Structural motif detection in high-throughput structural probing of RNA molecules

Diego Muñoz Medina, Sergio Pablo Sánchez Cordero

December 11, 2009

Abstract

In recent years, the advances in structural probing in RNAs have led to the development of high throughput experiments that generate vast amounts of data that is not feasible to curate manually. Analogous to the case of high-throughput genome sequencing, there is a need to analyze the structural results via computational methods. In this project, we propose an algorithm based on support vector machines (SVMs) to find structural motifs in data produced by an avant-garde method for RNA structural probing known as RNA tapestries. In this experimental method, every base of a specific RNA molecule is mutated iteratively, one by one, producing N number of mutants (where N is the length of the RNA, in nucleotides). Each mutant is then modified using different chemical treatments, such as DMS, CMCT, and SHAPE[3, 1]. The modification of the molecule occurs in sites where the RNA is unstructured (such as hairpins and loops for SHAPE chemical footprinting) or in specific base types (adenosine and cytosine for DMS, and guanine and uridine for CMCT), thereby, giving a strong fragment signal during reverse transcription check by gel electrophoresis. The final array of the modified mutants is given in form of a matrix T, where each entry, $T(i, j)$, is the signal strength of the ith nucleotide of the molecule in the jth mutant. Careful analysis of T may allow the recognition of patterns that are helpful in determining the structure of the RNA in question.

1 RNA Tapestries

Structural probing of RNAs by means of chemical modification is a quick and powerful way to determine important contacts of the RNA molecule in question. The main idea behind these methods is to induce RNA modifications using different chemical treatments. Modification of the molecule occurs as methylathion of unpaired nucleotides (for DMS and CMCT) or as backbone conformational changes in regions where it is not structured (for SHAPE analysis). The resulting RNA molecule is then reverse transcribed, which results in the production of fragments of DNA that stop at the modified sites and that can be subsequently analyzed and compared using standard gel electrophoresis techniques. The information given by these methods can be complemented by inflicting nucleotide mutations to the RNA and repeating the chemical footprinting analysis. Positions of the RNA that are modified after the mutations indicate strong structural dependence on the nucleotides near those locations. This technique can be further extended by repeating the process for many mutants and aligning the resulting signals to perform a comparative analysis between mutation effects on the RNA structure. Simple patterns, such as those describing stems, can be easily visualized and annotated by hand, but this is a time consuming and error prone process. Furthermore, more complex interactions (such as tertiary contacts) are not readily apparent and may be difficult to spot.

2 Methods

We demonstrate the use of SVMs with $l1$ regularization to detect structural motifs in RNA tapestries, using tapestry data from the medloop RNA motif, the signal recognition particle motif (SRPH8), tRNA Phe, and the Sarcin loop motif RNAs. These RNAs have well-known structures and can therefore be used as a training set for our SVM implementation.

2.1 Pre-processing the data

Since comparison of various RNA tapestries can be noisy(due to the variance in the strength of the chemical signal in each experiment), we compute global z-scores for each entry in the tapestries. When classifying a new RNA given its tapestry, we adjust the entries to global z-scores as well. Furthermore, since the experimental data contained chemical footprinting for the RNA sequence of interest plus primer sequences at its ends, we had to carefully locate the tapestry region that mapped to the RNA molecule in question and discard the primer sequences' data. This was done by standard methods that align the chemical footprinting data for each mutant (taking into account signal intensities and signal position in the sequence).

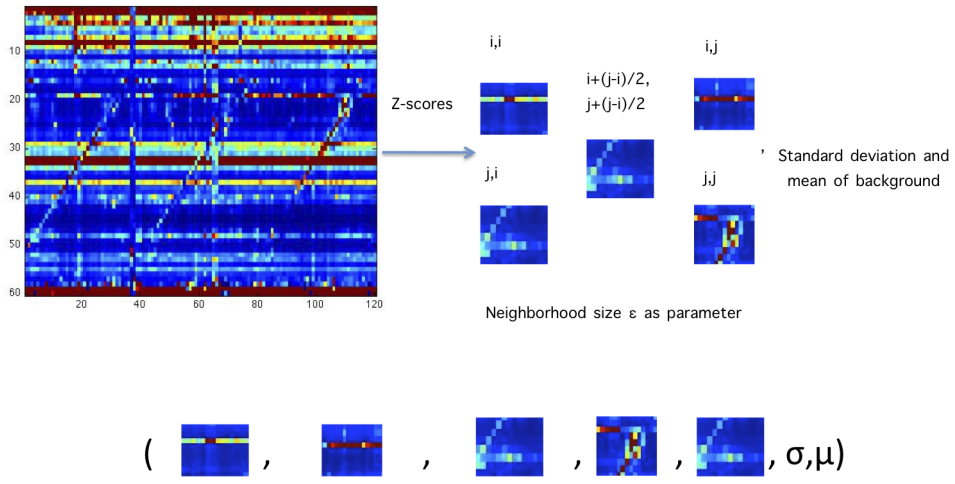

Figure 1: Defining the feature vector: An RNA tapestry as a heat map (left). We focus on the four corner boxes as well as the central box for each base pair. The central box is crucial for obtaining good results.

