BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 8 2014, pages 1049–1055
doi:10.1093/bioinformatics/btt757

Genome analysis

Advance Access publication December 27, 2013

Computational identification of protein binding sites on RNAs using high-throughput RNA structure-probing data

Xihao Hu[1], Thomas K. F. Wong[2,3], Zhi John Lu[4], Ting Fung Chan[5,6], Terrence Chi Kong Lau[7], Siu Ming Yiu[3] and Kevin Y. Yip[1,6,*]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, [2]Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong, [3]CSIRO Ecosystem Sciences, Canberra, ACT 2601, Australia, [4]MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China 100084, [5]School of Life Sciences, [6]Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong and [7]Department of Biology and Chemistry, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

Associate Editor: Inanc Birol

ABSTRACT

Motivation: High-throughput sequencing has been used to probe RNA structures, by treating RNAs with reagents that preferentially cleave or mark certain nucleotides according to their local structures, followed by sequencing of the resulting fragments. The data produced contain valuable information for studying various RNA properties.

Results: We developed methods for statistically modeling these structure-probing data and extracting structural features from them. We show that the extracted features can be used to predict RNA 'zip-codes' in yeast, regions bound by the She complex in asymmetric localization. The prediction accuracy was better than using raw RNA probing data or sequence features. We further demonstrate the use of the extracted features in identifying binding sites of RNA binding proteins from whole-transcriptome global photoactivatable-ribonucleoside-enhanced cross-linking and immunopurification (gPAR-CLIP) data.

Availability: The source code of our implemented methods is available at http://yiplab.cse.cuhk.edu.hk/probrna/.

Contact: kevinyip@cse.cuhk.edu.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 5, 2013; revised on December 8, 2013; accepted on December 13, 2013

1 INTRODUCTION

Next-generation sequencing technologies have created opportunities for studying many diverse properties of nucleic acids in a high-throughput manner. Recently, it has been used to probe RNA structures, which reveals interesting properties of RNA, including the largely unexplored mRNA structures (Kertesz *et al.*, 2010; Lucks *et al.*, 2011; Underwood *et al.*, 2010). One method involves treating RNAs with an enzyme that preferentially cleaves either double-stranded (such as RNase V1) or single-stranded (such as S1 nuclease and P1 nuclease) nucleic acids, and sequencing the resulting fragments (Kertesz *et al.*, 2010; Underwood *et al.*, 2010). Paired and unpaired nucleotides can then be deduced by comparing the distributions of read counts under the two different enzymatic treatments or by comparing with a control. Another method is based on Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) chemistry (Mortimer and Weeks, 2007). RNAs are treated with a SHAPE reagent, which chemically modifies the 2′-hydroxyl groups with reactivity at individual nucleotides depending on their local spatial disorder. The treated RNAs are then reverse-transcribed and subsequently sequenced. Because reverse transcription is blocked by the SHAPE adducts, the distribution of read counts can serve as an indicator of local structures (Lucks *et al.*, 2011).

For both methods, specialized algorithms have been proposed for processing the sequencing reads and analyzing the read counts (Aviran *et al.*, 2011; Kertesz *et al.*, 2010; Lucks *et al.*, 2011; Underwood *et al.*, 2010). These algorithms were designed to consider special properties of the corresponding experiments. For example, in SHAPE sequencing, reverse transcription could be stopped either by a SHAPE adduct or due to natural polymerase drop off. Different statistical models were designed for these processes for estimating the reactivity of the SHAPE reagent at each nucleotide.

In addition to these method-specific factors, read counts may also be affected by biases common to many sequencing protocols. For example, GC-rich regions may have more reads than AT-rich regions (Dohm *et al.*, 2008). Primer binding and amplification efficiency also depend on local sequences (Li *et al.*, 2010).

Here we show that statistical models that explicitly consider potential sequence-specific biases can be used to fit these high-throughput structure-probing data. The effectiveness of our models is demonstrated by a better goodness of fit to the data than some other models based on a cross-validation procedure.

To further explore the utility of our models, we show that features of structure-probing data extracted by our models can be used to locate zipcodes on yeast mRNAs, which are regions bound by the She complex in asymmetric localization (Shepard *et al.*, 2003). Previous studies have shown that localization activity depends on secondary structure (Chartrand *et al.*, 1999; Gonzalez *et al.*, 1999). A short sequence motif involving a

*To whom correspondence should be addressed.

CGA triplet in a loop and a conserved cytosine six bases away in another loop has also been shown to be necessary for bud localization of several RNAs (Olivier *et al*., 2005). Nonetheless, these criteria are not sufficient to identify all zipcodes (Shepard *et al*., 2003). It has been suggested that recognition and transport depend not only on the zipcodes but also on adjacent sequence and structural features (Jambhekar *et al*., 2005). We show that our extracted features could help distinguish zipcodes from other regions on the same mRNAs with good accuracy.

There exist experimental methods for transcriptome-wide identification of RNA regions bound by a specific RNA binding protein (RBP). The main idea is to crosslink RBPs with RNAs, followed by immunoprecipitation (CLIP) and microarray analysis or sequencing. The latter includes methods known as HITS-CLIP (Licatalosi *et al*., 2008), CLIP-seq (Sanford *et al*., 2009), PAR-CLIP (Hafner *et al*., 2010) and RIP-seq (Zhao *et al*., 2010).

