Nucleic Acids Research Advance Access published November 28, 2016

doi: 10.1093/nar/gkw1094

A high-throughput approach to profile RNA structure

Riccardo Delli Ponti[1,2], Stefanie Marti[1,2], Alexandros Armaos[1,2] and Gian Gaetano Tartaglia[1,2,3,*]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, 08003 Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain and [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluıs Companys, 08010 Barcelona, Spain

Received August 02, 2016; Revised October 05, 2016; Editorial Decision October 25, 2016; Accepted October 28, 2016

ABSTRACT

Here we introduce the Computational Recognition of Secondary Structure (CROSS) method to calculate the structural profile of an RNA sequence (single- or double-stranded state) at single-nucleotide resolution and without sequence length restrictions. We trained CROSS using data from high-throughput experiments such as Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE; Mouse and HIV transcriptomes) and Parallel Analysis of RNA Structure (PARS; Human and Yeast transcriptomes) as well as high-quality NMR/X-ray structures (PDB database). The algorithm uses primary structure information alone to predict experimental structural profiles with >80% accuracy, showing high performances on large RNAs such as *Xist* (17 900 nucleotides; Area Under the ROC Curve AUC of 0.75 on dimethyl sulfate (DMS) experiments). We integrated CROSS in thermodynamics-based methods to predict secondary structure and observed an increase in their predictive power by up to 30%.

INTRODUCTION

The structure of an RNA determines its interactions and functions (1,2). RNA structure can be studied using low-throughput techniques such as nuclear magnetic resonance (NMR) and X-ray crystallography. More recent approaches have started to exploit biochemical reactions to perform high-throughput profiling of the RNA structure: Parallel Analysis of RNA Structure (PARS) distinguishes double- and single-stranded regions using the catalytic activity of two enzymes, RNase V1 (able to cut double-stranded nucleotides) and S1 (able to cut single-stranded nucleotides) (3,4), while Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) (5,6) employs highly reactive chemical probes such as 1M6, NMIA (SHAPE) and NAI-N_3 (icSHAPE) to characterize RNA backbone flexibility. Another technique based on dimethyl sulfate (DMS) (7) is often used for *in vivo* probing of transcriptomes (8,9). DMS experiments are of high quality due to the smaller size of the $(CH_3O)_2SO_2$ probe, yet have low coverage, since the alkylating agent only reacts to adenine and cytosine.

Transcriptomic studies require intense experimental work that could be substantially reduced by using computational approaches. We built Computational Recognition of Secondary Structure (CROSS) to perform high-throughput predictions of transcript structure using the information contained in RNA sequences. The algorithm predicts the structural profile (single- and double-stranded state) of a transcript at single-nucleotide resolution using sequence information only and without sequence length restrictions.

We trained CROSS on data from high-throughput [PARS: yeast and human transcriptomes (3,4) and ic-SHAPE: mouse transcriptome (5)] and low-throughput [SHAPE: HIV RNA (10)] experiments as well as high-quality NMR/X-ray structures (11). We did not use DMS experiments because they do not provide information on the structural state of all the nucleotides (1,5). Each of the five models reflects the specificities of the experimental technique used to generate the data. Since each approach has practical limitations and a different range of applicability, we also evaluated different methods to integrate the five models into a single algorithm, *Global Score,* to provide a *consensus* prediction.

The core of CROSS is an artificial neural network yielding a propensity score ranging from −1 (bottom values; single-stranded RNA) to 1 (top values; double-stranded RNA). CROSS was designed to investigate large-scale data sets and to provide information that can be integrated in methods for prediction of RNA secondary structure (12) as well as interactions with other molecules (13).

MATERIALS AND METHODS

CROSS architecture

We trained CROSS models using an artificial neural network with one hidden layer and two adaptive weight matrices ω_k^i and Ω^k that are optimized using backpropagation.

*To whom correspondence should be addressed. Tel: +34 93 316 01 16; Fax: +34 93 396 99 83; Email: gian.tartaglia@crg.es

In our approach, we use the 4-mer notation to represent each nucleotide: A = (1, 0, 0, 0), C = (0, 1, 0, 0), G = (0, 0, 1, 0) and U = (0, 0, 0, 1). The input of our method (Supplementary Material: *Data sets*) is the vector F_i encoding the information on fragments of fixed length (Supplementary Material: *Selection of the optimal window*). The input information required to predict the structural state of a specific nucleotide was extracted using a sliding window spanning the precedent and subsequent 6 residues (i.e. 13 nucleotides; longer fragments do not substantially improve the method; Supplementary Material: *Selection of the optimal window*; Supplementary Table S1).

This input F_i is propagated to the first hidden layer of k nodes as

$$h_k = tanh(\omega_k^i F_i) \quad (1)$$

where $tanh(x)$ is the hyperbolic tangent of x and the sum follows Einstein's notation.

The score Π of the nucleotide in the center of the window is then given by

$$\Pi = tanh(\Omega^k h_k) \quad (2)$$

where the contributions h_k of the hidden layer are weighted by Ω^k.

To avoid over-fitting when optimizing ω_k^i and Ω^k, we varied the number of nodes proportionally to the size of the training set and performed a 5-fold cross-validation at each optimization. For $k = 20$ we obtain the performances reported in the Supplementary Figures S1 and S2.