2.2 Building the training set

First, let us define what an RNA tapestry is. An RNA tapestry is a $n \times n$ real-valued matrix T such that its kth column contains the chemical footprint data given by the chemical footprinting protocol when mutating a single nucleotide in the wild type RNA. The mutation is specified on a separate label vector L, that contains the index of the nucleotide modified and the base to which is was mutated. By definition, the first column of T contains the chemical footprint information of the wild type RNA, with no mutations. We use a pairwise criterion to build the training set, that is, for each pair of nucleotides i and j (given by their respective index in the RNA sequence), we will see if they interact in some manner (i.e. form a base pair). To do this, we will focus on a vecinity in the matrix T around the centers: $T(i,i)$, $T(i,j)$, $T(j,i)$ and $T(j,j)$. In other words, given a neiborhood parameter ϵ, we will look on the entries $T(k,l)$ where k and l are integers such that $k,l \in ([i-\epsilon, i+\epsilon] \cup [j-\epsilon, j+\epsilon]) \cap ([0,n])$. This follows the intuition that when mutating a nucleotide in a certain position, there will occur structural rearrangements in its vicinity that will be expressed as a particular signal in the tapestry. In this way, we focus our attention only in the "'boxes" centered at $T(i,j)$, $T(j,i)$, $T(i,i)$ and $T(j,j)$, with side size $2\epsilon+1$. Furthermore, if perturbating two positions separately result in a significant signal change in both positions then there is some evidence to believe that the nucleotides in those positions interact in some way. Our objective is to implement a SVM that is capable of predicting what changes in the chemical signal correlate with a physical interaction between the bases. Thus, we use all of the entries $T(k,l)$ as features to build our training set. In order to account for the the rest of the entries in the row k and l, we take the set B of all "background" entries $T(k',l')$ that lie in the rows i and j, but not in F, and obtain their mean and standard deviation. Finally, we append to each feature vector integer labels of the bases that were mutated. In this way, we can convert the tapestry T into n^2 feature vectors that give the pairwise tapestry information of the nucleotides in the RNA. The category of each feature vector is -1 if the respective bases i and j do not interact and 1 if they do. We repeat this process for each RNA tapestry that was include in the training set.

2.3 Tweaking the SVM parameters and reducing training set size by random sampling

We tried different parameters for our SVM and compared their performance. We varied the C trade-off parameter, tried linear, 2nd to 5th degree polynomials, gaussian kernels, and varied the ϵ vicinity value described above. We also varied the constraints on the condition $w^T x + b \geq 0$ to be $w^T x + b \geq \gamma$ to reduce the number of negative predictions. This last parameter is motivated by the small amounts of positive examples in our training set (from the n^2 possible base interactions, only a small fraction of them actually occur[1]). This observation also allowed us to reduce the training set size by performing random sampling of the negative training examples (this was done by discarding a negative training example with a certain probability which, in our case, was 0.6). Fortunately, this did not alter our results in any way and significantly reduced the complexity of the training

[1]If we are only considering secondary structure, the number of interactions is of order $O(n)$. This result derives from the fact that in secondary structure interactions a nucleotide can interact with at most one other nucleotide.

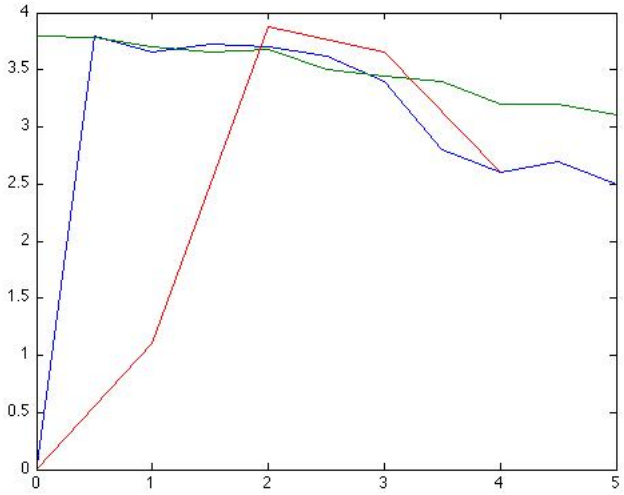

Figure 2: SVM parameters: Leave one out precisions of various SVM parameters: C trade-off parameter (blue), $-\gamma$ (green) and the neighborhood size ϵ (red). For comparison purposes, we fit γ from 0 to 5 for the figure's scale, although it was actually taken from 0 to -10. The precision given is before applying the filter.

set. We analyzed the performance of the SVM varying one parameter at a time (first the C parameter, followed by the kernel type, γ and, finally, ϵ). For each parameter value, we performed a cross validation similar to leave one out, but leaving out a complete set of vectors corresponding to a particular RNA in each trial.

