BIOINFORMATICS

Vol. 19 no. 7 2003, pages 865–873
DOI: 10.1093/bioinformatics/btg080

A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction

Sverker Edvardsson[1,†], Paul P. Gardner[2,4,†], Anthony M. Poole[3,4], Michael D. Hendy[2,4], David Penny[3,4] and Vincent Moulton[5,*]

[1]Department of Information Technology, Mid Sweden University, S-851 70, Sundsvall, Sweden, [2]Institute of Fundamental Science, [3]Institute of Molecular BioScience, [4]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand and [5]The Linnaeus Centre for Bioinformatics, Uppsala University, BMC Box 598, S-751 24, Uppsala, Sweden

Received on June 7, 2002; revised on November 1, 2002; accepted on November 26, 2002

ABSTRACT

Motivation: Noncoding RNA genes produce functional RNA molecules rather than coding for proteins. One such family is the H/ACA snoRNAs. Unlike the related C/D snoRNAs these have resisted automated detection to date.

Results: We develop an algorithm to screen the yeast genome for novel H/ACA snoRNAs. To achieve this, we introduce some new methods for facilitating the search for noncoding RNAs in genomic sequences which are based on properties of predicted minimum free-energy (MFE) secondary structures. The algorithm has been implemented and can be generalized to enable screening of other eukaryote genomes. We find that use of primary sequence alone is insufficient for identifying novel H/ACA snoRNAs. Only the use of secondary structure filters reduces the number of candidates to a manageable size. From genomic context, we identify three strong H/ACA snoRNA candidates. These together with a further 47 candidates obtained by our analysis are being experimentally screened.

Contact: vincent.moulton@lcb.uu.se

Supplementary Information: Tables 1–5 referred to in the text can be downloaded from http://RNA.massey.ac.nz/fisher/

INTRODUCTION

The number of genes identified that code for noncoding RNAs is growing rapidly (Eddy, 2001; Erdmann *et al.*, 2001; Meli *et al.*, 2001). While labor-intensive molecular biological approaches have been successful in identifying noncoding RNAs (Hüttenhofer *et al.*, 2001; Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee and Ambros, 2001), it is preferable to carry out initial RNA gene prediction *in silico*, as is common with protein-coding genes, e.g. (Delcher *et al.*, 1999).

Standard search methods such as BLAST (Altschul *et al.*, 1990) have been used in comparative searches of bacterial genomes for novel RNAs (Argaman *et al.*, 2001; Rivas *et al.*, 2001; Wassarman *et al.*, 2001) and in searches for novel small regulatory RNAs in animals and invertebrates (Pasquinelli *et al.*, 2000). In addition, programs for RNA gene finding are available; for example, the programs `tRNAscan-SE` (Lowe and Eddy, 1997), `QRNA` (Rivas and Eddy, 2001; Rivas *et al.*, 2001), and `RNAMotif` (Macke *et al.*, 2001) have been successfully applied in whole genome searches for novel RNAs.

The importance of primary sequence for the finding of new RNAs is clear, and was employed heavily in a comparative search for noncoding RNAs in *E.coli* (Rivas *et al.*, 2001). However, in general, standard homology searches are not suitable for finding RNAs. Thus successful searches have tended to use techniques such as neural networks (Carter *et al.*, 2001), pattern-based descriptors (Macke *et al.*, 2001) and covariance models (Eddy and Durbin, 1994; Lowe and Eddy, 1997, 1999) which incorporate RNA secondary structure information.

In this paper we investigate an alternative approach for incorporating secondary structure information into RNA searches. Secondary structure is amenable to mathematical analysis making minimum free-energy (MFE) structure prediction using algorithms such as dynamic programming possible. In consequence programs such as `VIENNA` (Hofacker *et al.*, 1994) and `Mfold` (Zuker *et al.*, 1999) can quite accurately predict secondary structure. Even so, Rivas and Eddy (2000) determined that a general search for noncoding RNAs in genomes

*To whom correspondence should be addressed.
† Both authors contributed equally to this work.

using MFE structure stability alone is unlikely to succeed since background noise is too high.

However, in (Collins *et al.*, 2000) the discovery of an RNase P candidate in the maize chloroplast genome was detected using an *ad hoc* combination of comparative genomics and MFE structure comparison. Encouraged by this result, we developed the RNA shape comparison techniques described in (Moulton *et al.*, 2000) and incorporated them into an algorithm that we present here which screens the budding yeast *Saccharomyces cerevisiae* genome (Goffeau *et al.*, 1996) for H/ACA snoRNAs. Our method is similar to that used by Lowe and Eddy (1999) in their successful computational screen of the *S.cerevisiae* genome for the related C/D snoRNAs, which employed a probabilistic model as opposed to MFE structure prediction.

METHODS

Our search strategy for novel snoRNAs in the *S.cerevisiae* or yeast genome uses known H/ACA snoRNAs to form primary and secondary structure models. Then we make a sequential search for novel snoRNAs in both directions of the yeast genome, passing candidate sequences obtained with the primary structure search through various secondary structure filters. The sequences that pass through all of these filters are then scored using both primary and secondary structure information.

