Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Motif models for RNA-binding proteins

Alexander Sasse[1], Kaitlin U Laverty[1], Timothy R Hughes[1,2,3] and Quaid D Morris[1,2,4]

Identifying the binding preferences of RNA-binding proteins (RBPs) is important in understanding their contribution to post-transcriptional regulation. Here, we review the current state-of-the art of RNA motif identification tools for RBPs. New *in vivo* and *in vitro* data sets provide sufficient statistical power to enable detection of relatively long and complex sequence and sequence-structure binding preferences, and recent computational methods are geared towards quantitative identification of these patterns. We classify methods by their motif model's representational power and describe the underlying considerations for RNA-protein interactions. All classical motif identification algorithms apply physically motivated architectures, consisting of a motif and an occupancy model, we call these explicit motif models. Recent methods, such as convolutional neural networks and support vector machines, abandon the classical architecture and implicitly model RNA binding without defining a motif model. Although they achieve high accuracy on held-out data they may be unsuitable to solve the ultimate goal of the field, using motifs trained on *in vitro* data to predict *in vivo* binding sites. For this task methods need to separate intrinsic binding preferences from cellular effects from protein and RNA concentrations, cooperativity, and competition. To tackle this problem, we advocate for the use of a 'three-layer' architecture, consisting of motif model, occupancy model, and extrinsic factor model, which enables separation and adjustment to cellular conditions.

Addresses
[1] Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada
[2] Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada
[3] Canadian Institute for Advanced Research, MaRS Centre, West Tower, 661 University Avenue, Suite 505, Toronto, ON M5G 1M1, Canada
[4] Department of Computer Science, University of Toronto, Toronto, ON M5T 3A1, Canada

Current Opinion in Structural Biology 2018, **53**:115–123

This review comes from a themed issue on **Protein–nucleic acid interactions**

Edited by **Eric Westhof** and **Dinshaw Patel**

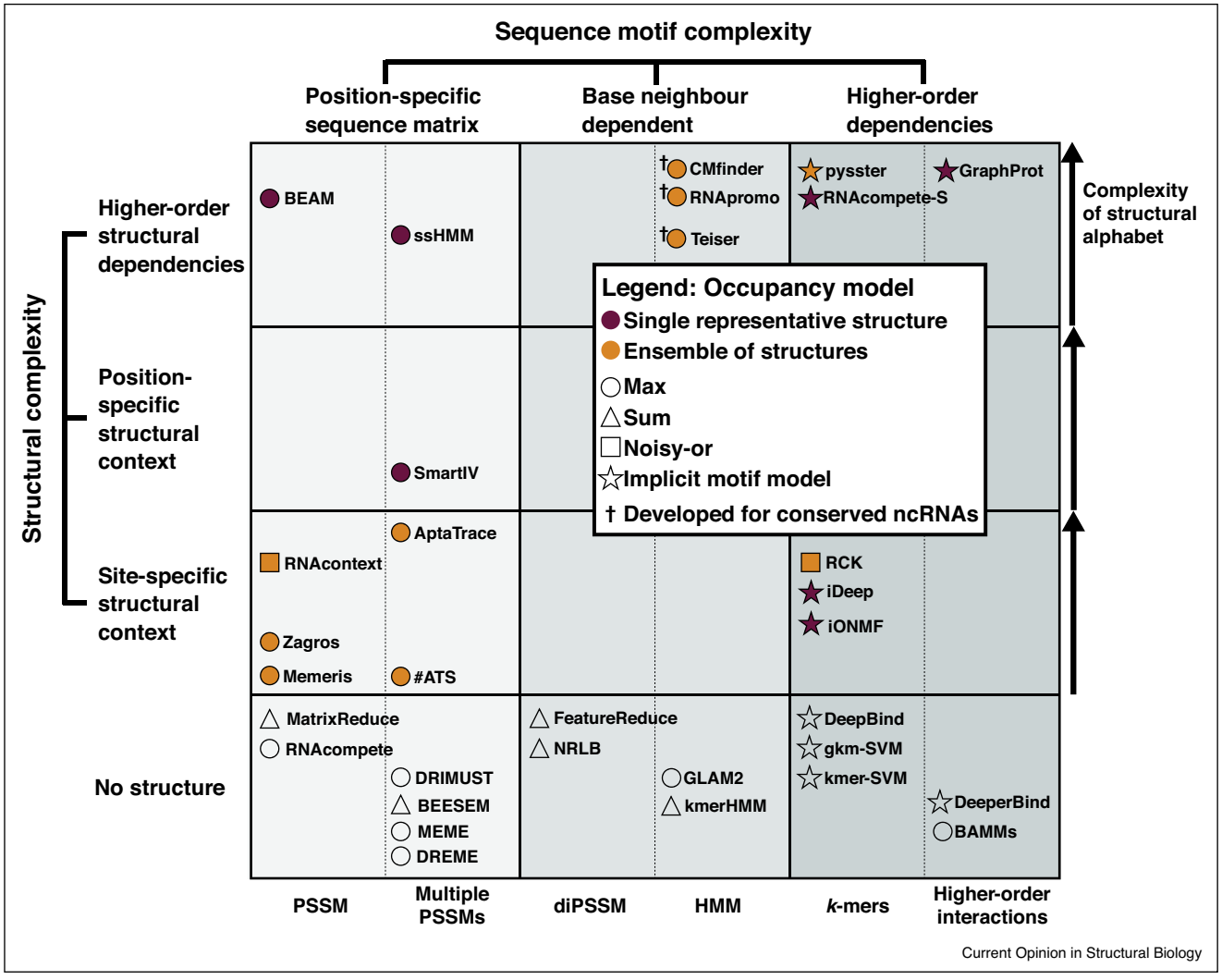
https://doi.org/10.1016/j.sbi.2018.08.001

