Inference of Biomolecular Interactions from Sequence Data

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Lukas Johannes Burger
aus Egg, Zürich

Basel, 2010

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Referent: Prof. Erik van Nimwegen

Korreferent: Prof. Massimo Vergassola

Expertin: Prof. Mihaela Zavolan

Basel, den 24.März 2009

Prof. Dr. Eberhard Parlow, Dekan

To my parents.

Introduction

This thesis describes our work on the inference of biomolecular interactions from sequence data. In particular, the first part of the thesis focuses on proteins and describes computational methods that we have developed for the inference of both intra- and inter-protein interactions from genomic data. The second part of the thesis centers around protein-RNA interactions and describes a method for the inference of binding motifs of RNA-binding proteins from high-throughput sequencing data.

The thesis is organized as follows. In the first part, we start by introducing a novel mathematical model for the characterization of protein sequences (chapter 1). We then show how, using genomic data, this model can be successfully applied to two different problems, namely to the inference of interacting amino acid residues in the tertiary structure of protein domains (chapter 2) and to the prediction of protein-protein interactions in large paralogous protein families (chapters 3 and 4). We conclude the first part by a discussion of potential extensions and generalizations of the methods presented (chapter 5).

In the second part of this thesis, we first give a general introduction about RNA-binding proteins (chapter 6). We then describe a novel experimental method for the genome-wide identification of target RNAs of RNA-binding proteins and show how this method can be used to infer the binding motifs of RNA-binding proteins (chapter 7). Finally, we discuss a potential mechanism by which KH domain-containing RNA-binding proteins could achieve the specificity of interaction with their target RNAs and conclude the second part of the thesis by proposing a novel type of motif finding algorithm tailored for the inference of their recognition elements (chapter 8).

Contents

Part I

Inference of Intra- and Inter-Protein Interactions from Genomic Data

Chapter 1

Bayesian network model for the characterization of aligned protein sequences

1.1 Introduction

The identification and characterization of functionally and structurally important elements in DNA, RNA and protein sequences is one of the main focuses of computational biology. The mathematical framework suited to describe a particular functional sequence element can differ strongly depending on the type of sequence (DNA, RNA or protein) and the particular problem. In the simplest case, a functional element can be described as a fixed string of a given length. This is at least approximately the case for miRNA target sites, where the presence of a sequence stretch that corresponds to the reverse complement of the first 8 nucleotides at the 5' end of the miRNA (or a one to two nucleotide shorter substring thereof) is of crucial importance for target site recognition [1,2]. If degeneracies are allowed at certain positions, strings are no longer a practical model as the number of possible strings grows very quickly. Degeneracies can more easily be described by regular expressions, or more generally, by position-specific weight matrices. Weight matrices model every position of the sequence with an independent probability distribution, that, unlike regular expressions, allows for a weighting of each nucleotide according to its frequency of occurrence. Weight matrix models have been successfully applied to many problems and are, for example, a generally accepted framework for the description of transcription factor binding sites (see e.g. [3]).

In cases where the secondary or tertiary structure of the sequence is of importance, such as in the case of RNA sequences with a particular structure (for example tRNAs) or protein domains, a weight matrix may still be too limited a model. For multiple

alignments of structured RNAs, it was noted a long time ago that particular pairs of positions, corresponding to base-paired residues in the respective structures, show strong correlations and can thus not be modelled independently (see e.g. [4,5]). This led to the development of so-called covariance models that probabilistically describe the sequence in terms of a set of single independent residues and a set of pairs of correlated residues [4].

However, as probabilities need to be estimated for the occurrence of all possible *pairs* of residues, covariance models require much larger training sets compared to weight matrix models. For RNA, due to the small alphabet size - there are only 4 different nucleotides and thus only 16 different pairs of nucleotides - this has not been a major issue, but for proteins, sensitive covariance analysis has for a long time been hindered by the fact that most protein families were two small to reliably estimate the frequency of occurrence of all $20^2 = 400$ possible pairs of amino acids. However, with the recent explosion of genomic data, the average number of homologous sequences per protein family has drastically increased and more and more protein families have become amenable to covariance analysis. Accordingly, much work has in recent years been dedicated to the detection of correlations between protein residues (see e.g. [6–16]). This work has shown that, like in RNA alignments, there is strong evidence for dependencies between pairs of positions in protein alignments. In particular, it has been shown that correlated pairs of residues, both within protein domain alignments and between alignments of interacting protein families, tend to lie in functionally important regions and tend to be close in the tertiary structure [9, 10, 13–16].

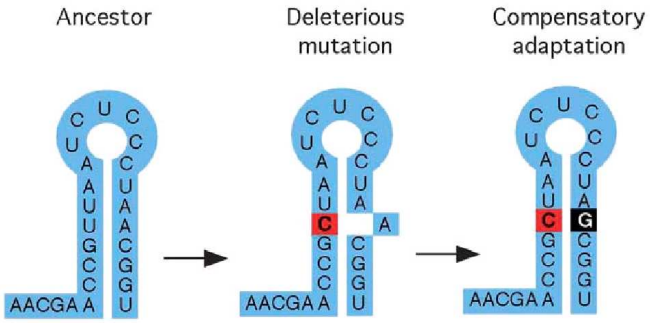


Figure 1.1: Illustration of a co-evolutionary event in a RNA sequence. A random mutation in the base pair U-A within the stem of a stable stem-loop structure results in a mismatched pair C-A, which destabilizes the structure. A compensatory mutation in the second nucleotide from A to G re-stabilizes the stem loop and compensates for the drop in fitness caused by the first mutation. Figure taken from [17].

In both RNA and protein sequences, correlations between residues have been attributed to co-evolutionary events, where the change in one residue must be compensated by a correlated mutation in a second residue in order to maintain the functionality of the sequence and thus the fitness of the organism [17–19]. For example,

as illustrated in figure 1.1, due to a random mutation, one nucleotide of a base pair within a functionally important RNA stem-loop may mutate, leading to a mismatched pair, which leads to the destabilization of the stem-loop and thus to a decrease in fitness. This drop in fitness can be compensated by a second, compensatory mutation in the second residue of the original base pair, which re-stabilizes the stem-loop. Interestingly, for proteins, it has been argued that compensatory mutations are more frequent than mutations that simply reverse the first random mutation, as there are typically several different residues at which compensatory mutations can occur [17].

The existence of correlated pairs of residues in protein alignments may suggest that protein sequences, similar to RNA sequences, can be well described in terms of sets of independent residues and sets of disjoint pairs of dependent residues. However, recently, there have been indications that there are important differences in the way RNA and protein residues co-evolve. Whereas RNA residues typically co-evolve as independent pairs [4, 5, 20], many co-evolving residues in proteins form chains or networks that may even connect residues that are distant in the tertiary structure [13, 14, 21, 22]. We and others [14, 16] have thus argued that a sound mathematical description of protein sequences must take into account the interconnectedness of co-evolving protein residues. To this end, we have developed a Bayesian network model that models the joint distribution of amino acids in a multiple alignment in terms of an underlying dependence tree structure and in this way can characterize the co-evolutionary patterns in proteins in a statistically sound and computationally tractable way. In the next section, we will give an intuitive explanation of this model. For mathematical details and calculations, the reader is referred to chapters 2, 3 and 4.

1.2 Mathematical framework

In principle, we would like to describe a set of aligned protein sequences in terms of the joint distribution of amino acids over all residues [1]. However, as the number of possible combinations of amino acids grows as 20^n, where n is the number of columns in the alignment (i.e. the number of residues of the proteins), and given the typical number of sequences per protein family in current databases (on the order of $10^2 - 10^3$ sequences for larger families [23]), it is for most families not feasible to go beyond the estimation of distributions of pairs of amino acids. We thus propose to describe the distribution of amino acids in protein alignments based only on conditional distribu-

[1] In the case of two interacting protein families, we join the two alignments so as to create an alignment of interacting protein pairs (cf chapters 3 and 4). This alignment is then modelled in the same manner as an alignment of single proteins. In a joint alignment of interacting protein pairs, the dependencies reflect constraints that are due to both the structure of the single proteins and due to the interaction.

tions of pairs of variables. In particular, our Bayesian model makes every column of the alignment conditionally dependent on exactly one other column. If the different columns of the alignment, corresponding to the distribution of amino acids in each position, are regarded as nodes in a graph and edges are drawn only between those nodes that are directly dependent on each other, then the resulting graph of the model is a spanning tree (see figure 1.2). A spanning tree has the property that it connects all the nodes of the graph without forming any cycles. Due to the tree structure of the underlying graph, the type of model that we propose is also called dependence tree model [24].

Given a spanning tree T, we can factorize the joint distribution of all residues $P(D_1, D_2, .., D_n|T)$ as

$$P(D_1, D_2, .., D_n|T) = P(D_r) \prod_{i=1, i\neq r}^{n} P(D_i|D_{\pi_T(i)}) \tag{1.1}$$

Here, D_i denotes the amino acids in position i, $\pi_T(i)$ stands for the node which node i depends on in the tree T (the 'father' node of node i), r denotes the root node of the tree and n is the total number of columns of the alignment. For example, in figure 1.2, node j is the root node as well as the father node of both nodes i and k. Note that independence of any node i, i.e. $P(D_i|D_{\pi_T(i)}) = P(D_i)$ is contained in this model. Writing the conditional probabilities as joint and marginal probabilities, equation 1.1 can be rewritten as

$$P(D_1, D_2, .., D_n|T) = \prod_{i=1}^{n} P(D_i) \prod_{i=1, i\neq r}^{n} R_{i\pi_T(i)} \tag{1.2}$$

with

$$R_{ij} \doteq \frac{P(D_i, D_j)}{P(D_i)P(D_j)} \tag{1.3}$$

R_{ij} is the ratio of the joint probability of the amino acids in columns i and j, divided by the marginal probability of the data in columns i and j separately. It is thus a measure of dependence between the two columns. It can be shown that in the limit of large amino acid counts, the logarithm of R is proportional to mutual information [25], a measure that is frequently used to determine co-evolving residues [6,9,10,13,21,26].

Noting that the second term of equation 1.2 is in fact the product of the R-values over all edges of the tree and thus independent of the choice of the root, we can rewrite this equation as

$$P(D_1, D_2, .., D_n|T) = \prod_{i=1}^{n} P(D_i) \prod_{e\in T} R_e \tag{1.4}$$

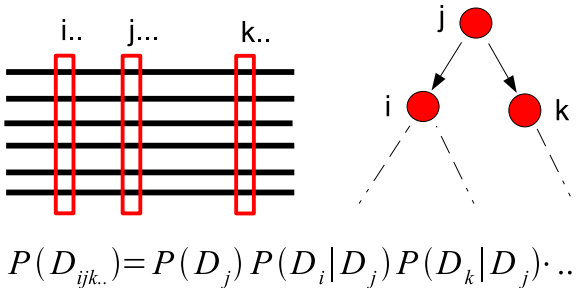


Figure 1.2: Illustration of the Bayesian network model. For the sake of simplicity, we only show three columns of the alignment, i, j and k. The random variables describing the distribution of amino acids in every column of the multiple alignment on the left corresponds to nodes in the graph on the right, which describes the dependencies between the variables. In our model, every node except for the root (here j) is dependent on exactly one other node in the tree (its father node), designated by the corresponding arrows. The resulting graph is a spanning tree, which allows for the factorization of the joint distribution of amino acids in all positions into conditional probabilities of pairs of variables.

where R_e is the dependency between the two nodes connected by edge e. This equation can be interpreted in a very intuitive way. The first product is equal to the probability of the data under a simple weight matrix model and describes the data independently for each residue. The second term, on the other hand, measures the amount of dependence along the edges of the spanning tree. In cases where there is strong dependence between most nodes that are connected by edges of the tree, most R values in the second term will be large ($\gg 1$) and the product over all edges larger than one, making the data more likely under a dependence tree model than a simple weight matrix model. On the other hand, if there is little evidence of dependence, most R values will be smaller than unity and, accordingly, the data is better described by a weight matrix model.

In most practical applications, the structure of the spanning tree T is not known. There are two ways to go about this problem. We can either *infer* the maximum likelihood tree T^* and approximate the probability of the data as

$$P(D_1, D_2, .., D_n) \approx \prod_{i=1}^{n} P(D_i) \prod_{e \in T^*} R_e \tag{1.5}$$

or we can calculated the probability of the data by summing over all possible spanning trees

$$P(D_1, D_2, .., D_n) = \prod_{i=1}^{n} P(D_i) \left(\frac{1}{|T|} \sum_{T} \prod_{e \in T} R_e \right) \tag{1.6}$$

where $|T|$ is the total number of spanning trees and we here assume a uniform prior over all spanning trees $\frac{1}{|T|}$.

It is relatively straightforward to infer the maximum-likelihood tree. Taking the logarithm of equation 1.4, we see that the log-likelihood of the data is equal to the sum of the $\log R$ values along the edges of the tree plus an additive term that is independent of the particular tree T,

$$\log P(D_1, D_2, .., D_n|T) = \sum_{e \in T} \log R_e + \sum_{i=1}^{n} \log P(D_i) \tag{1.7}$$

Thus, inferring the maximum-likelihood tree is equivalent to inferring the tree with the largest sum of $\log R$-values along its edges (for a derivation based on information-theoretical concepts, see [24]). This problem is a well-known problem in computer science (the so-called maximum spanning tree problem) and can be solved very efficiently, for example using Kruskal's algorithm [27].

The maximum-likelihood expression in equation 1.5 is a good approximation of the probability of the data if there is one dominating tree or a set of dominating trees with similar structure. However, for typical protein alignments with on the order of 10^2 to 10^3 sequences (see chapter 2), there are often many R values of similar magnitude and a maximum-likelihood estimate may easily discard certain 'true' edges. In this case, it is desirable to calculate the full probability of the data by summing over all possible spanning trees (equation 1.6). This is a difficult problem as the number of spanning trees $|T|$ grows super-exponentially in the number of nodes n, $|T| = n^{n-2}$. However, thanks to recent results in Bayesian network theory [28], the sum over all spanning trees only involves the calculation of a determinant and can thus be efficiently carried out,

$$\sum_{T} \prod_{e \in T} R_e = M(L) \tag{1.8}$$

where $L_{ij} = (\sum_k R_{ik})\delta_{ij} - R_{ij}$ is the Laplacian matrix of a graph with edge weights R_{ij} (δ_{ij} is the Kronecker delta, which is one if i equals j and zero otherwise) and M denotes any first minor of L, i.e. the determinant of the matrix L with any one column and row crossed out (the determinant is here independent of which row and column is removed).

Besides being useful for determining the amount of dependence in a sequence alignment (cf chapter 4), expression 1.8 also allows for the calculation of posterior probabilities of certain quantities of the model. In particular, the posterior probability of an edge is a very powerful quantity in contexts where we want to identify the pairs of columns of a multiple alignment that show most evidence of *direct* dependency (and do not depend indirectly on each other via other columns), as discussed in detail in the next chapter. The concept of the posterior probability is illustrated in figure 1.3. The posterior probability of an edge (i, j) is proportional to the sum of

the probabilities of all trees that contain the edge (i, j) and is calculated as

$$P((i,j)|D_1, D_2, .., D_n) = \frac{\sum_{T:(i,j)\in T} \prod_{e\in T} R_e}{\sum_T \prod_{e\in T} R_e} \tag{1.9}$$

where the sum in the numerator goes over all spanning trees that contain the edge (i, j). Intuitively speaking, the posterior probability of edge (i, j) is high if most trees of high likelihood contain this edge. Expression 1.9 can be easily calculated by noting that the sum over all trees that contain edge (i, j) is equal to the sum over all trees minus the sum over all trees that do *not* contain the edge (i, j) and thus

$$P((i,j)|D_1, D_2, .., D_n) = 1 - \frac{\sum_{T:(i,j)\notin T} \prod_{e\in T} R_e}{\sum_T \prod_{e\in T} R_e} \tag{1.10}$$

The sum over all trees without edge (i, j) is then given by expression 1.8 with R_{ij} set to zero (in this way all trees including this edge have zero weight).

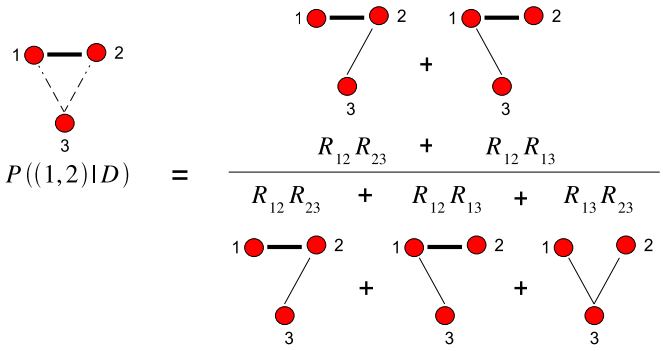


Figure 1.3: Illustration of the calculation of the posterior probability. For the sake of simplicity, we consider only three columns of the multiple alignment, depicted as nodes 1, 2 and 3 of the graph. In our model, the R-values describe the dependency between pairs of columns and the probability of the data given a particular tree is proportional to the product of the R-values along the edges of the tree. The posterior probability of an edge, here $(1, 2)$, is then given by the ratio of the sum of probabilities of all trees that contain the edge and the sum of the probabilities of all trees.

In the following chapters, we will describe two application of our Bayesian network model. In the first application (chapter 2), we use the posterior probabilities 1.9 to infer contacting pairs of residues in the tertiary structure of protein domains and show that due to the distinction of directly from indirectly dependent residues, our method significantly outperform previous methods. In a second application, we use both the maximum-likelihood expression 1.5 (chapters 3 and 4) and the full Bayesian expression 1.6 (chapter 4), applied to joint alignments of interacting proteins, to infer the specificity of interaction in large paralogous protein families.

Chapter 2

Disentangling Direct from Indirect Co-evolution of Residues in Protein Alignments

Lukas Burger and Erik van Nimwegen
PLoS Computational Biology, 6(1):e1000633, 2010

Predicting protein structure from primary sequence is one of the ultimate challenges in computational biology. Given the large amount of available sequence data, the analysis of co-evolution, i.e. statistical dependency, between columns in multiple alignments of protein domain sequences remains one of the most promising avenues for predicting residues that are contacting in the structure. A key impediment to this approach is that strong statistical dependencies are also observed for many residue pairs that are distal in the structure. Using a comprehensive analysis of protein domains with available three-dimensional structures we show that co-evolving contacts very commonly form chains that percolate through the protein structure, inducing indirect statistical dependencies between many distal pairs of residues. We characterize the distributions of length and spatial distance traveled by these co-evolving contact chains and show that they explain a large fraction of observed statistical dependencies between structurally distal pairs. We adapt a recently developed Bayesian network model into a rigorous procedure for disentangling direct from indirect statistical dependencies and we demonstrate that this method not only successfully accomplishes this task, but also allows contacts with weak statistical dependency to be detected. To illustrate how additional information can be incorporated into our method, we incorporate a phylogenetic correction, and we develop an informative prior that takes into

account that the probability for a pair of residues to contact depends strongly on their primary-sequence distance and the amount of conservation that the corresponding columns in the multiple alignment exhibit. We show that our model including these extensions dramatically improves the accuracy of contact prediction from multiple sequence alignments.

2.1 Introduction

The identification of functionally and structurally important elements in DNA, RNA and proteins from their sequences has been a major focus of computational biology for several decades. A common approach is to create a multiple alignment of homologous sequences, which places 'equivalent' residues into the same column and as such gives a hint of the evolutionary constraints that are acting on related sequences. In particular, so-called profile hidden Markov models [29] of protein families and domains have been highly successful in identifying sequences that have similar function and fold into a common structure, making them among the most important tools in functional genomics, see e.g. [30]. These hidden Markov models typically assume that the residues occurring at a given position are probabilistically independent of the residues occurring at other positions. At the time at which these models were developed, it was entirely reasonable to ignore dependencies between residues at different positions, since the amount of available sequence data was generally insufficient to estimate joint probabilities of multiple residues. However, currently the multiple alignments of many protein families and domains include hundreds and sometimes even thousands of sequences, making it possible to systematically investigate dependencies between the residues at different positions.

As the functionality of biomolecules crucially depends on their three-dimensional structures, whose stabilities depend on interactions between residues that are near to each other in space, it is of course to be expected that significant dependencies between residues at different positions will exist. Indeed such dependencies are evident for RNA (eg [4, 5]) and protein sequences [18, 19]. The existence of dependencies between residues at different positions is also supported by the observation of correlated mutations in which mutations at one residue tend to be compensated by a correlated mutation in a particular other residue [17–19].

Recently there has been a significant amount of work in which multiple alignments of single protein families have been used in order to predict pairs of residues that are functionally linked or interact directly in the tertiary structure (see eg [6, 8–13] and references therein). This work has shown that pairs of residues which show statistical dependencies are generally significantly closer in the structure than randomly chosen pairs. However, it has been repeatedly noted that there exist many highly statistically-dependent residues that are distant in space (eg [13, 14, 31]). Figure 2.1

illustrates these points. One of the most commonly used measures of dependency between two residues is the mutual information [5,8,13,25,32] between the distributions of amino acids occurring in the two corresponding alignment columns. We collected a comprehensive set of 2009 multiple alignments of protein domains from the Pfam database [23] for which a three dimensional structure was available (see *Materials and Methods*) and calculated, for each pair (ij) of columns in each alignment, the statistical dependency using a measure, $\log(R_{ij})$, which is a finite-size corrected version of mutual information (see *Materials and Methods*). Since the distribution of $\log(R)$ values for an alignment depends strongly on the number of sequences in the alignment, their phylogenetic relationship, and the length of the alignment, $\log(R)$ values cannot be directly compared across different alignments. Therefore, we calculated the mean and variance of $\log(R)$ values for each alignment and transformed the $\log(R)$ values to Z-values (number of standard deviations from the mean). Finally, for each alignment, we divided all pairs of residues into those that are contacting in the three-dimensional structure, and those that are distant in the structure, and calculated the distribution of Z-values for these two sets of residue pairs. As in previous work (e.g. [9, 33]) and as defined for CASP [34], two residues were considered in contact if their C_β distance (C_α for glycines) in the structure was smaller than 8. Combining the data from all alignments, the left panel of Figure 2.1 shows the fraction of all pairs of contacting residues (red) and distal residues (blue) larger than a given Z-value as a function of Z. The right panel shows, as a function of Z, what fraction of all residue pairs with at least this Z-value are contacting in the structure.

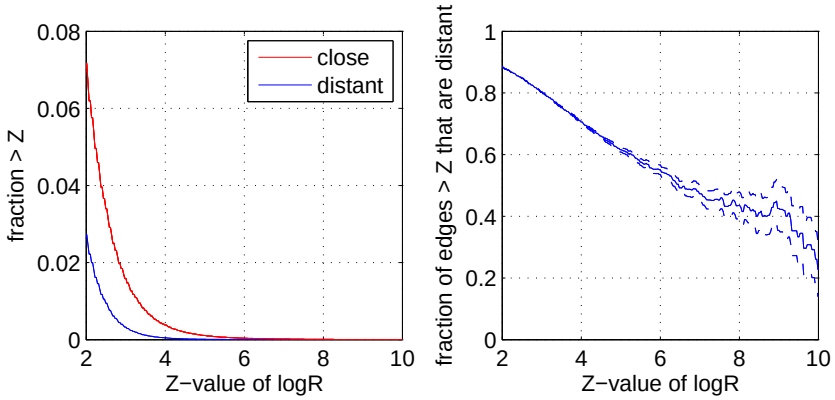


Figure 2.1: **Statistical dependencies of structurally close and distal residue pairs.** Left panel: Reverse-cumulative distribution of $\log(R)$ Z-values (horizontal axis) for structurally close (red) and distal (blue) residue pairs. Right panel: The fraction of all residue pairs that are distal in the structure as a function of their statistical dependency (Z-value).

The left panel of Figure 2.1 illustrates that, indeed, a higher fraction of contacting residues shows strong statistical dependencies than distal residues. However, we also see that the difference in the Z-distribution of close and distal pairs is only moderate.

Since there are generally many more distal pairs than close pairs, this implies that, even at high Z-values, the majority of residue-pairs are in fact distal in the structure (Figure 2.1, right panel). This result shows that simple measures of statistical dependency, such as mutual information, are poor at predicting which pairs of residues are directly contacting in the structure.

The main question is why so many structurally distal pairs show statistical dependencies in their amino-acid distributions that are stronger than those between directly contacting residues. First, whereas measures such as mutual information treat the sequences in the multiple alignments as statistically independent, in reality many of the sequences are phylogenetically closely related, which can cause 'spurious' statistical dependencies to appear between independent residue pairs which can be larger than the true statistical dependencies between contacting pairs. Several groups have investigated this confounding factor in contact prediction and several methods have been proposed for correcting these spurious phylogenetic correlations [6,8,12,13], which we will make use of below.

Although important, many strong statistical dependencies between distal residues remain even when spurious phylogenetic dependencies are corrected for (see below). Some of these distant dependencies have been suggested to be caused by homo-oligomeric interactions [13,16]. Thus, in this interpretation, some of the 'distal' pairs with strong statistical dependencies are in fact contacting in the homo-oligomer. Although it is not clear how many of the distal dependencies can be explained by this mechanism, it seems likely that only a relatively small number of residue pairs on the surface can be responsible for such homo-oligomeric interactions.

