To appear in *Non-coding RNAs*, J. Barciszewski and V. Erdmann (eds.) Landes Bioscience, Georgetown, TX 2003.

Computational Gene-finding for Non-coding RNAs

Peter Schattner
Center for Biomolecular Science & Engineering, 227 Sinsheimer Labs,
University of California,, 1156 High Street, Santa Cruz, CA 95064, USA
e-mail: schattner@cse.ucsc.edu

October 3, 2002

ABSTRACT:
Computer gene-finding programs have been quite successful at locating protein-coding genes in both prokaryotic and eukaryotic genomes. However these programs Ö which use genomic features such as long open-reading-frames and codon signatures Ö are not designed to identify non-coding RNA (ncRNA) genes. As a result ncRNA-specific gene-finders have been required.

The first successful attempts at computational ncRNA gene-finding focussed on ncRNAs with well-characterized primary sequences and/or secondary structures, such as tRNAs or methylation-guide snoRNAs. In addition user-configurable RNA-motif search programs were developed. These programs search for RNAs by looking for user-specified primary-sequence motifs and stable secondary-structures as indicated by increased Watson-Crick base-pairing or low calculated free energies. However, to date, these RNA-motif searching programs have had only modest success at finding ncRNAs.

Recently, computational ncRNA gene-finders have been developed which show promise of locating a much larger number of previously undetected ncRNAs. Some of the most successful are based on comparative sequence analysis between genomes of related species. Others exploit base-composition signatures of ncRNAs or use new methods for RNA sequence alignment and secondary-structure prediction. With these approaches, numerous previously undetected ncRNAs have been predicted and subsequently experimentally confirmed in species including *Escherichia coli* and the hyperthermophiles *Methanococcus jannaschii* and *Pyrococcus furiosus*.

This chapter will review the strategies employed in the principal computational ncRNA gene-finders. We will compare the successes of the different approaches as well as their limitations. Finally, we will consider the impact that these new computational methods are having on our picture of the world of ncRNAs.

Introduction

With the sequencing of the human genome, the scientific community has completed the first stage in compiling a complete "parts list" for the human body. The next phases – identifying the genomic location of the "parts" (i.e. the genes), discovering their functions, and determining how they are regulated are at a much less advanced stage. For the more familiar protein-coding genes there has been considerable progress in at least finding their genomic locations – by experimental means, such as by the building of cDNA libraries, as well as by computational methods. This task is not yet complete as is evidenced by the continuing debates as to the total number of human genes. But there at least appears to be a growing consensus of the approximate number of genes (at least to within a factor of two or three) as well as their genomic locations.

In the world of ncRNA genes there is nothing resembling such a consensus yet. Few authors speculate as to the total number of ncRNA genes, even in small genomes – and when they do, the estimates vary widely. For example, when two groups predicted and experimentally confirmed several novel ncRNAs in *Escherichia coli* in 2001, one group wrote[1] "we think it unlikely that there are many more than 50 sRNAs [i.e. small ncRNAs] encoded by the *E. coli* chromosome" while the other group[2] predicted that "a significant number of our 275 candidate loci do indeed correspond to independent ncRNA genes". And this is for the relatively compact *E. coli* genome whose complete genome sequence had been already known for four years.

One does not need to look far for reasons for this lack of consensus. Until recently, genome-wide screening for ncRNAs was quite limited. Traditional EST-based methods for RNA-screening were primarily designed to look for RNAs with lengths greater than 200 base pairs (bp) and with poly-A tails – i.e. for protein-coding mRNAs. To some extent the technology was deliberately skewed away from detecting ncRNAs because of the (somewhat self-fulfilling) beliefs that ncRNAs were few in number and not of much biological interest.

The last two years, however, have seen a significant increase of activity in identifying and characterizing ncRNAs. Experimental efforts[3,4] have yielded dramatic results – showing that ncRNAs are far more numerous and presumably have greater biological significance than had been anticipated. These results, though clearly exciting, are outside the scope of the present review; the reader is referred to the chapter of this volume by J.P. Bachellerie and J. Cavaille or the original papers.[3,4] Until recently, computational ncRNA searches have also had only limited success. In this case the reason has been less lack of interest, than the difficulty of developing effective gene-

finders. However, the last two years has also seen increasing success in ncRNA-gene discovery by new kinds of computational approaches.

Because of these developments, it seems timely to review the status of computational ncRNA gene-finding. I will summarize the different strategies which have been developed for this purpose– indicating the principal strengths and limitations of each . Particular attention will be paid to algorithms which have been developed within the last two years. In the next section we begin with a brief review of the basic strategies for ncRNA computational gene-finding and a comparison with the simpler case of gene-finding for protein-coding genes. The following two sections consider gene-finders targeting ncRNAs whose primary sequence and secondary structure are at least partially known and conserved. First we focus on customized programs that search for a single RNA class. Then we examine more general search programs that can be reconfigured by the user to target a variety of ncRNAs. Next we looks at the more difficult task of searching for ncRNAs when we have little or no idea of their consensus primary sequence or secondary structure. In the final section we summarize the accomplishments and limitations of these programs and speculate on their future development.

RNA sequence alignment, RNA secondary-structure prediction and the identification of RNA motifs in mRNA sequences will only be discussed to the extent that they have impacted ncRNA gene-finding. For further information on these topics the reader is encouraged to consult recent articles and reviews [5-9] and references therein.

Gene-finding for protein-coding genes and ncRNAs

We begin with a brief review of the methods used for finding protein-coding genes. In the recent article "Gene-finding approaches in eukaryotes."[10] Stormo notes that there are two main components to a gene-finder: the type of information – or *what* to look for, and the algorithm – or *how* to look for that information

Sequence information – signals, content statistics and similarity

Stormo groups sequence information into three basic classes: "signals", "content statistics" and similarity to known genes. Sequence signals may include promoters, terminators, poly-A-addition and transcription-factor binding sites, splice sites, start and stop codons and CpG islands. In addition, if characteristic distances are known between these individual signals, the distances themselves can serve as sequence signals.

Already here we begin to see the challenge of ncRNA gene-finding. Except for splice sites in the occasional multiple-exon ncRNA, only the usually weakly-conserved

promoter and terminator signals (and possibly other poorly known transcription binding sites) will be present in ncRNA genes.