As these large-scale datasets become available for more RBPs, it is interesting to ask whether different RBPs recognize similar features at their binding sites. It has recently been shown that a common set of features can be used to identify RNAs targeted by a group of different RBPs (Pancaldi and Bähler, 2011), although in the same study it is also shown that if the training set lacks any known targets of an RBP, the resulting statistical model is less capable of identifying its targets, indicating that the model may not have captured the binding site signals recognized by all RBPs. Here we use our extracted features from RNA structure-probing data to study this question in a different way, and ask if it is possible to build statistical models that can distinguish general RBP binding sites from other regions. Using a whole-transcriptome dataset of RBP binding sites, we show that such a model can be constructed with high distinguishing power using a small set of features, which supports the idea that different RBPs may share similar recognition signals of their targets.

2 MATERIALS AND METHODS

The overall workflow is illustrated in Supplementary Figure S1. The following sections provide details of different parts of the workflow.

2.1 RNA structure-probing data

We used a published transcriptome-wide structure-probing dataset in yeast (Kertesz *et al*., 2010) for our study, as we also had access to other types of yeast data needed for our study. The experiments that produced the data involved treatments of two enzymes, namely RNase V1 and S1 nuclease, which preferentially cleave phosphodiester bonds 3′ of double-stranded RNA and single-stranded RNA, respectively. The sequencing data from each enzymatic treatment were individually normalized, and the log ratio between the normalized V1 and S1 read counts at each nucleotide was defined as the PARS (Parallel Analysis of RNA Structure) score of the 5′ adjacent nucleotide (Kertesz *et al*., 2010). Strongly positive and negative PARS scores indicate a high chance for the nucleotide to be base-paired and single-based, respectively. For simplicity, we will call the number of reads attributed to the structure of a certain nucleotide its 'read count', although these reads start at its 3′ adjacent nucleotide.

2.2 Statistical models

For each enzymatic treatment, we propose a mixture of Poisson linear model to relate properties of each nucleotide and its observed read count. Our model is based on a Poisson linear model previously proposed for counting sequencing reads that start at a particular nucleotide (Li *et al*., 2010), which matches the situation of our data:

$$n_{ij} \sim \text{Poisson}(\mu_{ij}), \text{ where}$$
$$\ln \mu_{ij} = \ln \mu_i + \alpha + \sum_{k=1}^{K} \sum_{h \in \{A,C,G\}} \beta_{kh} \mathrm{I}(b_{ijk} = h) \quad (1)$$

In the equation, n_{ij} is the observed read count of nucleotide j of transcript i (i.e. the number of reads starting at the nucleotide), which is distributed according to a Poisson distribution with mean μ_{ij}. μ_{ij} in turn depends on μ_i, the unknown expression level of transcript i, and biases due to the local sequence within a window of size K centered on nucleotide j. Both the identity and position of the nucleotides within the window affect how they influence the read count of nucleotide j, and their total influence is assumed to take on a linear form with offset α and coefficients β_{kh} (no variables are defined for Uracils, as their values can be fixed at 0), where $\mathrm{I}(b_{ijk} = h)$ is 1 if the k-th nucleotide within the window is h, and 0 otherwise. The model is general enough to capture many types of biases due to local sequence features. The values of the parameters $\theta = \{\mu_i, \alpha, \beta_{kh}\}$ are to be determined such that the following log-likelihood function is maximized:

$$L(\theta) = \ln \prod_{i,j} \Pr(n_{ij}|\theta) = \sum_{i,j} \ln \frac{e^{-\mu_{ij}} \mu_{ij}^{n_{ij}}}{n_{ij}!} \quad (2)$$

For the structure-probing data we used, as an enzyme could have different probabilities of cleavage for different groups of nucleotides (such as paired versus unpaired bases), we extend the model by introducing a mixture of components, with each component representing a group of nucleotides. The resulting log-likelihood function is as follows:

$$\begin{aligned} L(\theta) &= \sum_{i,j} \ln \sum_{g \in G} \Pr(z_{ij,g} = 1) \Pr(n_{ij}|z_{ij,g} = 1, \theta) \\ &= \sum_{i,j} \ln \sum_{g \in G} \tau_g \frac{e^{-\mu_{ij,g}} \mu_{ij,g}^{n_{ij}}}{n_{ij}!}, \end{aligned} \quad (3)$$

where G represents the different components of the mixture model, $z_{ij,g}$ is a group membership variable, which equals 1 if nucleotide j of transcript i belongs to group g and 0 otherwise, $\tau_g \stackrel{\text{def}}{=} \Pr(z_{ij,g} = 1)$ is the prior probability that a nucleotide belongs to group g and $\mu_{ij,g}$ is the average read count of nucleotides of transcript i that belong to group g. Correspondingly, there is a separate set of parameters, α_g and $\beta_{kh,g}$ for each group. In this work we consider only two-component models, i.e. $G = \{1, 2\}$, as they correspond to some natural assumptions to be discussed below.

