Additional File 1: Supplementary information

1 The ssHMM

1.1 Schematic overview of our Gibbs sampling approach

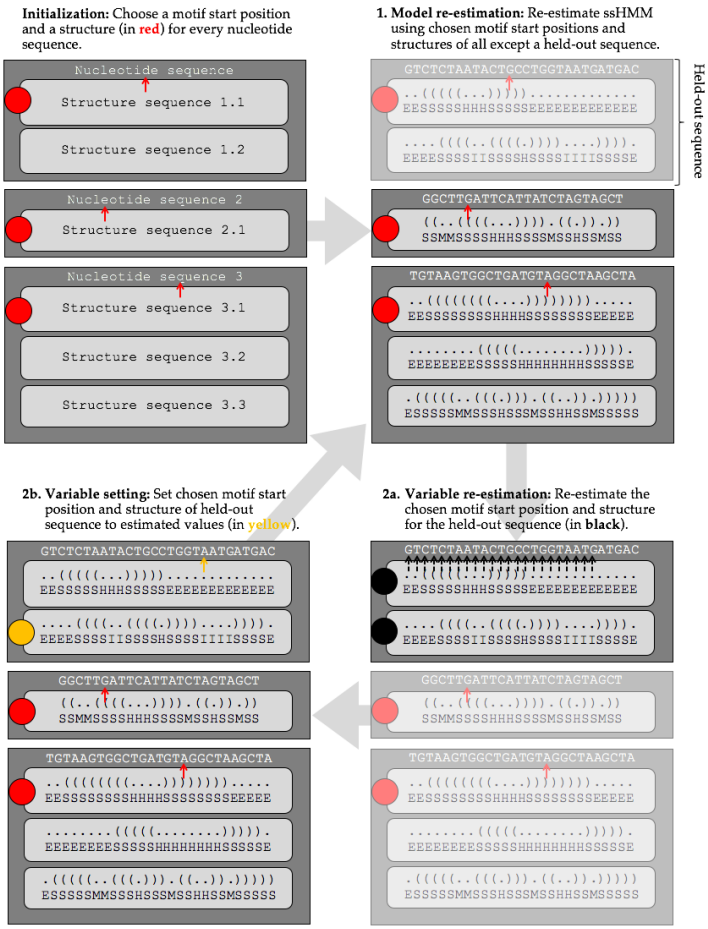


Figure S1: **Schematic overview of our Gibbs sampling approach** The dark gray boxes represent the RNA sequences. Each RNA sequence is characterized by its nucleotide sequence and several possible structure sequences (light gray). During initialization, a structure and a motif start (red) are set for each sequence. Every iteration consists of 3 steps: (1) One of the sequences is randomly chosen. Then, the ssHMM is re-estimated using all but this chosen sequence; (2) The two unknown variables of the held-out sequence are re-estimated. For this purpose, the probability of every combination of structure and motif start (black) given the ssHMM is calculated; (3) According to the distribution of these probabilities, a new structure and a new motif start is drawn (yellow). These three steps are repeated until termination.

1.2 Parameters of the motif finder

Our motif finder ssHMM can be customized by a set of six different parameters (see Table S1). Initialization has been discussed in the Methods section. The other parameters are described in the following.

1.2.1 Motif length

The ssHMM possesses a set of states for each motif position. Therefore, the length of the searched motif has to be defined prior to any training. For most RBPs, however, the size of the binding site is unknown.

We therefore recommend to try different biologically meaningful motif lengths n for an RBP (e.g. from 4 to 12) and to inspect the resulting average sequence-structure information content per position (see Sec. 3.7). In other words, we propose to run ssHMM several times with several n and, at the end, pick the model with best n according to an empirical rule. We evaluate the trained models both on the used motif length (n) and average per-position sequence-structure information content (i_n). Unfortunately, longer motifs tend to have a lower information content. To find a good compromise between n and i_n, we suggest the following simple heuristic for determining the best motif length b: $b = \text{argmax}_n \, i_n + (n * 0.15)$.

Such a heuristic prefers the higher motif length $n + 1$ over the lower motif length n only if the average information content i_{n+1} of its resulting trained model is no more than 0.15 less than for the shorter motif.

1.2.2 Block size

Gibbs sampling generates a Markov chain of samples from the multivariate probability distribution over all unknown variables. Each of these samples is correlated with nearby samples. To reduce this correlation and allow the algorithm to make progress faster, one can employ block sampling [1].

In the original Gibbs sampling algorithm, exactly one sequence is chosen in each iteration to have its unknown variables re-estimated. A blocked Gibbs sampler, in contrast, re-estimates the unknown variables of multiple sequences in one iteration. Thus, the ssHMM is not re-estimated for every single variable re-estimation. This can have two advantages: (1) a more stable optimization of the ssHMM and (2) an improved runtime.

The block size parameter determines the number of sequences for which the two unknown variables are estimated in each iteration. A block size of one is equivalent to

Table S1: The parameters of the motif finder.

Parameter	Description
Motif length	Length of the motif to find
Initialization	Type of initialization: random or Baum-Welch
Block size	Number of sequences for variable estimation in each iteration
Flexibility	Number of top variables (determined by conditional probabilities) to draw new variable from
Termination interval	Interval for checking termination condition
Termination threshold	Minimum difference in data log-likelihood needed to justify a continuation of the training process.

the original Gibbs sampling algorithm.

1.2.3 Flexibility

In the second estimation step of a Gibbs sampling iteration, the conditional probabilities of each combination of motif start position and structure are calculated. These conditional probabilities reflect how well a combination of motif start and structure fits the current ssHMM. According to the distribution that is proportional to the conditional probabilities, the new motif start position and structure can be drawn.

Since the drawing of the two variables is a random process, it can happen that very unlikely and possibly unfavorable variables are drawn (i.e. variables that do not fit the current ssHMM well). For this reason, we introduce a flexibility parameter f. It determines that the new motif start and structure is drawn according to the distribution proportional to only the top f conditional probabilities. Consequently, a small f makes the Gibbs sampling greedier as we draw from only the most likely (i.e. best) variables.

While the Gibbs sampler is completely greedy for $f = 1$ and runs into the adjacent local optimum, a larger f promises a more flexible walk through the search space. The original Gibbs sampling approach of drawing from all variables can be chosen by setting $f = 0$.

1.2.4 Termination condition

Like for any iterative algorithm, we need to define a termination condition. It is desirable to finish the execution when the algorithm has converged on variables, i.e. an ssHMM, that maximize the posterior probability $P(ssHMM|data)$ of the ssHMM given the data. We can rearrange the posterior probability using Bayes' theorem:

$$P(ssHMM|data) = \frac{P(data|ssHMM) * P(ssHMM)}{P(data)} \tag{1.1}$$

$P(data)$ does not depend on the ssHMM and the prior $P(ssHMM)$ is unknown which is why we assume that all possible models are equally likely. Therefore, the likelihood $P(data|ssHMM)$ can be used as a proxy for the posterior $P(ssHMM|data)$. The execution shall terminate when the likelihood $P(data|ssHMM)$ converges.

Concretely, the likelihood is computed as the joined log-likelihood $log(P(sequences, structures|ssHMM))$. This computation has a considerable runtime for a large set of sequences. Therefore, we introduce an interval parameter i. Only every i iterations, the joined log-likelihood is computed and compared to its three most recent values.

The log-likelihood improvement made in every iteration decreases over time. There is a point from which the improvement does not satisfy the required computational time anymore. The parameter t lets the user define the minimum difference in log-likelihood needed to justify a continuation of the training process. If the joined log-likelihood does not improve by at least t compared to any of its three last values, execution is terminated.

1.3 Motif output of the ssHMM

The main purpose of this motif finder is the recovery of interpretable sequence-structure motifs from experimental RBP data. It is therefore necessary to produce a clear and intuitive visualization of the motif that has been found by training the models described in this section. HMMs are graphical models that lend themselves to a visualization as a graph. The states and emissions of an HMM can be depicted as nodes while the

transition and emission probabilities can be represented as weighted edges between the nodes.

As an example, Figure S2 shows the output of our motif finder after it has been trained on a CLIP-Seq dataset of the RBP DGCR8. It depicts the final state of the ssHMM. The states of the ssHMM, including the start and the end state, are represented by rectangular boxes that are arranged into rows and columns. The columns represent the motif positions while the rows represent the five structural contexts of RNA.

The emissions of the ssHMM are the four nucleotides A, C, G, and T. Sequence logos have been successfully used for years to intuitively visualize probability distributions over these nucleotides. We therefore chose them as a means to visualize the emission probabilities of each state directly in the state's box. The heights of the nucleotide letters in the sequence logos depend on two properties: Firstly, the relative height of a letter reflects its prevalence in this state, i.e. its emission probability. Secondly, the total height of the stack of four nucleotides is scaled according to the information content of the emission probabilities in this state. Consequently, states with a high information content stand out because of their size while those with more uniform emission probabilities are smaller. Different colors for the four nucleotides make it very easy to immediately grasp the sequence motif defined by the ssHMM. For this concrete RBP, the sequence motif is UGGAA.