Introduction

RNA-binding proteins (RBPs) play myriad roles in co-transcriptional and post-transcriptional processes, influencing mRNA biogenesis, modification, stability, transport, and cellular localization [1]. Many RBPs recognize specific RNA sequences or structural patterns, computational models of these, called motifs, are typically derived from *in vitro* selection assays like RNAcompete [2••,3], RNA Bind-n-Seq [4•], or systematic evolution of ligands by exponential enrichment (SELEX) [5,6]. Sequences matching these motif models are enriched in binding sites in living cells ('*in vivo*'), identified by cross-linking and immunoprecipitation (CLIP) [7••]. But, these motifs are not perfect predictors of *in vivo* binding. Noise in experimental measurements may contribute partially, but differences are to be expected due to effects of the cellular environment, such as cooperation and competition with other RBPs, differences in abundance among cellular RNAs, localization of proteins and RNA, and large-scale RNA folding [8].

Here, we survey motif identification methods for RBPs and categorize them based on their modelling choices. Although most well-characterized RBPs recognize a primary sequence motif, analogous to a transcription factor binding site, some RBPs show a clear preference for folded RNA structures [9••,10••] and multi-partite sites [11••,12]. Mirroring this diversity, motif models vary considerably in which binding site features they consider. More complex motif models are necessary to capture some binding preferences. In our survey, we begin with motifs which consider only primary sequence, and then examine inclusion of RNA secondary structure. Classic motif models cannot be fit using binding data without specifying a *sequence occupancy* model that maps between their motif model and the probability of binding a given sequence. We also discuss the recent development of computational methods that model binding directly using, for example, convolutional neural networks (CNNs). These methods do not include an explicit motif model. For reference, Figure 1 places all of the methods on a grid that indicates the primary sequence model, the level of incorporation of RNA structure, and the occupancy model employed, illustrating that they vary considerably in what they represent. To provide context for the various levels of motif complexity, Figure 2 illustrates the binding preferences of a panel of well-studied proteins, spanning in complexity of the recognized site from those that bind single stranded RNA (ssRNA) (Figure 2A–C) to those that bind specific RNA structures (Figure 2D–G). Here, we do not attempt to evaluate the

Figure 1



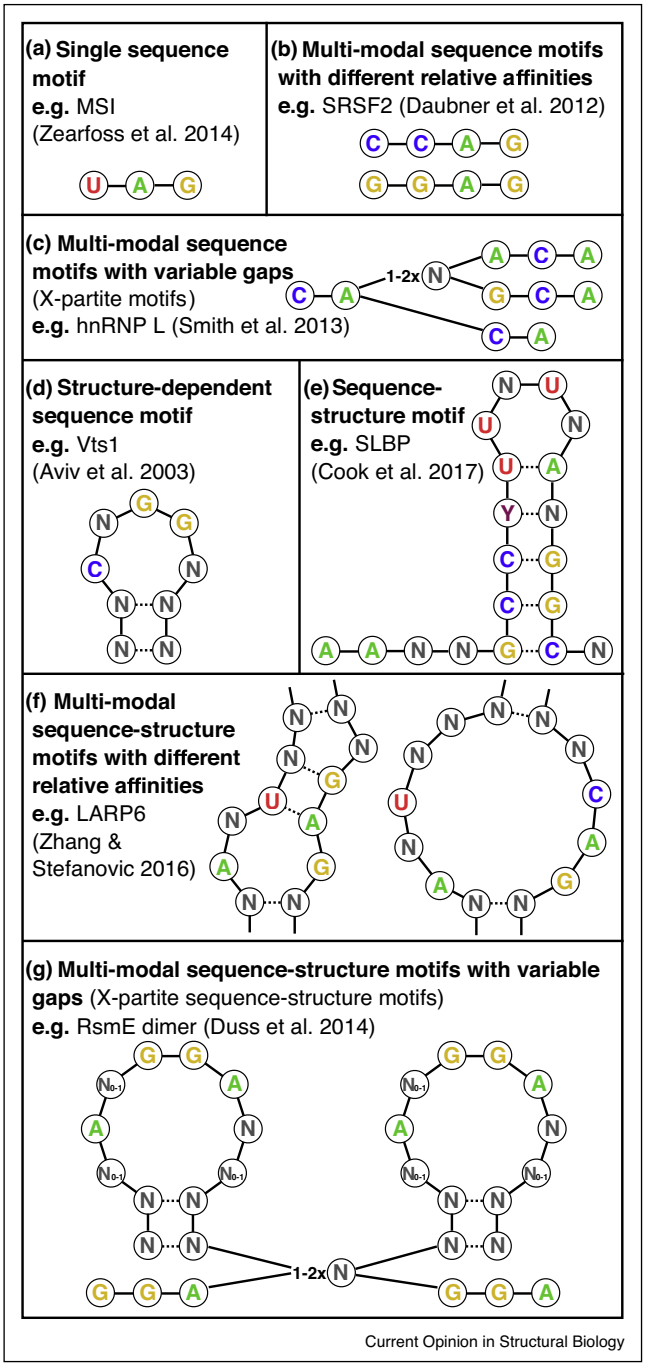
Classification of reviewed methods based on their sequence motif model, extent of structure incorporation, and occupancy model. Sequence motif models are divided into six different model types represented on the *X*-axis. RNA secondary structure is incorporated into the sequence motif models in three forms shown on the *Y*-axis. For each of the structure inclusions, a simple alphabet defining single or double strandedness, or an alphabet covering a wider range of secondary structures can be applied, roughly indicated by arrows on the right. Explicit motif identification methods make use of 3 types of occupancy models, represented by the shape to the left of the method name. Implicit motif models possess their own shape. RNA secondary structure is considered either as a single structure, or an ensemble of potential structural conformations, indicated by the color fill of the shape.

quality or accuracy of motif models, instead we seek to understand differences between algorithms and provide a framework for further assessment.