Training data set

SnoRNAs (small nucleolar RNAs) are named because of their localization to the eukaryote cell nucleolus. They fall into two families, the C/D box family and the H/ACA box family (reviewed in Weinstein and Steitz, 1999). Within the H/ACA family there is significant conservation of predicted MFE secondary structures, but very limited conservation of primary sequence Ganot *et al.*, 1997a,b).

The H/ACA box family guide site-specific isomerization of rRNA (Ni *et al.*, 1997; Ganot *et al.*, 1997a), whereby uridine (U) is converted to pseudouridine (Ψ) (reviewed by Ofengand and Fournier, 1998), see Figure 1. To date, 44 pseudouridines have been identified on yeast rRNAs (Table 1) and 17 H/ACA snoRNAs have been shown to guide 21 of these (Ofengand and Fournier, 1998; Samarsky and Fournier, 1999). Based on this data we suspect that perhaps 10–20 yeast H/ACA snoRNAs have yet to be identified.

We obtained a dataset of 16 yeast H/ACA snoRNA sequences from the Yeast SnoRNA Database (Samarsky and Fournier, 1999). These had been identified primarily by biochemical techniques (Ganot *et al.*, 1997a; Ni *et al.*, 1997) and are provided with demonstrated or predicted locations for H and ACA motifs and rRNA interactions. The sequences flanking the pseudouridylation sites in rRNA are obtained from (Ofengand and Fournier, 1998)

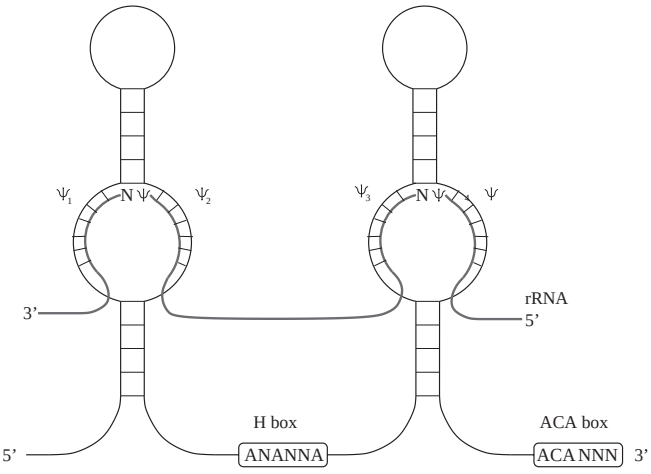

Fig. 1. Schematic of the consensus primary and secondary structural elements of the H/ACA box snoRNA. Note the hairpin-hinge-hairpin-tail secondary structure and the internal loop structures termed the pseudouridylation pockets (Ganot *et al.*, 1997b). The interaction of these pockets with rRNA is also shown. Ψ_i refers to the parts of the snoRNA that are complementary to the rRNA.

where information regarding pseudouridines in yeast rRNA is presented. We did not include snR9, snR30 or snR37 in our training data. For snR9, no capacity for guiding pseudouridylation has been assigned, and snR30 is involved in rRNA cleavage, not pseudouridylation. snR37 is 386nt long and does not compare well with the snoRNAs in the training set.

Primary structure search

The primary structure search algorithm sequentially identifies parts of the yeast genome harboring various primary structural motifs, separated as detailed in Figure 2. The algorithm first searches for an H-box. This motif is a sequence of the form $AN_1AN_2N_3N_4N_5$ with $N_i \in \{A, U, C, G\}$, $N_1 \neq C$, $N_3 \neq G$, and either $N_4 = A$ or $N_5 = A$. Once a candidate H-box is identified, it is scored using a probabilistic model that we constructed using the snoRNA dataset. In particular, we compute a similarity score between the putative H-box and each of the known H-boxes (presented in Table 2) using the frequencies of nucleotides at positions $(N_1N_2N_3N_4N_5)$ (presented in Table 3). The similarity between the putative H-box and each known H-box is computed as follows; the two sequences are placed one above the other, matches are given a score of 200, mismatches are scored according to the nucleotide frequencies at positions $(N_1N_2N_3N_4N_5)$ (e.g. if the putative sequence has a G in position N_1 which mismatches it is scored 81.25) and the scores are added, in a similar fashion to the profile matrix method used by PSI-BLAST (Altschul *et al.*, 1997). If the maximum

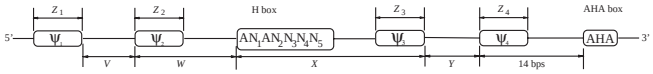

Fig. 2. Primary structure model used to search for putative snoRNAs consisting of an H-box, an ACA-box (here denoted AHA—see text) and four regions of complementarity to the rRNA subsequences flanking some pseudouridylation site on rRNA (denoted by $\Psi_1\Psi_2$ and $\Psi_3\Psi_4$). Our model requires: $X + Y + 14 \leqslant 142$; $16 \leqslant X \leqslant 70$; $Y \geqslant 30$; $3 \leqslant Z_3, Z_4 \leqslant 10$; $Z_3 + Z_4 \geqslant 9$; $20 \leqslant V \leqslant 100$; $11 \leqslant W \leqslant 17$; $3 \leqslant Z_1, Z_2 \leqslant 10$; $Z_1 + Z_2 \geqslant 9$.

