Identification and annotation of non-protein coding RNAs.

Kristin Reiche, Katharina Schutt, Kerstin Ullmann, Friedemann Horn,

Jörg Hackermüller

December 14, 2009

1 Introduction

Of the 3.3 billion bases of the human genome, only about 2% code for proteins. Since very recently, the remaining 98% have been considered to be 'junk' and functionless. However, large transcriptomic studies like ENCODE (ENCyclopedia Of DNA Elements) (1) or FANTOM (The Functional Annotation Of the Mammalian Genome) (2) have shown that around 90% of the genome is actively transcribed into RNA. The pilot phase of the ENCODE Project was focused on 1% of the human genome sequence and was organized as an international consortium of computational and laboratory-based scientists working to develop and apply high-throughput approaches for detecting all sequence elements that confer biological function. The studies advanced the collective knowledge about human genome function in several major areas. The ENCODE project provided convincing evidence that the genome is pervasively transcribed and the systematic examination of transcriptional regulation has yielded new understanding about transcription start sites, including their relationship to specific regulatory sequences and features of chromatin accessibility and histone modification. Beside this, the studies offered a more sophisticated view on chromatin structure and inter- and intra-species sequence comparison with respect to mammalian evolution has yielded new mechanistic and evolutionary insights concerning the functional landscape of the hu-

man genome (3). These findings stay in contrast to the so called 'central dogma of molecular biology', stating that DNA is transcribed into mRNA, which is then translated into protein, based on the assumption that only DNA that encodes a protein is transcribed into RNA. The potential of RNA as regulatory molecules has been revealed decades ago. Many classes of RNA regulators in species ranging from viruses to mammalian have been discovered being involved in the control of RNA stability, gene expression, tissue and cellular development, RNA modification, chromatin organization, alternative splicing, sub-cellular localization of proteins, heat shock sensing and other processes (4). RNAs, in particular non-coding RNAs and other functional RNAs, are characterized by their base sequence and higher order structural motifs. Short- and long-range base pair interactions that organize the molecules into domains are the main characteristics of those RNAs and provide a framework for functional interactions. RNA molecules fold into characteristic secondary and tertiary structures that are the base for their diverse functional activities. The RNA secondary structure consists of nucleotides that are either paired or unpaired, where pairing includes all base-base interactions. Most base pairings are adjacent and anti parallel with other base pairings to form secondary structures (5; 6). On the one hand secondary structures of RNAs are responsible for their diverse functional activities and help finding a certain functional classification for RNAs. On the other hand RNA structure, in particular secondary structures and their search and prediction algorithms, may help to classify RNAs without knowing functional details. Ongoing studies provide more and more evidence that the 'junk' non-protein coding RNAs (ncRNA) have important regulatory function and could explain the complexity of higher organisms, as the amount of ncRNAs increases dramatically during evolution, whereas the number of protein-coding genes remains relatively constant. Additionally, compared to mRNAs, ncRNAs are exceedingly cell-type specific expressed and regulated (1). A limited number of trans-acting small ncRNAs have been described in prokaryotes that appear mainly to regulate mRNA translation or stability. However, ncRNAs do not dominate genomic output in prokaryotes, representing only a small fraction of their genomes, which are generally dominated (80-95%) by protein-coding sequences (7; 8). In humans, and other eukaryotes, the number of identified functional ncRNAs has increased tremendously during the last few

years, but most ncRNAs and especially their function still remain to be elucidated. Taking their functional repertoire into account, ncRNAs can be divided into two classes: (1) housekeeping and (2) regulatory ncRNAs. Housekeeping RNAs are transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) involved in mRNA translation, small nuclear (snRNAs) involved in splicing, small nucleolar (snoRNAs) involved in the modification of rRNAs, RNase P RNAs, so called Ribozymes, telomerase RNA and others. Whereas housekeeping ncRNAs are generally constitutively expressed and are required for normal cell functions, the class of regulatory ncRNAs or riboregulators are expressed at certain stages of development, during cell proliferation or as a response to external stimuli (9). During the past few years microRNAs, ∼21nt long RNAs, which regulate mRNAs at the post-transcriptional level, have become extremely popular, as they seem to be involved in crucial biological processes, like development, differentiation and apoptosis (10). Initially, discovered in C.elegans (11; 12), they are today known to be expressed in plants and throughout the animal kingdom, including humans, and many of them are evolutionary conserved (13; 14). MiRNAs bind to partially complementary target sites in the 3'UTR of their target mRNAs which leads either to mRNA degradation or translational repression, depending on the grade of complementary. Another class of small ncRNAs, are the Piwi-RNAs (piRNAs). PiRNAs form RNA-protein complexes through interactions with Piwi proteins. These piRNA complexes have been linked to transcriptional gene silencing of retro-transposons and other genetic elements in germ line cells, particularly those in spermatogenesis (15). In contrast to the various classes of small ncRNAs, the long ncRNAs, with a length >200nt, lack satisfactory classifications. The broad functional repertoire of long ncRNAs includes epigenetic mechanisms, as well as transcriptional and post-transcriptional regulation. The large non-coding RNA Xist is the master regulator of X chromosome inactivation. Xist is negatively regulated by its antisense transcript Tsix. This repressive antisense transcription across Xist operates at least in part through the modification of the chromatin environment of the locus (16). Long-ncRNAs can act as co-factors, by modulating transcription factor activity, they can also effect global changes by interacting with basal components of the RNA polymerase II (RNAPII) dependent transcription machinery, and they can regulate RNAPII activity, e.g. by influencing

promoter choice. For example, in humans an ncRNA transcribed from an upstream region of the dihydrofolate reductase (DHFR) locus forms a triplex in the major promoter of DHFR to prevent the binding of the transcriptional co-factor TFIID (17). Most mammalian genes express antisense transcripts, which might constitute a class of ncRNA that is particularly adept at regulating mRNA dynamics. They have been shown to direct the alternative splicing of mRNA isoforms, for example Zeb2 (18), or alternatively the annealing of ncRNA can target protein effector complexes to the sense mRNA transcript in a manner analogous to the targeting of the RNA-induced silencing complex (RISC) to mRNAs by siRNAs (19). Computational and experimental identification as well as functional classification of ncRNAs will be described in the following sections.

2 ncRNA datasets

In this section we will briefly review available datasets of ncRNAs. In the field of non-coding RNAs computational prediction of functional ncRNAs and experimental identification of expressed RNAs were neck-to-neck to each other. Bioinformatic prediction delivered putative RNAs that showed signs of stabilizing selection, i.e. of functionality but lacked knowledge of expression. Experimental approaches delivered numerous expressed ncRNAs without being able to delineate their functional relevance. We will therefore present datasets of predicted RNAs as well as datasets of large scale experimental efforts for their identification. Due to the dynamics of this field which has recently developed, this review cannot be comprehensive and we mainly focus on the complex datasets derived for mammalian species, despite the many interesting reports on functional ncRNAs in prokaryotes and more basal eukaryotic model organisms.

2.1 Datasets of computationally predicted ncRNAs

Prediction of structured ncRNAs. Many of the long known, 'house-keeping' ncRNAs, like tRNAs, snRNAs or snoRNAs, exhibit pronounced secondary structure features which are under stabilizing selection, e.g. because a particular structure is required for interaction

with protein complexes. It was therefore suggested that functional RNA elements should have a secondary structure that is energetically more stable than expected by chance (20). Thermodynamic stability alone did, however, not prove to be statistically significant enough for ncRNA detection (21). Comparative approaches, in contrast, aimed to detect conserved RNA secondary structure in sequence alignments. `qrna`, which was the first successful algorithm of this type, compares which of three models describes a given pairwise sequence alignment best. A pair stochastic context free grammar (SCFG) is used to model RNA secondary structure evolution. A pair Hidden Markov Model describes protein coding sequence evolution and another pair HMM represent the null model of an unconstrained sequence (22). An up-to-date extension of `qrna` to multiple sequence alignments (MSAs) is `evofold`, which combines SCFGs to model RNA secondary structure with a phylogenetic tree to model substitution rates along the branches of the tree (23). Human genome wide predictions of structured RNAs using `evofold` are provided as a track in the UCSC genome browser.

