

Supervisors: Rune Lyngsoe, Jotun Hein

Project: Structural Analysis of Aptamers

RNA is mostly known as the messenger molecule in the translation of DNA to proteins. However, RNA also has structural and functional properties. Combined with the relative ease with which random or specific RNA molecules can be produced, this makes them ideal design molecules for e.g. targeting particular proteins. Mimicking natural selection in fast forward mode, from a huge set of random RNAs the ones that binds the best to the target molecule are selected. By several iterations of amplifying the current pool of RNAs and reselecting the ones binding the best, the end result is a pool of RNAs with very good affinity to the target. These molecules can then be sequenced to obtain *aptamers* for the target.

Really what we want to discover by these experiments are not the aptamers themselves, but rather the mode with which the aptamer binds to the target, or binding motif. This will reveal information that can be used to design more efficient molecules (usually not RNAs) for binding the target and possibly carrying out a biochemical function at the binding site.

The binding of an RNA molecule to the target is usually reliant on the overall structure of the RNA and the particular bases found in a few key positions. The structure, again, is to a large extent dominated by the formation of base pairings. These base pairings are similar to the ones observed in the DNA double helix formation. The set of base pairs is known as the secondary structure of the RNA, and reasonably successful techniques exist for computationally determining the secondary structure given an RNA sequence.

So one relevant formulation of the structural analysis of aptamers problem is this: group the aptamers in clusters where all aptamers in a cluster share a common secondary structure and zero or more constant positions. There are many ways to approach this problem. Our collaborators, Carla Griffiths and Dr. William James at the Dunn School of Pathology, are currently using software called RSmatch. This program finds sequences having structures similar to the structure of a given sequence. This means that the clustering is centered around a seed aptamer, which has some undesirable features. Moreover, RNA secondary structure prediction is not perfect. So the program also considers a host of suboptimal structures. One form this project could take would be to improve on the RSmatch approach: instead of clustering around a seed sequence, start by computing all pairwise distances between optimal and good suboptimal RNA secondary structures for the aptamers. Cluster the aptamers based on these distances. There are usually many good suboptimal structures for a given sequence. One approach to reduce the computational requirements would be to only include representative structures in the clustering. One method that can produce representative structures is Sfold (which only seems to be available as a web service, so other methods for obtaining representative structures may have to be explored).

A more ambitious approach would be predicting the possible secondary structures, not from the individual aptamers in isolation, but for the shared pool of aptamers. Some headway has been made toward predicting a shared structure for two or more RNA sequences. However, these methods do have serious drawbacks, most notably steep computational requirements. Moreover, they usually assume an evolutionary relationship between the RNA sequences, where alignment of the sequences also contribute to the structure prediction. Aptamers are not related by evolution, but only be shared function. It would therefore be desirable to develop an aptamer model taking into account that

- all the aptamers have identical start and end sequences, flanking a region with the random sequence
- aptamers with a common binding motif are assumed to have a shared, but unknown, secondary structure, and an unknown number of constant positions of unknown location
- not all the aptamers in the set do necessarily share the same binding motif

and reasonably efficient algorithms for inferring secondary structure and constant positions for real aptamer collections based on this model.

Suggested reading

Description of aptamer selection at www.lmb.uni-muenchen.de/groups/famulok/SELEX.html.

Lecture notes on RNA secondary structure prediction at www.stats.ox.ac.uk/~lyngsoe/rna.ps.

A method for aligning RNA secondary structures and its application to RNA motif detection by J. Liu., J.T. Wang, J. Hu, and B. Tian, *BMC Bioinformatics* **6**:89 (2005).

RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble by Y. Ding, C.Y. Chan and C.E. Lawrence, *RNA* **11**, pp. 1157–1166 (2005).

Pairwise local structural alignment of RNA sequences with sequence similarity less than 40% by J.H. Havgaard, R.B. Lyngsø, G.D. Stormo and J. Gorodkin, *Bioinformatics* **21**:1815–1824 (2005).

Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes by Z. Yang and W.J. Swanson, *Molecular Biology and Evolution* **19**, pp. 49–57 (2002).

RNA Sequence Evolution With Secondary Structure Constraints: Comparison of Substitution Rate Models Using Maximum-Likelihood Methods by N.J. Savill, D.C. Hoyle and P.G. Higgs, *Genetics* **157**, pp. 399–411 (2001).