Bayesian computational approaches for gene regulation studies of bioethanol and biohydrogen production

Abstract: It has recently become clear that regulatory RNAs play a major role in regulation of gene expression in bacteria. RNA secondary structures play a major role in the function of many regulatory RNAs, and structural features are often key to their interaction with other cellular components. Thus, there has been considerable interest in the prediction of the secondary structures for RNA families. A paper describing our new algorithm, RNAG, to predict consensus secondary structures for unaligned sequences using the blocked Gibbs sampler has been published[1]. This sampling algorithm iteratively samples from the conditional probability distributions: P(Structure | Alignment) and P(Alignment | Structure). Subsequent to publication of the RNAG paper we have employed the technology from RNAG in the development of an RNA motif finding algorithm. To develop and RNA motif finding algorithm, RGibbs, we capitalized on our long experience in DNA motif finding and RNA secondary structure prediction. We applied RGibbs to three data sets from the literature and compared it to existing methods: one for training and two others for tests sets. In both test sets we found RGibbs out performed existing procedures.

In so doing, it refines the models of both Alignment and Structure. This iterative algorithm has theoretical advantage in convergence time, which stems from our application of the collapsing theorem of Liu[2]. Our use of this theorem capitalizes on the grouping of high-dimensional random variables in both the structure and alignment spaces to accelerate convergence, and on efficient recursive computations available for each of these spaces. The resulting samples permit a characterization of the shape of the full posterior space. We use a hierarchical clustering method to characterize its shape, γ-centroid estimator[3] to generate a prediction from sampled structures, and credibility limits[4, 5] to characterize the uncertainty associated with each estimate. In addition, we find that sampled structures are compact around their ensemble centroids for all but two families, and that there are well separated classes of structures in at least 11 of the 17 families. Also, while the distances between the reference structures and the predicted structures were small, they are substantially larger than the variation among structures within clusters.

Subsequent to publication of the RNAG paper we have employed the technology from RNAG in the development of an RNA motif finding algorithm. To develop and RNA motif finding algorithm we plan to capitalize on our long experience in DNA motif finding and RNA secondary structure prediction. We expect that nearly all of the issues that we and others have confronted in DNA *ab-initio* motif finding will again need to be confronted in RNA motif finding with one major extension: the inference of RNA secondary structure. We now have a preliminary version of a new RNA motif finding algorithm, that we call RGibbs.

We have developed and implemented a preliminary version of RGibbs. To test it, we compared it to the well-known CMfinder algorithm and chose sequences from the 19 Rfam families that the CMfinder paper used to test their algorithm[6].

Because we found in our studies of RNAG that 10 diverse sequences were sufficient, we selected ten sequences, except only 9 in one family, from each family so as to minimize the similarities between these ten. As was done in the CMfinder paper, we embedded the Rfam sequences in 200 bp of flanking sequence. We predicted at most one site in each sequence with both CMfinder and RGibbs using the default parameter settings for both, where the defaults for RGibbs were those of RNAG. Predictions were counted as true if they overlapped the known Rfam target by at least 50%, and the predicted structure had at least one pair of nucleotides that where predicted to be paired. We found that both procedure do well in finding the Rfam targets, and have no false positives. In order to control false positives, both procedures include provisions to not make predictions in a sequence unless there is sufficient supporting evidence.

We conducted additional analyses of false positives, because control of false positives becomes increasing important when number of potential negatives increases, for example when there is more flanking sequence , an RNA-seq study returns many candidates, or in genome wide phylogenetic footprinting. Specifically, we did studies with two negative control sets. First we applied both procedures to the flanking sequences by removing the Rfam target sequences. CMfinder finds 4.5 (134/30) times more false positives than RGibbs. However, using flanking sequences as a control is problematic because there may be unreported structural elements within flanking regions. To obtain controls that avoid this problem we aligned the sequences including flanking and targets using a structurally unaware algorithm (Muscle), and shuffled the columns of this alignment 100 times. We then applied both procedures to each shuffled set of sequences. In this way we sought to preserve sequence conservation, but break up any real structures. Analyses of these data shows false positives in these shuffled data are 5.7 (142/25) fold higher on average in CMfinder than in RGibbs, and their ranges do not overlap. As in each of the 100 shuffled sequence RGibbs found fewer false sites than CMfinder the differences in false positive identification between the two are very unlikely under this permutation null. Also, the probability of RGibbs finding 167 structured targets in the test set under this permutation null is estimated from this sample to be ≤ 0.01.

Because adding sequences from related families almost always improves motif finding we plan to include sequences from related species in all of the applications with our collaborators and we will encourage users of RGibbs to do likewise. To investigate performance in this more appropriate circumstance we compared RGibbs to CMfinder using the twelve *Drosophila* species in fly base. For these we compared RGibbs with optimal weights described by Newberg et.al.[7] to CMfinder on a set of 36 Rfam families that only had sequence from *Drosophila* species. On average we found sequences from 10.8 of these species per family. Again we found that both procedures did very well at finding the true sites, and correspondingly because both procedures were limited to no more than one site per sequence, both procedures had a small number of false positives. However in the flanking controls CMfinder had 2.4 fold more false positives, and in the shuffled controls CMfinder reports many false positives and has 4.8 fold more false positives than RGibbs, and again the ranges of results from all the shuffles don't overlap. This study also showed that the use of optimal weights, instead of the relative weights that are the default for Infernal, played a major role in reducing the number of false positives predicted by RGibbs in these fly only data. Results when counting at the family level instead of at sequence level are very strongly similar.

These results show that combining RNAG with DNA motif finding methods provide a promising path for the development of improved methods for finding RNA motifs. Furthermore, because RGibbs is based on the CM model there is a clear path to the development of a combinatorial RNA motif finder.

1. Wei, D., L.V. Alpert, and C.E. Lawrence, *RNAG: a new Gibbs sampler for predicting RNA secondary structure for unaligned sequences.* Bioinformatics, 2011. **27**(18): p. 2486-93.
2. Liu, J., *The collapsed Gibbs sampler in bayesian computations with applications to a gene regulation problem.* Journal of American Statistical Association, 1994. **89(427)**: p. 958-966.
3. Hamada, M., et al., *Generalized Centroid Estimators in Bioinformatics.* PLoS One, 2011. **6**(2): p. e16450.
4. Webb-Robertson, B.J., L.A. McCue, and C.E. Lawrence, *Measuring Global Credibility with Application to Local Sequence Alignment.* PLoS CompBiol., 2008. **4**(5): p. e1000077.
5. Newberg, L.A. and C.E. Lawrence, *Exact calculation of distributions on integers, with application to sequence alignment.* J Comput Biol, 2009. **16**(1): p. 1-18.
6. Yao, Z., Z. Weinberg, and W.L. Ruzzo, *CMfinder--a covariance model based RNA motif finding algorithm.* Bioinformatics, 2005: p. btk008.
7. Newberg, L.A., L.A. McCue, and C.E. Lawrence, *The relative inefficiency of sequence weights approaches in determining a nucleotide position weight matrix.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article13.