Evaluating mixture models for building RNA knowledge-based potentials.

Adelene y L Sim, Olivier Schwander, Michael Levitt, Julie Bernauer

Published in final edited form as:
J Bioinform Comput Biol. 2012 April ; 10(2): 1241010. doi:10.1142/S0219720012410107.

EVALUATING MIXTURE MODELS FOR BUILDING RNA KNOWLEDGE-BASED POTENTIALS

Adelene Y. L. Sim[1,5], **Olivier Schwander**[2,6], **Michael Levitt**[3,7], and **Julie Bernauer**[4,8]

[1]Department of Applied Physics Stanford University, Stanford, CA 94305-4090, USA

[2]Combinatorial Models Team, Laboratoire d'Informatique (LIX), Ecole Polytechnique, 91128 Palaiseau, France

[3]Department of Structural Biology Stanford University, Stanford, CA 94305-5126, USA

[4]INRIA AMIB Bioinformatique, Laboratoire d'Informatique (LIX), Ecole Polytechnique, 91128 Palaiseau, France

Abstract

Ribonucleic acid (RNA) molecules play important roles in a variety of biological processes. To properly function, RNA molecules usually have to fold to specific structures, and therefore understanding RNA structure is vital in comprehending how RNA functions. One approach to understanding and predicting biomolecular structure is to use knowledge-based potentials built from experimentally determined structures. These types of potentials have been shown to be effective for predicting both protein and RNA structures, but their utility is limited by their significantly rugged nature. This ruggedness (and hence the potential's usefulness) depends heavily on the choice of bin width to sort structural information (e.g. distances) but the appropriate bin width is not known *a priori*. To circumvent the binning problem, we compared knowledge-based potentials built from inter-atomic distances in RNA structures using different mixture models (Kernel Density Estimation, Expectation Minimization and Dirichlet Process). We show that the smooth knowledge-based potential built from Dirichlet process is successful in selecting native-like RNA models from different sets of structural decoys with comparable efficacy to a potential developed by spline-fitting — a commonly taken approach — to binned distance histograms. The less rugged nature of our potential suggests its applicability in diverse types of structural modeling.

Keywords

Knowledge-based potential; mixture models; RNA structure

Correspondence to: Julie Bernauer.
[5]adelene@stanford.edu
[6]schwander@lix.polytechnique.fr
[7]michael.levitt@stanford.edu
[8]julie.bernauer@inria.fr

1. Introduction

Ribonucleic acids (RNAs) are important biological molecules that carry out a range of molecular processes, including protein synthesis, gene regulation and biochemical catalysis. In order to correctly function, RNA has to fold into a three-dimensional form that involves interactions between water, metal ions, small ligands and proteins. RNA, in general, tends to fold hierarchically such that its sequence defines its secondary structure consisting of a set of stable base-pairing interactions. This set of base-paired helices then rearrange to form its stable functional tertiary structure (or "native" state) under physiological conditions.

The knowledge of RNA structure is therefore crucial in enriching our understanding of RNA function. In particular, the higher the resolution at which we can analyze RNA structure, the more details we can extract, such as the importance of specific RNA sequences for particular functional roles. It was shown recently that with high-resolution structural information, it is possible to design RNA molecules with functions that mimic naturally occurring biological processes.[1]

Detailed structural information is typically obtained from experimental means such as X-ray crystallography and nuclear magnetic resonance (NMR) studies. Unfortunately, not all RNA molecules can be crystallized, and structural analysis by NMR is confined to small- or medium-sized RNA only. Other solution techniques circumvent these problems, but are still unable to provide high-resolution RNA structural information.

Simulations using physics-based potentials, conversely, give structural insights in atomic detail, but these potentials are limited in their accuracy, particularly in predicting native-like molecular structures of proteins and RNA.[1–3] Furthermore, the need to model solvent at the atomic level makes molecular modeling highly intractable, particularly for time scales of biological interest.

One approach to circumvent the modeling of solvent is to use information (e.g. dihedral and torsional angles[1,4,5] and inter-atomic distances).[6,7] from native structures of RNA to generate knowledge-based (KB) potentials. When building these potentials, it is assumed that solvent effects are implicitly contained in the structural information of the native RNA structures. The efficacy of the method depends on the information extracted, and how it is used to generate the KB potential. One source of information that is commonly used is inter-atomic pair-wise distances, and our recent work has shown that a KB potential built using distances is sufficient for screening native-like RNA, without the need to separate KB terms into physical components like other RNA KB potentials.[5,8,9]

Once these distances are mined, they can be converted to energy functions (see Sec. 2) by the potentials of mean-force method.[10,11] A common approach is to generate a histogram for each pair-wise distance (user-defined bin width) that is then normalized by a reference distribution. Next, these discretized normalized histograms are spline-fitted to give smooth functions for gradient-based calculations (e.g. molecular dynamics or energy minimization), resulting in a high-dimensional potential that is usually very rugged.[12] It is unclear if this ruggedness is an effect of improper binning, or true structural features.