An alternative strategy to the prediction of RNAs with secondary structure under stabilizing selection is `RNAz` which is based on thermodynamic RNA folding. It determines a structure conservation index (SCI) obtained by comparing folding energies of the individual sequences with the predicted consensus folding of a multiple sequence alignment as one classification criterion and a z-score measuring thermodynamic stability of the individual sequences as a second. Both measures are combined by a support vector machine that detects conserved and stable RNA secondary structures with high sensitivity and specificity (24). RNAz has been applied to detect ncRNAs in the human genome (25), data are available as bed files at `http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/ncRNA/`, for Urochordates (26), Nematodes (27), Drosophilids (28), yeast (29), Plasmodia (30) and teleost fishes (31). Estimation of false discovery rates (FDR) for alignment based methods like `RNAz` is intricate as it requires the randomization of alignments which is easily biased. A solution to this issue is `SISSIz`, which generates random alignments satisfying various constraints and has been used for more accurate FDR estimates of `RNAz` (32). `RNAz 2.0` has been recently released (33) and a web server is available for small scale predictions (34). Setting up your own `RNAz` screen is described in (35; 36) and may require computation of dedicated multiple

sequence alignments for which `NcDNAlign` provides a solution adjusted to the needs of `RNAz` (37).

Sequence alignment independent approaches. The above described approaches have the disadvantage that no reliable multiple sequence alignments may be available, because only distantly related genomes are available, or due to the rapid evolution of non-coding sequences. This can be overcome by relying on structural alignments, however, at the cost of a huge computational effort. Variants of the Sankoff algorithm, which solves RNA folding and sequence alignment simultaneously (38), have been used for ncRNA prediction e.g.: a classifier based on `Dynalign` (39), or a screen using `foldalign` to compare sequences between human and mouse which are not alignable on primary sequence level (40). While these and similar approaches can be used for ncRNA screens, a genome wide application is usually not feasible or only when outstanding computational resources are available. Apart from ncRNA identification, these approaches are useful for classification of ncRNAs, see the following chapter.

Secondary structure independent prediction of ncRNAs. While the different approaches for detecting structured ncRNAs delivered a plethora of potentially functional ncRNAs, not all to date known ncRNAs have easily detectable structural features or contain only few structured domains in a longer unstructured sequence, which is in particular true for most so far identified mRNA-like ncRNAs. Also, the overlaps between ncRNAs detected experimentally in ENCODE and those predicted by `evofold` and `RNAz` were surprisingly small (41). A recently published approach that successfully identified novel unstructured ncRNAs in *Drosophila melanogaster* relied on the detection of conserved intron positions in otherwise largely non-conserved sequence context (42).

2.2 Datasets of experimentally identified ncRNAs

We briefly introduce large scale studies that led to the identification of non-coding transcripts. Approaches aimed at finding individual ncRNAs cannot be covered, but we refer

to the databases listed in Section 4 which collect these transcripts. An experimental identification of ncRNAs has one important prerequisite: methods need to be unbiased, i.e. detection is not restricted to a particular subset of RNAs, like polyadenylated transcripts. Also, as many ncRNAs are – compared to mRNAs – expressed at low levels, approaches for identifying novel ncRNAs benefit from increased sensitivity. Over the last years, unbiased transcriptomics using tiling arrays and sequencing of cDNA and EST libraries have been most successful in finding new ncRNAs. More recently, transcriptome sequencing (RNAseq), approaches identifying RNAs bound to proteins and identification of ncRNAs based on epigenetic patterns proved to be of value.

ncRNA identification by transcriptomics approaches. Microarrays have been used since the 1990s for quantifying the expression of known transcripts. Phil Kapranov (43), Eric Schadt (44), Paul Bertone (45) and Viktor Stolc (46) and colleagues pioneered the application of tiling arrays to identify novel transcripts. Tiling arrays are oligonucleotide microarrays that unlike other arrays do not interrogate the sequences of specific transcripts but rather tile the genome of interest. I.e., probes are spaced in regular intervals throughout the genome. Typically, probes with low complexity or repetitive sequences are removed as their signal cannot be attributed to a specific genomic location. Apart from that and the variation of probe lengths for designs relying on longer probes, no probe design is possible for this type of arrays and hybridization free energies vary strongly between probes. Together with cross-hybridization effects this leads to rather noisy signals which require a stringent processing. Therefore, Kampa et. al. (47) proposed to use a stringent cutoff for considering individual probes as expressed and to require at least three consecutive probes to have signal above this cutoff to designate a region as expressed. With a probe pitch of 35nt in many whole genome tiling experiments, this results in the efficient detection of transcripts which are at least approximately 110 nt in size. In (48) arrays with a resolution of 5nt were used to extend the approach to the detection of small RNAs. However, with the need of using around 180 arrays per sample to accommodate the resulting number of probes this approach is hardly practical for whole genome applications. Different tiling array approaches have been

a mainstay of the pilot phase of the ENCODE project (49). Based on this data, ENCODE concluded that more than 90 % of the human genome is transcribed into RNA on at least one strand. Also, comparing expression data of several cell lines and tissues it was shown that on average expression patterns of ncRNAs are far more specific than those of mRNAs. Datasets of the ENCODE pilot phase and ongoing ENCODE studies are most easily accessible via the UCSC genome browser. A disadvantage of tiling arrays is the locality of information, i.e. transcript starts and ends can be inferred only indirectly based on segmentation of the expression signal into blocks of similar expression. Tiling array data are therefore ideally complemented by techniques like Cap-analysis gene expression (CAGE) tags, or paired end tags (PET).

With the rapid development of second generation sequencing techniques transcriptome sequencing (RNAseq) became a viable alternative to array based approaches, in particular for small RNAs (50). Compared to tiling arrays RNAseq methods have the advantage that sequences of any size that can be reliably mapped to the genome or assembled can be identified. On the other hand, library preparation for RNAseq may more biased than labeling and amplification for tiling arrays. Also, the capacity of today's sequencers may still be limiting for identifying differential expression of long ncRNAs in complex samples. See chapter XX for details on RNA-seq.

NcRNAs identified from cDNA clones. FANTOM, the functional annotation of the mouse transcriptome project initially identified numerous ncRNAs from sequencing full length cDNA clones (51; 52). Subsequent FANTOM studies supplemented cDNA data by massive identification of transcription start and termination sites (53) and most recently combining CAGE (Cap-analysis gene expression) with deep sequencing (deepCAGE) (54). In particular the latter technique alleviates one of the major down-sides of purely clone based approaches, which hardly capture differential expression.

Epigenetics based approaches. Recently Mitchell Guttman and colleagues introduced an approach for identifying ncRNAs that is orthogonal to transcriptomics (55): Studying the profiles of specific histone modifications in and around protein coding genes, they developed

an epigenetic signature of actively transcribed genes and used these patterns to identify a series of novel transcripts called lincRNAs.

3 Methods for ncRNA annotation

While genome databases offer a wealth of information about known and putative functions for protein-coding genes, functional information for novel non-coding RNAs (ncRNA) genes is almost non-existent. This is mainly explained by the lack of established software tools to efficiently identify the function and evolutionary origin of ncRNAs. Cellular functions of ncRNAs are either defined by structure motifs, sequence motifs or a combination of both, which is a consequence of their action in RNA-(m)RNA or RNA-protein complexes. In order to preserve the capability of interaction, evolutionary selection of ncRNA genes ensures conservation of these motifs. NcRNAs of the same evolutionary origin are defined to belong to the same *RNA family*. Members of the same ncRNA family do in general share strong sequence similarity, which decreases for ancestral homologies, while secondary structure may still be conserved. RNA families sharing the same evolutionary origin but no sequence similarity are defined to be members of the same *RNA clade*. Lastly, in case RNA families did not evolve from the same origin, but their similar cellular functions converged to similar secondary structures they are seen as members of the same *RNA class*.

To classify new ncRNA genes (either retrieved by genome-wide high-throughput experimental transcriptomics, or genome-wide computational screens) one may follow the outline described here. A selection of the most important software tools which address the problem of classifying ncRNA genes is given in table 1.

Sequence-based classification searches for ncRNAs that are conserved in their primary sequence and, thus, are ncRNAs exhibiting close homology as well as similar cellular functionality. With standard local sequence alignment tools like `blast` , good results can be achieved, already. However, sequence conservation in ncRNAs is mostly restricted to rather short regions of sequence motifs being functional that are interrupted by structure motifs.

