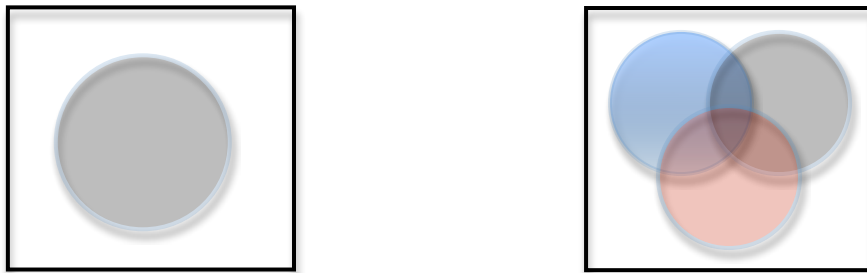


# RNA Structure, Evolutionary SCFGs and Classifiers

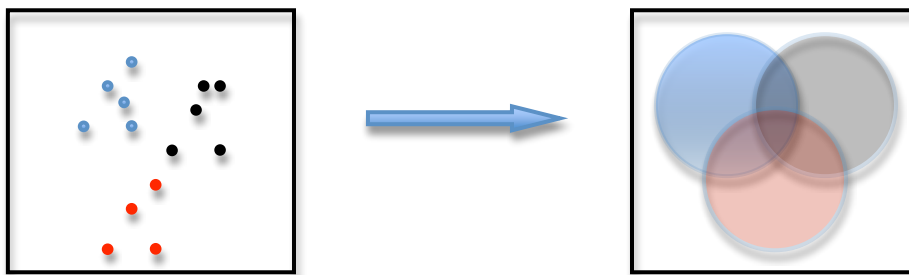
15.6.09

**Motivation and Background.** RNA molecules have proven to have a surprisingly large role in cellular functions of higher organisms. They are also present in much larger numbers than anticipated only a few years ago. Thus annotating and predicting the genome for RNA structures and functions is of great importance. In 1994 stochastic context free grammars (SCFGs) was introduced in two papers (Durbin and Eddy, 1994 and Sakikabara et al., 1994) to describe and analyze RNA structures. These grammars are “isomorphic” to earlier descriptions of RNA using cost functions to predicting structure by Zuker, Nussinov, Waterman and Smith in a series of papers around 1980. Knudsen and Hein (1999, 2003) coupled SCFG to molecular evolution to make a comparative predictor. Most RNA gene predictors try to classify a sequence into two classes: RNA gene – background functionless DNA. It is often of interest to classify an RNA gene/structure into functional classes. Typically, this is either done by homology (if two sequences are similar, they probably have the same function) or by some additional classifier algorithm (Batuwita and Palade, 2009 & Klingelhoefer, Moutsianas and Holmes, 2009).



On the left side we have a universe of sequences [square] and a subset, R [circle] that can be recognized as RNA sequence. SCFG tries to create model that describes R well as possible. We are often in a situation that we have several distinct classes of RNAs. In the illustration to the right it would be optimal to design 4 models that put as much probability on non-RNA, grey RNA, blue RNA, red RNA, respectively, and as little on the others as possible.

Using a set of SCFG trained individually [privately – without knowledge of the other RNAs] to recognize a set of distinct RNAs is probably quite efficient. However if there are strong similarities, but still differences of interests, then it isn't ideal to train privately. Training should be done to both recognize and dis-recognize other families.



We are given instances of RNAs belonging to different classes and want descriptors distinguishing the classes (including members not yet seen).

Dependent on the application area it could be convenient to train sets of grammars that are not independent, but constrained to be similar. If we for instance have sets of homologous miRNAs belonging to different types, then they will probably have almost identical structures within types, but similar with crucial differences between types. “Similar but different” should argue for the same in the grammars describing them.

One example of a SCFG RNA grammar is (designed by Bjarne Knudsen. Dowell and Eddy (2004) discusses alternatives):

$$S \rightarrow LS \mid L \qquad F \rightarrow dFd \mid LS \qquad L \rightarrow s \mid dFd$$

There are three variables  $[S, L, F]$ , each associated production rules [above]. The grammar becomes stochastic, when probabilities are assigned to the alternatives in each production rule. Thus there are two components to grammars: the rule set and the firing probabilities.

When  $dFd$  is chosen, it is interpreted as if the first  $d$  base pairs with the second  $d$ . When  $s$  is chosen, it is interpreted as a single nucleotide unpaired. Parameters are chosen by observing a set of sequences with structures. Different optimisation criteria can be chosen to choose parameters. Maximum likelihood (ML) where the both structures and sequences are observations is one possibility, ML where structure is conditional on sequence is used. If  $S(x)$  is the structure of a sequence  $x$ , then this would correspond the use  $P(S(x), x)$  and  $P(S(x)|x)$ , respectively. Since the objective of these methods are not parameter estimation, but structure prediction, it can be legitimate to use for instance expected correctness of prediction as optimisation criteria. See Eddy et al. (1995), Krogh (1997) and Do et al. (2006) for considerations of these issues. Do et al. (2006) additionally develops a more general approach than SCFGs.

