SnoReport

Computational identification of snoRNAs with unknown targets

Jana Hertel

Institute for Theoretical Chemistry, University of Vienna
Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig

February 22, 2007

Outline

1. Introduction
2. Materials and Methods
3. Results
4. Summary

Non-coding RNA

- codes for RNA genes
- transfer RNAs, ribosomal RNAs
  → involved in translation and gene expression
- micro RNAs, small nuclear and small nucleolar RNAs, ...
  → mainly essential regulatory functions within the cell
- imprecise defined or missing gene borders makes **identification of novel genes difficult**

Computational prediction of non-coding RNA genes

`RNAz`[1]

method: machine learning techniques to predict novel ncRNA genes

basis: multiple sequence alignment

features: thermodynamical stability and structural conservation

result: numerous putative ncRNA genes, many of them not annotated

next: **Annotation** to specific ncRNA class

- `RNAmicro`[2] - Detection of miRNAs
- **`SnoReport`- Detection of snoRNAs**

[1] Washietl *et. al.Fast and reliable prediction of noncoding RNAs*.Proc.Natl.Acad.Sci.U.S.A.2005

[2] Hertel & Stadler *Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data*. Bioinformatics 2006

SnoRNAs

- involved in processing and modification of other RNAs
- H/ACA, C/D box snoRNAs and scaRNAs
- guide and orphan genes

Detection **without using targets** and
with using conservation information

SnoReport

method: machine learning techniques (support vector machine)

basis: multiple sequence alignment or single sequence
no need of target sequences

features: sequence-structure based attributes and thermodynamical stability, structural conservation for alignments

purpose: predicting novel snoRNAs **and** distinguishing both major classes (H/ACA and C/D box snoRNAs)

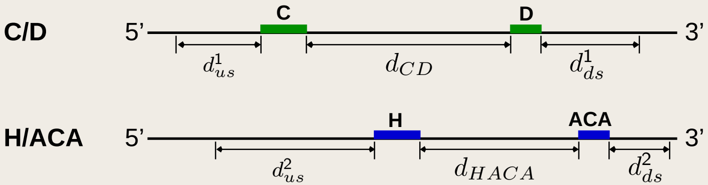
Data Sources

- Positive samples: H/ACA and C/D box snoRNAs from `snoRNABase`
- Negative samples: tRNAs, miRNAs, snRNAs, RNAse P, etc. from `Rfam`

	C/D		**H/ACA**	
	single	aligned	single	aligned
pos. samples	77	25	70	55
neg. samples	1486	535	231	223

SnoReport Workflow

1. finding characteristic sequence motifs (boxes)
2. truncate sequence according to box positions and estimated number of upstream and downstream regions
3. structure prediction, box positions prevented from pairing

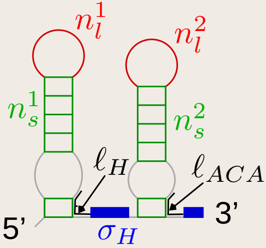


Classification only if match score of boxes > 0.5 and appropriate structure.

SnoReport Workflow - Feature vector

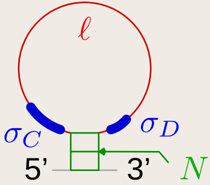
4. compute feature vector

H/ACA snoRNA



E_{diff}
GC content
ρ stemratio

C/D snoRNA



E_{diff}
GC content
L length

SnoReport Workflow - Classification

5. SVM classification: rbf kernel, probability estimates

- 2 classifications: H/ACA **and** C/D box snoRNA.
- Best classification probabilities for each class returned.

Alignment as Input

1. finding boxes in consensus sequence
2. truncate alignment
3. alignment structure prediction, box positions prevented from pairing
4. compute sequence and structural conservation features:

$$SCI = \frac{mfe_{cons}}{\overline{mfe}_{sgl}} \qquad S_{\xi} = -\frac{1}{\ell(\xi)} \sum_{i \in \xi} \sum_{\alpha = A,C,G,U} p_{i,\alpha} \ln p_{i,\alpha}$$

5. compute same features as for single sequences out of consensus sequence
6. SVM classification

Test Statistics

- 4 models

test: cross-validation with randomly distributed datasets

- using MSA increases statistical values

	C/D		H/ACA	
	single	aligned	single	aligned
sensitivity	0.65	0.92	0.82	0.98
specificity	0.98	0.99	0.96	0.99

- runtime independent of sequence length - truncation
- decelerating factor: number of sequences

Further Comparisons

- `SnoReport` applied to snoRNAs reported in recent publications
- Deng *et al.* 2006, *Caenorhabditis elegans*:
  41 C/D and 47 H/ACA + novel not further classified predictions
  - 20 C/D + 5 (2 missclassified)
  - 24 H/ACA + 3
- Yang *et al.* 2006, `snoSeeker`, *Homo sapiens*:
  21 C/D and 32 H/ACA box snoRNAs
  - 4 (2 confirmed) C/D
  - 19 (7 confirmed) H/ACA
- Zemann *et al.* 2006, *C. elegans, C. briggsae*:
  121 snoRNAs
  - 16/48 (novel), 20/28 (confirmed) and 8/11 (known) C/D
  - 5/11, 26/37 and 5/11 H/ACA

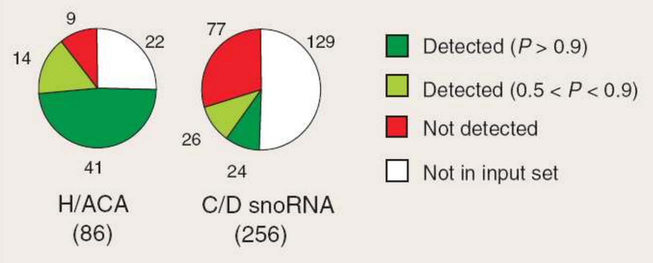
Further Comparisons ctd.

- Huang *et al.* 2005, nematodes:
  8/17 C/D and 11/16 H/ACA
- Accardo *et al.* 2006, *Drosophila melanogaster*:
  8/19 confirmed C/D box snoRNAs + 6 unconfirmed
- Liang *et al.* 2006, *Leishmania major*:
  22/62 C/D and 0/37 H/ACA-*like* box snoRNAs

`SnoReport` detected many of the snoRNAs from other approaches, mainly confirmed ones.
H/ACA-*like* snoRNAs in *Leishmania* quite different to the canonical ones in human and yeast.

Comparative Genomics Data

- RNAz based comparative genomics survey ⇒ ∼ 207000 alignments
- SnoReport: 1240 C/D and 1458 H/ACA box snoRNAs



Conclusions

- Recognition and classification of both major snoRNA classes
- `SnoReport` does not rely on targets in rRNA or snRNA
- Trained on mammalian data, `SnoReport` perfoms satisfactorily on nematodes and insects and even distant eukaryotes
- Suggestion of a large number of orphan snoRNAs hidden in mammalian genomes

Further work

- `SnoReport` designed to be easily retrained when more data comes available
- Recently published novel snoRNAs in other species than mammals will be used to create additional alignments
  ⇒ improve sensitivity on phylogenitical distant sequences (e.g. Leishmania)

Acknowledgement

Thanks to Peter F. Stadler and Ivo L. Hofacker

Thank you.

Financial support by the German *DFG* in the framework of the Bioinformatics Initiative (BIZ-6/1-2) and the SPP “Metazoan Deep Phylogeny”, as well as the Austrian GEN-AU project “non-coding RNA” is gratefully acknowledged.