8862–8873 *Nucleic Acids Research, 2012, Vol. 40, No. 18* *Published online 20 July 2012*
doi:10.1093/nar/gks684

Computational identification of new structured *cis*-regulatory elements in the 3′-untranslated region of human protein coding genes

Xiaowei Sylvia Chen* and Chris M. Brown

Department of Biochemistry and Genetics Otago, University of Otago, Dunedin 9054, New Zealand

Received April 19, 2012; Revised June 15, 2012; Accepted June 20, 2012

ABSTRACT

Messenger ribonucleic acids (RNAs) contain a large number of *cis*-regulatory RNA elements that function in many types of post-transcriptional regulation. These *cis*-regulatory elements are often characterized by conserved structures and/or sequences. Although some classes are well known, given the wide range of RNA-interacting proteins in eukaryotes, it is likely that many new classes of *cis*-regulatory elements are yet to be discovered. An approach to this is to use computational methods that have the advantage of analysing genomic data, particularly comparative data on a large scale. In this study, a set of structural discovery algorithms was applied followed by support vector machine (SVM) classification. We trained a new classification model (CisRNA-SVM) on a set of known structured *cis*-regulatory elements from 3′-untranslated regions (UTRs) and successfully distinguished these and groups of *cis*-regulatory elements not been strained on from control genomic and shuffled sequences. The new method outperformed previous methods in classification of *cis*-regulatory RNA elements. This model was then used to predict new elements from cross-species conserved regions of human 3′-UTRs. Clustering of these elements identified new classes of potential *cis*-regulatory elements. The model, training and testing sets and novel human predictions are available at: http://mRNA.otago.ac.nz/CisRNA-SVM.

INTRODUCTION

The translation of messenger ribonucleic acid (mRNA) is a precisely regulated process. There are a large number of *cis*-regulatory elements in mRNAs, most of them found in untranslated regions (UTRs). Both the 5′-UTR and 3′-UTR contain a multitude of translational control elements, many of which are structured (1–5). The 5′-UTR elements generally affect translational initiation, for example, upstream initiation codons and internal ribosomal entry sites (IRESs) (6). A more diverse range of regulatory mechanisms are associated with the 3′-UTR (4,5). Translational activation (7,8), repression (9), alternative splicing (10), mRNA stability (11) and localization (12) can all be controlled through *cis*-regulatory RNA elements in the 3′-UTR.

Some *cis*-regulatory elements on 3′-UTRs function at the primary sequence level and are included in several databases of regulatory RNA motifs (4,5,13,14). Some common elements include AU-rich (ARE) (15), CU-rich (16) or GU-rich (17) elements recruiting specific protein factors, which stabilize or destabilize the mRNA.

RNA structure has also been shown to play an important role in the function of some *cis*-regulatory elements. In viruses, complex RNA structures are frequently present in 3′-UTRs and are a common feature of translational control (18–20). In cellular genomes, there are diverse structured elements. There is evidence that a 100 base structure localizes the human vimentin mRNA, whereas a small and less thermodynamically stable AU-rich stem loop has been identified in the 3′-UTR of human Myc mRNA, that is likely to be the signal for the perinuclear localization of that mRNA (21). Some *cis*-regulatory elements are characterized by distinct secondary structures, such as the iron-responsive element (22), the gamma interferon inhibitor of translation (GAIT) element (23) and many large mRNA localization elements in *Drosophila* (24–27). A recent study also identified frequent small structural elements associated with stability of mammalian mRNAs (28). However, structures of *cis*-regulatory RNAs are often less conserved across distantly related species in comparison with non-coding RNAs.

*To whom correspondence should be addressed. Tel: +64 3 4797875; Fax: +64 3 4795201; Email: sylvia.x.chen@gmail.com

Translational control through *cis*-regulatory elements is most frequently associated with genes that require precise temporal and/or spatial regulation. In humans, localized translation is important for the function of certain mRNAs, including some transcription factors (21,29) and numerous neuronal mRNAs (30). Disease-associated mRNA localization mechanisms have also been characterized in humans (31,32). Extensive experimental studies have revealed localized mRNAs at genomic scale (30,33,34). However, despite the identification of numerous localized RNAs, the *cis*-regulatory elements, which potentially regulate localization and other post-transcriptional events, remain largely unknown. Gene ontological (GO) studies have revealed that, common *cis*-regulatory elements such as the ARE and GU-rich elements are over-represented in the 3′-UTRs of unstable mRNAs, and some of them have conserved secondary structures (21,29,35). Notably, genes involved in transcription regulation, cell cycle, apoptosis and RNA processing (36) are often tightly regulated to ensure a rapid response of transcription and translation in a given time and place of the cell. Therefore, it is likely that currently unknown regulatory RNA elements exist in these genes. On the basis of known elements, we expect that there are a large number of structured *cis*-regulatory RNA elements yet to be identified.

Various computational methods have been developed for RNA structural analysis. Folding a single sequence based on thermodynamic stability does not have enough specificity for genome-wide scans due to high false-positive rates. Therefore, genome scans for structural RNAs are generally performed on alignments. There are three major approaches to predict structurally conserved RNA elements (37,38): align-then-fold, fold-then-align and simultaneously fold-and-align. The first approach is the least computationally intensive; however, it relies highly on the quality of the sequence alignment and is not able to predict structures from alignments with low sequence similarities. On the contrary, the second approach eliminates the requirement of a prior sequence alignment and looks for the most stable common structure. However, this approach relies on the accuracy of secondary structure predictions of each RNA sequence. The third approach often uses a dynamic programming algorithm first introduced by Sankoff (39), which was implemented in a number of programs to look for common structures among sequences with low similarity, but this approach is often very computationally intensive.

A number of programs have been applied to genome-wide identification of novel structured RNAs (40,41), such as RNAz (42), RNA sampler (43), Evofold (44), RSmatch (45), Foldalign (46), CMfinder (47), RNApromo (48) and LocARNA (49). Given the complexity and variety of structured RNAs in a genome, the performance of different approaches differs considerably (43,45,47). Few studies have incorporated machine-learning algorithms such as support vector machine (SVM) or neural network to search for specific types of RNA structures with increased sensitivities and specificities (43,50,51).

In this study, we developed a new approach combining several different RNA structural prediction strategies, with a SVM classifier trained specifically for *cis*-regulatory RNAs, rather than ncRNA genes. Our new model CisRNA-SVM is able to distinguish *cis*-regulatory RNAs from random genomic sequences or shuffled sequences with higher sensitivities than existing methods. Combined with a comparative genomics approach, we then used our new model to search for novel structured *cis*-regulatory elements in the 3′-UTRs of human genes. Results are validated by the identification of known *cis*-elements, such as the selenocysteine insertion sequence (SECIS) element, iron responsive element (IRE) and histone 3′-UTR. New classes of novel structures are found by subsequent clustering analysis.

MATERIALS AND METHODS

Data sets for training and testing

The positive training alignments are the seed alignments of 98 filtered Rfam *cis*-regulatory elements (CisReg A, mRNA.otago.ac.nz/CisRNA-SVM) (52). Three sets of sub-alignments with different optimal and maximal pairwise identities were selected from each seed alignment using the rnazSelectSeq.pl tool from RNAz 2.0 package (50). The three sets are (i) optimal 50%, max 65%; (ii) optimal 70%, max 85% and (iii) optimal 85%, max 90%. This generated a total of 294 alignments, of which 233 were unique and used for training. The negative training set consists of random genomic alignments or dinucleotide shuffled CisReg A alignments. The negative control genomic alignments were obtained from 500 sequences immediately downstream (after the polyadenylation site) of protein-coding genes on chromosome 21 from University of California, Santa Cruz (UCSC) hg19 assembly (53). They were divided into 120-nt windows (40-nt sliding) resulting in 1488 alignments, 638 of which were included in the negative training set and the remaining 850 used in testing. Dinucleotide shuffled CisReg A set was generated using alignment-shuffling algorithm, Multiperm (54). Each of the 294 positive training alignments was shuffled twice, and one set was used in training, the other testing. In addition, a testing set consisting of a second group of 127 Rfam *cis*-regulatory elements was also used. This set contains additional *cis* regulatory elements including large 5′-IRES's (mainly viral) riboswitches (mainly bacterial 5′) and pseudoknots (mainly viral and coding; CisReg B, mRNA.otago.ac.nz/CisRNA-SVM). Alignments were selected from the CisReg B seed alignments with optimally 60% and maximum 80% pairwise identity.