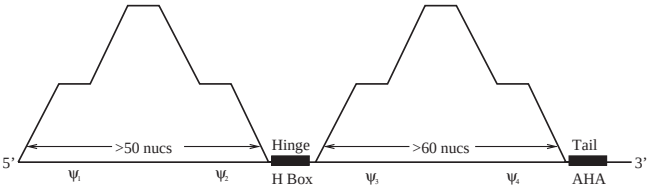

Fig. 3. Secondary structure model of H/ACA snoRNA. It consists of two 'mountains' with widths as indicated.

of these similarities exceeds the threshold value of 800 (obtained using a leave-one-out analysis), the H-box is accepted and this similarity score is recorded for the H-box. In addition, 200 is added to the similarity in case a complete match is obtained between the putative H-box and an H-box that occurs at least twice for the known snoRNAs (e.g. snR189 and snR34). Although such an H-box would be accepted without this bonus, the addition is made since the similarity is used later when scoring the final candidates.

After locating a high-scoring H-box, the algorithm searches downstream for Ψ_3 and Ψ_4 motifs. These are two sequences that are almost complementary to the sequences flanking a pseudouridylation site in the yeast rRNA, see Figure 1 (these motifs are listed in Table 1). Similar complementary motifs were also employed by Lowe and Eddy (1999) in their search for C/D snoRNAs. To look for a putative Ψ_3 motif, a known Ψ_3 motif is directly compared with the yeast genome. The comparison is considered a match if either the sequences are identical or there is at most one wobble, where a wobble corresponds to a C or an A in Ψ_3 lining up to an U or a G in the genome, respectively. The wobble corresponds to a non-canonical base pairing between the H/ACA snoRNA and the rRNA. Such pairings occur for the known snoRNAs. The same comparison is performed for the Ψ_4 motif. The lengths of the Ψ_3, Ψ_4 motifs (Z_3 and Z_4 in Figure 2), which were inferred by analyzing the snoRNA dataset, are required to be between three and ten bases, and the sum of their lengths must always exceed 8. If Ψ_3, Ψ_4 motifs are found in the correct locations (given by X and Y in Figure 2), then the algorithm continues to search for the ACA sequence. To reduce any confusion from now on we denote this sequence by AHA, where H can equal A, U or C. The AHA box is exactly 14 bases from the beginning of the Ψ_4 motif, a distance that is conserved for all known yeast snoRNAs (Ganot *et al.*, 1997b) and, if found, the complete H-AHA region is passed to the secondary structure filters described in the next section. Failure to locate a downstream motif in the above procedure in general results in a continuation of the sequential search for another H-box.

Secondary structure filters

The H-AHA region identified by the primary structure filter is passed through several secondary structure filters to reduce false positives.

A secondary structure model for yeast H/ACA snoRNA was derived using MFE structure prediction (Zuker and Steigler, 1981). For each known snoRNA two sequences consisting of the H-AHA region together with a sequence of length 100 or 120 bases upstream from the H-box were formed and then folded using the `RNAfold` function of the `VIENNA` v. 1.4 package (Hofacker *et al.*, 1994). The option '`no dangling ends`', improved the folds. Upstream lengths of 100 and 120 gave a good signal, even though these do not correspond exactly to those for the known snoRNAs. The dynamic length was necessitated because the 5′ end of a putative snoRNA sequence cannot be determined *a priori* in the yeast sequence.

The resulting structures were represented by mountain plots (see Moulton *et al.*, 2000), which are based on the representation of Hogeweg and Hesper (1984). This type of plot allows a simple connection between primary and secondary structure. The mountain plot consists of the points with x-coordinate k corresponding to the kth nucleotide and y-coordinate y_k equaling the number of base-pairs enclosing this nucleotide (see Figure 3). When we compare structures whose underlying sequences have different lengths, we normalize the corresponding mountain plots, scaling the x-coordinates to lie between 0 and 1 and the y-coordinates so that the total area under the graph equals one. In practice, mountain plots are represented by the vector containing the y-coordinates y_k corresponding to each nucleotide k, whereas normalized mountains plot are represented by vectors of a suitably large fixed length N, that contain the y-coordinates y_i of the normalized mountain plot at x-coordinates $\frac{i}{N}$, $1 \leqslant i \leqslant N$. To obtain these normalized vectors we employed splines.

Good similarity was observed between the normalized mountain plots of the known snoRNA dataset (Figure 4). The significant common structural features were incorporated into a secondary structure model consisting of two 'mountains' separated by a hinge region, the position of

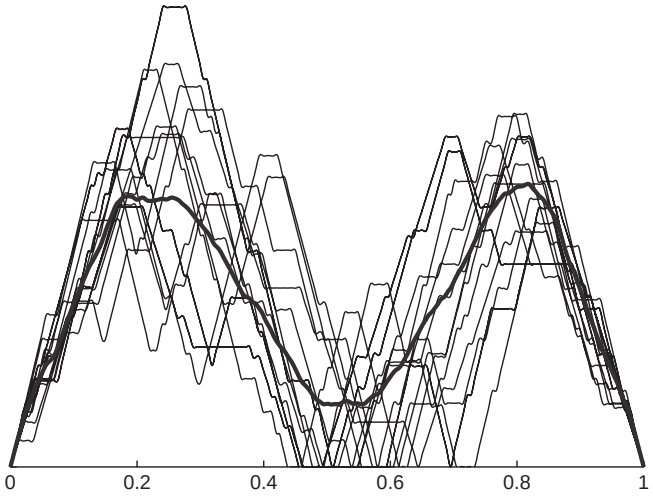

Fig. 4. Normalised mountain plots (L=100) for the 16 yeast snoRNAs. The thick line represents the mean structure for the whole snoRNA dataset.