The most promising approaches to detect homologous ncRNAs by sequence conservation are searches for fragmented patterns that do not require a conserved substring of sufficient length as seed region for the local alignment (56). Regions of poor conservation are simply treated as distance constraints between well conserved blocks.

Secondary structure-based classification, in contrast, resolves more distant homologous ncRNAs, because ancestral copies of the same ncRNA are likely to exhibit only low sequence conservation, but still recognizable conservation in secondary structure. Furthermore, classes of ncRNAs likely to be involved in the same cellular processes can be identified by structure-based classification, as they are expected to share secondary structure motifs. Such classification problems have in common that structural similarity must be evaluated in an efficient and reliable way among a large set of candidate structures.

Prerequisites for classification by secondary structure are reliable secondary structure models that reflect structure and sequence motifs that are central for the functional constraints of an ncRNA. Predicting the secondary structure from a set of evolutionary related RNA sequences outperforms predictions from single RNA sequences. RNA structures being functionally important are expected to evolve much more slowly than their underlying sequences. Exploiting patterns of sequence co-variations according to structure variation improves secondary structure prediction considerably. Most reliable predictions can be achieved by the Sankoff algorithm that optimizes sequence and structure conservation of related RNAs simultaneously (38). However, this is computationally too expensive for widespread application. Implementations of the Sankoff algorithm that are tractable for realistic input sizes are `FOLDALIGN` (57), `FoldalignM` (58), `Dynalign` (59), `PMcomp` (60), and `LocARNA` (61). Probabilistic approaches to simultaneously align and predict secondary structures are usually based on stochastic context free grammars (SCFGs) like `CMfinder` (62), `SimulFold` (63), and `Consan` (64). Efficient alternatives to those Sankoff-like methods are approaches that evaluate co-variations in sequence and structure on the basis of a given multiple sequence-alignment. Implementations of such methods are `RNAalifold` which predicts the minimum free energy secondary structure of a global sequence-alignment (65; 66), `Pfold` which combines an evolutionary model of the RNA sequences with a probabilistic model

of the secondary structures (67; 68), and `PETfold` which simultaneously uses evolutionary and energetic information (69). Predicting the common secondary structure from a set of pre-aligned RNA sequences has the problem that the pure sequence alignment must reflect sequence as well as structural conservation which is problematic for ncRNAs exhibiting sequence conservation less than 60% (70). For evolutionary distant ncRNAs hand-curation of the multiple-sequence alignments is still necessary to retrieve reliable consensus secondary structure predictions.

Specialized classification methods provide secondary structure models specifically designed for particular ncRNA families/classes. They focus on specific structure and sequence motifs occurring explicitly in the ncRNA family/class of interest. An alternative to predefined models are descriptor based approaches, that provide description languages allowing users to define combined sequence/structure models for particular ncRNA families/classes and to search for instances of the defined patterns in databases. Such methods mostly connect a pattern language with user-defined approximate rules, which rank the results according to their distance to the motif. The construction of specialized models is in general time-consuming and requires extensive expert knowledge about the RNA family/class of interest.

General purpose classification tools are, in contrast, not restricted to a specific ncRNA family/class, but detect new members for any predefined secondary structure motif. The most common tools require as input a sequence alignment in combination with structural annotation as a training set to automatically learn a statistical model. A recent evaluation of such methods is presented by Freyhult et. al. (71). The most common tools, `Infernal` (72) and `CMfinder` are based on covariance models, which are the stochastic context free grammar analogue of profile hidden markov models. They are specifically designed to model base pair interactions, and search in a database of candidate sequences for high-scoring matches to the RNA model. `Infernal` requires an alignment of the set of RNA sequences, while `CMfinder` is able to learn the covariance model from an unaligned set of RNA sequences. These approaches are extremely time-consuming, and pre-filtering steps, as provided by `RaveNnA`, are required. However, the recent version of `Infernal` might make pre-filtering unnecessary.

An alternative approach is followed by ERPIN (73). It takes a structure-annotated multiple alignment as input to construct an RNA descriptor in terms of log-odds-score profiles.
In contrast to descriptor based search methods the user has little effort with generating the statistical model, but, in consequence to that, cannot easily improve the model by incorporating expert knowledge.

Secondary structure clustering, on the other hand, identifies among a diverse set of ncRNAs those being members of the same secondary structure profile. It allows not only to assign candidates for functional ncRNAs to known RNA families, but also to reveal novel ncRNA families. These approaches use, in general, an efficient sequence-structural alignment method to compute all pairwise alignments in order to reveal pairwise structural similarities among the input set. Subsequently, all sequences are clustered according to a distance matrix derived from pairwise alignments, using, for example, an agglomerative clustering approach. Lastly, multiple alignments are computed for all clusters and the optimal number of clusters is retrieved (61; 58; 74).

Interaction-based classification focuses on computational identification of binding partners, i.e. the regulatory co-workers of ncRNAs. The majority of known ncRNAs are regulatory active by either base pairing with a target (m)RNA, or binding to protein complexes. Computational prediction of target (m)RNAs or target proteins, especially for novel ncRNA families, is essential in order to reveal their regulatory role and signal transduction pathways they may be involved in. Despite the importance of these facts, with exception of miRNAs, only few efficient software tools are available to predict RNA interaction targets with high sensitivity and specificity.

RNA target prediction either identifies RNA targets that simply show sense-antisense base pair interactions, which is mainly observed for small RNAs and their target mRNAs, or for targets exhibiting more complex interaction structures like co-folding of two long RNA molecules. Hybridization energies, structural accessibility of the target sites, as well as perfect Watson-Crick base pairing of short consecutive seed regions are important for sense-antisense base pair predictions. Software tools designed for this task can be separated into approaches that completely avoid internal base-pairing in both RNA strands (75; 76), and

tools that follow a more sophisticated approach which restricts base-pairing to sub-regions that are likely to remain unpaired in the internal secondary structure of both RNA molecules (77; 78). These approaches model general types of RNA-RNA interaction structures and do not specifically consider hybridization characteristics that are typical for the RNA classes under consideration. Specialized target prediction tools have mainly been developed for miRNAs, while little effort has been put into methods for other regulatory RNA classes. An exception is `RNAsnoop` a specialized target prediction tool for H/ACA box snoRNAs (**?**).

First attempts to predict complex RNA-RNA interactions concatenate both RNA molecules to one single strand and predict its minimum free energy structure. The drawbacks of this approach is that it is not able to predict important RNA-RNA interaction motifs, like kissing hairpin loops. Recently, two efficient methods have been published that compute the whole ensemble of interaction structures that are known from in-vivo RNA-RNA complexes (79; 80). Their implementations follow an energy minimization algorithm for RNA-RNA joint secondary structures proposed by Alkan et. al. (81).

MicroRNA target prediction: Though several hundred miRNAs have been validated in humans and other animals, only a small fraction of verified miRNA:target pairs exist, mainly due to lacking high-throughput experimental identification methods. However, several methods to predict miRNA targets computationally have been developed. The major challenge for such an in-silico target identification approach is the lack of complete complementarity between the miRNA and its target sequence, containing imperfect characteristics like mismatches, gaps and G:U wobbles. The seed match, almost complete complementary between nucleotides 2 and 7 of the miRNA, evolutionary conservation of the miRNA and target binding site, as well as binding of the miRNA to the 3'UTR of its target gene are features most of the prediction programs have in common. Most of these methods, like `miRanda` (now termed `Microcosm`) (82; 83), `PicTar` (84), `PITA` (85), TargetScan/TargetScanS (86; 87), `DIANA microT` (88), `MicroInspector` (89) and `miRtarget2` (90) primarily use sequence complementarity, thermodynamic stability calculations and evolutionary conservation among species to determine the likelihood of a productive miRNA:mRNA duplex for-

mation. Variations in these algorithms include for example the number of miRNA binding sites in one target, the accessibility of the target (`miRTarget2`, `PITA`), binding sites outside the 3'UTR (CDS+5'UTR) (`miRanda (Microcosm)`, `microInspector`), as well as additionally conserved nucleotides beside the seed region (`TargetScanS`). In addition, several target prediction methods are available, that do not rely on sequence conservation (`rna22` (91), `miTarget`, (92) `NbmiRTar` (93)), or the seed match only (`NBmiRTar`). `Rna22`, in contrast to all other programs, first identifies the putative target site in a given mRNA, before it then identifies the targeting sequence. Future knowledge gained from experimentally validated miRNA:target pairs is required for optimizing existing and implementing new algorithms and to lower the false positive rates of the programs ($\sim$30%).

