1

Regulatory Signals in Genomic Sequences

Sonja J. Prohaska, Axel Mosig, and Peter F. Stadler

Summary. Gene expression is a complex multiple-step process involving multiple levels of regulation, from transcription, nuclear processing, export, post-transcriptional modifications, translation, to degradation. Over evolutionary time-scales, many of the interactions determining the fate of a gene have left traces in the genomic DNA. Comparative genomics, therefore, promises a rich source of data on the functional interplay of cellular mechanisms. In this chapter we review a few aspects of such a research agenda.

1.1 Introduction

Gene expression is the process by which a gene's information is converted into a structural or functional gene product. This product is typically a protein, but might also be an RNA molecule or a complex of RNA and protein. A specific spatial and temporal distribution of these units within a cell is crucial for their function. The process of gene expression involves multiple steps, starting with the DNA in a state that makes the information accessible, transcription (DNA $\rightarrow$ RNA) and perhaps translation (RNA $\rightarrow$ protein), which is then followed by protein folding, posttranslational modification, and targeting. Once started, gene expression does not run through unaffected. Every step in the process is under tight control and actively regulates, or at least modulates, the flow through each checkpoint. Trapping intermediates at any step of the process may halt or even abort gene expression. Together, all regulatory effects from the gene to the functional gene product determine whether a gene product exceeds its threshold of expression to be effective. Therefore, the state of expression is not simply on or off.

In recent years it has become apparent that gene expression is a complex network comprising different, often often inter-twined, regulatory layers. A few of these mechanisms, such as the binding of transcription factors to the DNA, leave direct traces in the genomic sequences that can be detected and deciphered by comparative approaches. In other cases, gene regulation is afforded by *trans*-acting RNAs, first and foremost microRNAs. In this situation,

one first has to identify the transacting regulator before it becomes possible to search for the target that it regulates.

From a comparative genomics perspective, on the other hand, we can identify a plethora of evolutionary conserved DNA sequences that apparently do not code for proteins. Among these signals are also sizable regions with very high levels of sequence conservation and no reported function [29, 82, 127, 128]. The question that we at least begin to address in this chapter is how we can identify evolutionary conserved DNA, and how we can determine the mechanisms that they are involved in. While it is clear that comparative genomics cannot by itself elucidate the complete complexity of cellular regulation, it has also become clear in recent year that over evolutionary timescales, this regulatory network has left trace evidence at the DNA level. This requires, however, an understanding of the many distinct mechanisms.

Nucleic acid sequence motifs are, with a few examples such as self-splicing introns and some non-coding RNAs, not catalytically active. Additional proteins or (sometimes) ncRNAs are therefore involved that exert their regulatory function by binding either to the DNA or to the transcribed RNA at regulatory elements. Regulatory mechanisms that alter the components binding directly or indirectly to the sequence motifs are beyond the scope of this chapter. The generic mechanisms involving the motifs and their accessibility are listed in Table 1.1.

Changes at the nucleotide sequence level occur either by recombination/repair processes or covalent modification of single bases. Direct DNA modifications yield stable or even irreversible gene expression patterns in descending cells. Genomic imprinting reduces gene expression to one parental allele through DNA methylation. Once established, the methylation pattern is rather constant and hardly reversible.

Eukaryotic DNA is packed into a compact structure, the chromatin. Every 150 base pairs (bp), the linear DNA molecule is wrapped around a protein core in 1.65 turns, forming the nucleosome. Regulation at the epigenetic level concerns modifications of the histone protein core. The pattern of acetylation, (mono-, di-, and tri-)methylation, phosphorylation and other covalent modifications at about 40 different amino acids of the five different histone proteins (H4, H3, H2A, H2B, and H1) is also referred to as the histone code. Histone modification patterns are able to recruit specific protein complexes just like binding sites on the DNA or set the chromatin state. Heterochromatin or condensed chromatin is in the silent state, while euchromatin or open chromatin is transcribable mainly due to histone acetylation.

Insulators describe a phenotype rather than a single kind of element with a fixed mechanism of action. They have the ability to protect genes they surround from the influence either of outside enhancers or inactivating chromatin structures. An important part of the underlying mechanism might be the formation of insulator bodies. Insulator binding proteins form complexes that divide the chromatin into looped domains that are functionally isolated from one another. This could be a step toward regulation by discrete subnu-

Table 1.1. Overview of major regulatory modes in eukaryotic cells

Regulatory mechanism	Effect on gene expession	Example (organism)	Reference
DNA rearrangements	selective/irreversible	V(D)J-joining (human)	[105]
Site-specific recombination	selective/reversible	mating-type switching (yeast)	[26]
DNA amplification	enhancing	chorion genes (*Drosophila*)	[22]
DNA methylation of CpG dinucleotides	silencing/imprinting	parent-of-origin-specific silencing (human)	[120]
DNA demethylation by DNA repair	enhancing	glycosylase at polycomb genes (*Arabidopsis*)	[96]
Histone code	silencing/enhancing	everywhere	[81]
Heterochromatin barrier (fixed/flexible)	insulator/silencing/enhancing	USF binding at HS4 (chicken)	[146]
Enhancer blocker	insulator/silencing	su(Hw) (*Drosophila*), CTCF (human)	[146]
Enhancer promoter contact	insulator/silencing/enhancing	trithorax at Ubx promoter (*Drosophila*)	[146]
Nuclear matrix attachment region (MAR)	insulator/silencing/enhancing	lysozyme locus (chicken), tyrosinase locus (mouse)	[146]
Subnuclear compartment	silencing/enhancing	LCR at β-globin gene (human)	[145]
Gene competition for enhancers	silencing/enhancing	β-globin gene (human)	[131]
Chromatin remodeling	silencing/enhancing	SWI/SNF at PHO5 or PHO8 genes (yeast)	[15]
RNA-directed transcriptional gene silencing	silencing	X chromosome inactivation (human)	[7]
Promoters (TF binding sites)	basal	everywhere	[14]
Enhancer (TF binding sites)	enhancing	everywhere	[148]
Silencer/repressor (TF binding sites)	silencing	Ume6 at URS1-containing promoters (yeast)	[148]
Alternative transcription start sites	silencing/enhancing	IGF-1 (human)	[77]
Antisense transcripts	silencing/enhancing	frq gene (*Neurospora*)	[18]
Regulation of elongation phase	silencing/enhancing	Fkh at CLB2 locus (yeast)	[98]
Pre-mRNA processing (nucleus)	silencing/enhancing	everywhere	[68]
Alternative splicing (nucleus)	selective/silencing/enhancing	sex lethal (*Drosophila*)	[73]
Trans-splicing (nucleus)	selective/alteration	SL RNA at all genes (*Trypanosoma*	[78]
mRNA editing (nucleus)	alteration	ADAR at GluR mRNA (human)	[6]
Sequestration of mRNA (nucleus)	silencing	Rrp6 at exosomes (yeast)	[117]
Nonsense-mediated mRNA decay (nucleus)	silencing	pseudogenes	[144]
mRNA export	silencing/enhancing	EJC at spliced mRNAs (*Drosophila*)	[117]
RNA-directed mRNA degradation	silencing	DCL1 mRNA (*Arabidopsis*)	[80]
Degradation and stability of RNAs (cytoplasm)	silencing/enhancing	HuR at ARE-containing RNAs (human)	[147]
mRNA localization (cytoplasm)	silencing/enhancing	ASH1 (yeast)	[41]
Alternative translation start sites	silencing/enhancing	IRES	[69]
Scanning for translation start sites	silencing/enhancing	uORF at GCN4 (yeast)	[38]
Translation initiation regulators	silencing/enhancing	CPEB at CPE-containing mRNA (human)	[38]
RNA-directed translational gene silencing	silencing	lin-4 miRNA at lin-14 mRNA (*C. elegans*)	[38]

USF = upstream stimulatory factor, HS4 = hypersensitive site at the LCR of the β-globin gene, su(Hw) = suppressor of hairy wing protein, CTCF = CCCTC-binding factor, MAR = nuclear matrix attachment region, LCR = locus control region, SWI/SNF = remodeling complex, PH05 = repressible acid phosphatase precursor, PHO8 = repressible alkaline phosphatase precursor, TF = transcription factor, Ume6 = transcriptional regulator, URS1 = Ume6-binding site, frq = frequenin, Fkh = fork head protein, CLB2 = G2/mitotic-specific cyclin 2, ADAR = double-stranded RNA-specific adenosine deaminase, GluR = glutamate receptor, mRNA = messenger RNA, Rrp6 = exosome complex exonuclease, EJC = exon-junction complex, DCL1 = Dicer-like 1, HuR = Human-antigen R (stability factor), ARE = AU-rich element, ASH1 = daughter cells HO repressor protein, IRES = internal ribosomal entry site, uORF = upstream open reading frame, GCN4 = general control protein, CPEB = cytoplasmic polyadenylation element binding protein, CPE = cytoplasmic polyadenylation element, miRNA = micro RNA.

clear compartment. For example, actively expressed genes migrate to nuclear compartments enriched in RNAPol II, so called transcription factories, while inactive genes loop out. Such agglomerations may serve to raise the local concentration of associated components, favouring interactions that might not otherwise occur.