A third explanation that has been offered for the large number of distal pairs with strong statistical dependencies is that these dependencies are induced by *indirect* interactions that are mediated either by intermediate molecules [14, 21] or by chains of directly interacting residue pairs that run through the protein and connect distal pairs [21,22,35]. Indeed, for a small number of example domains, the existence of such chains of thermodynamically directly coupled residues has been demonstrated [21,22]. However, the connection between thermodynamic coupling and covariation is still under debate as there is little evidence that thermodynamic coupling of residues is limited to covarying positions [36].

In this paper, we comprehensively investigate to what extent statistical dependencies between distal pairs can be explained by indirect dependencies. The conceptual idea is illustrated in figure 2.2.

In this illustration, the letters reflect different residues, their distances in the figure reflect their distances in the three dimensional structure, i.e. only the pairs A-B, B-C, and D-E interact directly, and the strength of the statistical dependencies between the different pairs are represented by the thickness of the lines connecting them. Because the pairs A-B and B-C have very high statistical dependency, a strong dependency between A and C is *induced*, which is larger even than the statistical dependency of the

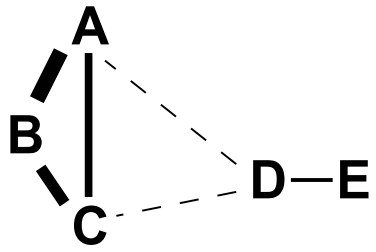


Figure 2.2: **Statistical dependencies between pairs of residues reflect both direct and indirect interactions**. The 5 letters (A through E) represent 5 residues and their distances in the figure reflect their distances in the three-dimensional structure. We assume that the pairs A-B, B-C, and D-E are in contact and interact directly. The thickness of the edges between pairs of nodes reflect the statistical dependencies between the corresponding columns in the multiple alignment.

directly interacting pair D-E. Any method that considers the statistical dependencies of each pair independently would thus erroneously assign higher confidence to the interaction of A-C than that of D-E.

It should be noted that mutual information and variants thereof have been used extensively for the inference of interacting nucleic acid pairs (see [5] for a review) in the secondary structures of RNA sequences. In these approaches too, the significance of the statistical dependency between a pair of potentially interacting positions is typically evaluated in isolation, i.e. independent of the dependencies between all other pairs. However, in contrast to protein structures, RNA secondary structures per definition consist of *disjoint pairs* of directly interacting residues, i.e. those that form Watson-Crick base pairs. Thus, for RNA secondary structures the 'percolation' of statistical dependencies to pairs that are distal in the structure cannot occur (ignoring tertiary structure).

Below we show that chains of statistically dependent contacts are very common in protein structures, explaining a significant fraction of observed dependencies between structurally distal pairs, and we characterize the distribution of lengths and distance traveled by such chains. We show that a Bayesian network model which we recently developed to predict protein-protein interactions [37] can be adapted to rigorously disentangle direct from indirect statistical dependencies between residues, and we demonstrate that such an approach much improves the prediction of pairs of residues that are in contact in the three-dimensional structure. We then investigate to what extent our Bayesian network algorithm can be further improved by incorporating a correction for the phylogenetic dependencies between sequences in the alignment [13], and by incorporating prior information regarding possible interactions. In particular we develop an informative prior that incorporates the observations that the probability for two residues to interact depends strongly on their distance in the primary sequence, and that highly conserved positions in the multiple alignment

tend to interact with a higher number of other residues. We show that incorporating these additional features into our Bayesian network model dramatically improves the accuracy of the predictions.

2.2 Results

2.2.1 Distant co-evolving pairs can frequently be explained by chains of co-evolving contacts

As mentioned above, it has been suggested that statistical dependencies between structurally distant residue pairs can be explained by chains of contacts that are all statistically dependent. However, the existence of such 'co-evolving chains' of contacts has only been demonstrated for a small number of examples [21, 22]. To examine comprehensively and systematically to what extent statistical dependencies between structurally distal residues can be explained by co-evolving chains of contacts we extracted, for each multiple alignment, all pairs of residues that showed high statistical dependency ($Z_{ij} > 4$). We then divided these 'co-evolving pairs' into co-evolving contacts and co-evolving distal pairs. As illustrated in Figure 2.3, we then determined for each distal pair whether there exists a chain of contacts that each show stronger co-evolution than the distal pair, i.e. $Z > Z_{ij}$ for all contacts in the chain.

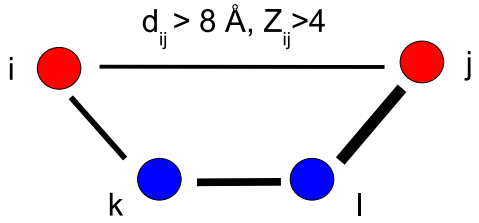


Figure 2.3: **Illustration of a chain that explains the dependency between two distant residues i and j.** The distance between the nodes illustrates the spatial separation and the thickness of the edges represents the strength of the dependence. Nodes i and j can be connected indirectly via a chain of contacts ($d < 8$) through nodes k and l (in blue) whose edges all have higher dependency (i.e. $Z_{ik} > Z_{ij}$, $Z_{kl} > Z_{ij}$ and $Z_{lj} > Z_{ij}$).

However, since our Z-values are in all likelihood only a very noisy measure of the true co-evolution of pairs, we expect that frequently one or more of the contacts in the chain may have a lower Z-value, even if their true co-evolution is higher than the co-evolution of pair (ij). We therefore also consider chains where some contacts (kl) have $Z_{kl} < Z_{ij}$ and define the total score $T(C)$ of a chain C as the sum of the difference in Z-value for all edges that have lower Z-value than the distal pair (ij),

i.e

$$T(C) = \sum_{(kl)\in C} (Z_{ij} - Z_{kl})\Theta(Z_{ij} - Z_{kl}), \tag{2.1}$$

where $\Theta(x)$ is the Heaviside-function which is one when $x \geq 0$ and zero otherwise. For each distal co-evolving pair, we determined the chain of contacts C that has minimal total score $T(C)$. Since pairs that are very distal per definition require longer chains, and since $T(C)$ generally grows with the length of the chain, we define the final score S of the best path for a given pair as the average score per contact, i.e. $S = T/n$, where n is the number of contacts in the best path.

The left panel of Figure 2.4 shows the cumulative distribution of the scores S of the best chains (blue curve). We see that for 6.5% of the distal co-evolving pairs, there exists a chain with score $S = 0$, i.e. where all contacts in the chain have $Z > Z_{ij}$. The median score of the best contact path is a little larger than $S = 1$, and the 25th and 75 percentiles occur at S-values of about 0.5 and 2 respectively. Note that, as all distal co-evolving pairs have $Z_{ij} > 4$, even at a score of $S = 2$ the contacts in the path have $Z > 2$ on average, meaning that they are still among the most significantly co-evolving pairs.

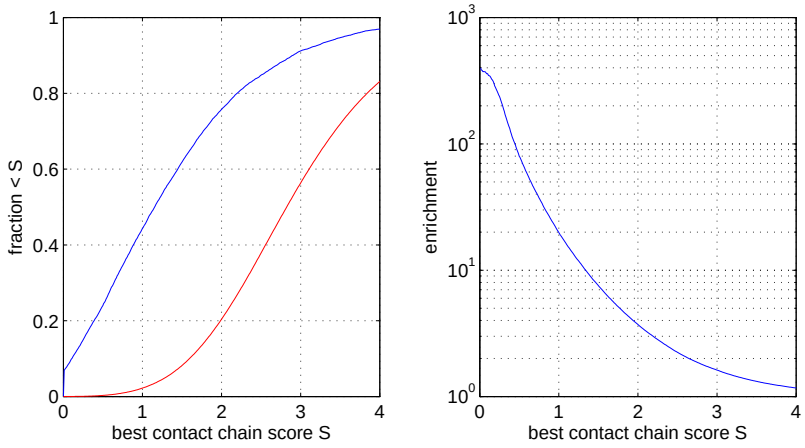


Figure 2.4: **Most distal co-evolving pairs can be explained by chains of co-evolving contacts**. Left panel: Cumulative distributions for the number of distal pairs (ij) $(d_{ij} > 8)$ that co-evolve $(Z_{ij} > 4)$ that can be explained by chains of co-evolving contacts as a function of the score S of the best chain (see text). The blue line shows the distribution for the true data and the red curve for the randomized data. Right panel: Ratio (fold-enrichment) of the fraction of distal co-evolving pairs that can be explained by chains versus the fraction that can be explained by chains from the randomized data. The vertical axis is shown on a logarithmic scale.

To assess the significance of the cumulative distribution S we performed a randomization test by randomly permuting the Z-values of all contacts of each domain 100 times and determining the S scores of the best paths that are obtained with these

permuted Z-values. The red curve in the left panel of Figure 2.4 shows the cumulative distribution of S-scores obtained in this randomized set and it is immediately clear that the S-scores are much higher for the randomized set. The right panel of Figure 2.4 shows, as a value of S, the ratio between the fraction of distal pairs that can be explained by a chain with score less than S for the real and the randomized data. Especially at low values of S the ratios are enormous. For example, at $S = 0.5$ the ratio is about 100, meaning that whereas about 25% of the distal pairs can be explained by chains in the real data, in the randomized data virtually no distal pairs can be explained, i.e. only 0.25%. But strong enrichment persists until much higher values of S. For example, at $S = 1.5$ about two-thirds of distal pairs can be connected by a chain, whereas the percentage is less than 8% for the randomized data.

2.2.2 Statistics of co-evolving contact chains

Our results show that, across essentially all protein domains for which multiple alignments and structures are available, chains of co-evolving contacts are common and explain a large fraction of statistical dependencies observed between structurally distal pairs. To gain insights in the nature of these co-evolving contact chains in protein structures, we selected all distal pairs that are explained by contact chains with scores $S < 1.5$ and obtained statistics on the number of steps and the spatial distance covered by these chains (Figure 2.5).

We see that the distance distribution of 'explainable' distal co-evolving pairs is roughly exponential with a length scale of about 8 Å. Since 'distal pairs' are by definition at least 8 apart, this means that the typical length scale covered by co-evolving contact chains is about 16Å. The right panel of Figure 2.5 shows the mean number of steps in the shortest co-evolving contact chain as a function of the structural distance of the co-evolving distal pair. With increasing spatial separation, the number of edges in the chain steadily increases from on average 2 steps at a separation of 8 to 15 steps at 50. Interestingly, the increase in the average number of steps as a function of distance is almost perfectly linear and corresponds to 3.25 ± 0.05 per step. We thus see that 'typical' co-evolving contact paths contain about $16/3.25 \approx 5$ steps, demonstrating that statistical dependencies typically percolate along paths with multiple steps. We also note that some chains are very long, consisting of up to 20 steps, connecting residues that are as far as 60 apart in the structure.

2.2.3 Bayesian network model

The insight that many of the statistical dependencies between structurally distal pairs result from chains of co-evolving contacts has important consequences for contact prediction methods. That is, any method that aims to predict contacting residues

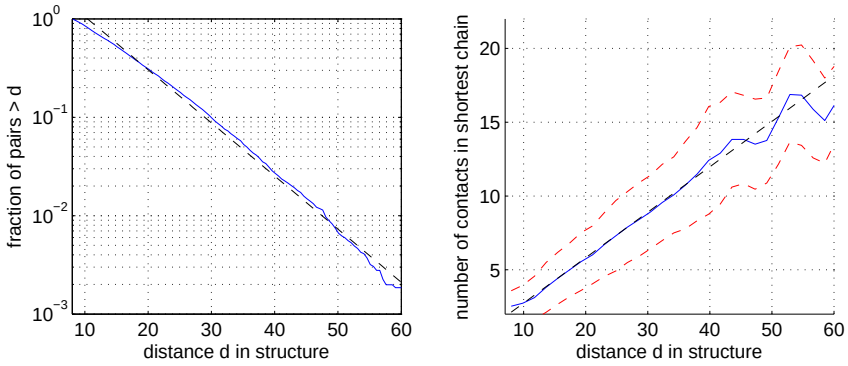


Figure 2.5: **Statistics of co-evolving contact chains**. Left panel: Reverse-cumulative distribution of the spatial distances between co-evolving pairs that can be explained by chains of co-evolving contacts of score $S < 1.5$. The vertical axis is shown on a logarithmic scale. The dotted line shows a fit to an exponential distribution $P(d > x) \propto e^{-x/8}$. Right panel: Number of steps in the shortest co-evolving contact chain as a function of the spatial distance of the co-evolving pair. The blue line shows the mean distance and the red dotted lines show mean plus and minus one standard deviation. The black dotted line shows a linear fit, the fitted slope of which corresponds to an increase in distance by 3.25 ± 0.05 per additional contact in the chain.

from statistical dependencies should clearly take into account indirect dependencies that are induced by such chains.

In [37] we developed a general Bayesian network model for calculating the probability of a multiple alignment of protein sequences taking into account dependencies between amino acids at all possible pairs of positions. We refer the reader to [37] for a comprehensive explanation of the method. Briefly, our model assumes that the sequences in a multiple alignment D (the data) are drawn from an (unknown) underlying joint probability distribution $P(x_1, x_2, \ldots, x_l)$ with l the width of the alignment and x_i the amino acid at position i. Profile hidden Markov models typically assume that the amino acids at different positions are independent so that one can write $P(x_1, x_2, \ldots, x_l) = \prod_{i=1}^{l} P_i(x_i)$, with $P_i(x)$ the probability distribution of amino acids at position i. Note that, since there are 20 amino acids (disregarding gaps), such models will have $19 \times l$ parameters in total. Our model of $P(x_1, \ldots, x_l)$ allows general dependencies, such that the probability for an amino acid at position i depends on the amino acids at other positions. Note that, if the residue at i is dependent on a residue at one single other position j, there are already $20 * 19 = 380$ parameters in the distribution $P(x_i|x_j)$, and that models with dependencies on two other positions, i.e. $P(x_i|x_j, x_k)$, would have 7600 parameters for each residue. Given the current amount of sequence data, it is certainly reasonable to consider models with single dependencies, but there is hardly ever enough data to meaningfully estimate 7600 parameters per position. Our model therefore only considers pairwise conditional dependencies of the form $P(x_i|x_j)$.

Any model that considers only pairwise conditional dependencies factorizes the joint probability $P(x_1, \ldots, x_l)$ as a product $P(x_1, \ldots, x_l) = \prod_{i=1}^{l} P(x_i|x_{\pi(i)})$, where $\pi(i)$ is the single other position which the residue at position i depends on (note that independence, i.e. $P(x_i|x_{\pi(i)}) = P(x_i)$ is contained in this general model). Our Bayesian network model is the most general model of this form. In particular, we do not attempt to estimate the conditional probabilities $P(x_i|x_j)$ but rather treat these conditional probabilities as nuisance parameters that we integrate out in calculating the likelihood of the alignment. In addition, and importantly, we do not consider only a single 'best' way of choosing which other position $\pi(i)$ each position i depends on, but rather we *sum* over all ways in which the dependencies can be chosen. Note that if we consider each column of the alignment as a node in a graph and connect each node i to the node it depends on, $\pi(i)$, then any consistent set of dependencies π, i.e. any set of dependencies π that does not introduce cycles in the graph, corresponds to a *spanning tree* of this graph. Thus, the sum over all consistent ways in which we can assign dependencies is in fact the sum over the set of all possible spanning trees of our graph. As explained in [37] and the *Materials and Methods* section, all integrals over the unknown conditional probabilities $P(x_i|x_j)$ can be performed analytically and, importantly, the sum over all spanning trees can be calculated as a matrix determinant using a generalization of Kirchhoff's theorem [28]. It is thus feasible to do inference with this general Bayesian network for a large number of multiple alignments, including alignments that are hundreds of columns wide.

2.2.4 Posterior probability of a pairwise interaction

In our model the joint probability of a multiple alignment is given as the sum over all possible spanning trees of node-dependencies, where each spanning tree is weighted according to the product of statistical dependencies across all edges in the tree (see *Materials and Methods*). Here the statistical dependence between any pair of positions (ij) is given by the ratio $R_{ij} = P(D_{ij})/[P(D_i)P(D_j)]$ of the joint probability of the alignment columns $P(D_{ij})$ and the product $P(D_i)P(D_j)$ of their marginal probabilities. Since the number of edges in any spanning tree is limited, there is a natural 'competition' in this model between the edges to be included in the spanning tree. Therefore, spanning trees with the highest statistical weight will only use edges whose statistical dependence can *not* be explained by chains of other edges with higher dependency, and edges between pairs with indirect statistical dependency will thus only appear in spanning trees with relatively low statistical weight. The posterior probability $P((ij)|D)$, given the data D, for a pair (ij) to interact directly can thus very naturally be quantified within our model by calculating the sum of the statistical weights of all spanning trees in which the edge between the pair (ij) exists. The calculation of this posterior is illustrated in Figure 2.6.

Note that in this calculation $P((ij)|D)$ depends on the statistical dependencies

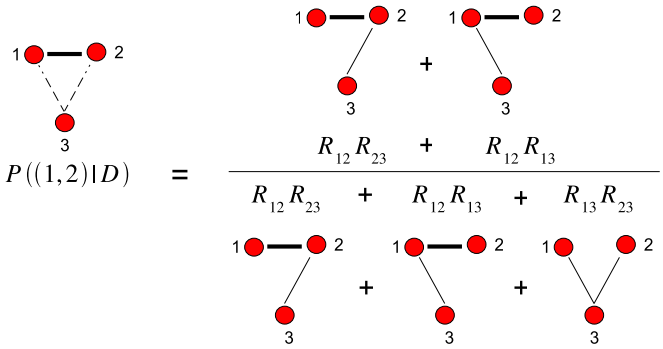


Figure 2.6: **Illustration of the calculation of the posterior probability**. For the sake of simplicity, we here show an example for an alignment with only 3 columns. The posterior probability for edge $(1, 2)$ is the statistical weight of all spanning trees that contain this edge relative to the weight of all possible spanning trees.

between all pairs of positions and that all possible spanning trees are included in the calculation. Roughly speaking, a high posterior $P((ij)|D)$ indicates that the edge (i, j) is included in most spanning trees that have high probability. In this way indirect dependencies are accounted for in a rigorous way, derived from first principles, and without any free parameters.

2.2.5 Posterior probabilities significantly improve contact predictions

To compare the performance of the traditional mutual information-based measurement with the predictions of our model, we calculated mutual information I_{ij}, our analogous measure $\log(R_{ij})$, as well as the posterior probabilities $P((ij)|D)$ for each pair of positions (ij) for each domain in our set of 2009 Pfam alignments with available three dimensional structure.

Different domains have widely varying widths and also widely varying numbers of sequences in the alignments. With regard to the former, it is well-known that the number of pairs that are in contact in three-dimensional protein structures increases with the length of the protein sequence. To compare prediction accuracies for proteins with different lengths, the consensus, also used by the CASP assessors [34], has been to compare the number of predictions per residue. However, although there is a large variation across domains, we find that the number of contacts scales slightly super-linearly, with an exponent of roughly 1.1 for all pairs of residues, and up to 1.6 if we consider only pairs of residues that are distal in the primary sequence (see Supplementary Figure 2.15). That is, the number of contacts per residue grows with the length of the domain, making it problematic to use predictions-per-residue as a common reference for domains of different length. We therefore decided to compare

prediction accuracies as a function of the number of predictions relative to the total number of contacts in the protein. In particular, we compare predictions for different proteins at the same *sensitivity*, i.e. the fraction of all true contacts that are predicted.

As mentioned previously, $\log(R)$ values typically increase with the number of sequences in the alignment and also depend on the phylogenetic distances of the sequences present in the alignment, such that $\log(R)$ values cannot be directly compared across different domains. Therefore, for each domain we produced three lists of predicted edges, one sorted by mutual information, one by $\log(R)$, and one by posterior probability $P((ij)|D)$. For different fractions x, we selected the top edges from each list such that the fraction of all true edges among the predictions (sensitivity) equals x, separately for each domain. For each value of x and all three measures, we then calculated the average positive predictive value, i.e. the fraction of all predicted edges that are truly in contact in the three-dimensional structure of the domain, by averaging over all domains. These results are shown in the left panel of Figure 2.7.

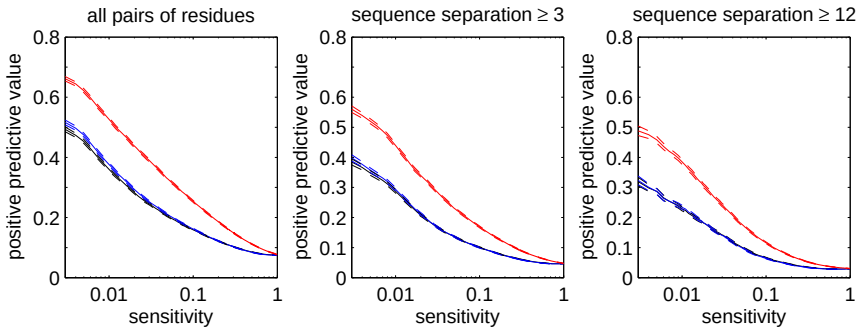


Figure 2.7: **Accuracy of contact predictions for all** 2009 **alignments**. Shown are the performances of mutual information (black), $\log(R)$ (blue), and the posterior probabilities (red). The vertical axis shows mean positive predictive value (PPV, solid line) plus and minus one standard error (dashed lines) as a function of sensitivity (horizontal axis, shown on a logarithmic scale). The left panel shows predictions for all residue pairs, the middle using only predictions for residues separated by at least 3 positions in the primary sequence, and the right panel for pairs separated by at least 12 positions.

Not surprisingly, residues that are close in the primary sequence are much more likely to contact each other in the structure than distant pairs, see [33] and figure 2.11 below. In particular, residues that are neighbors in the primary sequence are (by the definition used) *always* contacts and residues at distance 2 are contacting almost 90% of the time, whereas contacts between residues more distal in the primary sequence are relatively rare. Therefore, if one considers all contacts, the accuracy of the predictions is dominated by the large number of contacts between residues at primary sequence distances 1 and 2, which almost always exist, and are therefore not informative regarding protein structure. Therefore, the middle panel of Figure 2.7 shows the results when considering only pairs that are at least 3 residues apart

in primary sequence. In addition, following the practice established in the contact prediction literature, we also show results when considering only pairs at least 12 residues apart in primary sequence (Figure 2.7, right panel) and at least 24 residues apart (Supplementary Figure 2.16).

As expected, the accuracy of predictions for mutual information and $\log(R)$ are very similar and demonstrate that these two measures can be considered equivalent in this context (we will only refer to $\log(R)$ from hereon). Most importantly, Figure 2.7 shows that the predictions based on posterior probabilities (red curves) outperform the other methods by a large margin, i.e. with an almost 50% larger PPV at some sensitivities. This confirms that rigorous treatment of indirect dependencies strongly improves contact predictions. It should be noted, however, that at cut-offs where the positive predictive value is reasonably high, sensitivities are only on the order of one percent. It is thus clear that at high PPV, our method in its current form can only predict a minor fraction of all true interacting pairs, which is in accordance with results from previous studies [9, 13].

For completeness, we also considered the accuracy of prediction that would be obtained if, instead of summing over all possible spanning trees, we determine the maximum-likelihood tree and use only the links in this tree in our predictions, i.e. as done in [14]. As shown in Supplementary Figure 2.17, although this leads to an improvement over using $\log(R)$, the accuracy of the posterior probability measure by far outperforms the predictions based on the maximum-likelihood tree. This nicely demonstrates the value of summing over all possible spanning trees which is employed in the calculation of the posterior for a given edge.

2.2.6 The posterior removes indirect dependencies and predicts contacts with weaker statistical dependency

To demonstrate that our model successfully prevents the prediction of interactions between pairs with indirect dependency, we collected all distal pairs that showed significant statistical dependence ($Z > 4$) and ordered them by the score of the best co-evolving contact chain that can explain their statistical dependency, i.e. as shown in Figure 2.4. Figure 2.8 shows the reverse-cumulative distributions of the posteriors that these distal pairs obtain in our model for different cut-offs on the best path score S, as well as the distribution of posteriors of all contacting pairs with $Z > 4$.

First of all, we see that co-evolving contacts have dramatically higher posteriors than distal pairs in general, which confirms the improved accuracy of contact predictions that our method accomplishes. Moreover, we see that distal pairs that can be explained with the most strongly co-evolving contact chains, i.e. with the lowest scores S, obtain the lowest posterior probabilities. For example, less than 10% of the distal pairs with a chain at score $S = 0$ have a posterior larger than 0.2 and virtually

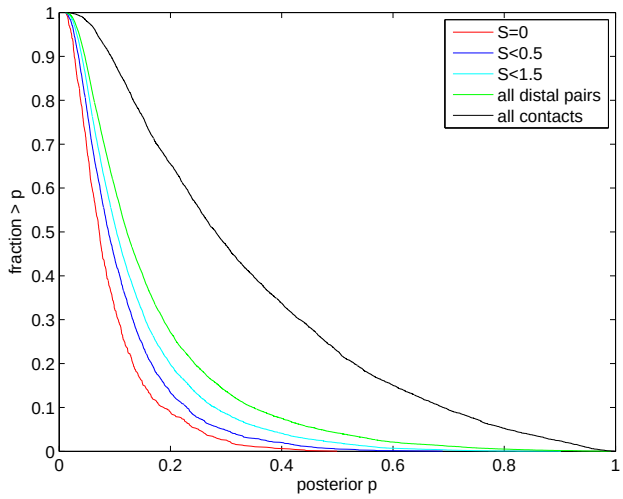


Figure 2.8: **Posteriors reflect the extent to which co-evolving pairs can be explained by contact chains**. Shown are the reverse cumulative distributions of the posteriors of distal co-evolving pairs ($Z > 4$) that can be explained by contact chains of scores $S = 0$ (red), $S < 0.5$ (dark blue), $S < 1.5$ (light blue), and for all distal co-evolving pairs (green). For comparison the reverse cumulative distribution of posteriors for co-evolving contacts ($Z > 4$) is also shown (black).

no pair has a posterior as large as 0.5. As the score S of the best chains increases, so generally do the posteriors. This confirms that the posterior as calculated by our model correctly captures the extent to which a statistical dependency is direct.