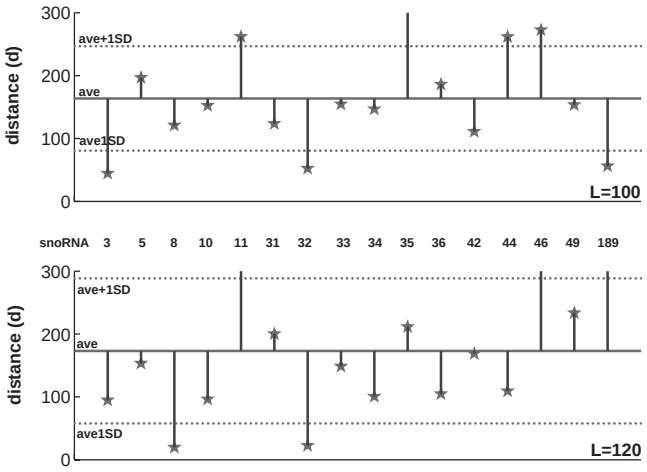

Fig. 5. Distance d from the 16 known snoRNAs to the mean snoRNA structure. The dotted lines represent the standard deviations (±1 SD). A candidate structure will pass if $d < 300$ for either $L = 100$ or $L = 120$.

which roughly corresponds to the H-box (Figure 3). As a preliminary coarse filter, the sequence comprising of the H-AHA region identified previously, with either $L = 100$ or 120 upstream bases, is folded. The resulting mountain plot is accepted only if it has a local minimum (corresponding to the hinge position) within ±11 bases of the H-box, the height of this minimum is at most 4 above the H-box, the width of the left mountain exceeds 50 bases (fulfilled automatically for the right mountain, see Figure 2), and above Ψ_3 and Ψ_4 the graph is high enough (>4) and also non-zero between these two motifs.

Those candidates displaying these coarse criteria are then passed through more-sensitive filters. The first filter computes a squared distance from its normalized mountain plot to a mean snoRNA structure ($d = \sum_{i=1}^{N}(y_i - \overline{y}_i^L)^2$, where y_i is the normalized structure and $\overline{y}_i^L$ is the mean normalized snoRNA structure taken over the training dataset). Figure 4 displays this mean snoRNA structure for the case $L = 100$. The distance d between a known snoRNA and the mean snoRNA is typically about 150 (see Figure 5) so that low values of d are not expected for candidate snoRNAs. Even though distances for candidate snoRNAs are expected to be about the same as for known snoRNAs (see Figure 5), a candidate snoRNA is still allowed to pass through this filter if $d < 300$.

A second filter uses the observation that known snoRNA structures whether obtained using the old (v1.3) or new (v1.4) folding parameters (Mathews *et al.*, 1999) provided in the `VIENNA` package were similar—a property that we did not observe in general for random sequences (data not shown). This may be because a small perturbation in parameters does not significantly change stable secondary structures. We implemented a stability filter that compares normalized mountain plots generated for candidate snoRNA sequences using both the old and new folding parameters. In particular, we compute the distances d_{old} and d_{new} for the 'old' and 'new' normalized mountain plots. Only candidate snoRNAs satisfying $|d_{old} - d_{new}| < 300$ are accepted.

Scoring the output

The last stage computes a score based on both primary and secondary structure for each candidate snoRNA. A score for the AHA-box is added to the H-box similarity score described earlier. The H in the AHA-box is scored according to: $A = 6.25$, $U = 18.75$, $C = 75$ and G is not allowed (based on frequencies from the training dataset; see Table 2). The scores are added as described in the section above and then transformed into a number $0 \leqslant P_1 \leqslant 1$.

As part of the score we also computed three other quantities P_2, P_3 and P_4 defined as follows (see Figure 2 and Table 4). If $X \leqslant 40$ then we put $P_2 = 1$, else $P_2 = 0.5$. Furthermore, if $66 \leqslant X + Y \leqslant 100$ then we put $P_3 = 1$, else $P_3 = 0.5$. The score is also based on the performance of the secondary structure. The average distance $\overline{d}$ is computed for the training dataset. For a putative snoRNA if $|d - \overline{d}| > \overline{d}$, then we put $P_4 = 0$, else $P_4 = (\overline{d} - |d - \overline{d}|)/\overline{d}$. Thus, the closer d is to the average $\overline{d}$, the higher the score. The putative snoRNA is only accepted if both $P_1 > 0.8$ and $P_4 > 0.65$ hold.