Primary sequence motif models

Primary sequence motif models for RBPs can be divided into six categories, shown on the horizontal axis in Figure 1. Position-specific scoring matrices (*PSSMs*), originally developed for transcription factors [13], are the least complex, and most widely used. PSSMs represent a single, preferred 'consensus' sequence, and model decreases in binding preference caused by single-base changes, assuming multiple base changes are independent (MatrixReduce, RNAcompete) [14,15]. Interconvertible versions of the PSSMs include position weight matrices (PWMs), position frequencies matrices (PFMs), and position-specific affinity matrix (PSAM) [16,17]. Some models use *multiple PSSMs* to capture distinct consensus sequences (MEME, DREME, DRIMUST, BEESEM) [18,19,20••,21]. PSSMs are sufficiently accurate models for many RBPs [13,22], but they lack the ability to detect dependencies between nucleotide positions and to differentiate between multi-modal or variably gapped binding specificities [23,24].

Figure 2



RNA motifs recognized by RBPs. **(a) Single sequence motif:** The protein binds to only one specific RNA sequence motif. **(b) Multi-modal sequence motifs with different relative affinities:** The protein binds to more than one specific sequence motif, each with a different binding affinity. **(c) Multi-modal sequence motifs with variable gaps:** The protein possesses multiple domains, two or more of which display a preference for a specific motif or motifs. The positions of these short motifs within a larger motif can vary due to flexible linkers between binding domains, creating variable gaps in the motif. **(d) Structure-dependent sequence motif:** The change in free energy upon binding is larger if the preferred sequence motif is located in a specific structural context. **(e) Sequence-structure motif:** The protein prefers to bind to an interdependent sequence and structural motif, here base-resolution structure is necessary to describe the motif. **(f) Multi-modal sequence-structure motifs with different relative affinities:** The protein binds to more than one sequence-structure specific motif, each with a different binding affinity. **(g) Multi-modal sequence-structure motifs with variable gaps:** Proteins possessing multiple RBDs or proteins forming homodimers and heterodimers are able to bind short sequence-structure motifs within a larger motif, these short motifs need to be within a certain 3-D distance. Binding affinity is influenced by the short motifs that the single domain/protein recognizes as well as the distance of these motifs to each other.

Models that account for nucleotide dependencies include *diPSSMs* and Hidden Markov Models (*HMMs*). The use of dinucleotide frequencies at each position in diPSSMs can model dependencies between neighboring bases (FeatureReduce, NRLB) [25,26]. HMMs also permit neighboring nucleotide dependencies, as well as multi-modal binding and variable gaps between binding sites (kmerHMM, GLAM2) [24,27]. A Bayesian Markov Model (BAMM) extends diPSSMs and HMMs by capturing tri-nucleotide or *higher-order interactions*, given sufficient data [23]. The most expressive models are those that assign different free energies to all RNA oligos up to a given length k (aka *k-mers*) (RCK, kmer-SVM) [28,29]. The k-mer models have an unwieldy number of parameters for larger values of k [30•,31]. The number of parameters can be reduced through the use of gapped k-mer models which permit representation of larger, fixed-sized binding sites, or through kernels used with max-pooling in convolutional neural networks (CNNs), which perform k-mer selection during training (gkm-SVM, DeepBind) [32,33,34••,35]. Recurrent or multiple convolutional layers enable modelling of even higher-order dependencies between fitted kernel activations (DeeperBind) [36].

Learning and representing RNA secondary structure

At least 30% of sequence-specific RBPs display a preference for a given structure or structural context for their binding site [2••,10••,28]. However, for many RBPs it remains unclear the extent to which RNA secondary structure is necessary for binding and the level of complexity of recognized structural features. Indeed, even in well-characterized genomes, most RBPs still lack primary sequence motifs, leaving open the possibility that they may bind specific structures. Figure 1 distinguishes three categories of complexity of the secondary structure parameters in RBP motif models: an overall structure context preference for a binding site (*site-specific structural context*), a position-specific structure context preference (*position-specific structural context*), or a higher-order, structure or sequence dependent, position specific structure preferences (*higher-order structural dependencies*).

Motif models like Zagros, Memeris, #ATS [9••,20••,37] include a site-specific structural context preference to

capture the overall preference for binding ssRNA displayed by some RBPs, like MSI, SRSF2 or hnRNP L (Figure 2A–C) or those proteins, like Vts1p (Figure 2D) which only bind their target site when it is in a hairpin loop. Other RBPs, like SNRPA, bind their target site both in internal loops and hairpin loops and, with lower affinity, in RNA external to any loop. Capturing this preference requires models able to assign different relative preferences to different contexts (e.g. AptaTrace, RNAcontext, RCK, iDeep) [28,38–40].

Models using position-specific structural context represent structural preferences with a PSSM-like model which assigns a relative structural context preference to each position in the binding site. These models can adequately capture the preferences of SLBP which recognizes the base identity of nucleotides both in a hairpin loop and a double stranded region [30•] (SLBP, Figure 2E). While some methods, like RNAcompete-S or Graphprot, use this to derive an approximation of their implicit motif model, the only method using it explicitly, SmartIV [41], currently only distinguishes between a paired and unpaired structural preference, and would not fully capture SLBP preferences.

