Statistical potentials for RNA-protein interactions optimized by CMA-ES

Takayuki Kimura  
Tokyo Institute of Technology

Nobuaki Yasuo  
Tokyo Institute of Technology

Masakazu Sekijima  
Tokyo Institute of Technology

Brooke Lustig  
San Jose State University, brooke.lustig@sjsu.edu

Follow this and additional works at: https://scholarworks.sjsu.edu/faculty_rsca

Part of the Artificial Intelligence and Robotics Commons, Molecular Genetics Commons, and the Organic Chemistry Commons

Recommended Citation
https://doi.org/10.1016/j.jmgm.2021.108044

This Article is brought to you for free and open access by SJSU ScholarWorks. It has been accepted for inclusion in Faculty Research, Scholarly, and Creative Activity by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.
Statistical potentials for RNA-protein interactions optimized by CMA-ES

Takayuki Kimura\(^a\), Nobuaki Yasuo\(^a,***\), Masakazu Sekijima\(^a,\)**, Brooke Lustig\(^b,\)*

\(^a\) Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan
\(^b\) Department of Chemistry, San Jose State University, One Washington Square, San Jose, 95192-0101, CA, USA

**Abstract**

Characterizing RNA-protein interactions remains an important endeavor, complicated by the difficulty in obtaining the relevant structures. Evaluating model structures via statistical potentials is in principle straightforward and effective. However, given the relatively small size of the existing learning set of RNA-protein complexes optimization of such potentials continues to be problematic. Notably, interaction-based statistical potentials have problems in addressing large RNA-protein complexes. In this study, we adopted a novel strategy with covariance matrix adaptation (CMA-ES) to calculate statistical potentials, successfully identifying native docking poses.

1. Introduction

1.1. RNA in biology

RNA-protein interaction is essential in many regulatory processes in cells [1]. Recently non-coding RNA (ncRNA) is emerging as an important class of biomolecules [2]. ncRNA can be classified by size and function as micro RNA (miRNA), short interfering RNA (siRNA), Piwi-interacting RNA (piRNA), and long non-coding RNA (lncRNA). Included are epigenetic processes such as DNA methylation, histone modification, and heterochromatin formation. Elucidating RNA-protein interactions, however, is not straightforward. Crystallizing RNA-protein complexes (RNP) is much more difficult than for proteins, in part at least due to the flexibility of the RNA chain [3,4]. Moreover NMR is limited by size considerations, and cryo-EM can require extensive sample preparation and related considerations [5].

1.2. Statistical potentials

Docking of specific RNA and protein chain(s) is a standard approach in exploring their interactions [6]. Statistical potentials can be utilized to evaluate such interactions, including explicitly evaluating RNA-protein docking poses [7,8]. Statistical potentials were introduced for polymers and proteins [8,9], and subsequently for DNA and RNA-protein complexes [10,11]. The application of statistical potentials is reported to be comparable in efficacy to those involving force-fields such as AMBER and CHARMM for proteins [12] and RNA-protein complexes [13,14]. Interestingly, coarse-grained statistical potentials, including by the Bujnicki group, have shown significant utility [15]. Perez-Cano, Fernandez-Recio and coworkers calculated, with some success over earlier methods, pairwise ribonucleotide-amino acid interaction potentials classified by accessibility to solvent. Here we utilize a similar approach for non-redundant RNA-protein complexes [16] of RNA targets.

In hierarchical approaches from an exhaustive set of coarse-grained structures [17], low energy conformations can be selected for additional structure building and analysis. Statistical potentials are also useful in efficiently assessing docking poses. Integrating RNA and/or protein chain modeling with docking can effectively predict RNA-protein binding poses [6]. A statistical potential can be calculated from the contact data of neighboring atoms [18–21]. Other than hydrogen bonds, \(\pi\)-interactions including cation-\(\pi\) interactions, and \(\pi\)-\(\pi\) interaction are known to exist in RNA-protein interfaces [22]. The \(\pi\)-interactions play significant roles in RNA-protein binding [23] and the energy is comparable to hydrogen bonds [24]. Therefore, in this study, statistical potentials are calculated from the frequency of hydrogen bonds and \(\pi\)-interactions.
RNA-protein interactions can be parsed into eighty classes involving twenty amino acid and four nucleotide types, and the probability for any particular class of RNA structure [16] having amino acid \( i \) and nucleotide \( j \) in contact is given as

\[
P(i, j) = \frac{N(i, j)}{\sum_i \sum_j N(i, j) \times \sum_i N(i)}\]

where \( N(i, j) \) is the number of such contacts that is summed over all such pairs, and \( N(i) \) and \( N(j) \) describe the number of respective amino acid and nucleotide contacts that are summed separately over all protein and RNA chains. The denominator that includes these individual terms can be described as a reference state, an important feature of the prediction.

