10. Bioinformatic Methods to Discover *Cis*-regulatory Elements in mRNAs

Stewart G. Stevens, Chris M. Brown

Cis-regulatory elements play a number of important roles in determining the fate of messenger RNAs (mRNAs). Due to these elements, mRNAs may be translated with remarkable efficiency, or destroyed with little translation. Untranslated regions cover over a third of a typical human mRNA and often contain a range of regulatory elements. Some elements along with their RNA or protein binding partners are well characterized, though many are not. These require different types of bioinformatic methods for identification and discovery. The most successful techniques combine a range of information and search strategies. Useful information may include conservation across species, prior biological knowledge, known false positives, or noisy high-throughput experimental data. This chapter focuses on current successful methods designed to discover elements with high sensitivity but low false-positive rates.

10.1 The Importance of *Cis*-regulatory Elements in mRNAs

Cells respond to the environment by changing gene expression. In human cells, gene expression is often regulated at both the transcriptional and translational levels. Steady-state levels of mRNAs and their proteins will alter due to the combined effects of this regulation. Regulation at the posttranscriptional level is critical for rapid responses to environmental factors. While there is correlation with the levels of mRNA, translational control is key to determining cellular levels of specific protein [10.1]. *Cis*-regulatory elements commonly affect mRNA stability or translational efficiency of the mRNA [10.2]. Figure 10.1 shows a simplified schematic of the regulatory elements found in mRNAs.

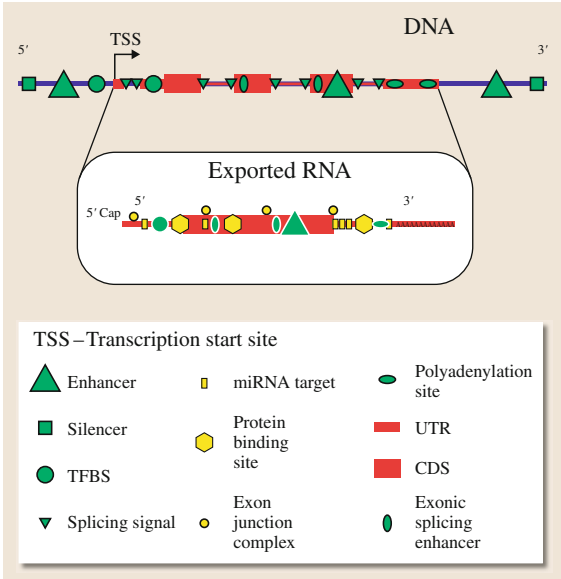

Fig. 10.1 Simplified schematic of regulatory elements in DNA and an mRNA transcribed from the corresponding genomic region. *White outlined* elements (e.g., TFBS) are present but nonfunctional in the mature mRNA (CDS: coding sequence)

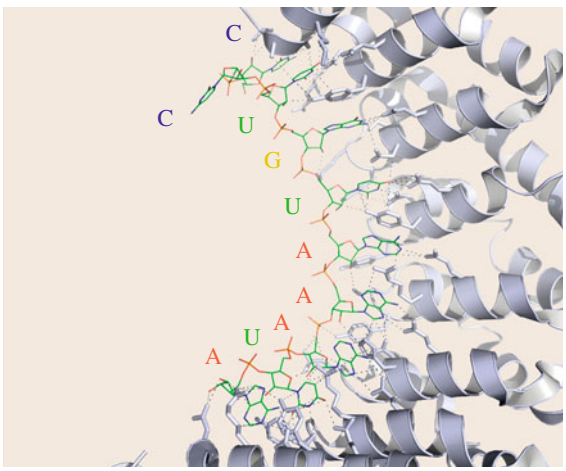

Fig. 10.2 PUF3p in complex with the Cox17 mRNA. The PUF3 binding site is not base paired. (Figure rendered using pyMol, PDB, 3K49)

Many different proteins potentially bind to RNAs. In yeast there are estimated to be over 600 RNA-binding proteins (Fig. 10.2) [10.3], and in humans, based on the occurrence of RNA binding motifs in proteins, it is likely that this number is substantially higher [10.4]. The most common domains are K homology (KH, Fig. 10.3c), RNA recognition motif (RRM, Fig. 10.3d), and double-stranded RNA (dsRNA, Fig. 10.3a) binding domains [10.5]. Many of these proteins are part of other structures, for example, ribosomes or splicing complexes, but some of these have additional roles in mRNA binding. In addition, there are many proteins that do not have obvious canonical RNA binding domains but bind specifically to groups of RNAs [10.6].

Some of these *trans*-acting binding proteins bind to specific *cis*-regulatory elements in mRNAs, providing a regulatory mechanism determining transcript fate. The translation, stability, and localization of an mRNA may vary in response to these interactions.

Interactions with microRNAs (miRNAs) are of particular importance in determining mRNA stability. Predicted target sites for miRNAs have been reported for over 30% of human genes. However, only a small proportion of these predictions are experimentally confirmed.

The *cis*-regulatory elements encoded in the transcript are most often found in the untranslated regions (UTRs). This bias may partly be because study of these regions is more tractable to experimental and bioinformatic analysis. There is however evidence that *trans*-acting factors such as miRNAs are easily displaced by the translational machinery and so operate more readily in the 3′ UTR [10.7]. Translation can be entirely repressed immediately after nuclear export, and so this displacement is not an issue for elements such as those found in the coding region of the mRNA for *ASH1* [10.8]. Some *cis*-regulatory elements such as the binding site for iron regulatory proteins certainly occur in the 5′ UTR [10.9].

Regulatory elements that act in the DNA may be also present in the mRNA. RNAs exported from the nucleus contain some of these elements that are nonfunctional in the mature transcript. This is particularly confounding for de novo element discovery in mRNAs. Knowledge of the genomic elements acting at the transcriptional level can help to resolve this.

Many types of structured and unstructured *cis*-regulatory mRNA elements act at the posttranscriptional level. Examples of such elements are the selenocysteine insertion sequence (SECIS), Histone3, PUF3, and iron responsive element (IRE). Computational methods have been developed to find each of these [10.10–12]. It is also useful to consider such well-characterized elements in order to estimate the variation that may exist in novel elements.

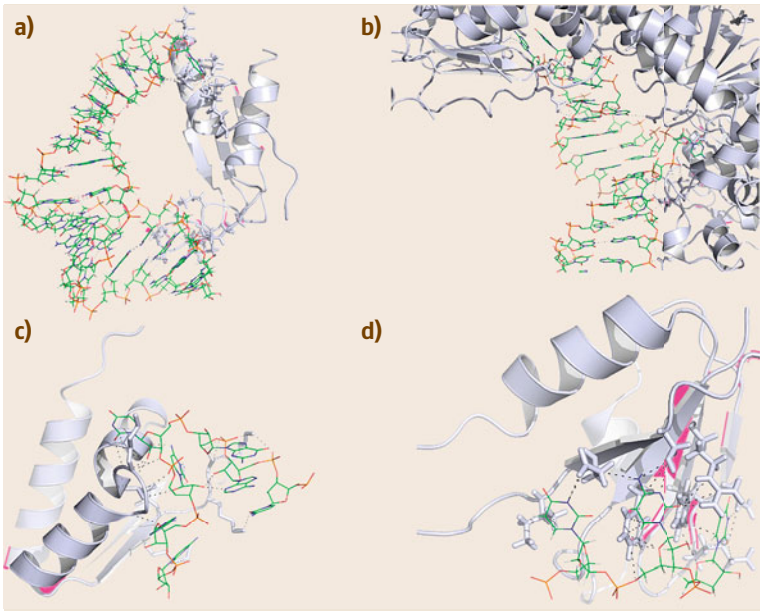

Fig. 10.3a–d Examples of RNA–protein interactions. **(a)** dsRNA binding domain in complex with Staufen (PDB 1EKZ). **(b)** Iron responsive element of ferritin mRNA in complex with an iron regulatory protein (PDB 2IPY). **(c)** KH domain (PDB 2ANR). **(d)** RRM domain (PDB 2L41). (All figures rendered using PyMol)

Methods developed for genes encoding structured noncoding RNAs (ncRNAs) may be applied to *cis*-regulatory elements [10.13, 14].

Most importantly, these well-characterized elements may be used for benchmarking the performance of prediction algorithms that may then be applied in the search for novel elements. For this purpose a subset of the Rfam database [10.15] (e.g., CisReg), or small parts of databases of known RNA secondary structures, e.g., RNA STRAND [10.16] or CompaRNA [10.17], can be used.

In the past, the limitations of methodology for determining *cis*-regulatory elements have meant that there was a large role for bioinformatic prediction. In the last few years new high-throughput RNA-Seq techniques combined with bioinformatics are now being developed [10.18–21].

10.2 Searching for *Cis*-regulatory Elements Using Bioinformatics

Elements in mRNAs are necessary for regulation of stability, translational control, and localization. They may have structural motifs critical to their function or be characterized by primary sequence alone. These *cis*-regulatory elements are the targets for miRNAs and protein binding sites. Bioinformatic analysis of mRNA can be useful in proposing new models and hypotheses for experimental testing and interpreting existing data. The aim of this analysis is to discover those regulatory sequences that regulate the fate of the mRNAs containing them.