To check the need for considering local sequence biases and for a mixture model, we also considered two simpler models for comparison purposes. The first one is the original Poisson linear model with only one component [Equation (1)]. The second one is a Poisson mixture model that does not consider local sequences, which is equivalent to our model with α_g and $\beta_{kh,g}$ all set to zero.

2.3 Optimization algorithms for data fitting

2.3.1 Fitting data from one enzymatic treatment Given a sequencing dataset with a particular enzymatic treatment, we used optimization algorithms to (locally) maximize the data likelihood of the different models. For the single-component Poisson linear model, we implemented the method described in Li *et al*. (2010). For our mixture model, we developed an algorithm based on the expectation-maximization (EM) framework (Dempster *et al*., 1977). EM considers three types of data, which in our case include the observed read counts $(\mathbf{X} = \{n_{ij}\})$, hidden group membership variables $(\mathbf{Z} = \{z_{ij,g}\})$ and model parameters $(\theta = \{\mu_{i,g}, \alpha_g, \beta_{kh,g}, \tau_g\})$. If the values of the hidden variables were observed, it would be easier to compute data likelihood and find

parameter values that maximize it. Because these values are actually unobserved, the EM procedure instead iteratively maximizes the expectation of the log-likelihood of the full data, $E_{\mathbf{Z}|\mathbf{X},\theta^{(t-1)}}[\ln \Pr(\mathbf{X},\mathbf{Z}|\theta)] = \sum_{\mathbf{z}} \left[\Pr(\mathbf{z}|\mathbf{X},\theta^{(t-1)}) \ln \Pr(\mathbf{X},\mathbf{z}|\theta)\right]$, where $\theta^{(t-1)}$ is the estimated parameter values in the $(t-1)$-th iteration. The procedure repeatedly derives the expression of this expected log-likelihood in the E-step and finds values of the parameters that maximize it in the M-step, until a certain stopping criterion is reached.

For our mixture model, it is possible to derive closed-form formulas for $\mu_{i,g}$ and τ_g that maximize the expected log-likelihood, but it is difficult for α_g and $\beta_{kh,g}$. Following Li *et al.* (2010), we searched for the optimal values of these two sets of parameters in turn, using closed-form formulas and numerical methods, respectively. A summary of the whole algorithm is given below. Detailed derivations can be found in the Supplementary Materials.

(1) Define variables $\bar{z}_{ij,g}^{(0)}$ and initialize each of them with a random value from (0, 1) such that $\sum_{g\in G} \bar{z}_{ij,g}^{(0)} = 1$.

(2) Initialize $\mu_{i,g}^{(0)}$ to $\frac{\sum_{j=1}^{l_i} \bar{z}_{ij,g}^{(0)} n_{ij}}{\sum_{j=1}^{l_i} \bar{z}_{ij,g}^{(0)}}$ and $\tau_g^{(0)}$ to $\frac{\sum_i \sum_{j=1}^{l_i} \bar{z}_{ij,g}^{(0)}}{\sum_i l_i}$, where l_i is the length of transcript i. For each iteration $t = 1, 2, \ldots$, repeat steps 3–8:

(3) Viewing $\mu_{i,g}^{(t-1)}$ as offsets and $\bar{z}_{ij,g}^{(t-1)}$ as weights, fit the generalized linear model with a log link function in the Poisson family to get $\alpha_g^{(t)}$ and $\beta_{kh,g}^{(t)}$.

(4) Redefine $\bar{z}_{ij,g}^{(t-1)}$ as $\frac{\tau_g^{(t-1)} \Pr(n_{ij}|z_{ij,g}=1,\theta^{(t-1)})}{\sum_{g'}\left[\tau_{g'}^{(t-1)} \Pr(n_{ij}|z_{ij,g'}=1,\theta^{(t-1)})\right]}$.

(5) Update $\mu_{i,g}^{(t)}$ to $\frac{\sum_{j=1}^{l_i} \bar{z}_{ij,g}^{(t-1)} n_{ij}}{\sum_{j=1}^{l_i} \bar{z}_{ij,g}^{(t-1)} \exp\left[\alpha_g^{(t)} + \sum_{k=1}^{K} \sum_{h\in\{A,C,G\}} \beta_{kh,g}^{(t)} \mathrm{I}(b_{ijk}=h)\right]}$.

(6) If $\mu_{i,1}^{(t)} > \mu_{i,2}^{(t)}$ for a transcript i, swap their values as well as those of $\bar{z}_{ij,1}^{(t-1)}$ and $\bar{z}_{ij,2}^{(t-1)}$ for all nucleotides j on i.

(7) Update $\tau_g^{(t)}$ to $\frac{\sum_i \sum_{j=1}^{l_i} \bar{z}_{ij,g}^{(t-1)}}{\sum_i l_i}$.

(8) Go to step 3 unless the deviance (defined in the Section 2.4) decreases by <0.01%.