The transition probabilities between the states are visualized as arrows. The thicker an arrow between two states, the more likely is a transition between the two. The most important transitions originate from the start state because they determine in which structural context the motif starts. Often, the motif remains in one particular context for its entire length (as can be seen for the hairpin and stem context in Fig. S2 but not for the internal loop context). To reduce clutter and increase clarity, unlikely transitions with a probability of less than 0.05 are completely omitted. This can be observed, for example, between the start state and both multiloop and exterior loop contexts. The RBP depicted in Fig. S2 preferably binds stem regions of RNA but to a lesser extent also hairpin loops and internal loops.

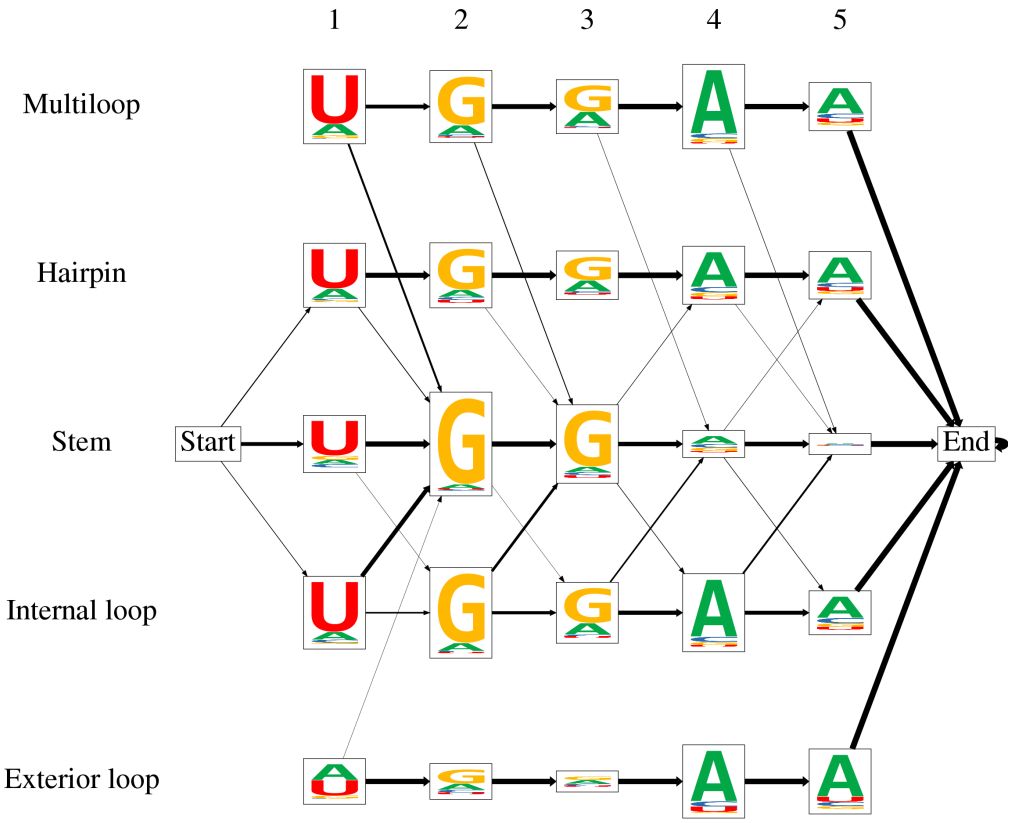


Figure S2: Output of the motif finder after training on experimental data from the RBP DGCR8. The figure visualizes the final state of the ssHMM. Each of its states, including the start and the end state, is represented by a box. The states are arranged into rows and columns where the columns represent the motif positions and the rows represent the five structural contexts of RNA. Each state's emission probabilities are visualized as a sequence logo. The relative heights of the nucleotides in a stack represent their emission probabilities. The total height of all four nucleotides is scaled according to the information content of the state's emission distribution. The transition probabilities between the states are visualized as arrows. The thicker an arrow between two states, the more likely is a transition between the two. To reduce clutter, no arrow is shown if the transition probability is less than 0.05.

2 Evaluation on synthetic datasets

2.1 Details on the synthetic datasets

Table S2: Properties of the 24 different synthetic datasets.

ID	Background seq.	IC per position	Structural context
H.A	uniformly random	1.0	100% hairpin
H.B	uniformly random	1.0	50% hairpin
H.C	uniformly random	1.0	10% hairpin
H.D	uniformly random	0.5	100% hairpin
H.E	uniformly random	0.5	50% hairpin
H.F	uniformly random	0.5	10% hairpin
H.G	3'UTR	1.0	100% hairpin
H.H	3'UTR	1.0	50% hairpin
H.I	3'UTR	1.0	10% hairpin
H.K	3'UTR	0.5	100% hairpin
H.L	3'UTR	0.5	50% hairpin
H.M	3'UTR	0.5	10% hairpin
S.A	uniformly random	1.0	100% stem
S.B	uniformly random	1.0	50% stem
S.C	uniformly random	1.0	10% stem
S.D	uniformly random	0.5	100% stem
S.E	uniformly random	0.5	50% stem
S.F	uniformly random	0.5	10% stem
S.G	3'UTR	1.0	100% stem
S.H	3'UTR	1.0	50% stem
S.I	3'UTR	1.0	10% stem
S.K	3'UTR	0.5	100% stem
S.L	3'UTR	0.5	50% stem
S.M	3'UTR	0.5	10% stem

2.2 Parameter optimization on synthetic datasets

Before commencing the benchmarks, we explored the influence of the different program parameters on the motif finder's performance. For all these tests, we set termination interval = 100 and termination threshold = 10. These two parameters have only a minor influence on the motif finder performance as they only determine when algorithm execution stops. Therefore, we did not perform a systematic search for their optimal values. A termination interval of 100, however, constitutes a good compromise between lower values with a higher computation overhead and higher values which could result in many unnecessary iterations being made after convergence.

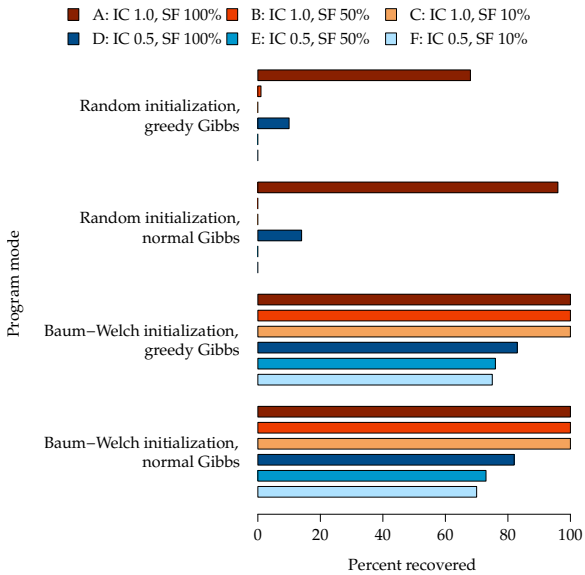


Figure S3: Recovery rates on synthetic datasets with random background. Shown is the percentage of successfully recovered motifs from synthetic datasets H.A to H.F. Six different program configurations are compared.

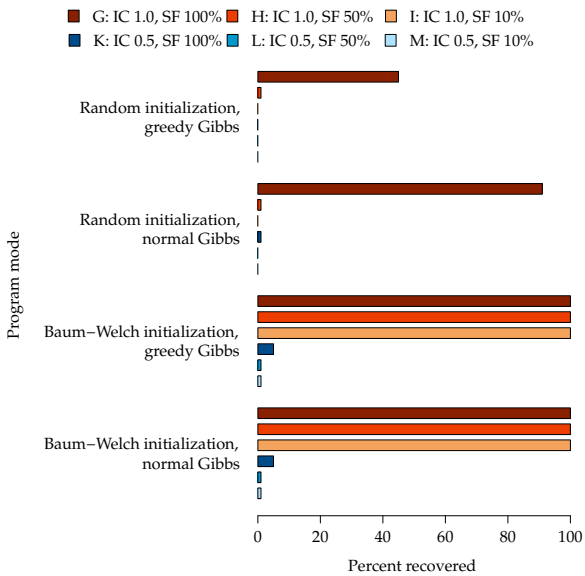


Figure S4: Recovery rates on synthetic datasets with random background. Shown is the percentage of successfully recovered motifs from synthetic datasets H.G to H.M. Six different program configurations are compared.

2.2.1 Initilization approach

Firstly, we evaluated the performance of four selected configurations of the motif finder on the 12 synthetic datasets with hairpin motifs (H.A - H.M). The four configurations differed by initialization approach and the flexibility parameter:

- Random initialization, flexibility 0 (original)
- Random initialization, flexibility 10 (greedy)
- Baum-Welch initialization, flexibility 0 (original)
- Baum-Welch initialization, flexibility 10 (greedy)

All 72 benchmarks were executed with motif length = 6, termination interval = 100, termination threshold = 10 and block size = 1. We performed three repetitions for each benchmark. The results clearly show that Baum-Welch initialization is superior to a random initialization (Figures S3 and S4).