Alignments of 3′-UTR sequences

The initial 3′-UTR alignments were obtained from the TargetScan database (www.targetscan.org). These had been extracted from the UCSC genome alignments. Species with almost identical sequences to the common model organisms were removed (i.e. chimpanzee and rat), and from the remaining species, only the ones have higher sequence coverage were selected for our analysis. In

total, 10 vertebrate species were included in the initial alignments (human, mouse, dog, horse, cow, opossum, platypus, lizard, chicken and frog). The alignments were converted to Clustalw format and separated into 120-nt/80-nt sliding windows using rnazWindow.pl tool from RNAz 2.0 package (50) to obtain six-sequence alignments with the human sequence in each alignment as the reference sequence.

Alignments and scores

Four programs were chosen for the training process: RNAalifold from the ViennaRNA-1.8.4 package (55), LocARNA-1.5.2 (49), Foldalign-2.1.1 (56) and Cove (57). They represented the three alignment strategies: align-then-fold, fold-then-align, simultaneous fold-and-align and covariance-based methods. Cove has been developed into the package Infernal (58), and this now requires an initial seed alignment, therefore is not suitable for a *de novo* search in unaligned sequence sets as in our study. However, the Infernal cmsearch tool was used later as a testing method for confirming the presence of known RNA elements. In addition, we assumed that all sequences in the training and testing sets were either definitely positive or negative; therefore, we did not use a more recent covariance-based *de novo* motif search tool CMfinder (47), which selects candidates with stable structures before an iterative search. This feature of CMfinder resulted in many known positive training alignments not producing models. Cove is a well-established method and is widely used in searching for small RNA structures, such as its implementation in tRNAscan-SE (59).

Feature scores were obtained from each of the four programs. Clustalw alignments were used for RNAalifold input, and the program was run with—noLP option to suppress single base-pair stems. For LocARNA, the multiple alignment version mlocarna was used. The feature scores for both RNAalifold and LocARNA are the minimum free energy of the aligned structure. Foldalign generated all-against-all pairwise alignments of the input multi-fasta alignment, and the average score was used as the feature score. The program coveb from the Cove-2.4.2 package was used, and the average bit score was taken as the feature score. Finally, the scores from all the programs were normalized by the length of alignment. The program APSI.java (www.mybiosoftware.com/tag/apsi) was used to compute the average pairwise sequence identity (APSI). Scripts giving the details of score processing can be found at mRNA.otago.ac.nz/CisRNA-SVM.

Support vector machine

The SVM classification model was made using Libsvm 2.9 (60) using SVM classification with radial basis function. The 233 unique alignments generated from the CisReg A set were positive instances, and the combined genomic (638) and shuffled alignments (294) were negative instances. The feature scores, GC content and APSI were scaled with set ranges, and a standard grid search and 5-fold cross-validation were performed for each model training procedure. As the ratio of positive training set to the negative training set was 1:4, a weighting ratio of 4:1 of positive:negative was initially used in all the training procedures. First, each feature score plus GC content and APSI was trained as individual SVM classification models to test the classification potential of the particular feature score. Then, the combination of all four structural alignment programs was tested against models with one program removed. Finally, scores that contributed positively in overall classification were incorporated into the final model. All the models were tested on three testing sets: CisReg B alignments, 850 different genomic alignments and the other set of 294 shuffled CisReg A.

Analysis of putative *cis*-regulatory elements

The predicted positive alignments were first combined to obtain non-overlapping alignments, and the human sequences from all alignments were extracted. Two sets of resultant sequences were obtained with SVM prediction probability ($P > 0.5$ and $P > 0.9$). The sequences were then analysed with Infernal 1.0 cmsearch (58) to test for the presence of known RNA elements. The result sequences with $P > 0.9$ were further analysed for potential new clusters of structures. Functional annotation with association to ontology was performed using DAVID bioinformatics resources 6.7 (61).

For clustering analysis, non-redundant human sequences were locally folded with RNALfold (55), and the sequence regions corresponding to the most stable local structures were extracted. The sequence regions were then aligned using MUSCLE (62), and the phylogenetic tree generated during the second iteration was used to obtain clusters of sequences with sufficient similarity. The tree was converted into an R-hclust object and subsequently cut at various depths to generate clusters. The clusters were realigned using MAFFT (63), option '% mafft-qinsi' and sorted based on average pairwise sequence identity. The final cutting depth used was determined empirically by the number of clusters with average pairwise identity greater than 0.5 that retained the positive control Histone 3′-element cluster.

RESULTS

Classification of true-positive *cis*-regulatory elements by the individual methods

The goal of our method was to distinguish *cis*-regulatory structures from both a shuffled background and other potential structures in random genomic regions. This tool could then be used to scan genomes. We did not require accurate structural predictions for this process. To make a model for the identification of new *cis*-regulatory RNA elements, a training set of known *cis*-regulatory elements were manually selected from the Rfam database (52,64). They all fulfilled the following criteria: had only one experimentally verified secondary structure, did not contain pseudoknots and did not have large complex structures. The resulting 98 elements (named CisReg A) are exclusive of riboswitches, pseudoknots and internal ribosome entry sites (IRESs). The set of 127 *cis*-regulatory elements excluded from set A was retained as a testing

set (CisReg B). Only well-characterized seed alignments from Rfam were used in this study. The negative training and testing sets were generated from genomic regions following 3′-UTRs and dinucleotide shuffled CisReg A alignments (see 'Materials and Methods' section).

Four RNA structural identification programs were selected to represent the three major strategies in searching for conserved RNA structures: 'align-then-fold', 'fold-then-align' and 'simultaneous fold-and-align'. The selected programs also represent several commonly used structural prediction approaches, including thermodynamic energy minimization combined with evolutionary conservation, the Sankoff algorithm and expectation maximization with covariance models. The selection was based on several criteria: the program is well established in the field, a numeric score is produced with the result, the program generates results on all the training and testing alignments and the program does not internally eliminate part of the training sequences based on rules regarding thermodynamic stability. Because of the large number of RNA structural prediction programs available, it is not feasible to test all of them. Therefore, based on the aforementioned criteria, we have chosen RNAalifold (55), LocARNA (49) and Foldalign (56) and Cove (57).

We used normalized scores in all the subsequent analyses (as described in 'Materials and Methods' section). First, RNAalifold, LocARNA, Foldalign and Cove were each tested for their ability to distinguish CisReg A set from the same set dinucleotide shuffled. Each program was run with the positive and negative data sets. Each program alone had only moderately different scores on the two data sets (Supplementary Figure S1). We then tested whether each program could better classify elements if two other general features were also used, the average GC content and average pairwise sequence identity (APSI). The three scores for each were input into a SVM (see 'Materials and Methods' section).

Including both GC and APSI made a major contribution to the classification success of the four programs (Supplementary Figure S2). Therefore, they were incorporated into the SVM in all the following analyses. Figure 1 shows the improved sensitivity and specificity of the additional scores and SVM method compared with each program on their own. The classification ability of each SVM model, as measured by the percentage of alignments classified as positive for each data set is summarized in Table 1.

CisRNA-SVM: a new classification model

We then tested to see whether combining scores from all four programs in one SVM classifier could improve the classification. First, the feature scores of all four programs were used as attributes in SVM training/testing, and subsequently four SVM models were tested with one attribute (score from each of the four programs) removed from the total SVM attributes. The sensitivity and specificity of the final SVM model and those of each program with the SVM classifier are shown in Figure 2, and the corresponding percentage of positive classification is listed in Table 2. The model (CisRNA-SVM) combining of all four programs had the overall best performance. It correctly classified 91.8% of the CisReg A set, whereas false positively classifying just 0.9% of the shuffled set.

