Mol. Cells, Vol. 20, No. 1, pp. 35-42

Molecules and Cells

A Combined Approach for Locating Box H/ACA snoRNAs in the Human Genome

Hae Seok Eo[1], Kwang Sun Jo[2], Seung Won Lee[2], Chang-Bae Kim[3], and Won Kim*

School of Biological Sciences, Seoul National University, Seoul 151-742, Korea;

[1] Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea;

[2] JLLAB, Inc., Seoul 135-933, Korea;

[3] National Genome Information Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-333, Korea.

(Received December 10, 2004; Accepted April 6, 2005)

A novel combined method for locating box H/ACA small nucleolar RNAs (snoRNAs) is described, together with a software tool. The method adopts both a probabilistic hidden Markov model (HMM) and a minimum free energy (MFE) rule, and filters possible candidate box H/ACA snoRNAs obtained from genomic DNA sequences. With our novel method 12 known box H/ACA snoRNAs, and one strong candidate were identified in 30 nucleolar protein genomic sequences.

Keywords: box H/ACA snoRNA; Hidden Markov Model; Minimum Free Energy.

Introduction

Small nucleolar RNAs (snoRNAs) are known as non-coding RNAs (Eddy, 1999; Lafontaine and Tollervey, 1998). They are important nucleolar elements in ribosome biogenesis on account of their roles in directing 2′-*O*-methylatations and pseudouridylations of the universal conserved regions of rRNAs by forming canonical duplexes during the post-transcriptional processing of pre-rRNA (Bachellerie and Cavaille, 1997; Bachellerie *et al*., 2002; Decatur and Fournier, 2003; Eddy, 2001; Kiss-Laszlo *et al*., 1996; Lane *et al*., 1995; Tollervey and Kiss, 1997; Tycowski *et al*., 1996).

Computational approaches have been developed for identifying snoRNAs, but the weak sequence similarities of the snoRNAs within gene families have been the main obstacle to developing effective methods (Lowe and Eddy, 1999). A probabilistic method based on a hidden Markov model (HMM) has been developed, and successfully identified box C/D snoRNAs (Lowe and Eddy, 1999). However, it could not be applied directly to box H/ACA snoRNAs because of their complex hairpin secondary structures. Recently, Edvardsson *et al*. (2003) introduced a method using minimum free energy (MFE) rules, but only 3 out of 50 candidates were reliably detected. Here we have developed a novel method combining a probabilistic model and MFE rules (Fig. 1). Three sieve systems filter out false positives, and thus accurately locate the box H/ACA snoRNAs with high reliability.

Materials and Methods

Algorithm The first sieve uses a hidden Markov model to identify conserved box motifs in each hinge region (Figs. 2 and 3A). Hidden Markov models which were originally used in speech recognition have satisfactorily identified the consensus features of biological sequences (Durbin *et al*., 1998). Using this primary sieve system, the restricted genomic regions that need to be screened can be identified. The primary sieve products are filtered further by minimum free energy rules. Homeomorphically irreducible trees (HITs) and energy dot plots of minimum free energy rules (Fontana *et al*., 1993; Hofacker, 2003; Hofacker *et al*., 1994; Zuker and Jacobson, 1995) are used to identify hairpin structures in the box H/ACA snoRNAs (Fig. 1). After the overall free energies of folding are calculated, only the foldings with minimum free energies and standard secondary structures are chosen, and verified by sequence alignment.

* To whom correspondence should be addressed.
Tel: 82-2-887-0752; Fax: 82-2-872-1993
E-mail: wonkim@plaza.snu.ac.kr

Abbreviations: HITs, homeomorphically irreducible trees; HMM, hidden Markov model; MFE, minimum free energy; snoRNAs, small nucleolar RNAs.

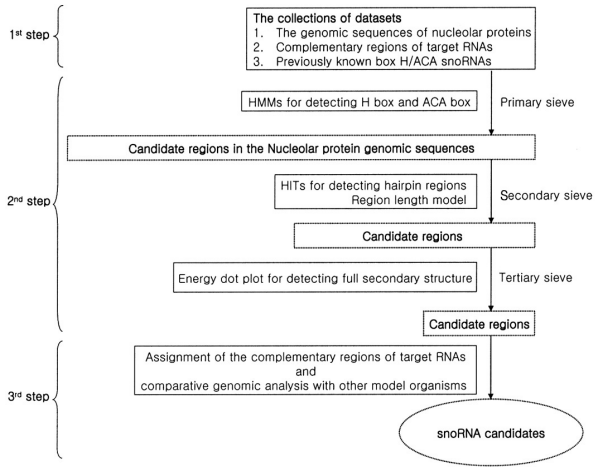


Fig. 1. Diagram of the sequence used for box H/ACA snoRNA identification.