To avoid issues associated with binning, in our aforementioned RNA KB potential, we made use of Dirichlet process mixture (DPM) models, which do not require user-defined bin sizes and instead build RNA potentials based on raw distance distributions directly (see Sec. 2 and Ref. 6 for more details). Here, as a follow-up, we attempted data-fitting using other mixture models (MMs) and compared their efficacy in capturing the raw distance distributions, and in selecting native-like RNA structures. As an additional test, we compared the quality of KB potentials generated using MMs to that generated using quintic spline-fitting, a commonly used alternative approach.[2]

Amongst the different MMs tested, the DPM and Kernel Density Estimation (KDE) models captured the true distance distributions well, while the Expectation-Minimization (EM) model resulted in significant deviation. We further compared the models in their ability to screen native-like RNA structures from sets of different decoys. The smooth KB potential from DPM is equally capable as the splined KB potential in selecting native-like RNA structures but the KB potential from KDE modeling fared much worse. Therefore DPM modeling appears to be the best method for smoothing KB potentials while preserving important structural information within the potentials for biological applications.

2. Materials and Methods

2.1. Data collection and potential construction

All distances were obtained from a non-redundant dataset specially curated for obtaining KB potentials.[6] Two different strategies were used to process distances compiled from a previously described five-atom per nucleotide coarse-grained model (total of 20 atom types)[6]: (i) building of histograms from raw data and (ii) estimating distance distributions using MMs. The histograms were calculated with an optimized bin size of 0.13 Å obtained by density estimation.[13] using the R software suite. MMs were built using three different techniques: EM, DPMs and KDE. Simplified versions of the mixture models were also built when applicable (see Sec. 2.2).

For the histogram-based model, counts are directly converted to energy potentials using the Lu and Skolnick formalism.[10,14] The total energy E of a given conformation can be expressed as the sum over potentials for all pairs of atoms i and j at distance d_{ij} apart[10]:

$$E = -kT\sum_{ij}\ln\left(\frac{p_{\mathrm{obs}}(d_{ij})}{p_{\mathrm{ref}}(d_{ij})}\right), \quad (1)$$

where T is the temperature (taken to be 300 K) and k the Boltzmann constant. $p_{obs}(d_{ij})$ and $p_{ref}(d_{ij})$ represent the observed and reference probabilities respectively for the atoms i and j separated by distance d_{ij}. Like in our previous study,[6] the reference distribution was chosen from the Samudrala and Moult model.[10] Due to numerical instabilities and inaccuracies in the low count region (below 2.5 Å), corrections were done as previously described.[2]

For the MMs, after each distance distribution was converted into a potential, the low count region was corrected by a linear approximation from the origin to the first descending

inflexion point (first observed basin). Smooth truncation at the 14Å cutoff distance is ensured by multiplying each potential by a negative sigmoid function.[6]

2.2. Mixture models

MMs were used to estimate the distance densities (reference probabilities) to get analytically differentiable potential functions. Each density is modeled as a mixture of univariate Gaussian distributions. This mixture has the general form:

$$P(d)=\sum_{i=1}^{N}\omega_i N(\mu_i,\sigma_i^2) \quad (2)$$

with

$$\sum_{i=1}^{N}\omega_i=1. \quad (3)$$

The MMs were obtained by estimating the parameters ω_i, μ_i and σ_i in order to maximize the quality of the approximation.

2.2.1. Expectation-Maximization—The EM algorithm is a common tool used to estimate the parameters of an MM.[15] The likelihood of the density estimation is maximized by iteratively computing the expectation of the log-likelihood using the current estimate of the parameters (E step) and by updating the parameters in order to maximize the log-likelihood (M step). The pitfall is that this method leads only to a local maximum of the log-likelihood. Another drawback is that the number of components has to be carefully chosen, and in some cases this choice is not straightforward.

2.2.2. Dirichlet process mixtures—Using an MM with an infinite number of components is a way to avoid having to choose the number of components. This can be done with a DPM model,[16] where a Dirichlet process is used to build priors for mixing proportions of components. DPMs were computed using the *fbm* package.[17] To get a finite mixture, the components are sorted according to their weights ω_i and a threshold is defined for the components to be kept. The main disadvantage of this approach is that a Dirichlet process needs to be evaluated by a Monte Carlo Markov Chain using the Metropolis algorithm, which is computationally costly.

2.2.3. Kernel density estimation—KDE[18] (also known as the Parzen windows method) uses one component (a Gaussian kernel) centered on each point of the dataset. All the components share the same weight and since the Gaussian means μ_i are directly obtained from the data points, the only remaining optimizable parameter is the standard deviations σ_i^2, which are chosen to be equal to a constant called the bandwidth. The critical part of the algorithm is the choice of the bandwidth and various methods can be used to automatically tune this parameter; we used the Sheather and Jones technique from the R software.[13] Since for this model there is one Gaussian component per point in the dataset, it is difficult to build the mixture from kernel density as it is large and basic operations can be slow

(evaluation of the density, random sampling, etc.), an issue further accentuated by the need to loop over all the components of the mixture during evaluation.

2.2.4. Simplification of mixture models—KDEs are very precise approximations of the density function but their size — the number of components in the mixtures — is very large. Conversely, models built with the EM algorithm are compact and usually precise (assuming a local minimum is found close to the global optimum). The simplification of MMs allows the building of a compact mixture from a very large but precise MM. The Bregman hard clustering algorithm was used to make the simplification[19,20] with the components of the mixtures compared by a k-means algorithm using the Kullback—Leibler divergence. We used pyMEF,[21] a generic library for handling mixtures of exponential families.

