J Biomed Sci (2008) 15:697–705
DOI 10.1007/s11373-008-9271-x

ORIGINAL PAPER

A computational screen for C/D box snoRNAs in the human genomic region associated with Prader-Willi and Angelman syndromes

Padmavati Sridhar · Hin Hark Gan · Tamar Schlick

Received: 18 March 2008 / Accepted: 10 July 2008 / Published online: 27 July 2008
© National Science Council Taipei 2008

Abstract Small nucleolar RNAs (snoRNAs) play a significant role in Prader-Willi Syndrome (PWS) and Angelman Syndrome (AS), which are genomic disorders resulting from deletions in the human chromosomal region 15q11–q13. To identify snoRNAs in the region, our computational study employs key motif features of C/D box snoRNAs and introduces a complementary RNA–RNA hybridization test. We identify three previously unknown methylation guide snoRNAs targeting ribosomal 18S and 28S RNAs, and two snoRNAs targeting serotonin receptor 2C mRNA. We show that the three snoRNA candidates likely possess methylation strands complementary to, and form stable complexes with, human ribosomal RNAs. Our screen also identifies 8 other snoRNA candidates that do not pass the rRNA-complementarity and/or hybridization tests. Two of these candidates have extensive sequence similarity to HBII-52, a snoRNA that regulates the alternative splicing of serotonin receptor 2C mRNA. Six out of our eleven candidate snoRNAs are also predicted by other existing methods.

Keywords RNA motif scanning · RNA–RNA hybridization · C/D box snoRNA · Prader-Willi syndrome · Angelman syndrome · Serotonin receptor 2C mRNA

P. Sridhar · H. H. Gan · T. Schlick (✉)
Department of Chemistry, New York University, 100 Washington Square East, New York, NY 10003, USA
e-mail: ts1@haifa.biomath.nyu.edu

T. Schlick
Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10021, USA

Introduction

Non-coding RNAs are abundantly expressed in cells and implicated in a growing list of genetic diseases [1, 2]. In particular, small nucleolar RNAs (snoRNAs) play a crucial role in Prader-Willi (PWS) and Angelman (AS) syndromes, which are neurodegenerative disorders caused by the deletion or lack of expression of genes in human chromosomal region 15q11–q13 [3–7]. SnoRNAs, found in the nucleolus, primarily guide chemical modification of ribosomal RNAs (rRNA) and small nuclear RNAs (snRNAs) [8]; some snoRNAs also have other functional roles [4]. They guide chemical modifications of the target through complementary base pairing. SnoRNAs come in two groups: C/D box snoRNAs guide 2-O′-methylation in rRNAs, and H/ACA box snoRNAs mediate pseudouridylation in rRNAs.

A number of C/D box snoRNA genes in PWS/AS region have been experimentally identified from cDNA libraries of total RNA extracted from various tissues [5]. The region hosts 47 and 24 copies of snoRNA HBII-52 and HBII-85, respectively, and a single copy of snoRNA HBII-13; additional snoRNA HBII-436, HBII-437, HBII-438A, and HBII-438B have also been identified [9]. Crucially, snoRNA HBII-52 regulates alternative splicing of the pre-mRNA of serotonin receptor 2C located on a different chromosome [4]. PWS patients do not express HBII-52, and defective pre-mRNA processing contributes to the disease. However, a recent deletion mapping suggests that the loss of all the HBII-52 snoRNAs in the region has no major role in PWS, although their role in the disease when in conjunction with the loss of other genes in the region cannot be ruled out [10]. Since experimental screens often miss small and less abundant noncoding RNAs, it is likely that not all snoRNAs in PWS/AS region have been identified experimentally. A computational search, such as that reported here, can help

provide a more comprehensive screen of disease-related C/D box snoRNA genes.

Current computational screens for snoRNAs have focused on non-mammalian genomes, including bacteria, yeast, and fruit fly [11–14]; several recent studies focused on mammalian snoRNAs [3, 5, 15, 16]. For example, the S. Eddy lab used a Snoscan program to screen for C/D box snoRNAs in *Saccharomyces cerevisiae* [11] to identify 22 new C/D box snoRNAs (termed sn50 to sn71). Snoscan uses conserved motif features of snoRNAs (i.e., sequence motifs and lengths) combined with a probabilistic scoring scheme to rank and predict snoRNAs. Recent assessment of RNA sequence/structure conservation in mammalian genomes also confirmed several novel human snoRNAs [17]. A similar approach has been developed using the snoReport program for identifying snoRNAs of unknown RNA or protein targets [18]. Yet another program snoSeeker has been used to identify both guide (C/D and H/ACA) snoRNAs and orphan snoRNAs [19].

Here, we screen for C/D box snoRNAs in human 15q11-q13 using a combination of search techniques for conserved sequence motifs (with pattern search program RNAMotif [20]), sequence complementarity to rRNAs, and assessment of the free energy of hybridization between putative methylation strands and 18S/28S rRNAs. This approach incorporates elements of existing computational screening methods plus the hybridization energy test to ensure that the candidate C/D box snoRNAs bind to the target rRNAs with high confidence. The hybridization energy test, similar to siRNA target prediction [21] is a new feature in computational screening of snoRNA. This approach uses a direct assessment of the RNA–RNA hybridization free energy from a standard RNA folding program rather than the heuristic probabilistic scores for base pairing (derived from complementary length, mismatches, canonical/non-canonical base pairing) employed in Snoscan and snoSeeker programs. All computational approaches share the capability of detecting snoRNAs expressed at low levels or in certain cell types, which are often missed in experimental screens.