Some motif models represent higher-order structural dependencies between neighboring positions or larger distances using either HMMs, sequence-structure *k*-mers, covariance models or multi-nucleotide frequencies (BEAM, ssHMM, RNAcompete-S, pysster) [30•,35,42,43]. For instance, a graph kernel models structural dependencies between base pairs on distant stretches of the sequence (Graphprot) [44•]. Instead of solely generating *k*-mers in sequential order, the graph kernel extracts short patterns from structurally connected nucleotides. Covariance models included in CMfinder, RNApromo and Teiser [45,46], use stochastic context-free grammars (SCFGs). They incorporate both distant and close sequence and structure dependencies enabling the modelling of multi-modal sequence and structure preferences (see, e.g. Figure 2F,G). [47] These complex preferences can result from the variable usage of multiple binding domains or the formation of heterodimers and homo-dimers leading to the recognition of variably gapped motifs; for example, the protein RsmE homodimerizes to bind to a bipartite motif that can vary in structure (Figure 2G) [47].

Fitting motif models

The data used to fit motif models generally consist of sets of short RNA sequences classified as bound and unbound or assigned a score reflecting experimentally-measured occupancy. Motif models can assign a score to each potential binding site in an RNA sequence. Often this score corresponds to an estimate of the change in free energy associated with binding the site. Motif model thus must be augmented with a *sequence occupancy* model to map from these free energy scores to the probability of binding an RNA sequence. In contrast, implicit motif models directly predict binding of an RNA sequence without an explicit motif representation; these representations must be inferred after the model is fit and are often approximations to the implied RBP binding preference.