1.3. Framework

RNA-protein interactions can be parsed into eighty classes involving twenty amino acid and four nucleotide types, and the probability for any particular class of RNA structure [16] having amino acid \( i \) and nucleotide \( j \) in contact is given as
Finally, the potential for each ribonucleotide-amino acid contact can then be calculated as
\[ \Delta G(i, j) = RT \ln \frac{P(i, j)}{(i, j)} \] (2)

The calculated potential was applied to poses from experiment and from docking simulations to evaluate the ability to separate native and native-like poses. Here a native pose is an experimentally characterized structure obtained from PDB. Native-like poses are the structures generated by docking with their respective RMSD less than 1-2Å. Note the free energy (raw statistical potential) for each nucleotide-amino acid contact is normalized such that

score of a pose = w \cdot x \quad (3)

Originally in this evaluation, the score of one binding pose was calculated as a sum of products of frequency of contacts \( F(i, j) \) and the potential \( \Delta G(i, j) \). This can be rewritten in Equation (3) as vectors \( w \) and \( x \), consisting of potentials and counts, respectively. If the score of a native or native-like pose is always lower than non-native poses, we can use the vector \( w \) to identify the native pose.

1.4. Interpretation via computer science

In terms of the computer science, Equation (3) shows that the score is an inner product of a weight vector \( w \), noted here as a statistical potential, and the frequency vector \( x \). Given a protein chain interacting with an RNA chain and the score of the native pose is lowest of all the poses, the weight vector is regarded as the normal vector of the plane that separates native pose and non-native poses (Fig. 1). Note that the weight vector does not include an intercept, denoted as \( b \) in Equation (4). Generally, an intercept determines the position of the boundary plane and is usually used in a general linear classification. Omitting the \( b \) intercept therefore provides generality to the scoring, free from the absolute position of the frequency vectors that can be problematic in interaction-based prediction when using the conventional expression

boundary plane = w \cdot x + b \quad (4)

In other words, the weight vector is the direction that helps to identify the native pose within the data set. The scores are used to compare the poses only within the data set. For example, in Fig. 2, when we have poses from RNA chain \( r \) and protein chain \( p \) of 1ABC (the data set 1ABC_p_r), the native pose cannot be identified until comparing it with other scores of the poses in 1ABC_p_r. It is impossible to identify the native pose just by the absolute value of the score. Therefore combining all the complexes data sets and applying general machine learning techniques will not work in this case.

If we define a native pose as positive and non-native ones as negative, a data set has only one positive data point, and accordingly the range of acceptable directions can be relatively large. Consequently, it might be possible to find a common direction or a range across multiple data sets. In this study, we optimized the direction that provided the best average of rankings of the native pose in each data set. Here rankings describe the score of the best native pose determined from PDB.

In summary, our scoring function consists of 80 (or 116) values and it is supposed to score a native pose as the highest or the lowest among non-native poses in each data set or the pair. Our approach here is to find the best scoring function or the common direction or potential vector

**Table 2**

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Training</th>
<th>Test</th>
<th>Interaction Type</th>
<th>Cross Validation Fold Number</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Original</td>
<td>Original</td>
<td>Hydrogen Bond</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Original</td>
<td>Original</td>
<td>Hydrogen Bond + π Interaction</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Original</td>
<td>Benchmarks</td>
<td>Hydrogen Bond + π Interaction</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Original + Benchmarks</td>
<td>–</td>
<td>Hydrogen Bond + π Interaction</td>
<td>–</td>
<td>See Table 3</td>
</tr>
</tbody>
</table>

Note the evaluation 1, and 2 involve cross validation. The original data is RNA-protein 131-data set. Benchmarks are data sets of Huang & Zou and Perez-Cano.
that provides the lowest average rank for the given data set. The search is performed by a numerical optimization algorithm. As such, this study can avoid the problems of reference state that is usually required in the calculation of statistical potentials. The model is different from general learning-to-rank problems [26, 27] because only the top ranking matters. It is different from general two-class classification in that the number of positive data (data from native pose in this study) considered is just one, and the model identifies the positive data point only by direction. The scoring process includes sorting and ranking, so the scoring process cannot be described by equations. Therefore this process is similar to the one in hyperparameter tuning [28, 29].