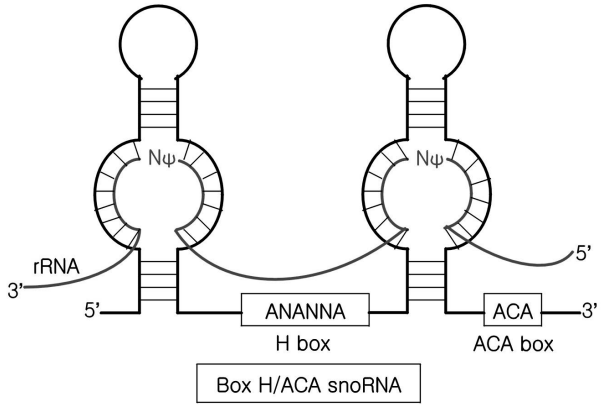


Fig. 2. Schematic diagram of box H/ACA snoRNAs. Box H/ACA snoRNAs have a conserved secondary structure, the hairpin-hinge (H box)-hairpin-tail (ACA box). Redrawn from Ganot *et al*. (1997). Ψ indicates pseudouridine.

Primary sieve system (hidden Markov model) An H box model based on a motif-based hidden Markov model was constructed and trained with the snoRNA dataset (Durbin *et al*., 1998; Krogh, 1998; Lowe and Eddy, 1999). A log-odd score system was introduced to remove abnormal structures below the range of scores of the snoRNA dataset (Fig. 3A). A software toolkit, Meta-MEME v3.2 was used (Grundy *et al*., 1997) for training. The same method was also used to locate ACA box motifs. In view of the canonical structure of box H/ACA snoRNAs, only regions with a sequence of 50–80 bases between the H box and the ACA box were selected as the primary sieve products. The upstream sequences from the H box were restricted to ten bases longer than the H box-ACA box regions. This was because, when this was done, the MFE rule in the secondary sieve system gave a good signal during the test.

Secondary sieve system (homeomorphically irreducible trees

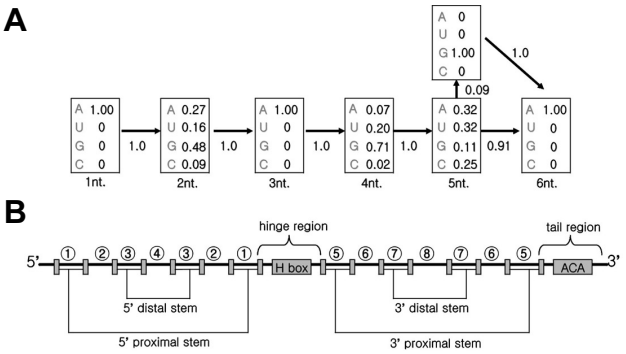
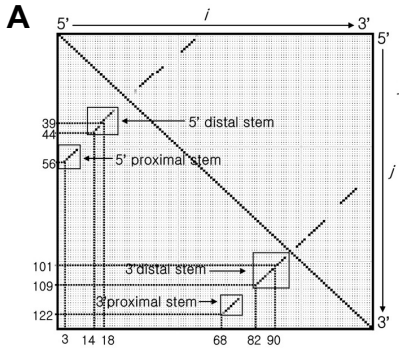


Fig. 3. The primary and secondary sieve system **A.** Hidden Markov model for H boxes of the snoRNA dataset. The probabilities of each nucleotide are contained in each column. The numbers of each arrow indicate transition probabilities. Log-odd scores for H boxes were calculated (Range score: 1.409–6.004; Mean score: 4.802). **B.** Region length model for box H/ACA snoRNA. ① 5′ proximal stem; ② 5′ pseudouridine pocket; ③ 5′ distal stem; ④ 5′ loop; ⑤ 3′ proximal stem; ⑥ 3′ pseudouridine pocket; ⑦ 3′ distal stem; ⑧ 3′ loop; 50 nt ≤ 2 (⑤ + ⑥ + ⑦) + ⑧ ≤ 80 nt; 10 nt ≤ ① + ③; 10 nt ≤ ⑤ + ⑦; 13 ≤ ① + ② ≤ 20; 13 ≤ ⑤ + ⑥ ≤ 20; 20 ≤ 2 × ③ + ④; 20 ≤ 2 × ⑦ + ⑧.

and the region length model) HITs based on the properties of the minimum free energy rule (Fontana *et al*., 1993; Hofacker, 2003; Hofacker *et al*., 1994) were applied to the upstream regions of each H box and ACA box to predict secondary structures. The self-complementary region can be identified as a single pair of matching brackets labeled 'P' and weighted by the number of base pairs. Correspondingly, a contiguous strand of unpaired bases is shown as a pair of matching brackets labeled 'U' and weighted by its length (Fontana *et al*., 1993). With the HITs algorithm and the region length model constructed from the snoRNA dataset, the false positives in the primary sieve products were eliminated (Fig. 3B).

Tertiary sieve system (energy dot plot) Finally, an energy dot plot matrix of minimum free energy (Hofacker *et al*., 1994; Zuker and Jacobson, 1995) was used to identify the complete secondary structures of the box H/ACA snoRNAs. Then a quantitative measure called *H*-num, discriminating between 'well-determined' and 'poorly-determined' structures on the energy dot plot (Zuker and Jacobson, 1995; 1998), was introduced as a decision guide (Fig. 4).