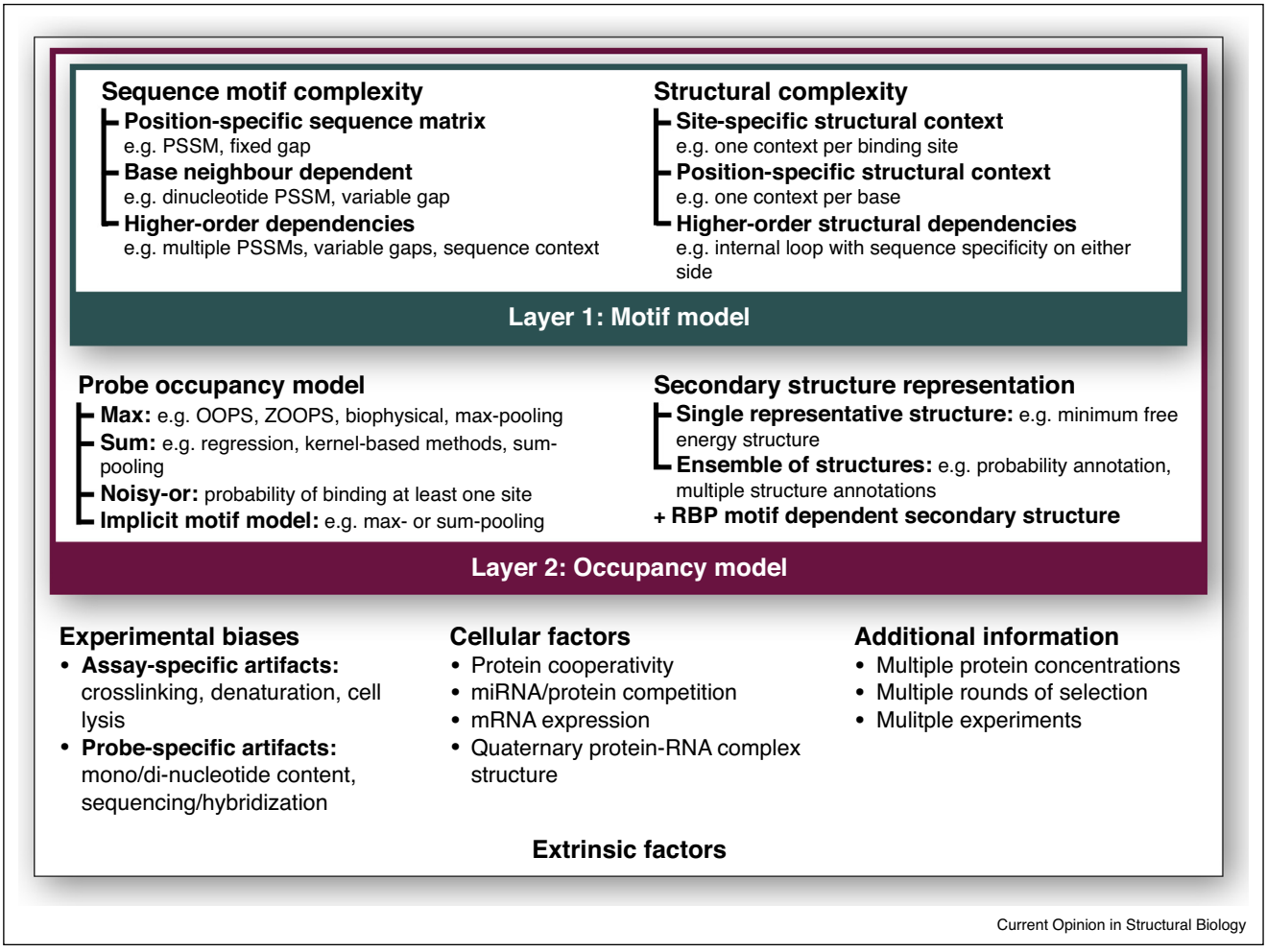
Implicit motif models

Implicit motif models based on non-linear machine learning methods, such as convolutional neural networks and Support Vector Machines, have recently displayed impressive performances predicting *in vitro* and *in vivo* [29,32,34••] binding. These methods can capture dependencies between individual nucleotides as well as non-additive interactions between two or more binding sites. However, in non-linear approaches, it is rarely possible to isolate an explicit motif model from those parameters that also capture cellular or experiment influences on binding. In CNNs, for example, visualized convolutional kernels resemble known RBP motifs and are interpreted as PSSMs. However, considering only the weights in the convolutional kernels can overlook linear or non-linear dependencies between kernels that are needed to characterize the binding preference of an RBP [48]. As such, it can thus be difficult to determine whether the resulting predictive performance is due to more accurately capturing the RBP binding preference, or to capture cellular conditions or experimental biases [49]. Because there is no explicit separation, good performance on held out data from one assay is no guarantee of good generalization to different ones.

Fitting explicit motif models

To fit explicit motif models, one must further specify how the motif scores of potential binding sites within an RNA sequence translate into a probability that the whole sequence is bound. We call this the *sequence occupancy model* (Figure 3). There are three main methods, *max*, *sum*, and *noisy-or*, to translate from the binding site occupancy into occupancy of the entire RNA sequence (see Table S1 for a mathematical definition of each). The max method sets RNA sequence occupancy to be the maximum motif score of its constituent binding sites. The max assumption is used by algorithms performing alignments of multiple sequences by expectation maximization (EM) (aka OOPS) [20••,24]. The sum method sets sequence occupancy to be the expected number of bound RBPs by adding all motif scores, assuming no steric or cooperative interactions. The sum model is generally used in biophysically motivated motif finding methods but also some implicit motif models which use linear or logistic regression and Support Vector Machines with linear kernels [29,30•,50]. The noisy-or model represents a middle ground between the max and the sum model but is rarely used. It uses the motif score to compute the probability that at least one site is bound, again assuming no steric or cooperative interactions [28,38]. Noisy-or and

Figure 3

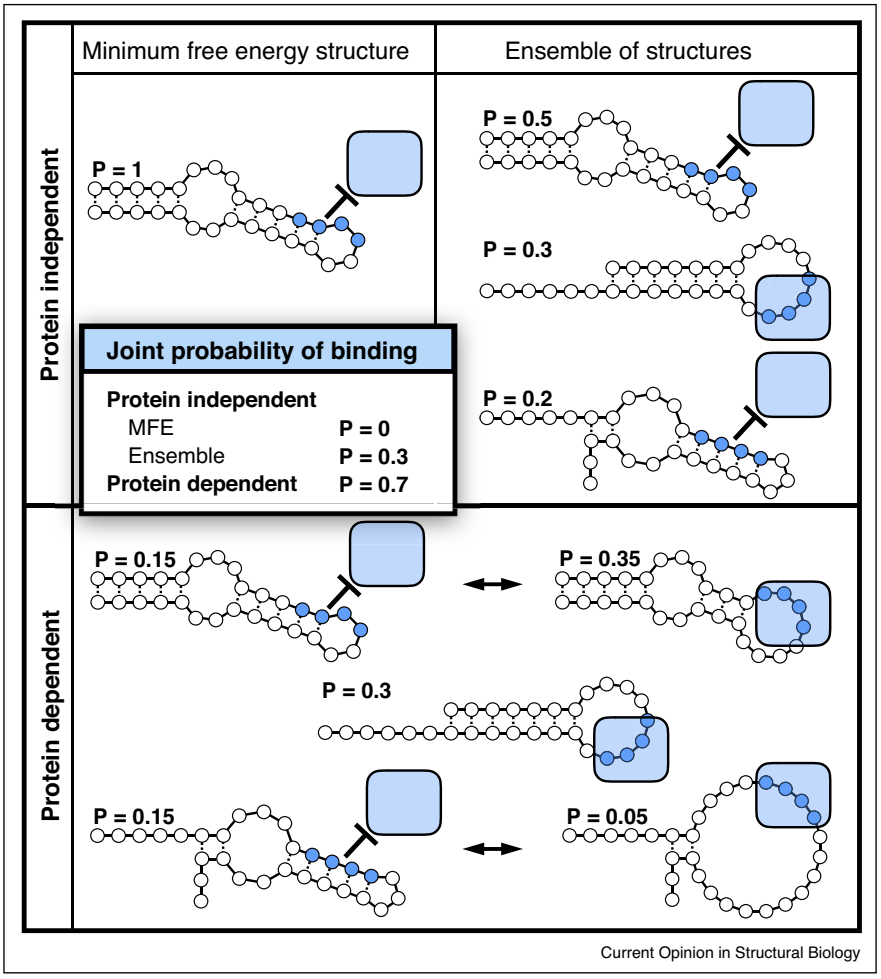


Schematic 'three-layer' architecture of RNA motif models. **Layer 1: The motif model** defines the gain of free energy upon the RBP binding to possible binding sites. Binding energy can be influenced by both sequence and structure patterns, which can be incorporated into the motif model at varying degrees of complexity. The occurrence of nucleotide dependencies within a motif can be represented by higher-order models. **Layer 2: The occupancy model** determines the likelihood of a probe being occupied by a protein given all its possible binding sites, competitive RNA probe abundance and protein concentration. **Layer 3: The extrinsic factors** influence the results of protein binding assays and need to be accounted for implicitly, or explicitly modeled in an additional layer. Extra measurements taken in some assays, such as the use of multiple protein concentrations, can provide further information and enable better generalization of the motif model.

sum models are most appropriate for short sequences, in long sequences, the accumulation of many low occupancy sites overwhelms the signal from high occupancy sites. On the other hand, the max assumption neglects potential impact of binding of multiple proteins in longer sequences.

