

Searching for ncRNAs in protist genomes.

Lesley J. Collins^{1*}, Thomas J. Macke² and David Penny¹

¹ Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Private Bag 11222, Palmerston North, New Zealand.

² Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

Keywords: Noncoding RNA, RNAmotif, U5 snRNA, RNase P, *Giardia lamblia*.

Introduction

Non-coding RNAs make transcripts that function as RNA, rather than encoding proteins e.g. ribosomal-RNA (rRNA) and transfer-RNA (tRNA) [1]. They often form part of RNA-protein complexes (Ribonucleoproteins or RNPs) and play vital roles in essential cellular processes such as protein metabolism and splicing. Searching databases for homologs based on sequence similarity is only useful for the most slowly evolving or large ncRNAs like ribosomal RNAs, and becomes much less reliable for other snRNAs. If there is also large evolutionary distance between the species that is being searched and the species for which the gene is known, sequence similarity methods often fail to uncover any potential gene candidates for further analysis and confirmation as ncRNA gene homologs [1].

Noncoding RNAs usually fold into characteristic secondary structures and also can contain small sequence motifs. RNAmotif [2] is a program that uses this information to find ncRNA gene candidates with the design of an appropriate “descriptor” to model secondary structure and sequence motifs. However, in the past, descriptors that had to take inter-species structural variation into account could run into problems with overloading of the results file.

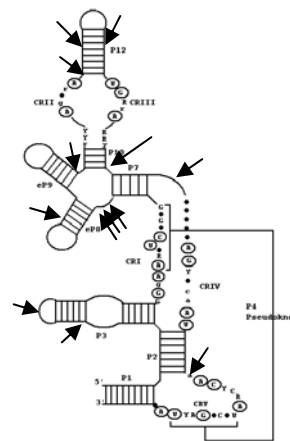
Here we show how the use of a user-defined scoring section, post-function commands and parallel implementation can help in reducing the problems associated with ‘looser’ descriptors. We describe the design and implementation of descriptors for two ncRNAs, the U5 snRNA and the eukaryotic RNase P RNA. These genes have conserved and variable sequence and structure areas which allowed the identification of gene candidates in some protist genomes such as *Giardia lamblia* [3] and *Encephalitozoon cuniculi* (a microsporidian).

Results and Discussion

Descriptors for the U5 snRNA were designed with a unique motif scoring section allowing important motif presence or absence to be seen at a glance in the results file. This is necessary because sometimes scoring regimes can add up individual motif scores in such a way that a sequence can be given a higher score even if an extremely important motif is missing. Spreadsheet sorting can also be used to detect sequences containing a certain motif. RNAmotif is a processor-intensive program. Even a short descriptor can run into problems when searching a large genome database and often the program will not run to completion in this situation. Parallel computing is one solution with large databases being split into smaller pieces, each piece run on a separate node, and then the results collated in a single result file. Getbest was also incorporated into the parallel implementation as a –post command, filtering the results from each worker node to give a more condensed results file. This reduced the space required for the results file and enabled realistic sequence analysis of the results

The U5 descriptors were tested by searching against the genomes of *E. cuniculi* and *Plasmodium falciparum* for which the U5 snRNA genes were already known. Application of this technique resulted in new U5 snRNA gene candidates from the genomes of *Ciona intestinalis* (sea squirt) and *Giardia lamblia*.

The RNase P RNA has a ‘reasonably’ conserved secondary structure with parts of this structure being highly conserved, even between all three kingdoms, and other parts are variable[4]. Figure 1 shows the consensus eukaryotic RNase P RNA secondary structure; circled nucleotides are conserved between all three kingdoms and arrows point to areas where helices may be inserted in some species. This secondary structure variability creates a challenge for designing a ‘working’ descriptor for the RNase P RNA. A descriptor was designed for part of the eukaryotic RNase P RNA secondary structure consisting of the P3-CRI-P7-P10 section of the whole secondary structure and was called the ‘P7 descriptor’. It was found that descriptors designed for the full eukaryotic secondary structure were computationally-prohibitive, requiring weeks to search the simplest databases. Results from the RNAmotif scan were then analyzed for downstream conserved elements (e.g. CRV pseudoknot pairing and CRIV consensus sequence) to find a viable candidate sequence.



Testing of this descriptor against a database of known RNase P RNAs from all three kingdoms (Bacterial, Archaea and Eukaryotes) showed that it had specificity for the Eukaryotic RNase P RNA. Searches with the P7 descriptor against *E. cuniculi* genome recovered 14 sequence areas with only one having a viable CRV consensus sequence. When this sequence was BLASTed against GenBank, rat and mouse RNase P RNAs were returned with low scores. Genome searches against *G. lamblia* recovered eleven sequence areas with the top score, each having the Eukaryotic CRI consensus sequence. Of these only one had a viable downstream CRV area and upon further examination generally fitted the Eukaryotic RNase P RNA consensus secondary structure having the eukaryotic consensus sequences for the CR (I-V) regions in correct locations.

The RNAmotif genome searching procedure described in this study cannot guarantee finding a particular ncRNA in a particular genome. Other factors such as the quality of the genomic data and the phylogenetic distance between the species from which the known ncRNAs (the ones used to design the descriptor) and the genome being searched may also play a large role in the search outcome. This procedure, however, does offer a new way of searching for some sometimes hard-to-find ncRNA genes. Finding these genes in protist genomes may help in our understanding of the evolution of RNA metabolism from the earliest Eukaryotes.

Acknowledgements:

Many thanks to the administrators of the Helix parallel processing facility at Massey University, Albany, New Zealand for their help and advice. Many thanks to Mitchell L. Sogin, and Andrew G. McArthur and their teams at the *Giardia lamblia* Genome Project (funded by the NIAID/NIH under cooperative agreement AI 043273), Marine Biological Laboratory at Woods Hole for access to non-public data. Other genomes were downloaded from GenBank. Thanks also to Anu Idicula, Alicia Gore and Trish McLenachan for performing preliminary experimental analysis on some of the ncRNA gene candidates. This work was supported by the NZ Marsden Fund.

References:

- [1] Eddy, S. (1991) Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, **2**, 919-929.
- [2] 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.
- [3] McArthur, A.G., Morrison, H.G., Nixon, J.E.J., Passamaneck, N.Q.E., Kim and 10 others (2000) The *Giardia* genome project database. *FEMS Microbiology Letters*, **189**, 271-3.
- [4] Frank, D.N., Adamidi, C., Ehringer, M.A., Pitulle, C. and Pace, N.R. (2000) Phylogenetic-comparative analysis of the eukaryal Ribonuclease P RNA. *RNA*, **6**, 1895-904.