The promoter is the assembly site for the basal transcription machinery right next to the transcription start site. Transcription factors (TFs) binding to enhancers facilitate recruitment of RNA polymerase to the promoter if physical contact can be established by cofactors. Silencers, on the other hand, circumvent such an interaction and therefore initiation of transcription. In general, binding sites for TFs are short (4–12 bp) and occur clustered upstream of the promoter sequence. While there are numerous examples where the context (i.e., order, orientation, distance, presence of certain TFs) of TF binding sites is functionally relevant, there is an equally large number of examples where the context is not relevant.

The following elongation of transcription and all regulatory steps at the RNA level that take place in the nucleus are coupled to a large extent. For example, nuclear export of RNAs is linked to the subnuclear compartment of transcription, transcription elongation, mRNA processing (splicing), and mRNA stability. Once the mRNA is exported to the cytoplasm, it is either degraded or translated, but it might also be stored for later use.

Translation of mRNA is the final step in gene expression that involves nucleic acid sequence elements in control. Not only upstream regulatory elements like secondary structures or upstream ORF may effect scanning of the small ribosomal subunit for the initiation codon. Close proximity of the 5' and 3' end of the mRNA allows protein binding sites located in the 3'-UTR to control translation initiation. In fact, most known regulatory sequences, and miRNA binding sites are found within the 3' UTR.

The regulatory mechanisms and phenomena described above leave more or less visible traces in their genome sequences. For some regulatory elements, the corresponding traces on DNA level are very well understood and have been studied in much detail. This chapter reviews known sequence characteristics of regulatory elements. Whenever one is available, we will give an overview over the corresponding computational methods for unveiling those traces in genomic sequences.

1.2 Gene Finding

The most conspicuous traces found in a genome sequence arise from protein coding regions. Since proteins are key players in the gene regulatory network, identifying the positions of the protein coding genes in whole genome sequences is an elementary step. Beside identifying the protein coding sequences, genome annotations serve a second purpose, namely to obtain those regions in the vicinity of the annotated genes that contain *cis*-regulatory elements,

Table 1.2. Estimated number of protein coding genes

Species	**Estimated No. of Genes**	Ref.
Homo sapiens	20,000-25,000	[132]
Drosophila melanogaster	12,000	[133]
Caenorhabditis elegans	19,000	[135]

such as transcription factor binding sites. Furthermore, gene annotations give an estimate of the number of players involved in regulatory networks.

Once a complete genome sequence is available, a first step typically is to identify protein coding genes in the sequence by computational means, a task commonly referred to as *gene prediction* or *gene finding*. Due to statistically noticeable features such as being grouped in coding triplets or protruding traits such as start or stop codons, protein coding genes typically show comparatively strong signals in genome sequences. Consequently, a number of well-established methods have contributed to detecting protein coding genes in genomes. The first type of gene prediction methods, so-called *ab initio* methods, are based on considering a single genome sequence in combination with a probabilistic model involving multiple characteristic traits of transcriptional, translational, or splicing sites that are typically visible on sequence level. Approaches such as `GENSCAN` [13] or `Genie` [72] incorporate this information into a hidden Markov model for unveiling genomic regions that have a striking probability of being protein coding.

While the accuracy of *ab initio* gene prediction methods turned out to be principally limited [43, 118], more reliable results can be obtained by *comparative gene prediction* approaches, which incorporate pairwise alignments of the underlying genomes produced by programs such as `Blastx`. Due to a very specific selectional pressure on the coding triplets of protein coding regions, predictions produced by programs such as `Twinscan` [67] or `Procrustes` [40] yield much more reliable results than *ab initio* methods.

As has been demonstrated by several studies, incorporating issues such as protein similarity or expressed sequence tags may enhance the reliability of gene prediction methods [54, 113, 150]. For surveys on gene prediction methods, we refer to [12, 34, 39, 71]. Gene prediction methods yield estimates of the number of (protein coding) genes, some of which are shown in Table 1.2.

While detecting protein coding genes appears to be a largely resolved problem, finding non-coding RNA genes is much more involved. For details on RNA gene prediction, see section 1.4.

1.3 Identifying *Cis*-Regulatory Elements

Once the protein coding genes and non-coding RNAs (see section 1.4) as the key players in the regulatory network and their coding regions are known,

Table 1.3. Major modes of transcription.

RNA polymerase	Promoter	Location relative to start site	Transcript	Function
Pol I	Core element UCE (upstream control element)	−45 to +20 −180 to −107	pre-rRNA (28S, 18S, 5.8S)	components of the ribosome; translation
Pol II	TATA-Box Initiator CpG islands no	−25 to −35 -100	mRNA snRNA (U1-4) LINEs	protein coding genes components of the spliceosome; mRNA splicing Retrotransposon
Pol III	Type 1: A-box, C-box Type 2: A-box, B-box Type 3: TATA-Box internal	+50 to +80 +10 to +60 −30 to −70	5S rRNA tRNA snRNA (U6) 7SL RNA SINEs	component of large ribosomal subunit translation components of the spliceosome; mRNA splicing component of the SRP (signal recognition particle); protein transport to ER (endoplasmatic reticulum) Retrotransposon

one is naturally interested in their *cis*-regulatory elements, that is, sequence elements associated with the gene to be regulated that serve as sequence-based "addresses" for their regulators. On the level of transcription regulation, the most striking sequence signals of a gene are given by the basal promoter and the proximal promoter. In mammals, 60% of all promoters colocalize with regions of high C+G content, known as CpG islands. A feature that can also be used to find unknown genes. In the immediate upstream region of the basal an proximal promoter, auxiliary binding sites can be located for further transcription factors, which are often observed to be organized in regulatory modules.

1.3.1 Polymerases and Associated Promoters

Transcription of DNA into RNA is performed by the three different types of RNA polymerases. For the modes of transcription associated with the different RNA polymerases, see Table 1.3. Each of the polymerases requires certain *cis*-acting elements in order to initiate transcription; due to its crucial relevance in

transcribing mRNA necessary for protein coding genes, much effort has been spent on studying the polymerase II core promoter as the minimal stretch of contiguous DNA sufficient for initiating transcription. In most (yet not all) polymerase II transcribed genes, the core promoter contains the *TATA box*, which is located 25 bases upstream of the transcription start site. The TATA box is a usually 6 nucleotide long sequence motif characterizing the binding site for the tata-binding-protein (TBP). TBP usually interacts with other transcription factors, whose binding sites are typically found within 40 nucleotides (nt) upstream to the transcription start site. For details on the RNA polymerase II core promoter, we refer to the survey by Butler and Kadonaga [14]. The action of the polymerase II promoter is often enhanced by several distal promoters organized in *cis*-regulatory modules (see section 1.3.3).

Polymerase I transcripts are also regulated by a a core promoter, which is separated by about 70 bp from a second elementary promoter element, the so-called *upstream control element* (UCE). In place of the TF for polymerase II, pol I requires two core transcription factors, namely UBF1 (*upstream binding factor*) binding to a GC-rich region and SL1 (*selectivity factor*).

Promoters for RNA polymerase III occur in several variants. First, they may consist of bipartite sequences downstream of the start point, with promoter element *boxA* separated from either of the promoter elements *boxC* or *boxB*. Second, some U snRNA genes are regulated by upstream type promoters involving an octamer binding site, a so-called *proximal sequence element* and a TATA box.

1.3.2 Identification of Transcription Factor Binding Sites

Transcription factors are known to bind to short, specific sequences of DNA. Experimental evidence obtained by techniques such as DNase footprinting [36] and gel-shift assays [37] suggests that protein-DNA binding of transcription factors involves a relatively short, contiguous DNA segment, whose length usually ranges between 8 and 15 nucleotides. Repositories of known transcription factors and their experimentally derived binding site motifs can be found in databases such as TRANSFAC [45] or JASPAR [123]. However, experimental determination of binding sites and their relevance *in vivo* takes a significant effort, so that numerous computational approaches have been proposed for determining candidates for TFBSs *in silico*.

While the length of the binding sites corresponding to one specific TF is observed to be essentially constant, the individual positions of the binding site sequences may vary up to a certain degree. Hence, to derive a suitable model of the sequences that a TF binds to, different notions of describing TFBS sequence variability have been proposed. Such models are also important in the context of computationally determining TFBSs based on comparative genomics approaches.

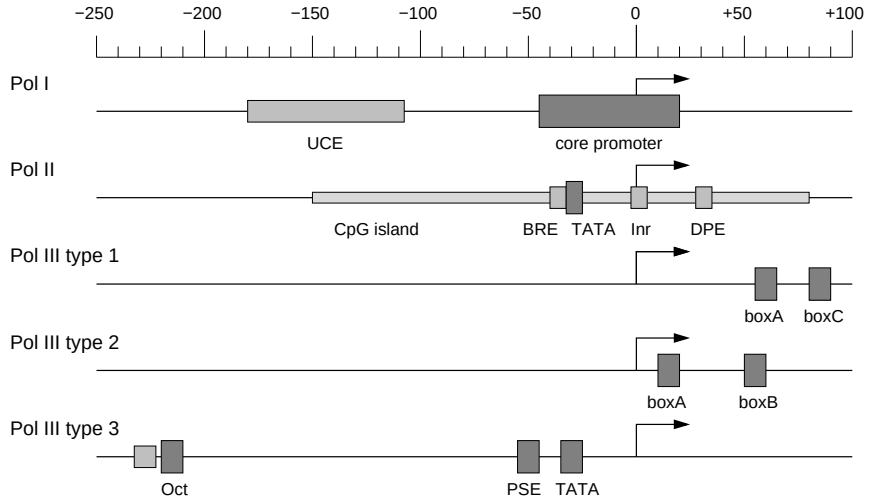


Fig. 1.1. Core motifs of the different promoter types. Motifs in dark gray are less dispensable than motifs in light gray. Any specific promoter may contain just a subset or, in the worst case, none of these motifs. UCE = upstream control element, BRE = TFIIB recognition element, Inr = initiator element, DPE = downstream core promoter element, Oct = octamer binding site, PSE = proximal sequence element. The arrow indicates the transcription start site at +1.