Consensus models

Since one technique might not be sufficient to capture structural properties of long transcripts (14), we evaluated different approaches to combine the five CROSS models (PARS-Human, PARS-Yeast, SHAPE-HIV, icSHAPE-Mouse and NMR/X-ray) into a *consensus* prediction. To this aim, we measured the performances of the models on an independent test set of 67 NMR/X-ray structures (15) for which SHAPE data are available (17 145 fragments in total), evaluating precision (PPV) and Area Under the ROC Curve (AUC; Supplementary Figure S3). Consistently with the type of information contained in the training set, we observed the best performances for the NMR/X-ray model (PPV: 0.69; AUC: 0.64) followed by HIV-SHAPE (PPV: 0.64; AUC: 0.63). Comparing the scores of the five models, we did not find strong correlations (Supplementary Table S2), except for PARS-Human and PARS-Yeast (Pearson's correlation = 0.50) that were trained on data obtained with the same experimental techniques (Figure 1; Supplementary Figures S1 and S2).

Z-Score

We combined the five CROSS models into a *Z-Score* variable. For each nucleotide in the sequence, the *Z-Score* is computed using the mean of the individual scores and the associated standard deviation: the double-stranded propensity is proportional to the *Z-Score*. We used this method to predict the structural profile of the *Xist* non-coding RNA (17 900 nt) and found an AUC of 0.75 on data from DMS experiments (16).

Global Score

We employed the scores of the five CROSS models to train a single classifier. The training set comprised 43 sequences (11 670 fragments) and the test set was composed of 24 transcripts (5 475 fragments; not in the training set of any of the CROSS models) (15). Among different classifiers the support vector machine with radial basis function kernel shows the best performances (Supplementary Table S3).

The *Z-Score* and *Global Score* predictions show a correlation of 0.85 (0.97 with a smoothing window of 200 nt) when applied to the *Xist* non-coding RNA (17 900 nt) (Supplementary Table S4). The high correlation indicates that the five models are assigned similar weights by *Global Score* and thus have similar performances. Since CROSS *Z-Score* and *Global Score* are correlated, we only provide *Global Score* on our webserver.

RNAstructure

We used *RNAstructure* with the *Fold* module and the *minimum free energy* flag to predict the best RNA secondary structure of each RNA sequence (17,18). To mimic experimental constraints in the *RNAstructure* algorithm, CROSS Global scores were normalized to lie in the range of SHAPE reactivities: first the scores were multiplied by −1, then linearly mapped to [0,1]. Scores >0.65 were then assigned a SHAPE reactivity of 1; scores <0.35 were assigned a reactivity of 0; scores >0.35 and <0.65 were linearly mapped to (0,1). We used the *Partition* and *Probability Plot* (with *-text* flag) modules of *RNAstructure* to compute the AUC based on the probabilities (17,18). We employed the package *Scorer* to calculate the positive predictive values (PPVs) and true positive rates (TPRs) for the specific structures.

Sequence patterns

We used DREME from the MEME suite (http://meme-suite.org/doc/dreme.html) to search for patterns in the positive and negative fragment sets (19). The flag *–n* was selected to specify a negative data set as comparison during the search of the motives.

RESULTS AND DISCUSSION

The CROSS algorithm

CROSS predicts the structural profile of an RNA sequence at single-nucleotide resolution and without sequence length restrictions. The algorithm is an artificial neural network with one hidden layer and two adaptive weight matrices to predict the structural state of a nucleotide considering its flanking residues (Materials and Methods: *CROSS architecture;* Supplementary Material: *Selection of the optimal window*). We built five independent models using data from SHAPE (Mouse and HIV transcriptomes (5,10): icSHAPE-Mouse and SHAPE-HIV) and PARS experiments (Human and Yeast transcriptomes (3,4): PARS-Human and PARS-Yeast) as well as data from NMR/X-ray studies (PDB database: NMR/X-ray) (Figure 1A). The training of each model was carried out on strong-signal sequences (Supplementary Material: *Data sets*) with the central nucleotide in