Sequence datasets To test its accuracy and applicability, the novel algorithm was applied to 30 nucleolar protein genomic sequences in the Locuslink of NCBI (Andersen *et al*., 2002; Pruitt and Maglott, 2001; Scherl *et al*., 2002) (Table 1). After collecting the data, pseudouridine maps of the target RNAs were constructed to assign pseudouridines to the candidates (Maden and Wakeman, 1988; Ofengand, 2002; Ofengand and Bakin, 1997) (Table 2). In addition, to train our algorithm, we derived box H/ACA snoRNA sequence datasets from yeast and human



level	length	i start	j start	h-num
1	3	78	113	1.0
1	4	14	44	1.0
1	4	90	101	1.0
1	6	3	56	1.0
1	6	18	39	1.0
1	7	68	122	1.0
1	7	82	109	1.0
3	2	1	61	2.0
4	2	1	59	2.0

Fig. 4. Tertiary sieve system. **A.** Energy dot plot of ACA19 snoRNA (accession number 38601882). The energy dot plot matrix function was ported from the mfold (Zuker, 2003). The top right triangle indicates the calculated probability matrix and the bottom left triangle indicates the current base pairs in the RNA structure (Matzura and Wennborg, 1996; Zuker, 2003). **B.** *H*-num table of ACA19 snoRNA. The start pairs of each stem have relatively low *H*-num values (1.0) compared to other 'poorly-determined' stems (2.0).

from DDBJ/EMBL/GenBank. After analysis of the primary and secondary structures of the snoRNA datasets, the structural differences between the snoRNAs of yeast and human were identified and yeast datasets were eliminated from the whole snoRNA datasets (Table 3).

Cut-off values of each sieve The primary and secondary structures of the snoRNA dataset were analyzed to adjust the cut-off values of each sieve system (Fig. 3) (Table 4). The hidden Markov model of H boxes was calculated by the log-odd score system and adopted as the cut-off values for the primary sieve. In the last sieve, the standard deviations (STD) of each snoRNA from the mean structure (Table 4) were calculated and used to eliminate false positives in the secondary sieve products. During the training phase, 49 out of 50 snoRNAs had STDs below 4 and only that of ACA 38, which has an abnormal structure between the H box and the ACA box, was 4.09. We conclude that secondary sieve products satisfying the cut-off value of STD < 4 are good candidate snoRNAs.

Table 1. The human nucleolar proteins screened.

Category	Nucleolus protein	Genomic sequence length (bp)	Accession No.
(A)	Ribosomal protein S12	3793	NT_025741.13
	Ribosomal protein S15a	8167	NT_010393.15
	Ribosomal protein L3	7539	NT_011520.10
	Ribosomal protein L4	6330	NT_010194.16
	Ribosomal protein L5	10671	NT_032977.79
	Ribosomal protein L18a	4197	NT_011295.10
	Ribosomal protein L21	5799	NT_024524.13
	Ribosomal protein L27a	3854	NT_009237.17
	Ribosomal protein L30 (reverse complement sequence)	4630	NT_008046.15
	Laminin receptor 1	6520	NT_022517.17
(B)	FUS interacting protein (serine-arginine)1	14682	NT_004610.17
	G-rich RNA sequence binding factor1	21147	NT_006216.14
	Histone 2, H4	1195	NT_004487.17
	Splicing factor, arginine/serine-rich2	4015	NT_010641.15
	Activator of basal transcription1	3898	NT_007592.14
(C)	Eukaryotic translation termination factor 1	37915	NT_034772.5
	Eukaryotic translation initiation factor 4A, iso1(EIF4A1)	6606	NT_010718.15
	Signal recognition particle 14 kDa	3947	NT_010194.16
	Eukaryotic translation initiation factor 5A	5605	NT_010718.15
	KIAA0111	12719	NT_024871.11
(D)	Dyskerin	15610	NT_025307.15
	Small nuclear ribonucleoprotein polypeptide D3	17691	NT_011520.10
	Nucleolar protein family A, member 1	10027	NT_016354.17
	Nucleophosmin 1	23831	NT_023133.12
	Nucleolar protein 5a (NOP56)	6585	NT_011387.8
(E)	DEAD box polypeptide (Asp-Glu-Ala-Asp)3	31874	NT_079573.2
	DEAD box polypeptide (Asp-Glu-Ala-His)9	49140	NT_004487.17
	DEAD box polypeptide (Asp-Glu-Ala-Asp)10	276634	NT_033899.7
	DEAD box polypeptide (Asp-Glu-Ala-His)15	57876	NT_006316.15
	DEAD box polypeptide (Asp-Glu-Ala-Asp)24	31089	NT_026437.11

(A) Ribosomal proteins; (B) Nucleotide binding and nucleic acid binding proteins; (C) Translation factors; (D) RNA modifying enzymes and related proteins; (E) Dead box proteins.

Table 2. (a) Large subunit rRNA pseudouridine map (*Homo sapiens*).