2.4 Prediction filtering

Since only a small set of RNAs was available and due to the noisy nature of the data, we designed a method to filter the resulting predictions. First, we scored each possible base pair $Bp(i,j)$, formed by the nucleotides positioned in i and j in the RNA sequence and where Bp is the base pair matrix of all possible base pairs. This was done in the following manner:

- Start by defining $Score(i,j) = 0$.
- If i and j are within a vicinity δ in Bp of a base pair interaction predicted by the SVM, then set $Score(i,j) := Score(i,j) + 2$.
- If $Bp(i,j)$ is a predicted base pair interaction but $Bp(j,i)$ is not, then set $Score(i,j) := Score(i,j) - 1$.
- If $Bp(i,j)$ is a Watson-Crick base pair (AU,GC) or a GU base pair, then set $Score(i,j) := Score(i,j) + 1$.
- Define the stem bonus $Sb(i,j)$ as the iterative sum of the positive scores in the diagonals starting at $Bp(i-1, j+1)$ and $Bp(i+1, j-1)$, stopping when a score of zero is encountered. Intuitively, these bonus corresponds to scoring a possible stem from which $BP(i,j)$ is a part of.

To filter the base pairs, we partitioned each column of the base pair matrix into m intervals. For each interval in each column, we discarded all of the predicted base pairs except the one with the highest $Score(i,j) + Sb(i,j)$. As the parameter m increases, our filter acts in a more selective manner, conserving fewer positive predictions. The case where m equals the sequence length allows only one prediction pair column, which allows at most one base interaction per nucleotide. This property aligns perfectly with the definition of the secondary structure, and we were pleased to find that, in the case of SRPH8, we could predict the full secondary structure setting m to be the length of the sequence.

3 Results

Given a full set of vectors that describe an RNA molecule, and following our previous definitions for positive/negative training examples and predictions, we expected to see only a small number of positive predictions. At first, our SVM greedily gave a negative prediction for every base pair. Indeed, given the nature of the testing set, this gave a spectacular error of less than 3% in most cases, but with horribly low precisions. We tried to include more positive predictions by decreasing the parameter γ, but still got poor results. Since we were

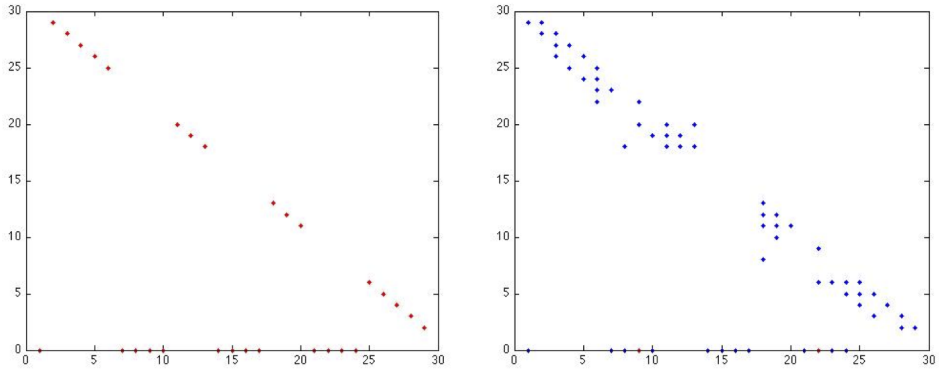

Figure 3: SVM base interaction predictions on the SRPH8 motif (right) and the correct secondary structure (left). After applying our filter, we can get the correct secondary structure.

constantly observing inconsistencies in the way that the training precision, testing precision, recall, and error were varying given increases in the training set size, our next approach was to include more features.