2.2.5. Comparison of models—Log-likelihood is the primary criterion used for evaluating the quality of the density estimation function. The best performing algorithm in terms of log-likelihood was chosen as the reference. We then compared each method to the reference using the Kullback—Leibler divergence evaluated by a Monte Carlo method available in PyMEF.[21] The purpose of this comparison is to evaluate the tradeoff between speedup and accuracy.

2.3. Biological evaluation

We evaluated the potentials with two quantitative criteria as previously published[6]: the number of decoys that scored lower than the native structure (N_{bad}) and the Enrichment Score *(ES).*[22] *ES* is based on the degree of overlap between structures that are in the top 10% scoring ($Etop_{10}$%) and those that have the best 10% root mean squared deviation (RMSD) to native ($Rtop_{10}$%). It is defined as:

$$ES = \frac{|E_{\text{top10\%}} \cap E_{\text{top10\%}}|}{0.1 \times 0.1 \times N_{\text{decoys}}}, \quad (4)$$

where $|E_{top10\%} \cap R_{top10\%}|$ is the number of structures in the intersection of $E_{top10\%}$ and $R_{top10\%}$. An ideal scoring function has an ES of 10, a random scoring function has as ES of 1 and an ES below 1 indicates a bad scoring function.

Both criteria provide insight on how the potentials perform in selecting native-like structures from a set of decoys. We evaluated the different models by scoring decoys generated by different means, namely position restrained replica exchange molecular dynamics,[23] normal modes[2] and fragment assembly (by FARNA).[8] Information about decoy generation has been described elsewhere.[6]

3. Results and Discussion

3.1. Comparison of mixture models

There are several ways to estimate which MM best represents the data and therefore is best for building a KB potential. In this case, visual inspection of the model fits to distance distributions provides insight, especially in determining if the distance distribution peaks are

correctly captured by the modes obtained from the MMs. Nonetheless, a quantitative measure is still required to distinguish between the quality of the different MMs that correctly model the first peak in distance distributions. One such measure is the log-likelihood criterion that is optimized in EM algorithms for a specific distance measurement set but unfortunately this measure cannot be directly compared between different pair-wise distance sets.

To provide a global comparison between MMs, we computed the ranks of each MM based on the log-likelihood criterion for each of the 210 pair-wise distance sets of the coarse-grained RNA representation (five atoms per nucleotide). We compared seven models: DPM (Dirichlet process model), KDE (Kernel Density Estimation), KDEs (Kernel Density Estimation followed by simplification), EM8 (Expectation Maximization with 8 components), EM8s (Expectation Maximization with 8 components followed by simplification), EM12 (Expectation Maximization with 12 components) and EM12s (Expectation Maximization with 12 components followed by simplification). For each distance measurement, the best performing MM in terms of log-likelihood is ranked 1 and the worst performing model is ranked 7. The average rank and the rank distribution then indicate the best performing model. Figure 1 shows a heat map of the ranks for each MM. Ranks based on log-likelihood show that DPM is by far the best model, ranking number 1 in more than 98% of the cases. The KDE-based models are next best, followed byEM8. The rank distributions of both KDE and EM8 are wider than that for DPM, suggesting that these models perform well on average but fail in capturing the distributions of some pair-wise distances. The global comparison also shows that k-means simplification drastically worsens the estimates; this could be due to a bad fit or to too few final components used during simplification. Interestingly, EM8 shows better results than EM12, indicating that an increase in the number of components in MMs does not always improve the fit to raw data.

In addition, we visually compared the modeled distributions to the raw distance sets. An example is provided in Fig. 2 for the gC4-gC4 distance set. DPM and KDE models capture the largest peaks and seem to be good coarse estimators of the gC4-gC4 distance distribution. Due to the small number of components used, EM models are easily computed, but this advantage comes at the expense of less accurate fit s to the data. Most notably, while the first peaks are well captured, the distance distributions at longer distances tend to be over-fitted (as judged by the excessive oscillations of the model fit), 1a problem accentuated by the fact that data at these distances tend to be more noisy than those at smaller distances (< 10 Å). For KDE, the model fit s to raw distance distributions appear to be more reasonable. However, KDE results in a complicated functional form that could slow down full energy calculations within any molecular modeling program. The functional form of DPM is simpler than that of KDE, but this does not appear to compromise the quality of model fit to data (Figs. 2 and 3). The benefit s of a simple functional form and the high quality of fit to raw data outweigh the disadvantage of the longer computational time needed to obtain DPM.

For further quantitative comparison, the Kullback—Leibler divergence was computed between all models for each distance set and the average results are shown in Table 1.

DPM and KDE models are relatively similar to each other, even for the simplified version of KDE, whereas EM based models are very different. Since DPM and KDE fit the data well but significant deviation was observed for EM, we focused only on the potentials from DPM and KDE for the following biological scoring assessment.