Instead of selecting all distal co-evolving pairs with contact chains below some score S, we also selected all co-evolving pairs with S scores larger than various cut-offs and determined the distributions of their posteriors. These distributions are shown in Supplementary Figure 2.18 and illustrate that distal co-evolving pairs with sufficiently large score S obtain posteriors comparable with those of co-evolving contacts. This suggests that the particular subset of distal co-evolving pairs that cannot be explained by any chain of contacts are likely true interacting residues, which may for example form contacts in the interaction surface of oligomers of the domain.

To further demonstrate that our Bayesian network model correctly distinguishes direct from indirect interactions, we also investigated the extent to which the posterior identifies structurally close pairs independent of the direct statistical dependency of the pair. We divided all pairs into bins according to their $\log(R)$ Z-value and calculated, for each bin, the distribution of structural distances of all pairs, and for the subset of pairs that have posterior probability larger than 0.2. Figure 2.9 shows, as a function of the Z-value of the pairs, the median, 25th, and 75th percentiles of the structural distance distributions of all pairs (blue) and those with posterior larger than 0.2 (red).

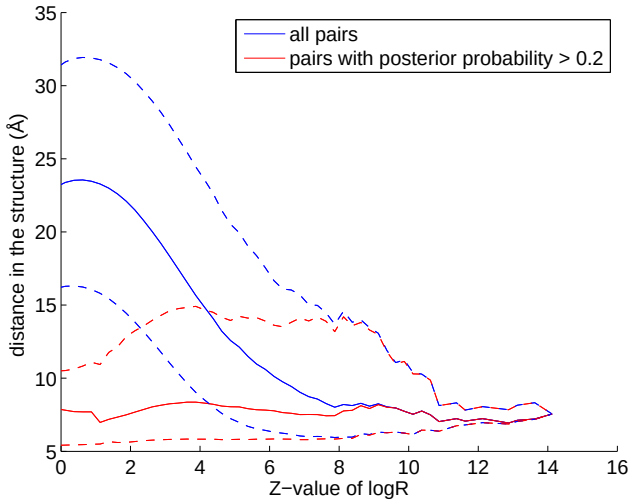


Figure 2.9: **The posterior predicts structurally close pairs independent of their direct statistical dependence**. The structural distance distribution (vertical axis) is shown for all pairs (blue) and for pairs with posterior probability larger than 0.2 (red) as a function of the Z-value of the $\log(R)$ statistic (horizontal axis). The solid lines show the medians of the distributions and the dashed lines the 25th and 75th percentiles.

At large Z-values the red and blue curves are essentially identical. In this regime, we are only looking at the most strongly dependent residues in each alignment and any spanning tree of high likelihood must contain edges between these pairs of residues, i.e. almost all of these edges have high posterior probabilities. However, already at Z-values as high as 8, the median distance of all pairs starts to increase rapidly, from roughly 8 to more than 20 at Z-value 0. This illustrates again that even at very high values of $\log(R)$ a substantial fraction of pairs are distal in the structure. In contrast, the subset of residues with high posterior probability remains close over the whole range of Z-values, down to Z-values of almost 0. In fact, strikingly, there is very little change in the distribution of structural distances for Z-values from 0 to 8. This is very significant because it demonstrates that, independent of the amount of direct statistical dependency between a pair of positions, a high posterior is indicative of close structural distance. Moreover, it demonstrates that our Bayesian network model can detect truly interacting pairs of residues even if they show only a small amount of statistical dependency.

2.2.7 The Bayesian network model with phylogenetic correction significantly outperforms existing methods

One of the key problems in contact prediction is the large number of distal pairs with high statistical dependency. In the foregoing sections we have shown that many of these distal co-evolving pairs are indirect, induced by chains of dependencies between contacting residues, and we have shown that our Bayesian network model can rigorously disentangle direct from indirect dependencies, thereby greatly improving contact predictions. In the remaining sections we develop a number of extensions of our basic method to further improve the predictions.

As mentioned in the introduction, the phylogenetic relationships of the underlying sequences is a major confounding factor when determining the statistical dependency between several residues (nicely explained in eg [8,12]) and it is a difficult task to 'subtract' from the apparent statistical dependency between two residues the part that is purely due to phylogeny. The best way to address this difficulty would of course be to construct a phylogenetic tree of all sequences in the multiple alignment and to explicitly model the evolution of the sequences along the tree, using an evolutionary model that takes dependencies between positions into account. Unfortunately, it appears that such a rigorous approach is computationally intractable for several reasons. First, one would either have to accurately reconstruct the phylogenetic tree, which is very challenging for large sets of sequences, or sum over all possible trees, which is computationally infeasible. The second issue is the evolutionary model. In our Bayesian network model, the conditional probabilities $P(x_i|x_j)$ are different at every pair (ij), introducing 380 parameters per pair, which are integrated over. However, for the evolutionary case analytic integration is no longer possible, which makes such models intractable. Indeed, models that treat dependencies between residues in an explicit phylogenetic setting [11, 14] consider much simpler evolutionary models in which only correlations in the overall *rates* of mutations at different positions are considered and not the specific identities of the mutations.

As an alternative to explicit phylogenetic methods, recently a number of simple *ad hoc* phylogenetic corrections have been proposed, which do not involve a reconstruction of the phylogenetic tree, which can be efficiently calculated, and which clearly improve contact predictions [12, 13]. One of these corrections, the so-called *average-product correction* APC has been shown to provide the most accurate contact predictions [13]. It is based on the idea that the statistical dependency between every pair of columns is the sum of a true statistical dependency and a background dependency due to the phylogenetic relationships. In the APC it is assumed that the background dependency is a product of independent factors associated with the two positions. Since a given position will interact with only a small fraction of other positions, the background dependencies can be estimated by calculating, for each column, its average statistical dependence with all other columns. The background

dependence for each pair is then subtracted to obtain a corrected statistical dependency. As described in *Materials and Methods*, we adapted the APC to our Bayesian model, essentially replacing $\log(R)$ with a corrected version $\log(R^c)$ that subtracts out the background dependency. These $\log(R^c)$ values can then be used, analogously to $\log(R)$ values, to determine corrected posterior probabilities (see *Materials and Methods*).

In Figure 2.10, we show the accuracy of our predictions using the corrected posterior probabilities (in blue) and compare it with predictions based on mutual information using the average-product correction APC (in black). The latter has been recently shown to outperform other existing methods [13]. The red curves show the performance of the method without the phylogenetic correction, i.e. as was shown in Figure 2.7. It is clear that the predictions based on posterior probability combined with the phylogenetic correction significantly outperform the current best methods. For example, considering pairs at primary sequence separation at least 3, the sensitivities at PPV of 0.5 are 0.5% for the uncorrected posterior, about 1% for the APC, and about 2% for the corrected posterior. The clear improvement in prediction accuracy is also evident for pairs with primary sequence separation of at least 24 amino acids (Supplementary Figure 2.19).

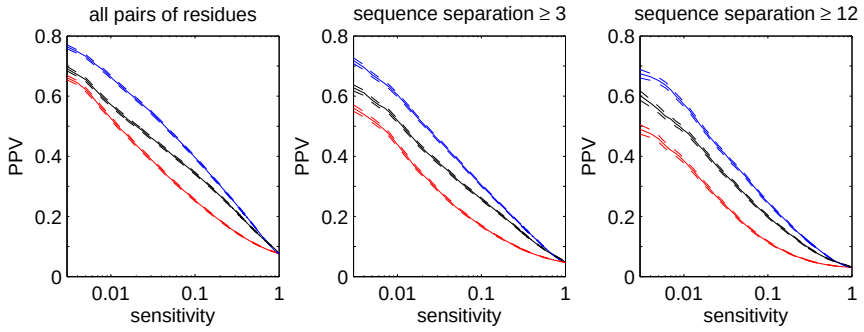


Figure 2.10: **Improved accuracy of contact predictions when a phylogenetic correction is included**. In blue, we show the performance of the phylogenetically-corrected posterior probabilities, in black the performance of the predictions based on the average-product corrected (APC) mutual information [13], and in red the performance of the posterior probability without phylogenetic correction. Curves were calculated as in figure 2.7.

Although Figure 2.10 combines results of the predictions on protein domains of differing sizes, the fact that the true interactions are a much smaller fraction of all possible interactions for long sequences makes the prediction task significantly harder for long sequences, see e.g. [38]. In Supplementary Figures 2.20, 2.21, 2.22 and 2.23, we show the performance of the various methods separately for short, medium length, and long sequences. We find that, independent of the length of the sequences, our method clearly outperforms current methods.

2.2.8 Co-evolution of residue pairs is independent of primary sequence separation

In protein structure prediction, where prediction of contacts at large sequence separations is particularly important [34], it is well-known that contact prediction accuracy generally decreases with increasing sequence separation ([33, 34], also seen in figure 2.10). This is a direct consequence of the fact that the fraction of contacts decreases rapidly as a function of sequence separation (roughly as $1/d$, where d is the primary sequence separation, see the left panel in figure 2.11), which makes the prediction problem much more difficult for contacts at large primary sequence separations. Vice versa, because contacts at large primary distances are rare, they are most informative for protein structure prediction [34].

The left panel of Figure 2.11 shows that there are several regimes in the distribution of contact-density at different primary sequence distances. First, residues at distance 1 and 2 are almost always contacts and thus contain very little information about protein structure. In contrast, at distances 3 and 4 the fraction of contacts has already dropped to roughly 50%, i.e. about 1 bit of information per contact, and the fraction then drops quickly, reaching about 5% at primary sequence separation 10. For distances between 10 and 30 the fraction stays roughly constant at 5% and for even larger distances it drops approximately as $1/d$.

Clearly, the information contained in Figure 2.11 regarding protein structures can be used to improve contact prediction, i.e. by assigning prior probabilities to different contacts based on their distance in primary sequence. However, before pursuing this we ask to what extent contacts at different primary sequence distances show statistical evidence of co-evolution. The almost ubiquitous contacts at primary sequence distances 1 and 2 are probably mainly the result of geometrical constraints, the contacts at intermediate distances are likely often part of the same secondary structure, and the very distal contacts might correspond to contacts between different secondary structure elements. Given the different nature of these contacts at different primary sequence separations, one might expect very different distributions of statistical dependencies, and this would clearly affect contact prediction.

To investigate this, we determined the distribution of the Z-values of corrected $\log(R^c)$ for all *contacts* at each primary sequence separation d (Figure 2.11, right panel). Interestingly, the distribution of statistical dependencies is almost *constant* across the entire range of primary sequence distances. The only significant deviation is a slight peak at sequence separation 4, corresponding to residues on the same side of alpha helices ([39] and data not shown), which apparently have slightly increased statistical dependency compared to other contacts. However, far more important for the purpose of predicting protein structure is that, with regard to the statistical dependency between alignment columns, all contacts appear to be essentially equal, so that the evidence of statistical dependency between residues can

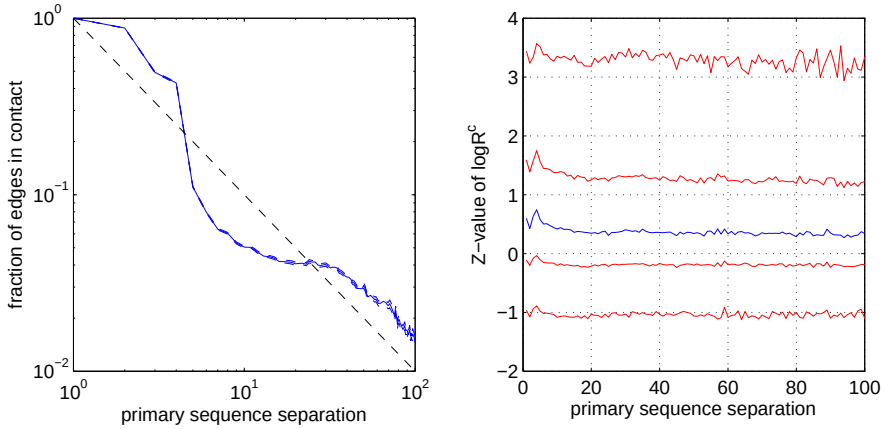


Figure 2.11: **Occurrence of contacts and co-evolution as a function of primary sequence separation**. Left panel: The fraction of residue pairs that are in contact in the structure as a function of primary sequence separation d. The solid blue line shows the mean, the dashed blue lines the mean $\pm$ one standard error. The dashed black line shows the function $1/d$. Right panel: The Z-value distribution of the $\log(R)$ statistics for all contacting pairs at different primary sequence separations. The blue line represents the median and the red lines represent the 5th, 25th, 75th and 95th percentiles, respectively. The Z-value was calculated with respect to the mean and standard deviation of the $\log(R)$ distribution of all pairs (including distal ones). In both panels only sequence separations up to 100 residues are shown as the curves become very noisy for larger sequence separations.

be treated completely independently of the prior information regarding which contacts are more or less likely to exist based on general structural considerations. From a biological and evolutionary perspective this result shows that, interestingly, different 'types' of contacts apparently lead to similar evolutionary constraints.

2.2.9 Influence of entropy on contact prediction

An important, but poorly understood issue in covariation-based contact prediction is the influence of conservation on prediction accuracy. The 'conservation' shown by a position in a multiple alignment can be most generally quantified by the entropy of the amino acid distribution in the column. It is well known that this column entropy can vary immensely along protein sequences, most probably due to functional and structural constraints. One would intuitively expect that a position that is contacting many other residues would generally have to satisfy more constraints and would thus be expected to show relatively low entropy.

To investigate this, we calculated, for each position in each domain, the column entropy and the number of contacts of the corresponding residue. As shown in the left panel of Figure 2.12 there is indeed a clear negative correlation between the column entropy and the number of contacts. For very low entropies, i.e. less than 1, the average number of contacts is constant and approximately 10.5. As the entropy increases from 1 to about 2.75 (which is close to the entropy of a uniform distribution of amino acids) the average number of contacts drops to almost 6. That is, very low entropy columns have on average almost twice as many contacts as high entropy columns. Since the number of residues in a sphere of 8 around the C_β atom of an amino acid (which is exactly our definition of a contact) is commonly used as a measure for how strongly a residue is buried in the core of the protein (e.g. [40]), the left panel of Figure 2.12 reiterates the well-known dependence between surface accessibility and conservation [41].

It is well appreciated in the literature that the variation of entropy across positions has important effects on predictions based on statistical dependencies. For example, a comparative study of different prediction methods has shown that commonly used co-variation measures differ in their sensitivity to per-site variability and generally, each method has highest accuracy within its specific preferred range of variability [9]. In analogy to our analysis of statistical dependency as a function of distance in primary sequence (Figure 2.11, right panel), we investigated how the statistical dependency that different contacts exhibit depends on the column entropies of the residues. As before, we transformed the $\log(R)$ values to Z-values and determined the Z-value distribution of all contacts as a function of the sum of the entropies of the corresponding columns (Figure 2.12, blue lines). We see that contacts indeed show a strong correlation between the sum of column entropies and statistical dependency. For low entropy columns the Z-values are mostly negative, and they become only

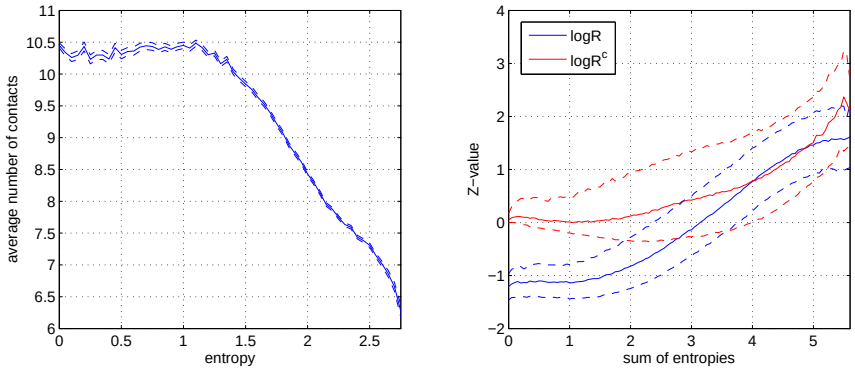


Figure 2.12: **Contact-degree and co-evolution as a function of positional entropy** Left panel: Average number of contacts of a residue (solid line) as a function of the entropy of its alignment column. The dashed lines denote mean ± one standard error. The right panel shows the Z-value distribution of both $\log(R)$ (blue) and $\log(R^c)$ (red) for all contacting pairs versus the sum of entropies of the corresponding columns. The solid lines denote the medians and the dashed lines the 25th and 75th percentiles.

positive at an entropy sum of about 3. It is thus clear that contact predictions that use mutual information ($\log(R)$) will preferentially predict contacts between residues of high entropy columns.

That mutual information and $\log(R)$ is low for contacts with low entropy columns is to a certain extent unavoidable. It is a basic result of information theory [25] that the mutual information between two variables cannot be larger than the minimum of the marginal entropies of the two variables. Intuitively, one could imagine a position that is so constrained by its function and its many contacts that only a single amino acid is viable at the position. Obviously, since this position shows no variation whatsoever it cannot display any signs of statistical dependency with any other column, even though it may contact many other residues. This is a basic limitation of using statistical dependency for contact prediction that cannot be avoided. However, it has been argued that modified versions of mutual information, such as the product or sum correction [13], besides correcting for the phylogenetic background signal, are also able to better identify co-evolution between less variable residues. The red lines in the right panel of Figure 2.12 show the mean and standard deviation of the Z-values of product-corrected statistical dependency $\log(R^c)$. We see that indeed, the correlation between the Z-values and the sum of column-entropies is significantly reduced when using $\log(R^c)$, and low entropy contacts no longer show negative Z-values on average.

Still, a clear correlation between the column-entropy sum and the statistical dependency remains even for $\log(R^c)$. On the one hand this may be the result of the inherent inability to 'detect' statistical dependency when columns are very conserved.

On the other hand, it is also conceivable that those positions that have low entropy, and that form many contacts, may generally show weaker statistical dependency *per contact*. For example, it could be argued that hydrophobic residues that lie in the core of the protein and thus contact many other residues are less variable because they need to remain on the interior and therefore do not allow for changes towards non-hydrophobic residues. Such residues may not be constrained so much by their contacting residues, but rather by the necessity to stay away from the solvent-exposed protein surface, leading to relatively weak statistical dependencies with the contacting residues.

2.2.10 Incorporation of prior information improves prediction accuracy

So far our Bayesian method assumes that a contact between any pair of positions is a priori equally likely. However, as seen in the previous sections, the probability for a contact to occur depends strongly on the primary sequence distance between the residues and the column-entropies of the residues. We therefore developed an 'informative prior' which makes the prior probability for a contact to occur depend on both of these variables. For a given pair of positions, let d be the distance in the primary-sequence of the two positions, and let H denote the sum of the column-entropies of these positions. As described in *Materials and Methods*, we estimated the fractions $f(d, H)$ of pairs at sequence distance d and entropy-sum H that are contacts and using these fractions constructed prior probability distributions that can be easily incorporated into our method.

Figure 2.13 shows the results of the contact predictions performed with our Bayesian network model incorporating the informative prior and using posterior probabilities (blue lines). For comparison the results using posteriors based on $\log(R^c)$ (the blue lines in Figure 2.10) are shown as well (red lines). We see that, for the set of all pairs, and all pairs that are at least $d \geq 3$ apart in primary sequence, the incorporation of the prior probability dramatically improves the predictions. For example, looking at all pairs, our method can predict roughly 40% of all existing contacts at a positive predictive value of 80%. If we restrict ourselves to non-trivial contacts, i.e. those with primary-sequence distance $d \geq 3$, we find that at a positive predictive value of 50% our method reaches a sensitivity of roughly 20%. For comparison, without the prior an approximately 10 times lower sensitivity is reached at the same positive predictive value.

Somewhat surprisingly, we find that the quality of the predictions for distal pairs $d \geq 12$ is slightly reduced by the incorporation of the prior, especially at low sensitivities. We speculate that this is a result of the fact that we constructed the prior distribution assuming that $f(d, H)$ is independent of the length of the domain itself.

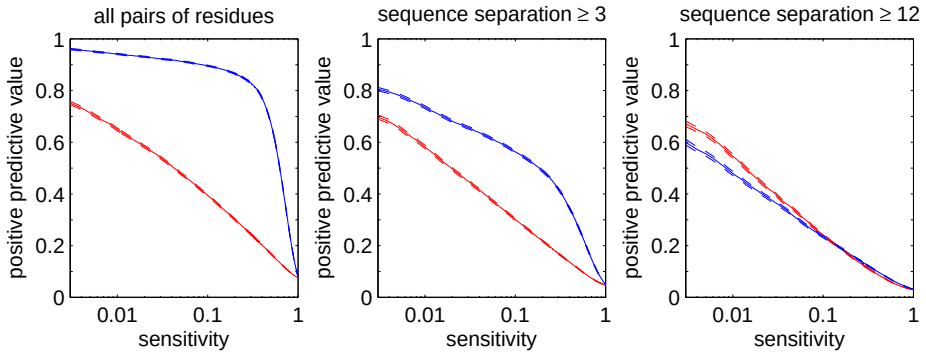


Figure 2.13: **Improved accuracy of contact prediction when an informative prior is included.** In blue, we show the performance of the posterior probabilities that take primary-sequence separation and column entropy into account. For comparison we show in red the performance of the posteriors with phylogenetic correction but uniform prior, which are the same as the blue lines in Figure 2.10.

This approximation breaks down most significantly when focusing on distal pairs because, whereas contacts at short primary distances occur in all domains, contacts at long primary distances are more common in long domains. However, it should be noted that, given that contacts at this primary-sequence distance are rare, one would most likely need to perform predictions at reasonably high sensitivity, i.e. 10% or more. In this regime, the performance with prior is comparable to or even a tiny bit better than without prior.

2.3 Discussion

One of the key problems in using co-evolution analysis to predict residue contacts is that so many structurally distal pairs show strong statistical dependencies [13, 14, 31]. A number of reasons have been proposed to explain this fact. One explanation is that sequences in multiple alignments are generally phylogenetically related and these phylogenetic relationships can induce strong apparent statistical dependencies between many pairs of columns. Although there is of yet no computationally tractable way for treating the phylogenetic dependencies in a rigorous manner, i.e. by explicitly modeling the evolution of the sequences including arbitrary dependencies, several procedures have been proposed that can correct at least for the main phylogenetic signal [6, 8, 13, 14]. Indeed the application of such methods has been shown to very significantly improve contact predictions [8, 13, 14].

Still, even with the current best phylogenetic corrections, strong statistical dependencies remain evident between many structurally distal pairs. One proposed explanation that has received little attention in the contact prediction literature is that statistical dependencies between distal pairs can be induced by the percolation

of statistical dependencies along chains of co-evolving contacts [21, 22]. Here we have shown that such chains of co-evolving contacts are indeed pervasive across all protein domains and that they explain many if not most of the distal co-evolving pairs. Statistical analysis shows that these chains travel on average 3.25 ± 0.05 per contact, and that the total distance covered by these chains is exponentially distributed with an average of 16, corresponding to a chain that consists of 5 contacts. Note that, whereas residues up to 8 apart are generally considered contacts, our results strongly suggest that the typical distance between co-evolving contacts is only 3.25. Another interesting observation is that, although it is likely that contacts between residues at different distances in primary sequences are different in nature, our analysis shows that the statistical dependency shown by contacts is completely independent of their primary-sequence separation. This is an important insight because it demonstrates that co-evolutionary analysis is equally informative about close and distal contacts.

We have adapted our recently evolved Bayesian network model [37] in order to assign, to any pair of positions, a posterior probability that they interact directly. This posterior probability rigorously takes into account all possible ways in which the statistical dependence between the pair can be explained in terms of chains of other co-evolving pairs. Analysis of the predictions of this model shows that it correctly detects distal pairs that can be explained by co-evolving contact chains, and that it also allows one to detect true interacting pairs that have only weak direct statistical dependency.

Recently Halabi et al [42] have shown that, by a spectral analysis of the matrix of statistical dependencies between positions, one can identify so called 'protein sectors': sets of positions that co-evolve significantly with each other, but that are relatively independent of the positions in other sectors. Since in [42] a rather simple measure of direct statistical dependency is used, we speculate that a much more accurate identification of protein sectors could be obtained by using statistical dependencies as assessed by our posterior probabilities.

While finishing the work in this study, a paper appeared that also aims to disentangle direct from indirect interactions [16]. Like our approach, [16] models the joint probability of sequences in the multiple alignment in terms of a set of pairwise interactions. What is appealing about the approach of [16] is that it is based on the more 'physical' assumption that an interaction energy is associated with each pairwise interaction such that a total interaction energy can be calculated for each sequence, and that the probability to observe a particular sequence is given simply by the Boltzmann distribution in terms of this total energy. However, the great disadvantage of this model is that its solution requires a heuristic approximation and is computationally very expensive to calculate. For example, in [16] the authors were forced to restrict themselves to only 60 positions in the alignment, and even then the calculations for a single alignment took several days. Therefore, an application of the approach of [16] on as large a scale as in this work, with thousands of multiple

alignments of up to several hundred positions, is not feasible. In addition, it is not clear how the approach of [16] could accommodate a phylogenetic correction, which would be necessary to obtain a competitive performance with this method.