1.5. Limitation of docking simulation

One possible limitation in calculating interaction-based statistical potentials is difficulty to reproduce binding modes of experiments in the docking simulations. For example, it happens quite often that an RNA chain has many contacts with multiple protein or RNA chains in the crystal.

However, in docking simulations, only one RNA chain and one protein chain are used for docking. In other words, docking simulation ignores the interactions from other than the two target chains. Therefore we need to assess the effect from those chains for each binding pair.

2. Methods

2.1. Contact data

First, 3D structural files of RNA-protein complex from X-ray crystallography were downloaded from Protein Data Bank (PDB) in the PDB format (1892 pdb files) with R-free value less than or equal to 0.3. Hydrogen bonds except for water bridged ones in the data set were then characterized by HBPLUS [30], divided into 80 classes based on the amino acid and the nucleotide, and the number of interactions in each class was counted per RNA-protein chain pair. HBPLUS identifies hydrogen bonds using distance and angle criteria involving the donor and acceptor atoms. We used the default criteria for the program. For example, a bound pair of protein chain A and RNA chain C in 2CV1 had eight hydrogen bonds between arginine and cytosine, and those eight interactions will be one component of the frequency vector. SNAP [31] characterizes interactions between RNA and protein from the 3D structure of the complex. The interactions include hydrogen bonds and π stacking interactions. In this study, we used the output of π stacking only. After executing SNAP for all the native poses, we found that the π interactions in the output were related to nine residues. Therefore the counts of π interaction were divided into 36 classes depending on the four bases, and nine residues (ARG, TRP, ASN, HIS, GLU, GLN, TYR, PHE, ASP). Generally, π interactions involve adjacent π systems [32].

---

Fig. 6. Progress of optimization for Evaluation 1–3 (A) and Evaluation 4 (B). For Evaluation 1 and 2, the plots are from the first of the five training sets (Test set 2–5 in Table 1) in the five-fold cross validation. Average ranking (ordinate) for each data set is plotted versus the iterations (abscissa) of calculations of an average ranking for an updated potential. All the weighting components of the initial potential were set to 0.01. Enlarged version around the convergence is shown on the right.
Previous analysis revealed π interaction between RNA and protein by cyclic residues [33, 34] and non-cyclic residues [35].

In order to mitigate the gap between docking software and actual crystal structures, we defined a new quality index called External Contact Ratio (ECR) to filter contact data. ECR is defined as a fraction of off-the-target hydrogen bonds of the two chains to the target bonds between the two chains. For example, when a target pair has 10 hydrogen bonds between the two chains, and have 30 hydrogen bonds with other chains than the ones between two target chains, the ECR is 3.0. Contacts with other chains were calculated [36] based on the crystal contacts obtained by Chimera (command: cryst, PDB entries with title including 'virus' or 'capsid' were not included because cryst command was not valid).

Data with ECR less than 2.0 was clustered (see Fig. 3) PISCES web server [37]. Thresholds used: minimum chain length 40, maximum free R-value 0.30, maximum resolution 2.5, and maximum R-factor 0.25. The chain quality was estimated based on the validation data in XML format at the PDB. In addition, we omitted chain pairs that had 30% or greater of the residues exceeding double the RSRZ, the Z score for Real Space R value [38]. Also, chain B of 4N2Q was omitted because 62% of the chain was not characterized. For clustered RNA-protein chain pairs, a pair with the best chain quality was selected as the representative of the cluster. A set of 131 allowed chain pairs was selected (Table 1).

In order to prepare non-native poses, ZDOCK [39] performs a full rigid-body search of docking orientations for protein-RNA docking, calculating statistical potential energies. Non-native poses were generated without any constraints. Only PDB chains from the complexes themselves were considered. For example, docking to a protein chain 3SNP_A (850-aa) with an RNA chain/fragment 3SNP_C (30-nt) using ZDOCK generates 3600 non-native poses (see Fig. 4). ZDOCK was conducted on TSUBAME 3.0 supercomputing system (Tokyo Institute of Technology, Japan).