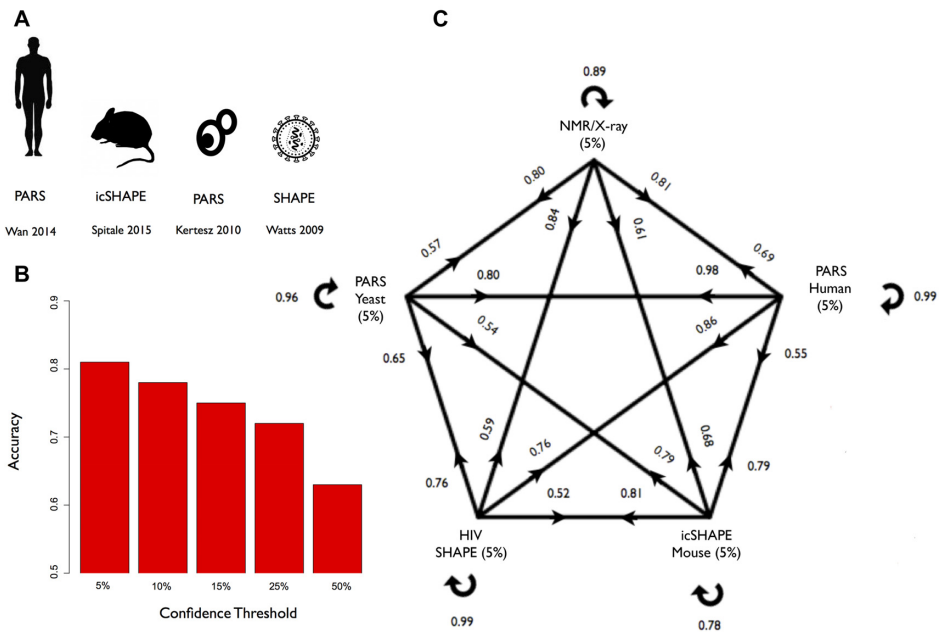

Figure 1. (**A**) To build the Computational Recognition of Secondary Structure (CROSS) models we used experimental data from four transcriptome-wide studies (*H. sapiens*, *M. musculus*, *S. cerevisiae* and HIV-1) as well as NMR/X-ray structures (3,5,4–11). Each model was trained on one data set (PARS-Human, PARS-Yeast, icSHAPE-Mouse, SHAPE-HIV and NMR/X-ray) and tested on the others. (**B**) Performances increase from low- (median) to high-confidence (top and bottom 5%) values of the CROSS scores distribution (Supplementary Figures S4 and S5). The plot illustrates the performances of the icSHAPE-Mouse model tested on the SHAPE-HIV data set. (**C**) High-confidence predictions: the arrows connect the training and testing sets along withrelative accuracies (cross-validation on training sets are marked with circular arrows). We used the same number of nucleotides with high (double-stranded) and low (single-stranded) propensity scores for comparison with experimental data. Negligible overlap exists between training and testing sets (Jaccard index < 0.002 between each couple of sets analyzed; Supplementary Table S5).

either single-stranded (negative cases) or double-stranded (positive cases) configuration. Each model was then tested on all the other data sets. Negligible overlap exists between training and testing sets (Jaccard index < 0.002 between each couple of sets analyzed; Supplementary Table S5).

From low- (top and bottom 50% of the CROSS score distribution) to high-confidence (top and bottom 5%) predictions, we observed an increase in the accuracies of our models, which indicates good ability to capture strong-signal regions. For instance, the accuracy of the icSHAPE-Mouse model applied to the SHAPE-HIV data set improves from 0.63 (low-confidence) to 0.81 (high-confidence; Figure 1B), and the same trend is found with respect to other data sets (Supplementary Figures S4 and S5). High-confidence predictions of PARS-Human (training fragments: 26 444; testing fragments: 77 476) and icSHAPE-Mouse (711 480 training fragments, 35 516 testing fragments) models on all the other sets reach accuracies of 0.77 and 0.76, PPVs of 0.80 and 0.77, TPRs of 0.76 and 0.76 and true negative rates of 0.79 and 0.77, respectively (Figure 1C; Supplementary Figures S1 and S2). As for PARS-Yeast, the accuracy, PPV and TPR are 0.64, 0.68 and 0.64, respectively. The model trained on NMR/X-ray data (29 428 training fragments, 77 176 testing fragments) shows an accuracy of 0.76, a PPV of 0.73 and a TPR of 0.79. SHAPE-HIV (fragments: 6 474 for training, 410 578 for testing) has an average accuracy, PPV and TPR of 0.66, 0.64 and 0.69.

We observed comparable cross-validation performances on the PARS datasets (area under the ROC curve AUC of 0.89 for PARS-Yeast applied to PARS-Human, and 0.90 for PARS-Human applied to PARS-Yeast), even though the experiments were carried out in different organisms and with slightly modified protocols, confirming the high quality of our predictions (Figures 2 and 3).

From low- (top and bottom 50% of the PARS score distribution) to high-confidence (top and bottom 1%) experimental values, we found a consistent increase in the performances of all models (Supplementary Tables S6 and S7), thus providing strong evidence on the reliability of CROSS predictions. For instance, the SHAPE-HIV model predicts the whole PARS-Human data set with an AUC of 0.70 and the top and bottom 1% of the scores are with an AUC of 0.80 (Supplementary Table S6). We note that very negligible overlap exists between yeast and human fragment sets (overlap: 0.001%; Jaccard index: 0.001; Supplementary Figure S6), which indicates that our method is not biased by specific sequences. On the same sets, approaches based on thermodynamic principles (15,18) show lower performances (Yeast: accuracies in the range 0.72–0.74, Human: accuracies in the range 0.67–0.69) than CROSS (Yeast: 0.80 accuracy using PARS-Human model; Human: 0.81 accuracy using PARS-Yeast model; Supplementary Table S8), indicating that our method is particularly useful for predictions on high-throughput data sets.

The HIV-1 case: correlation between *in silico* and *in vitro* data

The model built on PARS-Human is able to predict SHAPE-HIV data with an AUC of 0.75 (Figure 4A and B). Increasing the confidence threshold of SHAPE data (from

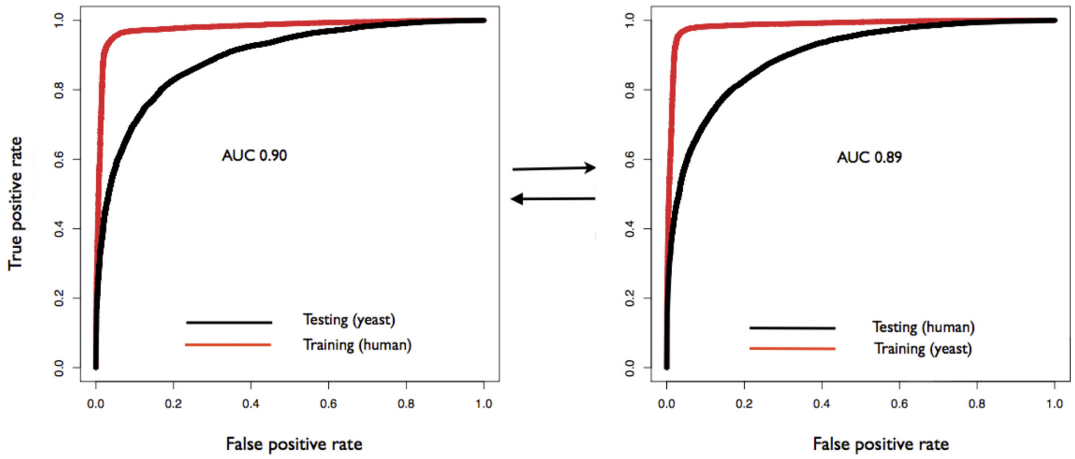

Figure 2. Receiver Operating Characteristic (ROC) curves reveal significant cross-validation performances on the complete data sets of PARS-Human (left panel; area under the ROC curve (AUC) = 0.90) and PARS-Yeast (right panel; AUC = 0.89).