Although the disentangling of direct and indirect statistical dependencies strongly improves contact predictions, and incorporating a phylogenetic correction further improves the performance, the predictions are still far from perfect. In particular, at reasonably high positive predictive value the sensitivity amounts to less than 10% of all true contacts. Although it is clear that contact predictions based only on statistical dependencies could be further improved, for example by a more rigorous treatment of the phylogenetic dependencies, we believe that it is unlikely that such improvements would dramatically enhance the performance. First of all, simple inspection of the data shows that a large number of the pairs that are contacts in the sense that they are less than 8 apart, really show no sign of co-evolution at all. That is, a large fraction of 'contacts' may simply not interact directly, and these obviously can never be detected using statistical dependence measurements. On the other end of the scale are residues that contact so many others that they are very strongly constrained, and show almost no variability in evolution. For such highly conserved residues it is also inherently impossible to identify their interaction partners using co-evolutionary analysis.

We thus believe that the largest further improvements to contact prediction are to be expected from incorporating information other than statistical dependency. To illustrate that additional information can be easily incorporated into our model, we developed an informative prior that takes into account that the likelihood of a contact to exist depends on the primary-sequence distance of the residues, and that highly conserved residues tend to have a higher number of contacts. The incorporation of even this simple additional information already leads to dramatic improvements in contact prediction. Clearly more powerful priors could be developed that take into account more sophisticated structural knowledge. In addition, in our current method we integrate over all possible joint probabilities for pairs of interacting residues, effectively assuming that all possible joint probability distributions are equally likely. Here too improvements could likely be made by taking into account prior knowledge on which joint probability distributions are more or less likely for interacting pairs of amino acids. Ultimately the most satisfying approach would be to combine our approach with direct structural modeling, i.e. somewhat along the lines of the approach taken in [43].

Following the plausible intuition that, the more different kinds of information are taken into account, the greater the prediction accuracy that can be obtained, several machine learning and statistical methods have been proposed that incorporate a much larger number of different features (see [33, 43, 44] and references therein). Besides primary sequence separation and conservation, these methods include features such as domain length, relative solvent accessibility, predicted secondary structure, the amino acid composition in short windows around the positions of interest, chem-

ical properties of the amino acids, and contact potentials. Due to varying training and test sets and varying standards of evaluation, it is very difficult to compare the performance of our method with these approaches. However, some principal differences between these methods and ours should be noted. First, all these methods rely on training sets to fit parameters, so that additional methods are required to avoid over-fitting, whereas our method is essentially without any tunable parameters and does not require any training sets. Second, some of these methods are rather *ad hoc* 'black box' methods, e.g. neutral networks [33] or support vector machines [44], that use partially redundant sets of features, from which it is typically hard to derive mechanistic insights. In contrast, our method is derived directly from first principles. In any case, the results that we have presented show that it is crucial to take indirect dependencies into account when incorporating co-evolution information. We have provided a rigorous method for doing so and it is clear that any contact prediction method that incorporates co-evolution information would strongly benefit from using our method for disentangling direct and indirect dependencies.

Whereas we have here applied our method to predict contacting residues in a single protein, it is straight forward to use the same method for predicting contacting residues between pairs of proteins that are known to interact. That is, given two set of orthologs proteins s_1 and s_2, for which it is known that each member of set s_1 interacts with the corresponding member of set s_2, we can simply concatenate the multiple alignments of s_1 and s_2 into one longer multiple alignment, and apply our method to this longer alignment.

More generally, our method provides a computationally tractable extension of weight matrix models to take into account arbitrary pairwise dependencies, and there are a number of more general applications that we envisage pursuing in the future. First, our method can be generally used to 'score' multiple alignments in a way that includes pairwise dependencies. This could be used to discover subfamilies within large multiple alignments or to generally refine multiple alignments. Since the performance of alignment-based contact prediction methods is expected to depend strongly on the quality of the alignments, such a refinement may further improve contact prediction. Finally, another attractive application is to develop a regulatory-motif finding algorithm that takes into account arbitrary pairwise dependencies between positions.

2.4 Materials and Methods

2.4.1 Domain sequences and structures

Domain alignments and the mappings from domains to available structures in the PDB database were downloaded from the Pfam database [23, 45]. We only used

Pfam A, which is the high-quality and manually curated part of Pfam [23]. For each Pfam domain with at least one known structure, we reduced the alignment to positions corresponding to match states of the corresponding Pfam hidden Markov model with no more than 20 percent gaps. The removal of columns with many gaps is necessary as gaps can cause spurious correlations (see below) and make it difficult to compare the phylogenetic background signal between different columns. We removed from each alignment all multiple copies of identical sequences as well as sequences that had more than 50 percent gaps with respect to the match states. Additionally, alignments containing less than 100 sequences or less than 50 columns were discarded. To keep computational times limited we also removed alignments with more than 400 columns. For each Pfam alignment, all corresponding PDB files were collected according to the iPfam annotation [45] and distances between pairs of residues were determined as the distance between the C_β atoms (C_α for glycines). In the case of NMR models, the minimal distances of all models contained in the PDB entry were chosen. If a Pfam domain was present in multiple protein structures or in several chains of one protein structure, we chose the median distance over all chains and structures. For some alignments the corresponding structure did not cover all columns in the alignment and we discarded the small number of examples where the coverage was less than 50%. This resulted in 2009 domains with structurally-defined distances between residues. Finally, distance in primary sequence was defined as the distance between the match states of the alignment.

2.4.2 Probabilistic model

Our Bayesian network model was described in detail in [37]. Briefly, given a single column i of the alignment with observed amino acid counts n_α^i, the probability $P(D_i|w^i)$ of the column is given in terms of the (unknown) probability distribution w^i, with w_α^i the probability that letter α occurs at position i, i.e. $P(D_i|w^i) = \prod_\alpha (w_\alpha^i)^{n_\alpha^i}$. Using a Dirichlet prior for w^i with parameter λ, we obtain the marginal probability of the column $P(D_i)$ by integrating over all possible distributions w^i. This integral can be performed analytically and the result can be expressed in terms of gamma functions:

$$P(D_i) = \frac{\Gamma(20\lambda)}{\Gamma(n+20\lambda)} \prod_\alpha \frac{\Gamma(n_\alpha^i + \lambda)}{\Gamma(\lambda)}, \tag{2.2}$$

where n is the number of sequences in the alignment. Similarly, the *joint* probability of the data D_{ij} in a pair of columns (ij) is given in terms of the number of times $n_{\alpha\beta}^{ij}$ that the combination of letters $(\alpha\beta)$ occurs at positions (ij), i.e.

$$P(D_{ij}) = \frac{\Gamma(20^2\lambda')}{\Gamma(n+20^2\lambda')} \prod_{\alpha\beta} \frac{\Gamma(n_{\alpha\beta}^{ij} + \lambda')}{\Gamma(\lambda')}. \tag{2.3}$$

Here, we set the parameter λ' of the Dirichlet prior for the joint probability distribution to 0.5. As shown in [28], in the context of a dependence tree model, consistency requires that λ equals $20\lambda'$.

The statistical dependence between columns i and j is quantified by the ratio

$$R_{ij} = \frac{P(D_{ij})}{P(D_i)P(D_j)}. \tag{2.4}$$

The connection of $\log(R)$ to mutual information is easily established by substituting equations (2.2) and (2.3) into the logarithm of R as given by (2.4) and using Stirling's approximation to the logarithm of the gamma function. We then find that approximately

$$R_{ij} \propto e^{nI_{ij}} \tag{2.5}$$

for large n, with I_{ij} the mutual information between columns i and j. Importantly, when determining the counts $n^{ij}_{\alpha\beta}$ and n^i_α in order to determine R_{ij}, we discard all pairs of residues within a given sequence where either α or β is a gap. Treating gaps as a 21 amino acid causes strong spurious correlations between residues that are close in primary sequence since gaps usually come in blocks (data not shown).

A *dependence tree* π specifies for each position i (except for the root of the tree) a parent position $\pi(i)$ which is the residue that i depends on. To keep the notation simple, we here use the symbol π to both denote the mapping from a node to its parent node and the dependence tree itself. It can be shown [37] that, given a dependence tree, the joint probability $P(D|\pi)$ of the entire alignment can be written as

$$P(D|\pi) = \prod_i P(D_i) \prod_{j \neq r} R_{j\pi(j)}, \tag{2.6}$$

where the first product goes over all positions and the second over all positions except for the root r.

Finally, the probability $P(D)$ of the whole alignment is given by summing over all possible dependence trees π

$$P(D) = \prod_i P(D_i) \left(\sum_\pi P(\pi) \prod_{j \neq r} R_{j\pi(j)} \right), \tag{2.7}$$

where $P(\pi)$ is the prior probability of a particular spanning tree π. The last product is in fact the product of the R-values over all edges of the tree given by π and is independent of the choice of the root. If the prior probability of a spanning tree can be written as a product of probabilities $W_{j\pi(j)}$ along each edge $(j, \pi(j))$ of the tree

$$P(\pi) = \prod_{j \neq r} W_{j\pi(j)} \tag{2.8}$$

then equation (2.7) can be rewritten as

$$P(D) = \prod_i P(D_i) \left(\sum_\pi \prod_{j \neq r} M_{j\pi(j)} \right) \tag{2.9}$$

with $M_{j\pi(j)} \doteq R_{j\pi(j)} W_{j\pi(j)}$. Thus, the weight of each edge is simply multiplied by its prior probability. The largest term in the sum of equation (2.9) is the *maximum spanning tree* when a weight $\log(M_{ij})$ is assigned to each edge (ij) and this maximum spanning tree can be easily determined [24].

The sum over spanning trees in (2.9) can be calculated using a generalization of Kirchhoff's matrix-tree theorem [28]. For this we need to calculate the Laplacian of the matrix M_{ij}, which is defined as

$$L_{ij} = \delta_{ij} (\sum_k M_{ik}) - M_{ij} \tag{2.10}$$

where the sum goes over all columns (or rows) of the M-matrix and δ_{ij} is the Kronecker delta function, which is one if $i = j$ and zero otherwise. We can then write the sum over all spanning trees as

$$\sum_\pi \prod_{i \neq r} M_{i\pi(i)} = \det(Q(L)) \tag{2.11}$$

where $Q(L)$ is the matrix L with one line and column removed (the determinant is independent of which line and column are removed). The summation over all spanning trees (there are n^{n-2} spanning trees for a full graph with n nodes) thus reduces to the calculation of a determinant, which can be done in a time proportional to n^3.

As discussed previously [37], the calculation of the determinant of the matrix M_{ij} is numerically very challenging since the entries M_{ij} vary over many orders of magnitude. In order to circumvent this problem, we rescale the entries of the matrix as suggested in [46]:

$$M_{ij} \to \beta \, (M_{ij})^\alpha \tag{2.12}$$

with $\alpha = \frac{K \log(10)}{\mathrm{logM_+} - \mathrm{logM_-}}$ and $\beta = -K \log(10) \frac{\mathrm{logM_+}}{\mathrm{logM_+} - \mathrm{logM_-}}$ where $\mathrm{logM_+}$ ($\mathrm{logM_-}$) is the logarithm of the maximal (minimal) entry of the matrix M_{ij}. This function maps all M values into the interval $\left[10^{-K}, 1\right]$, preserves the relative ordering of entries and does not exaggerate relative differences in belief [46]. The lower bound 10^{-K} ensures that the rescaled M-matrix remains numerically non-singular. K can be set according to the numerical precision of the machine and we set $K = 5$. We then use these rescaled M-values to calculate the posterior probabilities.

2.4.3 Calculating posteriors

Using expression (2.7), the posterior probability of a particular edge (kl) is given by

$$P((kl)|D) = \frac{P_{kl}(D)}{P(D)} \tag{2.13}$$

where

$$P_{kl}(D) = \prod_i P(D_i) \left(\sum_{\pi:(kl)\in\pi} \prod_{j\neq r} M_{j\pi(j)} \right) \tag{2.14}$$

which is the sum of the probabilities $P(D|\pi)P(\pi)$ for all spanning trees π that contain the edge (kl). This expression can be calculated by replacing the set of n nodes with a set of $(n-1)$ nodes, in which nodes k and l are contracted to one node, say kl, and the edge weights of this new node kl are given by $M_{kl,f} = M_{k,f} + M_{l,f}$ for all nodes $f \neq k, l$ [47]. Using this construction we can write the sum over all spanning trees containing edge (kl) as

$$P_{kl}(D) = \prod_i P(D_i) \left(M_{kl} \sum_{\pi'} \prod_{j\neq r} M_{j\pi'(j)} \right) \tag{2.15}$$

where the sum now goes over all spanning trees π' of the $(n-1)$ nodes. This sum over spanning trees can of course also be calculated as a determinant as described above. Roughly speaking, an edge (kl) will have high posterior if it occurs in the large majority of all spanning trees π that have high probability $P(D, \pi)$.

2.4.4 Phylogenetic correction

Due to the phylogenetic relatedness of the sequences in the alignment, there typically will be a statistical dependence between residues even in the absence of a functional linkage of these positions. Previous work [13] showed that this dependence can be corrected for (to some extent) by assuming that, due to phylogenetic relationships, each position has a certain amount of 'background' statistical dependence with other columns. Since each position interacts only with a small fraction of all other positions this background dependence can be estimated by calculating the average mutual information of that position with all the remaining positions. In [13], two types of corrections were proposed, a multiplicative one, named APC, and a additive one, named ASC. We here briefly review the derivation of these corrections.

The idea of the ASC is that the mutual information I_{ij} between positions i and j is the sum of the true mutual information I_{ij}^{true} and background mutual informations B_i and B_j, associated with positions i and j, i.e.

$$I_{ij} = I_{ij}^{\text{true}} + B_i + B_j. \tag{2.16}$$

We define average mutual informations as

$$\langle I_{i\cdot}\rangle = \frac{1}{m}\sum_{j=1}^{m} I_{ij}, \tag{2.17}$$

with m the number of columns of the alignment. Other averages like $\langle I_{\cdot\cdot}\rangle$, $\langle B\rangle$, and so on, are defined analogously. Note that, for notational simplicity, in these averages we have adopted the convention that $I_{ii} = 0$. We can then derive the equalities

$$\langle I_{\cdot\cdot}\rangle = \langle I^{\text{true}}_{\cdot\cdot}\rangle + 2\langle B\rangle, \tag{2.18}$$

and

$$\langle I_{i\cdot}\rangle = \langle I^{\text{true}}_{i\cdot}\rangle + B_i + \langle B\rangle. \tag{2.19}$$

If one assumes that, since true interactions are relatively rare, the averages $\langle I^{\text{true}}_{\cdot\cdot}\rangle$ and $\langle I^{\text{true}}_{i\cdot}\rangle$ are much smaller than $\langle B\rangle$, we can set $\langle I^{\text{true}}_{\cdot\cdot}\rangle \approx 0$ and $\langle I^{\text{true}}_{i\cdot}\rangle \approx 0$ and have

$$\langle B\rangle = \langle I_{\cdot\cdot}\rangle/2, \tag{2.20}$$

and

$$B_i = \langle I_{i\cdot}\rangle - \langle I_{\cdot\cdot}\rangle/2. \tag{2.21}$$

Finally, under these assumptions the true mutual information I^{true}_{ij} is then given by

$$I^{\text{true}}_{ij} = I_{ij} - \langle I_{i\cdot}\rangle - \langle I_{j\cdot}\rangle + \langle I_{\cdot\cdot}\rangle. \tag{2.22}$$

Motivated by this derivation, the ASC is defined as

$$I^{c}_{ij} = I_{ij} - \langle I_{i\cdot}\rangle - \langle I_{j\cdot}\rangle + \langle I_{\cdot\cdot}\rangle. \tag{2.23}$$

In the product correction APC we assume that the background mutual information between i and j can be written as a *product* of contributions of the two columns, i.e.

$$I_{ij} = I^{\text{true}}_{ij} + B_i B_j. \tag{2.24}$$

Assuming again that the true average mutual informations are small we find

$$\langle B\rangle^2 = \langle I_{\cdot\cdot}\rangle, \tag{2.25}$$

and

$$B_i = \frac{\langle I_{i\cdot}\rangle}{\sqrt{\langle I_{\cdot\cdot}\rangle}}. \tag{2.26}$$

Using this the APC version of the mutual information is given by

$$I^{c}_{ij} = I_{ij} - \frac{\langle I_{i\cdot}\rangle\langle I_{j\cdot}\rangle}{\langle I_{\cdot\cdot}\rangle}. \tag{2.27}$$

Since the APC performs better than the ASC we focused on adapting the APC for our Bayesian model. As mentioned above, the logarithms of the R values are the equivalent of mutual information in our model. Therefore, naively we would simply replace I_{ij} with $\log(R_{ij})$ in equation (2.27) above. However, whereas the mutual information naturally has a lower bound of zero, which is reached only for independent positions, $\log(R)$ is off-set with respect to mutual information and becomes *negative* for independent positions. Note also that all posterior probabilities are invariant under a global shift of all the $\log(R)$ values by a constant. Therefore, we substitute into equation (2.27) a shifted version of $\log(R)$ which is guaranteed to be non-negative. For each domain we determine the minimal value $\log(R_{\min})$ and define a shifted version of $\log(R)$ as

$$S_{ij} = \log(R_{ij}) - \log(R_{\min}). \tag{2.28}$$

Using these shifted $\log(R)$s we then define the corrected $\log(R)$ as

$$\log(R^c_{ij}) = S_{ij} - \frac{\langle S_{i\cdot}\rangle\langle S_{j\cdot}\rangle}{\langle S_{\cdot\cdot}\rangle}. \tag{2.29}$$

In our model with phylogenetic correction we simply replace each factor R_{ij} with R^c_{ij}.

2.4.5 Prior probability of spanning trees

Our Bayesian model easily allows for the incorporation of prior probabilities on each spanning tree via the edge probabilities $W_{j\pi(j)}$ in equation (2.9). Here, we use these edge probabilities to include the dependence on both the primary sequence separation of the positions in the pair (Figure 2.11), as well as the sum of the entropies of the corresponding columns (Figure 2.12). To estimate the fraction $f(d, H)$ of all pairs with sequence-separation d and entropy-sum H that are contacts, we separated all pairs of columns into entropy bins of width 0.2, spanning the whole range of entropies $[0, 2\log(20)]$ and compared the dependence on primary sequence separation within the different bins (Figure 2.14, left panel).

We see that, irrespective of the column entropy sum H, the fraction $f(d, H)$ has approximately the same shape as a function of d as the overall fraction of contacts $f(d)$ which we showed in Figure 2.11. We find that for distances $d = 4$ or less the fraction is virtually independent of entropy, i.e. $f(d, H) \approx f(d)$, while for larger distances the fractions $f(d, H)$ are roughly proportional to $f(d)$, with a proportionality constant that decreases with entropy H. That is, we assume the following general form for $f(d, H)$:

$$f(d, H) = \begin{cases} f(d) & \text{if } d <= 4 \\ f(d)g(H) & \text{if } d > 4 \end{cases} \tag{2.30}$$

We first estimated $f(d)$ directly from the observed fractions as shown in Figure 2.11 for all sequence separations up to $d = 50$. As $f(d)$ is proportional to $1/d$

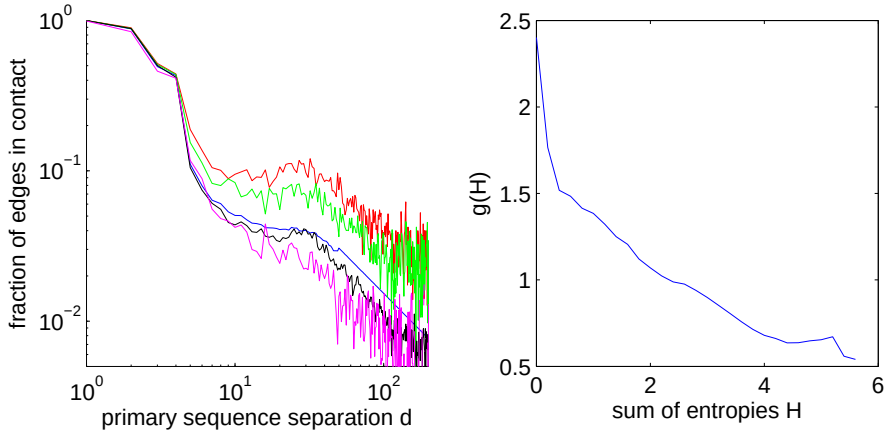


Figure 2.14: **Estimation of prior probabilities**. The left panel shows the dependence between the fraction of pairs that are in contact and primary sequence separation for all pairs (in blue) as well as for pairs whose sum of entropies lies in a given entropy bin (H $\in [0, 0.2)$ in red, H $\in [0.2, 0.4)$ in green, H $\in [3.4, 3.6)$ in black and H $\in [5.4, 5.6)$ in magenta). For the sake of clarity, only a few selected entropy bins across the entire range are shown. The right panel shows the estimated function $g(H)$, which describes how the probability of an edge to be a contact depends on the sum of entropies of the corresponding columns of the alignment (see text).

for sequence separations ≥ 50 and becomes very noisy for large sequence separations (data not shown), we approximate the curve as $f(d) = C/d$ for sequence separations ≥ 50 (blue line in Figure 2.14). The constant C is chosen so that the curve is continuous at $d = 50$. We then determined the function $g(H)$ by numerically maximizing, for each fixed entropy bin H_i, the likelihood of the data, which is given by

$$P(X) = \left[\prod_{e \in E} f(d_e)X\right] \left[\prod_{e \notin E} (1 - f(d_e)X)\right], \tag{2.31}$$

where the first product runs over all edges E with $d > 4$ and $H = H_i$ that are contacts, the second product over all edges with $d > 4$ and $H = H_i$ that are not contacts, and d_e stands for the primary sequence separation of edge e. The value X^* that maximizes the likelihood of the data determines the value of $g(H)$ for the bin H_i, i.e. $g(H_i) = X^*$. The resulting function $g(H)$ is shown in the right panel of figure 2.14. Clearly the probability of an edge decreases with the entropy-sum H, i.e. it drops by almost a factor of 5 from the lowest to the highest entropy edges.

Finally, in order to assign prior probabilities to different possible spanning trees, we assume a random graph model where each edge e occurs with a probability μ_e that is proportional to $f(d_e, H_e)$, with d_e the primary sequence separation, and H_e the entropy sum of edge e. Note that each spanning tree only contains $(l-1)$ edges for a domain of length l, and we thus have to ensure that our random graph model produces on average $(l-1)$ edges. As the expected number of edges in a random

graph is equal to the sum over all μ_e, we set μ_e to

$$\mu_e = (l-1)\frac{f(d_e, H_e)}{\sum_e f(d_e, H_e)}. \tag{2.32}$$

Let G be the full graph including all $\binom{l}{2}$ edges of a particular domain and let π be one particular spanning tree π. We can now write the prior probability of the tree as

$$P(\pi) = \prod_{e\in\pi} \mu_e \prod_{e\in G\setminus\pi} (1-\mu_e) \tag{2.33}$$

Here, the first product runs over all edges e in the tree π and the second one over all edges in G that are not in the tree π. Since the posteriors are independent of a global rescaling of all prior probabilities $P(\pi)$, we divide $P(\pi)$ by the probability of the graph that contains no edges, to obtain

$$P(\pi) \propto \prod_{e\in\pi} \frac{\mu_e}{1-\mu_e} \tag{2.34}$$

which is independent of the edges that are not contained in the tree. We can thus set the edge weights $W_{j\pi(j)}$ in equation 2.9 to

$$W_{j\pi(j)} = \frac{\mu_{j\pi(j)}}{1-\mu_{j\pi(j)}}. \tag{2.35}$$

Unfortunately, we cannot directly used $W_{j\pi(j)}$ to calculate the matrix entries $M_{j\pi(j)} = R_{j\pi(j)}W_{j\pi(j)}$ in equation 2.9. As discussed above, the R-values relate to mutual information I through $R \propto e^{nI}$, where n is the total number of sequences in the alignment. However, even when the phylogenetic correction is employed, because the n sequences contain many phylogenetically closely-related sequences, the number of *statistically independent* sequences is generally much smaller than n. Because of this, even the corrected R-values still significantly overestimate statistical dependence. To take this into account we define the matrix entries $M_{j\pi(j)}$ as

$$M_{j\pi(j)} = \left(R_{j\pi(j)}\right)^{\alpha} W_{j\pi(j)} \tag{2.36}$$

where α is a free parameter, which must lie between 0 (only prior information) and 1 (original R-values). Note that, through this transformation, we are assuming that instead of n independent sequences, there are only αn effectively independent sequences. The PPV-sensitivity curves for varying values of α are shown in Supplementary Figures 2.24, 2.25 and 2.26. For the curve in the main text, we chose $\alpha = 0.025$, so as to maximize the accuracy for pairs with $d \geq 3$ without a significant decrease in accuracy for pairs with $d \geq 12$.

