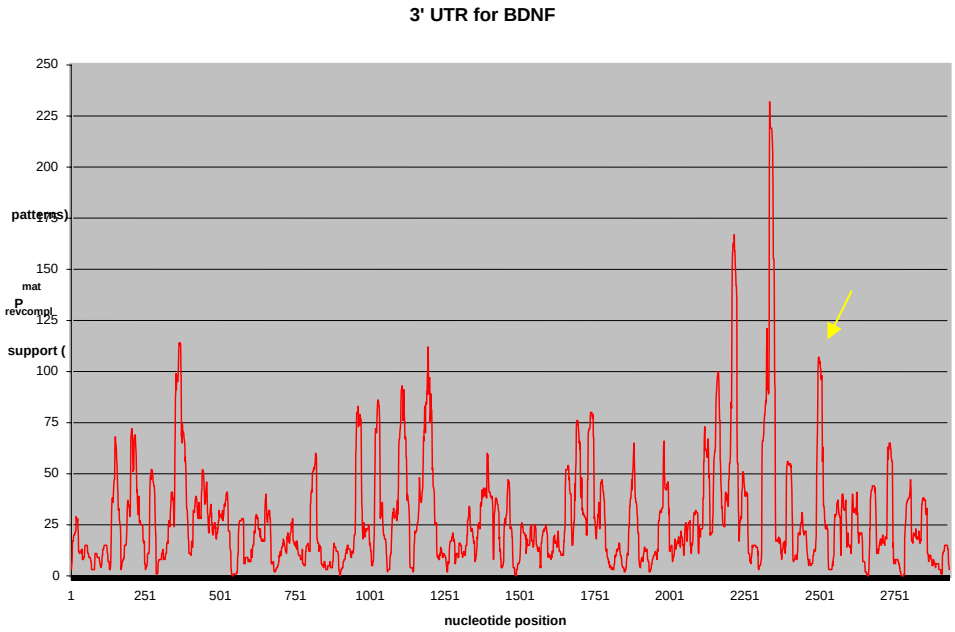

Figure 4. Plot showing the coverage of BDNF's 3' UTR by ${}^{revcompl}P_{mat}$ patterns. As in the case of COG-1 above, several clearly delineated target-islands approximately 22 nucleotides in length are evident. See also text.

We will focus on one of the more prominent islands, and in particular the one that is pointed to by the yellow arrow. The question "which of the known human miRNAs is likely to bind to this target-island?" has what appears to be a very definite answer: hsa-miR-213 will bind to a region within that island and form an RNA/RNA complex with a binding energy which is superior to that of the runner-up candidate by approximately 14 Kcal/mol (!). The predicted complex spans locations 2489 through 2510 inclusive and is shown here:

```
BDNF(ENSG00000176697)        hsa-miR-213
5′                    3′   5′                  3′
TCACAGTCACATGCTTGATGGT    ACCATCGACCGTTGATTGTACC
..(((((((.(((.((((((((    )))))))).)))))))))...
```

A picture of the RNA/RNA complex appears in Figure 5.

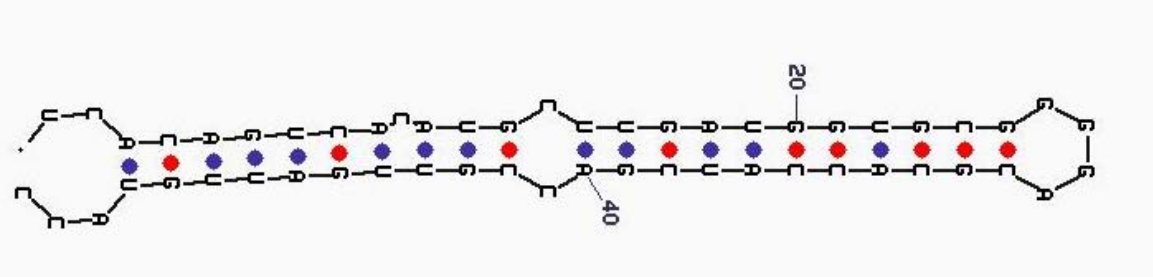

Figure 5. The RNA/RNA complex that is predicted to form between hsa-miR-213 and BDNF at locations 2489:2510. The folding and the picture were generated by MFOLD (Zucker 2003; Matthews *et al.* 1999).