There are many challenges in identifying *cis*-regulatory elements within RNAs. Their primary sequence patterns are often sparse. The binding sites for proteins may depend on just a few nucleotides with critical secondary structure. Determination of RNA structure is experimentally difficult, and prediction tools are often inaccurate. Regulatory elements are only sometimes conserved, and there are many distracting signals, such as elements operating at the transcriptional level. Some regulatory elements may be unique to a particular mRNA, but others such as the iron responsive element certainly operate within many mRNAs with divergent regulatory outcomes [10.12]. Even when effective bioinformatic models exist for a *cis*-regulatory element, the application of this model for the purposes of discovery will inevitably yield some false positives.

Some methods have been developed that attempt to discover new regulatory elements in mRNA sequences using only limited biological knowledge. However, the most successful methods utilize as much biological knowledge as is available in order to refine and inform the predictions. Despite the development and publication of several hundred methods (particularly for

Table 10.1 Commonly used and current tools and research directions

Name	Purpose	Type	Reference	URL
General tools				
UCSC Genome Browser	Visualize and download many different data for various genomes	Web	[10.22]	http://genome.ucsc.edu/
ncRNA Genome Browser	Version of the UCSC browser that includes many data tracks particularly aimed at RNA analysis	Web	[10.23]	http://www.ncrna.org/
Galaxy Suite	Web-based service that eases data acquisition, processing, and visualization by integrating many different tools	Web	[10.24]	http://galaxy.psu.edu/
Narrowing the search space				
RepeatMasker	Identify repeat elements	Command line via UCSC	[10.22, 25]	http://www.repeatmasker.org/
Transfac	Identify TFBS	Web	[10.26]	http://www.gene-regulation.com/
Jaspar		Web	[10.27]	http://jaspar.genereg.net/
STAMP	Binding motif/matrix comparison tool	Web Command line	[10.28]	http://www.benoslab.pitt.edu/services.html
Refseq	Gene annotation database that allows focus on UTRs	Via UCSC	[10.29]	http://www.ncbi.nlm.nih.gov/RefSeq/
Known regulatory elements				
Rfam	Contains covariance models of many known regulatory elements	Web	[10.15]	http://rfam.sanger.ac.uk/
Transterm	Contains pattern-based models of known regulatory elements	Web	[10.30]	http://mrna.otago.ac.nz/Transterm.html
UTRsite		Web	[10.31]	http://utrsite.ba.itb.cnr.it/
RBPDb	Database of protein binding sites	Web	[10.32]	http://rbpdb.ccbr.utoronto.ca/
TargetScan	Tools and database of predicted miRNA binding sites	Web Via UCSC	[10.33]	http://www.targetscan.org/
PicTar	Database of predicted miRNA binding sites	Web	[10.34]	http://pictar.mdc-berlin.de/
miRbase	Database of miRNAs	Web	[10.35]	http://www.mirbase.org/
ElMMo	Database of predicted miRNA binding sites – also allows searching based on mRNA expression profiles	Web	[10.36]	http://www.mirz.unibas.ch/ElMMo2/
Primary sequence analysis				
MEME	Tools for finding overrepresented patterns in primary sequences	Web Command line	[10.37]	http://meme.nbcr.net/
Weeder		Web Command line	[10.38]	http://www.pesolelab.it/
TEIRESIAS		Web Command line	[10.39]	http://cbcsrv.watson.ibm.com/Tspd.html

miRNA sites) only a few are commonly used (or cited). This may sometimes be because the expected or actual utility of the software is outweighed by the difficulty of installing and using it. In other cases the benefits of a particular tool may be outweighed by the familiarity of commonly used tools that do a similar job.

10.2.1 Summary of Tools and Data Sources

The most promising commonly used and current tools and research directions are discussed here (listed in Table 10.1). Lists with different foci can be found in the literature [10.40–42].

Table 10.1 (continued)

Name	Purpose	Type	Reference	URL
Secondary structure prediction [10.42]				
mfold/UNAfold	Predict secondary structure from primary sequences	Command line	[10.43]	http://mfold.rna.albany.edu/
RNAfold and RNAplFold		Web Command line	[10.44]	http://www.tbi.univie.ac.at/RNA/
RNAalifold	Predict secondary structure from alignments of primary sequences	Web Command line	[10.45]	http://www.tbi.univie.ac.at/RNA/
Dynalign/Multilign	Simultaneous alignment and folding of multiple similar RNAs to predict structure	Command line GUI	Mathews, 2010 #28239}	http://rna.urmc.rochester.edu/RNAstructure.html
Turbofold		Command line GUI	Mathews, 2010 #28239}	http://rna.urmc.rochester.edu/RNAstructure.html
Comparing secondary and tertiary structures				
RNAdistance	2-D structure comparison	Command line	[10.44]	http://www.tbi.univie.ac.at/~ivo/RNA/
RNAforester		Command line	[10.46]	http://bibiserv.techfak.uni-bielefeld.de/rnaforester/
iPARTS	3-D structure comparison	Web	[10.47]	http://bioalgorithm.life.nctu.edu.tw/iPARTS/
Searching for secondary structures				
RNAMotif	Search method to identify motifs that may be described structurally and/or by sequence	Command line	[10.48]	http://casegroup.rutgers.edu/
CMFinder	A tool that finds conserved motifs based on predicted structures using covariance models	Web Command line	[10.49]	http://wingless.cs.washington.edu/htbin-post/unrestricted/CMfinderWeb/CMfinderInput.pl
Combining primary and secondary structural search methods				
Infernal/cmsearch	Search method using covariance models built from sequence alignments to a consensus structure	Command line	[10.50]	http://infernal.janelia.org/
Scan_for_matches	Pattern-based search method	Command line	[10.51]	http://blog.theseed.org/servers/2010/07/scan-for-matches.html
Evidence for common regulation and tissue-specific expression				
GEO	Database of gene expression experiments	Web	[10.52]	http://www.ncbi.nlm.nih.gov/geo/
Publicly available combinatorial methods				
MEMERIS	An extension of MEME guided by predicted secondary structure	Command line	[10.53]	http://cs.stanford.edu/people/hillerm/Data/MEMERIS/
RNAz	A tool that finds conserved structural motifs in aligned sequences	Web Command line	[10.54]	http://www.tbi.univie.ac.at/~wash/RNAz/
FIRE	A tool that combines detection of overrepresented primary sequence patterns with other biological data	Web Command line	[10.55]	https://iget.princeton.edu/

10.3 Obtaining High-Quality mRNA Sequence Data to Analyze

Individual mRNA sequences can be obtained from the University of California, Santa Cruz (UCSC), Ensembl, or the National Center for Biotechnology Information (NCBI). Refseq annotations of UTRs are useful to focus on, as these avoid many elements acting at the transcriptional level and also patterns in coding sequences. The Refseq annotations are more conservatively curated than the Ensembl ones, which contain many predicted transcripts. The UTRs may be readily obtained from Ensembl using the web-based Biomart interface. Refseq transcripts along with their annotations can be obtained from NCBI or UCSC – the annotations must be processed to produce UTR sequences. Downstream statistical processing may be necessary to remove duplicate sequences and to remove redundant, very similar sequences (e.g., using CD-HIT-EST).

Multiple sequence alignments for several species can be obtained from the UCSC genome browser. The output is supplied in multiple alignment format (MAF). The alignments can be filtered to include only those sequences for which there is reasonable conservation. This process is simplified by using the online tool Galaxy [10.24].

10.3.1 Evidence for Common Regulation and Tissue-Specific Expression

mRNA Expression

In order to discover regulatory elements in a gene of interest it is useful to look for common sequences and structures in similarly regulated genes. Also, in the scenario where a regulatory element is known for a particular gene, it is sensible to search for similar elements that may be identified in coregulated genes. The gene expression omnibus [10.52] (GEO) contains mRNA expression data for multiple species and tissues under many experimental conditions. GEO also includes data from RNA immunoprecipitation chip (RIP-chip) experiments, e.g., with Staufen1 (GSE8438) or HuR (GSE29778) and more recently RIP-RNA-Seq data (e.g., Tdrd1 GSE29418).

The web interface provided by GEO allows the researcher to identify coregulated mRNAs via a link to *profile neighbors*. Expression data may also be downloaded and other statistical tools used to identify and quantify these neighbors, which is worth doing if a study of special interest is available. The profile neighbors provided by GEO are simply the 200 most closely expressed genes.

There are also some relevant sets of publicly available data from published work that are not included in GEO, e.g., co-localized genes for fruit-fly embryos [10.56] or the data from yeast RIP-Chip [10.3], which are provided as websites or supplemental data.

RNA Binding Protein Expression. The expression of RNA binding proteins in specific tissues can also be obtained from GEO. On the GEO website the protein expression data may be readily accessed using the advanced search option and specifying "Protein" for the "Sample Type" field.

miRNA Expression. Coexpression of miRNA and target may be an indication that a functional interaction occurs. Many miRNAs, like mRNAs, are expressed in specific tissues. Studies and methods that consider coexpression in the same tissues have been useful in identifying biologically relevant pairs. Large studies have determined the expression of most small RNAs in many tissues, for example, over 250 small RNA libraries from 26 different organ systems [10.57]. Several databases provide access to these expression data [10.58]. Some databases such as ElMMo combine both miRNA and mRNA data.

10.3.2 Narrowing the Search Space to Biologically Relevant Regions

In the search for novel *cis*-regulatory elements that operate at the posttranscriptional level, sequence elements that are likely to be false positives must be avoided. These include repetitive elements and elements acting purely at the transcriptional level.