No.	Pseudouridine locations	Sequences of Ψ regions
1	U1515	5′-UGAAC Ψ AUGCC-3′
2	U1561	5′-GUCCG Ψ AGCGG-3′
3	U1656	5′-UUCCC Ψ CAGGA-3′
4	U1662	5′-CAGGA Ψ AGCUG-3′
5	U1723	5′-GCGAA Ψ GAUUA-3′
6	U1758	5′-CAACC Ψ AUUCU-3′
7	U1761	5′-CCUAU Ψ CUCAA-3′
8	U1771	5′-AACUU Ψ AAAUG-3′
9	U1838	5′-GCCAC Ψ UUUGG-3′
10	U1840	5′-CACUU Ψ UGGUA-3′
11	U3606	5′-CCGAC Ψ GUUUA-3′
12	U3608	5′-GACUG Ψ UUAAU-3′
13	U3664	5′-UGAUU Ψ CUGCC-3′
14	U3684	5′-GAAUG Ψ CAAAG-3′
15	U3699	5′-GAAAU Ψ CAAUG-3′
16	U3703	5′-UUCAA Ψ GAAGC-3′
17	U3727	5′-GGGAG Ψ AACUA-3′
18	U3731	5′-GUAAC Ψ AUGAC-3′
19	U3733	5′-AACUA Ψ GACUC-3′
20	U3737	5′-AUGAC Ψ CUCUU-3′
21	U3739	5′-GACUC Ψ CUUAA-3′
22	U3787	5′-AUGAA Ψ GGAUG-3′
23	U3791	5′-AUGGA Ψ GAACG-3′
24	U3813	5′-GUCCC Ψ ACCUA-3′
25	U3820	5′-CCUAC Ψ AUCCA-3′
26	U3822	5′-UACUA Ψ CCAGC-3′
27	U3853	5′-GGGCU Ψ UGGCG-3′
28	U3889	5′-UGAGC Ψ UGACU-3′
29	U3928	5′-AGGUG Ψ AGAAU-3′
30	U4253	5′-CUUGA Ψ CUUGA-3′
31	U4256	5′-GAUCU Ψ GAUUU-3′
32	U4259	5′-CUUGA Ψ UUUCA-3′
33	U4272	5′-ACGAA Ψ ACAGA-3′
34	U4313	5′-ACCUU Ψ UGGGU-3′
35	U4321	5′-GGUUU Ψ AAGCA-3′
36	U4363	5′-CUGGC Ψ UGUGG-3′
37	U4380	5′-AGCGU Ψ CAUAG-3′
38	U4383	5′-GUUCA Ψ AGCGA-3′
39	U4391	5′-CGACG Ψ CGCUU-3′
40	U4402	5’-UUUGA Ψ CCUUC-3′
41	U4417	5′-UCGGC Ψ CUUCC-3′
42	U4431	5′-CAUUG Ψ GAAGC-3′
43	U4460	5′-GAUUG Ψ UCACC-3′
44	U4481	5′-GAACG Ψ GAGCU-3′
45	U4491	5′-UGGGU Ψ UAGAC-3′
46	U4512	5′-CAGGU Ψ AGUUU-3′
47	U4536	5′-AUGUG Ψ UGUUG-3′
48	U4539	5′-UGUUG Ψ UGCCA-3′
49	U4588	5′-GACAU Ψ UGGUG-3′
50	U4596	5′-GUGUA Ψ GUGCU-3′
51	U4633	5′-UACCA Ψ CUGUG-3′
52	U4649	5′-AUGAC Ψ GAACG-3′
53	U4927	5′-AACCA Ψ UCGUA-3′
54	U4956	5′-CGGGG Ψ UUCGU-3′
55	U4965	5′-GUAGG Ψ AGCAG-3′

Table 2. (b) Small subunit rRNA pseudouridine map (*Homo sapiens*).

No.	Pseudouridine location	Sequences of Ψ regions
1	U34	5′-GCUUG Ψ CUCAA-3′
2	U36	5′-UUGUC Ψ CAAAG-3′
3	U105	5′-UUAAA Ψ CAGUU-3′
4*	U109	5′-AUCAG Ψ UAUGG-3′
	U110	5′-UCAGU Ψ AUGGU-3′
5*	U119	5′-GUUCC Ψ UUGGU-3′
	U120	5′-UUCCU UGGUC-3′
6*	U218	5′-GUGCA Ψ UUAUC-3′
	U219	5′-UGCAU Ψ UAUCA-3′
7	U220	5′-GCAUU Ψ AUCAG-3′
8*	U571	5′-UCCAC Ψ UUAAA-3′
	U572	5′-CCACU Ψ UAAAU-3′
9	U573	5′-CACUU Ψ AAAUC-3′
10	U681	5′-CGUAG Ψ UGGAU-3′
11*	U688	5′-GGAUC Ψ UGGGA-3′
	U689	5′-GAUCU Ψ GGGAG-3′
12	U801	5′-GCGUU Ψ ACUUU-3′
13	U814	5′-AAAAA Ψ UAGAG-3′
14	U815	5′-AAAAU Ψ AGAGU-3′
15*	U822	5′-GAGUG Ψ UCAAA-3′
	U823	5′-AGUGU CAAAG-3′
16	U863	5′-AGGAA Ψ AAUGG-3′
17	U866	5′-AAUAA Ψ GGAAU-3′
18	U918	5′-AUGAU Ψ AAGAG-3′
19*	U966	5′-GAAAU Ψ CUUGG-3′
	U968	5′-AAUUC Ψ UGGAC-3′
20	U969	5′-AUUCU Ψ GGACC-3′
21*	U1003	5′-AGCAU Ψ UGCCA-3′
	U1004	5′-GCAUU Ψ GCCAA-3′
22	U1056	5′-GACGA Ψ CAGAU-3′
23	U1081	5′-GACCA Ψ AAACG-3′
24	U1174	5′-GAGUA Ψ GGUUG-3′
25*	U1238	5′-GCGGC Ψ UAAUU-3′
	U1239	5′-CGGCU AAUUU-3′
26*	U1243	5′-UUAAU Ψ UGACU-3′
	U1244	5′-UAAUU Ψ GACUC-3′
27	U1248	5′-UUGAC Ψ CAACA-3′
28*	U1367	5′-UCUGG Ψ UAAUU-3′
	U1368	5′-CUGGU Ψ AAUUC-3′
29*	U1444	5′-ACUUC Ψ UAGAG-3′
	U1445	5′-CUUCU Ψ AGAGG-3′
30	U1625	5′-AUUAU Ψ CCCCA-3′
31	U1643	5′-GGAAU Ψ CCCAG-3′
32*	U1690	5′-UGCCC UUGUA-3′
	U1691	5′-GCCCU Ψ UGUAC-3′
	U1692	5′-CCCUU GUACA-3′