Content statistics – i.e. non-random variations in base sequence – are also useful clues for finding protein coding genes. Especially in prokaryotes, open reading frame (ORF) length alone can serve as a statistically significant gene marker. In addition, codon statistics can be exploited. Species-specific variations in the selection among synonymous codons can be utilized. Since specific pairs of amino acids are often adjacent in proteins, constraints on more probable sequences of bases in a gene can be found. Moreover, since the third codon base is often degenerate, nonrandom relationships in mutations in homologous genes can be exploited. None of these codon-specific statistical variations are available in ncRNA gene-finding.

Finally, the rapid growth in the number of known protein sequences has made it increasingly likely that a new protein-coding gene will have at least some homology with an already-known protein. As a result, sequence-similarity methods (e.g. BLAST searches[11]) are often effective in gene hunting for protein-coding genes.

Again the situation with ncRNAs is more challenging. Far fewer ncRNA sequences are available in the databases. ncRNA sequences can be compared only at the nucleotide level – not as translated amino acids – and, except for ribosomal RNAs (rRNAs), ncRNA sequences are generally short. Consequently, distinguishing weakly conserved genes from random "hits" is more difficult when searching for ncRNAs than for protein-coding genes. Moreover, even in cases where there are large RNA families, sequence conservation is often at the secondary-structure level, i.e. what is conserved are base pairings rather than the individual base sequence. As a result, except for rRNAs or RNAs with well-conserved homologs in closely-related species, conventional sequence-similarity methods have had limited success in ncRNA gene identification. For other RNAs, different methods have been required.

The one class of sequence signals that ncRNAs do have, that are not present in protein-coding genes, are those that result from secondary-structure constraints. For example, many ncRNAs have specific base-pairings, computable low-free-energy folding patterns, unusual base-composition variations or characteristic cross-species patterns of mutations occurring in complementary pairs. As we will see, these secondary-structure signatures have been central in the design of many ncRNA gene-finders.

Algorithms

Nearly all of the current gene-finders for protein-coding-genes, such as Genscan,[12] Genie[13] and Glimmer,[14] use probabilistic algorithms – typically implementing some form of a Hidden Markov Model (HMM).[15] A central feature of these algorithms is a dynamic-programming protocol which is used to "train" the model – i.e. to maximize the selectivity between a training-set of verified genes and a negative training-set of similar sequences which do not represent genes.

However, HMMs are not able to model sequences with secondary structure and consequently are ill-suited for ncRNA gene-modeling. Moreover, as we shall see, fully probabilistic algorithms that can model secondary structure tend to have long CPU execution times and to require a larger training set than may be available. Consequently we will see a variety of deterministic and partially probabilistic algorithms – such as weight-matrix approaches[15,16] – in addition to the fully probabilistic gene-finders. In addition, with ncRNAs there is often only limited training data. Consequently, choices commonly must be made between training with only a small number of sequences for the specific ncRNA gene being sought, or using a larger number of training sequences, that however include a wider assortment of ncRNA genes.

A final issue in ncRNA-gene searches is determining precisely *where* to search. Should the gene-finding program work equally well on all genomes or only on those with specific properties (e.g. ones with high AT content)? Should the program scan the entire genome or only specific genomic regions? These considerations are generally irrelevant in conventional gene-finders. Gene-finders for protein-coding-genes usually search the entire genome, except for regions of multiply-repeated subsequences. And, aside from the fact that gene-finders are typically designed for either prokaryotic genomes or eukaryotic ones, protein-coding gene-finders generally work comparably well on most genomes. On the other hand, because of the often subtler signals in ncRNA gene searches, it is sometimes advantageous to restrict such searches either to only a portion of the genome (typically to the regions that do not overlap any known protein-coding gene) or to only a limited range of genomes with specific characteristics.

<u>Custom-designed ncRNA genefinders</u>

We begin with a discussion with programs that are custom-designed to find a single type of ncRNA. Among these the most successful have been gene-finders for the tRNAs and the methylation guide snoRNAs.

tRNA gene-finders

The earliest ncRNA gene-finders were custom programs designed to search for tRNAs. During the 1980's and early 1990's, increasingly detailed models of tRNA sequence and secondary structure were developed, [17-19] culminating in tRNAscan by Fichant and Burks[20] and the Pol3Scan linear search algorithm of Pavesi et al. [21] tRNAscan had a sensitivity (i.e. a "true positive" rate) of 95.1% with false positive rate of 0.37 / megabasepair(Mbp).[22] Pol3scan had a sensitivity of 98.6% with false positive rate of 0.23 / Mbp.[22] Both programs used weight matrices taken from the sequences of the known tRNA genes, though the motifs each program looked for were somewhat different. Pol3scan searched only for primary sequence signals – tRNA sequence motifs, transcriptional control elements and terminator sequences from eukaryotic tRNAs – while tRNAscan used tRNA sequence motifs combined with secondary structure patterns.

Although the sensitivities and specificities of tRNAscan and Pol3scan were impressive, they were not yet fully satisfactory for genomic scanning. For example, a false positive rate of 0.37 / Mbp would imply approximately 1100 false positives in the 3,000 MB human genome. Interestingly, the 7 known tRNAs missed by Pol3scan were (with a single exception) different from the 19 tRNAs missed by tRNAscan.[21] Considering that the two programs looked for somewhat different motifs, this result is not so surprising. But it did suggest that a more powerful gene-finder might be possible if the two programs were combined.

The next improvement in tRNA searching came with the introduction of the COVE program[23] in 1994. COVE is a reconfigurable gene-finder and will therefore be covered in the following section. However, COVE's principal success to date has been in searching for tRNAs, where it achieved a sensitivity of 99.8% with an estimated false positive rate less than 2 / *gigabase*[22] Unfortunately, this sensitivity came at a cost; COVE's algorithm could only scan at approximately 20 base pairs / sec[22] – far too slow for routine genomic screening .