Training

Testing	PARS yeast	PARS human	HIV Shape	icSHAPE	NMR/ X-ray
PARS yeast	0.96	**0.90**	**0.72**	0.55	0.60
PARS human	**0.89**	0.96	**0.75**	0.53	0.61
HIV Shape	**0.73**	**0.75**	0.99	0.56	0.62
icSHAPE	0.67	0.64	**0.70**	0.70	0.61
NMR/X-Ray	0.70	0.69	**0.71**	0.57	0.86

Figure 3. Performances on complete data sets. Testing performances with AUC > 0.70 are highlighted in bold (intra-set 5-fold cross-validations are in grey).

>0.5 for single-stranded, <0.2 for double-stranded to >1 for single-stranded and <0.1 for double-stranded) improves CROSS performances to an AUC of 0.80 (Figure 4B). We compared experimental and predicted values on fragments of 200 nucleotides, reporting an average correlation of 0.60 (peak of 0.86 in the region 3 800–4 000) for the HIV transcriptome (Figure 4A and C).

Recognition of complex patterns

CROSS is able to identify sequence patterns that cannot be captured by a position weight matrix approach. We searched the positive and negative fragment sets extracted from icSHAPE-Mouse and PARS-Human data for sequence patterns (Supplementary Table S9) using DREME (Materials and Methods: *Sequence patterns*) (19). In icSHAPE-Mouse sequences, the G/GC/ACGU/GC motif occurs with frequencies of 63% and 43% in the positive (556 645 fragments) and negative (355 740 fragments) sets (Supplementary Table S9), indicating poor discrimination. As for PARS-Human, the top motif in the positive fragment set GCU/GC/AG/G (71% frequency) is also non-specific (frequency of 47% in the negative set). This observation indicates that the neural network approach is particularly suitable to identifying complex patterns in biological sequences, which is key to discover trends in large data sets (20). We also note that CROSS models are sensitive to single point mutations: the signal drops progressively upon insertion of random mutations in the original sequences (PARS-Yeast; Supplementary Figure S7). As expected, mutations in the central position of the fragment produce the most dramatic reduction in the predictive power of the method (Supplementary Figure S8).

The consensus model *Global Score*

The consensus model *Global Score* was trained and tested on independent sets of NMR/X-ray structures (11 670 training fragments, 5 475 testing fragments; Supplementary Material: *Data sets;* Materials and Methods: *Consensus models*) (15,21). In the testing phase, single and double-stranded nucleotides were recognized with an AUC of 0.72 and a PPV of 0.74. Comparing the structures with experimental SHAPE data, we observed similar performances (AUC of 0.76 and PPV of 0.76 on the same data set; Figure 5). As PARS-Yeast and PARS-Human models show a 0.5 correlation (Supplementary Table S2), we decided to train the method without PARS-Yeast or PARS-Human. The procedure reduces *Global Score* performances (without PARS-Yeast: AUC from 0.72 to 0.65, PPV from 0.74 to 0.68; without PARS-Human AUC from 0.72 to 0.66, PPV from 0.74 to 0.65), which indicates that the methods should be used together.

Global score as experimental constraint for thermodynamic approaches

As previously done with experimental SHAPE data, we used *Global Score* as a constraint in *RNAstructure* (12). On the test set (15), *Global Score* increases the PPV of *RNAstructure* from 0.68 to 0.72, with remarkable improvements in 13 cases (from 0.44 to 0.72; Supplementary Table S10; Figure 6A and C; Supplementary Figure S9; Materials and Methods: *RNAstructure*), and decreases the PPV in three