* One of the identified uridines is pseudouridine

Table 3. Structural differences of snoRNAs between yeast and human.

	Yeast (20)*	Human (50)*
Average Length (nt.)[1]	93.6	62
Sample standard deviation[2]	18.97	4.49

( )*, Number of sample box H/ACA snoRNAs.
[1], Region between H box and ACA box.
[2], The formula of sample standard deviation.

$$S = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Results

A preliminary test with sampled nucleolar protein genomic sequences Nucleolar protein genes that had been identified as the host genes of snoRNAs were used as the targets for screening (Andersen *et al*., 2002; Eliceiri, 1999; Leung *et al*., 2003; Nag *et al*., 1993; Scherl *et al*., 2002) (Table 1).

In order to provide a preliminary test of the performance of our algorithm, the genomic sequences of three proteins (RPS15a, ETF1, NOP56) were taken at random. In the ETF1 genomic sequence, the total number of hits acquired by the primary sieve was 692. Since the length between the H box and the ACA box is variable (50–80 bases) (Fig. 3B), several ACA boxes were found in the same H box positions. The final number of hits was 555 after removing the hits in the same H box positions. By applying *homeomorphically irreducible trees* (HITs) to the primary sieve products to eliminate false positives with pairs less than 10 in each hairpin region (Fig. 3B), the candidate regions were reduced from 555 to 76. Finally the region length model of the secondary sieve left only one secondary sieve product (from 22580 nt. to 22703 nt. in ETF1 genomic sequence). The tertiary sieve, the energy dot plot, was applied to this one secondary sieve product. However, due to the relatively high value of the *H*-nums (poorly-determined) of each stem and the STD of 4.27, it was not caught in the tertiary sieve. In the NOP56 genomic sequence, the total number of hits acquired by the primary sieve was 36. Of the 48 H boxes searched by the hidden Markov model, only 36 hits included ACA boxes in their downstream regions. Only one candidate region was caught in the secondary sieve (from nt 2918 to nt 3049 in the NOP56 genomic sequence). The *H*-nums of each stem of the candidate region were relatively low, hence well-determined (Zuker and Jacobson, 1995), and the STD value was 1.93. To verify this candidate region, its sequence was aligned with the homolog in *Mus musculus* (Fig. 5). The 93 percent identity with its mouse homolog suggested that the candidate was a true positive. BLASTN searches identified the candidate as

Table 4. The degree of dispersion of the snoRNA dataset from the mean structure.