3.2. Assessing mixture models by scoring RNA structural decoys

While it is vital to quantitatively compare models to the raw distance distributions, it is also crucial to compare, from a biological perspective, the quality of the different KB potentials. In particular, the energy landscape of any molecule is high dimensional, and therefore it is not sufficient to compare the KB potentials just by looking at fit s to individual pair-wise distributions. We focus on one aspect of biology that KB potentials are useful for: selecting a near-native molecular structure from a set of decoys. To do this, we made use of RNA decoys generated as described previously,[6] and quantitatively compared the different KB potentials generated using KDE, DPM and spline-fitting (EM models were omitted due to the poor fitting to distance distributions as judged visually and using the log-likelihood measure).

Spline-fitting is a common approach to generate KB potentials,[2] hence for comparison to MMs, we generated a KB potential using the spline-fitting protocol as previously used for proteins.[2] Each distance distribution and reference distribution was normalized and binned into widths of 0.13 Å. This bin value was chosen as it is the optimum bandwidth of the KDE model (see above) and it is very close to the 0.1 optimal bin size obtained in a previous published study on proteins.[2] The negative logarithm of the ratio (distance distribution over reference distribution) was then spline-fitted to generate smooth KB potentials for each distance set (with appropriate corrections at low distances).[2] The spline-fitting results in smooth fit s to the data, but resultant KB potentials often were very rugged compared to those from MMs (see Fig. 3). This is a complexity that could have arisen due to true structural features and/or inappropriately chosen bin sizes. It was also shown for proteins that the final form of the potential and hence its efficacy is dependent on the chosen histogram bin size.[2,24]

From Table 2, we observed that the KB potential from KDE fit s performed significantly worse in screening RNA decoys than DPM and spline-fitting, a result that was not directly obvious from the KDE fit to the distance distributions. The KB potential generated from DPM fit s gives results that are comparable than those using the KB potential determined from spline-fitting; both potentials are similarly able to distinguish near-native structural models.

Since the potentials from DPM are less rugged than those obtained via spline-fitting, we expect to observe more substantial differences between both versions of potentials for energy minimization of structures. We plan to explore this in the future.

4. Perspectives and Conclusion

Our assessments of KB potentials derived from different MMs suggest that DPM modeling is an efficient approach to generate smooth, differentiable KB potentials of RNA that

preserve important biological information. Applications of traditional KB potentials (derived from spline-fitting) in biological structural modeling have often been limited by excessive ruggedness of KB potentials. Our study shows that the use of an appropriate MM (e.g. DPM), provides a less rugged KB potential with similar structure scoring properties. Hence the KB potential derived from DPM could plausibly be more versatile than the traditional version, thereby allowing extensive and effective applications in molecular modeling like minimization and sampling.

Acknowledgments

The authors would like to thank Frank Nielsen for his insightful comments and help on the EM-based models. This work is part of the “GNAPI Associate Team”. The authors thank the INRIA Équipe Associee program for financial support. A.Y.L.S. acknowledges support from the Agency for Science, Technology and Research (A*STAR), Singapore. M. L. is the Robert W. and Vivian K. Cahill Professor of Cancer Research and is supported by NIH GM063817. The authors acknowledge support from NSF award CNS-0619926 for computer resources.

References

1. Das R, Karanicolas J, Baker D. Atomic accuracy in predicting and designing non-canonical RNA structure. Nat Methods. 2010; 7:291–294. [PubMed: 20190761]
2. Summa CM, Levitt M. Near-native structure refinement using in vacuo energy minimization. Proc Natl Acad Sci USA. 2007; 104:3177–3182. [PubMed: 17360625]
3. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2010; 18:342–348. [PubMed: 18436442]
4. Frellsen J, Moltke I, Thiim M, Mardia KV, Ferkinghoff-Borg J, Hamelryck T. A probabilistic model of RNA conformational space. PLoS Comput Biol. 2009; 5:e1000406. [PubMed: 19543381]
5. Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA. 2009; 15:189–199. [PubMed: 19144906]
6. Bernauer J, Huang X, Sim AY, Levitt M. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. RNA. 2011; 17:1066–1075. [PubMed: 21521828]
7. Capriotti E, Norambuena T, Marti-Renom MA, Melo F. All-atom knowledge-based potential for RNA structure prediction and assessment. Bioinformatics. 2011; 27:1086–1093. [PubMed: 21349865]
8. Das R, Baker D. Automated de novo prediction of native-like RNA tertiary structures. Proc Natl Acad Sci USA. 2007; 104:14664–14669. [PubMed: 17726102]
9. Flores SC, Altman RB. Turning limited experimental information into 3D models of RNA. RNA. 2010; 16:1769–1778. [PubMed: 20651028]
10. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol. 1998; 275:895–916. [PubMed: 9480776]
11. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol. 1990; 213:859–883. [PubMed: 2359125]
12. Chopra G, Summa CM, Levitt M. Solvent dramatically affects protein structure refinement. Proc Natl Acad Sci USA. 2008; 105:20239–20244. [PubMed: 19073921]
13. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for Kernel Density Estimation. Royal Statistical Society. 1991; 53:683–690.
14. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins. 2001; 44:223–232. [PubMed: 11455595]
15. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc B (Methodological). 1977:1–38.