Tenfold cross-validation of this model indicated 92.4% accuracy. Notably, the CisRNA-SVM model also performed well in classifying the CisReg B set, which had rather different structural properties compared with the CisReg A set, indicating the model's strong potential for identifying new *cis*-regulatory elements not seen in the training set. In summary, our method classified 91.8% CisReg A alignments and 84.3% CisReg B alignments as positive, which corresponded to 96 of 98 CisReg A families and 108 of 127 CisReg B families.

The performance of the model was also tested on automatically processed 3′-UTR alignments containing either complete or partial known *cis*-regulatory elements. We selected a number of elements for this test: SECIS (RF00031), Histone3 (RF00032), IRE (RF00037), Vimentin3 (RF00109), GAIT (RF00179), G-CSF_SLIDE (RF00183) and DNA polymerase beta (DPB) (RF01455). All elements are also in CisReg A, but the training set does not always contain the human or mammalian sequences (e.g. Histone3), and their training seed alignments had length ranging between 26 nt and 100 nt. These elements have all been found in human 3′-UTRs. A number of vertebrate full-length 3′-UTR alignments (3 selenoprotein mRNAs, 3 histone mRNAs, the transferring receptor C mRNA, the vimentin mRNA, the ceruloplasmin mRNA and the deoxyribonucleic acid polymerase II mRNA) known to contain these elements were randomly selected and processed into 120-nt windows. In the resulting windows, most elements were wholly contained within a window, whereas a few windows had only partial elements. Figure 3 compared the SVM decision values of the training seed alignments and those of the testing UTR windows. Overall, the decision values for UTR windows were lower than the training seed alignments. This result was expected as the UTR windows had either flanking sequence outside the element or they only contained part of the element. In four of the seven cases, at least one element was predicted. However, this does indicate a loss in sensitivity on the automatically processed 3′-UTR alignments.

Genome-wide scans of UTRs for regions containing putative new *cis*-regulatory RNA elements

In total, 154 803 sub alignments with length of 120 nt were processed from vertebrate 3′-UTR alignments. Each alignment had the human sequence as the reference and contained five other vertebrate sequences selected based on their pairwise identity to the human sequence (optimal sequence identity of 70%; see section 'Materials and Methods'). Each alignment was analysed by the four programs to obtain the feature score, which were then used as input for SVM classification using the CisRNA-SVM model. The probability of each alignment being positive (Pp) was used as prediction result. By default ($Pp > 0.5$), 32 043 alignments were classified as positive and were merged to give 22 038 non-overlapping

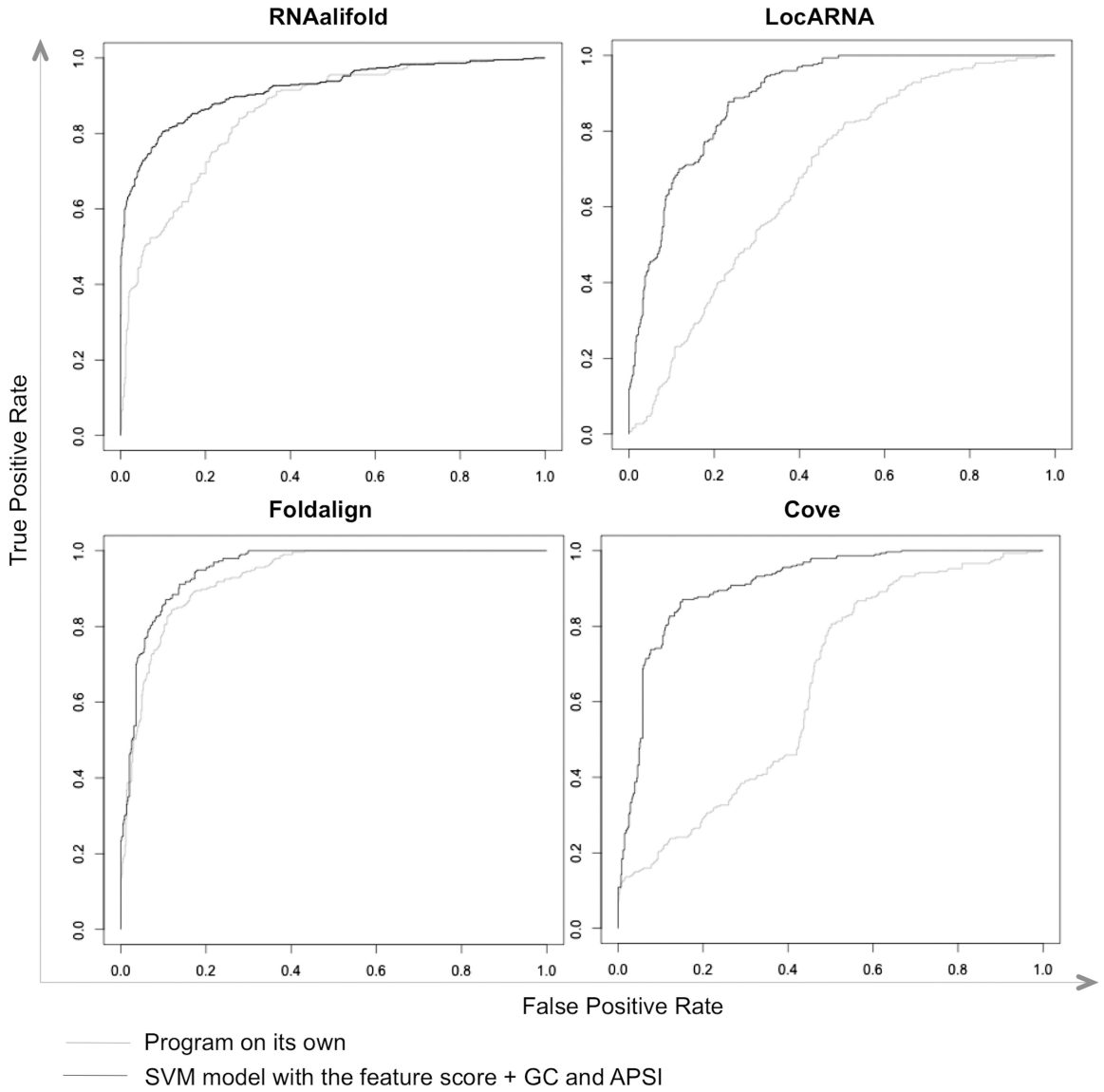

Figure 1. Performance of each structural prediction program with and without GC and APSI and an SVM classifier tested on the CisReg A and its control shuffled data set. The performance was measured by the true-positive rate versus false-positive rate on each data set using the R package ROCR. The average normalized score of alignments produced by running each program was used to compare with the SVM decision value produced by each SVM classifier for their ability of distinguish the CisReg A alignments and shuffled alignments preserving the dinucleotide frequencies.

Table 1. Performance of various SVM models measured by the percentage of alignments classified as positive

	CisReg A (%)	CisReg B (%)	Genomic (%)	Shuffled (%)
RNAalifold (R)	66.7	55.1	1.3	2.9
LocARNA (L)	54.8	41.7	1.5	8
Foldalign (F)	70.1	48	3.8	3.8
Cove (C)	70.1	54.3	5.8	6.3
R + L + F	73.5	61.4	1.4	3.1
R + L + C	89.1	75.6	3.6	0.9
R + F + C	91.8	77.2	2.9	1.2
L + F + C	85.7	72.4	3.4	0.7
CisRNA-SVM	91.8	84.3	3.8	0.9

alignments. Different *P* value cutoffs may be set to get higher confidence levels of prediction. At $Pp > 0.9$, there were only 4887 positive alignments, which when merged gave 4424 non-overlapping alignments. Given the false-positive rates obtained from SVM model training, at $Pp > 0.9$, the false-positive rates based on either shuffled alignments or genomic alignments were reduced to 0% and 0.6%, respectively.