In the algorithm, we first initialize variables $\bar{z}_{ij,g}$ to random values, and $\mu_{i,g}$ and τ_g to corresponding weighted means of them (steps 1–2). The algorithm then repeats steps 3–8 for iterations. In step 3, we fix the values of $\mu_{i,g}$ and $\bar{z}_{ij,g}$ using the estimates from the previous iteration, and solve the resulting Poisson regression problem by iterative reweighting least-square (Li *et al.*, 2010) to get α_g and $\beta_{kh,g}$. In step 4, the variables $\bar{z}_{ij,g}$ are formally defined as the expected group membership of nucleotide j of transcript i based on the parameter estimates in the previous iteration. In step 5, the values of $\mu_{i,g}$ are updated to ones that maximize the expected log-likelihood expression obtained in the E-step. In step 6, we define the group with a smaller mean as group 1, and swap the related parameters if necessary. In step 7, we update the values of τ_g to the ones that maximize the expected log-likelihood. Finally, in step 8, we check if the change of deviance is smaller than a threshold, to determine whether to stop the execution or to enter the next iteration. In practice, the deviance value converges quickly and the algorithm requires only a small number of iterations (Supplementary Tables S1 and S2).

For the mixture of Poisson model that does not consider local sequences, we estimated its parameters by modifying our algorithm with α_g and $\beta_{kh,g}$ fixed to zero.

2.3.2 Fitting data from both enzymatic treatments In the algorithm above, we compute an expected group membership value, $\bar{z}_{ij,g}$, for each nucleotide j. These variables were later used to derive structural features for our applications in two different ways. The first way was to fit each dataset (V1 and S1) independently and treat the variable from each as a separate feature. Another way was to coordinately estimate the parameters for both datasets, with additional constraints imposed on the variables. The details are given in the Supplementary Materials. Here we outline the high-level ideas.

We tested two constraints based on two corresponding assumptions. The first assumption is that the two mixture components correspond to unpaired and paired bases, respectively. Because RNase V1 and S1 nuclease have opposite preferences for these two types of bases, we would expect paired bases to have higher V1 and lower S1 read counts than unpaired bases. If we define $g = 1$ and $g = 2$ as the groups of unpaired and paired bases, respectively, we would expect for any transcript i, $\mu_{i,1} < \mu_{i,2}$ for the V1 dataset and $\mu_{i,1} > \mu_{i,2}$ for the S1 dataset. In any EM iteration, we imposed these two conditions as a constraint by swapping the membership values of the two groups in a dataset if it was violated. We call this approach coordinated model fitting with opposite group memberships.

We also tested a different assumption, that in general nucleotides with more reads in one dataset would also have more reads in the other. This could be caused by the accessibility of nucleotides in the 3D structure. In this case, if $g = 1$ and $g = 2$ represent the groups of less accessible and more accessible nucleotides, respectively, $\mu_{i,1}$ should be smaller than $\mu_{i,2}$ in both datasets for any transcript i. Again, we set these as a constraint and swapped group memberships if violated. We call this approach coordinated model fitting with consistent group memberships.

For the mixture of Poisson model not considering local sequences, we fitted the models for the V1 and S1 datasets independently, and used their $\bar{z}_{ij,g}$ variables as two separate features. We compared these features with those from our mixture of Poisson linear models in predicting RNA zipcodes (see below). As the one-component Poisson linear model does not have group membership variables, we did not use it to predict RBP binding sites.

Table 1 summarizes the models we compare in this study.

2.4 Evaluation of fitness to the sequencing data

We measured the goodness-of-fit of a model by its R^2, which is defined as $1 - \frac{d}{d_0}$, where d and d_0 are the deviance of the model and the corresponding null model, respectively (Cameron and Windmeijer, 1996). The deviance compares the likelihood of a model with a full model where each observation has its own set of parameters. The formulas of R^2 for the various models and their derivations are given in the Supplementary Materials.

To compare the data fitness resulted from the different models, we used a 5-fold cross-validation procedure as follows. Each time, we used the sequences and observed read counts of four-fifth of the genes in our dataset to perform model fitting and get the optimized parameter values. We then used the model with these fitted values to predict the read counts of the remaining one-fifth of genes based on their sequences only. The predicted and actual read counts were then compared to compute the R^2 values. The cross-validation procedure ensures that the reported average accuracy reflects the ability of a model in capturing the general properties of RNA structures rather than over-fitting the training data, as the latter would result in low testing accuracy. With 94 962 nt from 119 genes, the dataset was large enough to ensure the robustness of the 5-fold cross-validation procedure. For the mixture of Poisson model that does not consider local sequences, we report the total R^2 from cross-validation by taking the sums of d and d_0 as the total deviances of the model and the corresponding null model from the five testing sets, respectively. Our algorithms in general return models that locally maximize the corresponding likelihood functions, which are related to, but are not equivalent to, the goodness-of-fit function.