The total score P_{tot} of the candidate snoRNA is then computed using the formula

$$P_{tot} = 100 \left[\frac{w_1 P_1 + w_2 P_2 + w_3 P_3 + w_4 P_4}{w_1 + w_2 + w_3 + w_4} \right],$$

where w_1=10, w_2=2, w_3=1, and w_4=2. The values of the weights w_i were obtained by optimization using the Nelder–Mead method (see e.g. Kelley, 1999) on the training dataset. Only if $P_{tot} > 70$ is the structure accepted. All snoRNAs in the training dataset satisfy $P_{tot} \gg 70$ (see Figure 6).

Final processing

In case we target snoRNAs also having the complementary pair Ψ_1 and Ψ_2, a special procedure is called. This looks for the motifs Ψ_1 and Ψ_2, that fulfill the criteria $V \geqslant 20$ (the majority have V in the range 30–40) and $11 \leqslant W \leqslant 17$ (the majority have $W = 14$)—see Figure 2 and Table 4.

RESULTS

We have implemented the strategies and filters described above in a C program (`Fisher`). This is available via electronic mail [sverker.edvardsson@mh.se].

Ψ-pair assignments

In the known yeast H/ACA snoRNA dataset, no two snoRNAs have been demonstrated to guide the same pseudouridylation (Table 1). However, of the snoRNAs in our dataset, only 13 of 22 assigned Ψ-pairs perform the corresponding pseudouridylation (see Table 2) and snR3 is potentially capable of more than one pseudouridylation at the 3′ pocket ($\Psi_3\Psi_4$). We therefore examined the known snoRNAs for redundancy, using our primary structure engine to search for all $\Psi_1\Psi_2$- and $\Psi_3\Psi_4$-pairs within these (see Table 5). We used the following constraints: $25 \leqslant V \leqslant 45$ and $13 \leqslant W \leqslant 16$ (see Figure 2). Our algorithm locates all the assigned Ψ-pairs except 39, corresponding to snR34 (Table 2). The reason is the unusually large distance $W = 38$. A potential stem involving $24 + 2$ bases lies between snR34's Ψ_2 and the H-box. Despite this feature, it is reasonable to assume that functionally important spatial determinants are preserved (W. Decatur, J. Ni, and M. Fournier, pers. commun.). This is the only such situation known to exist for the yeast snoRNAs. Our examination of sequence complementarity between the rRNA and the $\Psi_1\Psi_2$ and $\Psi_3\Psi_4$ sequence pairs in the 5′ and 3′ pseudouridylation pockets of the known H/ACA snoRNAs reveals extensive potential for functional redundancy (Table 5). For instance, pseudouridylation of U_{1056} in the 25S rRNA subunit (23 in Table 1) is guided by snR44 (Ganot *et al.*, 1997a; Samarsky and Fournier, 1999), yet our analysis (Table 5) suggests that snR31, snR33, snR36 and snR49 are also potentially capable of guiding this pseudouridylation. Furthermore, we find that many of the known H/ACA snoRNAs can potentially guide more than 2 pseudouridylations.

A test scan through a randomized genome sequence

In order to investigate the performance of our search strategy with respect to false positives we created a randomized test genome sequence. To conserve the approximate frequencies of A, U, C and G, our test genome was created by copying a sequence of length 540 000 bases from the yeast genome. The sequence was shuffled using an algorithm that preserves dinucleotide frequencies (Workman and Krogh, 1999; Altschul and Erikson, 1985). For a sequence of length N, this is performed by randomly selecting pairs of triplets of the form XQY and XPY and then exchanging Q and P. This is repeated $10N$ times. We then added all 13 snoRNAs that have $\Psi_3\Psi_4$-pairs (Table 2) to create the final test genome. A complete scan through this genome took about a day on an AMD Athlon 1.4 Ghz which was reasonable for testing purposes.

The total number of hits obtained by the primary search was 66 600, which demonstrates that it is unrealistic to only consider the primary motifs of H/ACA snoRNAs. However, several of these were actually at the same H-box position. This redundancy occurs since the search can locate several different $\Psi_3\Psi_4$-pairs and AHA-boxes. For each H-box we only kept hits with highest primary score (i.e. $100(w_1P_1 + w_2P_2 + w_3P_3)/(w_1 + w_2 + w_3)$). After the initial secondary structure filters have been applied, we are left with 15 428 hits.

The scores P_{tot} for these hits are plotted in Figure 6. The squares indicate the scores obtained for the known snoRNAs, which were planted within the first 54 000 bases of the test genome. Out of the 15 428 hits, 2397 have total scores greater than 70. We observe in Figure 6 that the snoRNAs have scores well above most of the other hits. The snoRNA with the lowest score was ranked 192. Thus, to hit all of the known snoRNAs we need to accept 179 false positives. The final requirement that both ($P_1 > 0.8$) and ($P_4 > 0.65$) hold simultaneously, further reduced the number of false positives to 96. Thus, out of the 15 428 distinct hits, 96 false positives remained, giving a performance of $(15\,415 - 96)/15\,415 = 99.4\%$ (searching the reverse complemented test genome gave 99.3%).

Unfortunately, snR8 does not satisfy ($P_1 > 0.8$) and ($P_4 > 0.65$). Of course, this last filter could be relaxed in order to hit snR8, but then we would need to deal with many more false positives. After considerable testing, we concluded that the balance between the number of false positives and false negatives was acceptable.