When de novo element discovery is being pursued these distracting elements may overwhelm pattern prediction algorithms and so must be masked in input sequence. When searching for known elements these distractions should be considered at a later stage when assessing the likelihood of a putative hit.

Coding sequence contains confounding patterns arising from protein constraints and common combinations of amino acids in the translated product. They may be masked from analysis in a straightforward fashion where sufficient gene annotation is available. If this is not the case but sufficient protein information is available, it is possible to use tblastn (protein–nucleotide 6-frame translation) to map and

mask out likely coding regions. It must be noted that this approach will inevitably fail when the interesting *cis*-regulatory regions are actually in the coding region.

10.3.3 Distinguishing and Avoiding Repetitive Elements

Repetitive sequences arising from events such as viral retrotransposition are abundant in many genomes; For example, the Alu-element forms 10% of the human genome; furthermore, it is concentrated in gene regions and overlaps mRNAs [10.59]. Therefore, many human 3′ UTRs contain detectable Alu remnants of the 7S RNA.

This and other repeat elements are catalogued in Repbase [10.25]. These sequences may be masked out at an early stage of analysis using the RepeatMasker [10.60] program. Alternatively, the UCSC genome browser [10.22] may be used when assessing putative elements – this has a RepeatMasker track and allows intersection of uploaded candidate *cis*-regulatory elements to report overlaps.

It is possible that a novel common repeat element may be discovered as a putative regulatory element in a set of coregulated mRNAs. The choice of an appropriate negative control set of real RNAs (rather than simulated sequences) from the same species is important to avoid this.

10.3.4 Distinguishing and Avoiding Elements Acting at the Transcriptional Level and Other RNA Features

Elements acting in the DNA such as transcription factor binding sites (TFBSs) and enhancers may be misidentified as elements acting posttranscriptionally. If the genome of interest is well annotated or sufficient transcript sequence data are available, a genomic search can be avoided. This will go some way to avoiding transcriptionally acting elements.

The methods discussed in this chapter may be usefully restricted to untranslated regions (UTRs). This will avoid many false positives from genomic sequences and patterns associated with coding sequence.

TFBSs in Transcription Promoters that May Overlap 5′ Regions of the mRNA. TFBS databases such as Transfac [10.26] and Jaspar [10.27] contain both experimentally defined and computational predictions. As these databases contain many false positives it is inadvisable to mask or remove predicted TFBSs from an early stage of analysis.

The difficulties with predicted TFBSs are reviewed here [10.61]. One approach to improving TFBS prediction is to use conservation information, although this is controversial as binding sites may not be conserved in multispecies alignments [10.62].

The web-based tool STAMP [10.28] may be employed to determine whether a set of aligned predicted *cis*-regulatory elements coincides with any of those TFBSs in the public databases.

Enhancers that May Overlap 3′ Regions

ChIP-seq studies can be used to identify binding sites for known enhancers [10.63]. The cited review points out that there is conflicting evidence regarding the conservation of enhancer regions and TFBS; although many sites are conserved, there are a large number of nonconserved sites. Bearing these limitations in mind, the available data should still be considered; For example, a recent study first identified highly conserved noncoding sequences and then tested associated genes for tissue-specific expression during mouse embryogenesis [10.64]. Although the identified enhancer sites are clearly a subset of all such elements and very few occur within gene transcripts, putative *cis*-regulatory elements acting in mRNAs should be cross-checked against known enhancer regions.

10.4 Known Regulatory Elements

When a *cis*-regulatory element is characterized by a primary sequence, pairwise alignment methods or multiple ones, including hidden Markov models (HMMs), may be used to search for it in other RNAs. Pairwise alignment methods are best established and have many fast, readily available implementations such as the basic local alignment search tool (BLAST) [10.65]. On the other hand, HMMs which probabilistically model state transitions and thereby account for gaps in a nonarbitrary manner can have increased sensitivity, albeit at a computational cost [10.66].

An example where bioinformatic primary sequence analysis has been successful is in a study of the mRNA for the *Vg1* localized mRNA. This mRNA was shown to be localized to the vegetal cortex of *Xenopus laevis* oocytes by parts of a 340-base region in the 3′ UTR. Subsequently, this was shown to be bound by proteins that contain multiple KH domains. *Schnapp*'s group first identified using bioinformatics four repeated sequence elements, but experimental deletion showed that only one of these elements, *UUCAC* (E2), was critical for function [10.67].

Mowry's group also studied the localization of the mRNA for *Vg1* and took a different approach [10.68]. They systematically deleted sequences over the entire 340-nt region. Interestingly, the result they found was quite different – a different sequence, UUUCUA (VM1), was identified as critical, and this was supported using site-specific mutagenesis [10.68]. Although this sequence corresponded to one of the three (E1) elements identified in the Schnapp laboratory, their results showed those deletion constructs had reduced but not abolished localization.

The UUCAC, E2 element was also found to be required for localization of the mRNA for *vegT* [10.69]. Interestingly, both the *vegT* and the *Vg1* mRNAs had multiple (five) copies of this element. Subsequently a shorter more generalized motif, *CAC*, was postulated, repeats of which are present in the majority of RNAs localized to the vegetal cortex of *Xenopus laevis* oocytes [10.70].

This demonstrates the utility of bioinformatic analysis, although the requirement to find repeated clusters of short sequence required the development of a utility specialized to this task. For such a short motif (3–6 bases) multiple copies may be found by chance in any mRNA, and statistical tools have been developed to analyze this [10.71]. Functionally, multiple small dispersed E2 and VM1 elements provide, in combination, a specific binding site for the RNA-binding proteins in this case.

Notably many of the experimentally determined elements collected in the RBPDb are short (4–8 bases) and alone would not provide specificity. Computational tools that consider combinations of weakly informative sites have been used in other systems, e.g., for TFBS and miRNA targets. However, the functional relationships between multiple instances of the same or different sites are not usually known. For some sites, notably the important classes of AU-rich and CU-rich elements, programs to detect these sites operate by weighting multiple nearby repeats.

10.4.1 RNA Binding Protein Target Sites

Known protein binding sites are available from several public databases. The Rfam [10.15] database contains a growing group of covariance models for many *cis*-regulatory elements including RNA binding protein sites. Transterm [10.30] also contains patterns and descriptions of many known *cis*-regulatory elements – particularly protein binding sites. Nonredundant sets of sequences from NCBI or other uploaded sequences of interest may be searched for matches against these patterns. UTRsite [10.31] is a similar database of patterns. The RNA-Binding Protein DataBase [10.32] (RBPDb) catalogs proteins and their binding sites curated from the literature. Unstructured sequence motifs of binding sites may be downloaded and filtered by experiment type, species, and/or binding protein affinity.

These databases (Rfam, Transterm, UTRsite, and RBPDb) are useful in identifying known elements and may be used to find interesting candidates for testing. The reasoning behind this approach is to direct the identification of putative elements by their identity or similarity to known elements that have been experimentally demonstrated in other mRNAs. Another important usage of these databases is as a source of benchmarking datasets for any algorithms or pipelines designed to predict *cis*-regulatory elements de novo.

A note of caution is warranted. Database entries depend on manual curation based on literature review to remain current; For example, we have recently updated the Rfam model for the IRE [10.12] – the previous model was out of date and could not be used to identify many IREs that had been more recently experimentally demonstrated. It must also be considered that the models in Rfam, Transterm, and UTRsite inevitably vary in their sensitivity and specificity when identifying elements in target sequences.

Most of the entries in RBPDb are characterized by single sequences alone, and so if these data are to be used, the user must currently confine their search to be based on primary sequence only or build their own models to include predicted/demonstrated secondary structure. RBPDb does, however, contain some position weight matrices (PWMs) from systematic evolution of ligands by exponential enrichment (SELEX) experiments – the web interface may be employed to search mRNA sequences using these.

For some types of RNA–protein interactions it may be possible to predict either the binding from the protein structure, or the reverse. For a few classes of proteins this has been possible (e.g., PUF domain-containing

proteins), but research in this area is beyond the scope of this chapter [10.72, 73].

10.4.2 miRNA Target Sites

miRNAs have short ungapped seed sequences complementary to their target sites and act to downregulate expression. These targets are characterized by primary sequence with secondary structure normally impairing binding. Higher eukaryotic genomes may contain many hundreds of miRNAs (1,424 Human miRNAs in MiRBase release 17, 4/2011), and each one of those tested affects the expression of several hundred mRNAs. Some of the changes in expression are undoubtedly the result of indirect regulation; For example, transcription factors are regulated by miR-34a, so changes in miR-34a expression certainly have effects beyond the immediate targets [10.74].

There are numerous predicted miRNA binding sites, mainly in 3′ UTRs. It has been estimated that over 30% of human genes contain target sites. However, relatively few target sites have been validated experimentally. miRTarBase [10.75] listed the greatest number (2,819) of verified targets for the 269 human miRNAs that had been tested (4/2011). Databases of less reliable high-throughput data map over 150 000 targets to genomes [10.75].

Prediction of target sites for miRNAs is more straightforward than for proteins. However, there are many methods available which will detect different subclasses of sites with different accuracies; these have been recently reviewed [10.76], and ensemble methods that combine several different tools are available [10.77, 78]. Several examples of predictive tools are outlined below.