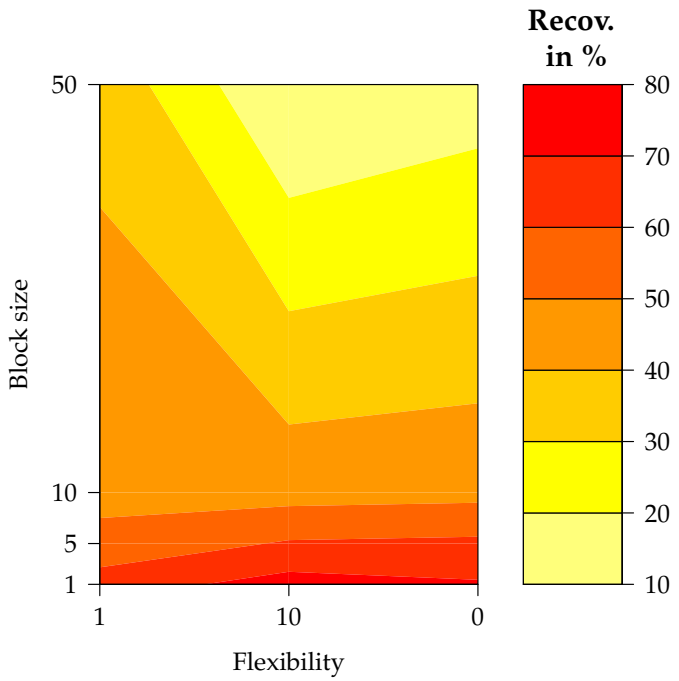


Figure S5: Recovery rates for different configurations of block size and flexibility. Color represents the percentage of successfully recovered motifs from synthetic dataset H.F. The warmer the color, the better the recovery rate.

2.2.2 Block size and flexibility

Secondly, we performed a systematic evaluation of different values for the block size and flexibility parameters. For that purpose, we used the synthetic dataset H.F (information content 0.5, hairpin fraction 10%) because it is the potentially most challenging among those with random background sequences. Since it contains motifs of length 6, we fixed the motif length parameter accordingly. As initialization, we chose the Baum-Welch variant as it generally yields superior results.

We ran the motif finder with all combinations of block sizes 1, 5, 10, and 50 and flexibilities 0 (original Gibbs), 10 (slightly greedy), and 1 (completely greedy). Each combination was executed with three repetitions.

The results (see Fig. S5) show that the motif finder's performance degrades with increasing block size. The original Gibbs sampling approach (block size = 1) that re-estimates the variables of exactly one sequence in each iteration performs best. No clear trend is visible, however, for the flexibility parameter. For small block sizes, the flexibilities 0 and 10 perform similarly well and substantially better than the greedy flexibility 1.

2.3 Program parameters for evaluation on synthetic datasets

All three analyzed tools as well as our method have multiple parameters that heavily influence their performance. Below, we describe how we chose the best parameters for each tool. Tables S3 to S6 list these parameters.

2.3.1 MEMERIS

Table S3: Chosen parameters for execution of MEMERIS

Parameter	Explanation	Value
-pi	Pseudocount to flatten structure prior	0, 1, 100
-w	Motif length	6
-bfile	Background distribution of nucleotides	uniform
-mod	Distribution of motif sites	one motif occurrence per sequence

MEMERIS inherits most parameters from *MEME* and adds one of its own: *pi*, which specifies the pseudocount that is used to flatten the prior probability distribution of the motif starts. In essence, it defines how strongly *MEMERIS* prefers motifs in a single-stranded context. The lower its value, the more does it prefer single-stranded motifs. A very high pseudocount of 10 or higher results in a *MEME*-like behavior which ignores structure completely. We ran *MEMERIS* with three different values for *pi*: 0, 1, and 100.

Besides that, we left most parameters at their default values. Motif length was naturally set to 6. A uniform background distribution was assumed because the sequences are too short to reliably estimate the background distribution of the characters from them. Lastly, the distribution of motif sites was set to OOPS which means that *MEMERIS* expects exactly one motif occurrence per sequence.

2.3.2 RNAcontext

RNAcontext has two important parameters. w specifies the range of motif lengths. A range of 4-7, for instance, means that the algorithm searches for motifs starting from length 4 until length 7. *RNAcontext*'s initialization procedure uses previously learned models for smaller motif lengths to initialize longer motif lengths. Therefore, they suggest to use, e.g. 4-6 instead of 6-6 when looking for motifs of length 6. We found,

Table S4: Chosen parameters for execution of RNAcontext

Parameter	Explanation	Value
-w	Motif length range	6-6
-s	Number of initializations	5

however, that w set to 6-6 performed much better than 4-6 which is why we use that setting. Depending on the initialization, *RNAcontext* can generate different results. The parameter *s* defines the number of different initializations that are tried to obtain the best result. For this parameter, we use the default value 5.

2.3.3 GraphProt

Table S5: Optimized parameters for GraphProt

Dataset	Epochs	Lambda	R	D	Bitsize	Abstraction
H.A	50	0.0001	1	5	14	3
H.B	50	0.0001	1	4	14	3
H.C	10	0.0001	2	4	14	3
H.D	40	0.0001	1	3	14	3
H.E	40	0.001	1	4	14	3
H.F	10	0.0001	4	4	14	3
H.G	50	0.0001	1	4	14	3
H.H	50	0.0001	2	4	14	3
H.I	40	0.0001	2	4	14	3
H.K	10	0.0001	1	2	14	3
H.L	50	0.001	1	3	14	3
H.M	50	0.0001	1	3	14	3

GraphProt has eight parameters which can be optimized in a dedicated parameter optimization step (program option -ls). We ran parameter optimization on the first sequence set of each dataset and used these optimized parameters for the entire dataset.

2.3.4 Our motif finder

Table S6: Optimized parameters of our motif finder (ssHMM)

Parameter	Value
Motif length	6
Initialization	Baum-Welch
Block size	1
Flexibility	1
Termination interval	100
Termination threshold	10

Our motif finder has six different parameters. See Sec. 2.2 for how we determined the best-performing values for these parameters.

2.4 Accurate recovery of structure preferences on synthetic datasets

The unique topology of our model and the incorporation of the full spectrum of structural RNA contexts allows ssHMM to recover structure motifs beside sequence motifs. To evaluate its ability to accurately detect the structural context of a binding site, we analyzed the synthetic datasets with hairpin motif. Figure S6 shows that the hairpin fractions recovered by ssHMM closely reflect the estimated hairpin fractions of the synthetic datasets. There is a striking Pearson correlation of 0.91 between the recovered and the estimated hairpin fraction. This confirms that ssHMM is able to recover both accurate sequence and structure motifs.

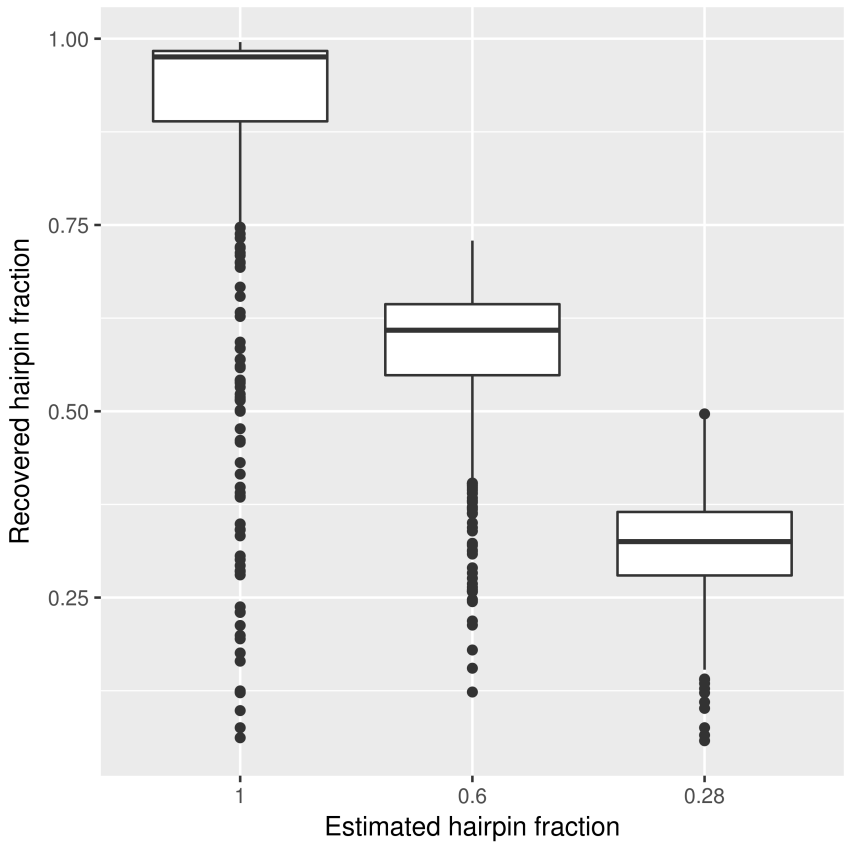


Figure S6: **Hairpin fractions recovered by ssHMM accurately reflect estimated hairpin fractions of the synthetic datasets** Boxplot of the recovered hairpin preference in the ssHMM models trained on synthetic datasets H.A to H.I. The results are grouped by the estimated hairpin fraction of the datasets (H.A, H.D and H.G: 100%; H.B, H.E and H.H: 60%; H.C, H.F and H.I: 28%). The recovered hairpin preference is defined as the transition probability between the start state and the H1 state in the trained ssHMM. Datasets H.K to H.M were excluded because of their low motif recovery rates (see Fig. 5 in the main text).