2.5 Supplementary Figures

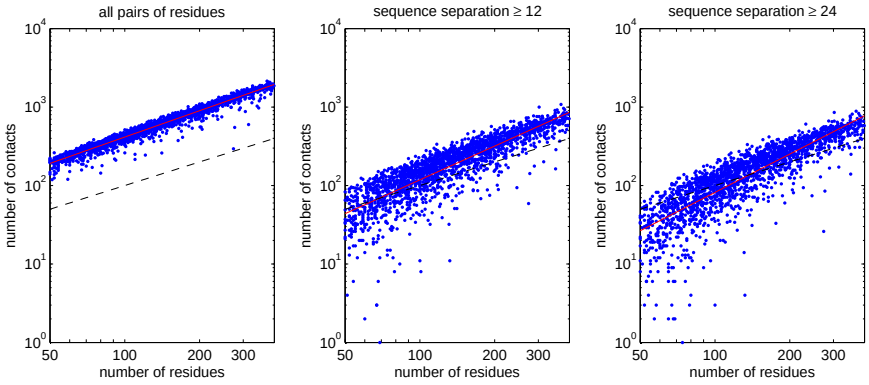


Figure 2.15: Number of contacts n versus the number of residues l per protein domain for varying separations in primary sequence. The red lines are the regression lines (in log-space), corresponding to the power-laws $n = 2.43l^{1.12}$, $n = 0.16l^{1.43}$ and $n = 0.05l^{1.62}$. The dashed black line corresponds to $n = l$.

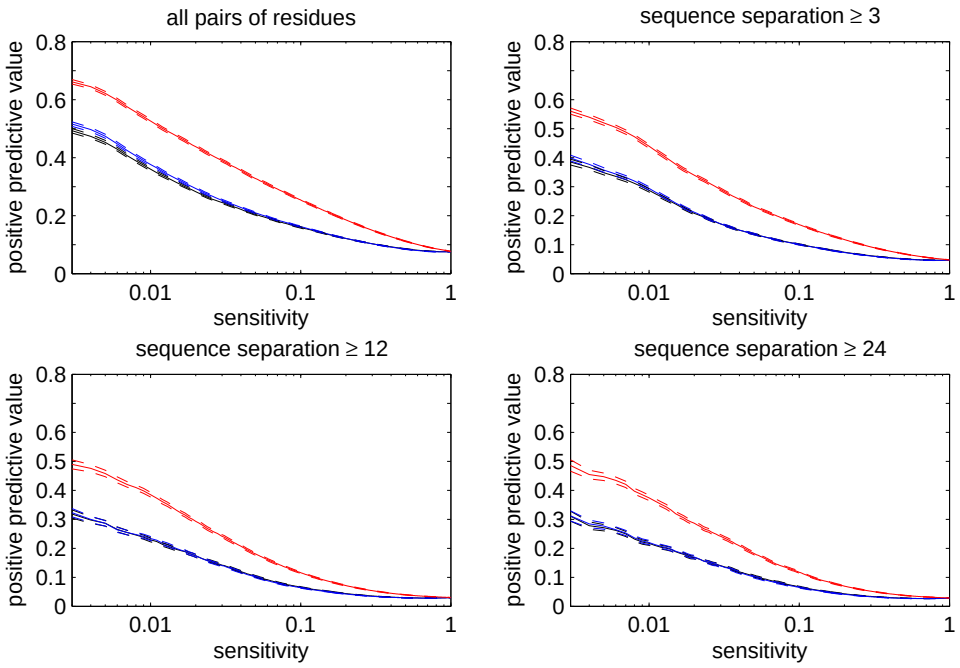


Figure 2.16: Accuracy of contact predictions for all 2009 alignments based on mutual information (black), $\log(R)$ (blue), and posterior probabilities (red). For different values of sensitivity, the corresponding number of predictions for each domain and each method were selected and their positive predicted value (PPV), i.e. the fraction of correct predictions, was calculated (vertical axis). Dashed lines indicate mean PPV plus/minus one standard error. The top left panel shows predictions for all residue pairs, the top right one using only predictions for residues separated by at least 3 positions in the primary sequence, the bottom left one for pairs separated by at least 12 positions, and the bottom right panel for pairs separated by at least 24 positions.

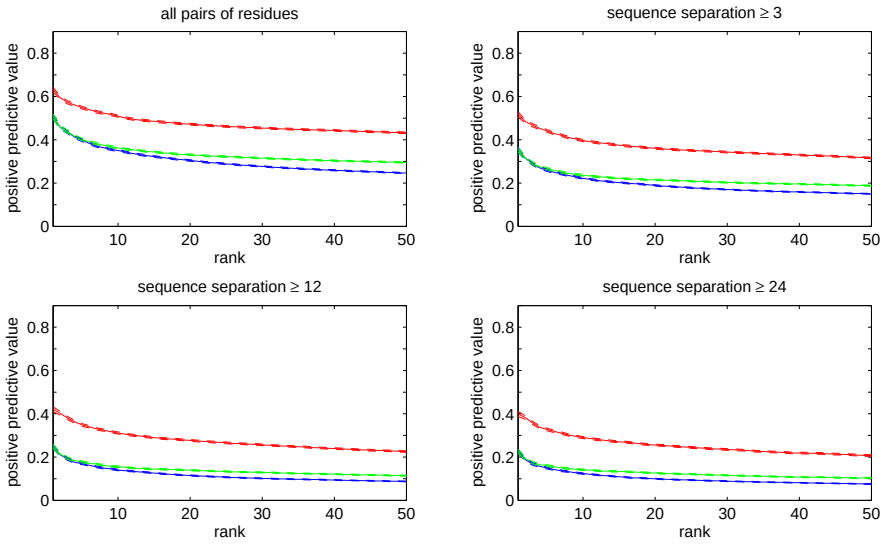


Figure 2.17: Comparison of prediction accuracy for $\log(R)$ (blue), for the $\log(R)$ values contained in the maximum-likelihood tree (green) and for the posterior probability (red). As the maximum-likelihood tree only predicts $l-1$ edges, where l is the number of columns of the alignment, the different measures cannot be directly compared in terms of sensitivity (there would be finite-length effects as predictions by the maximum-likelihood tree measure cannot reach a sensitivity of 1). Instead, we sort the predictions per domain and, for each fixed cut-off on the rank r, we show the average positive predictive value (solid lines) for all predictions with rank r or higher. The dashed lines indicate plus/minus one standard error. As the shortest domains in our dataset have length 50, all domains are included in the calculation of the green curve for ranks 1 to 49. The blue and green curve are identical for high ranks as all the highest-scoring edges are included in the maximum spanning tree. However, for decreasing ranks, the maximum-spanning tree discards edges that can be explained indirectly, which leads to an improvement in performance. Importantly, the posterior probability significantly outperforms the maximum-spanning tree predictions both for low *and* high ranks.

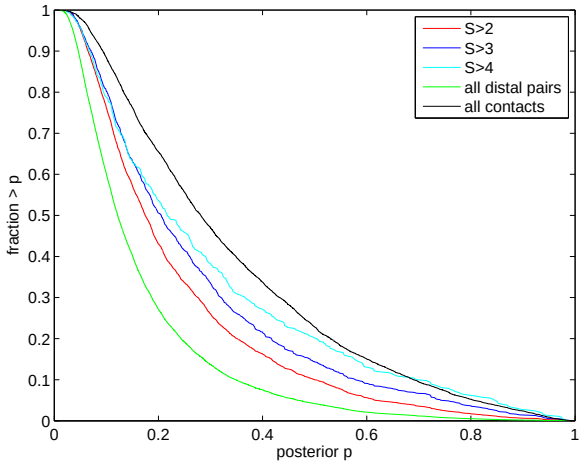


Figure 2.18: Posteriors reflect the extent to which co-evolving pairs can be explained by contact chains. Shown are the reverse cumulative distributions of distal co-evolving pairs ($Z > 4$) that cannot be easily explained by contact chains, i.e. where the best scoring chain has a score of $S > 2$ (red), $S > 3$ (dark blue), or $S > 4$ (light blue). For comparison the reverse cumulative distributions of posteriors for all co-evolving distal pairs (green) and all co-evolving contacts (black) are also shown.

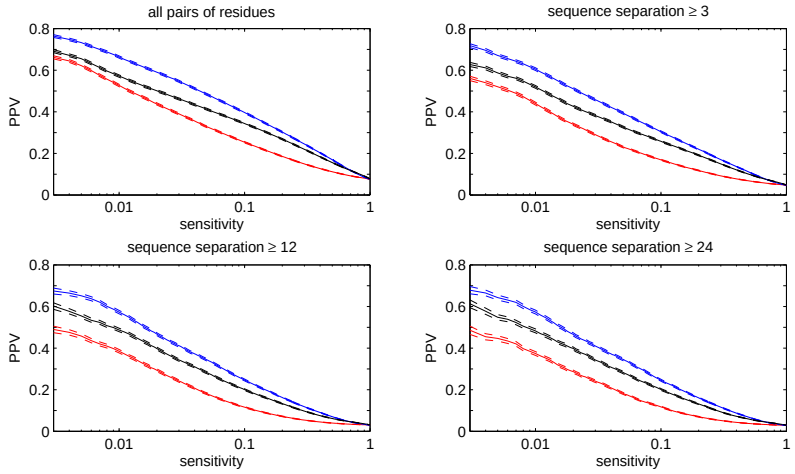


Figure 2.19: Accuracy of contact predictions for all alignments. In blue, we show the performance of the phylogenetically-corrected posterior probabilities, in black the performance of the predictions based on the average-product corrected (APC) mutual information, and in red the performance of the posterior probabilities without phylogenetic correction. Curves were calculated as described in the main text.

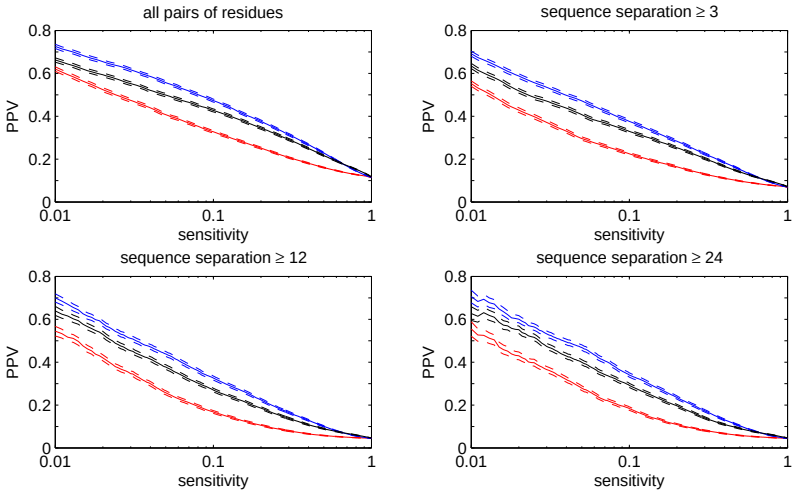


Figure 2.20: Same as figure 2.19, but for alignments of length 50 to 100.

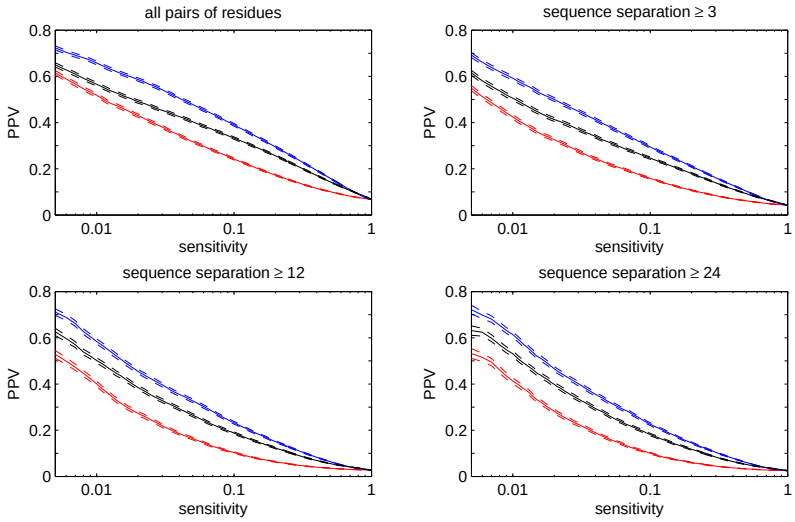


Figure 2.21: Same as figure 2.19, but for alignments of length 101 to 200.

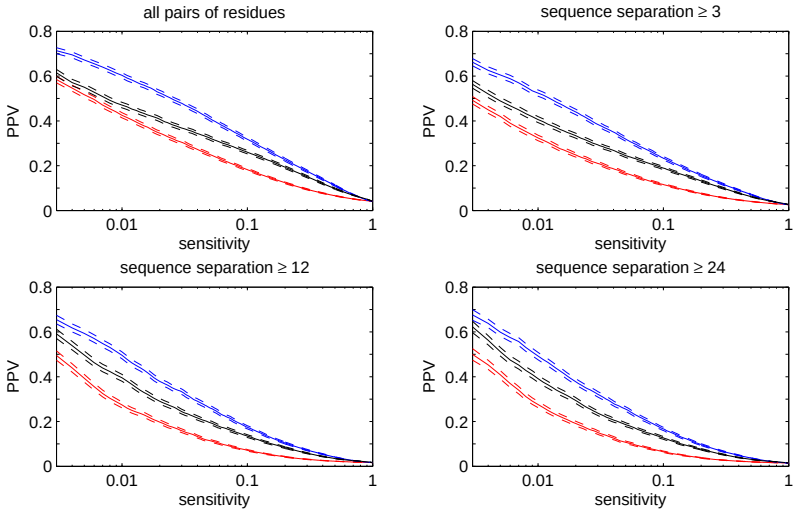


Figure 2.22: Same as figure 2.19, but for alignments of length 201 to 300.

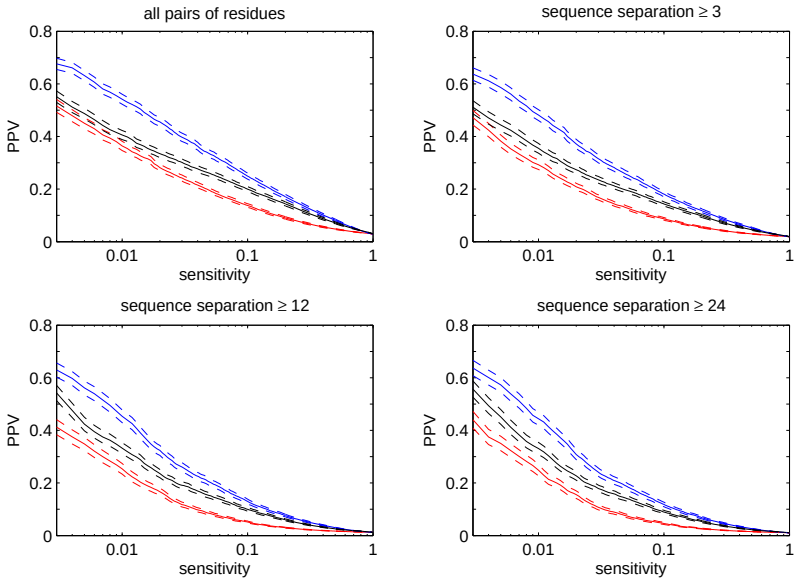


Figure 2.23: Same as figure 2.19, but for alignments of length 301 to 400.

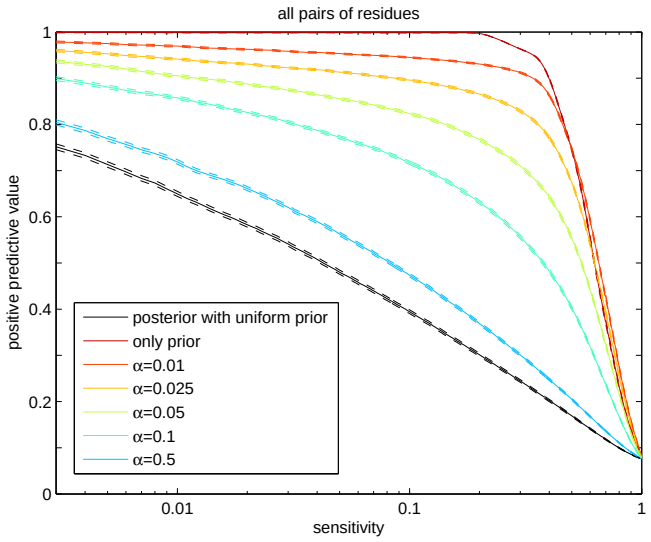


Figure 2.24: Accuracy of contact predictions including the informative prior for different values of the weighting parameter α, including the limit of using only the informative prior ($\alpha = 0$). The positive predictive value (vertical axis) is shown as a function of sensitivity (horizontal axis). Different colors correspond to different values of α (see legend) and dashed lines show mean plus and minus one standard error. For comparison, we also show the performance of the posterior when using no prior information (black). Note that the horizontal axis is shown on a logarithmic scale.

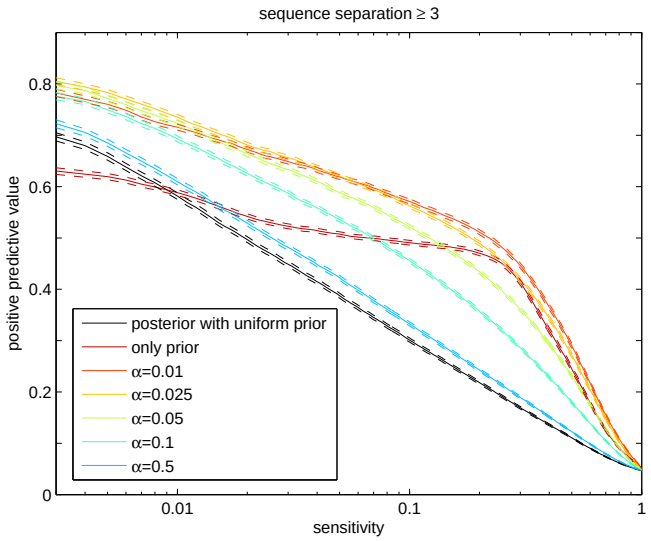


Figure 2.25: As in Supplementary Figure 2.24, but restricting the evaluation to pairs that are at least $d = 3$ apart in primary sequence.

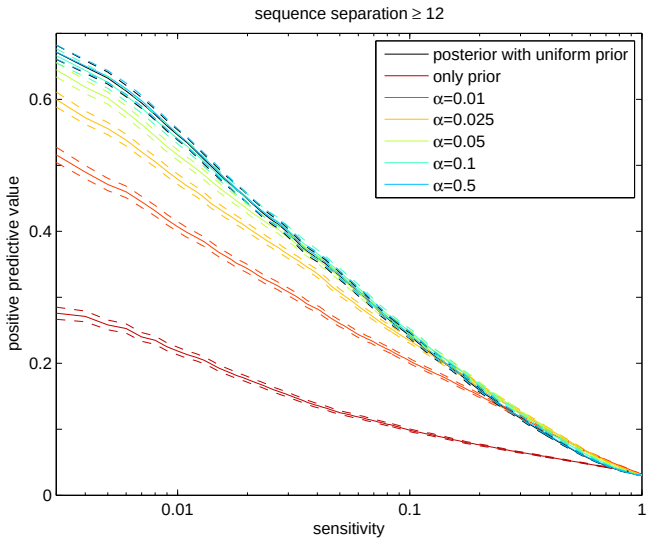


Figure 2.26: As in Supplementary Figure 2.24, but restricting the evaluation to pairs that are at least $d = 12$ apart in primary sequence.

Chapter 3

A Bayesian algorithm for reconstructing bacterial signaling networks

Lukas Burger and Erik van Nimwegen
published in *Algorithms in Bioinformatics, WABI 2006*

We present an algorithm, based on a Bayesian network model, for *ab initio* prediction of signaling interactions in bacterial two-component systems. The algorithm uses a large training set of known interacting kinase/receiver pairs to build a probabilistic model of dependency between the amino acid sequences of the two proteins and the algorithm uses this model to predict which pairs interact. We show that the algorithm can accurately reconstruct cognate kinase/receiver pairs across all sequenced bacteria. We also present predictions of interacting orphan kinase/receiver pairs in the bacterium *Caulobacter crescentus* and show that these significantly overlap with experimentally observed interactions.

3.1 Introduction

The automated prediction of protein-protein interactions on the basis of their amino acid sequences alone is one of the great challenges in computational biology. Here we present a first attempt at such an algorithm for the large class of bacterial two-component systems. Two-component systems consist of protein pairs in which one protein contains a histidine kinase domain that specifically transfers a phosphate to the receiver domain contained in the other protein. Since two-component systems are responsible for most signal transduction in bacteria [48, 49] successful computational

prediction of two-component system interactions would allow exhaustive reconstruction of signaling networks across all fully sequenced bacterial genomes.

There are several reasons that make two-component systems particularly attractive for computational modeling. Firstly, both member domains of this family of proteins, the histidine kinase and the receiver domain, exhibit a high degree of sequence similarity and they can be easily detected in fully-sequenced genomes using hidden Markov models. Second, two-component systems are very abundant in the bacterial and archeal kingdom, with many tens of interacting pairs in some genomes, and thousands of examples across all genomes, providing enough data for relatively subtle statistical modeling. Finally, for a significant fraction of all two-component systems, the interacting partners lie in the same operon on the genome, which allows us to easily extract a large number of examples of "known" interacting pairs.

Two component-systems are the main means through which bacteria sense and adapt to their environment [48]. In many cases the histidine kinase is a membrane-bound protein, with a sensor domain which responds to environmental cues and, on the cytoplasmic side, a kinase domain, which, upon activation of the sensor, binds ATP and autophosphorylates on a highly conserved histidine residue. The kinase domain very specifically interacts with its cognate response regulator by transferring the phosphate to a conserved aspartate residue in the regulator's receiver domain. Phosphorylation leads to the activation of the regulator's output domain, which often consists of a DNA-binding domain, enabling the regulator to act as a transcription factor. Due to the high sequence similarity among both kinases and regulators, it is unclear how the specificity of interaction comes about (see e.g. [50]). It is agreed that the residues close to the site of phosphorylation in the kinase, the conserved histidine residue, and the residues around the active site aspartates of the regulator are of great importance for the interaction [50, 51] but the origin of the specificity is currently not understood.

In this article, we will present an algorithm that uses a statistical model to predict interacting kinase/receiver pairs. We test the performance of the algorithm on reconstructing known cognate pairs from all sequenced bacteria and use it to predict interaction partners for orphan kinases in the Gram-negative bacterium *Caulobacter crescentus*, where orphans play an important role in the cell-cycle progression [52]. We will show that our predictions agree well with the experimental results. Although, in a previous work, statistical methods have been used to identify residues that are important for the interaction specificity [53], the method we describe in this paper is, to our knowledge, the first attempt to computationally predict interactions in two-component systems.

3.2 Outline of the algorithm

Our prediction algorithm operates in two steps. Comparison of the kinases from all sequenced bacteria shows that there are 7 major classes of domain architectures. Using a training set of cognate receivers for each class of kinases we build position-specific weight matrices (WMs) for the receivers of each class and use these to classify receivers. This allows us to predict, for each receiver, which type of kinase it will interact with. In the second step of our algorithm we aim to identify which kinase/receiver pairs within a class interact. To this end we again use training sets of cognate kinase/receiver pairs and identify pairs of amino acid positions in kinase and receiver that show significant mutual information. Using a network of such correlated positions we construct statistical models for the joint distribution of amino acids in interacting kinase/receiver pairs. The final "score" for a putative interacting pair is given by the ratio of the likelihood of their sequences given that they are an interacting pair and the likelihood assuming independence of their sequences. In order to reconstruct cognate kinase/receiver pairs genome-wide we use Monte-Carlo Markov sampling to sample all ways of assigning kinase/receiver pairs, sampling each assignment in proportion to the likelihood of the sequences of all interacting pairs in the assignment.

3.3 Classifying bacterial two-component systems

To gather an exhaustive collection of two-component system proteins we first collected a set of hidden Markov profiles from the Pfam database [23] that characterize two-component systems. These are the histidine kinases HisKA, HisKA_2 (or H2), HisKA_3 (or H3), and HWE_HK, the ATP-binding domain HATPase_c, the His-containing phosphotransfer domain Hpt, and the response regulator receiver domain Response_reg (or RR). We then collected all bacterial genomes from the NCBI database [54] and searched for matches to all these domains using the hmmpfam [55] program.

Whereas the response regulators are characterized by a single receiver domain Response_reg, the kinases are represented by 6 different domains. We find that almost all kinases exhibit one of 7 different domain architectures. These are, in order of their abundance, the HisKA (HisKA, HATPase_c), H3 (H3, HATPase_c), chemotaxis (Hpt, HATPase_c), long hybrid (HisKA, HATPase_c, RR, (RR), Hpt), short hybrid (HisKA, HATPase_c, RR, (RR)), HWE (HWE, HATPase), and Hpt(Hpt) kinases. Other architectures had less than 10 occurrences in the entire set of bacterial genomes. Note that both hybrid classes contain one (or sometimes even two) receiver domains themselves but are believed to almost always also interact with a receiver domain in another protein.

3.3.1 Multiple alignments

To produce multiple alignments of the receiver domains and of the kinases in each of the 7 classes we first used the program Hmmalign [55] for each domain. For the HisKA, chemotaxis, HWE and H3 kinase classes we constructed a full alignment by simply concatenating the alignments of the kinase and the HATPase_c domains. For the short hybrids we aligned the HisKA and HATPase_c domains and for the long hybrids only the Hpt domain.