Consequently the next efforts at tRNA searching involved developing faster algorithms. FASTRNAscan[24] ran faster than tRNAscan or Pol3scan and far faster than COVE. However, its sensitivity and specificity were worse than that of COVE. tRNAScan-SE[25] on the other hand managed to match the sensitivity and specificity of COVE while decreasing COVE's running time by a factor of 15,000. tRNAScan-SE accomplished this by first running tRNAscan and Pol3scan with more permissive "cutoff values" – in order to rapidly scan a genome for candidate hits. Subsequently the candidate-gene list was passed to COVE where the potential hits were subjected to COVE's stringent testing. Since COVE only had to test a relatively small proportion of the original

candidate sequence (1 – 10%), its slow execution speed was not a problem. tRNAScan-SE is now used routinely to rapidly screen newly sequenced genomes for tRNA genes, achieving sensitivities of 99.5% with an expected false positive rate of only 0.07 / gigabase.[22]

Unfortunately, it has not generally been possible to match the success of tRNAScan-SE in searches for other classes of ncRNAs. Probabilistic gene-finders like COVE generally work best when the RNA being sought is highly conserved in primary and secondary structure, and when sequences in multiple species are known and available for use in training. Consequently, tRNAs were ideal targets. Over 1000 tRNA sequences from multiple species have been in the databases[26] for many years. However, for many other types of ncRNAs only a few examples are known and there is little data available for model training.

Searches for methylation guide snoRNAs

One group of RNAs for which sufficient data has become available to develop successful custom gene-finders are the methylation-guide snoRNAs. An early attempt at snoRNA gene-finding used a combination of standard sequence pattern –recognition programs.[27] This approach however generates a large number of false-positive "hits". Consequently, the snoRNA gene search-space was limited to vertebrate introns, since previously identified snoRNAs had been found in these sequences. This approach resulted in the identification of 9 previously unknown methylation-guide snoRNAs.[27]

However, performing a genome-wide search for snoRNAs required a custom-designed, probabilistic search program. Such a program[28] – known as snoscan - was able to predict 22 previously undetected *S. cerevisiae* methylation-guide snoRNAs which were subsequently experimentally confirmed. For 12 of the 22 snoRNAs, the associated methylation site had previously not been known. Snoscan also facilitated identification of snoRNAs throughout the domain of Archaea, after a seed training set was biochemically isolated from a single species.[29]

Despite its successes, snoscan has its limitations. If the snoRNA methylation site is unknown – and hence can not be used as part of the snoRNA signature – the S/N ratio of the program decreases (i.e. the number of false positive increase.) Consequently snoscan has been difficult to apply to larger genomes with unknown methylation sites. In addition, snoscan is only designed to detect methylation-guide snoRNAs, Locating the less-well-conserved pseudouridylation-guide snoRNAs by computational means has not been accomplished to date.

Customized gene-finders for other classes of ncRNAs

Beyond tRNAs and methylation guide snoRNAs, there are few examples of custom ncRNA gene-finders which have successfully identified new RNA genes. Dandekar and Sibbald predicted several trans-splicing RNAs in a search of the EMBL database of which some candidates were later confirmed experimentally.[30] Lisacek et al[31] developed a weight-matrix-based program which successfully found 132 of 143 known group I catalytic RNAs; however, no new RNAs were predicted with this program. In addition, very recently, custom programs for locating microRNAs [C. Burge, personal communication] and tmRNAs[32] have been developed, with promising initial results

Reconfigurable ncRNA gene-finders.

In addition to the customized RNA gene-finders, user-configurable programs to search for RNA motifs have been developed since the late 1980's.[33,34] With these programs, the user specifies either a "descriptor file", a set of "production rules" or else a multiple-sequence alignment to describe the class of RNAs being searched for.

In programs using the descriptor-file approach, the descriptors typically include primary-sequence motifs, secondary-structure patterns, and gap-lengths between motifs . In most programs of this class, the user can also set additional search parameters such as the allowed number of mismatches in a motif or whether G-U base pairs should be accepted as matches in secondary-structure stems. A typical descriptor file is shown in figure 1.

In addition to the programs of refs 33-34, descriptor-file RNA search programs include RNAMOT,[35,36] RNABOB,[37] Overseer,[38] Patscan/Patsearch,[39] Palingol,[40] and RNAMOTIF.[41] With the exception of the recently introduced RNAMOTIF, these programs are all deterministic and consequently have limitations when searching for sequence motifs that are weakly conserved. RNAMOTIF, which is based on the earlier RNAMOT program, is an attempt to introduce the elements of probabilistic searching while maintaining the user-programmable descriptor-file interface of the earlier programs. Specifically, RNAMOTIF introduces user-supplied "score functions" that can incorporate statistical, thermodynamic or other information into the motif-evaluation procedure. Recently, RNAMOTIF has successfully searched for signal recognition protein (SRP) RNAs using an empirical scoring function based on observed biases in nucleotide and base-pair frequencies and loop lengths. Using this scoring function, RNAMOTIF was able to locate SRP RNAs in seven previously unannotated prokaryotic genomes.[41]

A different approach to user-configurability was taken by Searls and collaborators who introduced the concepts of "context-free-grammars"(CFGs) from the field of

computational linguistics to ncRNA gene-finding. CFGs are elegant models of ncRNA sequence and secondary-structure in which the descriptor file is replaced with a set of production rules (see figure 1). The production rules describe how to generate all the allowed structures of the model (e.g. the class of RNA structures). In this sense they are very similar to the production rules of "regular expressions" from computer science – or their probabilistic counterpart, Hidden Markov Models. However , CFGs can model more complex structures than regular expressions and as a result are able to model RNA secondary structure in addition to primary sequence. For additional details on CFGs in RNA structure modeling, the reader is referred to the original articles[42,43] and earlier reviews.[15] In practice, the CFG models did not predict any new ncRNAs; their importance has been more in laying the foundation for the stochastic-context free-grammar (SCFG) models that followed.

The third class of user-configurable RNA-motif and RNA-gene-finders rely on the input of sequence alignments of known RNAs to train the gene-finder, rather than using either descriptor-files or grammar production rules. The idea is that rather than have the user manually extract the critical features in a family of RNA sequences, the program will do it automatically.