Our computational analysis predicts three rRNA-targeting C/D box snoRNAs in the PWS/AS region. Eight other detected snoRNA candidates do not pass the rRNA hybridization tests. Two of these candidates have extensive sequence similarity to HBII-52, a snoRNA that regulates the alternative splicing of serotonin receptor 2C. One of the candidates has sequence segments with significant complementarity to rRNAs and serotonin receptor 2C mRNA, suggesting it may target multiple RNAs. Thus, we predict five snoRNAs from eleven candidate snoRNA sequences. Interestingly, six of the eleven candidates are also predicted to be box snoRNAs by other existing search algorithms (Snoscan, snoSeeker, snoReport), and three of these snoRNAs are the same ones predicted by our approach.

Materials and methods

Overall approach

We extract the conserved sequence and structural information from known mammalian and yeast C/D box snoRNAs. This information helps search for genomic sequences with the specified motif. We then assess the candidate sequences by verifying the existence in candidate snoRNAs' methylation strands of short sequences complementary to 18S and/or 28S rRNAs through sequence alignment and hybridization. Our search and verification steps are as follows:

1. Extract characteristics of C/D box snoRNAs motifs (see Table 1, Fig. 1), including conserved sequences (e.g., boxes C, C′, D, and D′), length of the sole helical stem, lengths of the two methylation strands, and distances between the conserved boxes.
2. Design motif descriptors based on snoRNA parameters and search for candidate snoRNAs using RNAMotif [20], an RNA motif search program.
3. Verify the existence of short sequences in methylation strands of candidate snoRNAs that are complementary to 18S and 28S rRNAs using sequence alignment and hybridization test. The significance of the identified candidates is rated based on the quality of sequence complementarity and hybridization energy.

In step 3, our use of snoRNA–rRNA hybridization energy test provides a stringent screening of snoRNA candidates not employed in existing approaches.

Motif descriptor of C/D box snoRNA

The secondary structure of a C/D box snoRNA has a 5′ helix of four to eight base pairs (bp) and a large loop of about 90 nucleotides (Fig. 1). The single stranded loop contains four conserved regions, labeled boxes C, D′, C′, and D. Boxes C and C′ have conserved sequence motif RUGAUGA (R for A or G). Boxes D and D′ have conserved nucleotides CUGA [11]. The 90 nt loop of most snoRNAs has two methylation strands that are complementary to one of four types of human rRNA: 5S rRNA, 5.8S rRNA, 18S rRNA and 28S rRNA. These complementary base pair interactions are indicative of the C/D box snoRNAs' roles in guiding methylation of rRNAs.

In addition to sequence motifs, length parameters are essential for characterizing snoRNAs [11]. The 13 human and mouse C/D box snoRNAs are summarized in Table 1. In our computational search, we use the following parameter ranges for motifs: sole helical stem (4–12 base pairs), methylation strands M1 (5–30 nt) and M2 (10–37 nt), D′ to C′ distance (2–30 nt), and total length (50–100 nt). In the helical stem, we allow at least 30% pairing

Table 1 The secondary structure lengths (nt) of 13 C/D box snoRNAs

	Accession #	H5 + SS	Box C–D′ (M1)	Box D′–C′	Box C′–D (M2)	H3 + SS	Total length
U101*	AY349602	7	10	19	11	5	72
U103A*	AY349604	7	16	5	19	5	72
U104*	AY349605	9	28	6	10	7	79
U105*	AY349606	13	12	5	24	13	85
U106*	AY349607	6	–	29	24	4	79
MBII-82*	AF357319	7	7	13	16	1	64
MBII-95*	AF357320	0	5	12	24	2	63
MBII-180*	AF357325	11	13	6	35	4	89
MBII-210*	AF357328	0	6	11	17	4	59
MBII-296*	AF357334	5	10	10	13	2	60
HBII-13**	NR001294	7	10	3	22	5	70
HBII-52**	NR001291	7	15	7	28	5	89
HBII-85**	NR003316	7	27	6	27	5	98
Range	–	0–13	6–28	3–29	10–35	1–13	59–98
Average		6.6	13.3	10.2	20.8	4.8	87.9

* Source: Genbank (www.ncbi.nlm.nih.gov); ** Cavaille et al. [5]

fraction on the 5′ end of the helix, which begins with GA; both Watson-Crick (GC, AU) and wobble (GU) base pairs are allowed. This generic descriptor for C/D box snoRNAs is displayed in Fig. 1.

Hybridization of methylation strands and rRNAs

Sequence alignment is the standard method to verify the existence of complementary snoRNA methylation strands M1 and M2 to rRNAs [11]. In addition to sequence alignment, we assess the free energy of hybridization to confirm that the methylation strand/rRNA complex is energetically favorable. We also perform competitive hybridization experiments to ensure that the methylation strands bind favorably to their target rRNA segments.