2.5 List of RNA zipcodes

To compare the effectiveness of different types of features in predicting RNA zipcodes, we collected experimentally verified zipcodes from two

Table 1. List of statistical models for fitting structure-probing data

Abbreviation	Description
PL	One-component Poisson linear model (Li *et al.*, 2010)
MP	Mixture of Poisson model not considering local sequences
MPL[a]	Mixture of Poisson linear model, fitting V1 and S1 independently
MPLC same[a]	Mixture of Poisson linear model, fitting V1 and S1 coordinately with consistent group memberships
MPLC oppo[a]	Mixture of Poisson linear model, fitting V1 and S1 coordinately with opposite group memberships

[a]Models proposed in this work.

published studies (Jambhekar *et al.*, 2005; Olivier *et al.*, 2005). After quality control and intersecting with our structure-probing data (Supplementary Materials), we obtained a list of 10 zipcodes on 6 genes (Table 2).

2.6 List of protein-RNA binding sites

We further checked if our features can be used to predict general RBP binding sites. We obtained a whole-transcriptome dataset of RBP-binding regions in yeast (Freeberg *et al.*, 2013), produced by global photoactivatable-ribonucleoside-enhanced cross-linking and immunopurification. We filtered out fragmented regions and focused on those between 10 and 40 bases, which is a range shown to include the majority of binding sites (Freeberg *et al.*, 2013). The resulting list contains 42 344 RBP-binding regions on 2972 genes. The nucleotide composition of these regions is shown in Supplementary Table S3. Because small read counts are more affected by noise, we considered only transcripts with RPM (reads per million) >1000 when training and testing the prediction models (Supplementary Table S4). We also obtained a set of regions bound by the Puf3p protein from the same study, with 831 binding regions on 668 genes.

2.7 Machine learning and prediction procedures

For the prediction of RNA zipcodes, we defined positive examples as nucleotides within these zipcodes, and negative examples as all other nucleotides from the same RNAs. For each nucleotide, we derived various types of features of it (Table 3). PARS is the log ratio of V1 and S1 read counts (Kertesz *et al.*, 2010). PARS2 is the square of PARS, which indicates whether a nucleotide is clearly single-based or base-paired. LogVS includes the logarithm of the V1 and S1 read counts as two separate features. ProbVS contains the expected group membership variables from one of our Poisson linear models (to be specified below). For both the one-component model and two-component models that fit V1 and S1 either independently or coordinately, there are two features per nucleotide. PredSS2 is the probability for a nucleotide to be base-paired according to RNAfold (Hofacker *et al.*, 1994). PredSS3 is an extended version of PredSS2, with two different labels for bases at the 5′ end and 3′ end of a base pair. SeqBinary contains 4 binary features that correspond to whether the nucleotide is A, C, G or U. In addition to the features defined per nucleotide, we also considered some aggregated features for the whole window, including the nucleotide frequencies (SeqRatio), dinucleotide frequencies (SeqDiNu) and GC content (SeqGC).

When producing the ProbVS features using the Poisson linear models, there is a user parameter, K, that describes the number of nucleotides to consider around the current nucleotide during model fitting. We tried multiple values of it and compared the corresponding results.

Table 2. List of zipcodes used in our prediction task

Zipcode	Gene	Location in gene	Length	Source
E1min	Ash1	635–683	49	(Jambhekar *et al.*, 2005)
E2A	Ash1	1109–1185	77	(Olivier *et al.*, 2005)
E2Bmin	Ash1	1279–1314	36	(Jambhekar *et al.*, 2005)
Umin	Ash1	1766–1819	54	(Jambhekar *et al.*, 2005)
EAR1-1	Ear1	1572–1621	50	(Olivier *et al.*, 2005)
ERG2N	Erg2	180–250	71	(Jambhekar *et al.*, 2005)
SRL1C	Srl1	419–596	178	(Jambhekar *et al.*, 2005)
TPO1N	Tpo1	2–178	177	(Jambhekar *et al.*, 2005)
WSC2C	Wsc2	1354–1384	31	(Jambhekar *et al.*, 2005)
WSC2N	Wsc2	418–471	54	(Jambhekar *et al.*, 2005)

Table 3. List of features used in our prediction tasks

Feature type	Description	Number of features
PARS	PARS score	1 per nucleotide
PARS2	Square of PARS score	1 per nucleotide
LogVS	Logarithm of V1 and S1 counts	2 per nucleotide
ProbVS[a]	Expected group membership variables	2 per nucleotide
PredSS2	Predicted base-pair probability	2 per nucleotide (1 d.f.)
PredSS3	Predicted base-pair probabilities (with directions)	3 per nucleotide (2 d.f.)
SeqBinary	Binary encoding of the RNA sequence	4 per nucleotide (3 d.f.)
SeqRatio	Nucleotide frequencies	4 per window (3 d.f.)
SeqDiNu	Dinucleotide frequencies	16 per window (15 d.f.)
SeqGC	GC content	1 per window

[a]Features extracted by our proposed models; d.f.: degree of freedom.

For general RBP binding sites, due to the experimental procedure used to produce the dataset we used, the identified binding sites tend to contain a large fraction of Uracils (Ting Han, personal communication; see also the 'Results' section). Using a negative training set uniformly sampled from non-binding regions would result in models that use the enrichment of Uracils as a core predictor, which is not useful in distinguishing RBP binding sites from other regions with similar Uracil contents. To overcome this issue, we constructed the negative set by random permutations of the nucleotides of the positive examples, with a 1:5 ratio of positive to negative examples.