Different observed binding sites corresponding to a given transcription typically show a high degree of sequence similarity; moreover, the observed binding site motifs have the same length ℓ. To capture the observed binding sites in one unique structure, we define a *binding site model of length* ℓ as a mapping $M\colon \Sigma^\ell \to \mathbb{R}\geq 0$ assigning a weight to each sequence of length ℓ over the DNA alphabet Σ. While $M(s)$ ideally should be related to the physical binding affinity of sequence s binding to the transcription factor modeled by M, $M(s)$ usually is obtained on the basis of the frequency of observed or putative binding sites in a given set of genomic sequences.

The concept of binding site models introduced above is too general in many situations: first, there are usually not enough data to derive reasonable weights for each DNA sequence of length ℓ and second, storing and retrieving the complete mapping M would be too expensive. Hence, several simplified models of TFBSs have been established. The most simple model of a binding site model is to derive a *consensus sequence*. In this model, each of the ℓ positions is associated with a subset of the DNA nucleotide alphabet. A given sequence fragment s of length ℓ is assigned score $M(s) = 1$ if at each position, the nucleotide of the fragment is contained in the corresponding nucleotide set of the model; otherwise, we have $M(s) = 0$. Alternatively, one can define $M(s)$ as the number of positions in s where the nucleotide in s is contained in the corresponding set in M.

Consensus sequences disregard major information contained in the sequences used for deriving the consensus model; namely, they do not take into account frequencies of occurrence. This is overcome in the most established way of TFBS modeling, namely *position weight matrices* (PWMs). In the PWM model (sometimes also referred to as a *position specific weight matrix*), each of the ℓ positions of the binding site is assumed to be distributed independently: for each position i, we are given a probability density function $pi\colon \Sigma \to [0,1]$ over the four nucleotides. Given $s := s1 \dots s\ell \in \Sigma^\ell$, this allows us to define

$$M(s) := p1(s1) + \cdots + p\ell(s\ell).$$

PWMs can be derived canonically from a collection of sequences $S1, \dots, SM \in \Sigma^\ell$: for $x \in \Sigma$ and $i \in \{1, \dots, \ell\}$, let $\nu(x,i)$ denote the number of sequences in which letter x occurs at position i. By setting $pi(x) := \nu(x,i)/M$, we indeed obtain a PWM model. In practice, the sequences $S1, \dots, SM$ are typically obtained either from a set of experimentally determined binding sites or from motif discovery methods.

PWMs, however, disregard any information about the correlation between sites that may be contained in the sequences that a matrix was derived from. As a remedy, Pudimat et al [112] have developed a more sophisticated way of modeling TFBSs based on parameter estimation in a Bayesian belief network. As opposed to most other approaches of TFBS modeling, their approach allows us to model correlations between the individual sequence positions. Another approach for modeling dependencies between positions in PWMs based on χ^2 statistics has been investigated in [33].

While obtaining models for TFBSs from experimental data is relatively easy, deriving them computationally from genomic sequences is a complex problem. Essentially all approaches are based on comparative genomics in the sense that they seek for motifs contained in each, or at least most, of K promoter regions belonging to K co-regulated (or orthologous) genes. The returned motifs usually result from optimizing a scoring function that measures how well a candidate motif statistically differs from global properties of the promoter sequences.

Among the earliest nontrivial approaches to extracting overrepresented short motifs as potential TFBSs, Hertz and Stormo [48] proposed a greedy algorithm. Their `CONSENSUS` approach starts with a position weight matrix derived from a single sequence of a fixed length, which is extended to a pairwise alignment of the same width by considering a best-matching subsequence of the second sequence. The algorithm proceeds by successively adding one subsequence of each remaining input sequence to obtain the final PWM, along with a p-value that allows us to assess the statistical significance of the result.

A different approach based on the expectation maximization (EM) algorithm is investigated in `MEME` [4], improving a previous approach by Lawrence and Reilly [75]. The EM-based approach starts with an a priori guess for a position weight matrix representing a binding site of fixed length ℓ, which is

then improved according to the input sequences in each of the subsequent iteration steps. A single iteration step works as follows: for each subsequence ℓ of the input sequences, the score of the current matrix is computed. After normalization, the matrix entries are updated by summing up the individual position contributions of each of the length ℓ subsequences, weighted by its corresponding normalized probability computed before. The resulting new matrix is then used as input for the next iteration step, until convergence of the process is observed.

`AlignACE` developed by Roth and Hughes's group [56, 119] is yet another approach to obtain PWMs from genomic regulatory sequences. `AlignACE` based on Gibbs sampling, enhancing approaches previously used for locally aligning motifs in protein sequences such as [74] in a way such that both strands of the input sequences are considered. Furthermore, single motifs that were found are masked iteratively to allow for the extraction of more than one binding site motif.

1.3.3 Discovering Regulatory Modules

As numerous studies demonstrate, transcription factors exhibit their function synergistically through complexes of several transcription factors activating or deactivating gene expression by binding to their corresponding binding sites [27, 152], which thus form the building blocks of regulatory modules. On the genome level, regulatory modules are characterized by binding sites being located close to each other, usually within a segment whose length does not exceed a few hundred nucleotides.

In recent years, a number of approaches have been developed in the context of discovering *cis*-regulatory modules. Kel-Margoulis et al. [65] propose a method based on identifying clusters with the property that pairwise distances between occurrences of TFBSs range within certain bounds; sets of binding sites that maximize a certain cluster score are searched by the means of a genetic algorithm. Other methods are based on probabilistic methods [109] or require (only sparsely available) knowledge about interactions between transcription factors such as the algorithm presented in [130].

Among the most established methods, Sharan et al. proposed an approach implemented in the program `CREME` [129], which is conceptually somewhat related to our approach. Given a set of candidate binding sites, `CREME` seeks to identify motif clusters of limited length that occur more than once in a set of regulatory sequences. However, the capabilities of the `CREME` approach is limited to discovering repetitive occurrences of modules that contain *precisely* the same set of binding sites. While biological data indeed indicate that functionally related modules share a significant number of binding sites, modules observed in a given regulatory region might as well contain occurrences of known binding site motifs, which are not functional in the given context. If this number of additional, non-shared binding sites is non-zero, the method

underlying CREME does not allow us to discover such functional modules reliably.

To overcome this shortcoming, the hypothesis underlying the bbq approach [99] is that CRMs are characterized by sharing a significant number of common binding sites, but do not necessarily contain precisely the same set of binding sites. More formally, we are given a set of candidate binding sites $s1, \ldots, sm$ together with a set of genomic sequences $T1, \ldots, TK$. The role of the genomic sequences Tj is taken by the regulatory regions of genes that are suspected to share a regulatory module (due to being orthologous or having a similar expression scheme), while the binding site motifs can be derived from databases such as TRANSFAC or JASPAR. Alternatively, these motifs can be derived from $T1, \ldots, TK$ using the motif discovery approaches discussed in section 1.3.2. Finally, an upper bound for the length L (specified as a number of nucleotides) of the regulatory module is given as an input parameter.

The bbq approach starts with determining the occurrences of each motif si in each Tj and associating a color i with binding site si. For each occurrence of si, an interval of length $(L - |si|)$ ending at the position of the occurrence is introduced, so that one finally obtains K arrangements of colored intervals. By "stabbing" into this arrangement, one obtains a cell in this arrangement. Such a cell is associated with a set of colors, which corresponds to a set of binding sites occurring within a genomic subsequence whose length is at most L nucleotides. Finally, attempting to stab a maximum number of common colors in each of the K arrangements leads to the so-called *best-barbecue problem*. This problem leads to a natural combinatorial and geometric optimization problem that is NP-complete in general.

1.3.4 Phylogenetic Footprinting

Just as genomic regions that code for proteins or functional RNAs, regulatory elements are also subject to stabilizing selection. They evolve much more slowly than adjacent nonfunctional DNA, so that one can observe conserved islands of regulatory regions within intergenic or intronic regions. These conserved islands are commonly referred to as *phylogenetic footprints*, which can be detected by comparison of the sequences surrounding orthologous genes in different species. The loss of phylogenetic footprints as well as the acquisition of conserved non-coding sequences in some lineages, but not others, can provide evidence for the evolutionary modification of *cis*-regulatory elements.