RNA secondary structure also influences site occupancy by competing with the RBPs to bind their preferred binding site [9••,38]. *In silico* prediction tools accurately predict RNA secondary structures sequences with less than fifty nucleotides and their predictions are sufficient to incorporate into occupancy models [51,52•]. Current motif finding methods incorporate secondary structure by one of three approaches: either a single, representative RNA structure, often the minimum free energy structure; ensemble-based estimates of the frequency of pre-specified structural contexts; or a set of potential structures derived from the ensemble of energetically plausible conformations (Figure 4, protein independent). A weakness of all current motif finding methods is that they predict the secondary structures based on the RNA free energy only, ignoring the influence of RBPs on the ensemble of structures. RBPs can bind target sites that are initially partially or fully inaccessible, if the resulting complex is thermodynamically preferred (Figure 4, protein dependent) [53•,54] Some methods implicitly model this interdependency under cellular conditions by

Figure 4



Three ways to consider RNA secondary structure in the occupancy model. **Protein independent case:** A single structure is typically determined by the minimum free energy structure of the RNA sequence. To model uncertainty in the secondary structure conformation, an ensemble of potential structures can be determined and incorporated. **Protein dependent case:** Energetically, the RNA secondary structure is dependent on protein binding. Protein binding is influenced by secondary structure while secondary structure is influenced by protein binding. The **joint probability of the protein binding** can change substantially based on the way RNA secondary structure is considered in the occupancy model.

inferring motifs based on patterns of evolutionary covariation or couplings [46,55,56]. Nevertheless, these methods generally place strong requirements on the conservation and number of aligned binding sites, making them unsuitable for analyzing high-throughput, experimental binding data for most RBPs.

Modeling extrinsic factors

Cellular (or experimental) conditions, such as concentrations of RNA and protein, impact both *in vivo* and *in vitro* binding, as can experimental biases unique to individual assays [57]. These extrinsic factors (Figure 3) must be accounted for when fitting motif models. To date, in *in vitro* assays, this is done by pre-processing or normalization [15,4•]. Correcting for extrinsic factors is more difficult for *in vivo* assays. As such, methods trained on *in vivo* data often use very stringent filtering and peak calling, as well as randomly sampled background sequences before deriving motifs [58]. Nonetheless, experimental biases from cross-linking, RNase specificity, and PCR amplification can be corrected by explicitly modelling of their effect [26,37,59]. Cellular conditions that affect binding include the localization and concentration the target mRNAs, the assayed RBP, and even the concentration of other cellular RBPs and ncRNAs that cooperate or compete for binding sites [7••,8]. Since cellular conditions are very hard to control, measure and model, motif inference from *in vivo* assays has proved extremely difficult.

One solution may be to explicitly model these extrinsic factors. This would lead to a 'three-layer' model whose

innermost layer is the explicit motif model, the second layer is the sequence occupancy model, and outer layer represents the cellular and experiment effects (Figure 3). However, an explicit model of cellular effects requires knowledge of all interactors and therefore so far is limited to well-characterized cellular conditions, such as Drosophila embryogenesis [60].

A more tractable solution might be using multi-task learning, in other words, training motif models simultaneously on *in vivo* and *in vitro* data and an exhaustive set of *in vivo* measurements for all involved RBPs (e.g. from ENCODE [7••]). For instance, iONMF is a method that accounts for dependencies between RBPs *in vivo* by adding binding sites from CLIP-seq experiments of other RBPs as additional features to the motif identification of each data set [31]. The additional binding sites permit the model to learn which peaks come from cellular effects, cooperation and competition with other proteins or intrinsic binding preferences of the protein. In theory, neural network architectures, similar to those already used in implicit models, could easily support this multi-task learning. However, they would need to be redesigned to mirror the three layer architecture to permit separation of the motif and sequence occupancy models from the parameters representing extrinsic factors, thus permitting easy adaptation to new cellular conditions.

Conclusions

Ultimately, motif models learned on *in vitro* data should permit accurate prediction of *in vivo* binding. Here, we argue that current implicit modelling frameworks do not permit adjustment of their model to unseen cellular conditions and, as such, are unlikely to generalize well to *in vivo* binding data if not trained on held-out data from the same set. Therefore, we advocate for a more biophysically-motivated three (or more) layer modelling architecture to merge *in vivo* binding site prediction with more complex *in vitro* motif identification tools.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.sbi.2018.08.001.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Glisovic T, Bachorik JL, Yong J, Dreyfuss G: **RNA-binding proteins and post-transcriptional gene regulation**. *FEBS Lett* 2008, **582**:1977-1986.

2. •• Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X *et al.*: **A compendium of RNA-binding motifs for decoding gene regulation**. *Nature* 2013, **499**:172-177.
RNAcompete presents a systematic analysis of the RNA motifs recognized by 205 RNA-binding proteins from 24 diverse eukaryotes. This represents currently the largest set of binding specificities and enables the authors to analyze their evolutionary conservation and relation.

3. Ray D, Kazan H, Chan ET, Castillo LP, Chaudhry S, Talukder S *et al.*: **Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins**. *Nat Biotechnol* 2009, **27**:667-670.

4. • Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB: **RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins**. *Mol Cell* 2014, **54**:887-900.

5. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR: **RBPDB: a database of RNA-binding specificities**. *Nucleic Acids Res* 2011, **39** D301–8.