The first examples of this class of programs were the stochastic context free grammars COVE by Eddy and Durbin[23] and the SCFG program of Sakakibara et al. [44] These programs are fully probabilistic extensions of the CFG models of RNAs. . The Sakakibara program requires a structurally-annotated, multiple-sequence alignment for training. COVE, on the other hand, can – with relatively "ideal" training data – be trained in three different ways: using a multiple-sequence alignment – with or without structural annotation – or just with a set of unaligned sequences.[45]. However, in more realistic situations, a structurally-annotated, multiple-sequence alignment for training is also important for COVE to perform well.[46]

The SCFGs have demonstrated advantages over their deterministic predecessors in cases where training data with many aligned sequences were available –e.g. for tRNAs. However, in most cases, such extensive training data is not available. In addition, SCFG's cannot describe non-planar RNA structures such as pseudoknots nor – at least in their current implementations - can they model non-exponential gap-length distributions. In addition. They are also complex; users of SCFG approaches typically face steep learning curves. Finally, and perhaps most importantly, the use of stochastic context free grammars has been limited by their being computationally expensive. Typical SCFG memory costs are of $O(N^3)$ and time costs are of $O(N^4)$ for a sequence of length N.[46] The speed limitation can sometimes be addressed by the use of a fast preprocessor, as was done by tRNAscan-SE. And recent work indicates that the memory demands of

SCFGs may also be decreased.[46]. However, to date, these technical limitations have limited the widespread application of SCFGs in ncRNA gene-finding.

To address some of the limitations of the SCFG's, a probabilistic model called ERPIN[47] was recently introduced. ERPIN uses weight matrices rather than a SCFG to probabilistically model an RNA sequence alignment. ERPIN requires a secondary-structure annotation along with a trusted multiple-sequence alignment for training. When such an annotated multiple-sequence alignment is available – e.g. with tRNAs – ERPIN performs well. In contrast to the SCFG's, ERPIN can handle RNA pseudoknots and its "reasonable run times"[47] are listed among its advantages compared to SCFGs. On the other hand, ERPIN has only limited capability for handling complex helix-stem "indels" – i.e. insertion and deletions of large structured regions within a base-paired stem. Consequently, ERPIN would be expected to have more difficulty than the SCFG's in handling RNAs with highly variable secondary structures such as Rnase-P RNA.

In any case, ERPIN as well as the SCFGs require multiple-sequence alignments – generally with structural annotations – for training. As a result, the future success of these methods will depend on improvements in the accuracy of RNA sequence alignment and RNA structure prediction. Within the last year, programs such as Dynalign[7] and Foldalign[8] have shown that when RNA sequence alignment and RNA structure prediction are performed simultaneously, the results can be significantly improved compared to when the two operations are performed separately. Foldalign has been used to create sequence alignments for use by COVE in mRNA motif-finding, with encouraging results.[48] In principal, this strategy of supplying more accurate training-alignments to probabilistic RNA motif-finders should improve the results of ncRNA gene-finding as well.

Although the idea of a user-configurable ncRNA gene-finder is appealing, these programs have had only limited success at actually finding new ncRNAs to date. Apart from tRNA searching and the application of RNAMOTIF to SRP RNA searches, there have been few confirmed ncRNA-gene predictions by programs of this class. Gaspin et al identified and experimentally confirmed 46 previously unknown *Pyrococcus* methylation-guide snoRNAs using the program Palingol along with genomic sequence comparisons[49] Another example was the identification of hammerhead ribozyme RNAs in schistosome satellite DNA by Cedergren and collaborators using RNAMOT.[50] In addition, Cedergren's group predicted other ncRNAs by running RNAMOT against the genomic databases – however, their papers do not indicate whether any of these putative ncRNAs were subsequently experimentally confirmed.[51,52] In fairness, it should be noted that most of the user-configurable RNA-motif finders were not designed primarily

to be ncRNA gene-finders. Instead, their main objective has been to detect sequence-motifs and secondary-structure motifs in mRNA, at which they have had more success.[53,54]

de novo ncRNA gene-finding – searching for genes without *a priori* knowledge of sequence or structure

As we have seen, when sufficient training data is available, computational searches for well-characterized ncRNAs can be quite successful. But what if one wants to search for RNAs for which there are few examples or none at all? At first this sounds impossible. How can one search for RNAs without knowledge of their primary sequence motifs or secondary structure? Yet it is possible to design such searches and, remarkably, in the last year such methods have succeeded in finding many new ncRNAs.

Algorithms to find completely unknown RNAs have fallen into three classes. The first group are based on finding stable secondary structures. Others exploit variations in ncRNA base-composition relative to the genomic background . Finally, some use genomic sequence comparisons among related species.

Structure- based de-novo *gene-finding*

The first *de novo* methods were based on secondary-structure computations. These methods exploited the observation that calculated thermodynamic free energies of ncRNAs are generally lower than those of random sequences with the same base composition. Hence the gene-finding program would segment the genome into fragments of the typical ncRNA-length (e.g. 100 or 200 base pairs) and compare the computed minimum free energies of the sequence fragments with randomized versions of the same sequences. If the calculated free energy of the sequence was significantly less than that of the randomly shuffled sequences, then one would predict the presence of an ncRNA gene.[55,56] One example of this approach was the program of Chen et al that searched for RNA pseudoknots.[57] Alternately one could look for genomic sequence fragments capable of being folded into specific types of RNA secondary structures in the spirit of RNA-folding programs such as Mfold[58] and ViennaRNA.[59] Looking for folding patterns – rather than computing free energies – had the advantage of generally being faster, while producing similar predictions.[60] Unfortunately, these methods have not led to the discovery of new ncRNAs. Moreover, computer experiments with known targets and randomized sequences suggested that secondary-structure computations by themselves would never be successful for *de novo* ncRNA gene-finding.[60]

Gene-finders using base-composition variations

The same paper[60] which demonstrated the limitations of using secondary-structure alone to search for ncRNAs, also suggested that (G+C)%, i.e. the percent of G and C bases in a sequence, might serve as a signature for the presence of an ncRNA gene. Subsequently, three groups[61-63] have successfully applied this idea to *de novo* gene-finding. Klein et al[61] and Schattner [62] searched for ncRNAs in thermophillic archaebacteria with high (A+T)% genomic backgrounds. Klein et al combined (G+C)% with the QRNA comparative genomics method[61] described below to search for ncRNAs in *M. jannaschii* and *P. furiosus*. Schattner examined variations in multiple base-composition statistics including (G+C)% , (G-C)% difference and dinucleotide frequency variations. Among these statistics (G+C)% and the frequency in the 'CpG' dinucleotide were observed to vary significantly between ncRNAs and the genome in the thermophile *M. jannaschii*. (Although the increased occurrence of CpG dinucleotides in *M. jannaschii* ncRNAs is reminiscent of the CpG islands of mammalian protein-coding-gene regions, there is currently no evidence indicating that they are in any way related.) Predictions from the two investigations in *M. jannaschii* were similar, though not identical. Northern blots performed by Klein[61] showed that 4 of the 6 *M. jannaschii* ncRNAs predicted by both approaches are in fact expressed. In addition, Klein et. al. predicted and experimentally verified 7 new ncRNAs in *P. furiosus.*

