**Supplementary Materials**

**Supplementary Materials and Methods**

**qBEAR notation**

To represent the motif model in a graphic way, we can report it in the Logo representation, with the software webLogo (1). The BEAR alphabet contains 85 characters, many of which are not alphanumeric and the yield in a Logo image is unfeasible, thus we defined a grouping for the structural characters, named quick BEAR (qBEAR). The mapping in table S1 shows the correspondence between BEAR and qBEAR characters (upper or lower case is equivalent for the latter).

|  |  |
| --- | --- |
| **BEAR notation** | **qBEAR notation** |
| abcde | Z – short stem |
| fghi | A – medium stem |
| = | Q – long stem |
| jklmnopqr | X – short loop |
| stuvwxyz | S – medium loop |
| ^ | W – long loop |
| !"#$%23456 | C – short internal loop |
| &'()7890 | D – medium internal loop |
| +> | E – long internal loop |
| [] | B – bulge |
| {} | G – bulge branch |
| : | T – branch |
| ABCDE | V – short stem branch |
| FGHI | F – medium stem branch |
| J | R – long stem branch |
| KLMNYZ~? | N – short internal loop branch |
| OPQRS\_/\ | H – medium internal loop branch |
| TUWYZ@ | Y – long internal loop branch |

**Table S1:** the qBEAR notation

**The BEAM motif model and algorithm**

BEAR is an alphabet for describing RNA secondary structure (38) replacing the classic dot-bracket notation with an 85-character alphabet distinguishing between different types of 2D elementary structures as well as their lengths. The purpose of this structural notation is dual. The expansion of the set of available characters while maintaining a string notation damps the problem of the small direct information provided by a 3-character alphabet: the dot-bracket could only distinguish between unpaired and paired nucleotides, but did not integrate information about bulges, internal loops or other structures in which an RNA is folded in. The second and most important aspect is the possibility to use alignments of curated structures to compute a substitution matrix of the BEAR characters in order to analyse secondary structures ensembles from an evolutionary point of view. We refer to this matrix as MBR (substitution Matrix of BEAR-encoded RNA secondary structures). The MBR allows us to compare structures not only at the topological level, but also at the evolutionary one.

We define a motif as a conserved (in the sense of the MBR) set of secondary structure elements present in a set of unaligned RNAs. The motif may be or not in any of the RNAs forming our subject ensemble but we restrict this presence to at most one occurrence per sequence, the so-called Zero or One Occurrence Per Sequence (ZOOPS) model. The choice of a ZOOPS model *in lieu* of a OOPS (exactly One Occurrence Per Sequence) is due to the nature of the data that we expect to analyse. Our experimental design is in fact targeted towards all those experiments that can cause variable amounts of false positives (such as post-processed CLIP-Seq data and similar assays) as well as smaller sets in which we expect noise of diverse nature and quantity.

As described in the Methods, a motif is defined as *M* BEAR-encoded sub-structures each having the same width, out of an initial dataset of *N* BEAR-encoded RNAs. The motif scoring function *BS* employs the MBR to evaluate the structural similarity of the structures. The choice of a ZOOPS model contributes to the system state space with about 2*N*possible states. This has the big drawback of making an exhaustive research unfeasible. BEAM follows a heuristically tweaked simulated annealing algorithm to look for the best score *BS*.Given *N* BEAR-encoded RNA secondary structures, a random initial state is drawn. From there, and for a fixed maximum number of steps (or until score convergence), a random perturbation (of the type *Add*, *Remove*, *Shift*, *Resize*, *Re-align*) is drawn from a uniform distribution. After every perturbation a new system state is reached and the score difference from the previous state is used to accept or reject the step, as for classical Metropolis rule depending on temperature and Boltzmann constant parameters.The “temperature” *T* regulates the amount of score variation considered significant and after every step the temperature is decreased in small steps up until *T=0*, ultimately resulting in a greedy algorithm. The Boltzmann constant *k* becomes an arbitrary constant needed to match the order of magnitude of score variations and is set accordingly.