We initially predict methylation strands using sequence alignment and then calculate the hybridization energy for the methylation strand/rRNA complex. Because the hybridization server (http://www.bioinfo.rpi.edu/applications/mfold/) [22] is limited to shorter sequences, we divide 18S and 28S rRNAs into five and nine equal sequence segments, respectively, of ~300 nt and each segment is hybridized with the two methylation strands (M1 and M2).

For each methylation strand, we identify the lowest energy of hybridization with the five 18S and nine 28S segments. Next, we determine the "target energy" defined as the hybridization energy between the methylation strand and the segment of 18S/28S containing the site predicted by sequence alignment to bind to the methylation strand. We then compute the "exact energy" of hybridization between the exact complementary sequences in 18S/28S and methylation strands, as identified by LALIGN server (http://www.ch.embnet.org/software/LALIGN_form.html). These three free energy values—the lowest energy, the target energy and the exact energy—are used to assess the quality of snoRNA sequence hits from RNAMotif. Consistency of the energy values implies that the methylation strand binds to the target rRNA sequence segment.

Criteria for snoRNA candidates

Based on sequence complementarity and methylation strand/rRNA hybridization energetics, we determine the significance or non-significance of snoRNA candidates using the following criteria: (1) complementary sequence length ≥9 nt, (2) complementary sequence match ≥75%, (3) and the target and exact free energies must be at least 75% of the lowest free energy. Criteria 1 and 2 are similar to those used in previous computational screens [11]. The new criterion 3 ensures that the candidate sequences bind to the complementary sequence segments identified in criteria 1 and 2. The sequence hits that satisfy criteria 1–3 are considered candidate C/D box snoRNAs.

Human chromosome 15q11–q13

We scan the larger 14.4 Mb human genomic region 15q11.1–q13.3 containing the classic PWS/AS region. The sequence was downloaded from Ensembl (http://www.ensembl.org); ~12% of the region is still not determined. The sequence's nucleotide composition is as follows: 28.5% (A), 21.5% (C), 21.5% (G), and 28.5% (T). In human 15q11–q13, the subregion BP2 to gene *HERC2* is the classical PWS/AS region. The BP1–BP2 subregion is also known to play a role in the

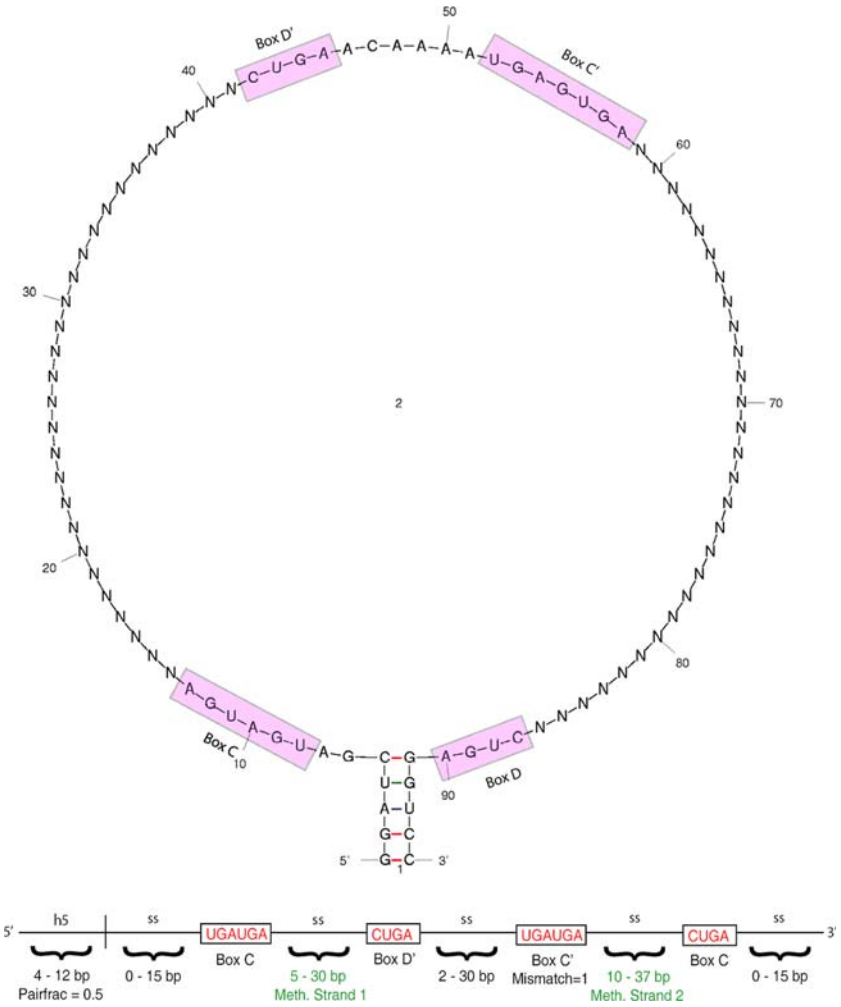

Fig. 1 (Upper) Secondary structure of a C/D box snoRNA. The nucleotides in methylation strands M1 and M2 were substituted with "NNN…" so as to prevent the two strands from base pairing on MFold. This single stranded loop contains four conserved regions, labeled boxes C, D′, C′, and D. Boxes C and C′ have conserved nucleotides rugauga. The r represents either a or g. Boxes D and D′ have conserved nucleotides cuga. The 90 nt loop of most snoRNAs has two methylation strands that are complementary to one of four types of human rRNA: 5S rRNA, 5.8S rRNA, 18S rRNA or 28S rRNA.(Lower) The snoRNA descriptor with motif parameters derived from 13 known C/D box snoRNAs (Table 1). It is used to search for snoRNAs in the 15q11.1–q13.3 region. Each of the 4 boxes and their conserved nucleotides were specified and 1 mismatch was allowed in Box C′. The approximate 90 nt loop was specified through several different single strands (ss), including the two methylation strands. In addition, the total length of the snoRNA was restricted to be between 50 and 100 nt

disease. Scanning the 14.4 Mb region for the C/D snoRNA motif requires only about 1 h of CPU time on a SGI workstation.