One of the conclusions of these two groups - that base-composition oriented gene-finding is primarily useful only with thermophiles - is somewhat discouraging, Nevertheless, the authors did suggest ways that the method may have wider applicability. Klein et al noted that ncRNAs may be found in non-thermophiles by first finding their homologs in a thermophillic species. Schattner observed that even in some non-thermophiles, such as *Caenorhabditis elegans,* significant base-composition variations between ncRNAs and the background exist. Although in *C. elegans* these variations are not sufficient to serve as an *de novo* gene-finder by themselves (as seen in figure 2), they may still be useful as a supplementary component of a gene-finder that also includes secondary-structure or primary-sequence motifs.

RNAGenie, developed by Carter et al,[63] incorporates base-composition variations – along with primary sequence motifs and free-energy calculations - in a "neural network" ncRNA gene-finder. Their work is particularly interesting since it was applied to *E. coli* and other non-thermophiles that might not be expected to be good candidates for a (G+C)% based gene-finder. Using their method, Carter et al find 370 putative novel ncRNAs in *E. coli*. Although no experimental testing of these candidate RNAs was performed in their original work, 13 of their predictions have been subsequently confirmed with Northern analysis.[63] In addition,, Carter et al listed 10 previously known

ncRNAs which had not been in their training set, seven of which were successfully found by RNAGenie. However, until a systematic verification of their predictions is performed, it will be unclear how many of their remaining 350 candidates are true ncRNAs and how many are simply false positives.

Gene-finding using comparative genomics

Perhaps the most exciting development in the area of *de novo* ncRNA gene-finding has come from three recent studies based on comparative genomics.[1,2,66] Each of these methods looks for regions of homology among two or more related genomes. The idea is that regions of biological importance – e.g. loci of ncRNAs – will be more conserved than regions that do not have any genes. So far these algorithms have been applied only to intergenic regions to avoid the large number of false positives likely to arise from homologous protein-coding genes.

In the method of Wasserman et al,[1] local cross-species sequence conservation was the only bioinformatic signature used. Since this resulted in a large number of putative "hits", Wasserman et al complemented their computational search with an experimental screen using micro-arrays. When applied to the *E. coli* genome (with comparisons to 5 related bacterial genomes), their method predicted 60 new ncRNAs of which 18 have been experimentally confirmed.[65]

In contrast, Argaman et al[66] combined the search for conserved sequences among bacterial genomes with a computational screen for nearby promoter and terminator sequence motifs. Since Argaman et al were searching in *E. coli* – where promoter and terminator sequence motifs are known and relatively well conserved – the method worked well. They made 24 predictions of novel ncRNAs of which 14 have been experimentally verified.[65] However, their method is difficult to apply to non-bacterial genomes for which promoter and terminator sequences are much less well conserved, or even to other bacterial genomes that have different promoter signatures.

Probably the most promising of the comparative approaches is the QRNA program of Rivas and Eddy.[2,64] Their method not only searches for regions of cross-species homology, but also examines the nature of the mismatches that occur among the aligned sequences. The key idea, illustrated in figure 3, is that if a region contains a protein-coding gene, then mismatches between homologous sequences should frequently correspond to a synonymous codon or a codon for a closely related amino acid. In contrast, if the region contains an ncRNA, then a higher percentage of substitutions should occur in complementary pairs such that the underlying ncRNA secondary structure is preserved despite the substitutions of the individual bases. Finally, if the region does not contain any gene, then the distribution of interspecies mismatches should

correspond to their background base frequencies. The appeal of this approach is that it should apply to any genome for which related sequenced genomes are available. Knowledge of promoter and terminator consensus sequences is not required. And since QRNA uses comparative information specific to RNA secondary-structure (in contrast to the methods of ref. 1 or ref. 66), it may be able to find ncRNAs while searching an entire genome - and not just the intergenic regions. QRNA has already been applied successfully to the *E. coli*, [2] *M. jannaschii* and *P. furiosus*[61] genomes and shows promise to being applicable to a wide range of additional genomes. Of course, since QRNA relies on secondary-structure signatures, it will have difficulty finding ncRNAs that have little or no secondary structure.

Current status and future prospects for computational ncRNA gene-finding

So just how good are the current computational gene-finders? And what are the prospects for improvement in the near future? As we have seen, in a few cases - such as tRNAs or methylation-guide snoRNAs – current programs work very well. However, in most cases the accomplishments have been more modest. For example, from our knowledge of the number of pseudouridylation sites in eukaryotic rRNA alone, we can be almost certain that there are dozens to hundreds, as yet unidentified, snoRNAs in essentially all eukaryotic genomes. For other classes of RNAs where we have no reliable information as to the total number of RNAs, we simply don't know whether the current programs are performing well or not. For example, were the 31 novel ncRNAs recently found in *E coli* [65] almost all of the ncRNAs that had previously escaped detection? Or are they only the "tip of the iceberg"? Estimates of the true number of *E coli* ncRNAs vary, but in reality the performance of current ncRNA gene-finders is still unknown.

In traditional machine-learning analysis, algorithm performance is generally evaluated by dividing the known examples into "training" and "testing" data sets. The program under evaluation is trained solely using the training data set and assessed with the testing data set. However, when few examples are known (the usual situation for ncRNA gene-finders), a modified procedure, "jack-knife testing" is typically used instead. In jack-knife testing, a single known example is sequentially removed from the training set and the program is trained with the remaining data. The program is then tested on its ability to find the omitted example. While eliminating the pitfalls of using the same data for training and testing, jack-knife testing still assumes that the known examples adequately represent the range of targets remaining to be found; otherwise, conclusions from jack-knife testing may be misleading. For example, Carter et al [63] demonstrated using jack-knife testing that RNAGenie had between 90.9% to 93.8% sensitivity in *E coli* depending on the threshold parameter they chose (see table 2 of ref. 63). On the other

hand, of the 31 novel ncRNAs found by refs. 1,2, and 66, only 13 were predicted by RNAGenie[63] suggesting a lower sensitivity at finding new ncRNAs.