Human sequences from the above two sets of alignments were extracted to be analysed for the presence of known structured 3′-UTR elements. First, all positively predicted sequences ($Pp > 0.5$) were searched using the models in the Rfam database 10.0 (64) using cmsearch (58) and compared with the total number of known elements in Rfam. Results were shown in

Supplementary Table S1. In total, the CisRNA-SVM method identified 39 of 62 histone 3′-UTR elements, 3 of 7 IREs and 7 of 15 SECIS elements. We also used the TransTerm set of *cis*-regulatory elements (5), and the result was similar. TransTerm also contains unstructured and non-Watson Crick structures, and our hits also included non-canonically structured G-quartet, as well as several apparently unstructured elements, such as the ARE and CU-rich elements.

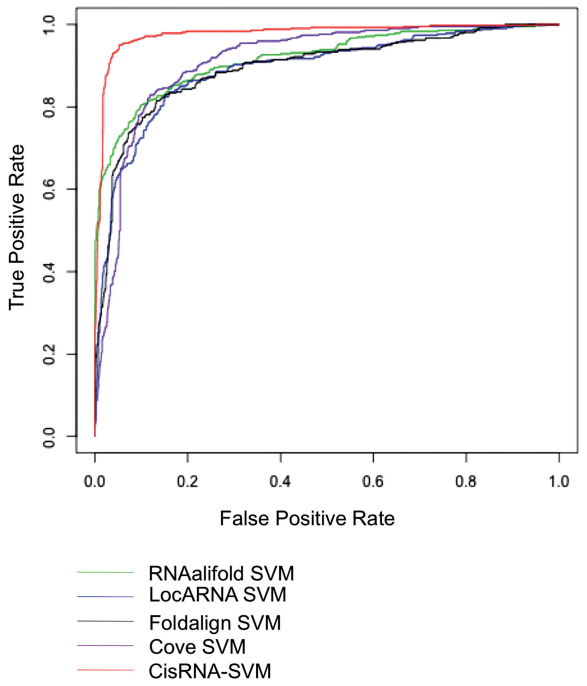

Figure 2. Performance of the combined CisRNA-SVM classifier versus the classifiers of individual programs. The true-positive rate and false-positive rate were calculated using the R package ROCR. In this test, four data sets were used: CisRegA (as positive), CisReg B (as positive), dinucleotide shuffled CisReg A (as negative) and genomic alignments downstream of protein coding genes (as negative). The training and testing of SVM classifiers were described in 'Materials and Methods' section. The classifiers based on individual programs were built with the feature score plus GC and APSI, and CisRNA-SVM was built with feature scores of all four programs, plus GC and APSI.

The genomic positions of the human predicted structured *cis*-elements are available at http://mRNA.otago.ac.nz/CisRNA-SVM. The sequences corresponding to Pp > 0.9 result set were analysed for local RNA structures using RNALfold from Vienna RNA package (55), and several top ranked stable structures based on minimum free energy were shown in Figure 4. The most stable local structures for all sequences are available at mRNA.otago.ac.nz/CisRNA-SVM.

To address the question as to weather these elements were found in particular classes of genes, for example highly regulated mRNAs, or placed across all classes of mRNAs, we mapped the genes to GO terms (see 'Materials and Methods' section). Table 2 lists the top functional groups for each main GO category. These results indicate a striking enrichment of structural elements in the 3′-UTRs of mRNAs encoding proteins involved in ion binding (24.6% of the total elements with *P* value 1.06E-15) and are components of the plasma membrane (13%, *P* value 2E-12). A significant proportion of the elements were found in the mRNAs of protein involved in transcription (12.3%, *P* value 2.54E-7), as well as cell cycle regulation, transport and localization.

New classes of structured *cis*-elements

Given the relatively small number of known *cis*-regulatory elements, mainly within a few classes, most of our predictions do not match known RNA classes. To further investigate the potential functional relations between these *cis*-elements, we applied a clustering method using

Table 2. Functional classification of genes containing predicted *cis*-regulatory elements with high confidence ($P > 0.9$)

Molecular function			Biological process			Cellular component		
Category	Percentage	*P*	Category	Percentage	*P*	Category	Percentage	*P*
GO:0043177 (ion binding)	24.6	1.06E-15	GO:0006350 (transcription)	12.3	2.54E-7	GO:0044459 (plasma membrane part)	13	2E-12
GO:0046872 (metal ion binding)	24	2.46E-15	GO:0046907 (intracellular transport)	4	1.37E-5	GO:0012505 (endomembrane system)	4.9	6.92E-9
GO:0043169 (cation binding)	24.2	3.95E-15	GO:0010942 (positive regulation of cell death)	2.7	1.54E-4	GO:0031226 (intrinsic to plasma membrane)	7.2	2.27E-8
GO:0046914 (transition metal ion binding)	16.4	1.46E-12	GO:0022403 (cell cycle phase)	2.6	1.71E-4	GO:0005887 (integral to plasma membrane)	7.1	2.66E-8
GO:0008270 (zinc ion binding)	13.7	5.01E-12	GO:0034613 (cellular protein localization)	2.6	1.94E-4	GO:0031947 (membrane-enclosed lumen)	10.6	3.3E-6

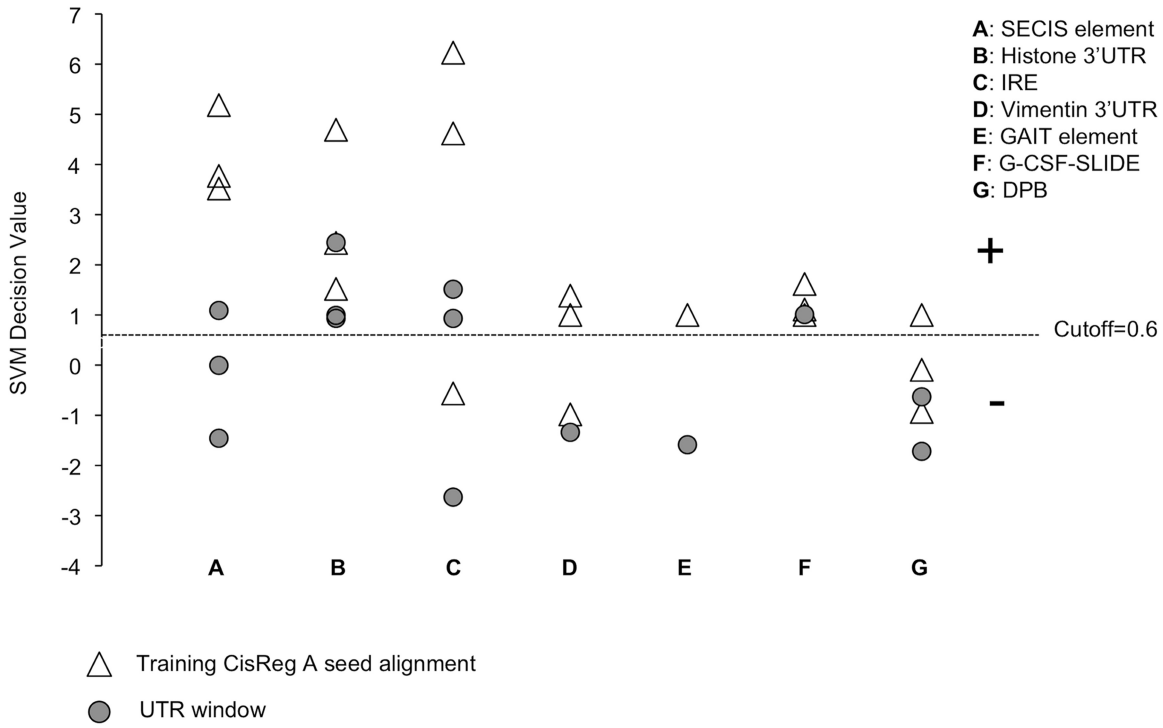

Figure 3. Performance of CisRNA-SVM on randomly selected UTRs containing known elements. A number of vertebrate UTR alignments were randomly selected to test the performance of CisRNA-SVM model on known elements with flanking sequences. The UTR alignments were sliced into 120 nt windows with 40 nt overlapping sequences. The windows containing known elements were three windows with1 SE*CIS* element each, three windows with 1 Histone 3'UTR, three windows with IRE elements (two with two IRE each and one with one IRE), one with Vimentin 3'UTR, one with GAIT, one with G-CSF-SLIDE and two with DPB. The triangles indicated the SVM decision values of the training seed alignments with the three different sequence identity ranges, as described in 'Materials and Methods' section. The solid circles indicated the SVM decision values of the UTR windows containing the known elements.

pairwise distances to search for new groups of *cis*-regulatory RNA motifs.