3 Evaluation on CLIP-Seq datasets

3.1 Full list of CLIP-Seq datasets

Table S7: Full list of CLIP-Seq datasets used in our analyses.

Protein	Data source	Genome	Cell line	Protocol	Ref.
Ago1/2/3/4	doRiNA	hg19	HEK293	PAR-CLIP	[2]
Ago2	doRiNA	hg19	HEK293	HITS-CLIP	[3]
Ago2	doRiNA	hg19	HEK293	PAR-CLIP	[3]
CAPRIN1	doRiNA	hg19	HEK293	PAR-CLIP	[4]
DGCR8	doRiNA	hg19	HEK293	PAR-CLIP	[5]
DICER	GEO	hg19	HEK293	PAR-CLIP	[6]
EIF4A3	doRiNA	hg19	HeLa	HITSCLIP	[7]
EWS	doRiNA	hg19	HEK293	PARCLIP	[8]
EZH2	GEO	mm9	ESC	PAR-CLIP	[9]
FMRP	doRiNA	hg19	HEK293	PAR-CLIP	[10]
FXR2	doRiNA	hg19	HEK293	PAR-CLIP	[10]
HuR	doRiNA	hg19	HEK293	HITS-CLIP	[3]
HuR	doRiNA	hg19	HEK293	PAR-CLIP	[3]
IGF2BP1/2/3	doRiNA	hg19	HEK293	PAR-CLIP	[2]
LIN28B	doRiNA	hg19	HEK293	PAR-CLIP	[11]
MOV10	doRiNA	hg19	HEK293	PAR-CLIP	[12]
Nova	Suppl. material	mm9		HITS-CLIP	[13]
PTBP1	GEO	hg18	HeLa	HITS-CLIP	[14]
PUM2	doRiNA	hg19	HEK293	PAR-CLIP	[2]
QKI	doRiNA	hg19	HEK293	PAR-CLIP	[2]
SRSF1	doRiNA	hg19	HEK293	HITS-CLIP	[15]
TAF2N	doriNA	hg19	HEK293	PAR-CLIP	[8]
TIA1	doRiNA	hg19	HeLa	iCLIP	[16]
YY1	GEO	mm9		HITS-CLIP	[17]
ZC3H7B	doRiNA	hg19	HEK293	PARCLIP	[4]

3.2 Length distribution and its influence on structure prediction

The CLIP-Seq datasets used in our analyses possess different length properties. Each dataset consists of thousands of binding sites determined by a CLIP-Seq experiment of the RBP under investigation. Figure S7 plots the length distribution of the binding sites for the five proteins shown in Table 2 and discussed in the Results section of the manuscript. While DICER, DGCR8, NOVA, and QKI binding sites are relatively short with median lengths in bps of 26, 31, 56 and 32, respectively, YY1 binding sites are much longer with a median of 164 bps.

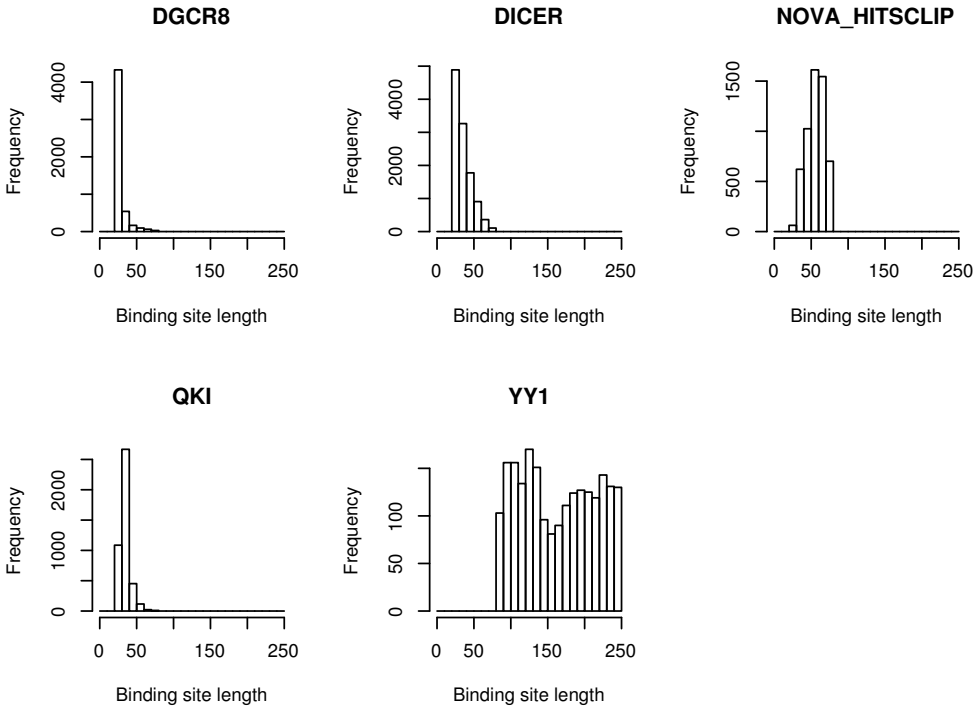


Figure S7: **Length distribution of selected CLIP-Seq datasets** The histograms show the length distributions of RBP-RNA binding sites as determined with the CLIP-Seq protocol. While DICER, DGCR8, NOVA, and QKI binding sites are short with median lengths of 26, 31, 56 and 32, respectively, YY1 binding sites are much longer with a median of 164.

To analyze the impact that sequence length has on the structure folding, we performed the following experiment: We elongated the binding sites for the five RBPs DICER, DGCR8, NOVA, QKI and YY1 by n base pairs up- and downstream where $n \in 10, 20, 40, 60$. Then, we performed structure prediction using the tool *RNAshapes* on all four classes of elongated binding site sequences. Figure S8 shows how the proportions of predicted structural contexts changed due to the elongation of binding sites.

One can see that a large fraction of base pairs in the only slightly elongated binding sites are predicted to be unpaired. This makes sense because of two reasons: Firstly, very short sequences are often unable to form secondary structures which leads to a larger number of unpaired bases. Secondly, secondary structure prediction is sensitive to the exact cutpoints of a nucleotide sequence. The two ends of an input nucleotide sequence only rarely match up exactly. More often, one or both of the ends are left unpaired in a 3' or 5' overhang even though the ends would find binding partners if the sequence was longer. Cosequently, structure prediction tools almost always predict unpaired bases at one or both ends of a nucleotide sequence. If the sequence length is increased, the proportion of these unpaired bases in relation to sequence length decreases.

Conversely, the other structural contexts become more prevalent with increasing elongation span and sequence length. This is particularly true for stem and multi-loop contexts whose proportions are monotonically increasing. Differences are most striking between an extension of 10 nt and 20 nt. With increasing elongation span, the differences become more subtle. Extending the binding sites by 20 nt as we did seems like a