The two structure/grammar problem would be given a set of strings,  $x_1, \dots, x_n$  and associated structures,  $S(x_1), \dots, S(x_n)$  to find the grammar that maximizes  $\prod_i P^A(S(x_i), x_i)$ . If two sets of strings (with structures) were give -  $(x, y)$  - it could be natural to define optimal grammars -  $A, B$  - as solving

$$\max_{A,B} [\prod_i P^A(S(x_i), x_i) - \prod_j P^A(S(y_j), y_j) + \prod_j P^B(S(y_j), y_j) - \prod_i P^B(S(x_i), x_i)]$$

This is easily generalized to multiple structures/grammars and adding the possibility of weighting the importance of recognising versus dis-recognising. The challenge is be to infer the two grammars in this situation.

Constraints on grammar similarity could be added through a metric on grammars combining differences on firing probabilities and rule sets. For instance

$$D(A,B) = |[A \setminus B] \cup [B \setminus A]| + c \sum_{ij} |p^A_{ij} - p^B_{ij}|$$

The  $c$  allows a weighting between the discrete and continuous differences between the grammars. The optimisation problem now has a penalty for choosing very dissimilar grammars,

$$\max_{A,B} [\prod_i P^A(S(x_i), x_i) - \prod_j P^A(S(y_j), y_j) + \prod_j P^B(S(y_j), y_j) - \prod_i P^B(S(x_i), x_i) - wD(A,B)]$$

The  $w$  is a weight for the dissimilarity function between grammars. The suggestions for these functions have been made only to conform with intuitive wishes for solutions and without consideration to computational issues. This optimisation problem is also chosen without reference to stochastic modelling of how structural/functional differences have arisen (possibly evolution).

Given a way representing classifying grammars, this would have to be combined with evolutionary models for single nucleotides and for paired nucleotides. This is a well-studied problem and one of the existing models could be adopted. Additionally the sequences would have to be aligned simultaneously or an alignment would have to be assumed given. The latter seems like the realistic starting point.

**Comment.** The project has been written after discussing two projects predicting microRNA function that used classifiers based on a series of selected features (Batuwita and Palade, 2009 & Klingelhoefer, Moutsianas and Holmes, 2009). JH felt this was a suboptimal approach for 3 reasons: i. it had no way of using evolutionary information, which has been the great advantage in most other annotation/classification of sequences. The present influx of data will mainly be homologous to existing sequences making evolutionary models a necessity. ii. Classifiers used no structural prediction, which in most other RNA analysis is the key component in analysis. iii. (related to ii) The lack of structure lead to no functional interpretation of differences between types. Clearly the proposed approach needs very different techniques from the classifier approaches mentioned above.

## References.

- Batuwita and Palade (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction *Bioinformatics* 25.8:989–995
- Do, Woods and Serafim Batzoglou (2006) CONTRAfold: RNA secondary structure prediction without physics-based models *Bioinformatics* 22.14:90–98
- Dowell and Eddy (2004) Design of Lightweight Stochastic Context-Free Grammars for RNA Secondary Structure Prediction *BMC Bioinformatics*
- Durbin-Eddy-Krogh-Mitchison (1998) *Biological Sequence Analysis* chapter 10 CUP
- Eddy, Mitchison & Durbin (1995) Maximum discrimination hidden Markov models of sequence consensus *J Comput Biol.*2(1):9-23
- Gardner et al. (2009) Rfam: updates to the RNA families database *NAR*
- Klingelhoefer, Moutsianas & Holmes (2009) Approximate Bayesian feature selection on a large meta-dataset offers novel insights on factors that effect siRNA potency *Bioinformatics*
- Knudsen & Hein (1999) Using stochastic context free grammars & molecular evolution to predict RNA secondary structure *Bioinformatics* 15.5:446-454
- Knudsen & Hein (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *NAR* 31(13), 3423–3428
- Krogh (1997) Two methods for improving performance of a HMM and their application for gene finding. *ISMB*
- Lowe and Eddy (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *NAR* 25(5):955-64.
- Nussinov and Jacobson (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *PNAS A.* 77(11):6309-13.
- Sakikabara et al. (1994) Stochastic context-free grammars for tRNA modeling *NAR* 22.23:5112-
- Waterman and Smith (1978) "RNA secondary structure: A complete mathematical analysis." *Mathematical Biosciences*, vol.41, pp.257–266
- Zuker & Stiegler (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information *NAR* 9.1:133-48