Results and discussion

Estimate of false negatives—tests of known C/D box snoRNAs

We calculate false negative rate by using our criteria 1–3 to screen previously identified 11 yeast snoRNAs, as shown in Table 2. Human snoRNAs from the PWS region are not suitable for this purpose because some of them do have sequence regions complementary to rRNAs. The methylation strands of yeast snoRNAs have extensive (90–100%) sequence complementarity to target segments in 18S rRNA. Significantly, the lowest, target and exact free energies of binding agree to within 90–100%, indicating that the methylation strands bind to their target segments in 18S rRNA. According to our criteria 1–3, 10 yeast snoRNAs are rated as significant candidates. The sole excluded case is snR53 which has a short 8 nt match with 18S rRNA and a large mismatched ($\sim$50%) target and exact hybridization energies for methylation strand 1 (M1). Even if criterion 1 (length $>$ 8 nt) is not used, criterion 3 (free energy match) would have eliminated snR53 as a good snoRNA candidate. Thus, our false negative rate is 1/11 or 9%.

Table 2 Hybridization analysis of yeast snoRNAs with 18S rRNA

Results								Evaluation
Yeast snoRNA	Lalign			Hybridization				Result
	% Overlap	nt/mm	rRNA location	Lowest		Target	Exact	
				Value	Part			
U14-M1	100	13/0	80–93	−19.5	1	−19.5 (100%)	−18.2 (93.33%)	Significant
U14-M2	92.9	14/1	458–472	−18.8	2	−18.8 (100%)	−18.1 (96.27%)	Significant
snR53-M1	87.5	8/1	18–26	−12.5	4	−6.8 (54.5%)	−6.1 (48.8%)	Not significant
snR54-M1	100	11/0	1030–1041	−15.3	4	−14.2 (92.8%)	−13.8 (90.1%)	Significant
snR55-M1	85.7	14/2	1325–1339	−20.9	4	−20.9 (100%)	−19.1 (91.3%)	Significant
snR56-M1	92.3	13/1	1490–1503	−18.8	4	−18.8 (100%)	−18.5 (98.4%)	Significant
snR57-M1	78.9	19/4	~1645	−19.7	5	−19.7 (100%)	−19.3 (97.9%)	Significant
snR70-M1	90	10/1	728–738	−16.2	1	−16 (98.7%)	−14.7 (90.7%)	Significant
snR74-M1	100	13/0	23–36	−19.4	1	−19.4 (100%)	−17.1 (88.1%)	Significant
snR77-M1	92.9	14/1	625–639	−21.3	2	−21.3 (100%)	−20.3 (95.3%)	Significant
snR79-M1	93.8	16/1	1060–1076	−20.9	3	−20.9 (100%)	−19.4 (92.8%)	Significant

Estimate of false positives—tests of C/D box snoRNAs motifs from randomized sequences

We also evaluate the sensitivity of our motif descriptor by calculating the expected number of sequence hits for random sequences using information theory and the motif scanning method; its specificity will be calculated using the hybridization test described below. The expected frequency from information theory is estimated as follows: the information of content of C/D Box snoRNA descriptor is 48 bits, assuming 4 bases pairs (2 bits each) and 20 conserved bases (2 bits each). The expected frequency is about one in 10^{14} nucleotides ($1/2^{48}$), implying that the probability of hits is very low.

Alternatively, we scan with the C/D Box snoRNA descriptor 20 randomized sequences of 14.4 Mb generated with the same nucleotide composition (28.8% A, 21.1% C, 21.2% G, 28.9% T) as found in the PWS/AS region. The expected number of hits with the snoRNA motif per 14.4 Mb is 12.65 (counting hits in both forward and reverse genome strands). The information theory estimate is considerably lower than the motif-scanning result because the variable aspects of the snoRNA descriptor are not accounted for. We then use our screening criteria 1–3 to test the significance of randomly selected 13 sequence hits from randomized sequences to estimate the overall false positive rate. For each snoRNA sequence hit, we test methylation strands M1 and M2 on 18S and 28S rRNAs for possible matches, yielding a total of 52 possible cases. For 18S rRNA, 3 out of 26 cases pass the screening criteria, yielding a false positive rate of 12% for a 14.4 Mb sequence. For 28S rRNA, the false positive rate is 31% (8/26). Thus, the overall false positive rate is 21%. Further reduction in the false positive rate can be achieved by examining sequence conservation patterns and similarity to known snoRNAs.