While the motifs detected by the discovery methods discussed in section 1.3.2 can be seen as one particular type of footprints, one can often observe conserved islands that are much longer, up to several hundred nucleotides, than individual binding sites. Therefore, phylogenetic footprinting tools are usually based on pairwise or multiple local alignment algorithms such as blastz [125] or Dialign [97]. The tools PipMaker and MultiPipMaker [125] (among others) process these alignments in order to provide information on significantly conserved regions. The Tracker tool [111] assembles individual

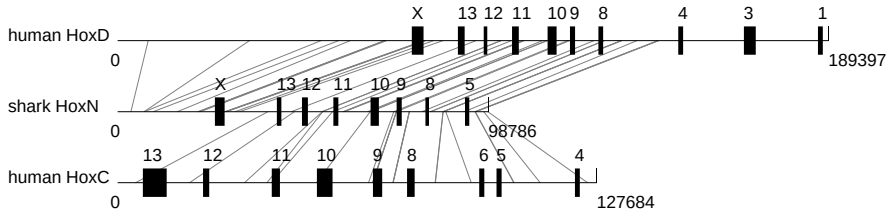


Fig. 1.2. Distribution of phylogenetic footprints in regulatory regions between the horn shark HoxN sequence and the human HoxC and HoxD sequences obtained by the `tracker` tool. Using this information for phylogenetic inference, this supports the hypothesis that the shark HoxN sequence is orthologous to the mammalian HoxD sequence [110]. Boxes indicate the location of the coding regions for the paralog groups 1 to 13, X denotes the *evx* gene. Lines conecting sequences represent phylogenetic footprints shared with the shark HoxN sequence.

pairwise `blastz` alignments into cliques of overlapping alignments. This results in the possiblity of listing alternative multiple local alignments if the pairwise matches are not consistent with one multiple alignment.

As demonstrated in [111], the analysis of sequence conservation of non-protein-coding DNA can be used to unveil the evolutionary origin of phenomena such as the duplication of Hox clusters in shark, human, and the duplicated zebrafish and Takifugu (Fig. 1.2). In this context, information contained in the regulatory regions yields insights that are not visible on the level of the corresponding protein coding regions.

1.4 Regulatory ncRNAs and RNA Motifs

1.4.1 Diversity of the RNA Inventory

Non-coding RNAs form a diverse group of transcripts with often poorly understood function. In contrast to protein-coding mRNAs there is little that they all have in common. One group, which itself is composed of a heterogeneous set of RNA families including tRNAs, the U6 snoRNA, the RNA component of the signal recognition particle, and a small number of less wellknown ncRNAs including 7SK RNA and Y RNAs is transcribed by RNA polymerase-III. Ribosomal RNAs, transcribed by pol-I, form a group by themselves. Almost all of these ncRNAs are evolutionarily very well conserved, and most of them are evolutionarily ancient.

In contrast, the majority of the known ncRNAs are transcribed by pol-II. These can be subdivided into messenger-RNA-like transcripts, such as *Xist*, which are typically spliced and polyadenylated, "structural ncRNAs" such as spliceosomal RNAs (snRNAs) and many microRNAs, which are neither spliced nor polyadenylated, and a class of functional RNAs that is processed from introns (in particular snoRNAs). Informally, it is useful to distinguish

a restricted group of "classical" ncRNAs containing the rRNAs, the pol-III transcripts listed above, spliceosomal RNAs, box-C/D and hox-H/ACA small nucleolar RNAs (snoRNAs), microRNAs, as well as telomerase RNA. As far as we know, these RNAs are evolutionarily old, they have distinctive RNA secondary structure, and most of them are reasonably well conserved at sequence level.

Recently, a number of small, non-mRNA-like ncRNAs have been found, for example in the nematode *Caenorhabditis elegans*, which does not appear to belong to one of the classical families, although at least some of them share the distinctive promoter features of tRNAs or pol-II transcribed snRNAs [28]. Bacterial genomes also contain a large and diverse set of small RNAs (sRNAs) in addition to the classical ncRNAs. A recent survey discusses 55 known *E. coli* sRNAs [47] and their conservation patterns within Enterobacteria. For a review of functional aspects of various bacterial sRNAs see [42]. An additional class of small anti-sense transcripts derived from UTRs is discussed in [64]. For a recent survey focusing on the regulatory effects of ncRNAs in eucaryotes, see [25].

The function of almost all mRNA-like ncRNAs remains unknown. The few well-studied examples, such as *Xist* or *H19*, have functions in imprinting [103].

Regulation by means of RNA can follow at least three distinct principles: `RNA switches` sense changes in temperature or chemical environment and react by conformational changes. *Cis*-acting RNA signals, often located in untranslated regions of mRNAs, are bound by proteins. *Trans*-acting RNAs, such as microRNAs, perform their function by binding to complementary nucleic acid sequence motifs.

1.4.2 RNA Secondary Structure Prediction and Comparison

From a theoretical perspective, computational RNomics draws much of its appeal from the fact that most quantities of interest can be computed exactly within the secondary structure model. In contrast to proteins, nucleic structures are dominated by a single, very specific type of interaction: the formation of Watson-Crick and wobble (G-U) base pairs. The resulting contact structures, which are predominantly stabilized by the stacking interactions of adjacent base pairs, are not only a convenient and routinely used representation [70, 104, 106, 124], they also quantitatively describe the energetics of RNA structure formation, and they form intermediate steps in the folding process itself.

Formally, a *secondary structure* is a set Ω of base pairs such that (1) each nucleotide position i is paired with at most one other nucleotide (i.e., Ω is a matching), and (2) base pairs do not cross, i.e., $(i,j),(k,l) \in \Omega$ with $i<j, k<l$ implies $j<k$ or $l<j$. The second condition ensures that two base pairs are either separated along the sequence or nested within each other. A secondary structure therefore can be seen as a circular matching. Drawing the bases along a circle, the base pairs form chords that do not cross. It follows

that RNA secondary structures can be dealt with by means of exact dynamic programming algorithms (Fig. 1.3).

A plethora of careful thermodynamic measurement confirmed that the energetics of RNA structures can be understood in terms of additive contributions of "loops" (Fig. 1.3) see [83, 84] and the references therein. Exact dynamic programming algorithms can be used to compute for example the minimum energy structure given any RNA sequence s [136, 153, 155]. The most frequently used implementations of these algorithms are `mfold` [153, 155] and the `Vienna RNA Package` [50, 53].

An RNA molecule, however, does not only form a single (ground state) structure; rather, there is an ensemble $\Sigma(s)$ of different structures Ψ that depend on the sequence s, which are populated in proportion to their Boltzmann factors $F(\Psi)/RT$. The partition function

$$Z = \sum \Psi \in \Sigma(s) \exp\left(-\frac{F(\Psi)}{RT}\right) , \tag{1.1}$$

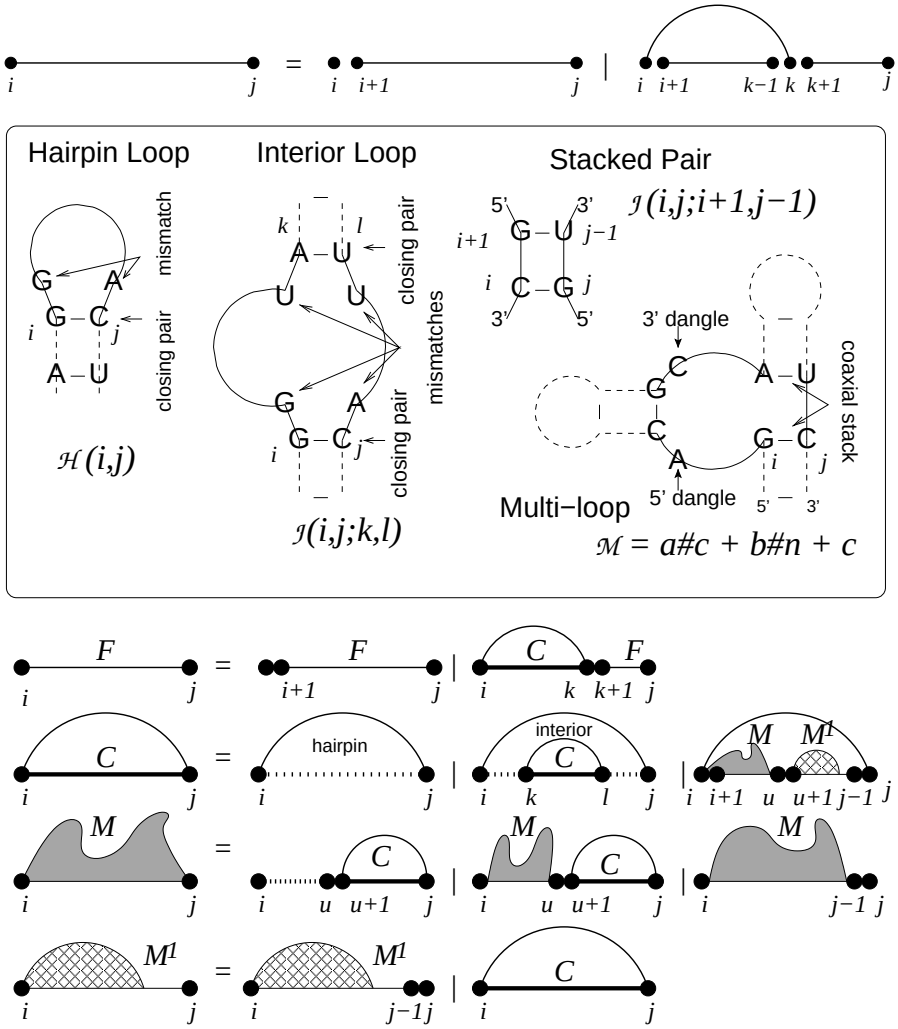
from which all thermodynamics quantities of interest can be readily derived, can be computed by the same type of dynamic programming approach [87].

