## Poster H-46

**Telomerase RNA searching using an Infernal-based pipeline**


**Authors:**
Ariane Machado Lima (*Instituto de Matemática e Estatística, Universidade de São Paulo*)
Sean R. Eddy (*Department of Genetics, Washington University*)
Hernando A. del Portillo (*Instituto de Ciências Biomédicas II, Universidade de São Paulo*)
Alan Mitchell Durham (*Instituto de Matemática e Estatística, Universidade de São Paulo*)

**Short Abstract:** Infernal is a tool that builds RNA probabilistic models from structural multiple alignments. We are developing an Infernal-based pipeline to search telomerase RNAs. In this work we make a preliminary assessment of this pipeline and of the phylogenetic reach of the vertebrate TERC alignment through cross-validation technique.

**Long Abstract:**

In recent years, the discovery of new classes of noncoding RNAs (ncRNAs) has motivated the search of genes of these ncRNAs in several organisms. However, the problem of ncRNA search is still open. Despite the availability of many tools and strategies used for searching coding sequences, such resources are not suitable for ncRNAs. One of the main reasons is that ncRNAs tend to conserve less sequence similarity than coding sequences, being characterized more for secondary structure conservation, which is essential for their functionality. Consequently, depending on the ncRNA family and the divergence level, instead of sequence similarity we may observe covariation of complementary bases. Therefore, the ncRNA research demands tools specific for this purpose.

An important ncRNA family, the TElomerase RNA Component (TERC) family, is part of the telomerase complex. In many organisms, this complex is responsible for the maintenance of the telomeres, a structure that forms the ends of linear chromosomes. The telomeres are involved in many roles, such as chromosome stability and avoidance of genetic information loss during replication cycles. The catalytic activity of the telomerase complex is performed by a reverse transcriptase that synthesizes telomeric DNA from a RNA template, the TERC. Similar to other ncRNAs, the TERC secondary structure is important for its function. Up to date only three TERC families have been characterized: ciliates, vertebrates and yeasts. For each one of these groups there is a characterized consensus secondary structure. However, due to a high mutation rate, TERCs across these groups are different in sequence, structure and length. This makes the TERC searching in more phylogenetically distant species a hard problem to be addressed. Particularly two organisms, Plasmodium falciparum and Caenorhabditis elegans, are being searched both by biological and computational methods. However, the TERC has not been found in these species yet. C.elegans is a model organism, specially for the study of cellular differentiation. TERC identification in this organism can help to understand if and how alterations in telomere length affect biological processes in C.elegans. P.falciparum is the parasite that causes the most lethal human malaria. Due to a high replication demanding inside the human host, telomerase must be essential for the survival of the parasite. This makes the TERC a target for drug development.

One of the most popular systems for ncRNA characterization and searching is Infernal. Infernal is a tool that uses Covariance Models (CMs) to characterize RNA families. CMs are a special case of Stochastic Context-Free Grammars, and so is able to detect nested dependences of arbitrary distances. This ability allows the characterization of RNA secondary structures, which can be defined by nested dependences between complementary base pairs. We are developing a pipeline to characterize and search TERCs using Infernal and a set of heuristics that perform genome pre-processing to select pre-candidate, candidate post-processing, and modifications on the generated CMs. We have also developed sets of parameter settings specific for TERC searching, as well as specific null models. The Infernal search algorithm, although efficient, is still slow for scanning long genomes. The pre-processing aims to decrease this running time and to reduce the false positive rate. Reducing the false positive rate is also the goal of the modifications on the Covariance Models. The pipeline, for instance, when applied on the human chromosome 3 (201Mb), detected only one candidate: the real known TERC (to the present knowledge, the TERC is a single gene).

Although we are working on a general TERC characterization, we are particularly interested in TERC identification in P.falciparum and C.elegans. Despite our efforts to adjust good TERC models, pipeline and infernal parameters, we could not detect the TERC in these organisms yet. Since P.falciparum and C.elegans are distant of the three TERC families mentioned above, we hypothesize their TERCs are too diverged from the TERCs known up to date. Therefore, in order to the phylogenetic reach of the TERC models, we are doing a cross-validation positive controls.

If our TERC models are not able to detect new TERCs in a phylogenetic distance necessary to reach P.falciparum and C.elegans, we have to build brigdes between them. The strategy is to search new TERCs in organisms phylogenetically closer to that ones with known TERCs in order to, step by step, enrich the existent models, making them more powerful to search the P.falciparum and C.elegans genomes. To develop this search strategy, the best candidate families to start would be ciliates and yeasts. However this approach is not viable. First, despite the phylogenetic closeness between P.falciparum and ciliates, there is a lack of sequenced genomes of species between them. Second, TERCs with completely characterized secondary structure from yeasts are from only one genus, Saccharomyces, and do not align with other TERC families, even with the other yeast TERCs, Kluyveromyces. In addition, yeast TERCs are very unusual, being approx. 4 times longer than vertebrate and 7 times longer than ciliate TERCs. Vertebrate TERCs, on the other hand, present the deepest alignment: 35 species, from shark to human. Furthermore, there are many vertebrate and superior-organism genomes that are sequenced and available for searching.

In this poster we make a preliminary assessment of our pipeline and of the phylogenetic reach of the vertebrate alignment through cross-validation technique. Iniatially, for each species, we remove its TERC from the vertebrate alignment and generated a CM model from the alignment having the remaining 34 species. For species whose genome is available, we scanned its genome, in order to have more accurate measures of sensitivity and specificity. For each species without available genome, we scanned an artificial genome made up of the concatenation of its GenBank sequences, including the TERC. If the TERC is identified, we follow removing the most similar TERC sequence from the alignment and repeat the search.

These results give us an idea about the phylogenetic distance that our pipeline can reach using the vertebrate TERC alignment.