Clustering analysis was performed on the human sequences of the high confidence set ($Pp > 0.9$). First, highly similar sequences (with identify >95%) were removed, resulting in 4409 non-redundant sequences. Regions of all the sequences containing the thermodynamically most stable structures were extracted, and the sequences were aligned using MUSCLE (62), and a phylogenetic tree was generated based on pairwise sequence distances. The resulting tree successfully generated one cluster including the known Histone 3′-UTR elements. The cluster was shown in Supplementary Figure S3. The tree was subsequently trimmed at various depths to obtain clusters of potential new elements.

To determine a feasible trimming depth, the Histone 3′-UTR cluster was used as a control. Several trimmings of the tree were tested, and the final cutting depth was chosen, so that the histone 3′-UTR cluster remained intact, while resulting in the maximum number of clusters with pairwise sequence identity > 0.5. In total, the trimming resulted in 4059 leaves, which corresponded to 134 clusters containing 2 sequences and 75 clusters containing 3 or more sequences. Figure 5 shows the structures of several large clusters with six or more sequences. The known histone 3′-UTR cluster, which also belonged to this group, is not shown.

Cluster (a) mostly contains genes involved in signal transduction, and they form a highly conserved AU-rich single stem-loop structure. AU-rich stem–loop structures had previously been found to regulate mRNA stability in a number of mRNAs (11,21,65). It is likely that this ARE is also associated with stability. Cluster (b) consists of ion binding and actin cytoskeleton-associated genes, and the consensus structure has a highly conserved GC-rich stem. Two sequences are missing the potential second stem; however, the middle region also shows significant sequence similarities. Cluster (c) exhibits the least sequence similarity among the four clusters, with a consensus uridine in the centre loop, the seven sequences form a conserved structure. Three of the seven protein genes in this cluster are involved in development. Cluster (d) consists of seven zinc finger protein genes and titin, which binds adenosine triphosphate and ankyrin and is involved in adult heart development. This cluster has the highest sequence similarity, which reflects the evolutionary relations among the zinc finger protein family.

A number of other highly ranked clusters can be found in Supplementary Text (alignments of top ranked clusters) with detailed description of the genes and GO information. The result indicated that the potential clusters of new *cis*-regulatory elements did contain some functionally related genes, however, were not exclusive to genes of the same function. The clusters were more likely to contain genes within broader functional categories such

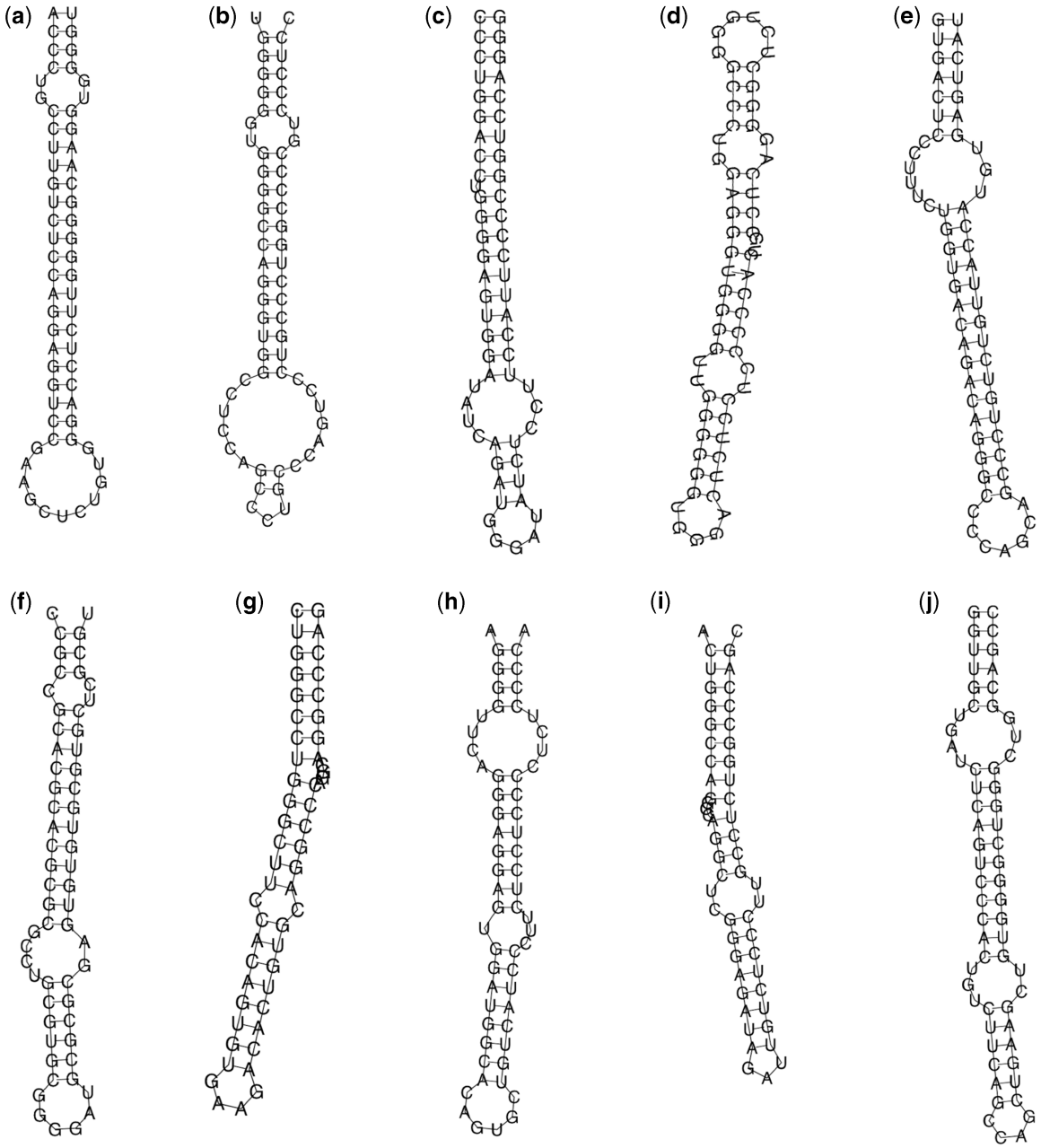

Figure 4. Predicted local structures of top ranked potential *cis*-regulatory elements based on MFE.(**a**) NM_000612: insulin-like growth factor 2 (somatomedin A); (**b**) NM_001080402: coiled-coil domain containing 61; (**c**) NM_022823: fibronectin type III domain containing 4; (**d**) NM_021241: widely interspaced zinc finger motifs; (**e**) NM_018129: pyridoxamine 5'-phosphate oxidase; (**f**) NM_004430: early growth response 3; (**g**) NM_182894: visual system homeobox 2; (**h**) NM_016162: inhibitor of growth family, member 4; (**i**) NM_018113: limb region 1 homolog (mouse)-like and (**j**) NM_015094: hypermethylated in cancer 2.

as ion binding, metabolite binding, regulation of signalling pathways or being associated with membrane.