These observations remind us that only experimental testing can confirm or refute the predictions of a computational gene-finder. Yet even when one uses experimental verification (e.g. Northern analysis or micro-array data) to assess computational gene-finders, one must proceed with caution. A negative result may simply indicate that the ncRNA isn't expressed under the specific cellular environment or the specific tissue type being assayed. On the other hand, even a positive identification on a Northern or micro-array may merely represent the presence of some other stable RNA such as an mRNA leader sequence. Of course, these experimental issues are not insurmountable and can be addressed by careful testing over multiple tissue types and cell environments .[3,4] However, they do remind us that even experimental results – when based on limited data – may not be sufficient to assess the performance of a computational gene-finder.

Despite these caveats, we may speculate a little on the future of ncRNA gene-finding. My belief is that all three classes of gene-finders - the customized gene-finders, the user-configurable motif-finders and the *de novo* programs – will continue to be important and will be used in a synergistic manner in the next few years. As additional genomes are sequenced, the comparative genomics gene-finders will be able to generate additional candidate ncRNAs. These candidates, along with those identified by the experimental screens,[3,4] will produce additional examples from the known classes of ncRNAs. These new examples will provide additional training data for the custom gene-finders, thereby improving their performance. Meanwhile better RNA-sequence-alignment and structure prediction programs should generate improved models of previously unknown classes of RNAs which can, in turn, be input into the reconfigurable RNA-motif-finders.

So, in the end, how many ncRNAs can we expect to find? One tantalizing hint may have come form the recent publication [67] of human-mouse sequence comparisons. This work showed that 66% of the strongly conserved, syntenic regions between mouse chromosome 16 and the corresponding human chromosomes do not overlap protein-coding exons. Intriguingly, the average length of these conserved syntenic regions is 189 base pairs[67]. How many of them encode ncRNA genes? No one knows. However, with the computational and experimental screening methods already available, the answer should become apparent soon. And if even a fraction do prove to be ncRNAs – as suggested by the recent screens in *E coli*[1,2,66] and other organisms[3,4] – then locating all these new ncRNAs should provide a window into an exciting modern RNA world far richer than many had previously believed it to be.

ACKNOWLEDGEMENTS:

I am grateful to Dr. Jan Barciszewski for encouraging me to write this article and Dr. Todd Lowe for a critical reading of the manuscript. I would also like to thank Drs. Sean Eddy and Daniel Gautheret for helpful comments on the manuscript.

REFERENCES

1. Wassarman KM, Repoila F, Rosenow C, et al Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev. 2001; 15: 1637-51.
2. Rivas E, Klein RJ, Jones TA, Eddy SR. Computational identification of noncoding RNAs in *E. coli* by comparative genomics Curr Biol. 2001; 11: 1369-73.
3. Huttenhofer A, Kiefmann M, Meier-Ewert S, et al. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. EMBO J. 2001;20:2943-53.
4. Tang TH, Bachellerie JP, Rozhdestvensky T, et al Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus.Proc Natl Acad Sci U S A. 2002; 99:7536-41.
5. Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.Bioinformatics. 1999 ;15:446-54.
6. Holmes I, Rubin GM.Pairwise RNA structure comparison with stochastic context-free grammars.Pac Symp Biocomput. 2002;:163-74
7. Mathews DH, Turner DH. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences.J Mol Biol. 2002 ;317:191-203
8. Gorodkin J, Heyer LJ, Stormo GD.Finding the most significant common sequence and structure motifs in a set of RNA sequences.Nucleic Acids Res. 1997 ;25:3724-32.
9. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs.Genome Biol. 2002;3(3):
10. Stormo GD. Gene-finding approaches for eukaryotes.Genome Res. 2000;10:394-7.
11. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol. 1990; 215: 403-10.
12. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA.J Mol Biol. 1997;268:78-94.
13. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models.Nucleic Acids Res. 1998;;26:544-8.
14. Kulp D, Haussler D, Reese MG, Eeckman FH. A generalized hidden Markov model for the recognition of human genes in DNA.Proc Int Conf Intell Syst Mol Biol. 1996;4:134-42
15. Durbin R, Eddy SR, Krogh A et al Biological Sequence analysis 1998; Cambridge Unversity Press
16. Gribskov M, McLachlan A, Eisenberg D Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A. 1987; 84:4355-8

17. Staden R. A computer program to search for tRNA genes. Nucleic Acids Res. 1980;8:817-25
18. Margalit H, Shapiro BA, Oppenheim AB, Maizel JV Jr. Detection of common motifs in RNA secondary structures. Nucleic Acids Res. 1989 ;17:4829-45.
19. Marvel CC. A program for the identification of tRNA-like structures in DNA sequence data. Nucleic Acids Res. 1986;14:431-5.
20. Fichant GA, Burks C. Identifying potential tRNA genes in genomic DNA sequences. J Mol Biol. 1991; 220: 659-71.
21. Pavesi A, Conterio F, Bolchi A, et al Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. Nucleic Acids Res. 1994; 22: 1247-56.
22. Data taken from Table 1 of ref 25.
23. Eddy SR, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Res. 1994; 22:2079-88.
24. el-Mabrouk N, Lisacek F. Very fast identification of RNA motifs in genomic DNA. Application to tRNA search in the yeast genome. J Mol Biol. 1996;264:46-55.
25. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997; 25: 955-64.
26. Sprinzl M, Horn C, Brown M, et al Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. 1998; 26, 148-53
27. Nicoloso M, Qu LH, Michot B, Bachellerie JP. Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2'-O-ribose methylation of rRNAs. J Mol Biol. 1996;260:178-95.
28. Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. Science 1999; 283: 1168-71.
29. Omer AD, Lowe TM, Russell AG, et al. Homologs of small nucleolar RNAs in Archaea. Science. 2000;288:517-22.
30. Dandekar T, Sibbald PR Trans-splicing of pre-mRNA is predicted to occur in a wide range of organisms including vertebrates. Nucleic Acids Res. 1990;1816:4719-25.
31. Lisacek F, Diaz Y, Michel F Automatic identification of group I intron cores in genomic DNA sequences. J Mol Biol. 1994; 2354;:1206-17
32. Laslett D, Canback B, Andersson S. BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. Nucleic Acids Res. 2002;30:3449-53.
33. Staden R. Methods to define and locate patterns of motifs in sequences. Comput Appl Biosci. 1988;4:53-60.
34. Saurin W, Marliere P. Matching relational patterns in nucleic acid sequences.Comput Appl Biosci. 1987 Jun;3(2):115-20.
35. Gautheret D, Major F, Cedergren R. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. Comput Appl Biosci. 1990;64;:325-31.