Identification of protein binding partners for an ncRNA is one important step towards understanding the cellular role of an RNA molecule. Computationally, this task is solved by either predicting protein-binding motifs of a given ncRNA, or predicting RNA-binding sites of a given protein. In both cases the corresponding binding partner is not known in general, and thus, the searching for a binding motif in either an RNA molecule or protein is only based on the sequence and structure information of the RNA or protein, respectively. Two main types of protein-RNA interactions are known (94; 95): (1) Interactions with the backbone of double-stranded RNA molecules and (2) interactions of single-stranded RNA bases that are accommodated in the protein binding pockets. While little knowledge is available to identify interactions concerning the backbone, different approaches exist to predict binding interactions with single stranded RNA bases (96). Many RNA-binding proteins contain domains, often also in multiple copies, that bind specifically to *single-stranded RNA*. Examples for RNA binding domains of proteins are the K homology domain (KH), the RNA recognition motif (RRM), the pumilio homology domain (PUF-HD) and the Zinc-finger binding domains (96; 97; 98; 99; 100; 101). RNA molecules interacting with protein domains contain sequence-specific motifs with a general structural property, such as being located within single-stranded regions of an arbitrary secondary structure, as for example the trans-activation response element (102) and the U1A polyadenylation inhibition element (103).

Software tools designed to predict RNA-binding domains in proteins can be separated in approaches incorporating knowledge about the tertiary structure of the protein (94), and in approaches that identify those nucleotides that may be located at the RNA-interaction interface from the amino-acid sequence of the protein of interest, while both the structure of the protein and the sequence of the target RNA is unknown. Examples for the latter approach are `RNAProB` (104), `BindN` (105), `RNABindR` (106), `PPRint` (107) and `PRINTR` (108). While most of the research focuses on predicting RNA-binding domains in proteins, only little research is done to predict protein-binding motifs in RNA sequences, with the exception of the works by Westhof and colleagues (109; 110; 111). Mostly, standard sequence motif search tools like `MEME` or `Gibbs sampler` (112; 113; 114) are utilized. The only tool that also includes information about the secondary structure of the RNA is `MEMERIS` (115). It guides motif search in RNA sequences towards single-stranded regions by considering information about paired and unpaired regions as a requirement for putative start positions for reasonable binding-motifs.

De-novo prediction

	Prob. Model	MSA	Energy	Remark	Citation
QRNA	y	pairwise	n	A pair-SCFG is used to model RNA secondary structure evolution. A pairwise alignment is classified to one of three evolution classes: structural RNA evolution, coding sequence evolution, or position-independent evolution.	(21)
EvoFold	y	y	n	Extension of QRNA to multiple sequence alignments (MSA).	(23)
RNAz	n	y	y	Alternative de-novo prediction based one thermodynamic RNA folding. It compares folding energies of the individual sequences with the predicted consensus folding of a MSA.	(24; 33)

Sequence based classification

	Local	MSA	Algorithm	Remark	Citation
blastn	y	n	heuristic Smith-Waterman	Fast local alignment tool to detect homology at the nucleotide level between a sequence pair. Detecting conserved subsequences of ncRNAs using blastn often forms a good starting point, while post-filtering steps reject all sequences not exhibiting specific structural features of the query ncRNA.	(116)
SSEARCH	y	n	full Smith-Waterman	It is much slower than blast but more sensitive, because no heuristics to speed up searches in large databases are implemented.	(117)
Probalign	y	n	partition function	In contrast to blast and SSEARCH it computes all sub-optimal alignments between a pair of sequences.	(118; 119)
GotohScan	(y)	n	semi-global alignment	Interprets the problem of homology searches as identifying the best match of the query ncRNA within a complete genome.	(120)
fragrep	n	-	pattern matching	Fragmented pattern search that treats sequence regions which are poorly conserved as distance constraints between significant conserved blocks.	(121)
ClustalW	n	y	progressive MSA	MSA is built from a phylogenetic tree either created by pairwise alignments or provided by the user. Produces good alignments of ncRNAs that show high to medium sequence homology (70), and forms the basis of the famous ncRNA prediction tool RNAz (24; 122).	(123)

Secondary structure prediction

	Energy	Sankoff	Prob. Model	MSA	Remark	Citation
UNAFold	y	n	n	n	One of the two standard software packages for secondary structure prediction for one or two single-stranded RNA sequences. Includes algorithms for free energy minimization, partition function calculations and stochastic sampling. It replaces and extends the original mfold package.	(124)
Vienna RNA Package	y	n	n	y/n	The second standard software package for secondary structure prediction for RNA sequences. Includes algorithms for prediction of minimum free energy structures, of sub-optimal secondary structures as well as of locally stable structure of long sequences, and calculation of partition functions. RNAfold predicts minimum free energy structure from a single sequence, and RNAalifold from a MSA.	(125; 66)
FOLDALIGN(M)	n	y	n	n	Implementation of a simplified version of the Sankoff algorithm (38) for local or global simultaneous secondary structure prediction and aligning a collection of RNA sequences.	(57; 58)
Dynalign	y	y	n	n	Combines free energy minimization and comparative sequence analysis to find a low free energy structure common to two sequences.	(59)
PMcomp	y	y	n	n	Uses McCaskill's approach (126) to compute base pairing probability matrices which incorporate information on the energetics of each sequence, and aligns these matrices to retrieve a consensus secondary structure model and alignment for two sequences.	(60)
(m)LocARNA	y	y	n	n	Same approach as used in PMcomp, however with decreased memory and run-time requirements. (m)LocARNA computes a multiple alignment progressively from pairwise LocARNA alignments.	(61)

	Pfold	n	n	y	y	Combines an evolutionary model of the RNA sequences with a probabilistic model for the secondary structures.	(67; 68)
	PETfold	y	n	y	y	Extends Pfold such that energetic information is incorporated into the structure prediction, too.	(69)
	Consan	n	y	y	n	Use pair stochastic context-free grammars (pairSCFGs) as a unifying framework for scoring pairwise alignment and folding.	(64)
	SimulFold	n	y	y	n	Predicts secondary structures (including pseudoknots), multiple sequence alignments, and an evolutionary tree from unaligned RNA input sequences simultaneously.	(63)
Secondary structure based classification	**Known class**	**Motif**	**Prob. Model**	**Model**	**Remark**		**Citation**
	tRNAscan-SE	n	y	CM	Standard software tool for scanning genomes for tRNA genes. It detects tRNA pseudogenes as well as unusual tRNA homologues such as selenocysteine tRNAs.		(127)
	RNAmicro	n	n	SVM	An SVM that recognizes conserved microRNA precursors in multiple sequence alignments.		(128)
	SnoReport	n	n	SVM	An SVM that recognizes two major classes of snoRNAs, box C/D and box H/ACA snoRNAs, in multiple sequence alignments. Classification of the multiple alignment is solely based on information about conserved sequence boxes and secondary structure constraints. It does not require any target information.		(129)
	Snoscan	n	y	CM	Implementation of a deterministic search algorithm and a probabilistic gene model to scan genomic sequences for C/D box snoRNA genes. It requires the sequence of the query, i.e. the snoRNA candidate, as well as the sequence of the target, i.e. RNAs that may base-pair and be modified by the query snoRNA sequence, as input.		(130)
	SnoGPS	n	y	CM	Analogical implementation of Snoscan for H/ACA snoRNA genes.		(131)
	RNAmmer	n	y	HMM	Fast computational predictor for the major rRNA species.		(132)
	General purpose						
	Infernal	n	y	CM	Searches sequence databases for RNA structure and sequence similarities. Requires a structure-annotated alignment of related RNA sequences as input.		(72)
	CMfinder	y	y	CM	Expectation maximization algorithm that captures secondary structure motifs within a set of unaligned RNA sequences.		(62)
	RaveNnA	n	y	HMM	Implementation of rigorous filters that accelerate CM searches for almost all known ncRNA families from the RFAM database and tRNA models in tRNAscan-SE.		(133)
	RSEARCH	n	y	CM	Alignments are scored by single nucleotide and base pair substitution matrices (RIBOSUM matrices) specifically designed for pairwise alignments of RNA sequences.		(134)
	FastR	n	n	pattern	Improves run-time by applying structural filters in order to eliminate unrelated sequences from the database, while true homologous RNAs are retained.		(135)
	ERPIN	n	n	lod-score profile	It requires an RNA sequence alignment as input and identifies related sequences using a profile-based dynamic programming algorithm. It is able to handle pseudoknots.		(136; 73)
	fragrep3	n	n	pattern	A pattern matching approach for RNA secondary structures combining features of statistical approaches and descriptor-based methods. While fragrep3 is conceptually similar to ERPIN, it treats poorly conserved regions as simple distance constraints.		(56)
	ExpaRNA	y	n	pattern	Another exact pattern matching approach to detect the longest collinear sequence of substructures common to two RNA sequences. Substructures common to both RNA sequences are treated as whole unit while variable regions are allowed between them.		(137)