Candidate C/D box snoRNAs

Table 3 shows the eleven sequences matching the C/D box snoRNA descriptor, Table 4 tabulates the lengths of their secondary structure motifs, and Fig. 2 indicates the relative locations of the candidates. Candidate C1 is found between BP1 and BP2, whereas C2 to C11 are found between BP2 and BP3, which contains the classical PWS/AS region. The snoRNA candidates C1–C8 are on the forward strand (+) and C9–C11 are on the reverse strand (−). The presence of C9–C11 candidate snoRNAs on the reverse strand of the BP2-*HERC2* sequence is an interesting finding because prior to this study, no known snoRNAs had been mapped to the reverse strand of this region. Table 4 shows that the overall distribution of secondary motif elements of the 11 sequence hits is fairly narrow compared with known snoRNAs in Table 1. For example, 9 hits have stems with 4–5 base pairs, 7 have M1 strands with 25–30 nt, 6 have M2 strands with 23–31 nt, and 9 with a total length of 80–100 nt. The candidate snoRNAs have similar total lengths and average structural parameters as known snoRNAs in Table 1. However, M1 strands for the candidates have an average of 22.5 nt compared with 13.3 nt for known snoRNAs.

Sequence complementarity and hybridization tests using criteria 1–3 suggest three good candidate snoRNA sequences: C8 targets both 18S and 28S rRNAs, and C10 and C11 target only 18S rRNA. The 18S targets of these good candidates are close but do not overlap with existing yeast methylation sites; C10-130 (130 is complementary

Table 3 Candidate C/D box snoRNA sequences

Candidate	Strand/Adjacent genes	Length (nt)	Sequence
1	+/NIPA1	102	gtctggatctgtgatgagatggggaaagtgggctcaggaggtctgaatctgtgatgagatgggggaagtgggctcaggaggtctggatctgagttggggatc
2	+/HBII-436	79	agcatgatgataaaacacactgatgatatcgtgataatatttgaaaccaaagaaggtaggactgactaaatacagtggt
3	+/HBII-436	91	tttgtgatgacattgacattcctactttggaatttctctgatgtttctctgatcattgccctgaatttgtgaatcatctctgatgattgca
4	+/HBII-436	97	ttattgatgacattaacatccccaatttgggatttctctgacatttttctcatgattatactggagttgggtttggggaaggaagatcactgagaga
5	+/HBII-436	84	aacttgatgaagcatgaaattaccagctggtctgatgggagatatgatgtggccctgaggtacagtaggtggtgcctgaagcat
6	+/HBII-85, HBII-52	86	gaggtgatgacttaaaaatcatgctcaataggattacgctgaggcccagccttggtgagaattttggaagaggatgctgagatccc
7	+/HBII-85, HBII-52	95	aaggtgatgacttaaaaatcatgctcaataggattacactgaggcccagcctaggtgataattttggaagaggatgctggaattctgaggtccct
8	+/HBII-438b, HBII-52	99	gaaaaaaggtgatgacttaaacatcatgcttaatagtattatgctgaagcccaagctaggtgagaattttggaagatgaggctgggatcctgagatcc
9	–/OCA2, GABRG3	99	catttgatgaaaacacctatcgcacagtctcctgggtttactgagcagtgataacaagaacgggtttataagacaagatctcaaaggagaaactgaagag
10	–/GABRG3	88	gcattgatgataggagtaagcaatatttctgatacaggcattgattagaatttctgaagaggattttcataatttctgacttgtaagc
11	–/HBII-13, HBII-438A	67	tcatttgatgattcctggctgaaacaattatgattatttccaaatactgattttccaaatcttgtca

Shown are the conserved box motifs, methylation strands (between first and second, and third and fourth boxes), genomic location and length. The good candidates are 6, 7, 8, 10, and 11

Table 4 The secondary structure lengths (nt) of the 11 candidate snoRNAs

CandidatesnoRNA	H5 + SS	Box C-D′ (M1)	Box D′-C′	Box C′-D (M2)	H3 + SS	Total length
C1	1	25	5	31	10	102
C2	4	9	8	24	14	79
C3	4	27	8	24	8	91
C4	4	27	11	31	4	97
C5	4	21	9	25	5	84
C6	**4**	**28**	**10**	**18**	**6**	**86**
C7	**4**	**28**	**13**	**23**	**7**	**95**
C8	**9**	**28**	**26**	**10**	**6**	**99**
C9	4	30	4	37	4	99
C10	**4**	**18**	**22**	**15**	**9**	**88**
C11	**5**	**7**	**8**	**10**	**17**	**67**
Range	1–9	7–30	4–26	10–37	4–17	67–102
Average	4.3	22.5	11.3	22.5	8.2	98.7

Good candidates are in bold

start location on 18S rRNA) is close to snR51 at 18S-Am100, and C11-560 to snR41 at 18S-Am541. Interestingly, several candidate sequences have high sequence similarity: C8 shares 85% global sequence identity with C6 and C7 although these candidates are not confirmed by the rRNA hybridization test. As will be shown below (Fig. 3), sequence analysis reveals that C6, C7 and C8 contain a complementary 18-nt segment to the pre-mRNA of