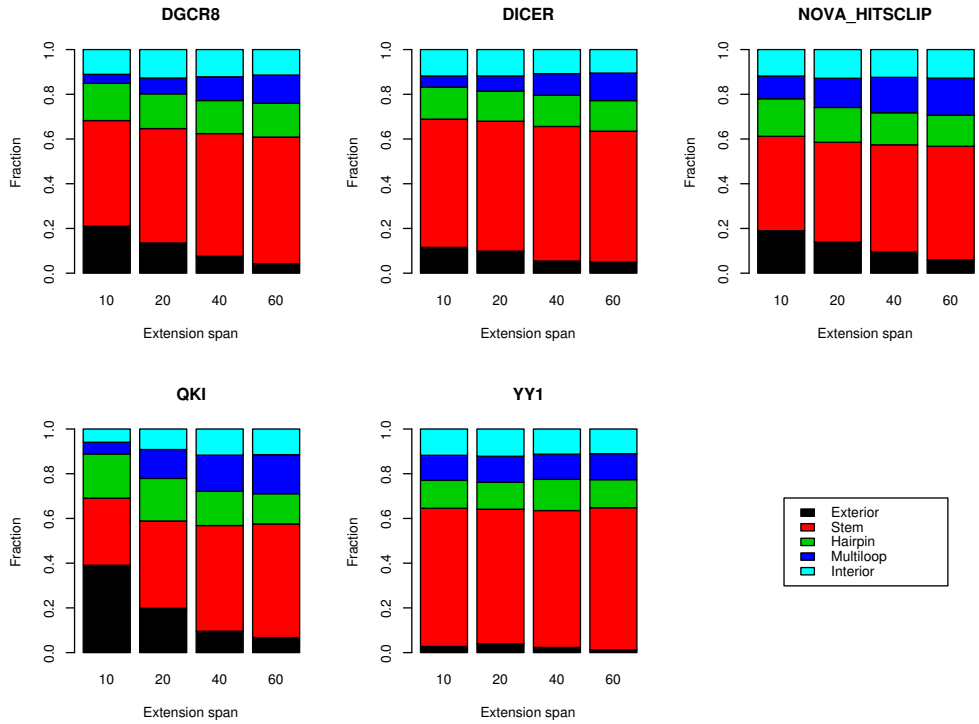
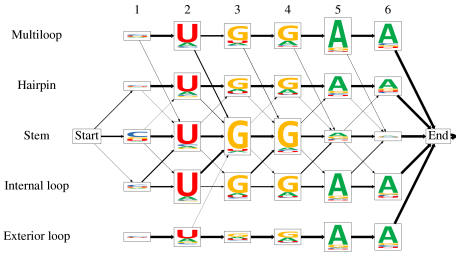


Figure S8: **Predicted structural contexts depending on extension of original binding sites (RNAshapes)** The original RBP binding sites as determined by CLIP-Seq were extended by 10, 20, 40, or 60 nt on each side (x-axis). The barplot visualizes how the predicted secondary structures changed as a consequence. The only slightly extended (extension span = 10) binding sites of DGCR8, DICER, NOVA and QKI contain many unpaired bases (exterior context). With increasing extension span, the stem and multi-loop contexts become more prevalent.

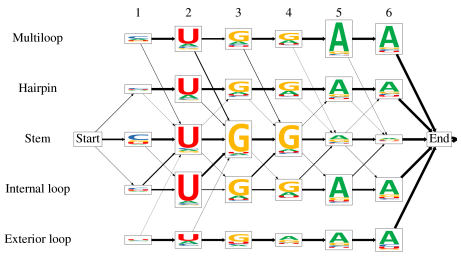
good choice but the users of our tool can customize the elongation span with a program parameter.

The YY1 protein is a special case again. Due to its a priori long binding sites, elongation does not have a large effect. Consequently, the proportions between predicted structural contexts do not change substantially (see Figure S8).

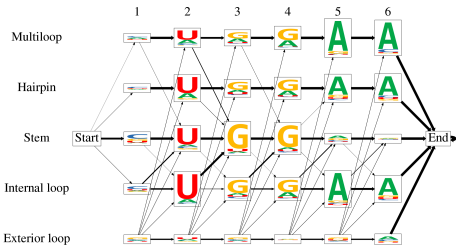
In a next step, we assessed the impact that elongating structure prediction input (as described above) had on the sequence-structure motifs recovered by ssHMM. Figures S9 to S13 show the trained sequence-structure motifs for the five proteins. Most importantly, the motif graphs (e.g. the arrows indicating the most probable transitions between different structures) remain overall conserved even though the predicted structures varied (see Fig. S8). For example, the strong preference of DGCR8 (Fig. S9) for the UGG motif in a stem context is always picked up, no matter what the length of the sequence subject to folding. Similar considerations hold for the other four analyzed proteins. Also the YY1 motif is robust to variations in sequence length. However, one notices the following things: 1) For QKI and NOVA, increasing the length of the sequence increases the chance to find the multiloop motif, as expected; 2) For QKI with an elongation by 10 nt, ssHMM picks up a probably unreliable exterior loop preference. This is probably due to its very short binding sites and can be fixed by elongation of



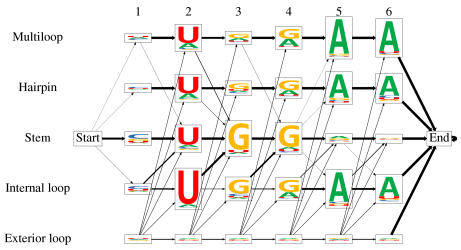
(a) extended by 10bp on each side



(b) extended by 20bp on each side

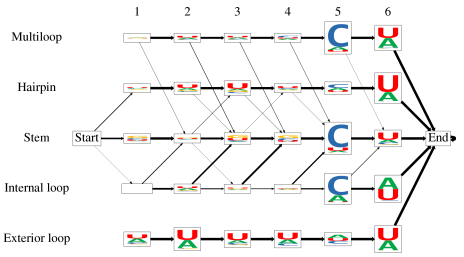


(c) extended by 40bp on each side

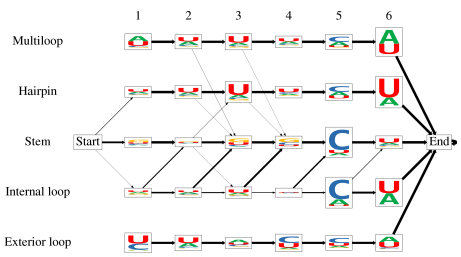


(d) extended by 60bp on each side

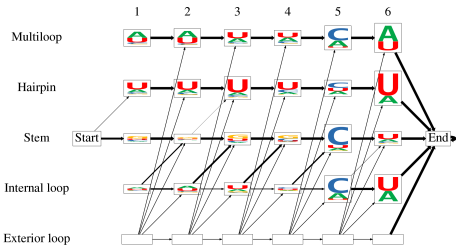
Figure S9: **DGCR8 motifs recovered by ssHMM with different extension spans** Original binding sites of DGCR8 were elongated by 10, 20, 40 and 60 nt up- and downstream prior to secondary structure prediction with *RNAshapes*. Then, the resulting structure sequences as well as nucleotide sequences of the original binding sites were used to train ssHMM. Shown are the trained sequence-structure motifs.



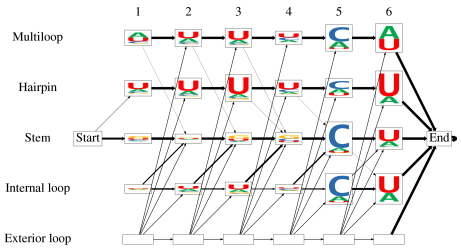
(a) extended by 10bp on each side



(b) extended by 20bp on each side

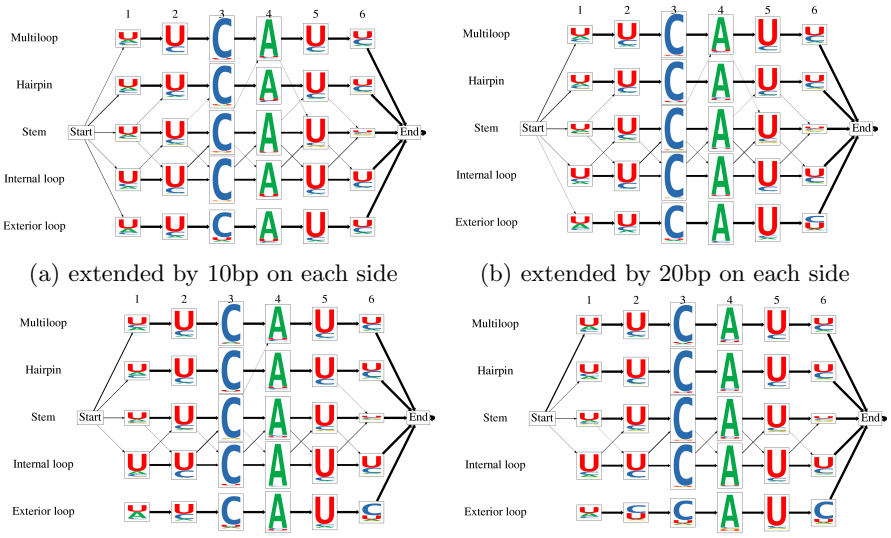


(c) extended by 40bp on each side



(d) extended by 60bp on each side

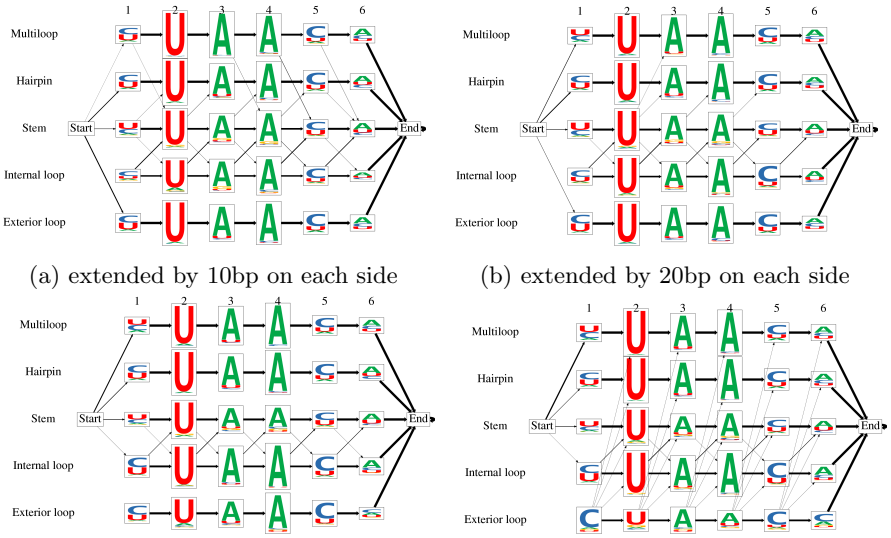
Figure S10: **DICER motifs recovered by ssHMM with different extension spans** Original binding sites of DICER were elongated by 10, 20, 40 and 60 nt up- and downstream prior to secondary structure prediction with *RNAshapes*. Then, the resulting structure sequences as well as nucleotide sequences of the original binding sites were used to train ssHMM. Shown are the trained sequence-structure motifs.



