Whelan Lab Journal Club: Seemann et al. (2011)

James E. Allen

7th April 2011

Members of Simon Whelan's lab at the University of Manchester participate in a regular journal club, where a paper with an evolutionary/phylogenetic slant is discussed. Here, I briefly describe a paper that I recently presented, and summarize our conclusions about the methodology and results. (I have attempted to represent the discussion and consensus of the group, but any inaccuracies are my own.)

[If referring to this document, please cite its location on the Monkeyshines website: http://www.monkeyshines.co.uk/blog/archives/534]

References

Seemann, S.E., Gorodkin, J. and Backofen, R. (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Research* **36**, 6355-6362.

PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. Stefan E. Seemann, Andreas S. Richter, Tanja Gesell, Rolf Backofen and Jan Gorodkin (2011) *Bioinformatics* 27: 2, 211-219. PubMed: 21088024 (Presented by James Allen, 31st March 2011)

The paper in a sentence: Multiple sequence alignments and methods of RNA structure prediction can be combined to detect interactions between non-coding RNA molecules, in order to better understand their function.

Background: The biological importance of ribosomal and transfer RNA (rRNA and tRNA) has long been recognised, but in the last 10 years other non-coding RNA molecules (ncRNA) have been shown to have a variety of biological roles. As with proteins, the structure of, and interactions between, ncRNA define their functions, which are generally poorly understood.

The paper in detail: There are many methods available for predicting RNA secondary structure, one of which (`PETfold`: Seemann, et al. 2008) forms the basis for this paper. Briefly, the idea of `PETfold` is to combine information about the thermodynamic folding potential with explicit evolutionary models that rely on the covariance of base pairs in multiple sequence alignments. Interactions between RNA molecules can be viewed as a relatively simple extension of the structure prediction of a single RNA molecule, if it assumed that the interactions are primarily defined by the same process of canonical base pairing that creates the stems in RNA structures.

The `PETcofold` program concatenates two alignments of RNA that are assumed to interact, then attempts to predict both the structure of each RNA and their interactions in a two-stage process. (The authors term this hierarchical folding, which perhaps implies the potential for more than two stages, but this does not seem to be possible in the current implementation.) In the first stage, the structure of each RNA is predicted independently, and base pairs that are reliably predicted (i.e. above a particular threshold) are marked as constrained. In the second stage, further pairs are predicted for the unconstrained bases, both within and between the two RNA molecules. Such a process allows for the modelling of pseudoknots, which may have an important role in the function of RNA molecules. Some members of our group questioned whether this process was a biologically plausible scenario of RNA folding and interacting, but until we know more about the mechanisms involved it is probably reasonable. In any case, there is no theoretical reason why the same

software cannot constrain and unconstrain base pairing in a more complex manner, to model biology more realistically.

The evaluation of `PETcofold` is somewhat problematic, as there is no well-defined dataset of RNA-RNA interactions against which to test the program. The authors gather a set of 32 interactions, of 13 different bacterial ncRNA molecules, based on experimental evidence, and use this to test different parameters of the model. It appears that by allowing only the most reliable base pairs in the first stage, the interacting sites are predicted with optimal accuracy, with a mean MCC of around 0.5. The MCC indicates the trade-off between sensitivity and positive predictive value (PPV), but these measures are not provided, so it is not clear in which area the program performs well (e.g. an MCC of 0.5 could arise from sensitivity=1 and PPV=0.25; from sensitivity=0.25 and PPV=1; or somewhere in between). Evaluating structure and interaction prediction was even trickier, with only four examples of interactions in the literature where the structure of two interacting molecules was also known. And one of these has to be discarded as being probably incorrect, so although `PETcofold` performs better than other programs, this is a rather small set on which to base general conclusions.

The authors find that very few interacting sites in their dataset show evidence of the covariance that their model is designed to detect, and thus use simulations to demonstrate that the model can take advantage of covariance if it does exist. However, the manner in which they introduce covariance, by multiplying an underlying tree by a factor of up to 200, is perhaps not ideal; this simulates a large amount of evolution across all sites in the RNA, which may affect the results.

It is rather optimistically stated in the discussion that `PETcofold` could be used to predict interactions between the results of genomic screens for ncRNA; however, given that such scans return thousands of results, the requisite pairwise combinations of predicted RNAs will be prohibitively computationally expensive.

Journal club conclusion: The program `PETcofold` is a potentially useful way to predict canonical base-pairing between RNA molecules, using covariance information. It is not yet clear whether there is, in practice, sufficient detectable covariance in such interactions, but if not, the thermodynamic aspect of the program, and the application of hierarchical folding, may result in predictions that are as good as, or better than, more complicated programs, and in less time.