To check the accuracy of the alignments we compared the Hmmalign alignments with alignments made by the ProbCons algorithm [56]. For each class 200 sequences were selected at random (or all if the class has less than 200 sequences) and aligned with ProbCons. We then selected all columns in the hmmalign alignments that contain less than 15% gaps and for which at least 80% of the amino acids in the column also align together in a single column in the ProbCons alignment. We call these columns the 'trusted positions'. Finally, we replaced each alignment with the alignment of only the trusted positions.

3.3.2 Cognate pairs and orphans

As mentioned in the introduction, many kinase/receiver pairs of genes occur next to each other or in the same operon on the DNA and it is believed that almost all of these form interacting pairs. Using this we constructed a training set of 'cognate' pairs as follows. We defined operons as maximal sets of contiguous genes on the same strand of the DNA with all intergenic regions between consecutive genes less than 100 bps in length. Whenever an operon contains only one kinase and one receiver these two were considered a cognate pair that we assume to interact.

Our analysis of all bacterial genomes resulted in 2165 cognate pairs of a HisKA kinase and a receiver, 415 cognate pairs with an H3 kinase, 113 cognate pairs with a chemotaxis kinase, 86 cognate pairs with a long hybrid kinase, 82 pairs with a short hybrid kinase, 22 pairs with an HWE kinase, and 19 pairs with an Hpt kinase.

Beside the cognate pairs there is an almost equal number of cases in which a kinase or a receiver occurs by itself in an operon. For virtually all of these 'orphan' kinases and receivers it is currently unknown what partners they interact with and one of the major aims of our algorithm is to predict interaction partners for these orphans.

3.3.3 Classification of response regulators

We found that response regulators that interact with different types of kinases show distinct amino acid compositions and these differences can be used to predict, for each receiver, what kind of kinase it will interact with.

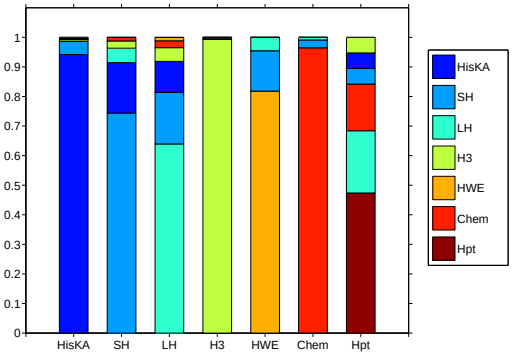


Figure 3.1: Predicted classification of receivers. Each bar represents the set of all receivers that are member of a cognate pair with kinases of a particular type. The color distribution in the bar shows what percentages of the receivers are classified with each class. The correspondence between color and kinase type is shown in the legend on the right.

We divided the multiple alignment of all cognate receivers into 7 sub-alignments corresponding to all receivers that interact with kinases of a particular class. For each of the 7 alignments we then constructed a position specific weight matrix

$$w_{i\alpha}^{c} = \frac{n_{i\alpha}^{c} + \lambda}{(n^{c} + 21\lambda)}. \tag{3.1}$$

Here $n_{i\alpha}^{c}$ is the number of times amino acid α occurs in column i of the alignment (gaps are treated as a 21st amino acid) of cognate receivers of class c, n^{c} is the total number of sequences in the alignment, and λ is the pseudocount resulting from the Dirichlet prior (we used $\lambda = 1/2$). $w_{i\alpha}^{c}$ is thus the estimated probability of seeing amino acid α in position i of a receiver of class c.

For each receiver sequence S we can now determine the posterior probability $P(c|S)$ that it belong to class c. We have

$$P(c|S) = \frac{P(S|w^{c})P(c)}{\sum_{c'} P(S|w^{c'})P(c')} \quad \text{with} \quad P(S|w^{c}) = \prod_{i \in \mathrm{TP}} w_{S_i i}^{c}. \tag{3.2}$$

Here S_i stands for the amino acid in the ith position of receiver sequence S. Note that the product only runs over all the trusted positions TP. We assumed a uniform prior $P(c) = 1/7$. When classifying a regulator whose sequence was used in the construction of the WM we removed its contribution from the counts $n_{i\alpha}^{c}$.

The results of the classification are shown in Fig. 3.1. The posterior probabilities for the 7 classes were calculated for each receiver and the receiver was assigned to the class with the highest posterior probability (which is often close to 1). The results show that for the three most abundant types (HisKA, H3, and chemotaxis kinases) the

classifier predicts almost perfectly which receivers interact with HisKA kinases, which with H3 kinases, and which with chemotaxis kinases. For the other, rarer classes the classification is still correct in the majority of the cases, except for the very rare Hpt kinases where slightly more than half are misclassified. The lower performance for the rarer classes is presumably due to the fact that the WM models for these classes are based on a relatively small number of examples.

The types of misclassifications match what is to be expected based on the domain architectures. Both chemotaxis and long hybrid kinases contain an Hpt domain and some of the receivers that interact with a single Hpt domain kinase are mistaken for a receiver that interacts with the Hpt domain of a chemotaxis or long hybrid kinase. Similarly, both long and short hybrids contain an HisKA domain and their receivers are sometimes mistaken for a receiver that interacts with a single HisKA domain kinase. Overall, the WM model predicts the correct type of kinase for 93% of the cognate receivers.

3.4 Predicting cognate interactions

Once we have classified the receivers according to the type of kinase they interact with the second step of our algorithm consists of predicting, for each class, which pairs of kinases and receivers interact. To do this we make alignments of all cognate kinase/receiver pairs in each class by simply concatenating the respective kinase and receiver alignments. We then build probabilistic "dependent" models for the joint amino acid sequences of cognate kinase/receiver pairs and "independent" models for the kinases and receivers independently. The algorithm then predicts interactions between kinase/receiver pairs whose sequences are more likely under the dependent than under the independent model.

3.4.1 Quantifying dependence between positions in kinase and receiver

Given the joint multiple alignment of kinase/receiver pairs in a particular class we quantify the dependence between all pairs of trusted positions (i, j), where the positions i and j may both be either in the kinase or in the receiver, using a measure closely related to mutual information. For each pair (i, j) we calculate the likelihood of the observed columns of amino acids under a model that assumes the amino acids at the two positions were drawn from two independent distributions and under a model that assumes general dependence between the two amino acids. In particular, for the independent model let p_α denote the probability to observe amino acid α at position i, and let q_β denote the probability to observe amino acid β at position j. For the dependent model, let $w_{\alpha\beta}$ denote the probability to observe the pair of amino

acids (α, β) at positions (i, j). Finally, let D_{ij} denote the columns of observed amino acids in the alignments at positions i and j, $n_{\alpha\cdot}$ the number of times α is observed at position i, $n_{\cdot\beta}$ the number of times β is observed at position j, and $n_{\alpha\beta}$ the number of times the pair of amino acids (α, β) is observed at positions (i, j).

Given the joint probability $w_{\alpha\beta}$ the probability of the data D_{ij} is given by

$$P(D_{ij}|w) = \prod_{\alpha\beta} (w_{\alpha\beta})^{n_{\alpha\beta}} \tag{3.3}$$

and under the independent models p and q the probability of the data is given by

$$P(D_{ij}|p, q) = P(D_i|p)P(D_j|q) = \left[\prod_{\alpha} (p_\alpha)^{n_{\alpha\cdot}}\right] \left[\prod_{\beta} (q_\beta)^{n_{\cdot\beta}}\right]. \tag{3.4}$$

Since the distributions p, q and the joint distribution w are unknown, they are nuisance parameters that we integrate out of the likelihood for the dependent and independent models. We use Dirichlet priors of the form $P(w) \propto \prod_{\alpha\beta} w_{\alpha\beta}^{\lambda-1}$ and integrate over the simplices $\sum_\alpha p_\alpha = \sum_\beta q_\beta = \sum_{\alpha\beta} w_{\alpha\beta} = 1$. We then obtain for the probability of the data under the dependent model

$$P(D_{ij}|\text{dep}) = \int P(D_{ij}|w)P(w)dw = \frac{\Gamma(21^2\lambda)}{\Gamma(n + 21^2\lambda)} \prod_{\alpha\beta} \frac{\Gamma(n_{\alpha\beta} + \lambda)}{\Gamma(\lambda)}, \tag{3.5}$$

and similarly for the probability of the data under the independent model

$$P(D_{ij}|\text{indep}) = \frac{\Gamma^2(21\lambda)}{\Gamma^2(n + 21\lambda)} \left[\prod_{\alpha} \frac{\Gamma(n_{\alpha\cdot} + \lambda)}{\Gamma(\lambda)}\right] \left[\prod_{\beta} \frac{\Gamma(n_{\cdot\beta} + \lambda)}{\Gamma(\lambda)}\right], \tag{3.6}$$

where $\Gamma(n)$ is the Gamma function. Finally, we quantify the amount of dependence between positions i and j by the log-ratio R_{ij} of likelihoods of the dependent and independent models

$$R_{ij} = \log\left[\frac{P(D_{ij}|\text{dep})}{P(D_{ij}|\text{indep})}\right]. \tag{3.7}$$

For our calculations we used the Jeffreys, or information theory prior with $\lambda = 1/2$. One can think of the quantity R as a finite-size corrected version of the mutual information that takes into account the larger model space of the dependent model [57].

3.4.2 Probabilities of kinase/receiver pairs under interacting and independent models

For each class, R_{ij} was calculated for each pair of trusted positions both between kinases and receivers and within the proteins themselves. Since the equation for R_{ij}

trades off the observed mutual information between the distributions of amino acids at positions i and j against the much larger model space for the dependent model, one generally finds that there are many more positions with R larger than zero for the large HisKA class than for the classes with much smaller numbers of sequences. To obtain a reasonable number of dependent positions for all classes we chose a stringent cut-off of $R = 50$ for the HisKA class and a more lenient cut-off of $R = 0$ for the other classes. For each class we then collected the set of 'significant positions' Ω^c that score over the threshold with at least one other amino acid.

The two-point correlation structure of the significant positions can be represented by a graph in which each node is a significant position and two nodes are connected if R for the two positions scores over the threshold. Interestingly, we find that this graph generally consists of a large connected component containing both kinase and receiver positions, plus a few small connected components containing either only kinase or only receiver positions. Since the positions in these small components do not contain information about the dependence between kinase and receiver we discarded them from the set Ω^c.

We now approximate the joint distribution of the significant positions in interacting kinase/receiver pairs using pairwise conditional probabilities between positions. The procedure is illustrated in Fig. 3.2. The multiple alignments of cognate kinase-receiver pairs are shown at the top with the significant positions as colored columns and the arcs indicating which pairs of columns correlate significantly. To factorize the joint probability of all significant positions we use a slightly modified version of the Chow-Liu algorithm [24] that reduces the correlation graph to a tree while maximizing the sum over the R values along the remaining edges. For example, in the bottom left of Fig. 3.2 the links 6 and 7 with the lowest R values were removed to yield a tree. Once a root is chosen (arbitrarily) each position i (except for the root) will have exactly one parent $\pi(i)$ and we factor the joint probability by assuming the amino acid at position i is only dependent on the amino acid at position $\pi(i)$. That is, if $S_{K,R}$ denotes the set of significant positions for kinase K and receiver R then the probability $P(S_{K,R}|c)$ of the sequences assuming that they are an interacting pair of class c is given by

$$P_{K,R}(S_{K,R}|c) = \prod_{i \in \Omega^c} P(S^i_{K,R}|S^{\pi(i)}_{K,R}, c) \quad \text{with} \quad P(S^r_{K,R}|S^{\pi(r)}_{K,R}, c) \equiv P(S^r_{K,R}|c), \tag{3.8}$$

for the root of the tree r. Here $S^i_{K,R}$ is the amino acid in the ith significant position of the kinase-regulator sequence and $\pi(i)$ is the parent of position i as defined by the tree. The probability $P(\alpha|\beta, c)$ to observe amino acid α at position i given that amino acid β occurs at position $\pi(i)$ is given by

$$P(\alpha|\beta, c) = \frac{n^c_{\alpha\beta} + \lambda w^c_{\alpha i}}{n^c_{\cdot\beta} + \lambda}, \tag{3.9}$$

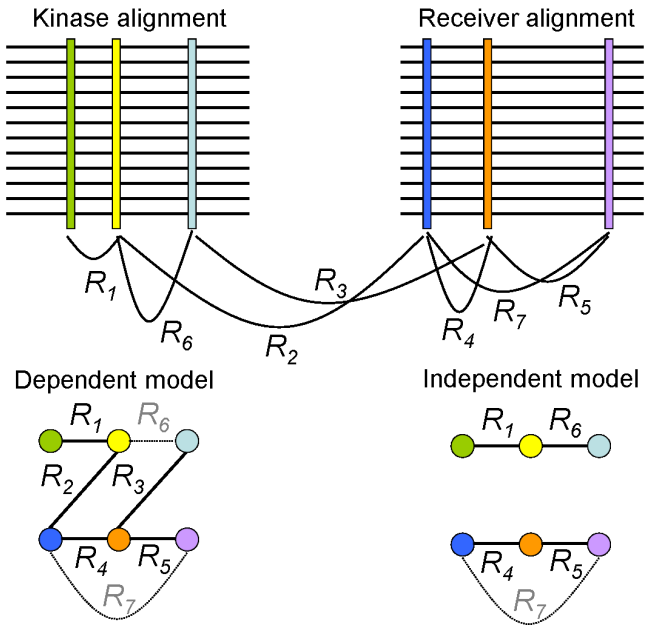


Figure 3.2: Multiple alignments of cognate (interacting) kinase/receiver pairs with significant positions shown as colored columns. The arcs show the pairs of positions that are significantly correlated. The correlation structure of the dependent and independent models are shown at the bottom. The edges that are removed by the Chow-Liu algorithm are shown as dotted lines.

where $n^c_{\alpha\beta}$ is the number of times the pair $\alpha\beta$ occurs at positions i and $\pi(i)$ of the cognate kinase-receiver pairs of class c, $n^c_{\cdot\beta}$ is the total number of times that β occurs at position $\pi(i)$, and λ is the pseudo-count of the Dirichlet prior (here we use a much larger $\lambda = 10$ to smooth fluctuations due to the small sample size). Note that we made the prior for the conditional probabilities proportional to the independent probability, i.e. the WM w^c, for class c.

In complete analogy we calculate the independent probabilities $P(S_K|c)$ of the kinase and $P(S_R|c)$ of the receiver, where we now only allow conditional dependence between positions within the kinase and positions within the receiver as in the bottom right of Fig. 3.2. Finally, we assign a "score" $Z(K, R|c)$ to the pair K, R which equals the logarithm of the likelihood ratio

$$Z(K, R|c) = \log\left[\frac{P(S_{K,R}|c)}{P(S_K|c)P(S_R|c)}\right]. \tag{3.10}$$

3.4.3 Results on reconstructing cognate pairs

For each genome and each class we collected all kinases in the class together with their cognate receivers. We then used Monte-Carlo Markov sampling to sample all ways of assigning one kinase to each receiver. Let a denote an assignment and let

$R(K, a)$ denote the receiver assigned to kinase K in assignment a. The probability of sampling a is then given by

$$P(a) \propto \exp\left[\sum_K Z(K, R(K, a))\right]. \tag{3.11}$$

We then measured what fraction of the time $f(R, K)$ during sampling each kinase K was assigned to each receiver R. Note that to calculate the scores $Z(K, R)$ the pair (K,R) in question was removed from the training set. The results are shown in Fig. 3.3. For different values of f we counted what fraction of true interacting pairs (i.e. cognate pairs) from all genomes have $f(R, K) > f$ (sensitivity) and also what fraction of all pairs that have $f(R, K) > f$ are true interacting pairs (specificity). The resulting sensitivity/specificity curves for the 4 most abundant kinase classes are shown in Fig. 3.3. As the figure shows, our model very accurately predicts which kinase interacts with which receiver. For example, more than 60% of all cognate pairs for all classes can be predicted at a specificity close to 1. Note that in a genome with n cognate pairs there are only n true interactions among n^2 possible interactions. This explains why the lowest specificities are obtained for the large HisKA class, i.e. the correct interactions have to be discovered in a much larger set of putative interactions for this class. Still, even for HisKA 75% of all true interactions are predicted at a specificity of about 75%.

3.5 Prediction of orphan interactions in *Caulobacter crescentus*

Although the previous section shows that our algorithm can accurately reconstruct interacting pairs for the cognate kinases and receivers, these predictions are not biologically novel since for cognate pairs the interacting partners could already be determined from their positions on the DNA. Therefore, we next applied our algorithm to predict interaction partners for orphan kinases and receivers. It is difficult to assess the performance of our algorithm in this context since only very few orphan interactions have been experimentally characterized. Moreover, most of the experimental work is done *in vitro* under conditions that are very different from those *in vivo* and it is not clear if observed interactions *in vitro* reliably reflect *in vivo* interactions.

We chose the bacterium *Caulobacter crescentus* as a test case since most experimentally known orphan interactions are from that organism [52, 58]. *C. crescentus* contains 40 orphan kinases of which 6 are in the class of HisKA kinases. Since all but two of the known interactions involve HisKA kinases we decided to focus on the 5 HisKA kinases for which at least one interaction has been experimentally characterized. There are 23 orphan receivers in *C. crescentus* and we determined the score

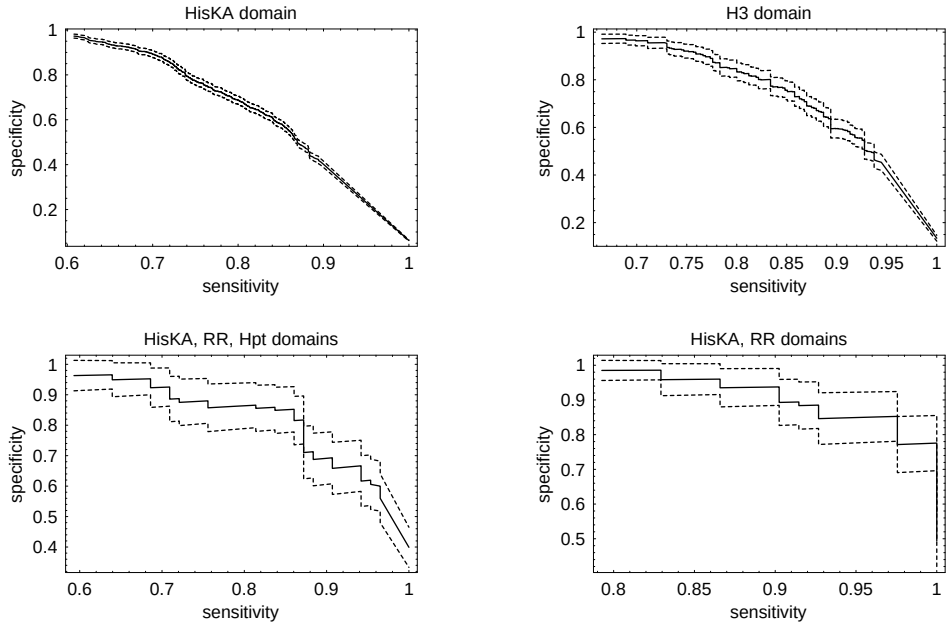


Figure 3.3: Sensitivity/Specificity curves for the 4 most abundant kinase classes. The solid lines give the estimated specificity and the dashed lines give two standard errors of the estimate.

Z for each orphan receiver with each of these 5 HisKA orphan kinases. The results are shown in table 3.1.

As shown in the table, 5 of the 8 experimentally observed interactions rank either immediately at the top or at the second position of the ordered list for each kinase. For DivJ the two known interactions with DivK and PleD occur at positions 7 and 8 of the list (of 23 receivers in total). The only known interaction not shown in the table is the interaction of DivL with CtrA which occurs at position 16 of DivL's list. To test the significance of these predictions we calculated p-values under a rank-sum test,i.e. by randomly permuting the ranks of the interaction scores. If we include the "bad" case DivL-CtrA, the probability of getting a set of predictions as good or better in ranks than ours is $p = 5 \cdot 10^{-4}$. Without CtrA, the p-value is $p = 3.5 \cdot 10^{-5}$.

In summary, in spite of the small number of experimentally determined orphan interactions the predictions of our algorithm show a significant overlap with the known interactions.

3.6 Conclusions

We have presented the first computational method for reconstructing bacterial signaling networks from the knowledge of amino acid sequences only. First, we have

kinase	regulator	interaction score	experimental evidence
DivL	DivK	3.75	yeast two-hybrid screen [59]
PleC	DivK	1.95	*in vitro* phosphorylation [60]
PleC	PleD	-0.47	*in vitro* phosphorylation [60]
CckN	CC1364 (CheYIII protein)	9.28	
CckN	DivK	8.47	yeast two-hybrid screen [59]
CenK	CC1842	7.38	
CenK	CenR	6.39	*in vitro* phosphorylation [60]
DivJ	CC3155 (CheYIII protein)	-0.51	
DivJ	CC0612 (NasT)	-1.75	
DivJ	CC3162	-2.17	
DivJ	CC1842	-2.28	
DivJ	CC3471	-2.38	
DivJ	CC1364 (CheYIII)	-2.65	
DivJ	DivK	-2.65	*in vitro* phosphorylation [60]
DivJ	PleD	-3.52	*in vitro* phosphorylation [60]

Table 3.1: Predictions for HisKA orphan kinases of *Caulobacter crescentus* for which at least one interaction has been experimentally characterized. For each kinase K the receivers are sorted by their score $Z(K, R)$ and the known interactions are indicated. We cut each list off to include the known interactions except for the interaction of DivL with the receiver CtrA, which occurs at position 16 in the list of DivL.

shown that the domain architectures of the kinases of bacterial two-component systems fall into 7 distinct classes and that, using position-specific weight matrices, one can accurately predict which of these kinase classes each receiver domain interacts with. Using training sets of known interacting kinase/receiver pairs we determined which positions in the kinase and the receiver show clear evidence of dependence between their amino acids. From this correlation structure we constructed a probabilistic model for the joint distribution of the amino acid sequences of interacting pairs, and 'independent' models for the distributions of amino acids in kinases and receivers separately. Finally, with these probabilistic models we predict kinase/receiver interactions across all sequenced bacterial genomes. We first tested our predictions on all cognate pairs of the training set using a 'leave-one-out' scheme. These tests show that the cognate interactions can be very accurately reconstructed using our model. Second, we predicted interactions between orphan kinase and receivers in *Caulobacter crescentus*, and compared these with the few interactions that have been characterized in the literature. This test showed a significant overlap between the known interactions and the predictions of our algorithm. Given the small number of examples involved we cannot yet assess if the very high performance observed on the cognates generalizes to the orphans but it is highly encouraging that for 4 of the 5 tested kinases observed interactions ranked at the first or second position of our list

of predictions. We believe that the large number of orphan interactions predicted by our algorithm across all sequenced genomes already form a valuable data-set for experimental investigation.

Chapter 4

Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method

Lukas Burger and Erik van Nimwegen
published in *Molecular Systems Biology 4:165, 2008*

Accurate and large-scale prediction of protein-protein interactions directly from amino acid sequences is one of the great challenges in computational biology. Here we present a new Bayesian network method that predicts interaction partners using only multiple alignments of amino acid sequences of interacting protein domains, without tunable parameters, and without the need for any training examples. We first apply the method to bacterial two-component systems and comprehensively reconstruct two-component signaling networks across all sequenced bacteria. Comparisons of our predictions with known interactions show that our method infers interaction partners genome-wide with high accuracy. To demonstrate the general applicability of our method we show that it also accurately predicts interaction partners in a recent dataset of polyketide synthases. Analysis of the predicted genome-wide two-component signaling networks shows that cognates (interacting kinase/regulator pairs which lie adjacent on the genome) and orphans (which lie isolated) form two relatively independent components of the signaling network in each genome. In addition, while most genes are predicted to have only a small number of interaction partners, we find that 10% of orphans form a separate class of 'hub' nodes that distribute and integrate signals to and from up to tens of different

interaction partners.

4.1 Introduction

A method that comprehensively and accurately predicts protein-protein interactions using only the amino acid sequences of proteins would essentially allow the reconstruction of genome-wide interaction networks directly from genome sequences. Automated prediction of protein-protein interactions from their amino acid sequences is therefore one of the great outstanding challenges in computational biology. Numerous approaches have already been proposed which, apart from the amino acid sequences themselves, use additional information as co-expression patterns, phylogenetic distributions of orthologous groups, co-evolution patterns, the order of genes in the genome, gene fusion and fission events, and synthetic lethality of gene knock-outs, see [61–63] for reviews. There are, however, serious shortcomings to the currently existing methods. For instance, many of the approaches cannot infer direct physical interactions, but indicate only general functional 'relationships' which may often be indirect and are difficult to validate. Some methods, such as those that rely on phylogenetic tree comparison, cannot be easily scaled up to large data-sets. In addition, accuracy in genome-wide predictions is a general problem. Because true interactions are only a small fraction of the large number of possible interactions genome-wide, even relatively low false-positive rates lead to high numbers of false positives compared to the number of true predictions, see e.g. [64]. Furthermore, since high-throughput experimental methods for mapping protein-protein interactions are notoriously noisy it is difficult to assess the reliability of computational predictions. This is especially a problem for transient protein-protein interactions such as those that take place during signaling. Yet these interactions are often most interesting because of their regulatory role.