RMSD was calculated with PyMOL [40]. First, protein alpha carbons were aligned for two protein chains, then RMSD was calculated on all atoms of the two RNA chains except for hydrogen atoms. The hydrogen bonds, obtained by HBPLUS, for both the native and the non-native poses were classified into 80 categories. As a result, one frequency vector from native pose and 3600 frequency vectors from non-native poses were prepared, and the dimension of the vectors are all 80. In evaluations including π interactions, the dimension is 116.

Two public benchmarks for RNA-protein binding were also used in this study. The benchmark data set from Perez-Cano L et al. [41], and the data set from Huang and Zou [42]. From Perez-Cano was the 62-set of unbound-bound cases that excluded the four unbound-pseudounbound ones. And Hung & Zou included 52 unbound-unbound cases and 20
unbound-bound ones. Note unbound-unbound cases are defined here as having available the apo form of the binding pairs, or an appropriate chain from another RNA-protein complex, while unbound-bound has only one of the two component chains also characterized experimentally. These unbound-pseudounbound cases typically refer to an RNA structure from another RNA-protein complex with protein sequence identity less than 35%. ECR distribution for the benchmark data set (Fig. 5) shows ECR output.

2.2. Optimization

We then optimized the statistical potentials. As the initial potential vector, every component was set to 0.1. In the optimization, for one pair (1 native frequency vector and 3600 non-native vectors), the potential vector was used to calculate 3601 inner products (Equation (3)). The ranking of the inner product for the native frequency vector, in an ascending order, is the ranking of the pair. The target function of the optimization takes a potential vector as the input, and the output is the average of the 131 rankings of native pose. As the function cannot be described by a closed-form equation, we adopted no-derivative optimization algorithm CMA-ES [15] (Evolutionary Strategy with Covariance Matrix Adaptation.) CMA-ES is well-known approach in optimizing an indifferentiable function with real number, and high-dimensional input. Pycma module [43] was used to optimize the function. Although this is a linear model and there is no risk of overfitting, we employed five-fold cross validation. In the first round of the cross validation, 80% of data was used for training and the remaining 20% was used for testing. In the second round, different 20% was used for testing and continue this to the fifth round. After the optimization, the obtained potential vector was normalized in order for the length to be 1.

We also scored the same data set with ITScore-PR [18], a benchmark scoring functions for RNA-protein complexes. ITScore-PR is structure-based with statistical potentials optimized by statistical mechanics-based iteration method. We adopted no minimization mode in scoring. We tested the original 131 data sets with 5-fold cross validation, and also tested the benchmarks of Huang & Zou and Perez-Cano with potentials optimized on all the original 131 data sets. In addition to the 131-set, Huang & Zou and Perez-Cano data sets were used for optimizing potentials and testing. Table 2 lists the four evaluations on different data sets or different bonding types. In Evaluation 3, 6 data sets (1M8V, 1G1X, 3HHZ, 3HL2, 2JGE, and 1HVU) were omitted, where 1M8V, 1G1X, 3HHZ, and 3HL2 were removed because of their respective high ECR (12.0, 11.5, 5.7, 4.8). According to the validation report of PDB, the chain R of 2JZE has bad chain quality. The 67% of chain R was not characterized in the PDB. The 80% of residues in chain C of 1HVU has at least one outlier with respect to our criteria of unusual geometric features for bond length, bond angle, Ramachandran, and other structure features [44]. In Evaluation 4, focusing on recalibrating the range of potentials, we just optimized potentials on four subsets of data. The original 131 data sets and the two bound-bound benchmark data sets were filtered four different ways to prepare the four subsets. Table 3 summarizes the differences of the subsets. One filtering inspects overall quality of validation reports at the PDB to determine the quality of each data set. If the data has any chain of which greater than or equals to 80% of residues had more than one outlier, we excluded the data. In addition, if at least one chain had missing residues for more than 50% of the chain,
the data was not included. For example, 2PY9 of Perez-Cano benchmark data set was omitted because 93% of residues in its chain r had at least one outlier. Another filtering was with ECR by hydrogen bonds. Any data set with ECR greater than 2.0 was omitted. In Evaluation 4, in addition to the conditions we used in Evaluation 1 to 3, we optimized the potential on the four subsets that were filtered differently.