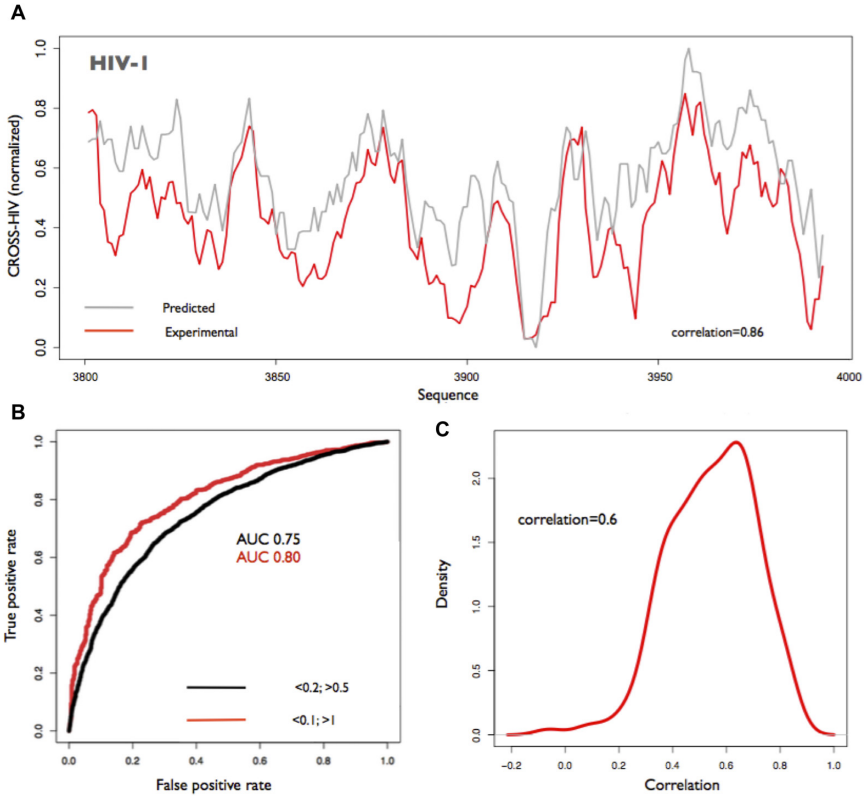

Figure 4. (**A**) Example of the secondary structure profile of the HIV transcriptome (nucleotides 3 800–4 000) calculated with the PARS-Human model. CROSS predictions show a correlation of 0.86 with the experimental Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) profile. (**B**) ROC curves of CROSS-Human applied to HIV-SHAPE data. The performances increase when selecting a high confidence threshold (>1 for single-stranded; <0.1 for double-stranded) on SHAPE experimental data. (**C**) Pearson correlations between experimental and predicted data for the HIV transcriptome calculated on 200-nucleotide regions using a smoothing window of 7 nucleotides. The average correlation is 0.6.

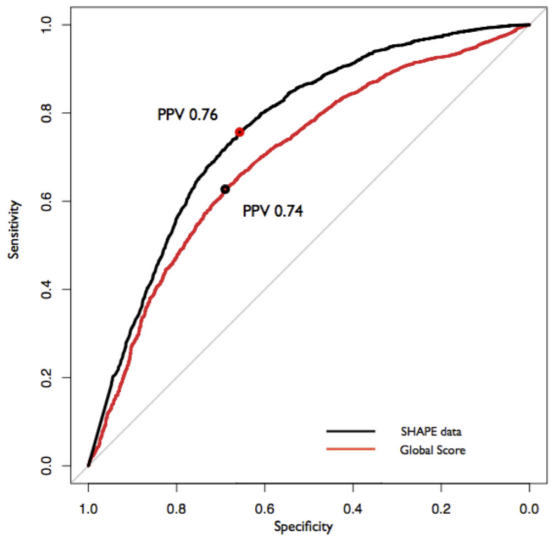

Figure 5. Performance comparison of SHAPE data and CROSS predictions (*Global Score*) on sequences with available structural information derived from NMR/X-ray data (21). The performances are calculated on 24 RNAs that were not employed for training (15). The precision is measured at Youden's cut-off.

cases for which real SHAPE data does not improve performances (PPV: Group II intron *O. iheyensis*: 0.97 with *RNAstructure* versus 0.84 with SHAPE data; HIV-1 5′ pseudoknot domain: 0.62 versus 0.55; SARS corona virus pseudoknot: 0.90 versus 0.75). To assess to what extent *Global Score* improves *RNAstructure* (Supplementary Table S10), we randomized the *Global Score* input and observed an overall PPV decrease to 0.64. Moreover, using the partition function computed with *RNAstructure*, we calculated the AUC for each structure with and without CROSS constraints and observed an improvement from 0.81 to 0.86 when CROSS is integrated in the algorithm (Figure 6B). On the test set (15), we found a similar trend using *RNAfold* (15) (the PPV increases from 0.67 to 0.70 using *Global Score* and the AUC remains at 0.85).

The X*ist* case and comparison with DMS experiments

Due to the complexity of the configuration space, the structural profile of sequences >1 000–1 500 nucleotides is extremely difficult to predict with thermodynamic approaches (22), which makes CROSS a valid alternative to study long non-coding RNAs (23). To illustrate CROSS performances on large RNAs, we predicted the structural profile of murine *Xist* non-coding RNA (17 900 nt) using the *consensus* of our five models (Materials and Methods: *Consensus models*; Figure 7A). *Xist* was analyzed using DMS probing (16), an independent technique not used in the training of CROSS (the transcript was not present in any training set of