16. Rasmussen CE. The infinite Gaussian mixture model. Adv Neural Inform Proc Sys. 2000; 12:554–560.

17. Neal, RM. Technical Report No. 9815. Department of Statistics, University of Toronto; 1998. Markov chain sampling methods for Dirichlet process mixture models.

18. Parzen E. On estimation of a probability density function and mode. Ann Math Stat. 1962; 33:1065–1076.

19. Banerjee A, Merugu S, Dhillon IS, Ghosh J. Clustering with Bregman divergences. J Mach Learn Res. 2005; 6:1705–1749.

20. Garcia V, Nielsen F, Nock R. Levels of details for Gaussian mixture models. Computer Vision, ACCV 2009. 2010:514–525.

21. Schwander O, Nielsen F. Statistical Signal Processing Workshop (SSP). 2011

22. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins. 2003; 53:76–87. 2003. [PubMed: 12945051]

23. Huang ES, Subbiah S, Tsai J, Levitt M. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. J Mol Biol. 1996; 257:716–725. [PubMed: 8648635]

24. Melo Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci. 2002; 11:43–48.

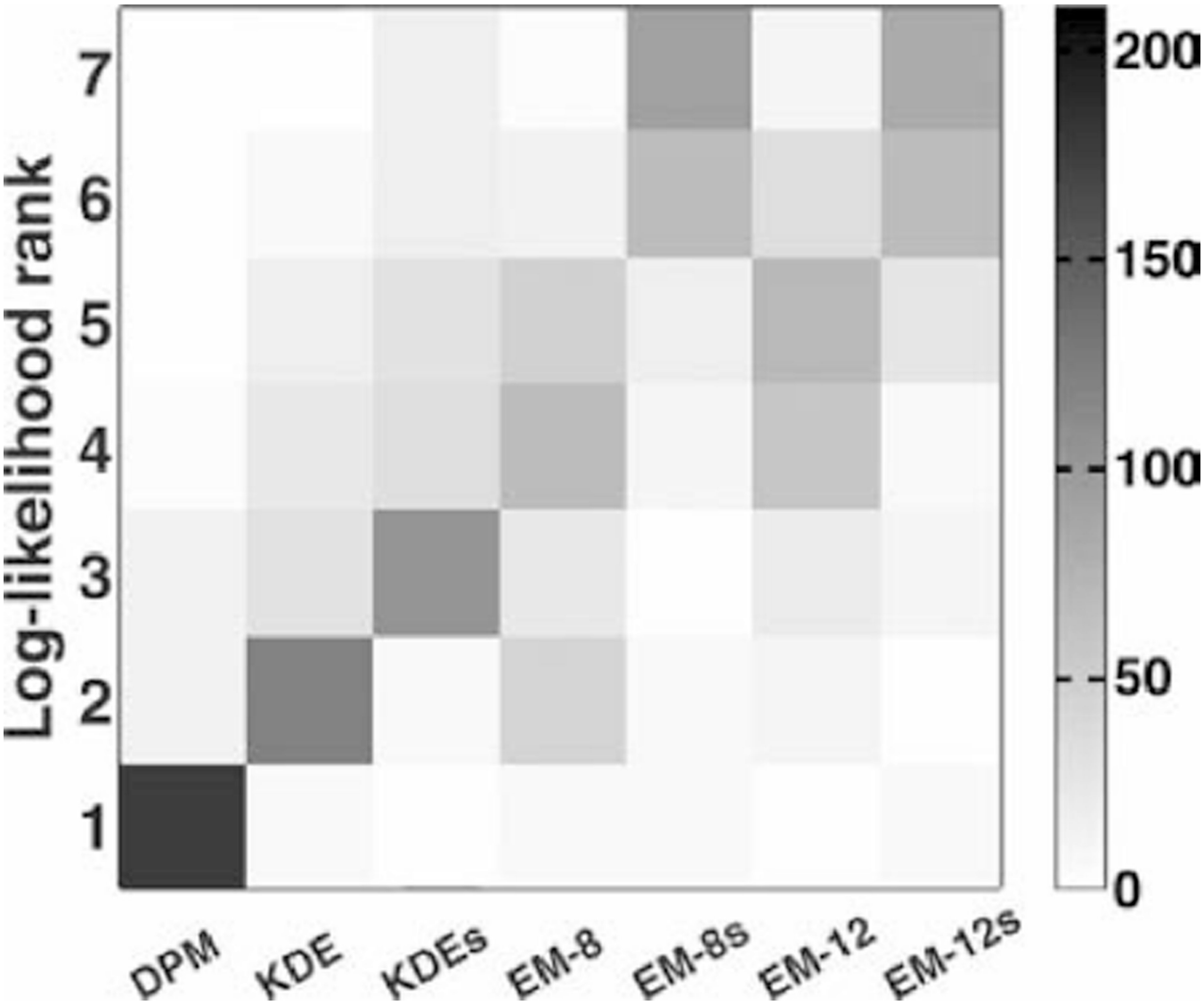

Fig. 1.

Evaluation of fit of mixture models to the complete set of raw pair-wise distance data (210 pairs for the 5-pt representation). All models were ranked based on the quality of fit (using the log-likelihood criterion; see text for details) for each pair-wise distance distribution. Results are shown as a heatmap, with low to high frequency color coded from white to black.