3.1 Adding the central box

After numerous failed attempts to improve our training and testing precisions by changing the SVMs parameters, we finally decided to include more features in our feature vector. Since we had observed that most structural motifs in the tapestries were visually manifested in the form of "'X"'s, we decided that it was not sufficient to focus on the boxes centered in the combined coordinates of each possible base pair. Therefore, we would also have to consider the changes ocurring in the middle of the region defined by the corners $T(i,i)$,$T(i,j)$,$T(j,i)$, and $T(j,j)$. In other words, we added the box centered at $T(i+\lceil\frac{j-i}{2}\rceil, j+\lceil\frac{j-i}{2}\rceil)$, whose size was also defined by the vicinity parameter ϵ. This gave a great boost in the performance of our SVM and we were happy to find that only using the medloop motif DMS and CMCT data was sufficient to obtain a training error of 0 using a linear kernel; in other words, the medloop motif training data inputted into our SVM was capable of predicting its own structure. This was an expected result, since the medloop motif data has a well-defined chemical footprint from which the structure can be quickly inferred by visual inspection.

3.2 Testing other SVM parameters

After enhancing our feature vector, we performed a leave one out type of cross validation, separating a whole set of vectors corresponding to a RNA in each trial. Setting each parameter to its default value (which would be zero for all parameters), we first optimized with respect to C, then with respect to the kernel type, then γ and, lastly, ϵ, taking the best value in each step. We found that the linear and second degree polynomial kernels performed well overall, while the higher degree polynomial kernels and the gaussian kernel tended to overfit the SVM[2].

3.3 Taking joint DMS and CMCT data

At first, we thought that DMS and CMCT data should be treated separately, building and testing training sets that included only DMS data or only CMCT data. This led to poor training precision and recall values, close to 0. However, we found that arbitrarily adding both types of data to the training and testing sets greatly improved the accuracy of the results, to the point where, when we apply the filtering method described above, we could get the complete secondary structure for the medloop and SRPH8 motif (sadly, we could not fully test this with the larger tRNA data, since we are missing the CMCT tapestry). This may follow from the intuitive idea that DMS and CMCT data complement themselves since they modify different nucleotides. To our surprise, by predicting using a training set of DMS and CMCT data, with a testing set only comprised of SHAPE data, we get noisy but visually good results (for example, the number of stems predicted is accurate). This indicates, that in all chemical footprinting methods, the RNA structure manifests itself in, qualitatively, the same manner.

[2]We varied the radial parameter of the gaussian kernel, but could not get a good intermediate between an extreme overfit and an extreme underfit, in which it could perform similar to the linear kernel.

4 Discussion and future work

We have described a SVM approach to analyze high-throughput chemical modification probings of RNA structure. This is a recent experimental approach that is much faster and cheaper than the crystalography method that is considered the gold standard when determining RNA structure. Its high-throughput nature gives us vast amounts of data that are ideal for machine learning analysis. Our current filtered SVM method can give an accurate estimate of the secondary structure. What is most striking is that there is no thermodynamics involved in our method, its findings are purely empirical[3]. This not only illustrates the power of the experimental method, but also gives great room for improvement by including thermodynamic rules or parameters in some manner. Currently, our classifier is trained with secondary structure interactions. We can easily extend the power of our classifier to tertiary contacts by adding known tertiary information for our training RNAs. In this case, it would be best to have a second SVM that predicts tertiary interactions. Finally, although our filtering approach was sufficient to extract the secondary structure, it could be replaced with another SVM that "cleans" the first SVM's predictions based on another set of features (secondary substructure patterns in the predicted base pairing plot are easily found by visual inspection, so there must be a way to estimate them using an automatic classifier). Furthermore, let us remember that we have a limited set of 4 RNAs, most of them missing one of the three chemical types of chemical footprinting (DMS, CMCT, and SHAPE). As more experimental data becomes available in the future, it can be easily integrated into our classifier for further improvement.

5 Acknowledgements

We deeply thank the Das lab for giving us the tapestry data. Special thanks to Rhiju Das for his insightful comments.

References

[1] Quentin Vicens Samuel M. Pearlman Michael Brenowitz Daniel Herschlag Alain Laederach, Rhiju Das and Russ B. Altman. Semi-automated and rapid quantification of nucleic acid footprinting and structure mapping experiments. *Nature Protocols*, 3:1395–1401, 2008.

[2] C. B. Do, D. A. Woods, and S. Batzoglou. Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22(14), July 2006.

[3] E. J. Merino, K. A. Wilkinson, J. L. Coughlan, and K. M. Weeks. Rna structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (shape). *J Am Chem Soc*, 127(12):4223–4231, March 2005.

[4] Marc Parisien and Francois Major. The mc-fold and mc-sym pipeline infers rna structure from sequence data. *Nature*, 452(7183):51–55, March 2008.

[5] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31:3406–3415, 2003.

[3] It is interesting to note that all *de novo* algorithms for RNA structure prediction involve thermodynamics in some way, either explicitly by following free energy minimization[4, 5], or implicitly by including thermodynamic parameters[2].