TargetScan [10.33] predicts miRNA target sites. Base pairing between the seed sequence at the 5′ end of the miRNA and the target mRNA and evolutionary conservation of the sites are the primary consideration. The software has been developed to additionally account for conservation at the seed region, minimum free energy of the hybridization including additional 3′ binding, flanking AU-rich sequence, proximity to additional miRNA target sites, and the position of the target site within the UTR [10.79, 80]. The UCSC genome browser is a straightforward way of accessing the TargetScan predictions. These have been calculated using conservation information based on multiple species alignment. The inclusion of this information reduces false positives and is probably a good idea for genome-wide analysis. If the target for analysis is restricted to several genes it may be worthwhile to consider nonconserved miRNA target sites by running the TargetScan program on unaligned sequence.

An alternative method is PicTar [10.34]. It allows the identification of target sites that have reduced conservation but are within mRNAs coexpressed with the targeting miRNA. The authors report that 30–50% of such sites are functional [10.81].

Another method, miRanda/mirSVR [10.82], has the distinction of allowing noncanonical G–U base pairs in the seed sequence; miRanda is also available on the web and has recently been extended to include a support vector regression algorithm – mirSVR [10.83]. This incorporates the relevant biological data, including expression data, into a scoring system and avoids a strict filter based on conservation.

An alternative method offering greater sensitivity to nonconserved target sites is ElMMo [10.36]. The algorithm uses Bayesian methods to assign priors calculated from the phylogenetic distribution of target sites for *each* miRNA. This allows miRNA-dependent adjustment of posterior probabilities for target sites with similar conservation patterns. The benefit of this is that conservation information is automatically tuned to each specific miRNA. The disadvantage is that it will be hard to match a nonconserved target site of an miRNA that has many widely conserved target sites. ElMMo is available via a web server and includes convenient filtering of mRNA targets to search using expression levels from numerous datasets, also providing Gene Ontology enrichment information on the identified targets.

10.5 De Novo Element Discovery

The discovery of new regulatory elements is a key goal for improved understanding of gene regulation. Recurring patterns in sequence and predicted structure may be detected and assessed for statistical significance. Some approaches for dealing with false positives caused by distracting sequence patterns have already been discussed. In de novo detection these steps are of particular importance – there is no model of even low specificity with which to begin.

There are many tools available for the detection of patterns in primary sequence, usually be-

cause they have been developed for finding DNA regulatory regions and this is then applied to RNA.

This is despite the demonstrated importance of structure in many regulatory elements. The reason for this lies in the difficulty of secondary structure prediction and that primary sequence patterns are still characteristic of structured elements.

10.5.1 Primary Sequence Analysis for Elements Lacking Significant Secondary Structure

A number of enumerative methods are available for de novo detection of primary sequence patterns. The patterns being sought are generally far shorter than the genomic and transcript sequences in which they

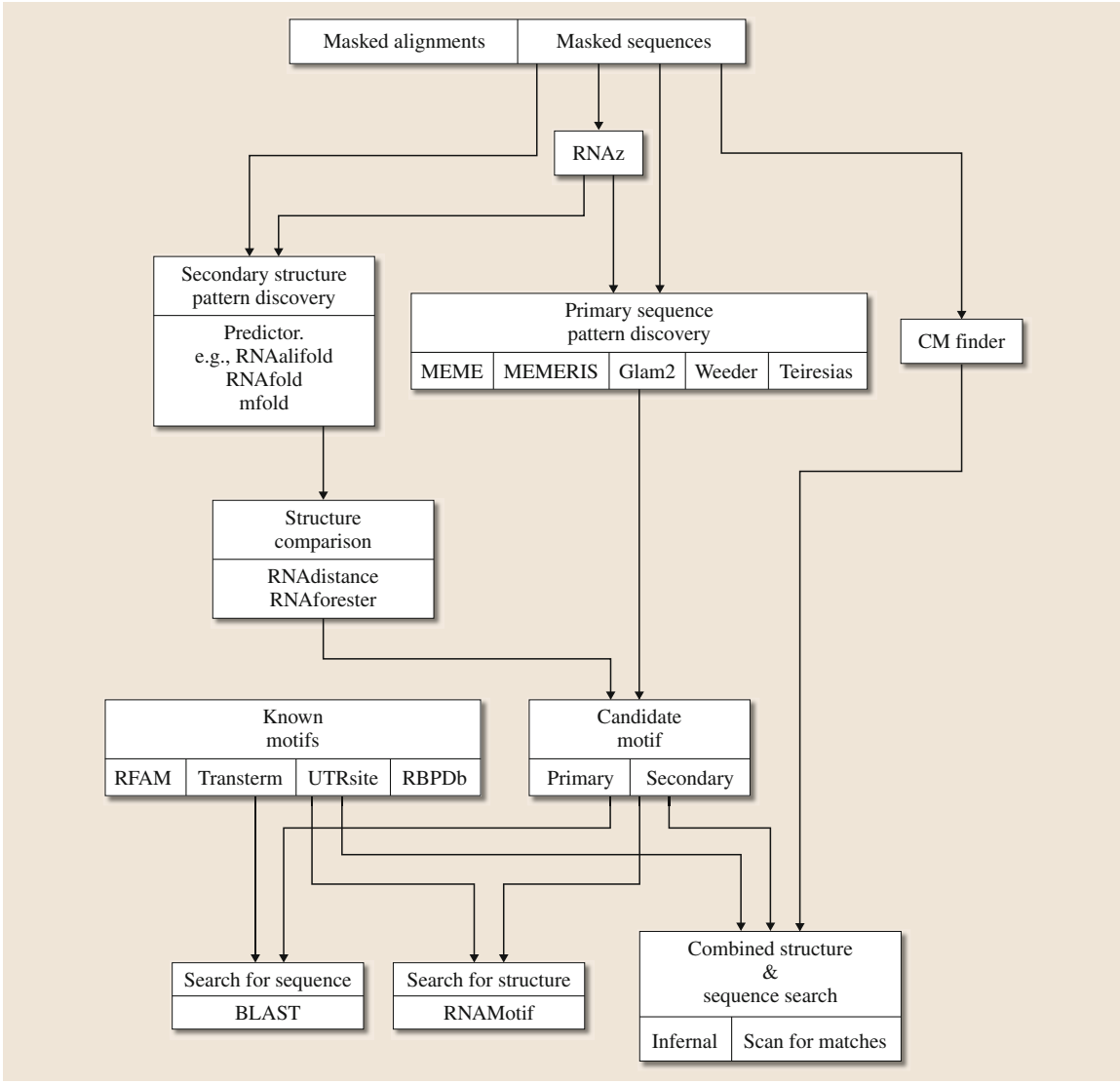

Fig. 10.4 Discovery of motifs starting from input sequences or alignments that are similarly regulated. The general goal of each component is shown, with some specific examples of currently available software named

are contained. Real primary sequence elements may be degenerate, gapped, redundant, and repetitive. Several differently patterned elements may be responsible for similar regulatory outcomes in different transcripts. All these factors contribute to the difficulty of detection.

Overrepresented patterns in a set of unaligned input sequences may be identified by multiple expectation maximization for motif elicitation (MEME) [10.37]. A background model may be provided for this analysis. The output motifs consist of position weight matrices (PWMs) showing the probability of a particular nucleotide at each position within the motif. MEME will not consider gaps in a motif. GLAM2 [10.84] allows gaps in the matched primary sequence, but it does not include these gaps in the output motifs.

An alternative approach offered by *Weeder* [10.38] involves building suffix trees from a set of input sequences. These are used to find all patterns of a set length, occurring in at least a certain number of sequences, with an upper limit on the number of mismatches (mutations). The program may be run in an automated way multiple times to detect patterns of different lengths.

The TEIRESIAS [10.39] algorithm is not restricted to searching for patterns of specific length and can detect gapped patterns. This is computationally more intensive, in both memory and processing requirements. The large number of results requires further processing for statistical significance.

The application of these methods is shown in overview in Fig. 10.4. Further methods are reviewed and benchmarked elsewhere [10.85].

10.5.2 Secondary Structure Prediction for Structured Element Discovery

Relatively few confirmed secondary or tertiary structures for *cis*-regulatory elements are available. Therefore, predictions of RNA structures are made computationally. High-throughput methods that may allow more structures to be determined experimentally are becoming available [10.86]. However, these methods are limited, and a combination of bioinformatic and high-throughput experiment has been successful [10.18].

In addition, it may be possible to predict the three-dimensional structures of RNAs using bioinformatics, which will become increasingly feasible as the number of known structures increases [10.87]. Packages are available to assist in tertiary structural prediction [10.88, 89]. These may be used with sequences of interest alone or in combination with available experimental data on similar structures.

Single Sequences

Predicting folding on individual sequences is a common technique. This may be done globally (for the entire mRNA) or locally on windows within a biologically relevant section (e.g., 80-base windows in 3′ UTRs). A global RNA fold prediction algorithm from the *Zuker* laboratory is implemented by the mfold [10.43] program. These methods are commonly used, and the paper associated with this program has over 700 citations in the literature. The Vienna RNA package provides a similar program, RNAfold [10.44]. Like mfold, this calculates predicted secondary structure for RNAs based on minimum free energies (MFE) using conformations derived from published values for stacking and destabilizing energies.