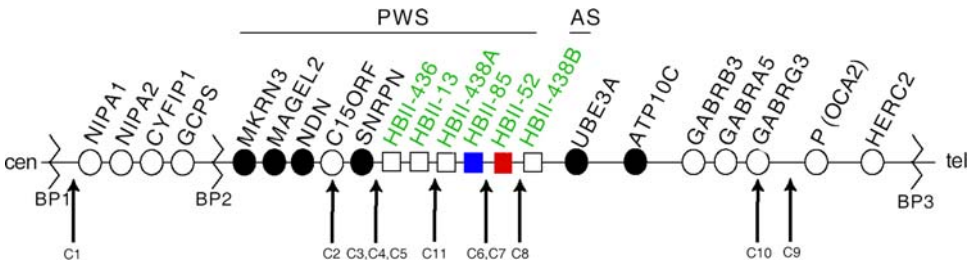

Fig. 2 The location of all eleven candidates in the 15q11q13 region. Candidate 1 is located right after break point 1 on the forward strand. Candidates 2–8 are located between break point 2 and C/D box snoRNA HBII-438B on the forward strand. Candidates 9, 10, 11 are located on the reverse strand. The good candidates are C6, C7, C8, C10, and C11

```
Candidate C6 with HBII-52: 100.0% identity in 47 nt overlap

          40        50        60        70        80
HB52  GAGGUGAUGACUUAAAAAUCAUGCUCAAUAGGAUUACGCUGAGGCCC
      ::::::::::::::::::::::::::::::::::::::::::::::::
C6    GAGGUGAUGACUUAAAAAUCAUGCUCAAUAGGAUUACGCUGAGGCCC
               10        20        30        40

Candidate C7 with HBII-52: 97.8% identity in 46 nt overlap

          40        50        60        70        80
HB52  AGGUGAUGACUUAAAAAUCAUGCUCAAUAGGAUUACGCUGAGGCCC
      :::::::::::::::::::::::::::::::::::: :::::::::
C7    AGGUGAUGACUUAAAAAUCAUGCUCAAUAGGAUUACACUGAGGCCC
               10        20        30        40

Candidate C8 with HBII-52: 86.5% identity in 52 nt overlap

               40        50        60        70        80
HB52  GAAGAGAGGUGAUGACUUAAAAAUCAUGCUCAAUAGGAUUACGCUGAGGCCC
      ::: : ::::::::::::::: ::::::::: ::::: :::: ::::: ::::
C8    GAAAAAAGGUGAUGACUUAAACAUCAUGCUUAAUAGUAUUAUGCUGAAGCCC
               10        20        30        40        50
```

Fig. 3 Sequence alignment of candidate and known snoRNAs. Sequence segments complementary to serotonin receptor 2C mRNA and Box D are highlighted in red and blue, respectively

serotonin receptor 2C. Intriguingly, we also find that the M1 methylation strands of C3 and C4 shared 82% sequence identity, hinting that they may target unknown mRNAs. Thus, sequence analysis and hybridization test indicate that at least 5 snoRNA candidates (C6, C7, C8, C10, C11) may target either rRNAs or mRNAs.

The three rRNA-targeting snoRNAs are found between BP2 and BP3 (Fig. 2). Significantly, the good candidate snoRNAs C8 and C11 are found in the classic PWS/AS deletion region within BP2–BP3: C8 between HBII-52 and HBII-438B on the forward strand, and candidate C11 between HBII-13 and HBII-438A on the reverse strand. Candidate C10 is located near *GABRG3*, toward the BP3 end and away from PWS and AS regions.

As shown in Table 5, the good candidates C8-M1, C10-M1 and C11-M2 show a high degree of consistency in their lowest, target and exact free energies, indicating that M1/M2 binds to its target rRNA segment. In general, the exact energy is less favorable than the target energy which in turn is less favorable than lowest energy. These energy discrepancies could be caused by additional base pairing interactions in the lowest energy since entire methylation strands are used for the hybridization test rather than target

Table 5 Candidate C/D box snoRNAs based on sequence alignment and hybridization tests with 18S and 28S rRNA

Candidate	% Overlap	nt/mm	rRNA location	Hybridization (kcal/mol)		
				Lowest	Target	Exact
18S						
C8-M1	85.7	21/3	40–60	−13.2	−13.2	−12.2
C10-M1	75	12/3	~130	−13.2	−13.2	−13.1
C11-M2	88.9	9/1	~560	−6.6	−6.6	−6.0
28S						
C8-M1	75	20/5	4320	−12.3	−11.4	−11.4

C1-M1 denotes methylation strand 1 (M1) of candidate C1, etc.; mm denotes the number of nucleotide mismatches

sequence segments as in the exact energy. Near perfect sequence complementarity tends to yield positive hybridization test, although this is not always that case because of possible favorable competing hybridization interactions. Thus, verification using both sequence alignment and hybridization tests increases confidence in the predicted snoRNA sequences.

Other candidate snoRNAs not confirmed to target rRNAs also display striking characteristics. For example, all four boxes (C, D′, C′, and D) of C1 are perfectly conserved, an important characteristic of C/D box snoRNAs, and both of its methylation strands (M1 and M2) have a striking (~90%) complementarity over 12–15 nt to 18S rRNA (not shown). For comparison, the known methylation strands for yeast snoRNA, snR53, has an 80% identity in a 10 nt overlap; previous studies have considered 80% to be a good marker for complementarity [11].