1.4.3 Suboptimal Structures and RNA Switches

Some RNA molecules exhibit two competing conformations, whose equilibrium can be shifted easily by molecular events such as the binding of another molecule. This can be used to regulate of gene expression, when the two mutually exclusive alternatives correspond to an active and in-active conformation of the transcript [46, 90]. While most riboswitches were found in bacteria, where they regulate several key metabolic pathways [11, 101], metabolite-binding RNA domains are also present in some eukaryotic genes [134]. An early computational study concluded that RNA switches are readily accessible in evolution and are therefore probably not exceptional instances of unusual RNA behaviour [35]. These findings, and the fact that riboswitches bind their effectors directly without the need of additional factors, suggest that riboswitches represent one of the oldest regulatory systems [139].

1.4.4 Detection of Functional RNAs in Genomic DNA

Large-scale efforts to uncover the human and mouse transcriptomes, using very different experimental techniques including tiling arrays [8, 21, 61, 63], cDNA sequencing [58, 102], and unbiased mapping of transcription factor binding sites [17], agree that a substantial fraction of these genomes is transcribed and that the majority of these transcripts do not code for proteins. It is still unclear at present, however, which fraction represents functional non-coding RNAs (ncRNAs), and which constitutes "transcriptional noise" [57].



$$
\begin{aligned}
Fij &= \min\{Fi+1, j,\ \ \min i < k \le jCik + Fk+1, j\} \\
Cij &= \min\{\mathcal{H}(i,j),\ \ \min i < k < l < jCkl + \mathcal{I}(i,j;k,l), \\
&\qquad \min i < u < jMi+1, u + M^1u + 1, j-1+a\} \\
Mij &= \min\{\min i < u < j(u-i+1)c + Cu + 1, j + b, \\
&\qquad \min i < u < jMi, u + Cu + 1, j + b,\ \ Mi, j-1+c\} \\
M^1ij &= \min\{M^1i, j-1+c,\ \ Cij + b\},
\end{aligned}
$$

Fig. 1.3. RNA folding in a nutshell. *Caption continued overleaf...*

Fig. 1.3, continued.
Top: The basic recursion for RNA folding is based on the observation that each structure either terminates in an unpaired base or in a base pair that then separates the structure into two *independent* parts: the one enclosed by the base pair, and the one outside.
Box: The standard energy model distinguishes three types of loops: hairpin loops with a single closing pair, interior loops (including bulges and the stabilizing stacking pairs) that are delimited by two base pairs, and multiloops at which the structure branches. For the latter the energy model assume additive contributions depending on the number of branches $\#c$ and the number of unpaired bases $\#n$ in the loop.
Middle: Using the loop-based energy model complicated the recursion since one now has to distinguish the different types of loops because of their distinct energy contributions. Instead of a single array storing the optimal energies Fij for substructure on the subsequence $x[i..j]$, one now need a few auxiliary arrays that correspond to restricted classes of structures. For instance, Cij is the optimal energy subject to the constraint that i and j form a base pair.
Bottom: We give the complete recursion for energy minimization in the loop-based energy model. Replacing minima by sums, and sums by products leads leads to the recursions for the partition function Z.

Genome-wide computational surveys of ncRNAs, on the other hand, have been impossible until recently, because ncRNAs do not share common signals that could be detected at the sequence level. An exception are bacterial genomes, where a purely sequence-based machine learning approach was fairly successful [122].

Most of the "classical" ncRNAs mentioned above, however, have characteristic (secondary) structures that are functional and hence are well conserved over evolutionary time scales. The stabilizing selection acting on the secondary structure causes characteristic substitution patterns in the underlying sequences: Consistent and compensatory mutations replace one type of base pair by another one in the paired regions (helices) of the molecule. In addition, loop regions are more variable than helices. These patterns not only have a significant impact on phylogenetic inference based on ribosomal RNA sequences (see, e.g., [62] and the references therein), but it also can be exploited for ncRNA detection in comparative computational approaches. Examples are the `alidot` [51] and `qrna` [116] programs. Related approaches predict consensus secondary structures for a set of aligned sequences [52, 107].

A second effect of stabilizing selection for RNA secondary structure is even easier to measure. It was first suggested by Maizel's group that functional RNA elements should have a more stable secondary structure than comparable random sequences [19, 76].

As demonstrated in [137], selection for structure implies that in the long run sequences evolve that are more robust against mutations, that is, for which a larger fraction of mutations does not lead to a change in the ground state structure. This effect can, for example, be detected in viral RNA structures [140]. Mutational robustness, however, is in turn strongly correlated with the

thermodynamic stability of the ground state structure [2, 149]. Thus we expect that the ground states of functional RNA structures should be thermodynamically more stable than expected by chance, independently of whether there is a direct selection pressure for thermodynamic stability or not. While this effect can indeed be demonstrated [23], it is not statistically significant enough for reliable ncRNA detection [115]. It can be quite large for specific classes of ncRNAs, in particular microRNAs, however [10, 141].

Combinations of thermodynamic stability and information on gene structure such as positions of rho-independent terminators were quite successful for ncRNA prediction in intergenic regions of prokaryotic genomes [16, 79]. Such methods cannot be employed in eukaryots because of their much larger genome size and the much more complex gene structures.

Sufficient statistical power for ncRNA detection in eukaryotic genomes can be obtained, however, by combining measures for both thermodynamics stability and structural conservation. An implementation of such a combined approach is the `RNAz` program [143]: A structure conservation index (SCI) is computed by comparing the predicted minimum free energies of the sequences in an alignment with a consensus energy, which is computed by incorporating covariation terms into a free energy minimization computation [52]. Thermodynamic stability is quantified by means of a z-score that measures the folding energy relative to shuffled sequences (a regression approach replaces time-consuming shuffling methods). A support vector machine then classifies an alignment as "structured RNA" or "other" based on z-score and SCI. The significance of the classification is quantified as "RNA-class probability" p.

Various computational screens [1, 16, 20, 79, 116, 122] predict several hundred ncRNA candidates. These predictions, however, show relatively little mutual overlap in general. Indeed, the majority of bacterial sRNAs was discovered based on computational predictions and subsequent experimental verification.

A `RNAz` survey based on the most conserved parts of the vertebrate genomes estimates that the ncRNA content of mammalian genomes is comparable to their proteincoding genes [142], and hence at least an order magnitude larger than in nematodes. In contrast, only a few thousand structured RNAs in the urochordate *Ciona intestinalis* [93] and in the nematode *C. elegans* [28, 94]. Only a few hundred ncRNAs appear to be present in the yeast *Saccharomyzes cerevisiae* [88]. This indicates that higher vertebrates have dramatically expanded their ncRNA inventory relative to their complement of protein coding genes. This is consistent with the assumption that the function of the ncRNAs is primarily regulatory [85, 86].

1.4.5 RNA-RNA Interaction

Algorithmically, the "co-folding" of two RNAs can be dealt with in the same way as folding a single molecule by concatenating the two sequences and using different energy parameters for the loop that contains the cut-point between the two sequences. A corresponding `RNAcofold` program is described in [53];

the `pairfold` program [3] also computes suboptimal structures in the spirit of `RNAsubopt` [149]. A restricted variant of this approach is implemented in the program `RNAhybrid` [114] as well as `RNAduplex` from the Vienna RNA package, see also [30, 154]: here secondary structures within both monomers are neglected so that only intermolecular base pairs are taken into account. The program `bindigo` uses a variation of the Smith-Waterman sequence alignment algorithm for the same purpose [49].

The most prominent application of RNA co-folding algorithms is the prediction of microRNA target genes [9, 70, 95, 104, 124, 151]. The biological activity of siRNAs and miRNAs is influenced by local structural characteristics of the target mRNA. In particular, the binding site at the target sequence must be accessible for hybridization in order to achieve efficient translational repression. Recent contributions [106, 124] suggest two significant parameters: the stability difference between 5' and 3' end of the siRNA, which determines which strand is included into the RISC complex [66, 126] and the local secondary structure of the target site [9, 70, 95, 104, 124, 151].

The energetics of RNA-RNA interactions can be understood in terms of two contributions: the free energy of binding consists of the contribution ΔGu that is necessary to expose the binding site in the appropriate conformation, and the contribution ΔGh that describes the energy gain due to hybridization at the binding site. The first term can be computed from a partition function computation as described above, and the second term is obtained through a version of the co-folding algorithm. Comparison with the partition function of the isolated systems and standard statistical thermodynamics can be used to explicitly compute the concentration dependence of RNA-RNA binding [30].

1.4.6 RNA-Protein Interaction

In recent years an increasing number of functional features has been reported in the untranslated regions of eukaryotic mRNA [60, 92, 108]. Well-known motifs include internal ribosomal entry sites (IRES) in viral as well as cellular mRNAs [55, 108, 121], and the AU-rich elements (ARE) [5, 89]. In many cases, secondary structure motifs are recognized by regulatory proteins with only highly degenerate, or no sequence constraints at all [91, 138]. In such cases, the thermodynamics of RNA folding can influence binding specificities.

