Structured non-coding RNAs

Prediction, Comparison, Annotation

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt
von Diplom Informatikerin *Jana Hertel*
geboren am 11. September 1981 in Leipzig

Die Annahme der Dissertation haben empfohlen:

1. Professor Dr. Peter F. Stadler (Leipzig, Deutschland)
2. Professor Dr. Hsien-Da Huang (Hsinchu, Taiwan)
3. Professor Dr. Daniel Gautheret (Paris, Frankreich)

Die Verleihung des akademischen Grades erfolgt auf Beschluss des Rates der Fakultät für Mathematik und Informatik vom 20.04.2009 mit dem Gesamtprädikat *summa cum laude*.

Structured non-coding RNAs

Prediction, Comparison, Annotation

Summary (Zusammenfassung)

Over the last decade, the idea of non-protein-coding RNAs as an abundant class of regulators in eukaryotic cells has gained more and more evidence and interest. Large scale transcriptome studies confirmed that the eukaryotic genome is characterized by a manifold mosaic of overlapping, bi-directional transcripts and a plethora of non-protein-coding gene products arising from the same locus. Even in basal Metazoa like Cnidaria and Placozoa and fungi more such transcripts are reported than previously thought.

Numerous experimental and computational approaches are developed with the aim to identify novel ncRNAs (small RNAs) and to study their functions. Bioinformatic approaches detect novel genes using sequence and structure comparison to known small RNA classes to determine the relationship of unknown sequences. Obviously, the combination of both experimental and computational methods is a powerful method to describe the RNA compartment of certain species.

This work focuses on the computational part of novel RNA gene-finding, reliable RNA annotation and evolutionary examinations. Furthermore, it is demonstrated that the modifications of already well-established algorithms enhances the detection of novel RNAs, their alignment and structure prediction.

The minimum free energy (*mfe*) algorithm for the prediction of an RNA secondary structure and its partition function variant are modified to return only canonical structures. Isolated basepairs that are not enclosed by an additional pair are not allowed. While this modification requires only minor changes in the *mfe* recursions, more complex changes are necessary to incorporate this feature into the partition function. It has been proved that computing canonical structures increases the accuracy of the prediction and, since fewer basepairs have to be considered in the partition function, there is also a moderate performance gain.

Many ncRNA classes can be categorized into families due to sequence similarity. Alignments of the members of RNA families are therefore a useful approach to infer their phylogenetic distance and evolutionary "behaviour". In the case of RNA, sequence alignments alone are not always sufficient for particular subsequent analyses. One of the established programs that consider both primary and secondary structure to create alignments of RNA genes, `LocARNA` (Will *et al.*, 2007), was modified to a scanning variant with an adjusted scoring function that acts as gene-finder. The correct alignment of certain RNAs that show high sequence variation in structured regions and almost no sequence variation in unstructured sequence motifs became a serious problem during this thesis. Therefore, `LocARNA` was further modified such that sequence motifs and structural constraints can be included in the algorithm. This modification significantly improves box H/ACA snoRNA alignments, for example, even if the sequences are *not* related. The `RNAsalsa` program was developed to produce structure-aware sequence alignments and individual reasonable secondary structures of large RNAs. It was primarily designed to infer phylogenetic relationship between species. Beyond primary and secondary structure, many functional small RNAs show characteristic motifs in their tertiary structure. The small tool `motifSearch.pl` was developed that scores such motifs and is able to find them in genomic sequences and alignments.

RNA gene-finding is a major task in this work. Based on sequence homology novel members can be assigned to known non-coding RNA families. The `GotohScan` program is introduced which implements a semi-global alignment algorithm with affine gap costs that detects even highly diverged homologs in distantly related species. Apart from such homology based approaches, novel RNA genes can be categorized into non-coding RNA classes by combinations of sequence and structure features. For micro RNAs and small nucleolar RNAs this problem is solved using machine learning in the `RNAmicro`

and `SnoReport` programs. These tools, notably contribute to the annotation of the RNA complement of eukaryotic species. Naturally, there is a plethora of other computational approaches for different tasks in RNA annotation. Since the jungle of bioinformatic methods is somewhat confusing for other scientists, an over-viewing chart on combining these tools in a reasonable order is provided. This recipe enables to comprehensively annotate the RNA complement of newly sequenced genomes.
According to this protocol several "real world" studies have been performed. In the pilot phase of the ENCODE project a large set of structured RNAs could be detected using `RNAz` (Washietl *et al.*, 2007) and `EvoFold` (Pedersen *et al.*, 2006), respectively. Beyond known ncRNAs a number of high-scoring candidates were predicted and some of them verified experimentally. This survey suggested that it is likely, that the number of functional RNAs in the ENCODE selected regions is even higher than the estimates of the two programs.
A number of smaller RNA annotation projects were performed in accordance to our RNA annotation recipe. The genomes of Drosophilids, *Leishmania*, flatworms, *Aspergillus* fungi and the placozoan *Trichoplax* were analysed. Two of these projects are introduced in this work. Non-protein-coding RNAs of *Trichoplax adhaerens* and *Aspergillus fumigatus* were detected computationally and partially experimentally verified by our co-authors in Hannover (lab of B. Schierwater) and Innsbruck (lab of A. Hüttenhofer), respectively.
In order to complete this work, the evolution of non-coding RNA genes is inferred at the end. In particular the innovation and duplication history of miRNAs is described on a huge set of miRNA families and an update of the illustrating example of the mir-17 miRNA cluster.
In summary, the combination of experimental verification, and computational prediction of novel RNA genes and the assignment of unknown sequences to their respective classes turns out to be a powerful way to annotate the RNA compartment of eukaryotic species. Genes that are expected but missed experimentally, often can be detected using alternative computational methods. While the computational searches often are exhaustive, these methods can be re-calibrated and, therefore may work more precisely in the future, when experimental support for the predictions becomes available. My general experience in the development of gene-finding and annotation methods increased significantly during this work. An enriched view on the problems raised new ideas to extend the introduced programs and methods such that they work more precisely.
All in all, this thesis demonstrates that deliberate bioinformatic methods that are carefully tested on known data highly contribute to a reliable RNA annotation in eukaryotic species. Basic methods and their variations comprise a powerful platform for the sensitive extraction of RNA class specific features and their classification; and for augmented alignments that are more "natural" than naive approaches. Their combination yield in a compelling set of programs that faithfully help to dig deep into the theoretical RNA world of the organisms.

References

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W. & Haussler, D. (2006). Classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*, **2**, e33.

Washietl, S., Pedersen, J., Korbel, J., Gruber, A., Hackermüller, J., Hertel, J., Lindemeyer, M., Reiche, K., Stocsits, C., Tanzer, A., Ucla, C., Wyss, C., Antonarakis, S., Denoeud, F., Lagarde, J., Drenkow, J., Kapranov, P., Gingeras, T., Snyder, M., Gerstein, M., Reymond, A., Hofacker, I. & Stadler, P. (2007). Structured RNAs in the ENCODE Selected Regions of the Human Genome. *Genome Res*, **17**, 852–864.

Will, S., Missal, K., Hofacker, I. L., Stadler, P. F. & Backofen, R. (2007). Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, **3**, e65.