	RNAbob	n	n	pattern	-	(138)
	HyPaL	n	n	pattern	Contains a library of annotated structural elements characteristic for certain classes of structural and/or functional RNAs and allows searching databases for those motifs. Allows user-defined approximate rules which rank matches according to their distance to the motif.	(139)
	RNAMotif	n	n	pattern	Relies on a flexible structure definition language which can specify any type of base-pairs and provides a user controlled scoring section that allows ranking of matches.	(5)

Interaction based classification	**miRNA-RNA**	**Access.**	**Seed**	**Energy**	**Cons.**	**Remark**	**Citation**
	RNAhybrid	n	y	y	n	Introduces specialized energy model for dimer hybridization. Especially designed to predict multiple potential binding sites of miRNAs in large target RNAs.	(140; 75)
	miRanda (Microcosm)	n	y	y	y	Whole-genome predictions of miRNA target genes. It also incorporates the degree of conservation of putative target sites in the prediction process.	(82; 83)
	PicTar	n	y	y	y	Identification of mRNAs likely to be the target for a combination of miRNAs using a scoring system based on a HMM.	(84)
	PITA	y	y	y	n	Quantified the effect of target site accessibility on microRNA-mRNA interactions by systematic experimentation. The model underlying PITA computes the difference between the free energy gained from the formation of the microRNA-target duplex and the energetic cost of unpairing the target to make it accessible to the microRNA.	(85)
	TargetScan(S)	n	y	y	y	Predicts targets of miRNAs by searching for conserved 8mer and 7mer sites that are complementary to the seed region of the miRNA.	(86; 87)
	DIANA microT	n	y	y	y	Employed a combined bioinformatics and experimental approach to identify rules important for miRNA-target identification that allow prediction of human miRNA targets.	(88)
	MicroInspector	n	y	y	n	A web-tool that uses a different view to the miRNA target prediction problem: Does a given mRNA sequence contain a binding site for any miRNA that originates from this organism and that is available in the database. Hence, the miRNA must not be known beforehand.	(89)
	rna22	n	y	n	n	Is a a pattern-based approach for the discovery of microRNA binding sites. It identifies putative microRNA binding sites without a need to know the identity of the specific targeting microRNA by defining sequence patterns common to known mature microRNAs.	(91)
	miRtarget2	y	y	y	(y)	Target prediction is done by an SVM which evaluates several features like seed conservation, GC content, accessibility, free energy and other. Conservation of target site is regarded, but not an requirement. SVM has been learnt by systematically studying public microarray data.	(90)
	miTarget	n	y	y	n	Another SVM based miRNA target prediction tool. Features include complementary to seed part and 3' region, free energy of total alignment as well as seed and 3' region, and position based features.	(92)
	NbmiRTar	n	y	y	n	A Naïve Bayes classifier that is based on sequence information and free energy.	(93)
	RNA-(m)RNA	**Accessibility**		**Scope (RNA/target)**		**Remark**	**Citation**
	UNAFold	-		global/global		Provides a tool that computes MFE secondary structure for two interacting RNA molecules. Only intermolecular basepairs are considered in duplexes.	(141)
	PairFold, MultiFold	-		global/global		Considers also intra-molecular pairing for RNA duplexes. MultiFold is the extension of the PairFold algorithm to multiple molecules.	(142)

RNAcofold	-	global/global		Provides also an extension of McCaskill's partition function algorithm to compute base pairing probabilities, realistic interaction energies, and equilibrium concentrations of duplex structures allowing intra-molecular base pairing.	(76)
RNAplex	n	global/local		Based on a simplified energy model to speed up target sites in large databases, no specifically designed for a particular target class.	(143)
RNAup	y	global/local		Includes target site accessibility, i.e. that the binding site is unpaired.	(78)
IntaRNA	y	global/local		Includes target site accessibility and user-definable seeds.	(77; 144)
RNArip	y	local/local		Enables the calculation of interaction probabilities for any given interval on the target RNA.	(80)
piRNA	y	local/local		Computes the interaction partition function over the whole ensemble of structures between two interacting RNAs.	(79)
RNA-protein	**Protein binding motif prediction in RNAs**				
	Accessibility			**Remark**	**Citation**
MEME	n			Standard sequence motif discovery tool. Discovers any type of sequence motif common to a set of DNA sequences. It is not especially designed to notice protein-binding sites in RNAs.	(112)
MEMERIS	y			Searches for sequence motifs that occur preferably in single-stranded regions by guiding the motif search to unpaired regions of the RNA sequence.	(115)
RNA-protein	**RNA binding site prediction in proteins**				
	Structure known	**Cons.**	**Method**	**Remark**	**Citation**
PPRint	n	y	SVM	RNA-binding sites are predicted from a protein sequence by training a SVM using a position-specific scoring matrix (PSSM) that reflects evolutionary conservation of the binding motif.	(107)
RNAProB	n	y	SVM	Here the PSSM is smoothed, such that it incorporates for each amino acid in the protein sequence the dependency effect from its neighboring amino acids.	(104)
BindN	n	n	SVM	Features like pK(a) value, hydrophobicity index and molecular mass of an amino acid are used to train a SVM from known DNA or RNA-binding residues.	(105)
RNABindR	n	n	Bayes classifier	Naïve Bayes classifier trained on a set of known protein-RNA complexes to predict which amino acids in a protein sequence are most likely to bind RNA.	(106)
PRINTR	y	y	SVM	Incorporates secondary structure prediction of the protein into the process.	(108)
RnaPred	y	y	-	Provide a classification of known RNA nucleotide and dinucleotide protein binding sites in order to identify common types of shared 3D physico-chemical binding patterns. Searches a complete protein for regions similar to known 3D consensus patterns of RNA-binding sites.	(94)

Table 1: A selection of the most common software tools for identification and functional classification of non-coding RNAs. The table is divided into five sub-tables describing software tools addressing de-novo prediction of ncRNAs, sequence based classification, secondary structure prediction, secondary structure based classification, and interaction based classification, respectively. **De-novo prediction:** Column headings *Prob. Model*, *MSA*, and *Energy* indicate if the described algorithm uses a probabilistic model for prediction, requires an multiple sequence alignment as input, and if thermodynamic RNA folding is regarded. **Sequence based classification:** Column headings *Local*, *MSA*, and *Algorithm* indicate if software tool computes local sequence alignments, a multiple sequence alignment, and the type of used algorithm, respectively. **Secondary structure prediction:** Column headings *Energy*, *Sankoff*, *Prob. Model*, *MSA* indicate if the prediction relies on minimization of the free energy, on predicting the alignment and the secondary structure simultaneously, on a probabilistic model for RNA secondary structures, and if a MSA is a prerequisite. **Secondary structure based classification:** Column headings *Motif*, *Prob. Model*, and *Model* indicate if the software tool is able to predict secondary structure motifs common to the input RNA sequences, if a probabilistic model is used, and the type of model used to describe the RNA motif. SVM: support vector machine, HMM: hidden markov model, CM: covariance model for RNAs. Software tools are separated in tools specifically designed for an RNA family/class (*known class*) or tools that are not restricted to a specific ncRNA family/class (*general purpose*). **Interaction based classification:** Column headings *Access.*, *Seed*, *Energy*, and *Cons.* for the miRNA-mRNA interaction sub-table indicate if the tools incorporate the structural accessibility of mRNA target, the complementarity to the seed region of a miRNA, the free energy, and the conservation of the target site into the prediction process, respectively. Column headings *Accessibility*, and *Scope* for the RNA-(m)RNA sub-table indicate if the tool checks if the binding region in the target RNA is likely to be unpaired, and the scope of interaction. Column headings *Accessibility*, and *Target known* in the RNA-protein sub-table indicate if the tool guides the motif search to single stranded regions of the RNA. Column headings *Structure known*, *Cons.*, and *Classification* for the RNA-protein sub-table indicate if tools incorporate the secondary or tertiary structure of the protein or solely base the prediction on sequence information, if conservation of binding sites is regarded, and the method used for classification.