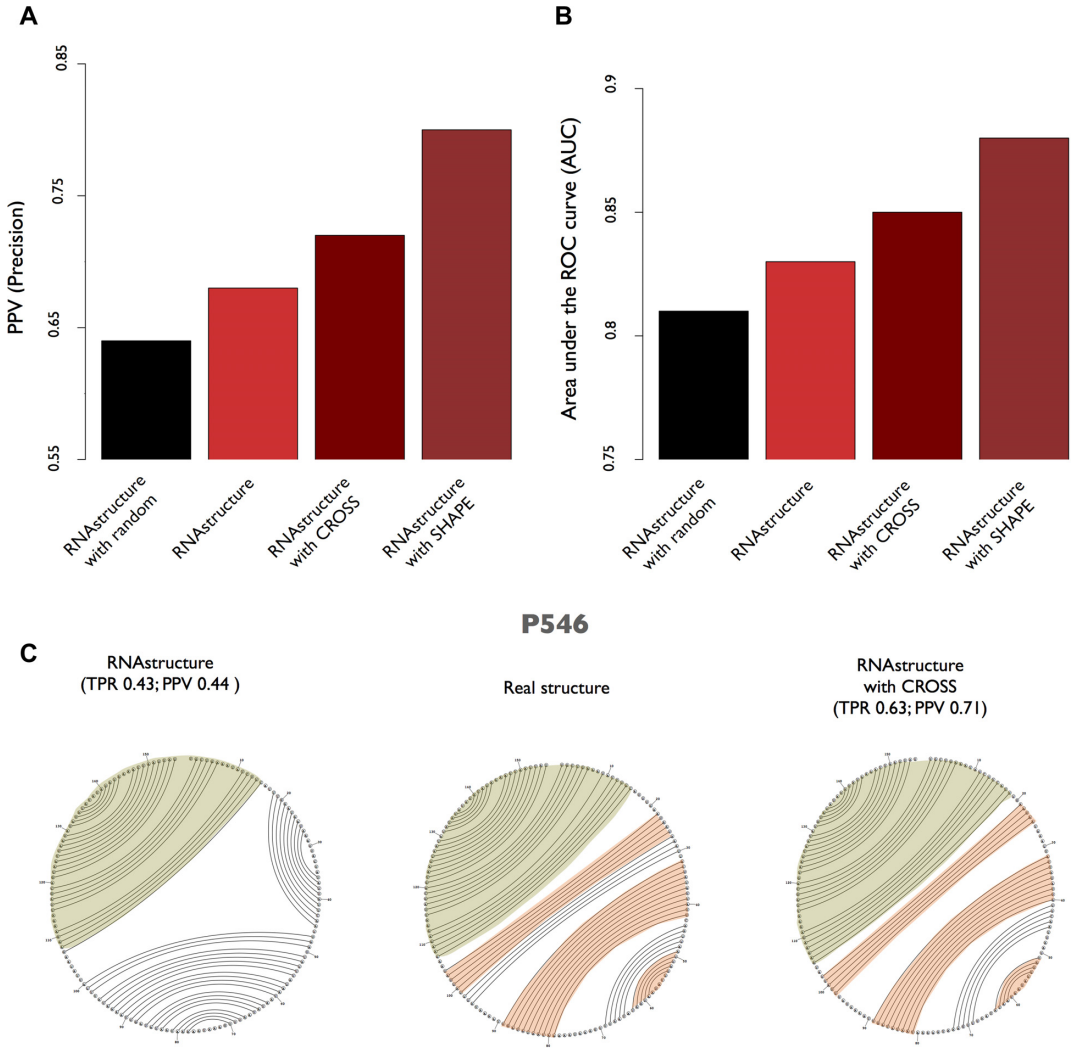

Figure 6. (**A** and **B**) Performances of *RNAstructure* (12,18) without constraints and using either CROSS *Global Score* predictions or SHAPE data. Precision (PPV) and area under the ROC curve (AUC) increase when CROSS predictions are employed as constraints in *RNAstructure*, indicating a global improvement of predictive power. Randomizing the CROSS signal (in the range of SHAPE data) decreases the performances of *RNAstructure.* (**C**) Prediction of the structure of the P546 domain bI3 group I intron using CROSS predictions as constraints to *RNAstructure*: Sensitivity (TPR) (without CROSS: TPR = 0.43; with CROSS: TPR = 0.63) and precision (without CROSS: PPV = 0.44; with CROSS: PPV = 0.71) are reported.

CROSS). Using the top and bottom 10% of the experimental DMS data on *Xist* profile (3 580 fragments removing regions with unreliable scores in Rep B) the Z-Score shows an AUC of 0.75 (Figure 7A, right corner). In agreement with DMS experiments (16), CROSS identifies the structural elements associated with repetitive regions Rep A, B and F and resolves their internal structures with correlations of 0.35, 0.46 and 0.75, respectively (see Supplementary Figure S10). Although the method slightly overestimates the structural content of Rep E, it is able to accurately predict its profile (correlation of 0.63, Supplementary Figure S10). While the sequences of Rep A and B are conserved across species and show a high degree of structural content, the 3′ region of *Xist* is variable (24) and predicted by CROSS to be more single-stranded.

Structural differences in human CDS, UTRs and lincRNAs

We employed CROSS to analyze the structural differences between human coding DNA sequences (CDSs), untranslated regions (UTRs) (total of 217 000 non-redundant sequences each with 3′ and 5′ UTRs; ENSEMBL 82) and long intergenic non-coding transcripts (14 000 non-redundant sequences; ENSEMBL 82; Supplementary Material; Figure 7B). In agreement with previous evidence (1), we predict that UTRs are more structured than CDSs (*P*-value < 2.2e-16; Kolmogorov–Smirnov). Long intergenic non-coding transcripts (see Supplementary Material: *Long intergenic non-coding RNAs*) are found to be less structured, as reported in other studies (25) (*P*-value < 2.2e-16; Kolmogorov–Smirnov). Indeed, long non-coding RNAs have complex regulatory abilities and their structure could be more flexible and less structured to provide a wide range of interactions (26). In agreement with previous data (27), we also observe that the 5′ UTR of the amyloid precursor