snoRNAs	5′ proximal[1]	5′ distal[2]	3′ proximal[3]	3′ distal[4]	Standard deviation[5]
1. ACA37	2	14	3	16	1.66
2. ACA34	2	14	4	20	0.71
3. ACA32	3	14	3	22	1.73
4. ACA31	1	13	6	17	1.80
5. ACA30	2	14	3	18	0.87
6. ACA27	2	15	5	19	0.50
7. ACA23	2	14	4	17	1.12
8. ACA21	1	16	3	17	1.32
9. ACA19	2	14	5	19	0.71
10. ACA17	1	13	3	15	2.35
11. ACA10	1	16	3	18	1.00
12. ACA9	1	19	4	18	2.12
13. ACA7	2	17	7	22	2.35
14. ACA6	2	15	3	18	0.71
15. ACA3	1	14	2	17	1.58
16. ACA2b	2	14	4	18	0.71
17. ACA2a	2	14	3	18	0.87
18. ACA1	2	19	2	17	2.45
19. ACA59	2	17	5	26	3.67
20. ACA55	1	12	4	17	1.87
21. ACA54	2	14	2	17	1.50
22. ACA51	1	16	7	14	3.00
23. ACA49	2	15	4	17	1.00
24. ACA39	1	13	5	24	2.78
25. ACA38	1	14	5	27	4.09
26. ACA33	2	20	2	15	3.35
27. ACA29	2	16	6	18	1.22
28. ACA18	2	16	5	20	0.87
29. ACA11	2	16	6	18	1.22
30. ACA60	2	14	7	21	1.87
31. ACA50	1	17	7	19	1.87
32. ACA46	2	14	2	17	1.50
33. ACA44	2	12	5	15	2.55
34. ACA42	1	13	5	20	1.32
35. ACA41	1	14	2	16	1.94
36. ACA36	1	17	5	22	1.94
37. ACA28	1	14	4	19	0.71
38. ACA25	3	15	4	18	0.71
39. ACA24	1	15	5	22	1.66
40. ACA20	2	15	4	21	1.00
41. ACA15	2	15	4	22	1.50
42. ACA14b	2	15	3	19	0.50
43. ACA14a	2	15	3	19	0.50
44. ACA13	3	15	4	19	0.50
45. ACA8	2	16	3	15	2.12
46. ACA5	2	15	4	18	0.50
47. ACA58	2	16	3	25	3.08
48. ACA56	2	12	3	17	1.87
49. ACA52	1	16	3	20	1.00
50. ACA48	1	16	8	22	2.60
Mean distance	2	15	4	19	0

[1], Distance between 5′ proximal stem and H box; [2], distance between 5′ distal stem and H box; [3], distance between 3′ proximal stem and H box; [4], distance between 3′ distal stem and H box; [5], degree of dispersion of snoRNA structures from the mean structure.

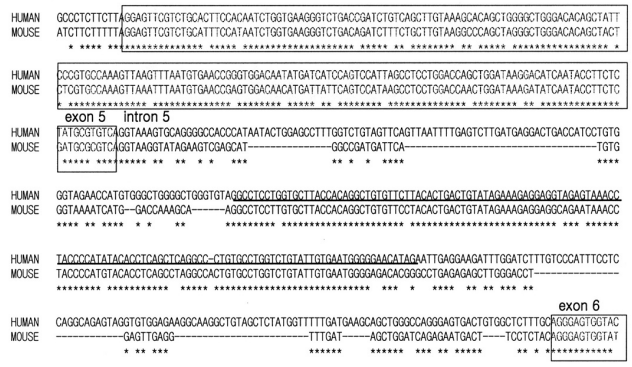


Fig. 5. Alignment of the human Nop56 genomic sequence (accession number NT_011387.8) with its mouse homolog (accession number NC_000068.3). The line below the human sequence indicates the candidate snoRNA (% identity: 93%). The sequence alignment was generated by the Clustal X (1.8) program.

ACA51 snoRNA. In the RPS 15a genomic sequence, although 142 hits with ACA boxes were detected by the primary sieve, no candidate regions were caught in the secondary sieve. Database searches confirmed that of the three nucleolar proteins, only the NOP56 genomic sequence contained a previously known snoRNA, ACA 51.

Searching for box H/ACA snoRNAs in nucleolar protein genomic sequences As a result of applying the algorithm to the 30 targets (Table 1), 13 candidates, including ACA 51 identified in the preliminary test, were caught by the last sieve (Table 5). None of the candidates were in the nucleotide-binding protein or dead-box protein genomic sequences. Before assigning the complementary regions of the target RNAs to the 13 candidates, BLASTN searches were performed to identify known snoRNAs among the candidates. Except for one candidate located in the intron 4 region of the *ribosomal protein L27a* gene, all the candidates turned out to be previously known snoRNAs (*Ribosomal protein S12*: ACA33; *Ribosomal protein L5*: U66; *Ribosomal protein L18a*: U68; *Ribosomal protein L21*: ACA27; *Ribosomal protein L27a*: ACA3; *Ribosomal protein L30*: U72; *Laminin receptor 1*: ACA6; *Eukaryotic translation initiation factor 4A*, *iso1*: U 67, ACA48; *Dyskerin*: ACA36, ACA56; *Nucleolar protein 5a*: ACA51). The candidate in the intron 4 region of the *ribosomal protein L27a* gene had the canonical primary and secondary structure. Moreover, alignment with the homologous region in the mouse yielded high sequence similarity (% identity of the region: 85%; % identity of the whole intron 4: below 40%) (Fig. 6). Also the pseudouridine at position 1367 of the SSU rRNA formed a bipartite duplex with the candidate (Table 2). These facts suggest it be a good candidate novel snoRNA.

SnoFront: Software implementation To reduce the manual processing required during the preliminary test, we developed a software pipeline called SnoFront implementing the primary and the secondary sieve algorithms. SnoFront uses training results from the snoRNA dataset, and locates box motifs. It also predicts the secondary structure of each hairpin in the input genomic sequences. SnoFront is available via electronic mail [ehs0328@snu.ac.kr].

Discussion

In the course of developing the method described above, we focused on reducing false positives and increasing

Table 5. Candidates detected in the objects.