4 Notes

4.1 Databases of known regulatory ncRNAs

RFAM: http://rfam.sanger.ac.uk

miRBase: http://www.mirbase.org

NONCODE: http://www.noncode.org

RNAdb: http://research.imb.uq.edu.au/rnadb/

snoRNA base: http://www-snorna.biotoul.fr/

snoRNAdb: http://lowelab.ucsc.edu/snoRNAdb/

4.2 Genome browsers

UCSC: http://genome.ucsc.edu/ UCSC mirror for ncRNAs: http://www.ncrna.org/glocal/cgi-bin/hgGateway Ensembl: http://www.ensembl.org/index.html

4.3 Useful web services

Vienna RNA Package: http://www.tbi.univie.ac.at/~ivo/RNA/, http://rna.tbi.univie.ac.at/

LeARN: http://symbiose.toulouse.inra.fr/LeARN/ (145)

Web repository for ncRNA community: http://www.ncrna.org/ SCOR (Database of 3D structural classification of ncRNAs): http://scor.berkeley.edu/

4.4 Software links in alphabetical order

BindN: http://bioinfo.ggc.org/bindn/

blast: http://blast.ncbi.nlm.nih.gov/Blast.cgi

CMfinder: http://bio.cs.washington.edu/yzizhen/CMfinder/

Consan: http://selab.janelia.org/software/consan/

DIANA MicroT: http://www.diana.pcbi.upenn.edu/cgi-bin/micro_t.cgi

Dynalign: http://rna.urmc.rochester.edu/dynalign.html

ERPIN: http://tagc.univ-mrs.fr/erpin/

evofold: http://users.soe.ucsc.edu/$\sim$jsp/EvoFold/

FOLDALIGN: http://foldalign.ku.dk//software/index.html

FoldalignM: http://foldalign.ku.dk/software/index.html

Gibbs sampler: http://bayesweb.wadsworth.org/gibbs/gibbs.html

Infernal: http://infernal.janelia.org/

IntaRNA: http://www.bioinf.uni-freiburg.de/Software/

LocARNA: http://rna.tbi.univie.ac.at/cgi-bin/LocARNA.cgi

MEME: http://meme.sdsc.edu/meme4_3_0/intro.html

MEMERIS: http://www.bioinf.uni-freiburg.de/$\sim$hiller/MEMERIS/

MicroInspector:http://bioinfo.uni-plovdiv.bg/microinspector/

miRanda(Microcosm): http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/

miRtarget2: http://mirdb.org/miRDB/

miTarget: http://cbit.snu.ac.kr/$\sim$miTarget/

NbmiRTar: http://wotan.wistar.upenn.edu/NBmiRTar/login.php

NcDNAlign: www.bioinf.uni-leipzig.de/Software/NcDNAlign/

qRNA: http://nbc11.biologie.uni-kl.de/framed/left/menu/auto/right/qrna/

PETfold: http://genome.ku.dk/resources/petfold/

Pfold: http://www.daimi.au.dk/$\sim$compbio/pfold/

PicTar: http://pictar.mdc-berlin.de/

PITA: http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html

PMcomp: http://www.tbi.univie.ac.at/RNA/PMcomp/

PPRint: www.imtech.res.in/raghava/pprint/

PRINTR: http://210.42.106.80/printr/

RaveNnA: http://bliss.biology.yale.edu/$\sim$zasha/ravenna/

RNA22: http://cbcsrv.watson.ibm.com/rna22.html

RNAalifold: http://rna.tbi.univie.ac.at/cgi-bin/RNAalifold.cgi

RNABindR: http://bindr.gdcb.iastate.edu/RNABindR/

RNAduplex: http://www.tbi.univie.ac.at/RNA/RNAduplex.html

RNAhybrid: http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/

RNAProB:

RNAsnoop: http://www.tbi.univie.ac.at/$\sim$htafer/RNAsnoop/RNAsnoop.html

RNAup: http://www.tbi.univie.ac.at/$\sim$ulim/RNAup/

RNAz: http://rna.tbi.univie.ac.at/cgi-bin/RNAz.cgi, http://www.tbi.univie.ac.at/~wash/RNAz/

SimulFold: http://people.cs.ubc.ca/~irmtraud/simulfold/

Stemloc: http://biowiki.org/StemLoc

TargetScan/TargetScanS: http://www.targetscan.org/;http://genes.mit.edu/tscan/targetscanS2005.html

References

[1] Consortium, A. F. B. *et al.* RNAs everywhere: genome-wide annotation of structured RNAs. *J Exp Zool B Mol Dev Evol* **308**, 1–25 (2007).

[2] Maeda, N. *et al.* Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet* **2**, e62 (2006).

[3] Consortium, E. N. C. O. D. E. P. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* **306**, 636–640 (2004).

[4] Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **8**, 413–423 (2007).

[5] Macke, T. J. *et al.* RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* **29**, 4724–4735 (2001).

[6] Tinoco, I. & Bustamante, C. How RNA folds. *J Mol Biol* **293**, 271–281 (1999).

[7] Mattick, J. S. The hidden genetic program of complex organisms. *Sci Am* **291**, 60–67 (2004).

[8] Mattick, J. S. & Makunin, I. V. Non-coding RNA. *Hum Mol Genet* **15 Spec No 1**, R17–R29 (2006).

[9] Prasanth, K. V. & Spector, D. L. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev* **21**, 11–42 (2007).

[10] Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).

[11] Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855–862 (1993).

[12] Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854 (1993).

[13] Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86–89 (2000).

[14] Kato, M. & Slack, F. J. microRNAs: small molecules with big roles - C. elegans to human cancer. *Biol Cell* **100**, 71–81 (2008).

[15] Seto, A. G., Kingston, R. E. & Lau, N. C. The coming of age for Piwi proteins. *Mol Cell* **26**, 603–609 (2007).

[16] Senner, C. E. & Brockdorff, N. Xist gene regulation at the onset of X inactivation. *Curr Opin Genet Dev* **19**, 122–126 (2009).

[17] Martianov, I., Ramadass, A., Barros, A. S., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666–670 (2007).

[18] Beltran, M. *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* **22**, 756–769 (2008).

[19] Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**, 155–159 (2009).

[20] Le, S. V., Chen, J. H., Currey, K. M. & Maizel, J. V. A program for predicting significant RNA secondary structures. *Comput Appl Biosci* **4**, 153–159 (1988).

[21] Rivas, E. & Eddy, S. R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16**, 583–605 (2000).

[22] Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**, 8 (2001).

[23] Pedersen, J. S. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**, e33 (2006).

[24] Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* **102**, 2454–2459 (2005).

[25] Washietl, S., Hofacker, I. L., Lukasser, M., Hüttenhofer, A. & Stadler, P. F. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* **23**, 1383–1390 (2005).

[26] Missal, K., Rose, D. & Stadler, P. F. Non-coding RNAs in ciona intestinalis. *Bioinformatics* **21 Suppl 2**, ii77–ii78 (2005).

[27] Missal, K. *et al.* Prediction of structured non-coding RNAs in the genomes of the nematodes caenorhabditis elegans and caenorhabditis briggsae. *J Exp Zool B Mol Dev Evol* **306**, 379–392 (2006).

[28] Rose, D. *et al.* Computational RNomics of drosophilids. *BMC Genomics* **8**, 406 (2007).