36. Laferriere A, Gautheret D, Cedergren R.An RNA pattern matching program with enhanced performance and portability. Comput Appl Biosci. 1994;102;:211-2.
37. Eddy S.R. RNABOB, http://www.genetics.wustl.edu/eddy/software/#rnabob
38. Winker S, Overbeek R, Woese CR, et al Structure detection through automated covariance search. Comput Appl Biosci. 1990;64;:365-71
39. Pesole G, Liuni S, D'Souza M. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance.Bioinformatics. 2000;16:439-50.
40. Billoud B, Kontic M, Viari A. Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database.Nucleic Acids Res. 1996;24:1395-403.
41. Macke TJ, Ecker DJ, Gutell RR, et al. RNAMotif, an RNA secondary structure definition and search algorithm.Nucleic Acids Res. 2001;29:4724-35.
42. Brendel V, Busse HG. Genome structure described by formal languages. Nucleic Acids Res. 1984;12:2561-8.
43. Dong S, Searls DB. Gene structure prediction by linguistic methods. Genomics. 1994;23:540-51.
44. Sakakibara Y, Brown M, Hughey R, et al Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res. 1994;2223;:5112-20.
45. Ref 15, chapter 10.
46. Eddy S.R. A memory efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure, BMC Bioinformatics, in press, 2002
47. Gautheret D, Lambert A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles.J Mol Biol. 2001;313:1003-11
48. Gorodkin J, Stricklin SL, Stormo GD Discovering common stem-loop motifs in unaligned RNA sequences.Nucleic Acids Res. 2001;29:2135-44.
49. Gaspin C, Cavaille J, Erauso G, Bachellerie JP Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. J Mol Biol. 2000;297:895-906.
50. Ferbeyre G, Smith JM, Cedergren R. Schistosome satellite DNA encodes active hammerhead ribozymes. Mol Cell Biol. 1998 Jul;18(7):3880-8.
51. Bourdeau V, Ferbeyre G, Pageau M, et al. The distribution of RNA motifs in natural sequences. Nucleic Acids Res. 1999 ;27:4457-67.
52. Ferbeyre G, Bourdeau V, Pageau M, et al. Distribution of hammerhead and hammerhead-like RNA motifs through the GenBank. Genome Res. 2000;10:1011-9.
53. Lescure A, Gautheret D, Carbon P, Krol A. Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif.J Biol Chem. 1999 ;274:38147-54.

54. Dandekar T, Hentze MW Finding the hairpin in the haystack: searching for RNA motifs. Trends Genet. 1995;112;:45-50.
55. Le SY, Chen JH, Braun MJ, et al Stability of RNA stem-loop structure and distribution of non-random structure in the human immunodeficiency virus (HIV-I). Nucleic Acids Res. 1988 10;:5153-68.
56. Le SY, Chen JH, Currey KM, et al A program for predicting significant RNA secondary structures. Comput Appl Biosci. 1988;4:153-9.
57. Chen JH, Le SY, Maizel JV. A procedure for RNA pseudoknot prediction. Comput Appl Biosci. 1992;8:243-8.
58. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981 10;9:133-48.
59. Schuster P, Fontana W, Stadler PF, et al .From sequences to shapes and back: a case study in RNA secondary structures. Proc R Soc Lond B Biol Sci.1994; 255:279-84.
60. Rivas E, Eddy SR. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics 2000; 16: 583-605.
61. Klein RJ, Misulovin Z, Eddy SR. Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc Natl Acad Sci U S A. 2002 ;99:7542-7.
62. Schattner P. Searching for RNA genes using base-composition statistics. Nucleic Acids Res. 2002;30:2076-82.
63. Carter RJ, Dubchak I, Holbrook SR A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res. 2001 Oct 1;2919;:3928-38
64. Rivas E, Eddy SR Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics. 2001;21;:8.
65. Eddy SR. Computational genomics of noncoding RNA genes. Cell. 2002 ;109:137-40., Table 1
66. Argaman L, Hershberg R, Vogel J, et al Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. Curr Biol. 2001; 11:941-50.
67. Mural R, Adams M, Myers E et al A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 2002; 296:1661-71.

FIGURE CAPTIONS:

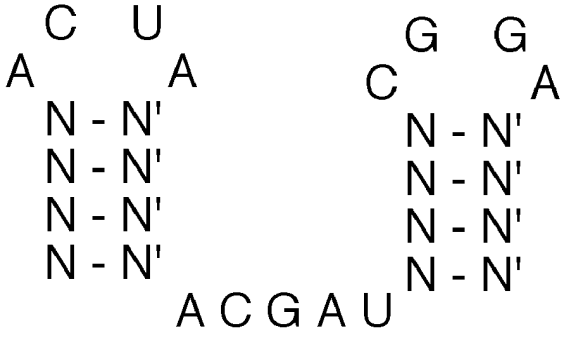
Figure 1: A simple hypothetical double stem loop as described by a descriptor file and a Context Free Grammar.
Figure 1A graphical representation of secondary structure. N indicates any one of the four bases. N' is the complement of N.
Figure 1B Description of structure using the RNAMOT file descriptor language.
Figure 1C Structure description using production rules of a Context-Free Grammar. Capital letters indicate CFG "Non-terminals". Note: The production rules for the second stem loop are omitted for brevity. For more details on the file descriptor languages see refs. 35-40. For more details on CFGs see ref 43 and chapter 10 of ref. 15

Figure 2: Separation of RNA's and genomic background using G+C%. Vertical axes indicate estimated relative number of 100 bp subsequences. Note that peak of curve for number of genomic sequences is truncated. RNA estimate assumes ratio of protein coding genes to ncRNA genes is approximately equal to that in *S. cerevisae*. Graphs are shown as normally distributed for purpose of illustration - actual distribution of G+C% may vary.
Figure 2A. *M. jannaschii* RNA and genome G+C% distributions are separated enough to enable discrimination between RNA and background populations
Figure 2B. *C. elegans* chromosome X. Although *C. elegans* ncRNA and genomic G+C% population means are significantly different, ncRNA distribution can not be distinguished from that of the background by G+C% alone.
(modified from ref 62 by permission, copyright 2002 Oxford University Press.)