3. Results and Discussion

3.1. Evaluation 1 - hydrogen bonds

In the cross validation of the original 131 data sets, we trained with five subsets of data, and an average ranking of the 131 complexes was obtained for each subset after optimizing potentials. The average quintile rankings were 1.02, 1.04, 1.06, 1.07, and 1.06. The average of these five was 1.05. The progress of the optimization (Fig. 6A) shows that the average ranking almost reached 1.0. Comparison with ITScore-PR is shown at Fig. 7. The average rankings of the five current tests were 1.54, 1.31, 1.15, 1.42, and 1.37. The average was 1.36. The worst ranking was 11 for 3RW6_A_H, and 6BL4_L_R. In the total 131 pairs, 117 pairs (89.3%) ranked their native pose as 1. Fig. 9A shows the optimized potentials in Evaluation 3. We used all of the obtained potential sets in the benchmark tests. In all of the four tests, our model achieved better average ranking than ITScore-PR. However, looking at the ranking distributions within the top 20, ITScore-PR predicted better in two cases (originally bound-bound and unbound-unbound Huang & Zou cases).

3.2. Evaluation 2 - hydrogen bonds and π interaction

In Evaluation 2, the original data was trained and tested by 5-fold cross validation for hydrogen bond and π interactions. Four of the five quintile optimizations achieved 1.0 as the highest average ranking (Fig. 6A). Evaluation 2 adopted 5-fold cross validation on the original RNA-protein 131 data sets, and frequencies included both hydrogen bonds and π interactions. Therefore the dimension of the potential and frequency vectors was 116. Fig. 7A shows the ranking of native poses and the rankings with ITSCore-PR. The percentage of the top 1 ranked predictions of our scoring model was slightly higher than for Evaluation 1 (from 90% to 95%). Therefore including π interactions increased the accuracy of native pose predictions.

3.3. Evaluation 3 - benchmark test

In Evaluation 3, a potential vector was optimized on all of the original 131 data sets, and tested on the two benchmarks. Evaluation 3 involved hydrogen bond and π interaction. The lowest rank was 1.008, and multiple sets of potential were obtained that achieved such. Fig. 9C shows the optimized potentials in Evaluation 3. We used all of the obtained potential sets in the benchmark tests. In all of the four tests, our model achieved better average ranking than ITScore-PR. However, looking at the ranking distributions within the top 20, ITScore-PR predicted better in two cases (originally bound-bound and unbound-unbound Huang & Zou cases).
3.4. Evaluation 4 - potential with 131 data sets and benchmarks

In Evaluation 4, we trained on our original 131 data sets and benchmarks to recalibrate the range of potentials (Fig. 6B). Four subsets were generated by changing filtering conditions on bound-bound data of Perez-Cano and Huang & Zou, where Fig. 10 shows the optimized potentials in Evaluation 4. The strongest 10 potentials in Subset 2 are shown in Table 4. We selected potentials of Subset 2 as representative for this study because the final average ranking was sufficiently good (1.04) and the data size was relatively large (169).

The 10 strongest potentials in Subset 2 of Evaluation 4 for hydrogen bond (left), π interaction (middle), and both types (right). Value of potential is shown in the parenthesis after each interaction name.

3.5. Validation of significant interactions

We have previously identified, in a comprehensive non-redundant set of 299 RNA-protein complexes, cyclic amino acids TRP, HIS, TYR,
and PHE (pi-stackers) as well as Pro, having increased percentage frequencies of non-H bond contacts when compared to H-bonds (Ma, C., Suwandi, E. and Lustig, B., unpublished results). In addition, increases in relative frequency were shown for ALA, ASN, GLN, GLY, ILE and quartet of bases (Ma, C., Suwandi, E. and Lustig, B., unpublished results). LEU has been shown to be a somewhat less significant participant as opposed to LYS [47, 48]. The importance of ARG has shown a strong preference in such interactions with U, A, C opposed to G [46]. The application of the linear scoring approach noted here shows significant advantage in characterizing the native structure with respect to decoy structures as compared to an approach requiring the normalization associated with a reference state. This leaves open the general utility of ECR filtering, where our pre-screening of the learning set does not offer any significant advantage with respect to ITScore-PR potentials [18]. However, our results are consistent with the notion that iterative methods that do not incorporate a reference state, including ITScore-PR, are advantageous [56]. It was clear in our earlier efforts that ECR was required to allow for reasonable convergence in learning set ranking scores and the addition of the limited set of \( \pi \) interactions further optimized the resulting potential. Future studies include adding other interactions, notably addressing salt-bridges specifically.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

This work included access to TSUBAME 3.0 supercomputing system at Tokyo Institute of Technology, Japan.

### References