Screening the yeast genome with Fisher

The yeast genome is approximately 12 Mb, which is about 20 times larger than our test genome, and we must search it

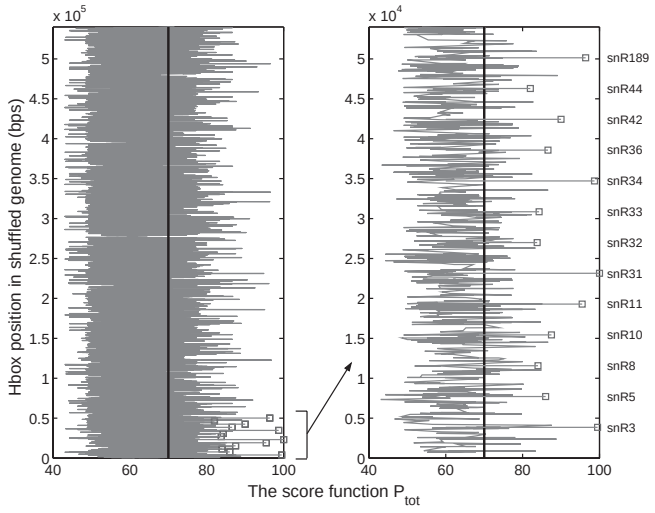

Fig. 6. The total score (P_{tot}) for the hits in the shuffled test genome. The left figure shows the whole test genome consisting of 540 000 bases. The snoRNAs were planted amongst the first 54 000 bases. This part is enlarged on the right. The total scores for the 13 Ψ_3 Ψ_4-snoRNAs are marked with squares. For a normal search we only accept hits with scores above 70 (marked with the bold line).

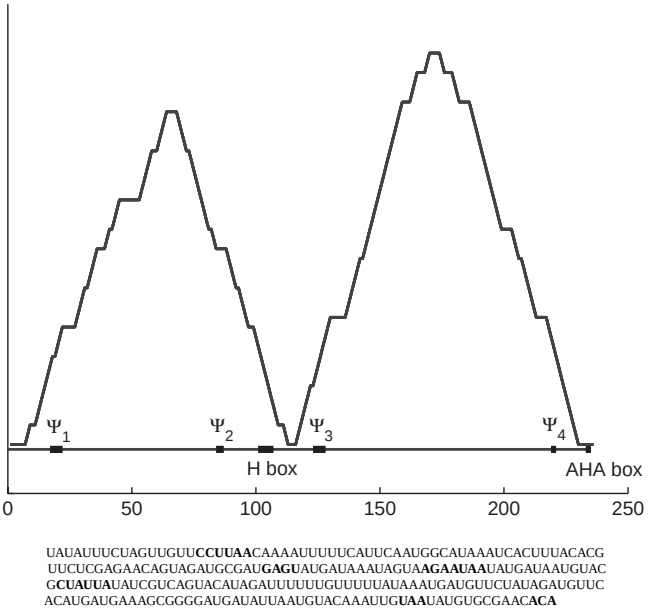

Fig. 7. An example of a typical hit in the yeast genome. This particular hit has both a $\Psi_1\Psi_2$ (32) and a $\Psi_3\Psi_4$-pair (2). Reading left to right, the bold motifs in the above sequence are: Ψ_1, Ψ_2, H-box, Ψ_3, Ψ_4 and AHA-box.

in both directions. Thus, from the results above we expect `Fisher` to yield perhaps ten quite highly ranked novel snoRNAs and about 4000 false positives.

In order to decrease this rather large number of expected false positives we created a reduced yeast genome sequence of approximately 3.5 Mb. This consisted of the NotFeature.fasta file (produced by removing all regions corresponding to ORFs listed in the yeast ORFs files), obtained by ftp from the *Saccharomyces* Genome Database (Cherry *et al.*, 2001), together with the known introns in yeast obtained from the Ares lab Yeast Intron Database version 2.0 (Davis *et al.*, 2000). This not only reduced the expected number of false positives to about 1000, but also saved significant CPU time (run time was about one week).

Instead of the approximately 1000 false positives/novel snoRNAs that we expected for the reduced yeast genome, we in fact found 579. These candidates were further examined and reduced in number by considering their scores and performing some manual processing, such as checking primary and secondary structures. We also checked the high ranking candidates regarding their genomic context. Amongst the 579 candidates, we only found 31 snoRNA structures having both a $\Psi_1\Psi_2$- and a $\Psi_3\Psi_4$-pair. To create a list of 50 candidates for experimental screening, we also added another 19 of our most interesting Ψ_3 Ψ_4-candidates.

We now discuss these 50 hits in more detail. In Figure 7 we present an example of a putative snoRNA that `Fisher` located in the yeast genome. The motifs: Ψ_1 Ψ_2, H-box, Ψ_3 Ψ_4 and the AHA-box are marked in bold. Its Ψ-pairs are Ψ_1 Ψ_2=32 and Ψ_3 Ψ_4=2 (see Table 1); these have not been previously assigned to any known snoRNA. The distances between the motifs are $V = 61$, $W = 17$, $X = 27$ and $Y = 91$ (see Figure 2). The secondary structure, that exhibits the typical double mountain, is also displayed in Figure 7. Encouragingly, the highest scoring candidates showed a clear over-representation of Ψ-pairs that are not assigned to known snoRNAs, whereas hits with lower scores more often had Ψ-pairs that are already assigned to known snoRNAs.