[29] Steigele, S., Huber, W., Stocsits, C., Stadler, P. F. & Nieselt, K. Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions. *BMC Biol* **5**, 25 (2007).

[30] Mourier, T. *et al.* Genome-wide discovery and verification of novel structured RNAs in plasmodium falciparum. *Genome Res* **18**, 281–292 (2008).

[31] Rose, D. *et al.* Duplicated RNA genes in teleost fish genomes. *J Bioinform Comput Biol* **6**, 1157–1175 (2008).

[32] Gesell, T. & Washietl, S. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* **9**, 248 (2008).

[33] Gruber, A., Findeiss, S., Washietl, S., Hofacker, I. & Stadler, P. RNAz 2.0: Improved noncoding RNA detection. In *PSB in press* (2010).

[34] Gruber, A. R., Neubck, R., Hofacker, I. L. & Washietl, S. The rnaz web server: prediction of thermodynamically stable and evolutionarily conserved rna structures. *Nucleic Acids Res* **35**, W335–W338 (2007).

[35] Washietl, S. Prediction of structural noncoding RNAs with RNAz. *Methods Mol Biol* **395**, 503–526 (2007).

[36] Washietl, S. & Hofacker, I. L. Identifying structural noncoding rnas using rnaz. *Curr Protoc Bioinformatics* **Chapter 12**, Unit 12.7 (2007).

[37] Rose, D., Hertel, J., Reiche, K., Stadler, P. F. & Hackermüller, J. NcDNAlign: plausible multiple alignments of non-protein-coding genomic sequences. *Genomics* **92**, 65–74 (2008).

[38] Sankoff, D. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. Appl. Math.* **45**, 810–825 (1985).

[39] Uzilov, A. V., Keegan, J. M. & Mathews, D. H. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics* **7**, 173 (2006).

[40] Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M. & Gorodkin, J. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common rna structure. *Genome Res* **16**, 885–889 (2006).

[41] Washietl, S. *et al.* Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* **17**, 852–864 (2007).

[42] Hiller, M. *et al.* Conserved introns reveal novel transcripts in Drosophila melanogaster. *Genome Res* **19**, 1289–1300 (2009).

[43] Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).

[44] Schadt, E. E. *et al.* A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol* **5**, R73 (2004).

[45] Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).

[46] Stolc, V. *et al.* Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. *Proc Natl Acad Sci U S A* **102**, 4453–4458 (2005).

[47] Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* **14**, 331–342 (2004).

[48] Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).

[49] Consortium, E. N. C. O. D. E. P. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

[50] Berezikov, E. *et al.* Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* **38**, 1375–1377 (2006).

[51] Kawai, J. *et al.* Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).

[52] Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cdnas. *Nature* **420**, 563–573 (2002).

[53] Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).

[54] Consortium, F. A. N. T. O. M. *et al.* The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**, 553–562 (2009).

[55] Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature* **458**, 223–227 (2009).

[56] Mosig, A., Zhu, L. & Stadler, P. F. Customized strategies for discovering distant ncRNA homologs. *Brief Funct Genomic Proteomic* (2009).

[57] Gorodkin, J., Heyer, L. J. & Stormo, G. D. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res* **25**, 3724–3732 (1997).

[58] Torarinsson, E., Havgaard, J. H. & Gorodkin, J. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics* **23**, 926–932 (2007).

[59] Mathews, D. H. & Turner, D. H. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol* **317**, 191–203 (2002).

[60] Hofacker, I. L., Bernhart, S. H. F. & Stadler, P. F. Alignment of RNA base pairing probability matrices. *Bioinformatics* **20**, 2222–2227 (2004).

[61] Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F. & Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* **3**, e65 (2007).

[62] Yao, Z., Weinberg, Z. & Ruzzo, W. L. CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445–452 (2006).

[63] Meyer, I. M. & Miklós, I. SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput Biol* **3**, e149 (2007).

[64] Dowell, R. D. & Eddy, S. R. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics* **7**, 400 (2006).

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

[66] Hofacker, I. L. RNA consensus structure prediction with RNAalifold. *Methods Mol Biol* **395**, 527–544 (2007).

[67] Knudsen, B. & Hein, J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **15**, 446–454 (1999).

[68] Knudsen, B. & Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* **31**, 3423–3428 (2003).

[69] Seemann, S. E., Gorodkin, J. & Backofen, R. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res* **36**, 6355–6362 (2008).

[70] Gardner, P. P., Wilm, A. & Washietl, S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* **33**, 2433–2439 (2005).

[71] Freyhult, E. K., Bollback, J. P. & Gardner, P. P. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res* **17**, 117–125 (2007).

[72] Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).

[73] Lambert, A. *et al.* The ERPIN server: an interface to profile-based RNA motif identification. *Nucleic Acids Res* **32**, W160–W165 (2004).

[74] Kaczkowski, B. *et al.* Structural profiles of human miRNA families from pairwise clustering. *Bioinformatics* **25**, 291–294 (2009).

[75] Rehmsmeier, M., Steffen, P., Hochsmann, M. & Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507–1517 (2004).

[76] Bernhart, S. H. *et al.* Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol* **1**, 3 (2006).

[77] Busch, A., Richter, A. S. & Backofen, R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* **24**, 2849–2856 (2008).

[78] Mückstein, U. *et al.* Thermodynamics of RNA-RNA binding. *Bioinformatics* **22**, 1177–1182 (2006).

[79] Chitsaz, H., Salari, R., Sahinalp, S. C. & Backofen, R. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics* **25**, i365–i373 (2009).

[80] Huang, F. W. D., Qin, J., Reidys, C. M. & Stadler, P. F. Partition function and base pairing probabilities for RNA-RNA interaction prediction. *Bioinformatics* **25**, 2646–2654 (2009).

[81] Alkan, C., Karakoç, E., Nadeau, J. H., Sahinalp, S. C. & Zhang, K. RNA-RNA interaction prediction and antisense RNA target search. *J Comput Biol* **13**, 267–282 (2006).

[82] Enright, A. J. *et al.* MicroRNA targets in Drosophila. *Genome Biol* **5**, R1 (2003).

[83] John, B. *et al.* Human MicroRNA targets. *PLoS Biol* **2**, e363 (2004).

[84] Krek, A. *et al.* Combinatorial microRNA target predictions. *Nat Genet* **37**, 495–500 (2005).

[85] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nat Genet* **39**, 1278–1284 (2007).

[86] Lewis, B. P., hung Shih, I., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).

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

[88] Kiriakidou, M. *et al.* A combined computational-experimental approach predicts human microRNA targets. *Genes Dev* **18**, 1165–1178 (2004).

[89] Rusinov, V., Baev, V., Minkov, I. N. & Tabler, M. MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res* **33**, W696–W700 (2005).

[90] Wang, X. & Naqa, I. M. E. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* **24**, 325–332 (2008).

[91] Miranda, K. C. *et al.* A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**, 1203–1217 (2006).

[92] Kim, S.-K., Nam, J.-W., Rhee, J.-K., Lee, W.-J. & Zhang, B.-T. miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics* **7**, 411 (2006).

[93] Yousef, M., Jung, S., Kossenkov, A. V., Showe, L. C. & Showe, M. K. Naïve Bayes for microRNA target predictions–machine learning for microRNA targets. *Bioinformatics* **23**, 2987–2992 (2007).

[94] Shulman-Peleg, A., Shatsky, M., Nussinov, R. & Wolfson, H. J. Prediction of interacting single-stranded RNA bases by protein-binding patterns. *J Mol Biol* **379**, 299–316 (2008).

[95] Draper, D. E. Themes in RNA-protein recognition. *J Mol Biol* **293**, 255–270 (1999).

[96] Auweter, S. D., Oberstrass, F. C. & Allain, F. H.-T. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res* **34**, 4943–4959 (2006).

[97] Messias, A. C. & Sattler, M. Structural basis of single-stranded RNA recognition. *Acc Chem Res* **37**, 279–287 (2004).

[98] Hall, K. B. & Stump, W. T. Interaction of N-terminal domain of U1A protein with an RNA stem/loop. *Nucleic Acids Res* **20**, 4283–4290 (1992).

[99] Spassov, D. S. & Jurecic, R. The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? *IUBMB Life* **55**, 359–366 (2003).