Sequence alignment of candidate and known C/D box snoRNAs in PWS/AS region

We have also analyzed sequence similarity between all candidate snoRNAs and the known human snoRNAs in the PWS/AS region. We find significant sequence similarity between C6–C8 and HBII-52, a C/D box snoRNA that has a

sequence segment complementary to serotonin 2C receptor mRNA [4] (Fig. 3); other candidates (C1–C5, C9–C10) do not have any significant sequence similarity with the known snoRNAs. As shown in Fig. 3, C6 and C7 sequences are virtually 100% identical over a 46–47 nt region of the HBII-52 sequence. Crucially, this region contains the 18-nt segment (AUGCUCAAUAGGAUUACG) that is complementary to serotonin receptor 2C mRNA. Further, the mRNA complementary segment of C6, C7 and HBII52 is immediately followed by Box D (CUGA). In the case of C7, the last nucleotide of the segment is an A rather than a G. C8 has a 87% sequence similarity to HBII-52 sequence over 52 nt but its mRNA complementary segment (AUGCU**U**AAU AG**U**AUUA**U**G) has three mutations (bold face) involving replacements by U's.

The above analysis shows that the sequences failing the rRNA hybridization tests do not mean they are poor snoRNA candidates. C6 and C7 target serotonin 2C receptor mRNA, and C1–C5 and C9 may target unknown mRNAs. Intriguingly, C8-M1 passes the 18S and 28S rRNA hybridization tests (Table 5) yet shows significant sequence similarity to HBII-52. This suggests that C8 contains sequence segments complementary to serotonin receptor 2C mRNA and rRNAs. It is plausible that C8 may regulate serotonin receptor 2C mRNA and 18S and 28S rRNAs via alternative splicing and methylation, respectively. Our findings suggest that 2–3 copies of HBII-52 snoRNAs may have been missed in experimental screens [5].

Comparison of snoRNA prediction algorithms

To further assess our snoRNA prediction approach, we compare it to three computational methods: Snoscan (server at http://lowelab.ucsc.edu/snoscan/), snoSeeker (server at http://genelab.zsu.edu.cn/snoseeker/), and snoReport (program downloaded from http://www.bioinf.uni-leipzig.de/Software/snoReport). The Snoscan [11] and snoSeeker [19] methods use probabilistic models and target sequences to identify snoRNAs, whereas the snoReport [18] approach uses support vector machine (SVM) classifiers based on a set of structural descriptors and does not use target information. We restrict our comparison to verification of known and candidate snoRNAs. Specifically, we consider verifying the following three snoRNA datasets: six HBII (13, 52, 85, 436, 437, 438a/b) snoRNAs [3], ten yeast snoRNAs in Table 2, and our 11 human candidate snoRNAs; the first two datasets have been experimentally verified to be snoRNAs. The results of our comparative analysis are presented in Table 6.

Table 6 reveals that the predictions significantly overlap for the known snoRNAs (yeast and HBII datasets) among the methods compared, with the exception of the snoReport approach. For the ten experimentally verified yeast snoRNAs in Table 2, our method predicts 9 snoRNAs, whereas Snoscan and snoSeeker predict 7 and 6 snoRNAs, respectively; for Snoscan, we use the default parameters and combined the results from yeast and mammalian probabilistic model options allowed by the server. The snoReport approach predicts only 2 snoRNAs; this lack of sensitivity may be due to the absence of target information. For the human HBII snoRNAs, our method predicts 3 out of 6 snoRNAs and the other three methods predict 5 snoRNAs; all 3 of our predicted snoRNAs are also found by the other methods. However, our approach does not predict hybridization of HBII-52 snoRNA with rRNAs; this snoRNA has been shown to bind the serotonin receptor 2C mRNA [4]. Thus, our approach shows good performance and agreement with other methods, especially for yeast snoRNAs. A perfect agreement among computational methods is not expected since they have different assumptions and inputs.

For our 11 human candidates (Table 3), our approach predicts 5 snoRNAs (C6, C7, C8, C10, C11) and snoSeeker predicts six "orphan" snoRNAs (C1, C2, C3, C6, C7, C8); orphan snoRNAs are defined as those that do not have a sequence segment antisense to rRNAs or snRNAs. Thus, both our approach and snoSeeker predict C6, C7 and C8 as snoRNAs. Significantly, as shown in Fig. 3, the C6, C7 and C8 snoRNAs have a 18 nt segment complementary to the serotonin receptor 2C mRNA. In addition to these snoRNAs, the existing methods predict C1, C2, and C3 to be snoRNAs: C1 is predicted by Snoscan, snoSeeker and snoReport, C2 by snoSeeker, and C3 by Snoscan and snoSeeker. Overall, our method predicts 5 out of 11 candidates, whereas existing methods predict 6 out of 11 candidates with 3 identical (C6, C7, C8) snoRNAs. This

Table 6 Comparison of snoRNAs predicted by different computational methods

Dataset/method	This work	Snoscan*	SnoSeeker	SnoReport
HBII snoRNAs (Ref. [5])	3/6 (HBII-13,85,438a)	5/6 (HBII-13, 52,85,436,438a/b)	5/6 (HBII-13,52,85,436,438a/b)	5/6 (HBII-13,52,85,436,438a/b)
Yeast snoRNAs (Table 2)	9/10 (U14;snR-54,55,56,57,70,74,77,79)	7/10 (U14;snR-53,54,56,74,77,79)	6/10 (U14;snR-53,54,56,77,79)	2/10 (snR-53,70)
Candidate C1-11 snoRNAs (Table 3)	5/11 (C6,7,8,10,11)	2/11 (C1,3)	6/11 (C1,2,3,6,7,8)	1/11 (C1)

*Snoscan uses default parameters and combined yeast and mammalian probability search models

result is consistent with our tests for experimentally verified yeast and human HBII snoRNA datasets.