Both the 50 candidates and the known snoRNAs are broadly distributed on the yeast genome, with all chromosomes possessing either known snoRNAs or candidates, or both. Chromosome XV is notable in that it carries the genes for four known snoRNAs, and is also the chromosome with the largest number of candidates located along its length. Most of our top candidates are located in chromosomes XII-XVI.

Three of our candidates were found to have especially interesting genomic locations. Two are located in the introns of the genes for the yeast ribosomal proteins, RPL43A and RPS11A (both genes contain one intron only). We consider this to be a strong indication that these two candidates are indeed snoRNAs, since the majority of

intronic snoRNAs are found in the introns of ribosomal protein genes (Maxwell and Fournier, 1995) and, in yeast, all intronic snoRNAs except one are in ribosomal or ribosome-associated proteins (Samarsky and Fournier, 1999). An examination of orthologous ribosomal proteins in other organisms revealed no additional information, though (Higa *et al.*, 1999) have demonstrated that the human and mouse *Rps11* genes house U35 (a C/D family snoRNA) in the third intron (the yeast *RPS11A* gene has only one intron). A third candidate was located in the ORF coding for the snoRNP U3 protein MPP10. This candidate is not intronic, and completely overlaps the coding sequence. This arrangement has been recently demonstrated for the C/D family snoRNA U86, in yeast (Filippini *et al.*, 2001). Given this demonstration of completely overlapping snoRNA-protein coding genes, and the fact that the host gene for our candidate is also involved in snoRNA-dependent rRNA processing, we consider this hit to be a good candidate for a *bona fide* snoRNA. This suggests that future genomic searches may require the entire genome sequence.

DISCUSSION

We have presented an algorithm for searching the yeast genome for H/ACA snoRNAs. It is reasonably fast and can be tuned to produce a manageable number of good candidates.

The method we describe could in principle be applied to any family of RNAs with low level conserved sequence and well-conserved secondary structure. However, it might be that it works well for H/ACA snoRNAs since the corresponding structure is quite simple; more studies need to be made to determine whether the method works for more complex structures. In any case, some of the methods we have developed might still be usefully incorporated into existing search strategies.

Two issues warranting discussion are the number of false hits and overtraining. For the test genome described in the results section our strategy had an performance of 99.4%, but this also required the introduction of one false negative (snR8). Since we did not include snR9, snR30, or snR37 in our training data, it is likely that our approach will not hit all known H/ACA snoRNAs, but it will hopefully recover most, as per the computational screen for yeast C/D snoRNAs (Lowe and Eddy, 1999). We are as yet unaware how iteration (adding verified candidates to the training dataset) will affect the ability of our method to identify new H/ACA snoRNAs, and it is not possible to predict how many iterations will be required to recover the majority of H/ACA snoRNAs in yeast. However, since it was found that the known snoRNAs were close to the top in the candidate list for the test genome, partial screening might be expected to effectively recover the majority of additional H/ACA snoRNAs.

In terms of immediate application of our algorithm to other organisms, the human genome provides an important data set for which genome sequence and a sizeable number of characterized H/ACA snoRNAs is available (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001). Preliminary work on the known human H/ACA snoRNAs indicates greater H-box and secondary structural homogeneity than for yeast. However, since the human genome is about 250 times longer than the yeast genome and since time complexity for folding a sequence with n bases is $O(n^3)$, the search may become too slow if too much secondary structure filtering is required. This could be offset by, for example, adapting the scanning algorithms described in Rivas and Eddy (2000) or by parallelizing the search. Perhaps more importantly with regards to folding, the accuracy of MFE structure prediction can depend quite heavily on the length of the subsequence of the genome that is being folded. Even so, we emphasize that the ability of the predicted structures to provide signal for discovery of new RNA family members is more important than their correctness.

In conclusion, as additional sequence data for yeasts becomes available (Souciet *et al.*, 2000), it should be possible to not only identify known snoRNAs in other yeasts using BLAST (Cliften *et al.*, 2001; Cervelli *et al.*, 2002), but also to evaluate a list of candidates by genome comparison. This has two implications. First, preliminary evidence that a candidate is a snoRNA can be gathered bioinformatically, as opposed to using labor-intensive experimental screening. Second, we can potentially reverse our approach and establish the site of pseudouridylation. While this does not replace the importance of experimentally determining the position of pseudouridylation, it does mean that our methods can in principle be applied in reverse order in cases where there is comparative data available but no experimentally determined pseudouridylation sites. We are currently developing this strategy, together with a comparative pseudouridylation map for rRNA alignments that may aid in assigning confidence to H/ACA snoRNAs identified by comparative genome analysis.

ACKNOWLEDGEMENTS

This work was supported by The Swedish Foundation for International Cooperation in Research and Higher Education (STINT), The Swedish Research Council (VR), and the New Zealand Marsden Fund. Thanks to M. Fournier and W. Decatur for sharing unpublished data and for helpful discussions. Thanks are also due to Alicia Gore for assistance in examining the genomic context of candidates and discussions, and Linus Sandegren for assistance with data gathering and analysis at an early stage of this project.

REFERENCES

Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, **11**, 941–950.