Estimating the Number of Precursors and 3' UTR Target-Islands

The above results have provided evidence for *rna22*'s ability to correctly discover *bona fide* miRNA precursors, localize the mature miRNA(s) within those precursors, correctly determine 3' UTR target-islands and finally predict miRNA/mRNA complexes. Naturally, the question arises: how many precursors are encoded by a given genome? how many target islands exist in the various genes' UTRs? how many of the genome's genes are under RNAi control? Clearly, computationally predicted answers suffer from two drawbacks: they are highly-dependent on the choice of thresholds, and they are just that, predictions. Table 5 below shows our initial estimates for several genomes of interest (using a threshold of 35 patterns and -28 Kcal/mol of binding energy). There is an evident increase in the number of predicted precursors and 3' UTR islands in direct correlation with the apparent complexity of the organism at hand.

Genome	# of precursors contained in RFAM 5.1	# of precursors (predicted)	# of matures (predicted)	# of 3' UTR target-islands (predicted)	# of affected transcripts (predicted)
HHV5	0	> 50	> 60	> 700	> 100
C. elegans	116	> 250	> 350	> 12,000	> 3,000
D. melanogaster	78	> 270	> 350	> 32,000	> 6,500
M. musculus	224	>3,500	> 4,000	> 90,000	> 8,000
H. sapiens	222	>4,000	> 4,500	> 120,000	> 10,000

Table 5. Computational estimates for the number of precursors, mature miRNAs and 3' UTR target-islands for several genomes.

DISCUSSION

We have presented a unified framework for studying RNAi and computationally answering many of the important questions that arise in this context. Our framework is based on a pattern-discovery scheme and has been shown to perform well on most publicly available data and using as a starting point a very small training dataset, RFAM 3.0, that contained 719 miRNA precursors.

The resulting algorithm, *rna22*, was put to the test using a very extensive set of experiments that were diverse in nature. *Rna22* demonstrated the ability to a) correctly predict previously unseen miRNA precursors, b) correctly predict previously reported binding regions for old and newly reported miRNAs, and c) correctly predict experimentally validated mRNA/miRNA complexes. We also used *rna22* to predict a mRNA/miRNA complex between BDNF and hsa-mir-213.

Finally, we used *rna22* to estimate the number of miRNA precursors encoded in several genomes as well as the number of 3' UTR target-islands and the number of affected transcripts in those genomes. Even though, they have been generated using what we consider to be rather stringent thresholds, our estimates are at odds with what has been reported to date. These estimates suggest that, across genomes, there exist a very extensive combinatorial mechanism for carrying out post-transcriptional gene regulation and that the RNA interference-based coordination and control of cellular processes is a very pronounced and wide-ranging mechanism.