DISCUSSION

In this study, we developed and used a method taking advantage of multiple RNA structural analysis programs combined with the classification capability of a SVM. With a specifically targeted training data set, we were able to distinguish structures that exhibit features of a defined class of *cis*-regulatory RNA elements from a genomic background. The method was shown to be considerably more sensitive and specific than single structural prediction programs. The sensitivity and specificity of the detection can also be adjusted by filtering on the SVM decision value or by functional classification to narrow down the candidates for experimental testing.

This method requires an initial sequence alignment. The quality of alignments has an impact on both the model building and classification. Here, we used alignments of vertebrate genomes made without consideration of RNA

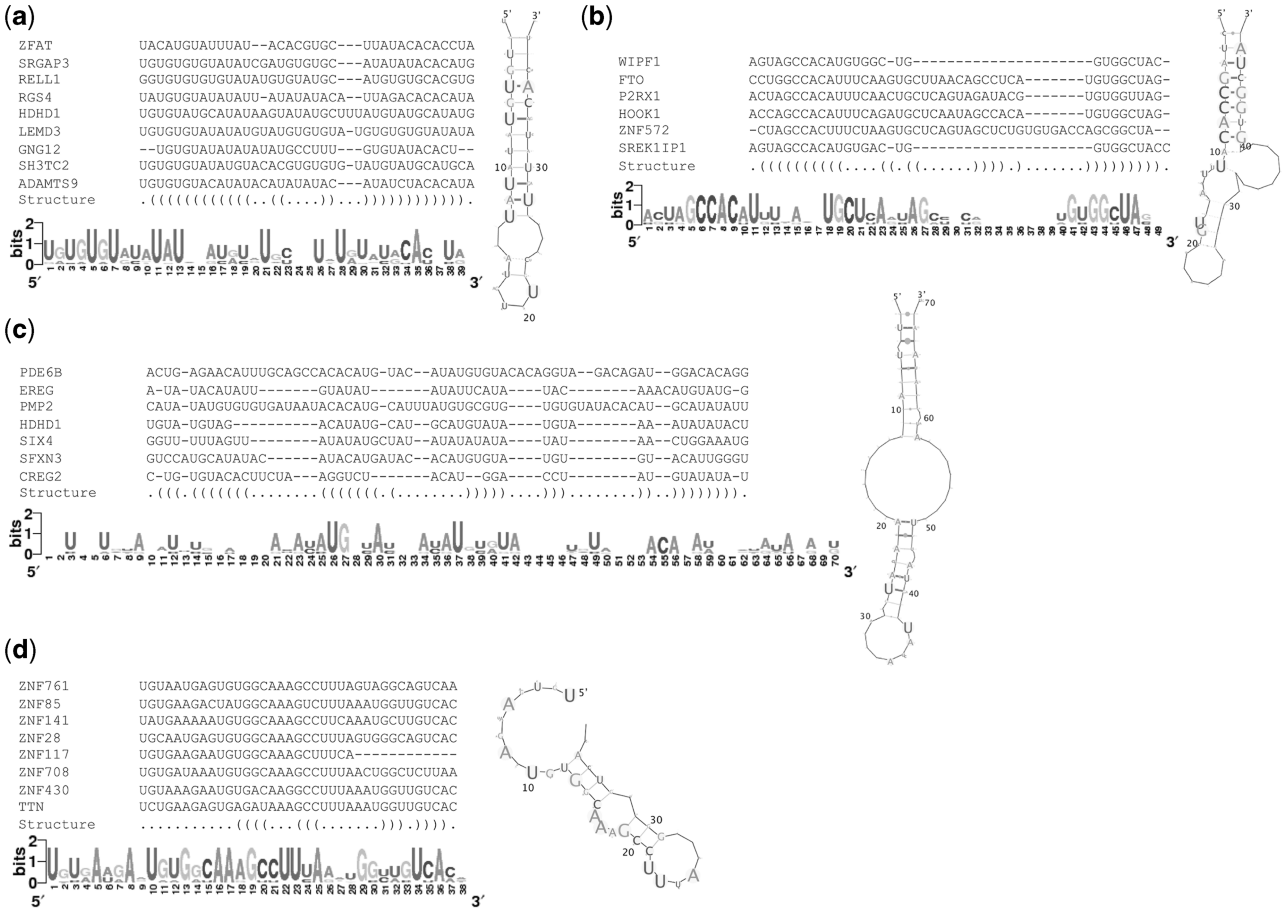

Figure 5. Putative large clusters of *cis*-regulatory structures. (**Cluster a**) ZFAT (NM_001029939): zinc finger and AT hook domain containing; SRGAP3 (NM_001033117): SLIT-ROBO Rho GTPase-activating protein 3; RELL1 (NM_001085400): RELT-like 1; RGS4 (NM_001102445): regulator of G-protein signalling 4; HDHD1 (NM_012080): haloacid dehalogenase-like hydrolase domain containing 1; LEMD3 (NM_014319): LEM domain containing 3; GNG12 (NM_018841): G protein, gamma 12; SH3TC2 (NM_024577): SH3 domain and tetratricopeptide repeats 2 and ADAMTS9 (NM_182920): ADAM metallopeptidase with thrombospondin type 1 motif. (**Cluster b**) WIPF1 (NM_001077269): WAS/WASL interacting protein family; FTO (NM_001080432): fat mass obesity associated; P2RX1 (NM_002558): purinergic receptor P2X, ligand-gated ion channel; HOOK1 (NM_015888): hook homolog 1 (*Drosophila*); ZNF572 (NM_152412): zinc finger protein 572 and SREK1IP1 (NM_173829): SREK1-interacting protein. (**Cluster c**) PDE6B (NM_000283): phosphodiesterase 6B, cGMP specific; EREG (NM_001432): epiregulin; PMP2 (NM_002677): peripheral myelin protein 2; HDHD1 (NM_012080): haloacid dehalogenase-like hydrolase domain containing 1; SIX4 (NM_017420): SIX homeobox 4; SFXN3 (NM_030971): sideroflexin 3 (SFXN3) and CREG2 (NM_153836): cellular repressor of E1A-simulated genes 2. (**Cluster d**) ZNF761 (NM_001008401): zinc finger protein 761; ZNF85 (NM_003429); ZNF141 (NM_003441); ZNF28 (NM_006969); ZNF117 (NM_015852); ZNF708 (NM_021269) and ZNF430 (NM_025189) and TTN (NM_133473): titin.

structure (53). Alternative alignments or realignments with a consideration of RNA structure may assist in detecting structures but might also bias the analysis to certain types of structured elements. Two of the 98 CisReg A set were not classified as positive by CisRNA-SVM: the potato virus X 3′-UTR element and the Hsp90 5′-UTR *cis*-regulatory element. Both of the elements had almost identical sequences in the training alignments and, therefore, did not provide useful information to the classifier. These might be improved with ongoing improvements to the Rfam database or possibly by including the predicted 'full' Rfam families.

We used genomic regions downstream of the 3′-UTR as an exacting negative set. The downstream sequences of protein coding genes may contain regulatory elements acting at the unprocessed pre-mRNA level or transcriptional level. CisRNA-SVM distinguished these from the *cis*-regulatory RNA elements we wish to detect. Surprisingly, given its training, the performance of CisRNA-SVM model on CisReg B set was still quite good (84.3%). Many pseudoknot elements are included in set B, and the programs used here do not predict pseudoknots. In addition, there are many large elements in the B set containing long unstructured regions, which fall into sub alignment windows.

Identification of *cis*-regulatory RNAs has been a difficult task mainly due to poor conservation, AU-richness and unusual structures. Our method is able to identify difficult-to-detect structures such as the IRE or the SECIS element, which contains four non-canonical A-G pairs. However, additional methodology will be required to recognize pseudoknots or unusual structures, e.g. G-quartets. Such methods could potentially be added to the set of software used for training or used for

post-processing of the putative elements discovered, as done here using Rfam and TransTerm elements (5,64). An alternative approach was recently used to detect enrichment of very common small ~17 base elements with four to five base stems associated with mRNA stability. Such small elements did not form part of our known element training set, and these would be unlikely to be detected by our method or by any method requiring thermodynamic stability (28).