Host gene	The number of candidates	The location in the genomic sequence	
Ribosomal protein S12	1	3044–3182 nt[1]	(Intron 5)[2]
Ribosomal protein L5	1	9058–9200 nt	(Intron 7)
Ribosomal protein L18a	1	3066–3200 nt	(Intron 3)
Ribosomal protein L21	1	4233–4359 nt	(Intron 4)
Ribosomal protein L27a	2	1836–1966 nt	(Intron 3)
		3047–3179 nt	(Intron 4)
Ribosomal protein L30	1	3732–3865 nt	(Intron 4)
Laminin receptor 1	1	2069–2229 nt	(Intron 2)
Eukaryotic translation initiation factor 4A, iso1 (*EIF4A1*)	2	2281–2425 nt	(Intron 3)
		5519–5667 nt	(Intron 9)
Dyskerin	2	6048–6186 nt	(Intron 8)
		12516–12656 nt	(Intron 12)
Nucleolar protein 5a	1	2918–3049 nt	(Intron 5)

[1], Location of candidate in the genomic sequence.
[2], Location of candidate in the host gene.

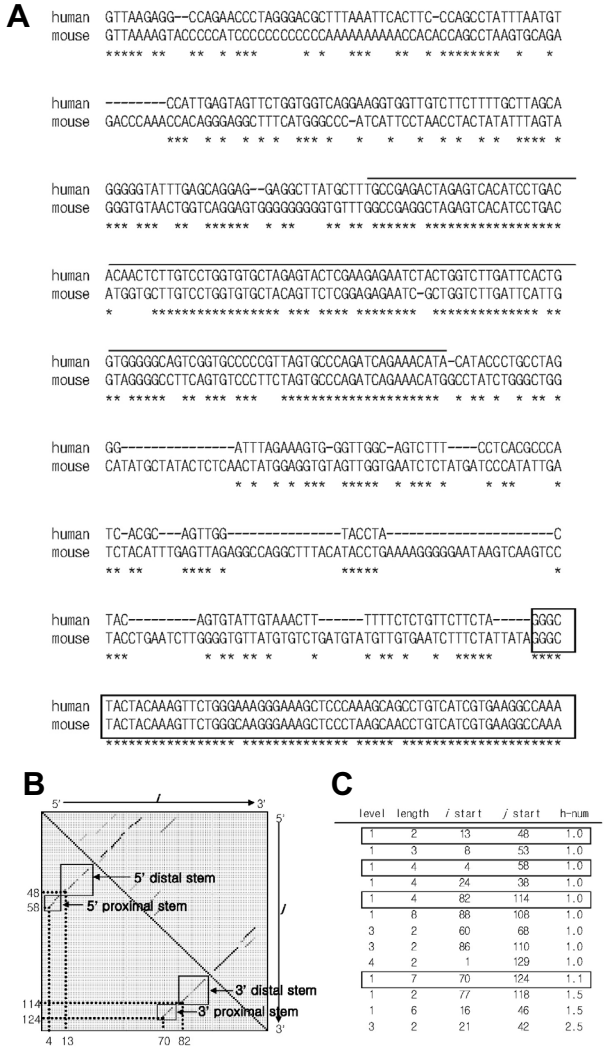


Fig. 6. Candidate in the *ribosomal protein L27a* gene. **A.** Alignment with the homologous gene of mouse (accession number NT_039445); Upper line on the sequence indicates the candidate region and the box on the sequence indicates the exon 5 region. The sequence alignment was generated by the Clustal X (1.8) program. **B.** Energy dot plot of the candidate. **C.** *H*-num table of the candidate; Boxes on each row indicate four "well-determined" stems.

efficiency. As a result, the final algorithm was able to identify one novel and 12 known box H/ACA snoRNAs in 30 nucleolar protein genomic sequences. However, the algorithm did not locate snoRNA, E2, in the *Laminin receptor 1* gene. E2 was caught by the secondary sieve, but just passed through the tertiary sieve because of its STD of 4.03. Although there are special cases with STDs > 4, such as ACA38, among the known snoRNAs (Table 4), we restricted the value to < 4 in order to reduce false positives more efficiently during application of the algorithm. However examination of the STDs of all the secondary sieve products revealed that all false positives among them had values above 4.20. So, to detect false negatives like E2, which have less canonical secondary structures, we would need to relax the tertiary sieve cut-off value.

Kiss *et al*. (2004) identified 61 novel putative box H/ACA snoRNAs and orphan snoRNAs which lack target pseudouridines in rRNAs and snoRNAs. Most of the host genes of the box H/ACA snoRNAs identified by Kiss *et al*. (2004) were found within the nucleolar protein genomic sequences constructed by Andersen *et al*. (2002) and Scherl *et al*. (2002). Approximately 350 proteins have been identified in the nucleolus, including proteins whose functions are not yet annotated (Andersen *et al*., 2002; Scherl *et al*., 2002). To search for novel snoRNAs that can fill the vacancies indicated by the target pseudouridine in rRNAs and other RNAs, the nucleolar protein genomic sequences will need to be examined by a combination of methods.

Acknowledgments This research was supported in part by a grant from the Korea Research Institute Bioscience and Biotechnology Research Initiative Program and the Korea Research Foundation (KRF-2002-070-C00080).