[100] de Moor, C. H., Meijer, H. & Lissenden, S. Mechanisms of translational control by the 3' UTR in development and differentiation. *Semin Cell Dev Biol* **16**, 49–58 (2005).

[101] Hudson, B. P., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol* **11**, 257–264 (2004).

[102] Kulinski, T. *et al.* The apical loop of the HIV-1 TAR RNA hairpin is stabilized by a cross-loop base pair. *J Biol Chem* **278**, 38892–38901 (2003).

[103] Clerte, C. & Hall, K. B. Global and local dynamics of the U1A polyadenylation inhibition element (PIE) RNA and PIE RNA-U1A complexes. *Biochemistry* **43**, 13404–13415 (2004).

[104] Cheng, C.-W., Su, E. C.-Y., Hwang, J.-K., Sung, T.-Y. & Hsu, W.-L. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics* **9 Suppl 12**, S6 (2008).

[105] Wang, L. & Brown, S. J. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* **34**, W243–W248 (2006).

[106] Terribilini, M. *et al.* Prediction of RNA binding sites in proteins from amino acid sequence. *RNA* **12**, 1450–1462 (2006).

[107] Kumar, M., Gromiha, M. M. & Raghava, G. P. S. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* **71**, 189–194 (2008).

[108] Wang, Y., Xue, Z., Shen, G. & Xu, J. PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids* **35**, 295–302 (2008).

[109] Leontis, N. B. & Westhof, E. The annotation of RNA motifs. *Comp Funct Genomics* **3**, 518–524 (2002).

[110] Leontis, N. B. & Westhof, E. Analysis of RNA motifs. *Curr Opin Struct Biol* **13**, 300–308 (2003).

[111] Hermann, T. & Westhof, E. Non-Watson-Crick base pairs in RNA-protein recognition. *Chem Biol* **6**, R335–R343 (1999).

[112] Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202–W208 (2009).

[113] Thompson, W., McCue, L. A. & Lawrence, C. E. Using the Gibbs motif sampler to find conserved domains in DNA and protein sequences. *Curr Protoc Bioinformatics* **Chapter 2**, Unit 2.8 (2005).

[114] Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).

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

[116] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990). URL `http://dx.doi.org/10.1006/jmbi.1990.9999`.

[117] Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195–197 (1981).

[118] Roshan, U. & Livesay, D. R. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* **22**, 2715–2721 (2006). URL `http://dx.doi.org/10.1093/bioinformatics/btl472`.

[119] Roshan, U., Chikkagoudar, S. & Livesay, D. R. Searching for evolutionary distant rna homologs within genomic sequences using partition function posterior probabilities. *BMC Bioinformatics* **9**, 61 (2008). URL `http://dx.doi.org/10.1186/1471-2105-9-61`.

[120] Hertel, J. *et al.* Non-coding rna annotation of the genome of trichoplax adhaerens. *Nucleic Acids Res* **37**, 1602–1615 (2009). URL `http://dx.doi.org/10.1093/nar/gkn1084`.

[121] Mosig, A., Sameith, K. & Stadler, P. Fragrep: an efficient search tool for fragmented patterns in genomic sequences. *Genomics Proteomics Bioinformatics* **4**, 56–60 (2006). URL `http://dx.doi.org/10.1016/S1672-0229(06)60017-X`.

[122] Gruber, A. R., Findei, S., Washietl, S., Hofacker, I. L. & Stadler, P. F. Rnaz 2.0: Improved noncoding rna detection. *Pac Symp Biocomput* **15**, 69–79 (2010).

[123] Chenna, R. *et al.* Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res* **31**, 3497–3500 (2003).

[124] Markham, N. R. & Zuker, M. Unafold: software for nucleic acid folding and hybridization. *Methods Mol Biol* **453**, 3–31 (2008). URL http://dx.doi.org/10.1007/978-1-60327-429-6_1.

[125] Hofacker, I. L. Rna secondary structure analysis using the vienna rna package. *Curr Protoc Bioinformatics* **Chapter 12**, Unit12.2 (2009). URL http://dx.doi.org/10.1002/0471250953.bi1202s26.

[126] McCaskill, J. S. The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers* **29**, 1105–1119 (1990). URL http://dx.doi.org/10.1002/bip.360290621.

[127] Lowe, T. M. & Eddy, S. R. trnascan-se: a program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Res* **25**, 955–964 (1997).

[128] Hertel, J. & Stadler, P. F. Hairpins in a haystack: recognizing microrna precursors in comparative genomics data. *Bioinformatics* **22**, e197–e202 (2006). URL http://dx.doi.org/10.1093/bioinformatics/btl257.

[129] Hertel, J., Hofacker, I. L. & Stadler, P. F. Snoreport: computational identification of snornas with unknown targets. *Bioinformatics* **24**, 158–164 (2008). URL http://dx.doi.org/10.1093/bioinformatics/btm464.

[130] Lowe, T. M. & Eddy, S. R. A computational screen for methylation guide snornas in yeast. *Science* **283**, 1168–1171 (1999).

[131] Schattner, P. *et al.* Genome-wide searching for pseudouridylation guide snornas: analysis of the saccharomyces cerevisiae genome. *Nucleic Acids Res* **32**, 4281–4296 (2004). URL http://dx.doi.org/10.1093/nar/gkh768.

[132] Lagesen, K. *et al.* Rnammer: consistent and rapid annotation of ribosomal rna genes. *Nucleic Acids Res* **35**, 3100–3108 (2007). URL http://dx.doi.org/10.1093/nar/gkm160.

[133] Weinberg, Z. & Ruzzo, W. L. Exploiting conserved structure for faster annotation of non-coding rnas without loss of accuracy. *Bioinformatics* **20 Suppl 1**, i334–i341 (2004). URL http://dx.doi.org/10.1093/bioinformatics/bth925.

[134] Klein, R. J. & Eddy, S. R. Rsearch: finding homologs of single structured rna sequences. *BMC Bioinformatics* **4**, 44 (2003). URL http://dx.doi.org/10.1186/1471-2105-4-44.

[135] Bafna, V. & Zhang, S. Fastr: fast database search tool for non-coding rna. *Proc IEEE Comput Syst Bioinform Conf* 52–61 (2004).

[136] Gautheret, D. & Lambert, A. Direct rna motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol* **313**, 1003–1011 (2001). URL http://dx.doi.org/10.1006/jmbi.2001.5102.

[137] Heyne, S., Will, S., Beckstette, M. & Backofen, R. Lightweight comparison of rnas based on exact sequence-structure matches. *Bioinformatics* **25**, 2095–2102 (2009). URL http://dx.doi.org/10.1093/bioinformatics/btp065.

[138] R., E. S. RNABOB: a program to search for RNA secondary structure motifs in sequence databases. *unpublished* .

[139] Grf, S., Strothmann, D., Kurtz, S. & Steger, G. Hypalib: a database of rnas and rna structural elements defined by hybrid patterns. *Nucleic Acids Res* **29**, 196–198 (2001).

[140] Krger, J. & Rehmsmeier, M. Rnahybrid: microrna target prediction easy, fast and flexible. *Nucleic Acids Res* **34**, W451–W454 (2006). URL http://dx.doi.org/10.1093/nar/gkl243.

[141] Dimitrov, R. A. & Zuker, M. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys J* **87**, 215–226 (2004). URL http://dx.doi.org/10.1529/biophysj.103.020743.

[142] Andronescu, M., Zhang, Z. C. & Condon, A. Secondary structure prediction of interacting rna molecules. *J Mol Biol* **345**, 987–1001 (2005). URL http://dx.doi.org/10.1016/j.jmb.2004.10.082.

[143] Tafer, H. & Hofacker, I. L. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* **24**, 2657–2663 (2008).

[144] Richter, A. S., Schleberger, C., Backofen, R. & Steglich, C. Seed-based intarna prediction combined with gfp-reporter system identifies mrna targets of the small rna yfr1. *Bioinformatics* (2009). URL http://dx.doi.org/10.1093/bioinformatics/btp609.

[145] Noirot, C., Gaspin, C., Schiex, T. & Gouzy, J. Learn: a platform for detecting, clustering and annotating non-coding rnas. *BMC Bioinformatics* **9**, 21 (2008). URL http://dx.doi.org/10.1186/1471-2105-9-21.