Here we present a novel probabilistic method for inferring interaction partners in families of homologous proteins, using only alignments of amino acid sequences. Of the existing methods for protein-protein interaction prediction, our method is most similar in spirit to the correlated mutations method of [7]. In their approach the assumption is made that, for interacting protein pairs, pairs of residues involved in the interaction will show correlated mutations. In particular, it is assumed that replacement of one of the interacting residues with a chemically highly dissimilar amino acid typically require the other residue to also change substantially. For a given pair of proteins orthologs from related genomes are collected and an *ad hoc* scoring scheme is used to identify pairs of positions that show significant correlation of their mutations across the orthologous pairs.

The similarity of this approach with ours is that we likewise assume that, for interacting protein pairs, there will be pairs of residues which show co-variation. However, whereas the method of Pazos *et al* only considers one pair of proteins together with their orthologs at a time, we consider multiple alignments of entire families of proteins (or protein domains) that are known to interact, which includes all paralogs and

orthologs at once. In addition, we use a rigorous Bayesian network frame-work to explicitly model the entire joint probability of all amino acid sequences in the multiple alignments. In this model the identity of each residue is probabilistically dependent on the identity of one other residue, which may either lie within the same protein or lie within the interacting partner. Our model also sums over all ways the residue dependencies can be chosen.

We demonstrate the power of our method by first applying it to bacterial two-component system proteins, which are responsible for most signal transduction in bacteria. Whereas much knowledge has been gained in recent years regarding the structure of transcriptional regulatory networks and metabolic networks, very little is known about the global structure of signaling networks in bacteria. Here we provide the first genome-wide reconstruction of two-component signaling networks across all sequenced bacterial genomes. By comparing our predictions with large sets of known interactions we demonstrate the high accuracy of our predictions. We further demonstrate the generality of the method by applying it to a recent dataset of about 100 polyketide synthases [15]. This application also illustrates that our method can predict interaction partners with high accuracy even for relatively small datasets. Finally, our genome-wide predictions of two-component signaling networks across all sequenced bacteria allow us to make an initial investigation of the structural properties of these networks across bacteria.

4.2 Results

4.2.1 General model

Our method in general operates on sets of multiple alignments of homologous proteins (or protein domains) for which it is known that members of one multiple alignment can interact with members of another multiple alignment. To explain the model we first describe it for the simplest possible case. In this situation, illustrated in Fig. 4.1, there are two (large) families of proteins or protein domains, typically with multiple paralogous members per genome, for which it is known that in each genome each member of the first family interacts with one member of the second family. The set of all possible 'solutions' for this problem corresponds to all possible ways in which we can assign, for each genome, each member of the first family to one member of the second family. In Fig. 4.1, the alignments of the two families are shown side by side, with sequences grouped per genome from top to bottom. An assignment of interaction partners a corresponds to a vertical ordering of the sequences within each genome such that the sequences on the same horizontal 'row' are assumed to interact. In this way an assignment a implies a *common* multiple alignment of all sequences of both families.

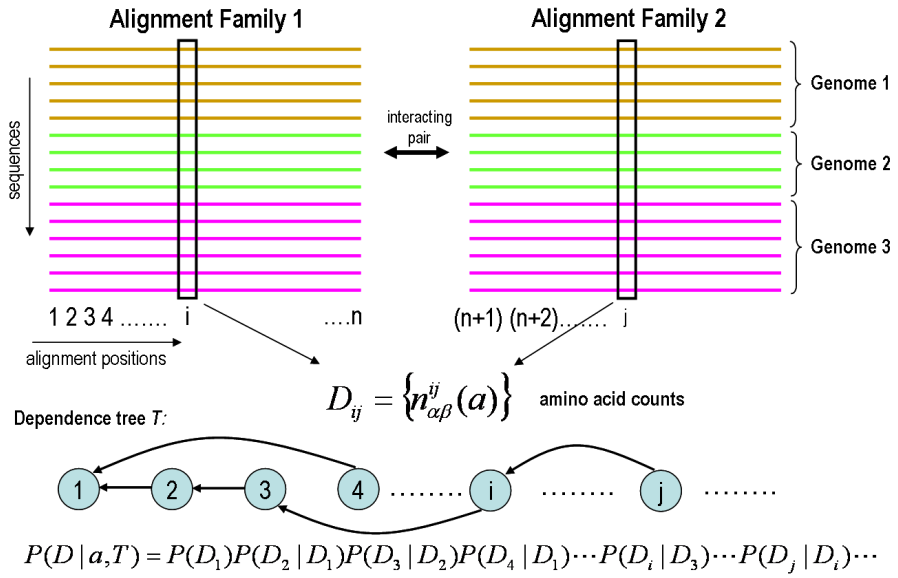


Figure 4.1: Illustration of the model used to assign a probability $P(D|a)$ to the joint multiple sequence alignment D of two protein families given an assignment a of interaction partners between them. Sequences from the same genome have the same color and horizontally aligned sequences are assumed to interact. The probabilities of pairs of alignment columns (ij) depend on the number of times $n_{\alpha\beta}^{ij}$ that amino acids $(\alpha\beta)$ occur in the corresponding columns. A dependence tree T and the corresponding factorization of the probability $P(D|a,T)$ of the entire alignment given the assignment and dependence tree is illustrated at the bottom of the figure.

We now calculate the probability $P(D|a)$ of observing the entire joint multiple-alignment D of the sequences of both families in assignment a. We assume that, for each alignment position i, the probability to observe amino acid α at that position depends on the amino acid β that occurs at one other position $j = \pi(i)$ (the 'parent' of i). A dependence tree T (see Fig 4.1) specifies the parent position $\pi(i)$ for each position i in the joint multiple alignment. The conditional probabilities $p^{ij}(\alpha|\beta)$ are unknown parameters that are integrated out of the problem. As shown in section 4.6, we can derive an explicit expression for the probability $P(D_i|D_j)$ of the entire alignment column i given alignment column j in terms of the counts $n^{ij}_{\alpha\beta}$, the number of times that the pair of amino acids $(\alpha\beta)$ is observed at the alignment columns (ij). The probability $P(D|a,T)$ of the data given dependence tree T is then the product of conditional probabilities $P(D_i|D_{\pi(i)})$ (see Fig. 4.1) over all positions. The unknown dependence tree T is a so called 'nuisance parameter' and probability theory specifies [65] that to obtain $P(D|a)$ we should *sum* $P(D|a,T)$ over all possible dependence trees. Using an uniform prior over trees, this amounts to averaging $P(D|a,T)$ over all dependence trees [28]. In cases where this summation is computationally intractable we can also approximate $P(D|a)$ by finding the dependence tree T^* that maximizes $P(D|a,T^*)$ (see section 4.6).

We sample the posterior distribution $P(a|D)$ over all possible assignments a using Markov chain Monte-Carlo sampling and keep track of the fraction $f(m,m')$ of sampled assignments in which proteins m and m' are interaction partners. In the limit of long sampling the frequencies $f(m,'m)$ give the posterior probabilities $P(m,m'|D)$, that m and m' interact. As explained in section 4.6 this approach can be extended in several ways, including allowing more than two paralogous families, and allowing for unequal numbers of members in the different families. These extensions are used for our predictions of two-component interactions below.

4.2.2 Application to two-component systems

Bacterial two-component systems (TCSs) are responsible for most of the signal transduction underlying complex bacterial behaviors [48, 49, 52]. Although a lot is known about the TCS signaling for specific subsystems in a few model organisms, the interaction partners for the vast majority of TCS genes have not been determined experimentally. Comprehensive predictions of TCS signaling interactions would thus provide important insights into how different bacteria respond to their environments, which regulons are under the control of which external signals, and which specific subsystems are connected by signaling pathways, with potentially important applications. For example, as TCS signaling is essential for host-pathogen interaction, insights in these interactions may have important applications related to human health. In addition, very little is currently known about the global structure of TCS signaling networks across bacteria. With about 400 fully-sequenced genomes available, com-

prehensive prediction of TCS signaling networks across all bacteria would thus also provide a significant data set for studying the global structure of signaling networks in bacteria.

In its simplest form, a two-component system consists of two proteins, a histidine kinase and a response regulator [48]. The histidine kinase is in many cases a membrane-bound protein containing an extracellular sensor domain, which responds to environmental cues, and a cytoplasmic kinase domain. The kinase domain autophosphorylates upon the activation of the sensor, interacts very specifically with the response regulator, and transfers the phosphate to the regulator's receiver domain. Phosphorylation typically leads to the activation of the regulator, which often acts as a transcription factor.

For several reasons two-component systems are particularly attractive for computational modeling. First, both histidine kinase and receiver domains exhibit significant sequence similarity and they can be easily detected in fully-sequenced genomes using hidden Markov models [23]. Second, because two-component systems are very abundant in the bacterial and archeal kingdom, with dozens of interacting pairs in some genomes and thousands of examples across all genomes, they provide enough data to detect subtle dependencies between the residues of interacting kinase/receiver domains. Finally, a significant fraction of all two-component systems form so-called *cognate pairs* in which a single kinase/regulator pair lies within one operon in the genome. It is generally assumed that such cognate pairs are interacting kinase/regulator pairs, which is supported experimentally for a substantial number of pairs, and there are, to our knowledge, no examples that contradict this assumption. Therefore, the cognate pairs provide a very large data-set of known interacting pairs that can be used to test the accuracy of the computational predictions. Additionally, they can be used as a 'training set' for predicting interactions between all other kinases and regulators, i.e. between 'orphan' kinases and regulators which do not occur within an operon with their interaction partner.

We gathered an exhaustive collection of two-component system proteins from 399 sequenced bacteria and multiply aligned all kinase and receiver domains. Whereas all receiver domains can be aligned in a single alignment, kinases show different domain architectures and we produced 7 separate multiple alignments for the 7 most abundant kinase domain architectures (see section 4.6). We also divided the kinases and regulators into cognate pairs and orphans.

4.2.3 Determining interacting residues

The HisKA class is by far the largest class of kinases, with 3388 cognate HisKA/regulator pairs, corresponding to 72% of all cognate pairs, and we first investigated the evidence for dependencies between the amino acid positions of the kinase and the receiver domains of this class. For each pair of positions (ij), where i lies in the kinase and j

in the receiver, we quantified the 'dependence' by the likelihood ratio R_{ij} between a model that assumes the amino acids at these positions are drawn from some joint probability distribution and a model that assumes they are drawn from independent distributions (see section 4.6). This measure R_{ij} for dependence between positions i and j is closely related to the mutual information of the observed distribution of amino acids in positions i and j, which in turn is related to the statistical coupling between positions introduced in [21]. As shown in the top left panel of Fig. 4.2, almost 15% of all pairs of positions have a positive $\log(R_{ij})$, which corresponds to over 1000 pairs. However, because our data set contains many examples of orthologous cognate pairs, we expect to see 'spurious' correlations that are just the result of evolutionary relationships between orthologous pairs.

To investigate whether the high observed $\log(R_{ij})$ values can be explained by phylogeny alone we performed the following randomization. We collected sets of orthologous cognate pairs into orthologous groups and identified pairs of orthologous groups that occur in the same genomes. We then swapped kinase/regulator assignments between such pairs of orthologous groups. Thus, each kinase is now assigned to a *wrong* receiver domain but the phylogenetic relations of all these 'false pairs' are exactly the same as the phylogenetic relationships of the true cognate pairs. If all correlations were due to phylogeny, the distribution of observed R_{ij} values for the false pairs should be the same as that of the true pairs. As the top left panel of Fig. 4.2 shows, the observed R_{ij} values for true pairs are much larger than can be explained by phylogeny. For example, only about 7% of false pairs show positive $\log(R_{ij})$ and there are no false pairs with $\log(R_{ij})$ larger than 235.

If the pairs of positions with large R_{ij} values reflect physico-chemical constraints, we may expect that they are in close physical contact during the interaction of kinase and receiver. Although no structure of a HisKA kinase/regulator pair is currently available, the structure of the sporulation histidine phosphotransferase Spo0B with the response regulator Spo0F [66] has been determined. Spo0B differs significantly in sequence from HisKA kinases, but can nonetheless be reasonably aligned to the HisKA Pfam-profile. We used the Spo0B/Spo0F structure together with the Spo0B/HisKA alignment to estimate the physical distances between all pairs of positions in HisKA kinase/receiver pairs. The top right panel of Fig. 4.2 shows that the pairs of positions with highest R_{ij} are significantly closer physically than other pairs (rank-sum test p-value 3×10^{-11}). In addition, Fig. 4.3 shows the pairs of amino acids with the highest R_{ij} values on the Spo0B/Spo0F complex (black lines).

It is striking that many of the positions that are predicted to depend on each other are indeed in close physical contact in the alpha-helices of the kinase and receiver domains (near the top right of the figure). Other interactions are predicted to occur between residues in an alpha-helix of the kinase domain and residues in loops of the receiver domain. A few of the predicted interactions are more puzzling: they involve residues not in close proximity but the R_{ij} values are too high to be explained

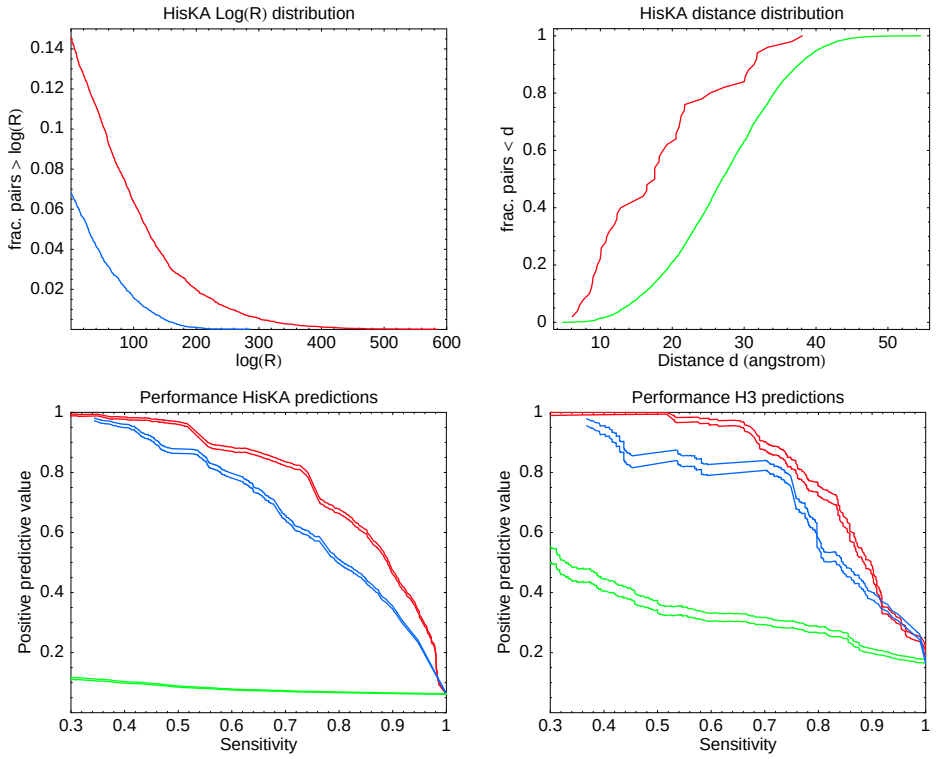


Figure 4.2: Analysis of cognate pairs for the HisKA and H3 kinase classes. **Top left panel:** The red line shows the tail of the reverse cumulative distribution of $\log(R_{ij})$ (dependency) values for pairs of positions in cognate HisKA kinase/receiver pairs. The blue line shows the tail of the $\log(R_{ij})$ distribution after randomizing kinase/receiver assignments in such a way that all phylogenetic relationships are maintained. **Top right panel**: The cumulative distribution of estimated (see text) distances between the amino acids in the co-crystal for the 50 pairs with highest R values (red line) vs all other pairs (green line). **Bottom left panel**: Sensitivities and positive predictive values of the predictions for cognate HisKA kinases and regulators. The red curves show the performance of the model in which $P(D|a,T)$ is averaged over all dependence trees, the blue curve shows the performance of the model $P(D|a,T^*)$ that uses only the best dependence tree, and the green line shows the performance of random predictions. All pairs of curves show estimated PPV plus and minus one standard error. **Bottom right panel**: Performance results as in the bottom left panel for cognate H3 kinases and regulators.

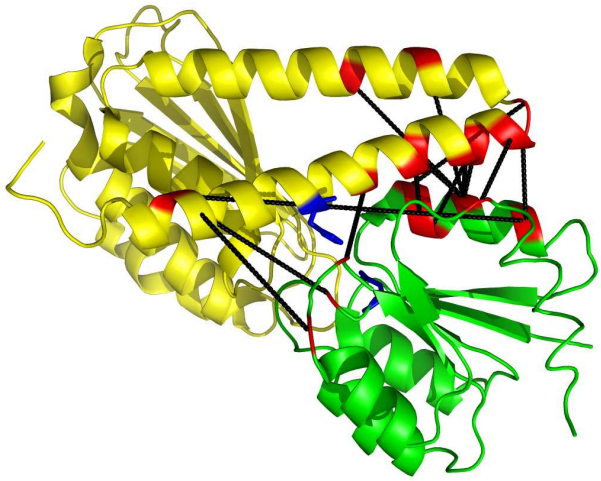


Figure 4.3: Complex of the histidine phosphotransferase Spo0B (yellow) with the response regulator Spo0F (green) [66]. Only one half of the Spo0B-dimer is shown. The site of autophosphorylation in Spo0B and the phosphorylation site in Spo0F are shown in blue. Out of the 20 HisKA/receiver pairs of residues with highest $\log(R_{ij})$, 17 are shown as black lines (3 cannot be displayed because the residues fall in gaps of the alignment with Spo0B). Amino acids marked in red are part of at least one of these 17 pairs.

by phylogenetic dependencies. Some of these may be due to structural differences between the Spo0B/Spo0F complex and the HisKA/receiver complex, to alignment errors, or to indirect dependencies. In summary, the control for phylogenetic signal, the distances between pairs with high R_{ij}, and their location on a related structure all support that our R_{ij} scores capture meaningful functional dependencies between individual pairs of positions in kinase and receiver.

4.2.4 Predicting cognate interactions

We next investigated how accurately the model can reconstruct known cognate pairs of HisKA kinases and their regulators. We collected the multiple alignments of all HisKA kinase domains and receiver domains from cognate pairs and sampled the space of all possible assignments, i.e. all ways in which each kinase from each genome can be assigned to one regulator from the same genome. We sorted all predicted pairs by their posterior probability and measured, as a function of a cut-off in posterior probability, the fraction of all true cognate pairs that are among the predictions (sensitivity) and the fraction of all predictions that correspond to true cognate pairs (positive predictive value). These results are shown in the bottom left panel of Fig. 4.2 both when approximating $P(D|a)$ using the tree with highest probability, i.e. $P(D|a) = \max_T P(D|a,T)$ (blue curves) and when averaging over all dependence trees $P(D|a) = \sum_T P(D|a,T)$ (red curves). In the first approach, the dependence

tree structure is calculated from the correctly paired cognate pairs prior to sampling whereas in the second approach, *no training set* is used at all. In both approaches, the cognate pairs are reconstructed with high accuracy, but averaging over dependence trees performs clearly the best. This is not surprising since, as mentioned above, averaging over dependence trees is the correct way of treating the nuisance parameter T. Using only the best tree may amount to overfitting.

At 60% sensitivity more than 95% (red curves) of the predictions correspond to true pairs. At a sensitivity of 75% the fraction of predictions that are true pairs is still higher than 80% (red curves). This high accuracy is very striking, particularly considering that the algorithm is not given a single example of a true interacting pair, but infers all the cognate pairs in all genomes in parallel by searching for assignments that maximize the amount of dependency observed between the kinase and receiver sequences. We also predicted interaction partners for all cognate kinases and regulators of the H3 class, which is the second most abundant class (Fig. 4.2, bottom right panel). In contrast to the HisKA class, for the H3 class there is a significant number of genomes with only a small number of H3 cognate pairs for which even random predictions would yield a reasonable fraction of correct predictions (green curves). However, it is still clear that our model reconstructs the cognate pairs with high accuracy, i.e. at a sensitivity of 80% more than 95% of the predictions (red curves) correspond to true pairs. In the supplementary material (section 4.7) we show analogous curves for the other (smaller) classes of kinases which all show high accuracy of predictions illustrating that the model can attain high accuracy on relatively small datasets. On the other hand, since for these smaller kinase classes there are often only a few cognate pairs per genome, the prediction problem is of course significantly easier. In summary, the results on cognate pairs suggest that, at least for cognate kinases and regulators, our algorithm can infer interaction partners *ab initio* with high accuracy.

4.2.5 Predicting orphan interactions

We are of course most interested in reconstructing those parts of bacterial two-component signaling networks that are currently not known, i.e. to predict interaction partners for the thousands of orphan kinases and regulators. The prediction of orphan interactions is more difficult for two reasons. First, although for cognate pairs the assumption that each kinase and each regulator interacts mainly with one partner is probably not unreasonable, for orphan kinases and regulators this is less likely to hold. Many genomes contain unequal numbers of kinases and regulators, suggesting that at least some must interact with multiple partners. Second, a given bacterium typically contains orphan kinases from multiple classes, and we thus also have to infer which kinase class each of the orphan regulators belongs to.

In order to predict orphan interactions we extended our model in several ways.

First, we treat the multiple classes of kinases in parallel. Second, to account for unequal numbers of orphan kinases and orphan regulators, for a given assignment some kinases and/or regulators may remain without an interaction partner and these are scored separately (see section 4.6). Finally, we add all the cognate pairs to the alignments of each class, with interaction partners correctly assigned, and keep these cognate pairs *fixed*. In this way the 'frozen' cognate pairs act as an training set for the orphan assignments. The algorithm again uses Markov chain Monte-Carlo to sample over all ways of assigning orphan receivers to classes, and all ways of assigning orphan interaction partners in each class. Due to numerical difficulties in the extension of our model to multiple classes (see section 4.6), we are unable to calculate the sum over all dependence trees with enough accuracy. Therefore, we use the cognate pairs to determine the best dependence tree and approximate $P(D|a)$ with $\max_T P(D|a,T)$.

To benchmark the performance of this extended model we first used it to predict interacting partners for all cognate kinases and receivers, running on all 7 classes in parallel. Since each cognate regulator is now allowed to switch dynamically between all 7 classes of kinases the search space of the extended model is much larger compared to the case in which each class is treated separately and we expect this to negatively affect the performance. As shown in the supplementary material (section 4.7), our predictions nonetheless remain quite accurate. Note also that for small classes, such as the HWE class, there is often only one kinase per genome and correct prediction amounts to identifying the regulator that belongs to the HWE class, which the extended model accomplishes with high accuracy.

Using our extended model, we then predicted orphan interaction partners genome-wide in all 399 bacteria. Currently very few orphan interactions have been measured experimentally. By far the most extensive knowledge is available for the interaction partners of HisKA orphan kinases in *Caulobacter crescentus* [59, 60, 67, 68]. Table 4.1 compares our orphan interaction predictions in *Caulobacter* with those in the literature.

Strikingly, for 10 of the 11 kinases with known interaction partners the top computational prediction corresponds to a known interaction. In fact, of the 22 predictions in the table, which includes *all* 16 known interactions for these kinases, only 5 are at odds with current experimental data. Since there are 29 different orphan regulators in *Caulobacter*, i.e. there are 29 interaction candidates for every kinase, this constitutes highly significant evidence that our method accurately predicts orphan interaction partners (p-value of $7.5 \cdot 10^{-18}$, see section 4.7). In the supplementary material (section 4.7), we also compare our orphan predictions with the few experimentally determined orphan interactions in *Helicobacter pylori*, *Bacillus subtilis*, and *Ehrlichia chaffeensis*.

kinase	regulator	posterior	se	exp evidence
CC0248	CC0247	1.0000	0.0000	putative cognate pair
CC0289	CC0294	0.9948	0.0015	*in vitro* phosphorylation [60]
CC2755	CC2757	0.8507	0.0585	putative cognate pair
CC2765	CC2766	1.0000	0.0000	*in vitro* phosphorylation [60]
CC2932	CC2931	0.9445	0.0059	putative cognate pair
CenK	CenR	0.9168	0.0545	*in vitro* phosphorylation [60]
CckN	DivK	0.3063	0.0357	yeast two-hybrid screen [59]
ChpT	CC3477	0.6074	0.0844	false positive, *in vitro* phosphorylation [67]
ChpT	CtrA	0.1965	0.0627	*in vitro* phosphorylation [67]
ChpT	CC2757	0.1281	0.0555	false positive, *in vitro* phosphorylation [67]
ChpT	CenR	0.0670	0.0450	false positive, *in vitro* phosphorylation [67]
ChpT	CpdR	0.0009	0.0008	*in vitro* phosphorylation [67]
DivJ	CtrA	0.4609	0.0451	*in vitro* phosphorylation [68]
DivJ	PleD	0.3854	0.0323	*in vitro* phosphorylation [60]
DivJ	DivK	0.0409	0.0078	*in vitro* phosphorylation [60]
DivL	DivK	0.5374	0.0582	yeast two-hybrid screen [59]
DivL	CC3477	0.1340	0.0514	not known
DivL	CtrA	0.1298	0.0233	*in vitro* phosphorylation [68]
PleC	DivK	0.0805	0.0145	*in vitro* phosphorylation [60]
PleC	CtrA	0.0020	0.0005	false positive, *in vitro* phosphorylation [60]
PleC	CC3477	0.0013	0.0007	false positive, *in vitro* phosphorylation [60]
PleC	PleD	0.0009	0.0002	*in vitro* phosphorylation [60]

Table 4.1: Comparison of our predictions for orphan HisKA kinases and orphan receivers with experimentally determined interactions in *C. crescentus*. For all orphan HisKA kinases (first column) with at least one known interaction, we show all predicted interaction partners (second column) ordered by posterior probability (third column) up to and including all the known interaction partners. The posterior probability has been averaged over 20 simulation runs, and its standard error is shown in the fourth column. Predictions supported by experimental data are shown in green, predictions not supported by the experimental data (false positives) in red, and predictions supported only by yeast two-hybrid data are shown in blue. Putative cognate pair means that, although we classified the kinase and regulator as orphans, they are less than 2 genes apart on the genome and are orthologous to cognate pairs in closely related genomes. These pairs are very likely to interact and are thus also considered as know interaction partners and colored in green.