References

Andersen, J. S., Lyon, C. E., Fox, A. H., Leung, A. K., Lam, Y. W., *et al.* (2002) Directed proteomic analysis of the human nucleolus. *Curr. Biol*. **12**, 1–11.

Bachellerie, J. P. and Cavaille, J. (1997) Guiding ribose methylation of rRNA. *Trends Biochem. Sci*. **22**, 257–261.

Bachellerie, J. P., Cavaille, J., and Hüttenhofer, A. (2002) The expanding snoRNA world. *Biochimie* **84**, 775–790.

Decatur, W. A. and Fournier, M. J. (2003) RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.* **278**, 695–698.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) Markov chains and hidden Markov models; in *Biological Sequence Analysis*, pp. 46-79, Cambridge University Press, Cambridge UK.

Eddy, S. R. (1999) Noncoding RNA genes. *Curr. Opin. Genet. Dev*. **9**, 695–699.

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

Edvardsson, S., Gardner, P. P., Poole, A. M., Hendy, M. D., Penny, D., *et al.* (2003) A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction. *Bioinformatics* **19**, 865–873.

Eliceiri, G. L. (1999) Small nucleolar RNAs. *Cell. Mol. Life Sci*. **56**, 22–31.

Fontana, W., Konings, D. A. M., Stadler, P. F., and Schuster, P. (1993) Statistics of RNA secondary structures. *Biopolymers* **33**, 1389–1404.

Ganot, P., Caizergues-Ferrer, M., and Kiss, T. (1997) 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.

Grundy, W. N., Bailey, T. L., Elkan, C. P., and Baker, M. E. (1997) Meta-MEME: Motif-based hidden markov models of protein families. *Comput. Appl. Biosci*. **13**, 397–406.

Hofacker, I. L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*. **31**, 3429– 3431.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Chemical Monthly* **125**, 167–188.

Kiss, A. M., Jady, B. E., Bertrand, E., and Kiss, T. (2004) Human box H/ACA pseudouridylation guide RNA machinery. *Mol. Cell. Biol*. **24**, 5797–5807.

Kiss-Laszlo, Z., Henry, Y., Bachellerie, J., Caizergues-Ferrer, M., and Kiss, T. (1996) Site specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell* **85**, 1077–1088.

Krogh, A. (1998) An introduction to hidden Markov models for biological sequences; in *Computational Methods in Molecular Biology*, Salzberg, S. L., Searls, D. B., and Kasif, S. (eds.), pp. 45–63, Elsevier, Amsterdam.

Lafontaine, D. L. J. and Tollervey, D. (1998) Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. *Trends Biochem*. *Sci*. **23**, 383–388.

Lane, B. G., Ofengand, J., and Gray, M. W. (1995) Pseudouridine and $O^{2'}$-methylated nucleosides. Significance of their selective occurrence in rRNA domains that function in ribosome-catalyzed synthesis of the peptide bonds in proteins. *Biochimie* **77**, 7–15.

Leung, A. K. L., Andersen, J. S., Mann, M., and Lamond, A. I. (2003) Bioinformatic analysis of the nucleolus. *Biochem. J.* **376**, 553–569.

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

Maden, B. E. H. and Wakeman, J. A. (1988) Pseudouridine distribution in mammalian 18S ribosomal RNA. A major cluster in the central region of the molecule. *Biochem. J*. **249**, 459–464.

Matzura, O. and Wennborg, A. (1996) RNAdraw: an integrated program for RNA secondary structure calculation and analysis under 32-bits Microsoft Windows. *Comput. Appl. Biosci*. **12**, 247–249.

Nag, M. K., Thai, T. T., Ruff, E. A., Selvamurugan, N., Kunnimalaiyaan, M., *et al.* (1993) Genes for E1, E2, and E3 small nucleolar RNAs. *Proc. Natl. Acad. Sci. USA* **90**, 9001–9005.

Ofengand, J. (2002) Ribosomal RNA pseudouridines and pseudouridine synthases. *FEBS Lett*. **514**, 17–25.

Ofengand, J. and Bakin, A. (1997) Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryote, prokaryotes, archaebacteria, mitochondria and chloroplasts. *J. Mol*. *Biol*. **266**, 246–268.

Pruitt, K. D. and Maglott, D. R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140.

Scherl, A., Coute, Y., Deon, C., Calle, A., Kindbeiter, K., *et al.* (2002) Functional proteomic analysis of human nucleolus. *Mol. Biol. Cell* **13**, 4100–4109.

Tollervey, D. and Kiss, T. (1997) Function and synthesis of small nucleolar RNAs. *Curr. Opin. Cell Biol*. **9**, 337–342.

Tycowski, K. T., Smith, C. M., Shu, M. D., and Steitz, J. A. (1996) A small nucleolar RNA requirement for site-specific ribose methylation of rRNA in *Xenopus*. *Proc. Natl. Acad. Sci. USA* **93**, 14480–14485.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. **31**, 3406–3415.

Zuker, M. and Jacobson, A. B. (1995) 'Well-determined' regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. *Nucleic Acids Res.* **23**, 2791–2798.

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