Summary and conclusions

We have used a computational method for finding snoRNAs that combines sequence analysis and a novel hybridization energy test. In contrast to previous computational studies, the hybridization test we introduced ensures that the predicted methylation strands bind to their rRNA targets, therefore increasing the confidence of the identified candidate snoRNAs. The application of our method to the PWS/AS region of the human genome leads to eleven snoRNA candidates, three of which pass the hybridization test. These good snoRNA candidates binding rRNAs do not overlap with previously identified snoRNAs, suggesting that all genes important to PWS/AS have not been identified by experimental techniques [5]. Of the snoRNA candidates that do not pass the rRNA complementarity and hybridization tests, two are found to target serotonin receptor 2C mRNA via sequence alignment analysis. All of these five rRNA and mRNA-targeting snoRNA candidates are located in the PWS region and vicinity. Other candidates may target unknown mRNAs; these are called "orphan" snoRNAs. We also confirmed that six out of our eleven candidate snoRNAs are predicted by other existing search algorithms, three of which are the same snoRNAs (C6, C7, C8) predicted by our approach. Such computationally detected snoRNA candidates require experimental confirmation. Furthermore, our hybridization test can be generally incorporated and automated with motif scanning for genome-wide studies.

Acknowledgments This research was supported by a Joint NSF/NIGMS Initiative in Mathematical Biology (DMS-0201160), the Human Frontier Science Program (HFSP), and NSF (EMT-0727001).

References

1. Costa FF (2005) Non-coding RNAs: new players in eukaryotic biology. Gene 357(2):83–94
2. Eddy SR (2001) Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2(12):919–929
3. Cavaille J, Vitali P, Basyuk E, Huttenhofer A, Bachellerie JP (2001) A novel brain-specific box C/D small nucleolar RNA processed from tandemly repeated introns of a noncoding RNA gene in rats. J Biol Chem 276(28):26374–26383
4. Kishore S, Stamm S (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science 311(5758):230–232
5. Cavaille J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B et al (2000) Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. Proc Natl Acad Sci USA 97(26):14311–14316
6. Nicholls RD, Knepper JL (2001) Genome organization, function, and imprinting in Prader-Willi and Angelman syndromes. Annu Rev Genomics Hum Genet 2:153–175
7. Epstein CJ, Erickson RP, Wynshaw-Boris AJ (2004) Inborn errors of development the molecular basis of clinical disorders of morphogenesis. Oxford University Press, Oxford
8. Smith CM, Steitz JA (1997) Sno storm in the nucleolus: new roles for myriad small RNPs. Cell 89(5):669–672
9. Runte M, Huttenhofer A, Gross S, Kiefmann M, Horsthemke B, Buiting K (2001) The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. Hum Mol Genet 10(23):2687–2700
10. Runte M, Varon R, Horn D, Horsthemke B, Buiting K (2005) Exclusion of the C/D box snoRNA gene cluster HBII-52 from a major role in Prader-Willi syndrome. Hum Genet 116(3):228–230
11. Lowe TM, Eddy SR (1999) A computational screen for methylation guide snoRNAs in yeast. Science 283(5405):1168–1171
12. Accardo MC, Giordano E, Riccardo S, Digilio FA, Iazzetti G, Calogero RA et al (2004) A computational search for box C/D snoRNA genes in the Drosophila melanogaster genome. Bioinformatics 20(18):3293–3301
13. Yuan GZ, Klambt C, Bachellerie JP, Brosius J, Huttenhofer A (2003) RNomics in Drosophila melanogaster: identification of 66 candidates for novel non-messenger RNAs. Nucleic Acids Res 31(10):2495–2507
14. Speckmann WA, Li ZH, Lowe TM, Eddy SR, Terns RM, Terns MP (2002) Archaeal guide RNAs function in rRNA modification in the eukaryotic nucleus. Curr Biol 12(3):199–203
15. Fedorov A, Stombaugh J, Harr MW, Yu S, Nasalean L, Shepelev V (2005) Computer identification of snoRNA genes using a Mammalian Orthologous Intron Database. Nucleic Acids Res 33(14):4578–4583
16. Vitali P, Royo H, Seitz H, Bachellerie JP, Huttenhofer A, Cavaille J (2003) Identification of 13 novel human modification guide RNAs. Nucleic Acids Res 31(22):6543–6551
17. Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol 23(11):1383–1390
18. Hertel J, Hofacker IL, Stadler PF (2008) SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics 24(2):158–164
19. Yang JH, Zhang XC, Huang ZP, Zhou H, Huang MB, Zhang S et al (2006) snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. Nucleic Acids Res 34(18):5112–5123
20. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R (2001) RNA Motif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 29(22):4724–4735
21. Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH (1999) Predicting oligonucleotide affinity to nucleic acid targets. RNA 5(11):1458–1469
22. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31(13):3406–3415