Consider a (protein) ligand that can bind to certain set RNA$*$ of structural conformations a given RNA molecules:

$$\text{Ligand} + \text{RNA}* \rightleftharpoons \text{Ligand} \cdot \text{RNA}$$

The law of mass action implies that the concentrations [RNA$*$], [Ligand], and [Ligand $\cdot$ RNA] of free accessible RNA, free protein, and complex are related through the dissociation constant

$$Kd = \frac{[\text{RNA}*]\ [\text{Ligand}]}{[\text{Ligand} \cdot \text{RNA}]} \tag{1.2}$$

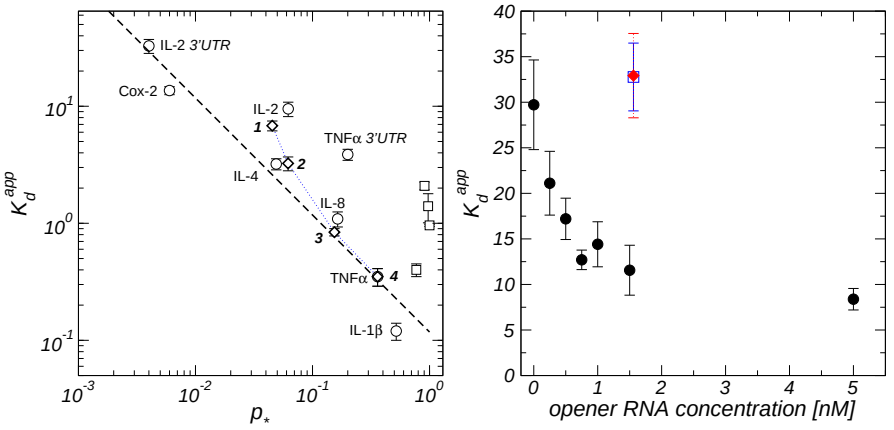


Fig. 1.4. Left: Apparent dissociation constants for *HuR*-mRNA complexes at $23.5°C$ for natural ARE and UTR sequences (○), artificial molecules (□), and designed mutants of the tumor necrosis factor α 3'UTR (◊) [44]. Right: Effect of a complementary opener of length $N0 = 20$ on *in vitro* *HuR*/RNA affinities. The apparent affinity of recombinant HuR to IL-2 3'UTR was determined in the presence and absence of the opener Op3 (black circles) and of the negative controls with 1D-FIDA detection. Data redrawn from [89].

Writing $A(s) \subseteq \Sigma(s)$ for the accessible structures of our RNA molecule s we obtain

$$[\mathrm{RNA}*] = p* \, [\mathrm{RNA}] \tag{1.3}$$

where $p*$ is the fraction of accessible secondary structures, which can be computed as a ratio of two partition functions

$$p* = \sum \Psi \in A(s) p(\Psi) = \frac{1}{Z} \sum \Psi \in A(s) \exp\left(-\frac{F(\Psi)}{RT}\right) = \frac{Z*}{Z}\,. \tag{1.4}$$

$Z*$, the partition function of all RNAs with suitable structure can be computed by dynamic programming [87, 100] or by means of stochastic backtracking and sampling [31, 32].

Using conventional methods to measure RNA protein interactions, only the total concentration of unbound RNA, [RNA], can be measured. Hence, only the apparent dissociation constant $Kd^{\mathrm{app}} = Kd/p*$ can be determined experimentally. The theory therefore predicts structure dependence of the measured values of Kd^{app}. Under the assumption that the true value of Kd depends only on the ligand and the sequence-structure motif that binds the ligand, we can predict sequence-dependent variations in RNA-ligand binding affinity by means of a computational analysis of the ensemble of RNA structures. In [44, 44] it has been shown that the interaction of the HuR protein with ARE-carrying mRNAs indeed follows this scheme.

An immediate consequence of this mechanism is the possibility of using small RNA "modifiers" to modulate the binding affinities of RNAs and ligands by binding to their target RNA in such a way that it alters the local structure at the ligand binding site. The *HuR*-mRNA interaction again serves as a well-studied *in vitro* example for such a technique [44, 89], Fig. 1.4. The regulation of *HuR*-ARE-mediated export and RNA stability *in vivo*, however, remains enigmatic. There is only the single ubiquitously expressed protein *HuR* (and a handful of tissue specific relatives such as the neuronal specific homologue *HuD*) that upregulates the export and stability of potentially thousands of ARE-carrying mRNAs. It is tempting to speculate that modifying RNA "openers" could be involved in target gene-specific regulation of *HuR* activity.

1.5 Conclusion

In this chapter we have discussed at least some of the regulatory mechanism that leave traces at the DNA level. A significant fraction of the non-repetitive DNA of higher eukaryotes is subject to stabilizing selection. It has been estimated, for example, that about 5% of the human genome is under stabilizing selective pressure [24, 59], while less than 2% are protein-coding genes. It is a major challenge for bioinformatics to elucidate the meaning of the remaining conserved DNA.

The information about at least a large part of the regulatory circuitry of a species is accessible by means of comparative genomics. Without a large body of independent experiments, however, we have little chance to decode this information. The first, and maybe crucial step, beyond identifying the DNA footprints themselves is to discriminate between regulatory elements that exert their function at the DNA level, *cis*-acting elements that function at the mRNA level, and non-coding RNAs.

We have reviewed here some of the currently available computational approaches that can be used to detect and analyze such elements. Few general tools are available. A subclass of non-coding RNAs and *cis*-acting mRNA elements, for example, can be recognized because of its conserved secondary structure. On the other hand, at present there is not even a way to distinguish protein binding sites on the genomic DNA from those on the mRNA, unless specific knowledge about a particular sequence motifs is available from experiments.

References

1. L. Agarman, R. Hershberg, J. Vogel, G. Bejerano, E. G. Wagner, H. Margalit, and S. Altuvia. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, 11:941–950, 2001.

2. L. Ancel Meyers, J. F. Lee, M. Cowperthwaite, and A. D. Ellington. The robustness of naturally and artificially selected nucleic acid secondary structures. *J. Mol. Evol.*, 58:681–691, 2004.
3. M. Andronescu, Z. Zhang, and A. Condon. Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.*, 345:987–1001, 2005.
4. T. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, 3:21–29, 1995.
5. T. Bakheet, M. Frevel, B. R. Williams, W. Greer, and K. S. Khabar. ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. *Nucl. Acids Res.*, 29:246–254, 2001.
6. S. Barlati and A. Barbon. RNA editing: a molecular mechanism for the fine modulation of neuronal transmission. *Acta Neurochir Suppl*, 93:53–57, 2005.
7. E. Bayne and R. Allshire. RNA-directed transcriptional gene silencing in mammals. *Trends Genet*, 21:370–373, 2005.
8. P. Bertone, V. Stoc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306:2242–2246, 2004.
9. E. A. Bohula, A. J. Salisbury, M. Sohail, M. P. Playford, J. Riedemann, E. M. Southern, and V. M. Macaulay. The efficacy of small interfering RNAs targeted to the type 1 insulin-like growth factor receptor (IGF1R) is influenced by secondary structure in the IGF1R transcript. *J. Biol. Chem.*, 278:15991–15997, 2003.
10. E. Bonnet, J. Wuyts, P. Rouzé, and Y. Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20:2911–2917, 2004.
11. S. Brantl. Bacterial gene regulation: from transcription attenuation to riboswitches and ribozymes. *Trends Microbiol.*, 12:473–475, 2004.
12. M. Brent and R. Guigó. Recent advances in gene structure prediction. *Curr Opin Struct Biol*, 14:264–272, 2004.
13. C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268:78–94, 1997.
14. J. Butler and J. Kadonaga. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev*, 16:2583–2592, 2002.
15. B. Cairns. Chromatin remodeling complexes: strength in diversity, precision through specialization. *Curr Opin Genet Dev*, 15:185–190, 2005.
16. R. J. Carter, I. Dubchak, and S. R. Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucl. Acids Res.*, 29:3928–3938, 2001.
17. S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 116:499–509, 2004.
18. J. Chen, M. Sun, L. Hurst, G. Carmichael, and J. Rowley. Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. *Trends Genet*, 21:326–329, 2005.

19. J. H. Chen, S. Y. Le, B. Shapiro, K. M. Currey, and J. V. Maizel Jr. A computational procedure for assessing the significance of RNA secondary structure. *Comput. Appl. Biosci.*, 6:7–18, 1990.
20. S. Chen, E. A. Lesnik, T. A. Hall, R. Sampath, R. H. Griffey, D. Eker, and L. Blyn. A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*, 65:157–177, 2002.
21. J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard, and T. R. Gingeras. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308:1149–1154, 2005.
22. J. Claycomb and T. Orr-Weaver. Developmental gene amplification: insights into DNA replication and gene expression. *Trends Genet*, 21:149–162, 2005.
23. P. Clote, F. Ferré, E. Kranakis, and D. Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11:578–591, 2005.
24. G. M. Cooper, M. Brudno, E. A. Stone, I. Dubchak, S. Batzoglou, and A. Sidow. Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res*, 14:539–48, 2004.
25. F. F. Costa. Non-coding RNAs: New players in eucaryotic biology. *Gene*, 357:83–94, 2005.
26. J. Dalgaard and S. Vengrova. Selective gene expression in multigene families from yeast to mammals. *Sci STKE*, 2004(256), 2004.
27. E. Davidson. *Genomic Regulatory Systems*. Academic Press, San Diego, 2001.
28. W. Deng, X. Zhu, G. Skogerbø, Y. Zhao, Z. Fu, Y. Wang, L. He, Housheng Cai, H. Sun, C. Liu, B. L. Li, B. Bai, J. Wang, Y. Cui, D. Jai, Y. Wang, D. Du, and R. Chen. Organisation of the *Caenorhabditis elegans* small noncoding transcriptome: genomic features, biogenesis and expression. *Genome Res.*, 16:20–29 2006.
29. E. T. Dermitzakis, A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier, and S. E. Antonarakis. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*, 302:1033–1035, 2003.
30. R. A. Dimitrov and M. Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, 87:215–226, 2004.
31. Y. Ding, C. Y. Chan, and C. E. Lawrence. Sfold web server for statistical folding and rational design of nucleic acids. *Nucl. Acids Res.*, 32(Web Server issue):W135–141, 2004.
32. Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucl. Acids Res.*, 31:7280–7301, 2003.
33. X. Feng, W. Lin, D. Minghua, S. Fengzhu, and Q. Minping. An efficient algorithm for deciphering regulatory motifs. In These Proceedings, 2005.
34. J. Fickett. Finding genes by computer: the state of the art. *Trends Genet*, 12: 316–320, 1996.
35. C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. Design of multi-stable RNA molecules. *RNA*, 7:254–265, 2000.
36. D. Galas and A. Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucl. Acids Res.*, 5:3157–3170, 1978.
37. M. Garner and A. Revzin. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the