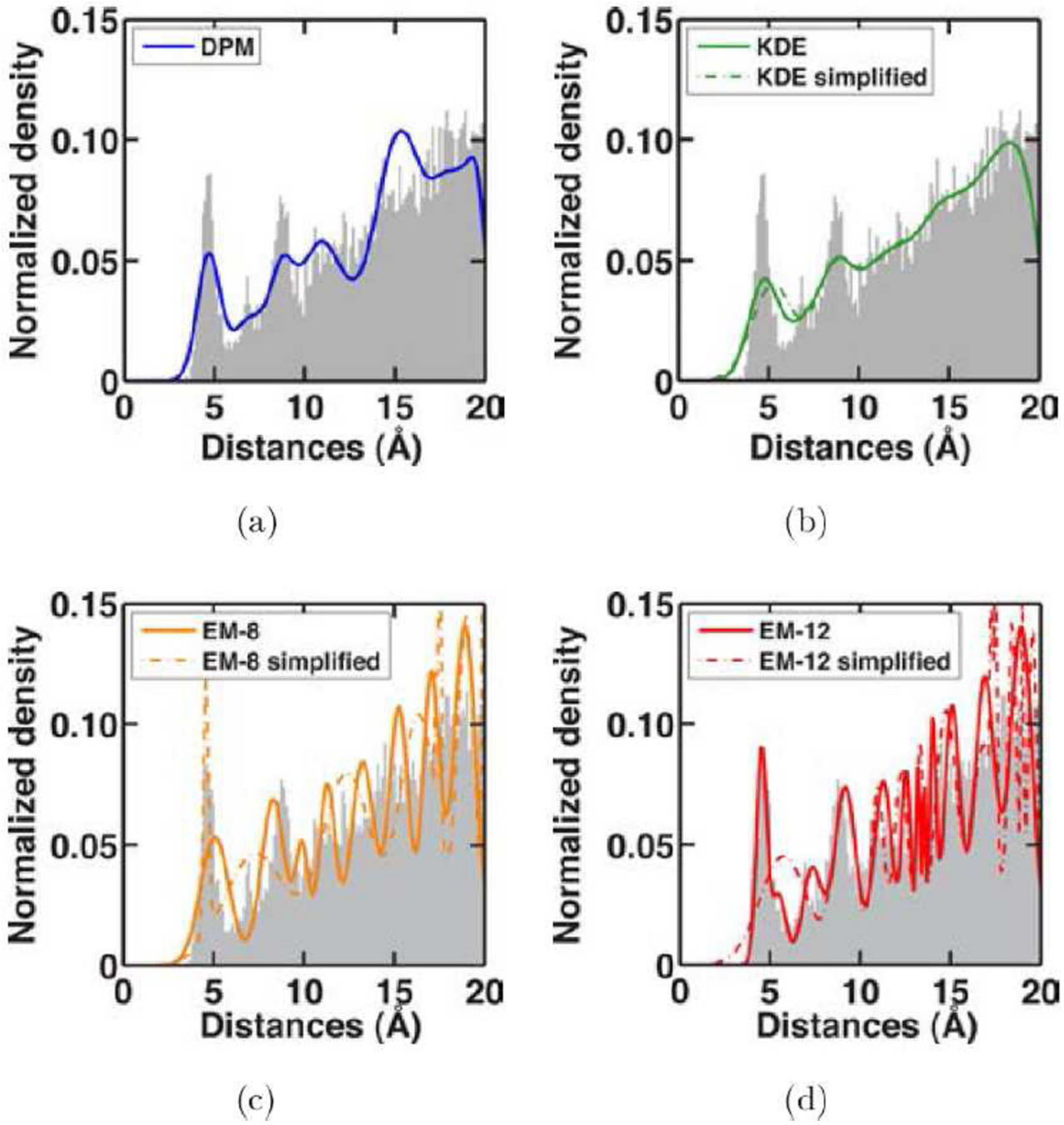

Fig. 2.

Raw distance distribution for gC4—gC4 (grey) compared to different mixture models: DPM (a), KDE (b), EM-8 (c) and EM-12 (d). EM fit s are quick to obtain but are less than ideal representations of the raw distribution. KDE is quick and yields a good fit, however, it has a more complicated functional form than DPM, therefore requires more computational time to evaluate the energy of each molecular structure. DPM is the ideal model due to the quality of fit to the raw data and the simplicity of its functional form. These advantages offset the longer time needed to obtain the DPM fit.