(a) extended by 10bp on each side (b) extended by 20bp on each side

(c) extended by 40bp on each side (d) extended by 60bp on each side

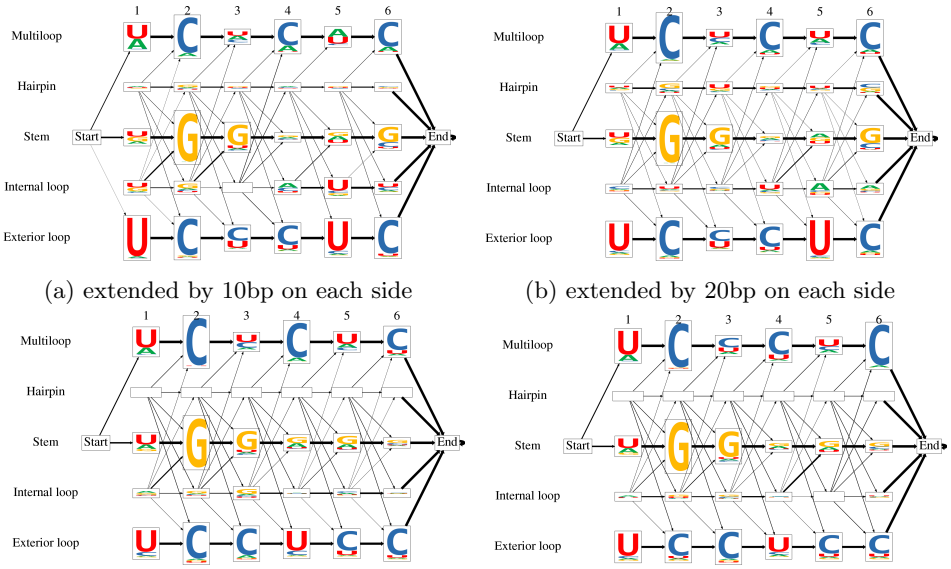
Figure S11: **NOVA motifs recovered by ssHMM with different extension spans** Original binding sites of NOVA were elongated by 10, 20, 40 and 60 nt up- and downstream prior to secondary structure prediction with *RNAshapes*. Then, the resulting structure sequences as well as nucleotide sequences of the original binding sites were used to train ssHMM. Shown are the trained sequence-structure motifs.



(a) extended by 10bp on each side (b) extended by 20bp on each side

(c) extended by 40bp on each side (d) extended by 60bp on each side

Figure S12: **QKI motifs recovered by ssHMM with different extension spans** Original binding sites of QKI were elongated by 10, 20, 40 and 60 nt up- and downstream prior to secondary structure prediction with *RNAshapes*. Then, the resulting structure sequences as well as nucleotide sequences of the original binding sites were used to train ssHMM. Shown are the trained sequence-structure motifs.



(a) extended by 10bp on each side

(b) extended by 20bp on each side

(c) extended by 40bp on each side

(d) extended by 60bp on each side

Figure S13: **YY1 motifs recovered by ssHMM with different extension spans** Original binding sites of YY1 were elongated by 10, 20, 40 and 60 nt up- and downstream prior to secondary structure prediction with *RNAshapes*. Then, the resulting structure sequences as well as nucleotide sequences of the original binding sites were used to train ssHMM. Shown are the trained sequence-structure motifs.

20 nt and more. All in all, these results show that ssHMM is generally robust to length variations of the binding site sequences subject to RNA structure folding.

3.3 Classification performance comparison

As described in Material and Methods, we compared the classification performance of ssHMM and three different setting of *MEMERIS* (pi=0, pi=1 and pi=100) on 23 CLIP-Seq datasets. Two datasets, SFRS1 and DICER, were omitted from the analysis because *MEMERIS* needed more than a full week to process them. The classification accuracy of both tools was evaluated using the Area under the Precision-Recall curve (AUCPR).

Table S8 lists the AUCPR for *MEMERIS* and ssHMM on the 23 protein datasets. In all settings, ssHMM outperformed *MEMERIS* on at least 15 out of 23 datasets, while *MEMERIS* outperformed ssHMM only on 7 dataset. On average, the increases in AUCPR of ssHMM over *MEMERIS* were considerably larger than the decreases, with several gains higher than 10%. These results demonstrate that the full sequence-structure model of ssHMM yields a substantial benefit over a sequence-only model (*MEMERIS* with pi=100). The superiority of ssHMM over *MEMERIS* is also consistent over the other two *MEMERIS* settings. For the majority of proteins, the structural preference incorporated into the model of ssHMM helps to distinguish binding from non-binding sites. However, there are a number of datasets (particularly PTBP1, TAF2N, IGF2BP123, EZH2, FMRP and AGO2 (PARCLIP)) for which *MEMERIS* performs better than ssHMM.

For precision-recall curves of all four tools (including *RNAcontext* and *GraphProt*) refer to Additional File 4.

Table S8: ssHMM substantially outperforms *MEMERIS* on the majority of our CLIP-Seq datasets. The table lists the area under the Precision-Recall curve for ssHMM and three different settings of *MEMERIS* ($pi = 0$, $pi = 1$ and $pi = 100$). Because ssHMM employs Gibbs sampling and results may vary from run to run, the 5th column shows the mean and standard deviation from 3 independent executions of ssHMM. The last three columns visualize the differences between ssHMM and *MEMERIS* (with parameter $pi = 0$, $pi = 1$ and $pi = 100$, respectively).

Protein	MEMERIS (pi=0)	MEMERIS (pi=1)	MEMERIS (pi=100)	ssHMM	ssHMM – MEMERIS (pi=0)	ssHMM – MEMERIS (pi=1)	ssHMM – MEMERIS (pi=100)
TIA1	69.4%	63.5%	64.8%	75.4% ± 0.8%	6.0%	12.0%	10.7%
AGO1234	50.6%	47.8%	47.6%	55.7% ± 0.5%	5.1%	7.9%	8.1%
PUM2	63.3%	69.3%	71.5%	77.0% ± 0.6%	13.7%	7.8%	5.5%
QKI	56.1%	65.1%	69.8%	72.2% ± 0.7%	16.1%	7.1%	2.4%
MOV10	52.0%	50.7%	51.7%	56.4% ± 0.3%	4.5%	5.7%	4.7%
HuR_HITSCLIP	66.9%	65.1%	65.2%	70.5% ± 0.4%	3.7%	5.5%	5.3%
EIF4A3	54.3%	59.8%	61.7%	64.3% ± 0.7%	10.0%	4.5%	2.6%
LIN28B	49.8%	49.8%	49.9%	53.4% ± 0.5%	3.6%	3.6%	3.5%
AGO2_HITSCLIP	46.1%	49.5%	52.1%	53.0% ± 1.0%	6.9%	3.5%	0.9%
NOVA_HITSCLIP	61.2%	70.6%	71.5%	73.3% ± 0.2%	12.1%	2.7%	1.8%
YY1	78.4%	80.6%	82.4%	82.4% ± 1.9%	4.1%	1.9%	0.1%
EWSR1	60.8%	58.7%	59.3%	60.1% ± 0.2%	-0.7%	1.4%	0.8%
ZC3H7B	51.8%	51.7%	50.1%	52.6% ± 0.5%	0.7%	0.9%	2.5%
CAPRIN1	52.0%	52.4%	52.3%	52.9% ± 0.6%	0.9%	0.6%	0.7%
HuR_PARCLIP	75.2%	73.3%	72.5%	73.5% ± 0.8%	-1.7%	0.2%	1.0%
FXR2	53.3%	54.8%	55.5%	54.8% ± 0.8%	1.5%	0.0%	-0.7%
DGCR8	66.8%	67.8%	67.2%	67.7% ± 0.1%	0.9%	-0.1%	0.5%
AGO2_PARCLIP	57.1%	56.2%	58.3%	54.7% ± 0.5%	-2.3%	-1.5%	-3.5%
FMRP	62.7%	64.1%	63.9%	61.3% ± 0.2%	-1.3%	-2.7%	-2.6%
EZH2	66.1%	66.9%	67.2%	63.9% ± 0.9%	-2.2%	-3.0%	-3.3%
IGF2BP123	61.0%	61.0%	60.8%	56.7% ± 0.5%	-4.3%	-4.3%	-4.1%
TAF2N	63.8%	63.3%	63.8%	58.7% ± 0.3%	-5.1%	-4.6%	-5.1%