In a simulated annealing, the “cooling” of the system (i.e. decreasing the parameter T linearly by a fixed amount) allows an initial freedom of exploring the conformational space, which becomes more restricted as T decreases, ideally focusing around the global maximum. Once convergence, or the maximum number of allowed iterations, is reached, a search through the input set is done with the created motif model, in order to catch RNAs that were possibly left out during the model creation.

We additionally added a feature to avoid falling in global *maxima* created by a significant amount of homogeneous noise, for example due to frequently recurring small secondary structures (moreover, it cannot be excluded that the input set contains two or more motifs characteristic of the input RNAs). Hence, we added an additional layer of computation with the insertion of the “masks”, which hide motifs identified in previous runs. In this way, prevalent motifs do not eclipse anymore the ones with fewer representatives, allowing the identification of motifs that have lower score but that can be more discriminative. To estimate this discriminatory ability, a p-value is reported representing the probability of having a score higher than the model mean score per sequence given the hypothesis that the background RNAs have the same Beam scores distribution if scored against the motif, more detail in the Supplementary Materials.

**Background distributions**

We built a set of background distributions from the Rfam families for the calculation of statistical parameters. In order to avoid possible confounding effects due to different characteristics of RNAs we divided the ensemble in groups of RNA molecules characterized by similar length and structure content (2), defined as the fraction of paired nucleotides in the RNA secondary structure (Figure S1). We divided the whole set in 15 bins (3 in length and 5 in structure content) each one containing approximately 2000 RNAs. This choice was made in order to guarantee a reasonable size for the background and keep a certain level of heterogeneity in every bin. These bins are from now on called L-S bins (as for Length-Structure bins).

**Figure S1**. L-S bins. Every point corresponds to an RNA from the seed set of Rfam. The separation between bins is made with equally numbered quantiles, in order to generate the most possible heterogeneous background for small datasets.

**Matthews Correlation Coefficient (MCC) estimation**

To establish the discriminative power of a found motif during our tests on Rfam, we measure the number of instances of the motif in the background dataset. Given a background dataset, the putative motif (represented by its Position Frequency Matrix, or PFM of BEAR characters) is scored against every sequence in the background and against all RNAs in the original input set, even those that were not included in the putative motif, by aligning every RNA in the background and input sets with the PFM using a sliding window to find the local ungapped alignment that provides the highest score, in terms of MBR. The lowest individual score of the RNAs composing the putative motif is considered the threshold to identify other motif instances. Hence, a RNA in the background or in the input dataset is considered a *Positive* if its score is higher than the lowest individual score of every sequence composing the motif, while it is considered a *Negative* otherwise. *False Positive (FP)* is then every *Positive* belonging to the background set. *True Positive (TP)* is every *Positive* from the input set (even those RNAs not ending up in the identified motif in the BEAM run, but that are nevertheless assumed to contain a motif instance). *True Negative (TN)* is every *Negative* from the background set. *False Negative (FN)* is every *Negative* belonging to the input set (i.e. those RNAs in the input set for which the PFM of the motif identified by BEAM is not able to identify a motif instance). Note that *FN* is actually condensing both a measure of how many input RNAs do not contain the motif as well as the possible amount of noise in the input. In other words, since there is no guarantee that each RNA in the input set contains a motif instance, we are knowingly underestimating the BEAM accuracy. *MCC* is calculated according to the following equation:



**Generation of artificial large datasets**

To demonstrate the capability of BEAM of working with CLIP-Seq like datasets (thousands of RNAs after filtering, removal of redundancy, etc.) we took a series of RNAs contained in the dataset GSE37114 in mouse. This is a large dataset (about 180k RNAs) of LIN28A interactors (44). Of all the RNAs, a subset of 10k was chosen until we had a “ground” set without any significant motif (MCC ~ 10-2). We then proceeded to the insertion of a known motif (from now on denoted as the gold structure) with this pipeline: (i) Fold the ground set of 10k sequences with RNAfold with default parameters; (ii) For every RNA, spot the “less affecting” zone of the fold where to insert a particular structure that BEAM will look for; (iii) Construct 10 different datasets at different density levels for each gold structure by inserting its sequence (i.e. the gold sequence) into the ground set and refolding the resulting sequence; (iv) Repeat until 100 sets of 10 datasets are built.