The UNAFold [10.90] software is a development of mfold and further predicts hybridization and melting profiles. Like RNAfold, it produces dotplots showing pairing probabilities over the sequence (Fig. 10.5). The dotplots from both programs include the pairing probabilities corresponding to suboptimal (predicted) folding of the input sequence. Suboptimal structures are discussed later.

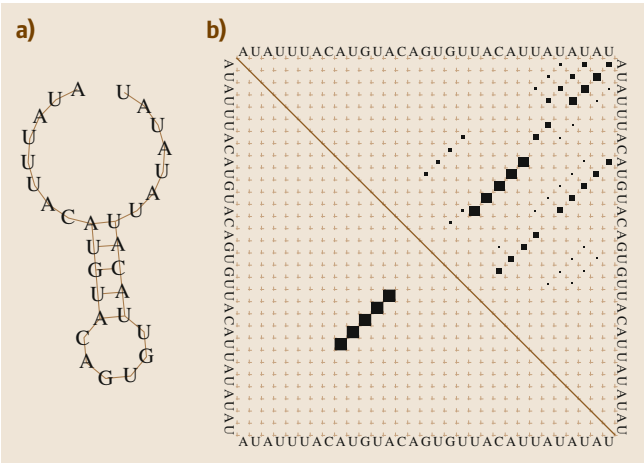

Fig. 10.5a,b Predicted secondary structures for the IRE in human *CDC14A*. **(a)** Optimal MFE structure from RNAfold (−2.9 kcal/mol). The dot plot **(b)** represents the ensemble of structures. Some suboptimal structures contain predicted pairs in a lower stem, their presence being more consistent with other IRE structures. The C–G base pair in the loop observed in the experimental structure **(b)** is not predicted by MFE methods

Local folding methods are likely to be more appropriate for small *cis*-regulatory elements. Several algorithms which are particularly suitable have been developed. These perform better on known *cis*-regulatory elements [10.91]. This is an area of active development and testing; current methods include RNAplfold and Rfold [10.92, 93]. In addition, methods that combine structure prediction with comparison with known secondary structures of homologous sequences are being developed [10.94].

RNA structure is dynamic. Binding interactions with a structured *cis*-regulatory element are likely to further influence the structure of that element. A limited set of thermodynamic measurements provides the basis of the RNA structure predicting algorithms. Furthermore, these methods do not account for pseudoknots, G-quartets, and other structures. The predicted structures cannot be expected to entirely model the complex molecular interactions found in the in vivo environment. Approaches using homology to known secondary and tertiary structures that are not necessarily the MFE structures should assist in this [10.87, 94, 95].

Some elements may contain pseudoknots, but the prediction of pseudoknots remains computationally slow and is only tractable for shorter RNA sequences. The pknots [10.96] program and HotKnots [10.97] are methods that may be employed on targeted regions if such structures are expected [10.98]. There are *cis*-regulatory elements containing pseudoknots, for example, frameshift elements. Further examples may be found in general and specific databases [10.15, 99].

Predicted structures do not always match natively observed structures. An example is the iron responsive element (IRE) found in the human mRNA for *CDC14A*. Both the UNAfold and RNAFold program find the same MFE. Additional base pairs are seen in the dotplot, some of which would make a lower stem observed in all IREs tested. Some pairs are not predicted at all by MFE approaches, e.g., the C–G in the apical loop (Fig. 10.5).

When there is existing information about the structure of a particular *cis*-regulatory element, these constraints may be provided to the folding programs. This allows the estimation of the MFE in a candidate element even when this is not the lowest resulting from prediction. Other sources, such as gene expression and phylogenetic information, may then be used in combination to arrive at strong candidates for experimental testing.

Benchmarking on ncRNA genes has shown both sensitivity and specificity of RNA structure prediction using MFE methods to be limited (22–63% and 20–60%, respectively) [10.100]. Newer algorithms improve on this [10.101]. Although secondary and tertiary structure is a factor in RNA interaction, the difficulty of experimentally determining these structures and of accurately predicting them must be always borne in mind.

These programs do not allow noncanonical bases, e.g., U–U or A–G, which have been observed in many experimentally determined RNA structures. Several algorithms do allow predictions that include these noncanonical pairs [10.89, 101]. Though these are considerably slower, they may be used with short structured *cis*-regulatory elements.

Multiple Sequences

Calculating the consensus structure for aligned sequences can overcome some of the shortcomings in the accuracy of MFE calculations for single sequences. This approach depends on the ability to obtain a reasonable alignment of the primary sequences. Methods available include RNAalifold [10.45] – part of the Vienna RNA package.

Covariance (e.g., an A–U base pair being exchanged by some other pairing) or compatible mutation (e.g., a G–C base pair being exchanged by a G–U pair) help tools such as RNAalifold to provide an optimal structure consistent with the alignment. However, too many variations in the primary sequence will make alignment at this level impossible.

A further class of algorithm simultaneously folds and aligns input sequences. The approach is computationally more intensive, though may be useful where the primary sequence alignments have limited similarity [10.41]. Dynalign, Multilign, and Turbofold are all part of the RNAstructure package [10.14]. The original Dynalign works with two sequences – Multilign operates on multiple sequences. Turbofold does not produce an alignment but presents separate structures for each of the sequences in the input, rather than one consensus structure. Structures are predicted based on pairing probabilities in each sequence severally, combined with the pairing probabilities in a consensus model [10.102].

Foldalign and FoldalignM [10.103] will produce local and global alignments along with structure predictions. An interesting feature of Foldalign is that it will attempt local alignments over the input sequences (based on structure and primary sequence) and then report the best alignments found. It is therefore also an element detection tool and not limited to structural prediction. Additional discussion and comparison of RNA

fold prediction algorithms can be found in recent reviews [10.14, 87, 104].

10.5.3 Comparing Secondary Structures

It is useful to compare two structures to assess their similarity. In some cases the best information about a regulatory element may be its secondary structure, and so this is key in finding similar elements. RNAdistance, RNAforester, and Cofolga2mo [10.105] allow these comparisons for simple secondary structures. A simplification of the problem is to compare overall shapes, e.g., stem loop or cloverleaf [10.106].

The Vienna RNA package provides RNAdistance [10.44], which allows not only comparison of pairs of structures but also simultaneous comparison of multiple structures, providing a comparison matrix for all input structures. The output quantifies the differences between the structures. A pipeline on the Vienna RNA servers, *Structure Conservation Analysis*, includes this method. Input is an alignment and RNAalifold is used to predict MFE structures that are compared with predicted structures of individual sequences.

The RNAforester [10.46] algorithm builds tree-like data structures that represent RNA secondary structure. These can then be used to build multiple alignments of different RNA structures. Thresholds may be applied determining whether a particular structure is sufficiently similar to form part of an aligned group. This allows the degree of similarity between structures to be assessed as well as the grouping of RNAs into structurally determined families. Another useful output of this tool is an alignment of input sequences that is wholly determined by the given structures. This can be useful to build a seed for a covariance model.

10.5.4 Searching for Secondary or Tertiary Structural Elements

When there is good evidence for structure but the specificity of the primary sequence in the regulatory element is largely or completely unknown, a search based on structure alone is required. RNAMotif [10.48] allows the creation of a pattern which has no or little information about primary sequence. The resulting matches may be used to find additional sequences for testing; alignments of these will hopefully allow the incorporation of primary sequence information into the model for the regulatory element in question. Three-dimensional (3-D) motifs (for example, *G-bulges* from a lysine riboswitch can be searched for with RMdetect [10.95].

10.5.5 Combining Primary and Secondary Structural Search Methods

A covariance model (CM) is a stochastic context-free grammar (SCFG) that can be used to model the consensus sequence and structure of RNAs. Given an alignment to a consensus secondary structure, not only nucleotide residues at single-stranded positions but also base pairings, insertions, and deletions are scored. The Infernal [10.50] software package provides the tools to build covariance models and to search for matches to the model over a target sequence. The resulting *bit score* is the log-odds ratio of the probability of the target matching the model to the probability of target matching random sequence. This methodology is key to the Rfam database which catalogs RNA families using these models, showing their paralogs and homologs [10.15].

When sufficient information exists about a *cis*-regulatory element such that known examples may be meaningfully aligned to a consensus structure, a covariance model may be constructed. This model may be used to search for matches within other mRNA sequences – as has been done in the case of the IRE [10.12] and other *cis*-regulatory elements in the Rfam database. This can result in new candidates for experiment testing.

An alternative to using covariance models is to build a pattern corresponding to a motif. This is the approach taken by Transterm and UTRsite. A useful tool for interpreting and searching using such patterns is scan_for_matches [10.51]. The pattern descriptions used by this tool can incorporate structural information. Such a pattern does not depend on being able to construct an alignment of known sequence elements. However, complex patterns can be difficult to construct, and the ability to make a good pattern depends on prior knowledge of a motif, including permissible variations at different points. The output from scan_for_matches does not include the statistical information provided by the Infernal software's method.

10.6 Combinatorial Methods

Figures 10.6 and 10.4 give an overview of how the tools discussed here may be used together in the search for *cis*-regulatory elements. Some combinations of tools are sufficiently novel to be considered new methods in their own right. These methods are discussed here.

Unpaired bases are more likely to be involved directly with RNA interaction, and certain protein interactions. Consequently, primary sequence patterns found in unpaired regions are of particular interest. This can be seen for example in the IRE, where the unpaired nucleotides have been shown to interact with the iron regulatory proteins. MEMERIS [10.53] is an extension of the MEME algorithm. A script that comes with the package first uses the Vienna RNA tools to predict secondary structure across the input sequences (which must be in FASTA format with all sequence on one line). Altering the prior probabilities for putative motif start sites directs the MEME algorithm towards predicted single-stranded positions. The weakness of not being able to detect gapped patterns is inherited from MEME.