3.4 Classification performance with only best shape

The structure prediction tools that we use compute several highly probable secondary structure conformations for each RNA nucleotide sequence. By default, ssHMM samples over all of them to obtain the best sequence-structure motif. To elucidate whether sampling over all structures is superior to sampling over the optimal one only, we performed a comprehensive comparison and executed ssHMM with both sampling variants on all 25 CLIP-Seq datasets. Because of the non-deterministic nature of the Gibbs sampler, we ran ssHMM three times for each setting. Figure S14 shows using the example of YY1 that training on all shapes leads to a model that is better fit to the training data. When using only the optimal shape for training, the likelihood of the data given the trained model is substantially lower.

We also analyzed whether the sampling variant had any influence on the classification performance. Similarly to the classification setting that is described in the manuscript, we analyzed the Areas under the precision recall curve for all 25 CLIP-Seq datasets (see Fig. S15). The results were ambiguous and differed a lot between the proteins. To filter for significant differences, we performed Welch's t-test on the two sets of AUCPRs for each protein. After applying Benjamini-Hochberg correction to the p-values, only the differences for NOVA and SFRS1 came out to be significant (p-values 0.04 for both). In both cases, sampling over all shapes produced a higher AUCPR than using only the optimal shape.

We conclude that sampling over all shapes leads to a better model fit but does not have a major impact on classification performance. Sampling over all shapes is the default setting of ssHMM but the user can decide to use only the optimal shape with the command-line parameter *only_best_shape* in *train_seqstructhmm*.

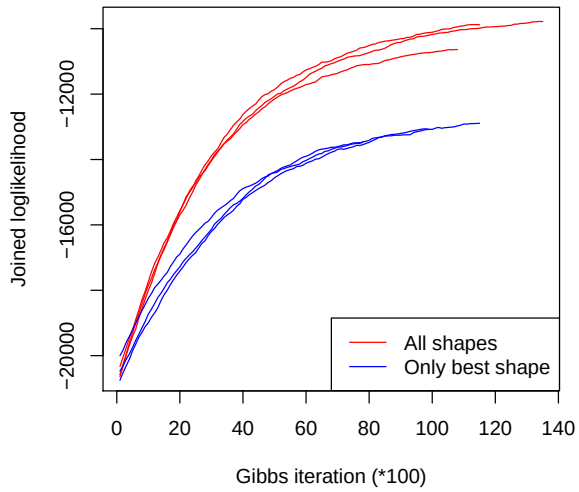


Figure S14: **Sampling over all shapes produces a model with better fit** The joined log-likelihood of the training data (YY1 CLIP-Seq dataset) given the current model is plotted over training time (in iterations). The three runs of ssHMM using all shapes reach a substantially better fit than the three runs of ssHMM using only the optimal shape.

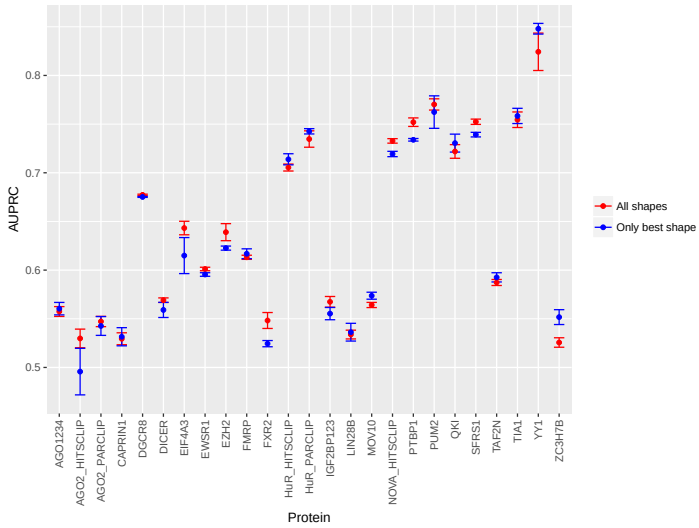


Figure S15: **Areas under the precision recall curve (AUCPR) for sampling over all shapes versus using only the optimal shape** For all 25 CLIP-Seq datasets, points and error bars visualize the mean and standard deviation of the AUCPR, respectively. The values were computed from three independent training runs of ssHMM for each setting.

3.5 Program parameters for analysis of CLIP-Seq datasets

Tables S9-S11 list all parameters of *MEMERIS*, *RNAcontext*, and ssHMM that were used in the analysis of the CLIP-Seq datasets. For *MEMERIS* and ssHMM, different configurations were used in the classification setting versus the motif retrieval for Additional File 2. For *GraphProt*, we used the default parameters.

Table S9: Chosen parameters for execution of MEMERIS

Parameter	**Explanation**	**Value for class.**	**Value for motif retr.**
-pi	Pseudocount to flatten structure prior	0, 1, 100	1
-w	Motif length	6	
-bfile	Background distribution of nucleotides	uniform	
-mod	Distribution of motif sites	one motif occurrence per sequence	

Table S10: Chosen parameters for execution of RNAcontext

Parameter	**Explanation**	**Value**
-w	Motif length range	6-6
-s	Number of initializations	5

Table S11: Optimized parameters of our motif finder (ssHMM)

Parameter	**Value for class.**	**Value for motif retr.**
Motif length	6	
Initialization	Baum-Welch	
Block size	1	
Flexibility	10	0, 10
Termination interval	100	
Termination threshold	10	5, 10

3.6 Fisher's exact test

Table S12 lists the adjusted p-values from Fisher's exact test on loglikelihoods of positive and negative test sequences (and structures). Initially, all 25 protein datasets were split into training and test data and a motif ssHMM was trained for each protein on the training data portion. Then, the trained models were used to compute loglikelihoods of the sequences from the test portion. To find the optimal classification cutoff on these loglikelihoods, we chose the cutoff with the lowest p-value as obtained by Fisher's exact test. Finally, the p-values of the optimal cutoffs for all proteins were adjusted using Benjamini & Hochberg correction [18]. All p-values (see Table S12) were below a significance threshold of 0.05 which demonstrates that ssHMM can distinguish between real binding sites and background sites.

Table S12: Adjusted p-values from Fisher's exact test on loglikelihoods of positive and negative test sequences.

Protein	Adjusted p-value
Ago1/2/3/4	1.24e-03
Ago2 (HITS-CLIP)	2.54e-03
Ago2 (PAR-CLIP)	4.98e-04
CAPRIN1	1.24e-02
DGCR8	<1e-16
DICER	4.78e-15
EIF4A3	<1e-16
EWSR1	<1e-16
EZH2	<1e-16
FMRP	<1e-16
FXR2	3.14e-05
HuR (HITS-CLIP)	<1e-16
HuR (PAR-CLIP)	<1e-16
IGF2BP1/2/3	3.43e-10
LIN28B	7.21e-03
MOV10	<1e-16
Nova	<1e-16
PTBP1	<1e-16
PUM2	<1e-16
QKI	<1e-16
SRSF1	<1e-16
TAF2N	<1e-16
TIA1	<1e-16
YY1	<1e-16
ZC3H7B	9.39e-05

3.7 Computation of motif information content

We measured the ability of ssHMM to retrieve informative motifs given a set of binding site sequences by computing the information content of the retrieved motif model. Three variants of the motif information content were computed on three different alphabets A:

- Information content of the sequence motif ($A = \{A, C, G, U\}$)
- Information content of the structural motif (if applicable, $A = \{E, I, S, H, M\}$)
- Information content of sequence and structure combined (if applicable, $A = \{A, C, G, U\} \times \{E, I, S, H, M\}$)

Depending on the underlying alphabet A, the information content of a binding motif position can range from 0 to $\log_2 |A|$. Consequently, the maximum information content per position of a nucleotide sequence motif is $\log_2 4 = 2$. The maximum information content per position of a structural motif with $|A| = 5$ is $\log_2 5 \approx 2.32$ and of a sequence-structure motif it is $\log_2(4 * 5) \approx 4.32$. To calculate the information content of a motif position, the frequency f_s of each symbol $s \in A$ is required (e.g. from a PPM) [19]. Then, the Shannon entropy H and small-sample correction e of that position are defined as

$$H = -\sum_{s \in A} f_s * log_2(f_s) \tag{3.1}$$

and

$$e = \frac{1}{ln2} * \frac{|A| - 1}{2n}. \tag{3.2}$$

Finally, the information content of that position can be computed as [19]

$$R = log_2(|A|) - (H + e). \tag{3.3}$$

We distinguish two ways of calculating the information content: 1) Information content from the top sequences, and 2) Information content directly from the model.