6. Tuerk C, Gold L: **Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase**. *Science* 1990, **249**:505-510.

7. •• Van Nostrand *et al.*: **A large-scale binding and functional map of human RNA binding proteins**. *bioRxiv* 2017.
The paper presents a large-scale functional analysis of post-transcriptional gene regulation and editing through in corporation of in vitro and in vivo binding with gene expression and alternative splicing data sets provided by the ENCODE project.

8. Änkö M-L, Neugebauer KM: **RNA–protein interactions in vivo: global gets specific**. *Trends Biochem Sci* 2012, **37**:255-262.

9. •• Li X, Quon G, Lipshitz HD, Morris Q: **Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure**. *RNA* 2010, **16**:1096-1107.
The method #ATS demonstrates that*in silico* predicted secondary structure significantly improves the accuracy of predicting in vivo binding for the majority of sequence specific RBPs.

10. •• Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T *et al.*: **Sequence, structure, and context preferences of human RNA binding proteins**. *Mol Cell* 2018, **70** 854–867.e9.
The paper determines sequence, structure and context preferences of 78 human RNA-binding proteins in large-scale RNAbind-n-seq assays. The paper provides evidence that proteins possess conserved core motifs but regulate gene expression through differences in sequence and structure context specificities.

11. •• Afroz T, Cienikova Z, Cléry A, Allain FHT: **One, two, three, four! How multiple RRMs read the genome sequence**. *Methods Enzymol* 2015, **558**:235-278.
The paper reports high-resolution structures of single and multi RRM-RNA complexes and lays out the numerous ways how these domains establish specific binding preferences to complex sequence and sequence-structure motifs in tandem RRMs and higher-order RNPs.

12. Maris C, Dominguez C, Allain FH-T: **The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression**. *FEBS J* 2005, **272**:2118-2131.

13. Zhao Y, Stormo GD: **Quantitative analysis demonstrates most transcription factors require only simple models of specificity**. *Nat Biotechnol* 2011, **29**:480-483.

14. Foat BC, Morozov AV, Bussemaker HJ: **Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE**. *Bioinformatics* 2006, **22** e141–9.

15. Ray D, Ha KCH, Nie K, Zheng H, Hughes TR, Morris QD: **RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins**. *Methods* 2017, **118–119**:3-15.

16. Bailey TL, Elkan C: **Unsupervised learning of multiple motifs in biopolymers using expectation maximization**. *Mach Learn* 1995, **21**:51-80.

17. Stormo GD: **Biological sequence analysis: probabilistic models of proteins and nucleic acids. Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison**. *Q Rev Biol* 2000, **75**:313-314.

18. Leibovich L, Paz I, Yakhini Z, Mandel-Gutfreund Y: **DRIMust: a web server for discovering rank imbalanced motifs using suffix trees**. *Nucleic Acids Res* 2013, **41** W174–9.

19. Ruan S, Joshua Swamidass S, Stormo GD: **BEESEM: estimation of binding energy models using HT-SELEX data**. *Bioinformatics* 2017, **33**:2288-2295.

20. •• Hiller M, Pudimat R, Busch A, Backofen R: **Using RNA secondary structures to guide sequence motif finding towards single-stranded regions**. *Nucleic Acids Res* 2006, **34**:e117.
MEMERIS represents the first method which used RNA secondary structure to guide sequence motif finding algorithms towards single stranded regions. It uses expectation maximization to find the sequence motif and guides the sequence alignment towards regions with high probabilities of being single stranded.

21. Bailey TL: **DREME: motif discovery in transcription factor ChIP-seq data**. *Bioinformatics* 2011, **27**:1653-1659.

22. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR *et al.*: **Evaluation of methods for modeling transcription factor sequence specificity**. *Nat Biotechnol* 2013, **31**:126-134.

23. Siebert M, Söding J: **Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences**. *Nucleic Acids Res* 2016, **44**:6055-6069.

24. Frith MC, Saunders NFW, Kobe B, Bailey TL: **Discovering sequence motifs with arbitrary insertions and deletions**. *PLoS Comput Biol* 2008, **4**:e1000071.

25. Riley TR, Lazarovici A, Mann RS, Bussemaker HJ: **Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE**. *Elife* 2015, **4** http://dx.doi.org/10.7554/eLife.06397.

26. Rastogi C, Rube HT, Kribelbauer JF, Crocker J, Loker RE, Martini GD *et al.*: **Accurate and sensitive quantification of protein-DNA binding affinity**. *Proc Natl Acad Sci U S A* 2018, **115**: E3692-E3701.

27. Wong K-C, Chan T-M, Peng C, Li Y, Zhang Z: **DNA motif elucidation using belief propagation**. *Nucleic Acids Res* 2013, **41**:e153.

28. Orenstein Y, Wang Y, Berger B: **RCK: accurate and efficient inference of sequence- and structure-based protein-RNA binding models from RNAcompete data**. *Bioinformatics* 2016, **32**:i351-i359.

29. Fletez-Brant C, Lee D, McCallion AS, Beer MA: **kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets**. *Nucleic Acids Res* 2013, **41** W544–56.

30. • Cook KB, Vembu S, Ha KCH, Zheng H, Laverty KU, Hughes TR *et al.*: **RNAcompete-S: combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection**. *Methods* 2017, **126**:18-28.

31. Straar M, itnik M, Zupan B, Ule J, Curk T: **Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins**. *Bioinformatics* 2016, **32**:1527-1535.

32. Ghandi M, Lee D, Mohammad-Noori M, Beer MA: **Enhanced regulatory sequence prediction using gapped k-mer features**. *PLoS Comput Biol* 2014, **10** e1003711.