Given the importance of structure in *cis*-regulatory elements, another useful approach is to identify sequence regions likely to have conserved structure. RNAz [10.54] was originally developed to detect structured ncRNAs in genomic sequence, but this method is also of potential interest in *cis*-regulatory element discovery and can successfully detect several known elements. The software takes a sequence alignment as input and uses RNAalifold to calculate the MFE of consensus structures arising over a sliding window. MFE values for these same sequences are also severally calculated using RNAfold. The consensus MFE in ratio to the average single sequence MFE gives a structure conservation index (SCI). Additionally, a z-score is calculated, which represents the deviation of the MFE score from random sequence of similar composition and length. The z-score and the SCI are used as inputs to a support vector machine together with the number of aligned sequences and the pairwise identity of the sequences – this results in a probability value for the occurrence of a conserved structural motif. RNAz may be run from the command line or on the Vienna RNA servers. Also the ncRNA [10.23] site has a version of the UCSC genome browser that includes a track for sites predicted by RNAz.

For multiple sequences with common function, automated production of covariance models for elements with similar sequence and structure is provided by the CMFinder [10.49] program, which takes short unaligned sequences (< 500 bp) as input. Based on MFE, the algorithm first selects a number of candidates from within these sequences. The candidates are

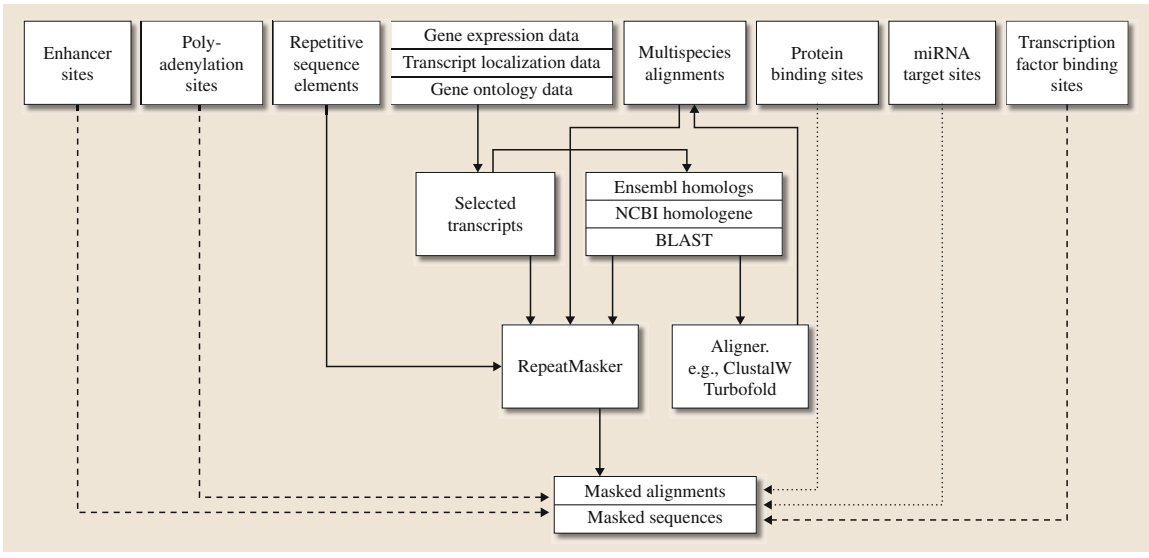

Fig. 10.6 Preparing and obtaining sequences and alignments. *Solid lines* indicate basic approach. *Dashed lines* indicate additional steps appropriate to genome-scale analysis or de novo element discovery. *Dotted lines* indicate additional processes dependent on the specific problem

aligned according to predicted secondary structure and an expectation-maximization algorithm used to refine a covariance model that identifies elements within the candidate sequences distinct from a background distribution. The input sequences are then rescanned using this covariance model, and the top hits are included as candidates. The authors of CMFinder note that identifying larger motifs is problematic and go some way to addressing this by attempting to merge smaller motifs as a final step.

The importance of incorporating biological data into the element discovery process has already been discussed. Selection of sequence for pattern discovery together with appropriate background models is an important first step in many analyses. The finding informative regulatory elements (FIRE) [10.55] pipeline automates this process by combining the detection of primary sequence patterns with other biological data – either discrete or continuous data (e.g., gene expression data). FIRE starts by looking for 7-mer *seed* motifs, but these can be extended one base in either direction. The initial *seeds* are systematically modified using degenerate International Union of Pure and Applied Chemistry (IUPAC) codes to arrive at a motif most significantly associated with the other biological data. This pipeline also offers the convenience of displaying gene ontology terms with enriched association to the identified motifs.

10.7 Conclusions and Future Prospects

The importance of posttranscriptional regulation is becoming increasingly apparent as large-scale proteomic data become available. Transcripts are translated with a wide range of efficiencies, giving differing numbers of functional proteins per message.

It has now become almost routine to measure transcript levels at a genome scale using microarrays or next-generation sequencing. The expectation was that transcript abundance would provide good estimates of protein abundance in the cell. However, the early studies done over a decade ago that suggested that mRNA levels might predict less than 45% of protein levels have been reiterated by recent studies. This indication of widespread posttranscriptional control is seen in many organisms.

High-throughput *wet-lab* studies and analysis of regulatory elements will facilitate discovery of elements with widely conserved functions. However, it should also be noted that some key elements might only be found in a small number of messages or species, e.g., human-specific miRNA targets, or the targets of other noncoding RNAs. These exceptional elements are a challenge for both bioinformatic and wet-lab studies and may also be of critical importance for cell growth and development, and have applications in biotechnology.

At least some of the variation in the amount of protein translated from individual mRNAs will be mediated by *cis*-regulatory elements in the mRNAs. This chapter has outlined current bioinformatic methods available for their discovery. Development of new methods and the use of high-throughput data on a genome-wide scale, particularly comparative genomic, proteomic, and high-throughput transcriptomic data, will facilitate this.

References

10.1 C. Vogel: Translation's coming of age, Mol. Syst. Biol. **7**, 498 (2011)

10.2 S.A. Tenenbaum, J. Christiansen, H. Nielsen: The post-transcriptional operon, Methods Mol. Biol. **703**, 237–245 (2011)

10.3 D.J. Hogan, D.P. Riordan, A.P. Gerber, D. Herschlag, P.O. Brown: Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system, PLoS Biology **6**, e255 (2008)

10.4 M.J. Moore: From birth to death: The complex lives of eukaryotic mRNAs, Science **309**, 1514–1518 (2005)

10.5 P.A. Galante, D. Sandhu, R. de Sousa Abreu, M. Gradassi, N. Slager, C. Vogel, S.J. de Souza, L.O. Penalva: A comprehensive in silico expression analysis of RNA binding proteins in normal and tumor tissue: Identification of potential players in tumor formation, RNA Biology **6**, 426–433 (2009)

10.6 N.G. Tsvetanova, D.M. Klass, J. Salzman, P.O. Brown: Proteome-wide search reveals unexpected RNA-binding proteins in *Saccharomyces cerevisiae*, PLoS One **5**, e12671 (2010)

10.7 D.P. Bartel: MicroRNAs: Target recognition and regulatory functions, Cell **136**, 215–233 (2009)

10.8 P. Chartrand, X.H. Meng, S. Huttelmaier, D. Donato, R.H. Singer: Asymmetric sorting of ash1p in yeast results from inhibition of translation by localization elements in the mRNA, Mol. Cell. **10**, 1319–1330 (2002)

10.9 M.W. Hentze, S.W. Caughman, T.A. Rouault, J.G. Barriocanal, A. Dancis, J.B. Harford, R.D. Klausner: Identification of the iron-responsive element for the translational regulation of human ferritin mRNA, Science **238**, 1570–1573 (1987)

10.10 S. Castellano, V.N. Gladyshev, R. Guigo, M.J. Berry: SelenoDB 1.0: A database of selenoprotein genes, proteins and SECIS elements, Nucleic Acids Res. **36**, D332–338 (2008)

10.11 M. Davila Lopez, T. Samuelsson: Early evolution of histone mRNA 3′ end processing, RNA **14**, 1–10 (2008)

10.12 S.G. Stevens, P.P. Gardner, C. Brown: Two covariance models for iron-responsive elements, RNA Biology **8**, 792–801 (2011)

10.13 R. Backofen, S.H. Bernhart, C. Flamm, C. Fried, G. Fritzsch, J. Hackermüller, J. Hertel, I.L. Hofacker, K. Missal, A. Mosig, S.J. Prohaska, D. Rose, P.F. Stadler, A. Tanzer, S. Washietl, S. Will: RNAs everywhere: Genome-wide annotation of structured RNAs, J. Exp. Zool. B **308**, 1–25 (2007)

10.14 D.H. Mathews, W.N. Moss, D.H. Turner: Folding and finding RNA secondary structure, Cold Spring Harb. Perspect. Biol. **2**, a003665 (2010)