When predicting whether a nucleotide is within a RBP binding site, we used not only its features but also features of the nucleotides around it, to capture any local feature patterns. We denote by w the total number of nucleotides the features of which were considered when making predictions for a nucleotide. We tested multiple values of this window size w. When $w = 10$, we had 783, 8881 and 19 725 nt as positive examples in the prediction of RNA zipcodes, general RBP binding sites and binding sites of Puf3p, respectively.

We used Random Forest (Breiman, 2001) to perform training and predictions, based on the implementation in the *R* package *randomForest* (http://cran.r-project.org/web/packages/randomForest/index.html). We used the default values of all parameters, except that we set parameter

'sampsize' to 5000 when there were >5000 training data points, to reduce model training time.

Prediction performance was evaluated by cross-validation procedures and quantified by the area under the receiver operator characteristics (AUC), calculated by the R package ROCR (http://cran.r-project.org/web/packages/ROCR/index.html). For zipcodes, each time we kept the zipcodes of one gene for testing and used all the others to perform model training. The testing results were then combined to compute an overall AUC of the model. For transcriptome-wide protein binding sites, we performed 5-fold cross-validation with four-fifth data used as training and one-fifth for testing, for five disjoint random left-out sets of genes.

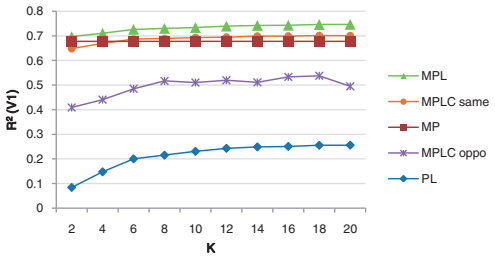

Fig. 1. Goodness of fit of the different models to the V1 read counts

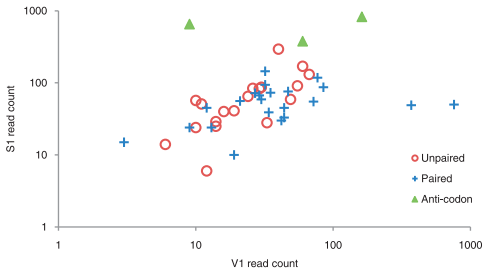

Fig. 2. Relationships between read counts and tertiary structures of tRNA. Positions with zero read counts due to non-unique read mapping are omitted

3 RESULTS

3.1 Goodness of fit of different models

To check if our proposed models are appropriate for structure-probing data, we computed R^2 at different values of K, the number of nucleotides around the target nucleotide considered by the models when predicting the read count of it. The results (Fig. 1 and Supplementary Fig. S2) show that for both V1 and S1 read counts, mixture models fit the data much better than the single-component Poisson linear models, even though the single-component Poisson linear models (PL) contain more parameters than the mixture models without considering local sequence biases (MP). Considering local sequences (MPL) provides some additional goodness of fit. The R^2 value of all models did not change much over a wide range of values of K. Comparing the two-component Poisson linear models when V1 and S1 data were fitted either independently or coordinately, the R^2 value was highest when the two sets of data were fitted independently (MPL), which is expected because there were no additional constraints imposed on the parameter values of the two models. Importantly, the R^2 value was only slightly dropped when the models were fitted coordinately with consistent group memberships (MPLC same). In contrast, the R^2 value was much lower when the membership variables of the two models were set to be opposite (MPLC oppo). These results suggest that besides expression levels, the most dominant factor that governs the read count of a nucleotide is likely something that stays the same in the two settings.

We hypothesized that one such factor is the accessibility of a nucleotide, which is related to the 3D structure of the RNA. Because 3D structures of mRNAs were largely unavailable in databases of molecular structures such as PDB (Berman *et al.*, 2000), we could not comprehensively test this hypothesis. Here we illustrate its possibility using tRNAs. We aligned all tRNAs with data in the structure-probing dataset to the full alignment of tRNAs in Rfam (Gardner *et al.*, 2011) (ID:RF00005). One of the sequences in our set had an exact match with the sequence in a structure in PDB (ID:486D—E). We took this structure, and calculated the solvent accessible surface area of each nucleotide using the POPS web server (Cavallo *et al.*, 2003). We found that the read counts from the V1 and S1 datasets were highly correlated (Fig. 2). In contrast, paired and unpaired nucleotides in general do not have significantly larger V1 and S1 read counts, respectively.

This example alone cannot prove that in general our extracted features are related to RNA 3D structures. Nonetheless, regardless of their exact biological interpretations, we found that our features are practically useful in predicting RBP binding sites, as shown below.

3.2 Using extracted features to predict RNA zipcodes

3.2.1 Number of bases required to capture sequence biases We first tested the use of our features in predicting RNA zipcodes. Our prediction framework involves two user parameters, namely K, the number of nucleotides considered in modeling read count biases due to local sequences, and w, the number of nucleotides the features of which to be considered in zipcode prediction. We first fixed w to two particular values (40 and 100), and compared the accuracy of our three two-component Poisson linear models at different values of K (Supplementary Fig. S3 and Fig. 3). For all three models, prediction accuracies were between AUC = 0.6 and 0.8, which are substantially higher than random predictions (AUC = 0.5). Consistent with the data fitting results, the models that fit V1 and S1 independently (MPL) or coordinately with consistent group memberships (MPLC same) were better than the one with opposite group memberships (MPLC oppo) in identifying zipcode regions. As the accuracy did not change much with different values of K, we fixed $K = 2$ for the remaining tests to minimize program execution time.