We compared our method to two popular genome-wide ncRNA prediction programs, RNAz (50) and Evofold (44,66). Intersecting the genomic positions of predicted RNA structures with those of Rfam elements showed that our result contained mostly *cis*-regulatory RNAs that our model was trained on, whereas the other two programs also predicted many microRNAs and snoRNAs. As our approach specifically targeted *cis*-regulatory RNAs rather than ncRNAs in general, our result did not have large overlap with ncRNAs predicted by either RNAz or Evofold, as shown in Supplementary Figure S4. However, the common 3′-UTR structures predicted by RNAz and Evofold only consist of 23% and 10% of RNAz and Evofold predictions on the 3′-UTR, indicating clear difference of underlying structural prediction algorithms. Our result overlaps more with RNAz predictions, possibly due to the fact that RNAz also implements a SVM but was trained on ncRNA families different from our training families (50).

Our method predicted several thousand *cis*-regulatory-like secondary structures in 3′-UTRs with high probability of SVM prediction ($Pp > 0.9$). This sequence set contains previously known 5 SECIS elements and 11 Histone 3′-elements. These sequences represent a rich source of elements for further testing. To facilitate this, we have provided the data as bed files that can be mapped to the human genome.

Functional annotation of the putative elements discovered here indicates an enrichment of mRNAs encoding transcription factors, metal ion-binding protein and plasma membrane proteins. It is well established that known *cis*-regulatory elements, particularly unstructured elements, are enriched in genes with transcriptional or translational regulatory activities (67,68). Strikingly, these novel elements also have that functional association.

A number of elements clustered into groups with sequence and structural similarities. It is, therefore, clear that many of the several thousand elements predicted here are not individual elements but would form small functionally similar groups. This idea is based on the finding of small numbers of instances of well-known structured *cis*-regulatory elements in the human genome, such as the 7 IREs, 62 histone 3′-elements and 24 SECIS elements (64,69). In our clustering analysis, we successfully clustered the histone 3′-UTR element, and subsequently used it as a control to obtain other clusters of sequences based on pairwise sequence similarity. The clustering analysis showed that several large clusters contain, however, not exclusively, members of closely related functional groups, supporting the notion that some of these could be co-regulated. Although this clustering method would not detect conserved structures with low sequence similarity, which may be the case for some *cis*-regulatory elements. However, based on the current knowledge, such cases are relatively rare.

Overall, our study has developed an effective way for searching structured *cis*-elements exhibiting features of the previously known *cis*-elements and proposes many for further experimental testing. CisRNA-SVM is designed to detect subtle structures of *cis*-regulatory RNAs but could be retrained to search for other types of RNA structures, for example small non-coding RNA genes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1–4 and Supplementary Text.

FUNDING

Human Frontier Science Foundation [RGP0031 2009 to Ian Macara, Anne Spang and C.M.B.]; University of Otago Research Grant. Funding for open access charge: University of Otago Research Grant.

Conflict of interest statement. None declared.

REFERENCES

1. Chatterjee,S. and Pal,J.K. (2009) Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. *Biol. Cell.*, **101**, 251–262.
2. Chen,J.M., Ferec,C. and Cooper,D.N. (2006) A systematic analysis of disease-associated variants in the 3′ regulatory regions of human protein-coding genes II: the importance of mRNA secondary structure in assessing the functionality of 3′ UTR variants. *Hum. Genet.*, **120**, 301–333.
3. Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, REVIEWS0004.
4. Pesole,G., Liuni,S., Grillo,G., Licciulli,F., Mignone,F., Gissi,C. and Saccone,C. (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.*, **30**, 335–340.
5. Jacobs,G.H., Chen,A., Stevens,S.G., Stockwell,P.A., Black,M.A., Tate,W.P. and Brown,C.M. (2009) Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res.*, **37**, D72–D76.
6. Gray,N.K. and Wickens,M. (1998) Control of translation initiation in animals. *Annu. Rev. Cell. Dev. Biol.*, **14**, 399–458.
7. Lau,A.G., Irier,H.A., Gu,J., Tian,D., Ku,L., Liu,G., Xia,M., Fritsch,B., Zheng,J.Q., Dingledine,R. *et al.* (2010) Distinct 3′UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF). *Proc. Natl. Acad. Sci. USA.*, **107**, 15945–15950.
8. Reveal,B., Yan,N., Snee,M.J., Pai,C.I., Gim,Y. and Macdonald,P.M. (2010) BREs mediate both repression and activation of oskar mRNA translation and act in trans. *Dev. Cell.*, **18**, 496–502.
9. Fabian,M.R., Sonenberg,N. and Filipowicz,W. (2010) Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.*, **79**, 351–379.
10. Mangone,M., Manoharan,A.P., Thierry-Mieg,D., Thierry-Mieg,J., Han,T., Mackowiak,S.D., Mis,E., Zegar,C., Gutwein,M.R., Khivansara,V. *et al.* (2010) The landscape of C. elegans 3′UTRs. *Science*, **329**, 432–435.
11. McGray,A.J., Gingerich,T., Petrik,J.J. and Lamarre,J. (2011) Regulation of thrombospondin-1 expression through AU-rich