33. Agius P, Arvey A, Chang W, Noble WS, Leslie C: **High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions**. *PLoS Comput Biol* 2010, **6** http://dx.doi.org/10.1371/journal.pcbi.1000916.

34. •• Alipanahi B, Delong A, Weirauch MT, Frey BJ: **Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning**. *Nat Biotechnol* 2015, **33**:831-838.
DeepBind represents the first method that uses a convolutional neural network to predict RNA binding in vitro and in vivo. Subsequent deep learning approaches apply very similar architectures and only slightly increase the performance compared to this method.

35. Budach S, Marsico A: **pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks**. *Bioinformatics* 2018 http://dx.doi.org/10.1093/bioinformatics/bty222.

36. Wang MD, Hassanzadeh HR: *DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins [Internet]*. 2017 http://dx.doi.org/10.1101/099754.

37. Bahrami-Samani E, Penalva LOF, Smith AD, Uren PJ: **Leveraging cross-link modification events in CLIP-seq for motif discovery**. *Nucleic Acids Res* 2015, **43**:95-103.

38. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q: **RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins**. *PLoS Comput Biol* 2010, **6**:e1000832.

39. Dao P, Hoinka J, Takahashi M, Zhou J, Ho M, Wang Y *et al.*: **AptaTRACE elucidates RNA sequence-structure motifs from selection trends in HT-SELEX experiments**. *Cell Syst* 2016, **3**:62-70.

40. Pan X, Shen H-B: **RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach**. *BMC Bioinform* 2017, **18**:136.

41. Polishchuk M, Paz I, Kohen R, Mesika R, Yakhini Z, Mandel-Gutfreund Y: **A combined sequence and structure based method for discovering enriched motifs in RNA from in vivo binding data**. *Methods* 2017, **118–119**:73-81.

42. Heller D, Krestel R, Ohler U, Vingron M, Marsico A: **ssHMM: extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data**. *Nucleic Acids Res* 2017, **45**:11004-11018.

43. Pietrosanto M, Mattei E, Helmer-Citterich M, Ferrè F: **A novel method for the identification of conserved structural patterns in RNA: from small scale to high-throughput applications**. *Nucleic Acids Res* 2016, **44**:8600-8609.

44. • Maticzka D, Lange SJ, Costa F, Backofen R: **GraphProt: modeling binding preferences of RNA-binding proteins**. *Genome Biol* 2014, **15**:R17.

45. Yao Z, Weinberg Z, Ruzzo WL: **CMfinder – a covariance model based RNA motif finding algorithm**. *Bioinformatics* 2006, **22**:445-452.

46. Rabani M, Kertesz M, Segal E: **Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes**. *Proc Natl Acad Sci U S A* 2008, **105**:14885-14890.

47. Duss O, Michel E, Yulikov M, Schubert M, Jeschke G, FH-T Allain: **Structural basis of the non-coding RNA RsmZ acting as a protein sponge**. *Nature* 2014, **509**:588-592.

48. Shrikumar A, Greenside P, Kundaje A: **Learning important features through propagating activation differences**. *arXiv* 2017. 1704.02685.

49. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean C, Snoek J: *Sequential regulatory activity prediction across chromosomes with convolutional neural networks [Internet]*. 2017 http://dx.doi.org/10.1101/161851.

50. Varani G, Chen Y: *Faculty of 1000 evaluation for Systematic discovery of structural elements governing stability of mammalian messenger RNAs [Internet]. F1000 – Post-publication peer review of the biomedical literature*. 2012 http://dx.doi.org/10.3410/f.717597903.793052805.

51. Lorenz R, Bernhart SH, Siederdissen CHZ, Tafer H, Flamm C, Stadler PF *et al.*: **ViennaRNA Package 2.0**. *Algorithms Mol Biol* 2011, **6**:26.

52. • Hofacker IL: **Vienna RNA secondary structure server**. *Nucleic Acids Res* 2003, **31**:3429-3431.

53. • Forties RA, Bundschuh R: **Modeling the interplay of single-stranded binding proteins and nucleic acid secondary structure**. *Bioinformatics* 2010, **26**:61-67.

54. Gaither J, Lin YH, Bundschuh R: **RBPBind: quantitative prediction of protein–RNA interactions**. *arXiv* 2016:01245.

55. Nielsen MM, Tehler D, Vang S, Sudzina F, Hedegaard J, Nordentoft I *et al.*: **Identification of expressed and conserved human noncoding RNAs**. *RNA* 2014, **20**:236-251.

56. Seemann SE, Mirza AH, Hansen C, Bang-Berthelsen CH, Garde C, Christensen-Dalsgaard M *et al.*: **The identification and functional annotation of RNA structures conserved in vertebrates**. *Genome Res* 2017, **27**:1371-1383.

57. Wheeler EC, Van Nostrand EL, Yeo GW: **Advances and challenges in the detection of transcriptome-wide protein–RNA interactions**. *Wiley Interdiscip Rev RNA* 2018, **9** http://dx.doi.org/10.1002/wrna.1436.

58. Chakrabarti AM, Haberman N, Praznik A, Luscombe NM, Ule J: **Data science issues in studying protein–rna interactions with CLIP technologies**. *Annu Rev Biomed Data Sci* 2018, **1** http://dx.doi.org/10.1146/annurev-biodatasci-080917-013525.

59. Gillen AE, Yamamoto TM, Kline E, Hesselberth JR, Kabos P: **Improvements to the HITS-CLIP protocol eliminate widespread mispriming artifacts**. *BMC Genomics* 2016, **17**:338.

60. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U: **Predicting expression patterns from regulatory sequence in Drosophila segmentation**. *Nature* 2008, **451**:535-540.