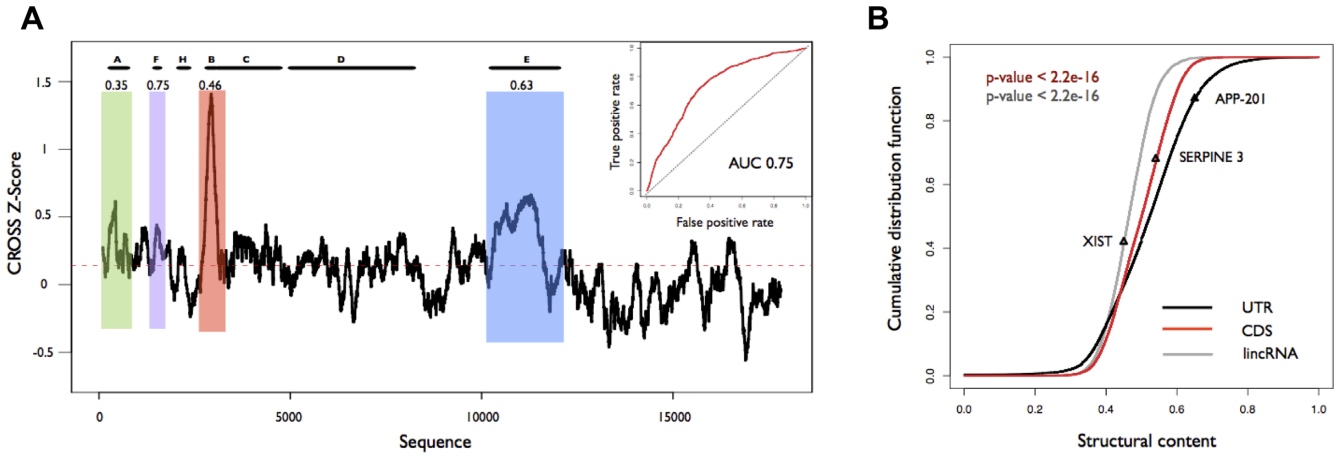

Figure 7. (**A**) CROSS *Z-Score consensus* prediction of the secondary structure profile of murine *Xist* long non-coding RNA (a 200 nt window is employed for smoothing). Structured regions, in correspondence to known repetitive domains (Rep A, B and F), are highlighted and the correlations with dimethyl sulfate (DMS) data (16) are reported on top. A detailed view of the CROSS and DMS profiles for Rep A, B, F and E is provided in Supplementary Figure S10. We note that Rep B contains regions with insufficient sequencing data to determine DMS reactivity (16) that were excluded from the analysis. Our predictions indicate lower structural content at 3′, in line with previous reports indicating poor sequence conservation (only Rep A and B are highly conserved) (24). The ROC curve of the *Z-Score* predictions on high-confidence DMS data (10% top and bottom nucleotides, 3 580 fragments) is reported in the corner (AUC 0.75). (**B**) Predictions of human coding DNA sequences (CDSs), untranslated regions (UTRs) and long-intergenic non-coding RNA (lincRNA) (ENSEMBL version 82; total number of transcripts: 50 000; 14 000 lincRNA isoforms). We predict that the UTRs are more structured than the CDSs, in agreement with previous studies (*P*-value < 2.2e-16, Kolmogorov–Smirnov) (1). For each set we show a known example [the APP 5′ UTR is more structured, as shown in previous studies (27); SERPINE3 has a structured CDS in agreement with PARS data (4); *Xist* structural content is in agreement with DMSdata (16)].

protein APP transcript is highly structured (>65% double-stranded). Similarly, the mRNA of serpin peptidase inhibitor SERPINE3 is predicted to be highly structured (>55%), as reported in PARS screenings (4). We predict that 45% of *Xist* is structured in domains, as revealed by DMS profiling (16).

CONCLUSION

The study of large transcripts requires intense experimental work that could be substantially reduced by using computational approaches to characterize their structural features (16). Methods based on thermodynamic principles (18,28) can be employed for RNAs < 1 000–1 500 nucleotides and do not work for larger molecules because of the complexity of the configuration space (22). In our approach, we use local sequence properties of RNAs, which is key to perform fast high-throughput profiling of sequences, since the computational load scales linearly with the sequence length. Therefore, CROSS allows the prediction of the structural profile without sequence length restrictions.

We built CROSS using data from SHAPE (5,6) and PARS (3,4) studies as well as NMR/X-ray experiments. Models based on PARS and icSHAPE experiments show the highest predictive power with an average accuracy of 0.77 and 0.76, and a positive predictive value PPV of 0.8 and 0.77. The different algorithms can be used independently or combined together to obtain insights into the secondary structure of a transcript. Since each technique has its specificities and biases, the combination of multiple approaches is recommendable to achieve a better understanding of structural properties (14).

On high-throughput experimental data sets CROSS outperforms *RNAstructure* (18) and *RNAfold* (15) (CROSS: accuracy of 0.80 for PARS-Yeast and 0.81 for PARS-Human; *RNAstructure* and *RNAfold*: 0.72–0.74 for PARS-Yeast, 0.67–0.69 for PARS-Human). Yet, previous studies indicate that thermodynamic methods have a higher predictive power when the information derived from SHAPE experiments is integrated (12). Comparing SHAPE experiments and CROSS predictions on RNA molecules for which NMR/X-ray data are available (15), we found similar performances with an average precision of 0.74 (CROSS) and 0.76 (SHAPE), and an area under the receiver operating characteristics of 0.72 (CROSS) and 0.76 (SHAPE). Thus, CROSS can be considered an *in silico* alternative to SHAPE experiments (5,6) and its integration in *RNAstructure* (17,18) shows performances (PPV: 0.72; AUC: 0.85) that are comparable to those achieved using real SHAPE data (PPV: 0.80, AUC: 0.88).

Since CROSS is fast (less than 2 min to profile a transcript of 20 000 nucleotides), it can be used for high-throughput predictions of the RNA secondary structure. We used CROSS to investigate profiles of sequences taken from CDSs as well as untranslated regions UTRs > 200 000 isoforms (calculated in <72 h) reporting a structural content that is compatible with what is available in literature (1). We also studied the structural profile of *Xist* and identified specific regions in agreement with DMS experiments [correlations of 0.63 and 0.75 for Rep E and Rep F (16)].