REFERENCES

1. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acid Res.* 25, 3389-3402.
2. Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D. MicroRNAs and other tiny endogenous RNAs in C. elegans. Curr Biol. 2003 May 13;13(10):807-18.
3. Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell. 2003 Apr 4;113(1):25-36.
4. Chang S, Johnston RJ Jr, Frokjaer-Jensen C, Lockery S, Hobert O. MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode. Nature. 2004 Aug 12;430(7001):785-9.
5. Cogoni, C., N. Romano and G. Macino. 1994. Suppression of gene expression by homologous transgenes. Antonie Van Leeuwenhoek 65:205-209.
6. Doench JG, Sharp PA. Specificity of microRNA target selection in translational repression. Genes Dev. 2004 Mar 1;18(5):504-11. Epub 2004 Mar 10.
7. Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev. 2001 Jan 15;15(2):188-200.
8. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. 2003;5(1):R1. Epub 2003 Dec 12.
9. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998 Feb 19;391(6669):806-11.
10. Floratos, A., I. Rigoutsos, L. Parida, G. Stolovitzky and Y. Gao. 1999. Sequence homology detection through large-scale pattern discovery. In Proceedings Third Annual ACM International Conference on Computational Molecular Biology (RECOMB '99), Lyon, France.
11. Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J. Computational and experimental identification of C. elegans microRNAs. Mol Cell. 2003 May;11(5):1253-63.
12. Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res. 2004 Jan 1;32 (Database issue):D109-11.
13. Hamilton AJ, Baulcombe DC. A species of small antisense RNA in posttranscriptional gene silencing in plants. Science. 1999 Oct 29;286(5441):950-2.
14. Johnston RJ, Hobert O. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. Nature. 2003 Dec 18;426(6968):845-9.
15. Guo S, Kemphues KJ. par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. Cell. 1995 May 19;81(4):611-20.
16. Hobert O. Common logic of transcription factor and microRNA action. Trends Biochem Sci. 2004 Sep;29(9):462-8. Review.
17. Hofacker, I. L. W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast Folding and Comparison of RNA Secondary Structures. Monatsh.Chem. 125: 167-188 (1994).
18. Hutvagner G, Zamore PD. A microRNA in a multiple-turnover RNAi enzyme complex. Science. 2002 Sep 20;297(5589):2056-60. Epub 2002 Aug 01.
19. Hutvagner G, Zamore PD. RNAi: nature abhors a double-strand. Curr Opin Genet Dev. 2002 Apr;12(2):225-32. Review.
20. Huynh, T. and I. Rigoutsos, "The Web Server of IBM's Bioinformatics and Pattern Discovery Group: 2004 update." *Nucleic Acids Research*, 32 (Web Server issue):10-15, July 2004.
21. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol. 2004 Nov;2(11):e363. Epub 2004 Oct 05.
22. Johnston RJ, Hobert O. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. Nature. 2003 Dec 18;426(6968):845-9. Epub 2003 Dec 14.
23. Khvorova A, Reynolds A, Jayasena SD. Functional siRNAs and miRNAs exhibit strand bias. Cell. 2003 Oct 17;115(2):209-16.
24. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A. A combined computational-experimental approach predicts human microRNA targets. Genes Dev. 2004 May 15;18(10):1165-78. Epub 2004 May 06.
25. Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T. New microRNAs from mouse and human. RNA. 2003 Feb;9(2):175-9.

26. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of novel genes coding for small expressed RNAs. Science. 2001 Oct 26;294(5543):853-8.
27. Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of Drosophila microRNA genes. Genome Biol. 2003;4(7):R42. Epub 2003 Jun 30.
28. Lau NC, Lim LP, Weinstein EG, Bartel DP. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science. 2001 Oct 26;294(5543):858-62.
29. Lee RC, Ambros V. An extensive class of small RNAs in *Caenorhabditis elegans*. Science. 2001 Oct 26;294(5543):862-4.
30. Lee Y, Jeon K, Lee JT, Kim S, Kim VN. MicroRNA maturation: stepwise processing and subcellular localization. EMBO J. 2002 Sep 2;21(17):4663-70.
31. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005 Jan 14;120(1):15-20.
32. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003 Dec 26;115(7):787-98.
33. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. The microRNAs of Caenorhabditis elegans. Genes Dev. 2003 Apr 15;17(8):991-1008.
34. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. Vertebrate microRNA genes. Science. 2003 Mar 7;299 (5612):1540.
35. Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U. Nuclear export of microRNA precursors. Science. 2004 Jan 2;303(5654):95-8.
36. Martinez J, Patkaniowska A, Urlaub H, Luhrmann R, Tuschl T. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell. 2002 Sep 6;110(5):563-74.
37. Mathews, D.H., J. Sabina, M. Zuker and D.H. Turner. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure *J. Mol. Biol.* 288, 911-940 (1999).
38. Moss EG, Lee RC, Ambros V. The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. Cell. 1997 Mar 7;88(5):637-46.
39. Napoli, C., C. Lemieux, and R. Jorgensen. lntroduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes *in trans*. The Plant Cell, Vol. 2, 279-289, April 1990.
40. Ohler U, Yekta S, Lim LP, Bartel DP, Burge CB. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA. 2004 Sep;10(9):1309-22.
41. Olsen PH, Ambros V. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev Biol. 1999 Dec 15;216(2):671-80.
42. Pearson W.R. (1996) Effective protein sequence comparison. *Methods Enzymol.* 266, 227-258.
43. Pfeffer S, Zavolan M, Grasser FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, Tuschl T. Identification of virus-encoded microRNAs. Science. 2004 Apr 30;304(5671):734-6.
44. Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, Macdonald PE, Pfeffer S, Tuschl T, Rajewsky N, Rorsman P, Stoffel M. A pancreatic islet-specific microRNA regulates insulin secretion. Nature. 2004 Nov 11;432(7014):226-30.
45. Rajewsky N, Socci ND. Computational identification of microRNA targets. Dev Biol. 2004 Mar 15;267(2):529-35.
46. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA. 2004 Oct;10(10):1507-17.
47. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature. 2000 Feb 24;403(6772):901-6.
48. Riek R.P., I. Rigoutsos, J. Novotny and R. M. Graham. (2001) Non-a-helical elements modulate polytopic membrane architecture. *J. Mol. Biol.* 306, 349-362.
49. Rigoutsos, I. and A. Floratos. 1998. Combinatorial pattern discovery in biological sequences: the *Teiresias* algorithm. Bioinformatics, 14(1):55-67.
50. Rigoutsos, I. and A. Floratos. 1998. Motif discovery without alignment or enumeration. In Proceedings 2nd Annual ACM International Conference on Computational Molecular Biology (RECOMB), New York, NY.