Escherichia coli lactose operon regulatory system. *Nucl. Acids Res.*, 9:3047–3060, 1981.
38. F. Gebauer and M. Hentze. Molecular mechanisms of translational control. *Nat Rev Mol Cell Biol*, 5:827–835, 2004.
39. M. Gelfand. Prediction of function in DNA sequence analysis. *J Comput Biol*, 2:87–115, 1995.
40. M. Gelfand, A. Mironov, and P. Pevzner. Gene recognition via spliced sequence alignment. *Proc Natl Acad Sci U S A*, 93:9061–9066, 1996.
41. G. Gonsalvez, C. Urbinati, and R. Long. RNA localization in yeast: moving towards a mechanism. *Biol Cell*, 97:75–86, 2005.
42. S. Gottesman. The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.*, 58:303–328, 2004.
43. R. Guigó, E. Dermitzakis, P. Agarwal, C. Ponting, G. Parra, A. Reymond, J. Abril, E. Keibler, R. Lyle, C. Ucla, S. Antonarakis, and M. Brent. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc Natl Acad Sci U S A*, 100:1140–1145, 2003.
44. J. Hackermüller, N.-C. Meisner, M. Auer, M. Jaritz, and P. F. Stadler. The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: A quantitative model. *Gene*, 345:3–12, 2005.
45. T. Heinemeyer, E. Wingender, I. Reuter, H. Hermjakob, A. E. Kel, O. V. Kel, E. V. Ignatieva, E. A. Ananko, O. A. Podkolodnaya, F. A. Kolpakov, N. L. Podkolodny, and N. A. Kolchanov. Databases on transcriptional regulation: TRANSFAC, TRRD, and COMPEL. *Nucl. Acids Res.*, 26:364–370, 1998.
46. T. M. Henkin and C. Yanofsky. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcribtion termination/antitermination decision. *BioEssays*, 24:700–707, 2002.
47. R. Hershberg, S. Altuvia, and H. Margalit. A survey of small RNA-encoding genes in *Escherichia coli*. *Nucl. Acids Res.*, 31:1813–1820, 2003.
48. G. Hertz and G. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.
49. N. Hodas and D. Aalberts. Efficient computation of optimal oligo-RNA binding. *Nucl. Acids Res.*, 32:6636–6642, 2004.
50. I. L. Hofacker. Vienna RNA secondary structure server. *Nucl. Acids Res.*, 31:3429–3431, 2003.
51. I. L. Hofacker, M. Fekete, C. Flamm, M. A. Huynen, S. Rauscher, P. E. Stolorz, and P. F. Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.*, 26:3825–3836, 1998.
52. I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.
53. I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.
54. P. Hooper, H. Zhang, and D. Wishart. Prediction of genetic structure in eukaryotic DNA using reference point logistic regression and sequence alignment. *Bioinformatics*, 16:425–438, 2000.
55. I. Huez, L. Créancier, S. Audigier, M. Gensac, A. Prats, and H. Prats. Two independent internal ribosome entry sites are involved in translation initiation of vascular endothelial growth factor mRNA. *Mol. Cell. Biol.*, 18:6178–6190, 1998.

56. J. Hughes, P. Estep, S. Tavazoie, and G. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296:1205–1214, 2000.
57. A. Hüttenhofer, P. Schattner, and N. Polacek. Non-coding RNAs: hope or hype? *Trends Genet.*, 21:289-97 2005.
58. T. Imanishi and *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biology*, 2:0856–0875, 2004.
59. International Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.
60. G. H. Jacobs, O. Rackham, P. A. Stockwell, W. Tate, and C. M. Brown. Transterm: a database of mRNAs and translational control elements. *Nucl. Acids Res.*, 30:310–311, 2002.
61. J. M. Johnson, S. Edwards, D. Shoemaker, and E. E. Schadt. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.*, 21:93–102, 2005.
62. H. Jow, C. Hudelot, M. Rattray, and P. G. Higgs. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Mol. Biol. Evol.*, 19:1591–1601, 2002.
63. D. Kampa, J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana, and T. R. Gingeras. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.*, 14:331–342, 2004.
64. M. Kawano, A. A. Reynolds, J. Miranda-Rios, and G. Storz. Detection of 5'- and 3'-UTR-derived small RNAs and *cis*-encoded antisense RNAs in *Escherichia coli*. *Nucl. Acids Res.*, 33:1040–1050, 2005.
65. O. Kel-Margoulis, T. Ivanova, E. Wingender, and A. Kel. Automatic annotation of genomic regulatory sequences by searching for composite clusters. In *Proc. Pac. Symp. Biocomput.*, pages 187–198, 2002.
66. A. Khvorova, A. Reynolds, and S. D. Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115:209–16, 2003.
67. I. Korf, P. Flicek, D. Duan, and M. Brent. Integrating genomic homology into gene structure prediction. *Bioinformatics*, 17 Suppl 1:140–148, 2001.
68. A. Kornblihtt, M. de la Mata, J. Fededa, M. Munoz, and G. Nogues. Multiple links between transcription and splicing. *RNA*, 10:1489–1498, 2004.
69. M. Kozak. Alternative ways to think about mRNA sequences and proteins that appear to promote internal initiation of translation. *Gene*, 318:1–23, 2003.
70. R. Kretschmer-Kazemi Far and G. Sczakiel. The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides. *Nucl. Acids Res.*, 31:4417–4424, 2003.
71. A. Krogh. Gene finding: putting the parts together. In M. Bishop, editor, *Guide to human genome computing*, pages 261–274. Academic Press, 2nd edition, 1998.
72. D. Kulp, D. Haussler, M. Reese, and F. Eeckman. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol*, 4:134–142, 1996.
73. L. Lareau, R. Green, R. Bhatnagar, and S. Brenner. The evolving roles of alternative splicing. *Curr Opin Struct Biol*, 14:273–282, 2004.
74. C. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, and J. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

75. C. Lawrence and A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7:41–51, 1990.
76. S. V. Le, J. H. Chen, K. M. Currey, and J. V. Maizel Jr. A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.*, 4:153–159, 1988.
77. D. LeRoith and C. Roberts. The insulin-like growth factor system and cancer. *Cancer Lett*, 195:127–137, 2003.
78. X. Liang, A. Haritan, S. Uliel, and S. Michaeli. *trans* and *cis* splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot Cell*, 2:830–840, 2003.
79. J. Livny, M. A. Fogel, B. M. Davis, and M. K. Waldor. `sRNAPredict`: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucl. Acids Res.*, 33:4096–4105, 2005.
80. A. Mallory and H. Vaucheret. MicroRNAs: something important between the genes. *Curr Opin Plant Biol*, 7:120–125, 2004.
81. R. Margueron, P. Trojer, and D. Reinberg. The key to development: interpreting the histone code? *Curr Opin Genet Dev*, 15:163–176, 2005.
82. E. H. Margulies, M. Blanchette, D. Haussler, and E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Res.*, 13:2507–2518, 2003.
83. D. H. Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10:1178–1190, 2004.
84. D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.
85. J. S. Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25:930–939, 2003.
86. J. S. Mattick. RNA regulation: a new genetics? *Nature Genetics*, 5:316–323, 2004.
87. J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.
88. J. P. McCutcheon and S. R. Eddy. Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucl. Acids Res.*, 31:4119–4128, 2003.
89. N.-C. Meisner, J. Hackermüller, V. Uhl, A. Aszódi, M. Jaritz, and M. Auer. mRNA openers and closers: A methodology to modulate AU-rich element controlled mRNA stability by a molecular switch in mRNA conformation. *Chembiochem.*, 5:1432–1447, 2004.
90. E. Merino and C. Yanofsky. Regulation by termination-antitermination: a genomic approach. In A. L. Sonenshein, J. A. Hoch, and R. Losick, editors, Bacillus subtilis *and its closest relatives: From Genes to Cells*, pages 323–336. ASM Press, Washington D.C., 2002.
91. A. C. Messias and M. Sattler. Structural basis of single-stranded RNA recognition. *Acc. Chem. Res.*, 37:279–287, 2004.
92. F. Mignone, C. Gissi, S. Liuni, and G. Pesole. Untranslated regions of mRNAs. *Genome Biology*, 3, 2002. reviews0004.1-0004.10.
93. K. Missal, D. Rose, and P. F. Stadler. Non-coding RNAs in *Ciona intestinalis*. *Bioinformatics*, 21(S2):i77-i78, 2005.