10.15 P.P. Gardner, J. Daub, J. Tate, B.L. Moore, I.H. Osuch, S. Griffiths-Jones, R.D. Finn, E.P. Nawrocki, D.L. Kolbe, S.R. Eddy, A. Bateman: Rfam: Wikipedia, clans and the "decimal" release, Nucleic Acids Res. **39**, D141–145 (2011)

10.16 M. Andronescu, V. Bereg, H.H. Hoos, A. Condon: RNA STRAND: The RNA secondary structure and statistical analysis database, BMC Bioinformatics **9**, 340 (2008)

10.17 K. Rother, M. Rother, M. Boniecki, T. Puton, J.M. Bujnicki: RNA and protein 3-D structure modeling: Similarities and differences, J. Mol. Model. **17**, 2325–2336 (2011)

10.18 J.G. Underwood, A.V. Uzilov, S. Katzman, C.S. Onodera, J.E. Mainzer, D.H. Mathews, T.M. Lowe, S.R. Salama, D. Haussler: FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing, Nat. Methods **7**, 995–1001 (2010)

10.19 D.P. Riordan, D. Herschlag, P.O. Brown: Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome, Nucleic Acids Res. **39**, 1501–1509 (2011)

10.20 Y. Wan, M. Kertesz, R.C. Spitale, E. Segal, H.Y. Chang: Understanding the transcriptome through RNA structure, Nat. Rev. Genet. **12**, 641–655 (2011)

10.21 S. Kishore, S. Luber, M. Zavolan: Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression, Brief Funct. Genomics **9**, 391–404 (2010)

10.22 W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler: The human genome browser at UCSC, Genome Res. **12**, 996–1006 (2002)

10.23 T. Mituyama, K. Yamada, E. Hattori, H. Okida, Y. Ono, G. Terai, A. Yoshizawa, T. Komori, K. Asai: The Functional RNA Database 3.0: Databases to support mining and annotation of functional RNAs, Nucleic Acids Res. **37**, D89–92 (2009)

10.24 J. Goecks, A. Nekrutenko, J. Taylor: Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol. **11**, R86 (2010)

10.25 J. Jurka: Repbase update: A database and an electronic journal of repetitive elements, Trends Genet. **16**, 418–420 (2000)

10.26 V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, E. Wingender: TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes, Nucleic Acids Res. **34**, D108–110 (2006)

10.27 J.C. Bryne, E. Valen, M.H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, A. Sandelin: JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update, Nucleic Acids Res. **36**, D102–106 (2008)

10.28 S. Mahony, P.V. Benos: STAMP: A web tool for exploring DNA-binding motif similarities, Nucleic Acids Res. **35**, W253–258 (2007)

10.29 K.D. Pruitt, T. Tatusova, W. Klimke, D.R. Maglott: NCBI Reference Sequences: Current status, policy and new initiatives, Nucleic Acids Res. **37**, D32–36 (2009)

10.30 G.H. Jacobs, A. Chen, S.G. Stevens, P.A. Stockwell, M.A. Black, W.P. Tate, C.M. Brown: Transterm: A database to aid the analysis of regulatory sequences in mRNAs, Nucleic Acids Res. **37**, D72–76 (2009)

10.31 G. Grillo, A. Turi, F. Licciulli, F. Mignone, S. Liuni, S. Banfi, V.A. Gennarino, D.S. Horner, G. Pavesi, E. Picardi, G. Pesole: UTRdb and UTRsite (RELEASE 2010): A collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs, Nucleic Acids Res. **38**, D75–80 (2009)

10.32 K.B. Cook, H. Kazan, K. Zuberi, Q. Morris, T.R. Hughes: RBPDB: A database of RNA-binding specificities, Nucleic Acids Res. **39**, D301–308 (2010)

10.33 B.P. Lewis, C.B. Burge, D.P. Bartel: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, Cell **120**, 15–20 (2005)

10.34 A. Krek, D. Grun, M.N. Poy, R. Wolf, L. Rosenberg, E.J. Epstein, P. MacMenamin, I. da Piedade, K.C. Gunsalus, M. Stoffel, N. Rajewsky: Combinatorial microRNA target predictions, Nat. Genet. **37**, 495–500 (2005)

10.35 A. Kozomara, S. Griffiths-Jones: miRBase: Integrating microRNA annotation and deep-sequencing data, Nucleic Acids Res. **39**, D152–157 (2010)

10.36 D. Gaidatzis, E. van Nimwegen, J. Hausser, M. Zavolan: Inference of miRNA targets using evolutionary conservation and pathway analysis, BMC Bioinformatics **8**, 69 (2007)

10.37 T.L. Bailey, N. Williams, C. Misleh, W.W. Li: MEME: Discovering and analyzing DNA and protein sequence motifs, Nucleic Acids Res. **34**, W369–373 (2006)

10.38 G. Pavesi, G. Mauri, G. Pesole: An algorithm for finding signals of unknown length in DNA sequences, Bioinformatics **17**(Suppl. 1), S207–214 (2001)

10.39 I. Rigoutsos, A. Floratos: Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm, Bioinformatics **14**, 55–67 (1998)

10.40 A.D. George, S.A. Tenenbaum: Web-based tools for studying RNA structure and function, Methods Mol. Biol. **703**, 67–86 (2011)

10.41 R.S. Hamilton, I. Davis: Identifying and searching for conserved RNA localisation signals, Methods Mol. Biol. **714**, 447–466 (2011)

10.42 Wikipedia: List of RNA structure prediction software (2012), available at http://en.wikipedia.org/wiki/List_of_RNA_structure_prediction_software

10.43 M. Zuker, P. Stiegler: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, Nucleic Acids Res. **9**, 133–148 (1981)

10.44 I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, P. Schuster: Fast folding and comparison of RNA secondary structures, Monatsh. Chem./Chem. Mon. **125**, 167–188 (1994)

10.45 S.H. Bernhart, I.L. Hofacker, S. Will, A.R. Gruber, P.F. Stadler: RNAalifold: Improved consensus structure prediction for RNA alignments, BMC Bioinformatics **9**, 474 (2008)

10.46 M. Hochsmann, B. Voss, R. Giegerich: Pure multiple RNA secondary structure alignments: A progressive profile approach, IEEE/ACM Trans. Comput. Biol. Bioinform. **1**, 53–62 (2004)

10.47 C.W. Wang, K.T. Chen, C.L. Lu: iPARTS: An improved tool of pairwise alignment of RNA tertiary structures, Nucleic Acids Res. **38**, W340–347 (2010)

10.48 T.J. Macke, D.J. Ecker, R.R. Gutell, D. Gautheret, D.A. Case, R. Sampath: RNAMotif, an RNA secondary structure definition and search algorithm, Nucleic Acids Res. **29**, 4724–4735 (2001)

10.49 Z. Yao, Z. Weinberg, W.L. Ruzzo: CMfinder – A covariance model based RNA motif finding algorithm, Bioinformatics **22**, 445–452 (2006)

10.50 E.P. Nawrocki, D.L. Kolbe, S.R. Eddy: Infernal 1.0: Inference of RNA alignments, Bioinformatics **25**, 1335–1337 (2009)

10.51 M. Dsouza, N. Larsen, R. Overbeek: Searching for patterns in genomic data, Trends Genet. **13**, 497–498 (1997)

10.52 T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, R.N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, A. Soboleva: NCBI GEO: Archive for functional genomics data sets – 10 years on, Nucleic Acids Res. **39**, D1005–1010 (2010)

10.53 M. Hiller, R. Pudimat, A. Busch, R. Backofen: Using RNA secondary structures to guide sequence motif finding towards single-stranded regions, Nucleic Acids Res. **34**, e117 (2006)

10.54 S. Washietl, I.L. Hofacker, P.F. Stadler: Fast and reliable prediction of noncoding RNAs, Proc. Natl. Acad. Sci. USA **102**, 2454–2459 (2005)

10.55 O. Elemento, N. Slonim, S. Tavazoie: A universal framework for regulatory element discovery across all genomes and data types, Mol. Cell. **28**, 337–350 (2007)

10.56 E. Lecuyer, H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T.R. Hughes, P. Tomancak, H.M. Krause: Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function, Cell **131**, 174–187 (2007)

10.57 P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A.O. Kamphorst, M. Landthaler, C. Lin, N.D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foà, J. Schliwka, U. Fuchs, A. Novosel, R.U. Müller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D.B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H.I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C.E. Rogler, J.W. Nagle, J. Ju, F.N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M.J. Brownstein, A. Bosio, A. Borkhardt, J.J. Russo, C. Sander, M. Zavolan, T. Tuschl: A mammalian microRNA expression atlas based on small RNA library sequencing, Cell **129**, 1401–1414 (2007)

10.58 D. Betel, M. Wilson, A. Gabow, D.S. Marks, C. Sander: The microRNA.org resource: Targets and expression, Nucleic Acids Res. **36**, D149–153 (2008)

10.59 M.A. Batzer, P.L. Deininger: Alu repeats and human genomic diversity, Nat. Rev. Genet. **3**, 370–379 (2002)

10.60 A. Smit, R. Hubley, P. Green: *RepeatMasker Open-3.0*. (1996–2010), available at http://www.repeatmasker.org

10.61 S. Hannenhalli: Eukaryotic transcription factor binding sites–modeling and integrative search methods, Bioinformatics **24**, 1325–1331 (2008)