elements in the 3′UTR of the mRNA. *Cell. Mol. Biol. Lett.*, **16**, 55–68.
12. Martin,K.C. and Ephrussi,A. (2009) mRNA localization: gene expression in the spatial dimension. *Cell*, **136**, 719–730.
13. Huang,H.Y., Chien,C.H., Jen,K.H. and Huang,H.D. (2006) RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. *Nucleic Acids Res.*, **34**, W429–W434.
14. Bekaert,M., Firth,A.E., Zhang,Y., Gladyshev,V.N., Atkins,J.F. and Baranov,P.V. (2010) Recode-2: new design, new search tools, and many more genes. *Nucleic Acids Res.*, **38**, D69–D74.
15. Barreau,C., Paillard,L. and Osborne,H.B. (2005) AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res*, **33**, 7138–7150.
16. Wein,G., Rossler,M., Klug,R. and Herget,T. (2003) The 3′-UTR of the mRNA coding for the major protein kinase C substrate MARCKS contains a novel CU-rich element interacting with the mRNA stabilizing factors HuD and HuR. *Eur. J. Biochem.*, **270**, 350–365.
17. Vlasova,I.A., Tahoe,N.M., Fan,D., Larsson,O., Rattenbacher,B., Sternjohn,J.R., Vasdewani,J., Karypis,G., Reilly,C.S., Bitterman,P.B. *et al.* (2008) Conserved GU-rich elements mediate mRNA decay by binding to CUG-binding protein 1. *Mol. Cell.*, **29**, 263–270.
18. Barends,S., Bink,H.H., van den Worm,S.H., Pleij,C.W. and Kraal,B. (2003) Entrapping ribosomes for viral translation: tRNA mimicry as a molecular Trojan horse. *Cell*, **112**, 123–129.
19. Stupina,V.A., Meskauskas,A., McCormack,J.C., Yingling,Y.G., Shapiro,B.A., Dinman,J.D. and Simon,A.E. (2008) The 3′ proximal translational enhancer of Turnip crinkle virus binds to 60S ribosomal subunits. *RNA*, **14**, 2379–2393.
20. Tzanetakis,I.E., Tsai,C.H., Martin,R.R. and Dreher,T.W. (2009) A tymovirus with an atypical 3′-UTR illuminates the possibilities for 3′-UTR evolution. *Virology*, **392**, 238–245.
21. Chabanon,H., Mickleburgh,I., Burtle,B., Pedder,C. and Hesketh,J. (2005) An AU-rich stem-loop structure is a critical feature of the perinuclear localization signal of c-myc mRNA. *Biochem. J.*, **392**, 475–483.
22. Selezneva,A.I., Cavigiolio,G., Theil,E.C., Walden,W.E. and Volz,K. (2006) Crystallization and preliminary X-ray diffraction analysis of iron regulatory protein 1 in complex with ferritin IRE RNA. *Acta Crystallogr. Sect F Struct. Biol. Cryst. Commun.*, **62**, 249–252.
23. Sampath,P., Mazumder,B., Seshadri,V. and Fox,P.L. (2003) Transcript-selective translational silencing by gamma interferon is directed by a novel structural element in the ceruloplasmin mRNA 3′ untranslated region. *Mol. Cell. Biol.*, **23**, 1509–1519.
24. Crucs,S., Chatterjee,S. and Gavis,E.R. (2000) Overlapping but distinct RNA elements control repression and activation of nanos translation. *Mol. Cell.*, **5**, 457–467.
25. dos Santos,G., Simmonds,A.J. and Krause,H.M. (2008) A stem-loop structure in the wingless transcript defines a consensus motif for apical RNA transport. *Development*, **135**, 133–143.
26. Van De Bor,V., Hartswood,E., Jones,C., Finnegan,D. and Davis,I. (2005) gurken and the I factor retrotransposon RNAs share common localization signals and machinery. *Dev. Cell.*, **9**, 51–62.
27. Hamilton,R.S., Hartswood,E., Vendra,G., Jones,C., Van De Bor,V., Finnegan,D. and Davis,I. (2009) A bioinformatics search pipeline, RNA2DSearch, identifies RNA localization elements in *Drosophila* retrotransposons. *RNA*, **15**, 200–207.
28. Goodarzi,H., Najafabadi,H.S., Oikonomou,P., Greco,T.M., Fish,L., Salavati,R., Cristea,I.M. and Tavazoie,S. (2012) Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, **485**, 264–268.
29. Dalgleish,G., Veyrune,J.L., Blanchard,J.M. and Hesketh,J. (2001) mRNA localization by a 145-nucleotide region of the c-fos 3′-untranslated region. Links to translation but not stability. *J. Biol. Chem.*, **276**, 13593–13599.
30. Gumy,L.F., Yeo,G.S., Tung,Y.C., Zivraj,K.H., Willis,D., Coppola,G., Lam,B.Y., Twiss,J.L., Holt,C.E. and Fawcett,J.W. (2011) Transcriptome analysis of embryonic and adult sensory axons reveals changes in mRNA repertoire localization. *RNA*, **17**, 85–98.
31. Brown,V., Jin,P., Ceman,S., Darnell,J.C., O'Donnell,W.T., Tenenbaum,S.A., Jin,X., Feng,Y., Wilkinson,K.D., Keene,J.D. *et al.* (2001) Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell*, **107**, 477–487.
32. Kao,D.I., Aldridge,G.M., Weiler,I.J. and Greenough,W.T. (2010) Altered mRNA transport, docking, and protein translation in neurons lacking fragile X mental retardation protein. *Proc. Natl Acad. Sci. USA.*, **107**, 15601–15606.
33. Furic,L., Maher-Laporte,M. and DesGroseillers,L. (2008) A genome-wide approach identifies distinct but overlapping subsets of cellular mRNAs associated with Staufen1- and Staufen2-containing ribonucleoprotein complexes. *RNA*, **14**, 324–335.
34. Mili,S., Moissoglu,K. and Macara,I.G. (2008) Genome-wide screen reveals APC-associated RNAs enriched in cell protrusions. *Nature*, **453**, 115–119.
35. Nury,D., Chabanon,H., Levadoux-Martin,M. and Hesketh,J. (2005) An eleven nucleotide section of the 3′-untranslated region is required for perinuclear localization of rat metallothionein-1 mRNA. *Biochem. J.*, **387**, 419–428.
36. Lee,J.E., Lee,J.Y., Wilusz,J., Tian,B. and Wilusz,C.J. (2010) Systematic analysis of cis-elements in unstable mRNAs demonstrates that CUGBP1 is a key regulator of mRNA decay in muscle cells. *PLoS One*, **5**, e11201.
37. Mathews,D.H., Moss,W.N. and Turner,D.H. (2010) Folding and finding RNA secondary structure. *Cold Spring Harb. Perspect. Biol.*, **2**, a003665.
38. Hamilton,R.S. and Davis,I. (2011) Identifying and searching for conserved RNA localisation signals. *Methods Mol. Biol.*, **714**, 447–466.
39. Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
40. Freyhult,E., Moulton,V. and Gardner,P. (2005) Predicting RNA structure using mutual information. *Appl. Bioinformatics*, **4**, 53–59.
41. Gardner,P.P. (2009) The use of covariance models to annotate RNAs in whole genomes. *Brief. Funct. Genomic Proteomic*, **8**, 444–450.
42. Washietl,S., Hofacker,I.L., Lukasser,M., Huttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.
43. Xu,X., Ji,Y. and Stormo,G.D. (2009) Discovering cis-regulatory RNAs in Shewanella genomes by support vector machines. *PLoS Comput. Biol.*, **5**, e1000338.
44. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
45. Khaladkar,M., Liu,J., Wen,D., Wang,J.T. and Tian,B. (2008) Mining small RNA structure elements in untranslated regions of human and mouse mRNAs using structure-based alignment. *BMC Genomics*, **9**, 189.
46. Torarinsson,E., Sawera,M., Havgaard,J.H., Fredholm,M. and Gorodkin,J. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.*, **16**, 885–889.
47. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
48. Rabani,M., Kertesz,M. and Segal,E. (2008) Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl Acad. Sci. USA.*, **105**, 14885–14890.
49. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
50. Gruber,A.R., Findeiss,S., Washietl,S., Hofacker,I.L. and Stadler,P.F. (2010) Rnaz 2.0: improved noncoding RNA detection. *Pac. Symp. Biocomput.*, **15**, 69–79.

51. Tran,T.T., Zhou,F., Marshburn,S., Stead,M., Kushner,S.R. and Xu,Y. (2009) De novo computational prediction of non-coding RNA genes in prokaryotic genomes. *Bioinformatics*, **25**, 2897–2905.
52. Lange,S.J., Maticzka,D., Mohl,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.
53. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al*. (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*., **39**, D876–D882.
54. Anandam,P., Torarinsson,E. and Ruzzo,W.L. (2009) Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. *Bioinformatics*, **25**, 668–669.
55. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*., **31**, 3429–3431.
56. Torarinsson,E., Havgaard,J.H. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932.
57. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*., **22**, 2079–2088.
58. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
59. Schattner,P., Brooks,A.N. and Lowe,T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res*., **33**, W686–W689.
60. Chang,C. and Lin,C. (2001) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*, **2**, 1–27.
61. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc*., **4**, 44–57.
62. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
63. Katoh,K. and Toh,H. (2008) Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics*, **9**, 212.
64. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al*. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*., **39**, D141–D145.
65. Sarnowska,E., Grzybowska,E.A., Sobczak,K., Konopinski,R., Wilczynska,A., Szwarc,M., Sarnowski,T.J., Krzyzosiak,W.J. and Siedlecki,J.A. (2007) Hairpin structure within the 3′UTR of DNA polymerase beta mRNA acts as a post-transcriptional regulatory element and interacts with Hax-1. *Nucleic Acids Res*., **35**, 5499–5510.
66. Parker,B.J., Moltke,I., Roth,A., Washietl,S., Wen,J., Kellis,M., Breaker,R. and Pedersen,J.S. (2011) New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res*., **21**, 1929–1943.
67. Keene,J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet*., **8**, 533–543.
68. Gruber,A.R., Fallmann,J., Kratochvill,F., Kovarik,P. and Hofacker,I.L. (2011) AREsite: a database for the comprehensive investigation of AU-rich elements. *Nucleic Acids Res*., **39**, D66–D69.
69. Castellano,S., Gladyshev,V.N., Guigo,R. and Berry,M.J. (2008) SelenoDB 1.0: a database of selenoprotein genes, proteins and SECIS elements. *Nucleic Acids Res*., **36**, D332–D338.