94. K. Missal, X. Zhu, D. Rose, W. Deng, G. Skogerbø, R. Chen, and P. F.Stadler. Prediction of structured non-coding RNAs in the genome of the nematode *Caenorhabitis elegans*. *J. Exp. Zool.B: Mol. Dev. Evol.*, 2006. DOI: 10.1002/jez.b.21086
95. V. Mittal. Improving the efficiency of RNA interference in mammals. *Nat. Rev. Genet.*, 5:355–365, 2004.
96. H. Morgan, F. Santos, K. Green, W. Dean, and W. Reik. Epigenetic reprogramming in mammals. *Hum Mol Genet*, 14(Spec No 1):47–58, 2005.
97. B. Morgenstern. `DIALIGN 2`: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218, 1999.
98. A. Morillon, J. O'Sullivan, A. Azad, N. Proudfoot, and J. Mellor. Regulation of elongating RNA polymerase II by forkhead transcription factors in yeast. *Science*, 300(5618):492–495, 2003.
99. A. Mosig, T. Bıyıkoğlu, S. J. Prohaska, and P. F. Stadler. Discovering cis regulatory modules by optimizing barbeques. *Theor. Comput. Sci.*, 2005. submitted.
100. U. Mückstein, H. Tafer, J. Hackermüller, S. B. Bernhard, P. F. Stadler, and I. L. Hofacker. Thermodynamics of RNA-RNA binding. In *German Conference on Bioinformatics 2005*, A. Torda, S. Kurtz, M. Rarey (eds.), Lecture Notes in Informatics P-71, Gesellschaft f. Informatik, Bonn, pp.3–13, 2005.
101. E. Nudler and A. S. Mironov. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, 29:11–17, 2004.
102. Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, I. Nikaido, N. Osato, R. Saito, H. Suzuki, and *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420:563–573, 2002.
103. M. J. O'Neill. The influence of non-coding RNAs on allele-specific gene expression in mammals. *Hum Mol Genet.*, 14:R113–120, 2005.
104. M. Overhoff, M. Alken, R. K. Far, M. Lemaitre, B. Lebleu, G. Sczakiel, and I. Robbins. Local RNA target structure influences siRNA efficacy: A systematic global analysis. *J. Mol. Biol.*, 348:871–881, 2005.
105. P. Pan, M. Lieber, and J. Teale. The role of recombination signal sequences in the preferential joining by deletion in DH-JH recombination and in the ordered rearrangement of the IgH locus. *Int Immunol*, 9:515–522, 1997.
106. J. S. Parker, S. M. Roe, and D. Barford. Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, 434:663–666, 2005.
107. J. Parsch, J. M. Braverman, and W. Stephan. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics*, 154:909–921, 2000.
108. G. Pesole, F. Mignone, C. Gissi, G. Grillo, F. Licciulli, and S. Liuni. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*, 276:73–81, 2001.
109. A. Philippakis, F. He, and M. Bulyk. Modulefinder: a tool for computational discovery of cis regulatory modules. In *Proc. Pac. Symp. Biocomput.*, pages 519–30, 2005.
110. S. J. Prohaska, C. Fried, C. T. Amemiya, F. H. Ruddle, G. P. Wagner, and P. F. Stadler. The shark HoxN cluster is homologous to the human HoxD cluster. *J. Mol. Evol.*, 58:212–217, 2004.

111. S. J. Prohaska, C. Fried, C. Flamm, G. Wagner, and P. F. Stadler. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. *Mol. Phyl. Evol.*, 31:581–604, 2004.
112. R. Pudimat, E. Schukat-Talamazzini, and R. Backofen. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics*, 21: 3082–3088, 2005.
113. M. Reese, G. Hartzell, N. Harris, U. Ohler, J. Abril, and S. Lewis. Genome annotation assessment in Drosophila melanogaster. *Genome Res*, 10:483–501, 2000.
114. M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10:1507–17, 2004.
115. E. Rivas and S. R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16:583–605, 2000.
116. E. Rivas, R. J. Klein, T. A. Jones, and S. R. Eddy. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.*, 11:1369–1373, 2001.
117. M. Rodriguez, C. Dargemont, and F. Stutz. Nuclear export of RNA. *Biol Cell*, 96:639–655, 2004.
118. S. Rogic, A. Mackworth, and F. Ouellette. Evaluation of gene-finding programs on mammalian sequences. *Genome Res*, 11:817–832, 2001.
119. F. Roth, J. Hughes, P. Estep, and G. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*, 16:939–945, 1998.
120. S. Rousseaux, C. Caron, J. Govin, C. Lestrat, A. Faure, and S. Khochbin. Establishment of male-specific epigenetic information. *Gene*, 345:139–153, 2005.
121. R. R. Rueckert. Picornaviridae: The viruses and their replication. In N. Fields, D. Knipe, and P. Howley, editors, *Virology*, volume 1, pages 609–654. Lippincott-Raven Publishers, Philadelphia, New York, third edition, 1996.
122. P. Sætrom, R. Sneve, K. I. Kristiansen, O. Snøve jr., T. Grünfeld, T. Rognes, and E. Seeberg. Predicting non-coding RNA genes in *Escherichia coli* with boosted genetic programming. *Nucl. Acids Res.*, 33:3263–3270, 2005.
123. A. Sandelin, W. A. Pär Engström, W. Wasserman, and B. Lenhard. Jaspar: an open access database for eukaryotic transcription factor binding profiles. *Nucl. Acids Res.*, 32:D91–D94, 2004.
124. S. Schubert, A. Grunweller, V. Erdmann, and J. Kurreck. Local RNA target structure influences siRNA efficacy: Systematic analysis of intentionally designed binding regions. *J. Mol. Biol.*, 348(4):883–93, 2005.
125. S. Schwartz, Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, , and W. Miller. `PipMaker` — a web server for aligning two genomic DNA sequences. *Genome Research*, 4:577–586, 2000.
126. D. Schwarz, G. Hutvagner, T. Du, Z. Xu, N. Aronin, and P. Zamore. Asymmetry in the assembly of the RNAi enzyme complex. *Cell.*, 115:99–208, 2003.
127. S. Shabalina, A. Ogurtsov, V. Kondrashov, and A. Kondrashov. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet*, 17:373–376, 2001.
128. S. A. Shabalina and A. S. Kondrashov. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet Res*, 74:23–30, 1999.

129. R. Sharan, I. Ovcharenko, A. Ben-Hur, and R. M. Karp. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. In *ISMB (Supplement of Bioinformatics)*, pages 283–291, 2003.
130. S. Sinha, E. van Nimwegen, and E. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19, 2003.
131. G. Stamatoyannopoulos. Control of globin gene expression during development and erythroid differentiation. *Exp Hematol*, 33:259–271, 2005.
132. L. Stein. Human genome: end of the beginning. *Nature*, 431:915–916, 2004.
133. F. Sterky and J. Lundeberg. Sequence analysis of genes and genomes. *J Biotechnol*, 76:1–31, 2000.
134. N. Sudarsan, J. E. Barrick, and R. R. Breaker. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, 9:644–647, 2003.
135. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, 282:2012–2018, 1998.
136. D. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Annu. Rev. Biophys. Biophys. Chem.*, 17:167–92, 1988.
137. E. van Nimwegen, J. P. Crutchfield, and M. A. Huynen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, 96:9716–9720, 1999.
138. G. Varani. RNA-protein intermolecular recognition. *Acc. Chem. Res.*, 30:189–195, 1997.
139. A. G. Vitreschak, D. A. Rodionov, A. A. Mironov, and M. S. Gelfand. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Gen.*, 20:44–50, 2004.
140. A. Wagner and P. F. Stadler. Viral RNA and evolved mutational robustness. *J. Exp. Zool. (Mol. Dev. Evol.)*, 285:119–127, 1999.
141. S. Washietl and I. L. Hofacker. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, 342:19–30, 2004.
142. S. Washietl, I. L. Hofacker, M. Lukasser, A. Hüttenhofer, and P. F. Stadler. Genome-wide mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in human. *Nature Biotech.*, 23:1383-1390, 2005.
143. S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, 102:2454–2459, 2005.
144. J. Weischenfeldt, J. Lykke-Andersen, and B. Porse. Messenger RNA surveillance: neutralizing natural nonsense. *Curr Biol*, 15:559–562, 2005.
145. A. West and P. Fraser. Remote control of gene transcription. *Hum Mol Genet*, 14 Spec No 1:101–111, 2005.
146. A. West, M. Gaszner, and G. Felsenfeld. Insulators: many functions, many mechanisms. *Genes Dev*, 16:271–288, 2002.
147. C. Wilusz and J. Wilusz. Bringing the role of mRNA decay in the control of gene expression into focus. *Trends Genet*, 20:491–497, 2004.
148. G. Wray, M. Hahn, E. Abouheif, J. Balhoff, M. Pizer, M. Rockman, and L. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20:1377–1419, 2003.
149. S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.
150. R. Yeh, L. Lim, and C. Burge. Computational inference of homologous gene structures in the human genome. *Genome Res*, 11:803–816, 2001.

151. K. Yoshinari, M. Miyagishi, and T. K. Effects on RNAi of the tight structure, sequence and position of the targeted region. *Nucl. Acids Res.*, 32:691–9, 2004.
152. C. H. Yuh, H. Bolouri, and E. H. Davidson. Genomic cis-regulatory logic: Experimental and computational analysis of a sea urchin gene. *Science*, 279:1896–1902, 1998.
153. M. Zuker. Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, 10:303–310, 2000.
154. M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.*, 31:3406–15, 2003.
155. M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, 9:133–148, 1981.