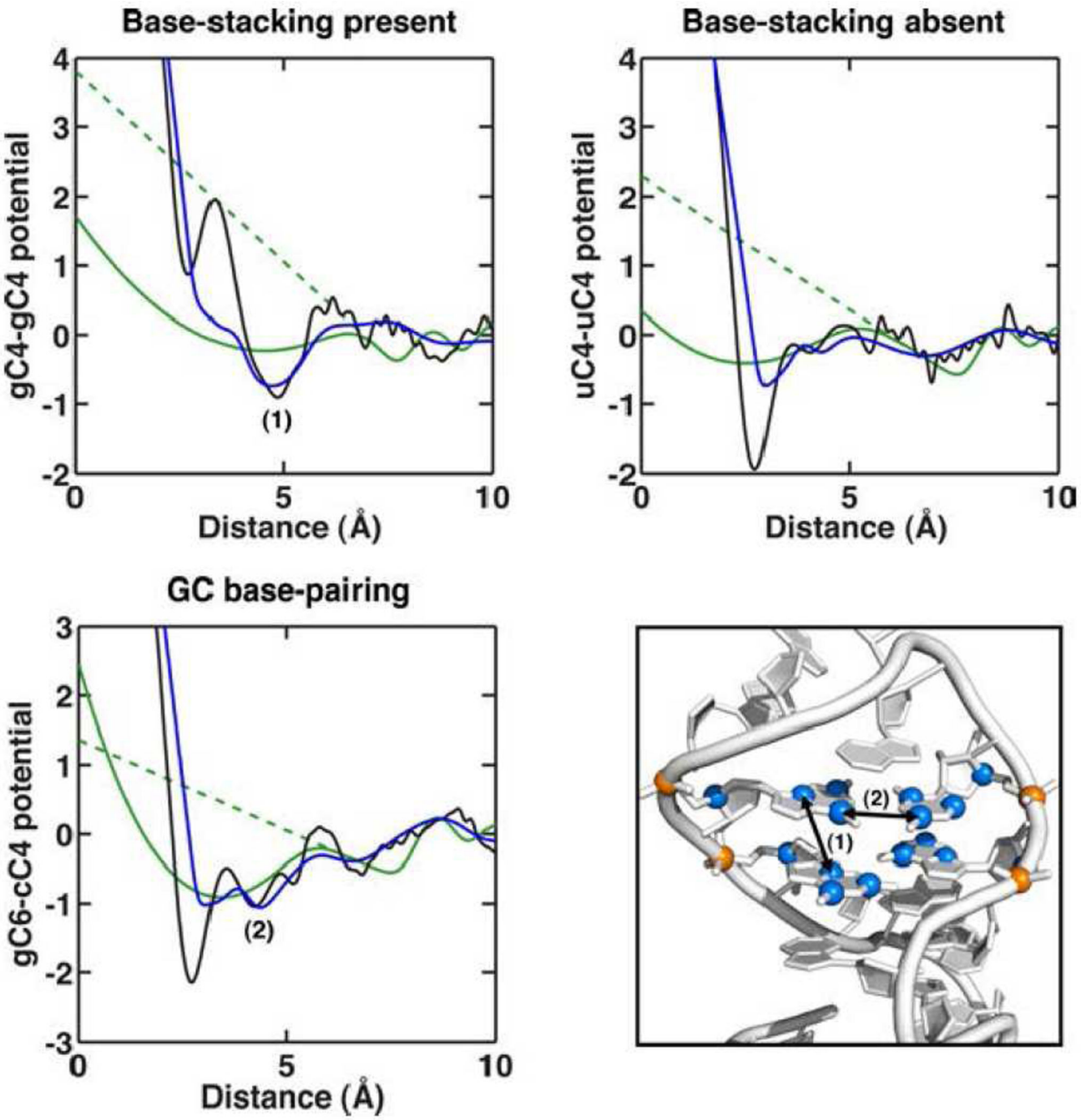

Fig. 3.

Comparison of the potentials obtained for the gC4—gC4, uC4—uC4 and gC6—cC4 distances using DPM with low count correction (blue), KDE without low count correction (green), KDE with low count correction (dashed green) and spline-fitting (black). The KDE models appear to be excessively smoothed since some distance potential basins are absent. The KB potential from spline-fitting is significantly more rugged than that from the DPM model. The spline-fitted potential wells are also more pronounced (particularly for the first

well), which could be an artifact of the spline-fitting model. Distances corresponding to (1) and (2) labeled in the potential graphs are shown in the atomic figure.

Table 1

Means and standard deviations of the Kullback—Leibler divergences between the different models computed for 210 distance sets. DPM and KDE are very similar to each other, regardless of simplification. EM models are significantly different and are also dissimilar to each other. Simplification introduces a noticeably large variation in the results.

KL	KDE	KDEs	DPM	EM8	EM12	EM8s	EM12s
KDE	0.0 ±0.0	0.008 ±0.0	0.018 ± 0.003	0.094 ±0.012	0.177 ±0.114	0.154 ±0.02	0.142 ± 0.032
KDEs	0.033 ±0.014	0.0 ± 0.0 0.039 ± 0.008	0.058 ± 0.023	0.138 ±0.035	0.234 ±0.177	0.196 ± 0.063	0.148 ± 0.015
DPM	0.021 ±0.002		0.0 ±0.0	0.128 ±0.214	0.139 ±0.34	0.133 ± 0.186	0.093 ± 0.017
EM8	0.279 ±0.026	0.269 ± 0.017	0.266 ±0.014	0.0 ±0.0	0.356 ± 0.261	0.323 ± 0.035	0.325 ±0.046
EM12	0.328 ±0.014	0.324 ± 0.011	0.308 ±0.007	0.336 ±0.009	0.0 ± 0.0	0.364 ± 0.018	0.364 ±0.013
EM8s	1.034 ± 2.854	0.864 ± 1.778	0.856 ± 1.539	1.068 ±3.227	1.109 ± 3.242	0.0 ± 0.0	0.864 ± 1.986
EM12s	0.871 ± 3.334	0.784 ±2.551	0.754 ±2.172	0.887 ±3.029	0.94 ±3.574	0.749 ± 1.648	0.0 ±0.0

Table 2

Summary of the ability of different KB potentials to screen native-like RNA models from three types of molecular decoys (position-restrained replica-exchange molecular dynamics, normal modes and fragment assembly by FARNA).