3.7.1 Information content from the top sequences

GraphProt computes sequence and structure logos from the 1,000 highest-scoring k-mer nucleotide sequences and structure profiles. From these 1,000 sequences, the frequency f_s of each symbol $s \in A$ in each motif position can be calculated by counting. These frequencies serve as input to compute the information content as described above.

To ensure comparability, we followed a similar procedure to obtain the information content of sequence motifs produced by our motif finder. We calculated the information content on the 1,000 sequences with the best score given our trained motif model. First, the estimated motif start and best structure of each of these sequences was obtained from the trained model. They pointed directly to each sequence's motif occurrence which made it possible to align the 1,000 motif occurrences (sequence and structure separately). Then, the frequencies were counted and the three different types of information content (sequence, structure and combined sequence-structure) were computed.

3.7.2 Information content directly from the model

RNAcontext directly produces a PPM representing its inner model. To compare against *RNAcontext*, we obtained a second set of information contents directly from the ssHMM. First, a sequence motif (or PPM) was extracted from the trained ssHMM by averaging over all paths in the model. The average was weighted by the transition probabilities so that more likely structural contexts have a bigger impact on the sequence motif than less likely contexts. Then, the information content was computed.

The information contents calculated from the top sequences are naturally larger than those directly from the model. The reason is that those sequences, which are most likely given the model, are more homogeneous than the set of all sequences. Consequently, the resulting motif is more clearly defined.

4 Comparison of runtime

To assess how the runtimes of *GraphProt*, *RNAcontext*, *MEMERIS* and the ssHMM progress with increasing input size, we took runtime measurements with *GNU time 1.7*, the Linux timekeeping tool. We measured the runtime (User time) of all four tools on ten different datasets containing 200, 400, ... and 2,000 RNA sequences. The datasets were sampled from the synthetic dataset H.D. For each tool and dataset, 3 independent measurements were taken and the mean of the three values was reported. All measurements were taken on a Linux machine with 8 cores running at 3.40GHz (Intel i7-3770) and 7.6 GB RAM.

We measured the total runtimes required to obtain motifs from sequence files. Thus, the runtimes include both structure prediction and sequence logo computation if these are separate from the training step. Figure S16 shows the results for ssHMM and *GraphProt* only. The results of all four tested tools are plotted in the Results section of the paper.

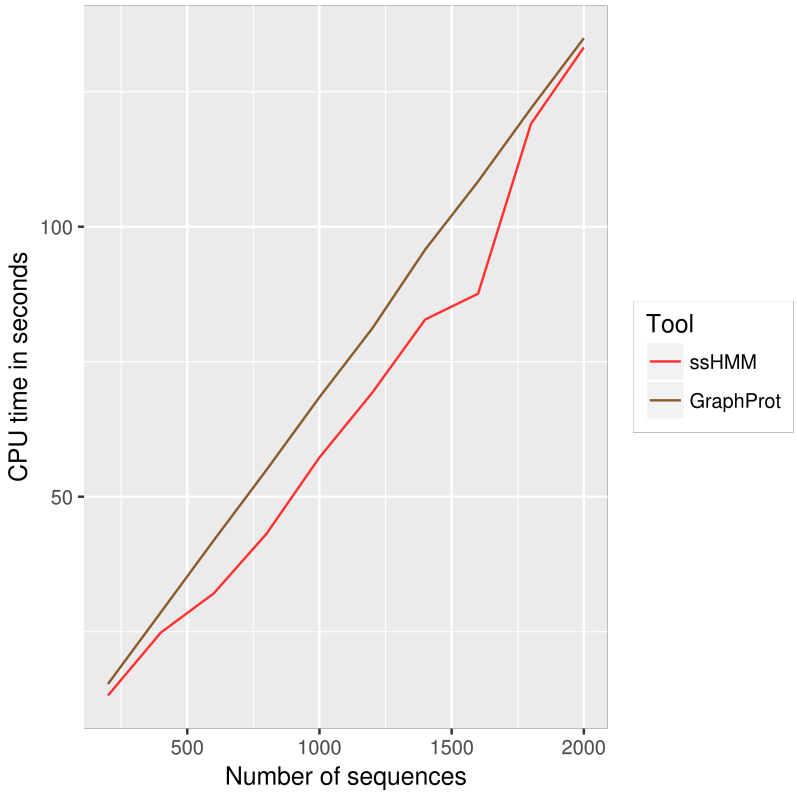


Figure S16: ssHMM scales approximately linearly on the input size. The CPU time in seconds (y-axis) is plotted against the number of input sequences (x-axis). Only the two fastest tools *GraphProt* and ssHMM are shown.

Bibliography

[1] Jensen, C.S., Kjrulff, U., Kong, A.: Blocking gibbs sampling in very large probabilistic expert systems. International Journal of Human-Computer Studies **42**(6), 647–666 (1995)

[2] Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G.S., Dewell, S., Zavolan, M., Tuschl, T.: Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. Cell **141**(1), 129–141 (2010)

[3] Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., Zavolan, M.: A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nature Methods **8**(7), 559–564 (2011)

[4] Baltz, A.G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., et al.: The mrna-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular Cell **46**(5), 674–690 (2012)

[5] Macias, S., Plass, M., Stajuda, A., Michlewski, G., Eyras, E., Cáceres, J.F.: Dgcr8 hits-clip reveals novel functions for the microprocessor. Nature Structural & Molecular Biology **19**(8), 760–766 (2012)

[6] Rybak-Wolf, A., Jens, M., Murakawa, Y., Herzog, M., Landthaler, M., Rajewsky, N.: A variety of dicer substrates in human and c. elegans. Cell **159**(5), 1153–1167 (2014)

[7] Saulière, J., Murigneux, V., Wang, Z., Marquenet, E., Barbosa, I., Le Tonquèze, O., Audic, Y., Paillard, L., Crollius, H.R., Le Hir, H.: CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. Nature structural & molecular biology **19**(11), 1124–1131 (2012)

[8] Hoell, J.I., Larsson, E., Runge, S., Nusbaum, J.D., Duggimpudi, S., Farazi, T.A., Hafner, M., Borkhardt, A., Sander, C., Tuschl, T.: RNA targets of wild-type and mutant FET family proteins. Nature Structural & Molecular Biology **18**(12), 1428–1431 (2011)

[9] Kaneko, S., Son, J., Shen, S.S., Reinberg, D., Bonasio, R.: Prc2 binds active promoters and contacts nascent rnas in embryonic stem cells. Nature Structural & Molecular Biology **20**(11), 1258–1264 (2013)

[10] Ascano, M., Mukherjee, N., Bandaru, P., Miller, J.B., Nusbaum, J.D., Corcoran, D.L., Langlois, C., Munschauer, M., Dewell, S., Hafner, M., Williams, Z., Ohler, U., Tuschl, T.: FMRP targets distinct mRNA sequence elements to regulate protein expression. Nature **492**(7429), 382–386 (2012)

[11] Hafner, M., Max, K.E., Bandaru, P., Morozov, P., Gerstberger, S., Brown, M., Molina, H., Tuschl, T.: Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition. Rna **19**(5), 613–626 (2013)

[12] Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F., Paro, R.: Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. Nucleic Acids Research **40**(20), 160–160 (2012)

[13] Zhang, C., Frias, M.A., Mele, A., Ruggiu, M., Eom, T., Marney, C.B., Wang, H., Licatalosi, D.D., Fak, J.J., Darnell, R.B.: Integrative Modeling Defines the Nova Splicing-Regulatory Network and Its Combinatorial Controls. Science **329**(5990), 439–443 (2010)

[14] Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D.L., Sun, H., Fu, X.-D., Zhang, Y.: Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. Molecular Cell **36**(6), 996–1006 (2009)

[15] Sanford, J.R., Wang, X., Mort, M., VanDuyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J., Liu, Y.: Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Research **19**(3), 381–394 (2009)

[16] Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N.M., Rot, G., Zupan, B., Curk, T., Ule, J.: iCLIP Predicts the Dual Splicing Effects of TIA-RNA Interactions. PLoS Biol **8**(10), 1000530 (2010)

[17] Sigova, A.A., Abraham, B.J., Ji, X., Molinie, B., Hannett, N.M., Guo, Y.E., Jangi, M., Giallourakis, C.C., Sharp, P.A., Young, R.A.: Transcription factor trapping by rna in gene regulatory elements. Science **350**(6263), 978–981 (2015)

[18] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289–300 (1995)

[19] Schneider, T.D., Stormo, G.D., Gold, L., Ehrenfeucht, A.: Information content of binding sites on nucleotide sequences. Journal of Molecular Biology **188**(3), 415–431 (1986)