3.2.2 Number of features required to predict zipcodes In Supplementary Figure S3 and Figure 3, we see that prediction accuracies were higher for $w = 100$ than $w = 40$. To see if it is generally true that a large value of w is needed for accurate prediction of zipcodes, we compared the performance of our two-component Poisson linear model with separate V1 and S1 fittings at various values of w. We observed increasing prediction accuracy until around $w = 80$ (the ProbVS curve in Fig. 4), which

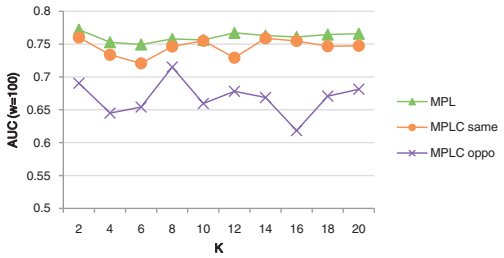

Fig. 3. Accuracy of the features extracted from our two-component Poisson linear models in predicting RNA zipcodes with respect to different values of K, number of nucleotides considered in modeling read count biases due to local sequences when w is fixed to 100

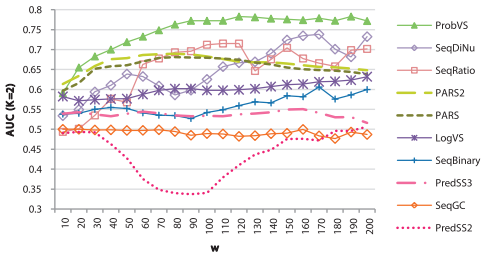

Fig. 4. Accuracy of various types of features in predicting RNA zipcodes, at different values of w, the number of neighboring nucleotides to be considered when classifying a target nucleotide

matches the average length of the zipcodes in our set of examples (Table 2).

3.2.3 Comparing different types of features We next compared the ability of different types of features in identifying zipcodes (Fig. 4). Among the structural features compared, those obtained from structure-probing sequencing data (ProbVS, LogVS, PARS and PARS2) achieved higher accuracy than those from computational predictions alone (PredSS2 and PredSS3). Within the former group, the features extracted by our two-component Poisson linear model (ProbVS) produced the highest accuracy for almost all values of w tested. Nucleotide and dinucleotide frequencies (SeqRatio and SeqDiNu) worked fairly well with large w values, but were not as strong as our structural features (ProbVS). The other sequence features (SeqBinary and SeqGC) performed poorly. We have also devised a method to formally quantify the amount of uncertainty reduced from random predictions by each set of features, and observed the same trend as these AUC values (Supplementary Fig. S4).

As an example, Supplementary Figure S5 shows the V1 and S1 read counts along the SRL1 RNA, and the probability for each nucleotide to be within a zipcode as predicted by three models. It is seen that the actual zipcode region SRL1C is not particularly single-stranded or double-stranded according to the V1 and S1 read counts, but has a high count of both in general. Our structural features were able to capture this trend and identify the zipcode with high accuracy.

3.3 Whole-genome prediction of RBP binding sites

We then further tested if the same approach could predict general RBP binding sites on RNAs. We first checked the nucleotide composition of the RBP-bound regions in our dataset, and as expected found a higher fraction of Uracils in the bound regions as compared with the overall composition of the RNAs (Supplementary Table S3), which supports our use of the negative set with the same nucleotide composition as the positive examples (see 'Materials and Methods' section).

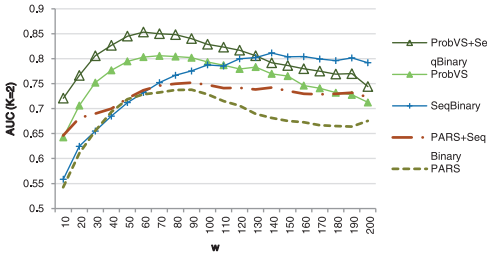

Fig. 5. Cross-validation accuracy of sequence and structural features in predicting general protein binding sites on highly expressed RNAs. ProbVS was based on independent fitting of V1 and S1 data

Figure 5 shows the cross-validation results of sequence features and the two of the best structural features in zipcode predictions. The positive and negative examples could well be separated by our features extracted from structure-probing data, with a top AUC of 0.8. PARS also achieved an AUC of close to 0.75. Interestingly, the accuracy of the model with sequence features alone increased steadily as w increased, until reaching a peak AUC of ~0.8 at $w = 140$. It thus appears that general RBP binding sites may contain some complex sequence patterns at the flanking regions. We retrained our prediction models using both sequence and structural features at the same time, and found that the resulting accuracy was improved for both our extracted features and the PARS scores, reaching a top AUC of ~0.85 for ProbVS + SeqBinary. We also predicted binding sites of Puf3p, an RBP with data available from the same study, and observed similar trends (Supplementary Fig. S6).