Our predictions of structural features will facilitate the design of experimental studies on long transcripts by revealing the structural state of their regions. The calculations can be employed to shed light on the evolution of RNA molecules and on their interactions with other molecules. Our approach can be also exploited to improve the predictive power of algorithms such as for instance *cat*RAPID, which computes the interaction propensity of protein and

RNA molecules (29). We envisage that the combination of CROSS with thermodynamics-based approaches will be the key ingredient to improve predictions of RNA structure.

AVAILABILITY

CROSS is freely available at http://service.tartaglialab.com/new_submission/cross.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Philipp Germann, Irene Julca, Davide Cirillo and the other members of our group for useful comments.

FUNDING

The research leading to these results has received funding from European Union Seventh Framework Programme [FP7/2007-2013]; European Research Council [RIBOMYLOME_309545 to GGT]; Spanish Ministry of Economy and Competitiveness [BFU2014-55054-P to GGT]; AGAUR [2014 SGR 00685 to GGT]; Spanish Ministry of Economy and Competitiveness, European Research Development Fund ERDF, 'Centro de Excelencia Severo Ochoa 2013-2017' [SEV-2012-0208]. Funding for open access charge: European Research Council [RIBOMYLOME_309545 to GGT]; Spanish Ministry of Economy and Competitiveness [BFU2014-55054-P to GGT]. The authors also thank the CRG fellowship to SM.
Conflict of interest statement. None declared.

REFERENCES

1. Mortimer,S.A., Kidwell,M.A. and Doudna,J.A. (2014) Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.*, **15**, 469–479.
2. Tartaglia,G.G. (2016) The grand challenge of characterizing ribonucleoprotein networks. *Front. Mol. Biosci*, **3**, doi:10.3389/fmolb.2016.00024.
3. Kertesz,M., Wan,Y., Mazor,E., Rinn,J.L., Nutter,R.C., Chang,H.Y. and Segal,E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.
4. Wan,Y., Qu,K., Zhang,Q.C., Flynn,R.A., Manor,O., Ouyang,Z., Zhang,J., Spitale,R.C., Snyder,M.P., Segal,E. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706–709.
5. Spitale,R.C., Flynn,R.A., Zhang,Q.C., Crisalli,P., Lee,B., Jung,J.-W., Kuchelmeister,H.Y., Batista,P.J., Torre,E.A., Kool,E.T. *et al.* (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, **519**, 486–490.
6. Wilkinson,K.A., Merino,E.J. and Weeks,K.M. (2006) Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.*, **1**, 1610–1616.
7. Cordero,P., Kladwang,W., VanLang,C.C. and Das,R. (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry*, **51**, 7037–7039.
8. Rouskin,S., Zubradt,M., Washietl,S., Kellis,M. and Weissman,J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–705.
9. Wells,S.E., Hughes,J.M., Igel,A.H. and Ares,M. (2000) Use of dimethyl sulfate to probe RNA structure in vivo. *Methods Enzymol.*, **318**, 479–493.
10. Watts,J.M., Dang,K.K., Gorelick,R.J., Leonard,C.W., Bess,J.W., Swanstrom,R., Burch,C.L. and Weeks,K.M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–716.
11. Andronescu,M., Bereg,V., Hoos,H.H. and Condon,A. (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340–349.
12. Deigan,K.E., Li,T.W., Mathews,D.H. and Weeks,K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 97–102.
13. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
14. Novikova,I.V., Hennelly,S.P. and Sanbonmatsu,K.Y. (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res.*, **40**, 5034–5051.
15. Lorenz,R., Luntzer,D., Hofacker,I.L., Stadler,P.F. and Wolfinger,M.T. (2016) SHAPE directed RNA folding. *Bioinformatics*, **32**, 145–147.
16. Fang,R., Moss,W.N., Rutenberg-Schoenberg,M. and Simon,M.D. (2015) Probing Xist RNA structure in cells using targeted structure-seq. *PLoS Genet.*, **11**, e1005668.
17. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
18. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129–138.
19. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME Suite. *Nucleic Acids Res.*, **43**, W39–W49.
20. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotech.*, **33**, 831–838.
21. Wu,Y., Shi,B., Ding,X., Liu,T., Hu,X., Yip,K.Y., Yang,Z.R., Mathews,D.H. and Lu,Z.J. (2015) Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Res.*, **43**, 7247–7259.
22. Lange,S.J., Maticzka,D., Möhl,M., Gagnon,J.N., Brown,C.M. and Backofen,R. (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.*, **40**, 5215–5226.
23. Ulitsky,I. and Bartel,D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26–46.
24. Nesterova,T.B., Slobodyanyuk,S.Y., Elisaphenko,E.A., Shevchenko,A.I., Johnston,C., Pavlova,M.E., Rogozin,I.B., Kolesnikov,N.N., Brockdorff,N. and Zakian,S.M. (2001) Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res.*, **11**, 833–849.
25. Wan,Y., Qu,K., Ouyang,Z., Kertesz,M., Li,J., Tibshirani,R., Makino,D.L., Nutter,R.C., Segal,E. and Chang,H.Y. (2012) Genome-wide measurement of RNA folding energies. *Mol. Cell*, **48**, 169–181.
26. Rinn,J.L. and Chang,H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–166.
27. Gsponer,J. and Babu,M.M. (2012) Cellular strategies for regulating functional and nonfunctional protein aggregation. *Cell Rep.*, **2**, 1425–1437.
28. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA Websuite. *Nucleic Acids Res.*, **36**, W70–W74.
29. Agostini,F., Zanzoni,A., Klus,P., Marchese,D., Cirillo,D. and Tartaglia,G.G. (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, **29**, 2928–2930.