			Number of Structures below Native Energy (N_{bad})				Enrichment Score (ES)			
Decoy generation method	**RNA**	**Experimental method**	**Spline**	**DPM**	**KDE**	**KDE raw**	**Spline**	**DPM**	**KDE**	**KDE raw**
(A) Position-restrained REMD	1duq	X-ray	0	1	2536	1044	8.4	8.0	0.1	2.9
	1f27	X-ray	0	0	2051	432	8.2	7.9	0.3	2.2
	1msy	X-ray	0	13	1447	536	6.8	6.3	0.1	4.0
	1nuj	X-ray	0	0	1154	266	7.6	7.0	0.8	5.7
	434d	X-ray	0	0	2655	479	8.4	7.9	0.1	3.4
(B) Normal modes	1duq	X-ray	0	0	490	35	6.6	6.6	1.2	6.0
	1esy	NMR	0	0	470	44	4.6	5.0	0.0	3.8
	1f27	X-ray	0	0	490	0	5.8	5.4	0.2	3.2
	1i9v	X-ray	0	0	480	235	3.0	4.4	0.0	3.6
	1kka	NMR	0	8	374	102	4.6	5.2	0.2	4.8
	1msy	X-ray	0	0	490	0	5.4	4.6	0.0	3.6
	1nuj	X-ray	0	0	468	0	6.4	6.0	5.4	4.6
	1qwa	NMR	0	0	468	161	2.4	3.0	0.0	1.2
	1x9k	X-ray	0	0	480	0	1.8	2.4	0.2	2.2
	1xjr	X-ray	0	0	480	0	7.8	7.2	1.4	4.0
	1ykq	X-ray	0	0	481	99	4.6	5.6	0.0	4.8
	1zih	NMR	0	0	484	10	4.6	4.6	0.2	3.6
	28sp	NMR	0	0	489	0	4.6	4.6	0.0	5.0
	2f88	NMR	0	0	486	200	4.8	5.0	2.6	4.8
	434d	X-ray	0	0	476	0	6.8	7.0	0.4	6.8
(C) FARNA	157d	X-ray	0	0	444	0	2.8	2.5	1.2	2.5
	1a4d	NMR	39	1	500	0	1.7	2.0	2.7	1.0
	1csl	X-ray	0	0	262	0	1.5	1.5	1.3	1.2
	1dqf	X-ray	0	0	390	61	2.8	3.3	2.3	1.5
	1esy	NMR	502	419	126	80	2.7	2.5	0.3	1.8

Decoy generation method	RNA	Experimental method	Number of Structures below Native Energy (N_{bad})				Enrichment Score (ES)			
			Spline	DPM	KDE	KDE raw	Spline	DPM	KDE	KDE raw
	1i9x	X-ray	0	0	340	2	1.8	1.8	1.5	1.7
	1j6s	X-ray	211	244	315	0	0.5	0.8	0.8	1.2
	1kd5	X-ray	0	0	380	2	1.2	1.8	1.3	1.0
	1kka	NMR	505	502	110	501	0.7	1.3	0.7	1.0
	1l2x	X-ray	0	9	447	67	0.5	0.5	0.0	0.5
	1mhk	X-ray	0	58	505	226	1.2	1.0	1.0	0.3
	1q9a	X-ray	161	302	505	280	1.0	1.5	1.7	0.8
	1qwa	NMR	501	434	108	492	1.3	1.7	1.0	1.0
	1xjr	X-ray	0	2	492	0	1.3	1.8	1.5	2.0
	1zih	NMR	205	92	503	47	5.3	5.8	3.5	5.7
	255d	X-ray	0	5	492	259	0.8	1.5	1.7	1.0
	283d	X-ray	0	0	497	10	1.3	1.2	0.8	1.3
	28sp	NMR	488	463	503	78	1.5	1.7	0.8	1.7
	2a43	X-ray	0	180	501	300	2.2	1.7	0.7	1.3
	2f88	NMR	147	13	504	90	2.5	2.7	3.0	1.7
AVERAGE VALUES	(A)	5	0	3	1969	551	7.9	7.4	0.3	3.6
	(B)	15	0	1	474	59	4.9	5.1	0.8	4.1
	(C)	20	138	136	396	125	1.7	1.9	1.4	1.5
	X-Ray	27	70	80	708	200	8.2	8.3	1.8	5.6
	NMR	13	68	44	443	57	3.2	3.4	1.3	2.8
	All	40	69	69	622	153	3.7	3.8	1.0	2.8

Note: In all cases, the KB potential by KDE (with and without low count corrections; "KDE raw" indicates the latter) performed significantly worse than the KB potentials from spline-fitting and DPM. Both potentials from spline-fitting and DPM yield comparable results in decoy screening.