Figure 3: QRNA sequence alignment for protein-coding, structural-RNA-coding and non-coding sequences. Three pairwise alignments of identical composition with identical base substitutions can be classified by distinctive patterns of mutation caused by different selective constraints. The figure indicates how each alignment is scored according to the model that best fits the pattern of mutations: one position at a time for the position-independent model, one codon at a time for the protein-coding model (integrated overall six possible reading frames) and as a combination of base paired positions and single positions for RNA (integrated over all possible secondary structures). For more details see ref. 64.
(modified from ref 64 by permission, copyright 2001 Sean R. Eddy.)

Figure 1: A simple hypothetical double stem loop as described by a descriptor file and a Context Free Grammar

A:



B:

```
H1  s1  H1  s2  H2  s3  H2

H1  4:4  0
s1  A C U A
s2  A C G A U
H2  4:4  0
s3  C G G A
```

C:

$B \rightarrow H_A \; S \; H_B$

$H_A \rightarrow a\,H_{A1}\,u \mid c\,H_{A1}\,g \mid g\,H_{A1}\,c \mid u\,H_{A1}\,a$

$H_{A1} \rightarrow a\,H_{A2}\,u \mid c\,H_{A2}\,g \mid g\,H_{A2}\,c \mid u\,H_{A2}\,a$

$H_{A2} \rightarrow a\,H_{A3}\,u \mid c\,H_{A3}\,g \mid g\,H_{A3}\,c \mid u\,H_{A3}\,a$

$H_{A3} \rightarrow a\,L\,u \mid c\,L\,g \mid g\,L\,c \mid u\,L\,a$

$L \rightarrow a\,L_1 \; ; L_1 \rightarrow c\,L_2 \; ; L_2 \rightarrow u\,L_3 \; ; L_3 \rightarrow a$

$S \rightarrow a\,S_1 \; ; S_1 \rightarrow c\,S_2 \; ; S_2 \rightarrow g\,S_3 \; ; \; S_3 \rightarrow a\,S_4 \; ; S_4 \rightarrow u$

Figure 2: Separation of RNA's and genomic background using G+C%.

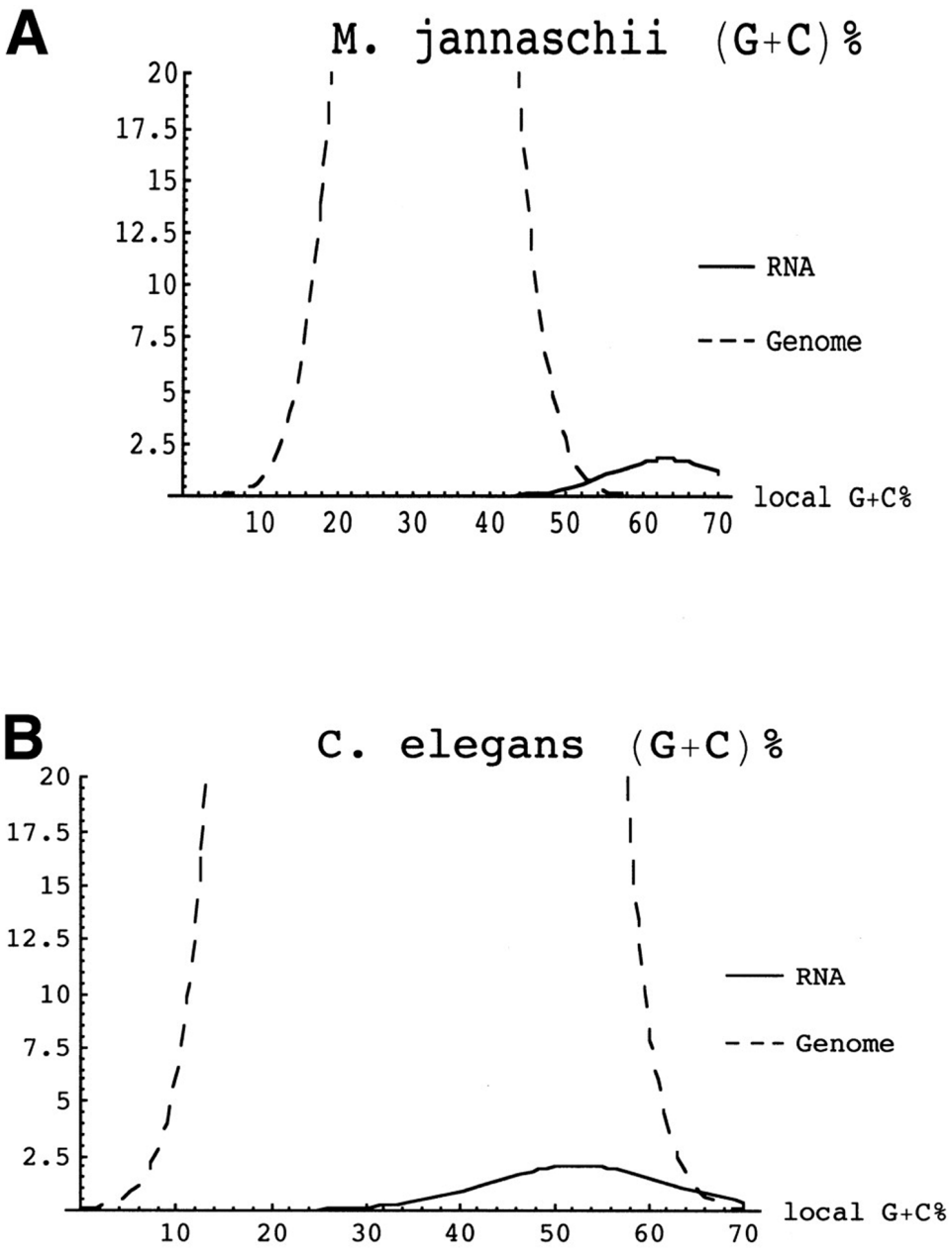


Figure 3: QRNA sequence alignment for protein-coding, structural-RNA-coding and non-coding sequences