We defined the less affecting zone as the area of the fold that was more distant from every non-branching hairpin. In the BEAR encoding it translates in finding for every folded RNA the longest region of “:” characters and choosing the middle point of this stretch. This choice was made to reduce the possibility that the subsequent structure would be affected by flanking regions that would otherwise fold in other ways.

We automatically retrieved from Rfam curated structures (as defined in *Dataset preparation*) to be used as gold structures to be inserted in the ground set. In these tests we used single hairpin motifs, defined as non-branching hairpins satisfying the following regular expression:

ΣΥΣ(ΥΣ){0,3}Υ?ΛΥ?(ΣΥ){0,3}ΣΥΣ

where Σ is a series of adjacent nucleotides in a stem, Υ is a series of adjacent nucleotides in an internal loop or bulge, and Λ is a series of adjacent nucleotides in a loop. The gold structure was tested to have at least less than 90% of structure identity from all the other structures chosen in previous iterations.

The gold sequence was inserted in the ground set at 10 different levels of density, with a percentage of insertion varying from 10% to 100% of RNAs affected (e.g. in the 20% set, we had 0.2\*10000=2000 randomly selected ground set RNA primary sequences inserted with the gold sequence in the less affecting zone). The whole resulting datasets were then folded again *de novo*. In this way, we aim to minimize the folding inaccuracies introduced by state-of-the-art secondary structure predictors. By not including the structure directly and letting the underlying sequence fold independently from the expected gold structure, however in different nucleotides contexts, we aim to introduce a variability in the resulting structures that is more similar to real cases.

We repeated this process for a total of 100 automatically retrieved gold structuresinserted in the ground set, resulting in 1000 datasets, 100 for each level of insertion (from 10% to 100%), each of which simulated different level of noise in real datasets. The 100 gold structures are reported in the Supplementary Table 1.

BEAM was launched on all the datasets (with parameters –s 10000 –S 15000 –w 10 –W 100 –C 3 –M 1 –R 1). The background set was another subset of 10k RNA from LIN28A not containing any significant motif.

The subsequent analysis was carried as follows: The MCC of a run was computed by adding the prior knowledge on RNAs holding the insert and by classifying as: TP those RNAs in the dataset that host the insert and were found in the reported motif; FP those RNAs that were found in the reported motif, and that are both contained in the background set AND the dataset but without an insertion; FN those RNAs in the dataset which received the insert but were NOT found in the reported motif; TN those RNAs that were NOT found in the reported motif, and that are both contained in the background set AND the dataset but without an insertion

Moreover, by having the gold structure we could compute the distance of the reported motif from the reference by using a premetric defined as:

Here *PFM* is the motif model PFM of BEAR characters and gold is the BEAR string representing the gold structure. *W* is the width of the gold string of characters *goldi* and ω the width of the model PFM of frequencies *PFMik*. Finally *Α* is the set of BEAR characters and MBR(j,k) is the substitution matrix value between BEAR characters *j* and *k*. *D* is 0 if and only if the model PFM contains a subsection that is identical to the gold string due to the property of the substitution matrix of having the highest scores on its diagonal.

A third value we looked at, in addition to MCC and distance, was the folding accuracy, defined as the mean accuracy over all RNAs inserted in a dataset scored against the gold structure. In detail:

*foldni* is the i-th character of the n-th inserted RNA and Δ is the difference in length between the RNA and the gold structure. The folding accuracy *facc* is *1* if all the inserted RNAs contain a structure that is identical to the reference structure.

This measure is useful to distinguish between optimal and non-decidable cases when analysing our performances (if the folding was too different from the expected structure the problem lies in the secondary structure predictor).