51. Rigoutsos, I., A. Floratos, C. Ouzounis, Y. Gao and L. Parida. 1999. Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. Proteins: Structure Function and Genetics, 37(2):264-277.
52. Rigoutsos, I., Y. Gao, A. Floratos, and L. Parida. 1999. Building dictionaries of 1D and 3D motifs by mining the unaligned 1D sequences of 17 archaeal and bacterial genomes. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB '99), Menlo Park, California. AAAI Press.
53. Rigoutsos, I., A. Floratos, L. Parida, Y. Gao and D. Platt. 2000. The emergence of pattern discovery techniques in computational biology. Metabolic Engineering, 2(3):159-177.
54. Rigoutsos, I., T. Huynh, A. Floratos, L. Parida and D. Platt. 2002. Dictionary-driven protein annotation. Nucleic Acids Res., 30:3901-3916.
55. Rigoutsos, I., P. Riek, R. M. Graham and J. Novotny. 2003. Structural details (kinks and non-conformations) in transmembrane helices are intrahelically determined and can be predicted by sequence pattern descriptors. Nucleic Acids Research, 31:4625-31.
56. Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD. Asymmetry in the assembly of the RNAi enzyme complex. Cell. 2003 Oct 17;115(2):199-208.
57. Schwarz DS, Hutvagner G, Haley B, Zamore PD. Evidence that siRNAs function as guides, not primers, in the Drosophila and human RNAi pathways. Mol Cell. 2002 Sep;10(3):537-48.
58. Shibuya, T. and I. Rigoutsos. 2002. Dictionary-driven microbial gene finding. Nucleic Acids Res., 30:2710-2725.
59. Slack FJ, Basson M, Liu Z, Ambros V, Horvitz HR, Ruvkun G. The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. Mol Cell. 2000 Apr;5(4):659-69.
60. Smith T. F., Waterman M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195-197.
61. Stark A, Brennecke J, Russell RB, Cohen SM. Identification of Drosophila MicroRNA targets. PLoS Biol. 2003 Dec;1(3):E60.
62. Tomari Y, Matranga C, Haley B, Martinez N, Zamore PD. A protein sensor for siRNA asymmetry. Science. 2004 Nov 19;306(5700):1377-80.
63. Yekta S, Shih IH, Bartel DP. MicroRNA-directed cleavage of HOXB8 mRNA. Science. 2004 Apr 23;304(5670):594-6.
64. Yi R, Qin Y, Macara IG, Cullen BR. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev. 2003 Dec 15;17(24):3011-6.
65. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31 (13), 3406-15, (2003).