4 DISCUSSION

4.1 Information contained in structure-probing data

In this study, we have shown that read counts and derived quantities obtained from structure-probing data are affected by a number of factors, including expression levels, cleavage preference of the enzymes involved and biases due to local sequences.

We found that a mixture model provided substantially better goodness-of-fit to the structure-probing data we studied than a single-component model. The components we identified did not correspond well to paired and unpaired bases, as indicated by a smaller R^2 value when group memberships were set to be opposite for the two sets of data than when they were set to be consistent. Instead, we hypothesize that the components may better reflect local accessibility of individual bases in the 3D structure. New experimental data and analyses are required to prove this hypothesis.

We have also shown there is a clear difference in terms of both data fitness and prediction accuracy of zipcodes between models that consider local sequences and those that do not.

Taking these results together, we conclude that structure-probing data need to be carefully processed to extract useful features. Taking simple ratios of the read counts from two different enzymatic treatments or between an experiment with the corresponding control in a nucleotide-by-nucleotide manner could eliminate factors that stay largely the same in the experiments being compared, but at the same time some useful information may also be removed, such as the solvent accessibility of each nucleotide.

4.2 Signals for recognizing RNA zipcodes

We have shown that structural features were able to identify RNA zipcodes with high accuracy. The prediction models were most accurate when the features of a large number of (~80) nucleotides were used. Unlike transcription factors that bind DNA with strong sequence motifs, sequence signatures proposed for RNA zipcodes have not been able to provide a complete model (Shepard *et al*., 2003). Our results suggest that the recognition of RNA zipcodes by the She complex may involve more complex features from the secondary and tertiary structures of RNAs. The large number of nucleotides needed for strong prediction suggests that recognition may be mediated by a large amount of weak signals.

4.3 General features of protein-RNA binding sites

The encouraging performance of our models in identifying general protein binding sites suggests that there are some general features recognized by different RBPs. A next step is to test whether a small set of common features is shared by most RBPs, or there exist different classes of RBPs each recognizing different features.

We found that both sequence and structural features could predict general protein binding sites with high accuracy. The exact relationship between these two types of features is still not clear. New insights are needed to elucidate how proteins interact with both RNA sequences and structures.

ACKNOWLEDGEMENTS

We would like to thank John Kim, Mallory Freeberg and Ting Han for sharing their unpublished data with us and for their helpful comments on our work.

Funding: X.H. and K.Y.Y. are partially supported by the Direct Grant for Research 2050479 of The Chinese University of Hong Kong.

Conflict of Interest: none declared.

REFERENCES

Aviran,S. *et al.* (2011) Modeling and automation of sequencing-based characterization of RNA structure. *Proc. Natl Acad. Sci. USA*, **108**, 11069–11074.
Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
Cameron,A.C. and Windmeijer,F.A.G. (1996) R-squared measures for count data regression models with applications to healthcare utilization. *J. Bus. Econ. Stat.*, **14**, 209–220.
Cavallo,L. *et al.* (2003) POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.*, **31**, 3364–3366.
Chartrand,P. *et al.* (1999) Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle *in vivo*. *Curr. Biol.*, **9**, 333–338.
Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–38.
Dohm,J.C. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
Freeberg,M.A. *et al.* (2013) Pervasive and dynamic protein binding sites of the mrna transcriptome in Saccharomyces cerevisiae. *Genome Biol.*, **14**, R13.
Gardner,P.P. *et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.
Gonzalez,I. *et al.* (1999) ASH1 mRNA localization in yeast involves multiple secondary structural elements and Ash1 protein translation. *Curr. Biol.*, **9**, 337–340.
Hafner,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and MicroRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, **125**, 167–188.
Jambhekar,A. *et al.* (2005) Unbiased selection of localization elements reveals cis-acting determinants of mRNA bud localization in Saccharomyces cerevisiae. *Proc. Natl Acad. Sci. USA*, **102**, 18005–18010.
Kertesz,M. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.
Li,J. *et al.* (2010) Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biol.*, **11**, R50.
Licatalosi,D.D. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
Lucks,J.B. *et al.* (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq). *Proc. Natl Acad. Sci. USA*, **108**, 11063–11068.
Mortimer,S.A. and Weeks,K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.*, **129**, 4144–4145.
Olivier,C. *et al.* (2005) Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol. Cell. Biol.*, **25**, 4752–4766.
Pancaldi,V. and Bähler,J. (2011) In silico characterization and prediction of global protein-mRNA interactions in yeast. *Nucleic Acids Res.*, **39**, 5826–5836.
Sanford,J.R. *et al.* (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.*, **19**, 381–394.
Shepard,K.A. *et al.* (2003) Widespread cytoplasmic mRNA transport in yeast: Identification of 22 bud-localized transcripts using DNA microarray analysis. *Proc. Natl Acad. Sci. USA*, **100**, 11429–11434.
Underwood,J.G. *et al.* (2010) FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Mehtods*, **7**, 995–1001.
Zhao,J. *et al.* (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, **40**, 939–953.