As expected, looking at single sets aggregate results, we can see that the distribution of single runs spans along almost all the MCC/distance versus insert density space. However, by using a different color code for different *facc*  a trend emerges, both in the measure of MCC and distance of the found motif from the correct structure (Figure S2). To better visualize this result we plotted the same data by binning over different values of facc. In this way we could evaluate the performances of our method having isolated a possible form of bad benchmark that is actually due to out-of-the-software inaccuracies. Separately aggregating these in four classes yielded a comprehensive view of the test results.

By looking at the results coming from a folding accuracy between 75% and 100%, we see the real performances of BEAM. Good results in terms of MCC (that is, retrieving most of the correct structures and including the least possible False Positives from the background) come from 100% of insert percentage down to 20%. Moreover, from the distance plot we can tell that at the same percentages of inserts we are retrieving a collection of structures that are mostly similar, but not identical (distance ≠0), to the inserted motif (this is due to the dataset construction process, and it ensures us the presence of a variability which is inevitably present in other datasets). Good results in terms of MCC come from the second group too (50% < facc < 75%) but with a higher detach from the golden standard (distance ~0.4, meaning that it can bear up to 40% alteration of structure in terms of the similarity matrix with respect to the “correct” one.) The consequence is that BEAM is able to retrieve the correct structures in large datasets (104 sequences) with up to 70-80% of noise and the identification is solid up to 40% of structure alteration in terms of BEAR alignment score due to folding inaccuracies or mutations.

**Figure S2**. Benchmark plots for large artificial datasets. Left panel: MCC over percentage of inserts. The data shows the high dependence on folding accuracy. Where folding accuracy is within 50% from correct structures BEAM successfully retrieves the gold motif (MCC > 0.9) in datasets with up to 80% of noise. Right panel: Distance over percentage of inserts. We see how folding accuracy is related to the distance of the retrieved motif from the intended gold structure. This is consistent since we expect the mean distance to grow as the folding accuracy falls. All errors are computed using a sample standard deviation.

**P-value**

To evaluate a motif discriminatory ability to characterize the input set from which it was found we compute a p-value, with the following procedure:

1. Every sequence in the background set is aligned to the model by means of the similarity matrix MBR and the score is noted.
2. The distribution of these scores is assumed to be Gaussian, and it constitutes the background distribution.
3. The motif model mean score per sequence is assumed to be a random variable π with the same distribution as the background.
4. We compute the probability that, under these assumptions, π assumes a value that is higher than the one seen with the motif model, a p-value.

**Table S2:** BEAM parameters detailed description

|  |
| --- |
| -f input file |
| -g backGround, to compute p-value |
| -w Min motif Width (10) |
| -W Max motif Width (50) |
| -s minSteps (10000) |
| -S MaxSteps (15000) |
| -M masks - number of masks(motifs) to be computed (1) |
| -R runs - number of runs to be tried before choosing the mask |
|  |
| -T Starting Temperature (100) |
| -r cooling Rate (0.001) |
| -o output folder |
| -C clean mode (1,2,3) : 1. Exclude negative partials 2. Exclude partials under the 50% of the partial mean 3. Exclude partials under 90% of the mean. (3) |
| -b F (false) - weight less (mbr[:][:]=0 ) the branch-branch alignments |
| -v t Verbose, otherwise mostly silent |
| -k keep all the runs, otherwise show best |
| -n model\_limit (100) – influences the maximum number of structures that form a model. It is advisable not to go over this limit, for computational time reasons |
| -h print help and exit |

**Figure S3**. Positional preference for DoRiNA datasets. This collection of histograms shows positional preference for the motifs found with DoRiNA derived datasets, one for each RBP. The zero of the x-axis corresponds to the start of the binding site as reported in the BED files for every RNA.

**References**

1. Crooks,G., Hon,G., Chandonia,J. and Brenner,S. (2004) NCBI GenBank FTP Site\nWebLogo: a sequence logo generator. *Genome Res*, **14**, 1188–1190.

2. Mattei,E., Pietrosanto,M., Ferrè,F. and Helmer-Citterich,M. (2015) Web-Beagle: a web server for the alignment of RNA secondary structures: Figure 1. *Nucleic Acids Res.*, **43**, W493–W497.