Altschul,S.F. and Erikson,B.W. (1985) Significance of nucleotide-sequence alignments—a method for random sequence permutation that perserves dinucleotide and codon usage. *Mol. Biol. Evol.*, **2**, 526–538.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Carter,R.J., Dubchak,I. and Holbrook,S.R. (2001) A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res.*, **29**, 3928–3938.

Cervelli,M., Cecconi,F., Giorgi,M., Annesi,F., Oliverio,M. and Mariottini,P. (2002) Comparative structure analysis of vertebrate U17 small nucleolar RNA (snoRNA). *J. Mol. Evol.*, **54**, 166–179.

Cherry,J.M., Ball,C., Dolinski,K., Dwight,S., Harris,M., Matese,J.C., Sherlock,G., Binkley,G., Jin,H., Weng,S. and Botstein,D. (2001) '*Saccharomyces* Genome Database' (downloaded 11/1/2001 from ftp://genome-ftp.stanford.edu/pub/yeast/yeast_NotFeature/) (visited 18/2/2002 http://genome-www.stanford.edu/Saccharomyces/).

Cliften,P.F., Hillier,L.W., Fulton,L., Graves,T., Miner,T., Gish,W.R., Waterson,R.H. and Johnston,M. (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, **11**, 1175–1186.

Collins,L., Moulton,V. and Penny,D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, **51**, 194–204.

Davis,C.A., Grate,L., Spingola,M. and Ares,Jr,M. (2000) Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res.*, **28**, 1700–1706. http://www.cse.ucsc.edu/research/compbio/yeast_introns/currentDBv2/intronsAround.fa (downloaded 11/1/2001).

Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Eddy,S.R. (2002) Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.

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

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.

Erdmann,V.A., Barciszewska,M.Z., Szymanski,M., Hochberg,A., de Groot,N. and Barciszewski,J. (2001) The non-coding RNAs as riboregulators. *Nucleic Acids Res.*, **29**, 189–193.

Filippini,D., Renzi,F., Bozzoni,I. and Caffarelli,E. (2001) U86, a novel snoRNA with an unprecedented gene organization in Yeast. *Biochem. Biophys. Res. Comm.*, **288**, 16–21.

Ganot,P., Bortolin,M.L. and Kiss,T. (1997a) Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, **89**, 799–809.

Ganot,P., Caizergues-Ferrer,M. and Kiss,T. (1997b) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev.*, **11**, 941–956.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.

Higa,S., Yoshihama,M., Tanaka,T. and Kenmochi,N. (1999) Gene organization and sequence of the region containing the ribosomal protein genes RPL13A and RPS11 in the human genome and conserved features in the mouse genome. *Gene*, **240**, 371–377.

Hofacker,I.L., Fontana,W., Bonhoeffer,S. and Stadler,P.F. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, **125**, 167–188.

Hogeweg,P. and Hesper,B. (1984) Energy directed folding of RNA sequences. *Nucleic Acids Res.*, **12**, 67–74.

Hüttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachellerie,J.P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.

International Human Genome Sequencing Consortium (2001) Initialsequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Kelley,C.T. (1999) *Iterative Methods for Optimization, Frontiers in Applied Mathematics*, SIAM, 18, Philadelphia.

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

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

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

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.

Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAmotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.

Mathews,D.H., Sabina,J., Zucker,M. and Turner,H. (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Maxwell,E.S. and Fournier,M.J. (1995) The small nucleolar RNAs. *Annual Reviews of Biochemistry*, **35**, 897–934.

Meli,M., Albert-Fournier,B. and Maurel,M.C. (2001) Recent findings in the modern RNA world. *Int. Microbiol.*, **4**, 5–11.

Moulton,V., Zuker,M., Steel,M., Pointon,R. and Penny,D. (2000) Metrics on RNA secondary structures. *J. Comp. Biol.*, **7**, 277–292.

Ni,J., Tien,A.L. and Fournier,M.J. (1997) Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell*, **89**, 565–573.

Ofengand,J. and Fournier,M.J. (1998) The pseudouridine residues of ribosomal RNA: number, location, biosynthesis, and function. In Grosjean,H. and Benne,R. (eds), *Modification and Editing of RNA*. ASM Press, pp. 229–253.

Pasquinelli,A.E. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.

Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.

Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.

Rivas,E., Klein,R.J., Jones,T.A. and Eddy,S.R. (2001) Computational identification of noncoding RNAs in *E.coli* by comparative genomics. *Curr. Biol.*, **11**, 1369–1373.

Samarsky,D.A. and Fournier,M.J. (1999) A comprehensive database for the small nucleolar RNAs from *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **27**, 161–164. http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html.

Souciet,J.-L. *et al.* (2000) Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett.*, **487**, 3–12.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wassarman,K.M., Repoila,F., Rosenow,C., Storz,G. and Gottesman,S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, **15**, 1637–1651.

Weinstein,L.B. and Steitz,J.A. (1999) Guided tours: from precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.*, **11**, 378–384.

Workman,C. and Krogh,A. (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.*, **27**, 4816–4822.

Zuker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide in RNA biochemistry and biotechnology. In Barciszewski,J. and Clark,B.F.C. (eds), *NATO ASI Series*. Kluwer Academic Publishers.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.