10.62 D. Schmidt, M.D. Wilson, B. Ballester, P.C. Schwalie, G.D. Brown, A. Marshall, C. Kutter, S. Watt,

C.P. Martinez-Jimenez, S. Mackay, I. Talianidis, P. Flicek, D.T. Odom: Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding, Science **328**, 1036–1040 (2010)

10.63 A. Visel, E.M. Rubin, L.A. Pennacchio: Genomic views of distant-acting enhancers, Nature **461**, 199–205 (2009)

10.64 L.A. Pennacchio, N. Ahituv, A.M. Moses, S. Prabhakar, M.A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K.D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B.L. Black, O. Couronne, M.B. Eisen, A. Visel, E.M. Rubin: In vivo enhancer analysis of human conserved non-coding sequences, Nature **444**, 499–502 (2006)

10.65 S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman: Basic local alignment search tool, J. Mol. Biol. **215**, 403–410 (1990)

10.66 S.R. Eddy: Profile hidden Markov models, Bioinformatics **14**, 755–763 (1998)

10.67 J.O. Deshler, M.I. Highett, B.J. Schnapp: Localization of Xenopus Vg1 mRNA by Vera protein and the endoplasmic reticulum, Science **276**, 1128–1131 (1997)

10.68 D. Gautreau, C.A. Cote, K.L. Mowry: Two copies of a subelement from the Vg1 RNA localization sequence are sufficient to direct vegetal localization in Xenopus oocytes, Development **124**, 5013–5020 (1997)

10.69 S. Kwon, T. Abramson, T.P. Munro, C.M. John, M. Kohrmann, B.J. Schnapp: UUCAC- and vera-dependent localization of VegT RNA in Xenopus oocytes, Curr. Biol. **12**, 558–564 (2002)

10.70 S. Choo, B. Heinrich, J.N. Betley, Z. Chen, J.O. Deshler: Evidence for common machinery utilized by the early and late RNA localization pathways in Xenopus oocytes, Dev. Biol. **278**, 103–117 (2005)

10.71 B.B. Andken, I. Lim, G. Benson, J.J. Vincent, M.T. Ferenc, B. Heinrich, L.A. Jarzylo, H.Y. Man, J.O. Deshler: 3′-UTR SIRF: A database for identifying clusters of short interspersed repeats in 3′ untranslated regions, BMC Bioinformatics **8**, 274 (2007)

10.72 P.P. Tam, I.H. Barrette-Ng, D.M. Simon, M.W. Tam, A.L. Ang, D.G. Muench: The Puf family of RNA-binding proteins in plants: Phylogeny, structural modeling, activity and subcellular localization, BMC Plant Biol. **10**, 44 (2010)

10.73 I. Tuszynska, J.M. Bujnicki: DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking, BMC Bioinformatics **12**, 348 (2011)

10.74 M. Kaller, S.T. Liffers, S. Oeljeklaus, K. Kuhlmann, S. Roh, R. Hoffmann, B. Warscheid, H. Hermeking: Genome-wide characterization of miR-34a induced changes in protein and mRNA expression by a combined pulsed SILAC and microarray analysis, Mol. Cell. Proteomics **10**(M111), 010462 (2011)

10.75 S.D. Hsu, F.M. Lin, W.Y. Wu, C. Liang, W.C. Huang, W.L. Chan, W.T. Tsai, G.Z. Chen, C.J. Lee, C.M. Chiu, C.H. Chien, M.C. Wu, C.Y. Huang, A.P. Tsou, H.D. Huang: miRTarBase: A database curates experimentally validated microRNA-target interactions, Nucleic Acids Res. **39**, D163–169 (2011)

10.76 M. Thomas, J. Lieberman, A. Lal: Desperately seeking microRNA targets, Nat. Struct. Mol. Biol. **17**, 1169–1174 (2010)

10.77 F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, T. Li: miRecords: An integrated resource for microRNA-target interactions, Nucleic Acids Res. **37**, D105–110 (2009)

10.78 H. Dweep, C. Sticht, P. Pandey, N. Gretz: miRWalk-database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes, J. Biomed. Inform. **44**, 839–847 (2011)

10.79 R.C. Friedman, K.K. Farh, C.B. Burge, D.P. Bartel: Most mammalian mRNAs are conserved targets of microRNAs, Genome Res. **19**, 92–105 (2009)

10.80 A. Grimson, K.K. Farh, W.K. Johnston, P. Garrett-Engele, L.P. Lim, D.P. Bartel: MicroRNA targeting specificity in mammals: Determinants beyond seed pairing, Mol. Cell. **27**, 91–105 (2007)

10.81 K. Chen, N. Rajewsky: Natural selection on human microRNA binding sites inferred from SNP data, Nat. Genet. **38**, 1452–1456 (2006)

10.82 B. John, A.J. Enright, A. Aravin, T. Tuschl, C. Sander, D.S. Marks: Human MicroRNA targets, PLoS Biol. **2**, e363 (2004)

10.83 D. Betel, A. Koppal, P. Agius, C. Sander, C. Leslie: Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites, Genome Biol. **11**, R90 (2010)

10.84 M.C. Frith, N.F. Saunders, B. Kobe, T.L. Bailey: Discovering sequence motifs with arbitrary insertions and deletions, PLoS Comput. Biol. **4**, e1000071 (2008)

10.85 M. Tompa, N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, V.J. Makeev, A.A. Mironov, W.S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, Z. Zhu: Assessing computational tools for the discovery of transcription factor binding sites, Nat. Biotechnol. **23**, 137–144 (2005)

10.86 E. Westhof, P. Romby: The RNA structurome: High-throughput probing, Nat. Methods **7**, 965–967 (2010)

10.87 E. Westhof, B. Masquida, F. Jossinet: Predicting and modeling RNA architecture, Cold Spring Harb. Perspect. Biol. **3**, a003632 (2011)

10.88 F. Jossinet, T.E. Ludwig, E. Westhof: Assemble: An interactive graphical tool to analyze and build RNA architectures at the 2-D and 3-D levels, Bioinformatics **26**, 2057–2059 (2010)

10.89 M. Parisien, F. Major: The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data, Nature **452**, 51–55 (2008)
10.90 N.R. Markham, M. Zuker: UNAFold: Software for nucleic acid folding and hybridization, Methods Mol. Biol. **453**, 3–31 (2008)
10.91 S.J. Lange, D. Maticzka, M. Mohl, J.N. Gagnon, C.M. Brown, R. Backofen: Global or local? Predicting secondary structure and accessibility in mRNAs, Nucleic Acids Res. **40**, 5215–5216 (2012)
10.92 S.H. Bernhart, U. Muckstein, I.L. Hofacker: RNA accessibility in cubic time, Algorithms Mol. Biol. **6**, 3 (2011)
10.93 H. Kiryu, G. Terai, O. Imamura, H. Yoneyama, K. Suzuki, K. Asai: A detailed investigation of accessibilities around target sites of siRNAs and miRNAs, Bioinformatics **27**, 1788–1797 (2011)
10.94 M. Hamada, K. Yamada, K. Sato, M.C. Frith, K. Asai: CentroidHomfold-LAST: Accurate prediction of RNA secondary structure using automatically collected homologous sequences, Nucleic Acids Res. **39**, W100–W106 (2011)
10.95 J.A. Cruz, E. Westhof: Sequence-based identification of 3-D structural modules in RNA with RMDetect, Nat. Methods **8**, 513–519 (2011)
10.96 E. Rivas, S.R. Eddy: A dynamic programming algorithm for RNA structure prediction including pseudoknots, J. Mol. Biol. **285**, 2053–2068 (1999)
10.97 J. Ren, B. Rastegari, A. Condon, H.H. Hoos: HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots, RNA **11**, 1494–1504 (2005)
10.98 S. Bellaousov, D.H. Mathews: ProbKnot: Fast prediction of RNA secondary structure including pseudoknots, RNA **16**, 1870–1880 (2010)
10.99 M. Bekaert, A.E. Firth, Y. Zhang, V.N. Gladyshev, J.F. Atkins, P.V. Baranov: Recode-2: New design, new search tools, and many more genes, Nucleic Acids Res. **38**, D69–D74 (2010)
10.100 P.P. Gardner, R. Giegerich: A comprehensive comparison of comparative RNA structure prediction approaches, BMC Bioinformatics **5**, 140 (2004)
10.101 C.H. zu Siederdissen, S.H. Bernhart, P.F. Stadler, I.L. Hofacker: A folding algorithm for extended RNA secondary structures, Bioinformatics **27**, i129–136 (2011)
10.102 A.O. Harmanci, G. Sharma, D.H. Mathews: TurboFold: Iterative probabilistic estimation of secondary structures for multiple RNA sequences, BMC Bioinformatics **12**, 108 (2011)
10.103 E. Torarinsson, J.H. Havgaard, J. Gorodkin: Multiple structural alignment and clustering of RNA sequences, Bioinformatics **23**, 926–932 (2007)
10.104 C.M. Reidys, F.W. Huang, J.E. Andersen, R.C. Penner, P.F. Stadler, M.E. Nebel: Topology and prediction of RNA pseudoknots, Bioinformatics **27**, 1076–1085 (2011)
10.105 A. Taneda: An efficient genetic algorithm for structural RNA pairwise alignment and its application to non-coding RNA discovery in yeast, BMC Bioinformatics **9**, 521 (2008)
10.106 S. Janssen, R. Giegerich: Faster computation of exact RNA shape probabilities, Bioinformatics **26**, 632–639 (2010)