4.3 Prediction of interactions between polyketide synthases

Polyketide synthases (PKSs) are a family of bacterial proteins with extraordinary biosynthetic capabilities. Depending on very specific protein-protein interactions, they form multi-protein chains in which the order of the PKS proteins determines the order of monomers of the synthesized polyketide product. PKSs are of particular interest as, through genetic engineering of new PKS chains, they can potentially be used to achieve combinatorial biochemistry in the laboratory [69].

The specificity of PKS interaction is believed to be determined by a small number of residues in the head (N-terminal) and tail (C-terminal). Here we focus on a dataset of 149 interacting head-tail pairs published very recently [15]. Analysis of this dataset has shown [15] that both head and tail sequences can be phylogenetically clustered into three groups (H1 through H3 and T1 though T3), and that interacting pairs only occur between proteins from corresponding groups. Group membership can thus be used to predict which head and tail pairs are likely to interact.

We apply our method without any modification (i.e. as described in section 4.2.1) to the above-mentioned dataset. That is, we consider heads and tails as the protein families 1 and 2 (see Fig. 4.1) and sample over all possible ways of assigning every head to exactly one tail within the same genome. This implies that heads of PKSs within one pathway are allowed to interact with tails of PKSs of a different pathway as long as they belong to the same genome, which is a harder and probably more biologically relevant problem than the one considered in [15]. The results are shown in the left panel of Fig. 4.4. The red curve shows the performance of our model in which the probability of the data is averaged over all possible dependence trees, the blue curve shows the performance of a classification model that only takes into account the phylogenetic group information of the sequences (see section 4.7) and the green curve shows the performance of random predictions. Note that although our model does not take into account any prior information about the phylogenetic grouping of heads and tails, it clearly outperforms the classification model used in [15].

In [15], it was also shown that within the largest group of interacting head-tail pairs (the H1-T1 group containing 90 pairs) there are a number of amino acid residue pairs that lie close in the NMR structure of an interacting head-tail pair and that show significant evidence of co-evolution. However, attempts in [15] to use these pairs of positions to predict interactions within the H1-T1 subclass yielded results that were only slightly better than random. In contrast, as shown in the right panel of Fig. 4.4, our model shows excellent prediction accuracy on the H1-T1 subclass. This demonstrates that at least for some protein families our model obtains accurate predictions on datasets with less than one hundred sequences.

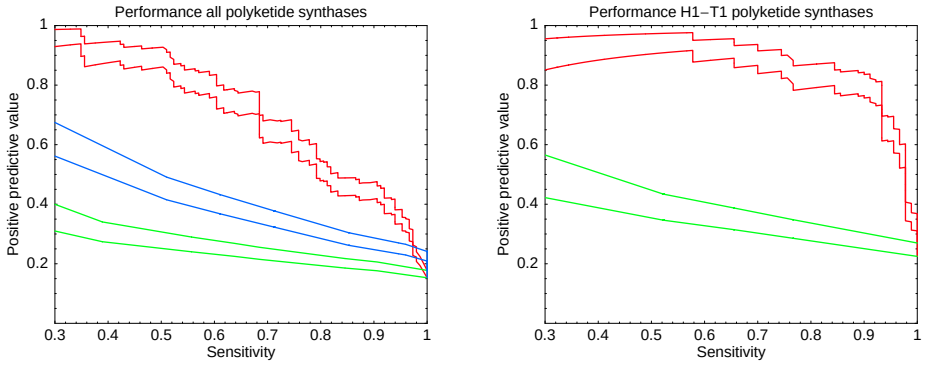


Figure 4.4: Performance of predicted head/tail interactions for polyketide synthases. **Left panel:** Sensitivities and positive predictive values of the predictions for all polyketide synthases in the dataset of [15]. The performance of our model in which $P(D|a,T)$ is averaged over all dependence trees is shown in red. The blue curve shows the performance if only the class information of heads and tails is used (see section 4.6) and the green line shows the performance of random predictions. All pairs of curves show estimated PPV plus and minus one standard error. **Right panel**: Same as the left panel, but predictions restricted to the H1-T1 subclass.

4.4 The structure of two-component signaling networks across bacteria

Our genome-wide predictions of TCS signaling interactions allow us, for the first time, to investigate and compare the structure of TCS signaling networks across bacteria. However, in our cognate predictions above, we assumed each cognate to interact with only one other cognate, and the orphan predictions also assumed that orphans interact only with each other. As explained in section 4.6, to ensure that the network predictions are as comprehensive and unbiased as possible, we used a static scoring scheme that treats cognates and orphans equally (allowing for interactions between orphans and cognates) and allows an arbitrary number of interaction partners per protein.

Before investigating the predicted interactions we first investigated how the number of TCS *genes* of different types varies across genomes. As was shown in [70], the total number of TCS genes varies significantly between bacteria and scales approximately as the square of the number of genes in the genome, i.e. whenever the total number of genes doubles the total number of TCS genes roughly quadruples. Figure 4.5 shows the total number of cognates and orphans across genomes (left panel) and the number of orphan kinases and orphan receivers (right panel). There is a remarkably large variation in the relative number of orphans and cognates, i.e. there are examples of genomes with tens of cognate pairs without any orphans, and vice versa genomes that have tens of orphans and no cognates. In addition, there appears to be little correlation between the number of cognates and the number of orphans. We also

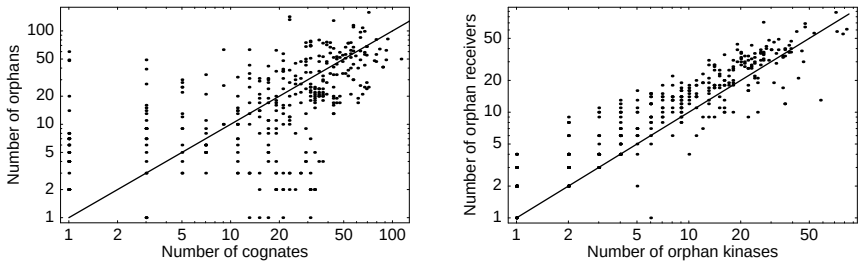


Figure 4.5: Total numbers of cognates, orphans kinases, and orphan regulators across 399 sequenced bacterial genomes. Left panel: The total number of cognates (horizontal axis) versus the total number of orphans (vertical axis). Right panel: The number of orphan kinases (horizontal axis) versus the number of orphan regulators (vertical axis). Each dot in each panel corresponds to a genome. All axes are shown on logarithmic scale. To be able to show genomes with zero genes in one or more of the categories 1 was added to each count, i.e. one on the axis corresponds to a count of zero.

find no discernible correlation between the number of orphan kinases and the number of cognate regulators, or the number of orphan regulators and cognate kinases (data not shown). In contrast, as noted before [71], there is a clear correlation between the number of orphan kinases and the number of orphan regulators in a genome (right panel of figure 4.5) . These statistics provide a first suggestion that orphan kinases and orphans regulators might predominantly interact with each other rather than with cognates.

To investigate this further we analyzed how the total number of predicted interactions depends on the number of TCS genes of different kinds. We distinguish four types of interactions: cognate-cognate interactions between cognate kinases and cognate receivers, orphan-orphan interactions between orphan kinases and orphan receivers, cognate-orphan interactions between cognate kinases and orphan receivers, and orphan-cognate interactions between orphan kinases and cognate receivers. For a genome with C cognate pairs, K orphan kinases, and R orphan receivers there are, respectively $T = C^2$ cognate-cognate, $T = KR$ orphan-orphan, $T = CR$ cognate-orphan, and $T = KC$ orphan-cognate interactions possible. For each genome we determined the fractions f_{cc}, f_{oo}, f_{co}, and f_{oc} of all possible interactions in each class that are predicted to occur. For each category we sorted the genomes by the total number of interactions T of that category, and by calculating running averages of the fractions (see section 4.6) we determined the dependence of the fractions f_{cc}, f_{oo}, f_{co}, and f_{oc} on the total number of possible interactions T (Fig. 4.6). If each possible interaction had a constant probability of being predicted, then the observed fraction of interactions would be independent of the total number of possible interactions T. In contrast, Fig. 4.6 shows that all fractions decrease as a function of the total number of possible interactions T. To a reasonable approximation all four fractions fall as a power-law of the total number of possible interactions T, with exponents -0.4 for

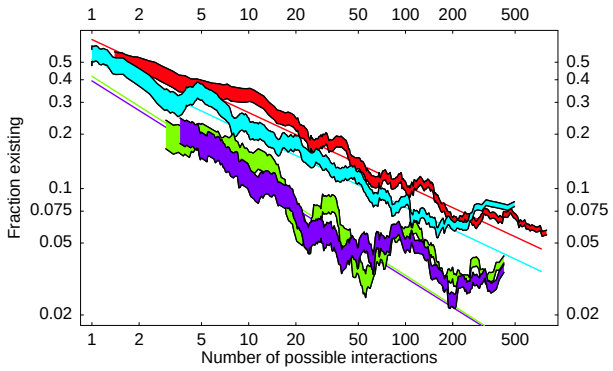


Figure 4.6: The fractions of interactions between cognates (red), between orphan kinases and orphan regulators (light blue), between cognate kinases and orphan regulators (green), and between orphan kinases and cognate regulators (purple) that are predicted to exist (vertical axis), as a function of the total number of possible interactions (horizontal axis). Both axes are shown on logarithmic scales. The values on the vertical axis were obtained by ordering genomes by the total number of interactions of each type, and taking running averages over 25 consecutive genomes. The widths of the curves correspond to two standard errors. The straight lines are power-law fits to the raw data and are given by $f_{cc} = 0.63T^{-0.4}$, $f_{oo} = 0.50T^{-0.38}$, $f_{co} = 0.41T^{-0.55}$, and $f_{oc} = 0.39T^{-0.55}$.

cognate-cognate and orphan-orphan interactions, and -0.55 for cognate-orphan and orphan-cognate interactions.

To investigate the consequences of this scaling for TCS network structure as a function of genome-size, let us first focus on cognate-cognate interactions. For a genome with N cognate pairs there are $T = N^2$ possible interactions of which a fraction $T^{-0.4}$ exist. The total number of cognate-cognate edges thus scales as $T^{0.6} = N^{1.2}$. That is, as the number of cognate pairs increases, the total number of interactions between cognates grows just a bit faster than linear. This implies that, although the total amount of cross-talk between cognates is small, the amount of cross-talk grows with the number of cognate pairs. In particular, the average number of interaction partners per cognate gene grows as $N^{0.2}$. To give an idea of the order of magnitude, for a genome with 4 cognate pairs the power-law fit predicts a total of 3.5 interactions, i.e. essentially one interaction per gene. For a genome with 40 cognate pairs a total of 56 cognate-cognate interactions are predicted, which amounts to 16 cross-talks on top of the 40 cognate interactions. For orphan-orphan interactions the numbers are very similar.

The power-law fits show that the fractions of cognate-orphan and orphan-cognate interactions decrease even faster with T. Consider for simplicity genomes with N cognate pairs, N orphan kinases, and N receivers. The total number of cognate-orphan and orphan-cognate interactions grows as $N^{0.9}$ in such genomes. Since this is slower than linear, it in particular implies that the average number of cognate-orphan and orphan-cognate interactions *per gene* decreases as $N^{-0.1}$. Apart from

decreasing more rapidly with N, Fig. 4.6 also shows that cognate-orphan and orphan-cognate interactions are much less frequent than cognate-cognate and orphan-orphan interactions.

In summary, all our observations support the idea that orphans and cognates form two relatively separate TCS signaling networks, i.e. cognate-orphan and orphan-cognate interactions are relatively rare, and whereas the number of orphan-orphan and cognate-cognate cross-talks per gene increases with increasing network size, the number of cognate-orphan and orphan-cognate interactions per gene decreases with network size. As we saw above (Fig. 4.5), this idea is also supported by the correlation in the number of orphan kinases and orphan receivers, and the absence of correlations between the numbers of cognates and numbers of orphans.

To provide additional evidence that orphans and cognates form relatively separate TCS signaling networks, we mapped orthology relations of cognates and orphans across the 399 sequenced genomes (see sections 4.6 and 4.7). We find that, whenever both genes of a cognate pair have orthologs in another genome, the two orthologs are also a cognate pair in this genome 99.1% of the time. In 0.6% of the cases the orthologs of the cognate pair are both orphans, and in the remaining 0.3% of the cases one ortholog is a cognate and the other an orphan. In cases where only the kinase of the cognate pair has an ortholog the orthologous kinase is a cognate 79% of the time. Similarly, if only the receiver of the cognate pair has an ortholog, then this orthologous receiver is a cognate 78% of the time. Finally, orthologs of orphan kinases are orphans 86% of the time, and orthologs of orphan receivers are orphans 80% of the time. Thus, although both cognate and orphan TCS genes undoubtedly share a common phylogenetic ancestry, our results intriguingly suggest that on shorter evolutionary time scales orphans and cognates evolve relatively separately from each other, and support our finding that the orphans and cognates form two relatively separate interaction networks.

To shed some light on the difference between orphans and cognates, we determined the connectivity, i.e. the number of predicted interaction partners, for each TCS protein and calculated the distribution of connectivities separately for all orphans and all cognates. Figure 4.7 shows the reverse cumulative distribution of kinases (left panel) and regulators (right panel). The figure shows striking differences between the connectivity distributions of cognates (red) and orphans (blue). First, for both kinases and regulators, the reverse cumulative distribution initially falls rapidly and roughly exponentially. In this regime, which includes roughly 90% of all genes, the connectivity distributions of cognates and orphans are very similar, although there are slightly more cognates with at least 1 predicted interaction partner than orphans. However, for the remaining 10% of genes the connectivity distributions of cognates and orphans are very different. In particular, there is a much larger number of orphans with high connectivity. For all four curves, but especially clearly for the orphans, there are two regimes in the distribution: one corresponding to relatively

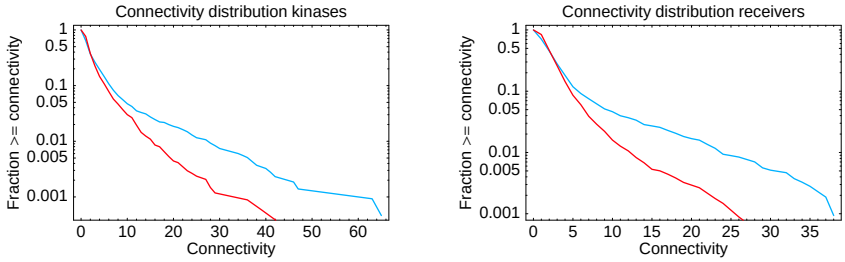


Figure 4.7: Reverse cumulative connectivity distributions of kinases (left panel) and receivers (right panel). The fraction of genes with at least a given number of interaction partners (connectivity) is shown as a function of the connectivity. Cognates are shown in red and orphans in blue. The vertical axis is shown on a logarithmic scale.

low connectivity genes, which includes about 90% of all genes, and a second regime of high connectivity genes which covers the remaining 10%. It thus appears that, to a rough approximation, there are two types of two-component system genes. Most kinases and regulators interact with only a few (less than 5) partners, but about 10% interacts with a large number of partners. The kinases in this class thus distribute a signal to a large number of downstream regulators, and the regulators in this class integrate a large number of input signals. Most of these 'hub' kinases and regulators are orphans.

4.5 Discussion

We have presented a novel general Bayesian network model for predicting interactions between families of interacting protein domains directly from amino acid sequences. Our method incorporates several important methodological advances. First, the model does not require any training sets, but predicts interactions *ab initio* by sampling the space of all possible interaction assignments. For each interaction assignment the probability of the data is derived from first principles, i.e. without any tunable parameters, and *sums* over all possible ways in which a tree of dependencies can be assigned to pairs of residues both within and between the interacting proteins [28]. The latter is an important feature of the model. One might think that dependencies between residues within one protein are immaterial for the interaction with the other protein and that equal or even better performance could be obtained by simply summing the dependencies of only those pairs of residues that go *between* the two interacting proteins. This is however not the case as the following example illustrates. Imagine two residues r and r' in the first protein that both show clear dependence on a single residue q in the other protein, but that show even larger dependence on *each other*. Obviously, in this case it would be wrong to assume that the observed dependencies of q with both r and r' are evidence that both r and r' interact

directly with q. Rather, q presumably interacts only with one of the these residues (say r) and the apparent dependency with r' is a result of the strong dependency of r and r' with each other. In contrast, if r and r' were to show no dependency, then the observed dependency of q with both r and r' *would* provide evidence that both residues interact with q. That is, the 'meaning' of the dependency between any pair of residues depends subtly on the dependency that these residues have with all other residues and summing over dependence trees is the probabilistically correct way of taking all dependencies into account. Other important features are that we assign interaction partners for all proteins from all genomes in parallel, thereby maximizing the algorithm's ability to detect subtle sequence dependencies, and the use of Markov chain Monte-Carlo sampling to automatically obtain a measure of the reliability of each prediction.

Here we have applied our method to two bacterial protein families, two-component system signaling proteins and polyketide synthases, which provide quite different challenges. In the case of the TCSs we have thousands of examples, allowing for the detection of subtle statistical signals. However, since the kinases naturally divide into subfamilies and receivers do not, receivers need to be both classified and matched to their interaction partners at the same time. In the case of the polyketide synthases, we are dealing with only on the order of 100 homologous proteins, which makes the detection of dependencies between amino acid residues much more difficult and requires careful statistical modeling. The fact that our algorithm successfully predicts interaction partners for both datasets demonstrates the generality of the method.

Our predictions of two-component interactions provide the first reconstruction of genome-wide signaling networks across all currently sequenced bacteria and our results suggest that these predictions have high accuracy (Fig. 4.2 and table 4.1). All predictions for each genome are available at the SwissRegulon website (http://www.swissregulon.unibas.ch/cgi-bin/TCS.pl). Our predictions allow us to perform a first analysis of the structure of TCS signaling networks across bacteria. First, we find that the average connectivity per gene increases slowly but significantly with the number of nodes in the network. Intriguingly, we find that cognates and orphans form two relatively independent groups, with cognates interacting predominantly with cognates and orphans predominantly with orphans. The latter observation is supported by an analysis of orthology relations which showed that, at least on shorter evolutionary time scales, cognates and orphans evolve relatively independent of each other. Another significant finding is that, whereas 90% of TCS genes have a relatively small number of interaction partners, 10% of orphans form a distinct class of 'hub' nodes in the signaling networks which have large numbers of interaction partners.

The finding that cognate and orphan TCSs form two relatively independent groups is further supported by a recent study by Alm et al. [71]. They showed that kinases that have been horizontally transferred are more likely to be found in an operon with

a response regulator than kinases that have been created by lineage specific expansion. This may partly explain the preferential cognate-cognate interaction as cognate kinases tend to be transferred with their interaction partners. However, it does not explain why 'new' orphan kinases that have been created by duplication, evolve interaction specificity towards orphan regulators and rarely interfere with cognate systems. One may argue that cognate pairs form simple linear stimulus-response pathways that form a functional unit and are expressed (and transferred between genomes) as such. In contrast, TCS signaling in complex behaviors involving multiple in- and outputs, would typically necessitate independent expression of the different components, especially if the processes involve temporal regulation of the interactions. This is in agreement with experimental evidence in *Caulobacter*, where orphans with generally multiple interactions control cell-cycle progression [58] and in *Bacillus subtilis*, where they are involved in sporulation [50]. In addition, our predictions suggest that indeed orphans are more likely than cognates to have high connectivity. However, it is clear that much more investigation is necessary to understand the reasons behind these global differences in interaction propensity between orphans and cognates.

There are many other examples to which our method can now be applied, i.e. whenever there are two or more protein families or protein domains that interact we can apply the method to multiple alignments of these protein families/domains. Some examples to which the method can be applied in an essentially unaltered way are e.g. ABC 'half transporters' [72] or certain subfamilies of cytokines and their receptors [73]. Our results for the family of polyketide synthases suggest that accurate predictions can also be obtained for fairly small protein families with on the order of 100 homologous sequences. However, the minimal number of sequences needed for reliable predictions is very difficult to estimate as it depends on many different factors. One of them is the entropy of the amino acid distribution at different positions in the alignments, which has a strong influence on the strength of the co-evolutionary signal. For example, if only charged amino acids appear at two particular residues and positively charged amino acids preferably pair with negatively charged amino acids and vice-versa, then only a very small number of sequences is needed to detect a dependence (the size of the alphabet is effectively reduced). In general it is probably safe to say that for any successful application at least a few dozen examples are needed, and that a thousand examples should always be sufficient. In any case, as new sequences are becoming available at an ever increasing pace we expect many protein families to become amenable to our analysis in the coming years.

Finally, the concept of dependence tree models may have very general applications. For example, hidden Markov models of protein domains and protein families score multiple alignments by assuming each alignment column is drawn from a weight matrix column that represents the propensities for different amino acids to occur at that position [23]. Our Bayesian network model provides a generalization of such scoring models to take into account dependencies between all pairs of positions in the

Name	Architecture	no.cognates	no.orphans
HisKA	HisKA, HATPase_c	3388	2158
H3	HisKA_3, HATPase_c	636	183
His_kinase	His_kinase, HATPase_c	245	23
Long hybrid	HisKA, HATPase_c, RR, (RR), Hpt	132	286
Short hybrid	HisKA, HATPase_c, RR, (RR)	126	985
Chemotaxis	Hpt, HATPase_c	89	77
Hpt	Hpt	37	192
HWE	HWE or HisKA_2, HATPase_c	34	162

Table 4.2: Pfam domain combinations of the most abundant kinase architectures and the number of times they occur in all 399 genomes. RR stands for the receiver domain profile Response_reg. Both the short and long hybrid architecture can contain one or two receiver domains.

alignment. Our method can thus be very generally applied to multiple alignments of protein sequences, e.g. to infer interactions between residues, to discover subfamilies, and generally to improve multiple alignments of protein domains and families.

4.6 Materials and methods

We extracted the sequences of an exhaustive collection of two-component system proteins from 399 sequenced bacterial genomes in NCBI [1] using histidine kinase and response regulator profiles from the Pfam database [23]. Whereas there is only one Pfam profile for the receiver domains of response regulators, there are 7 different kinds of kinase domains and kinases show a variety of domain combinations. The large majority of kinases falls into one of the 8 domain architectures shown in table 4.2. Multiple alignments of all 8 kinase classes and the entire set of receiver domains were produced using the program hmmpfam (`http://hmmer.wustl.edu/`). For the long hybrid class, we aligned only the Hpt domain as the interaction is believed to take place mainly between this domain and the cognate receiver domain [48]. The ATP-binding domain (HATPase_c) was not aligned as it does not seem to be important for the kinase-receiver interaction [59].

We defined operons as maximal sets of contiguous genes on the same strand of the DNA with all intergenic regions between consecutive genes less than 50 bps in length. Whenever an operon contained only one kinase and one regulator this pair was considered a cognate pair. Kinases(Regulators) that did not sit in an operon with any regulators(kinases) were considered orphan kinases(regulators). We made separate alignments for the 8 sets of receiver domains from cognate regulators that interact with each of the 8 kinase domain architectures. As shown in the supplementary

[1] *ftp.ncbi.nlm.nih.gov/genomes/Bacteria*

material (section 4.7), in accordance with previous results [49, 74], we observe that receiver domains that interact with different types of kinases show distinct amino acid compositions which can be used to predict what kind of kinase each receiver will interact with. Those results also indicated that Hpt and long hybrid receivers are very similar, and for the remainder of the analysis we fused these two classes into a single class.

4.6.1 Bayesian network model

We discuss first the simplest model setting: There are two families of proteins (or protein domains) X and Y that interact and we have multiply aligned all members of families X and Y from all sequenced genomes. We assume each member x of family X has precisely one interaction partner y of family Y in the same genome. An *assignment* a of interacting pairs can be thought of as specifying a joint multiple alignment D of the two families in which interacting members are aligned horizontally (Fig. 4.1).

We calculate the probability $P(D|a)$ of the entire joint alignment given the assignment a and our model assumptions. Let D_i denote the alignment column at position i in the joint alignment, i.e. i runs from 1 to $L = L_X + L_Y$, with L_X and L_Y the lengths of the family X and Y alignments. We first calculate the probability $P(D_i|w)$ of the data D_i in column i given a *weight matrix* (WM) column w:

$$P(D_i|w) = \prod_\alpha w_\alpha^{n_\alpha^i} \tag{4.1}$$

where w_α is the probability of seeing amino acid α at this position, and n_α^i is the number of times amino acid α occurs in column i. Since we do not know the WM we integrate over all possible WMs. Using a Dirichlet prior $P(w) \propto \prod_\alpha w_\alpha^{\lambda-1}$, we have

$$P(D_i) = \int_{\sum_\alpha w_\alpha = 1} P(D_i|w)P(w)dw = \frac{\Gamma(21\lambda)}{\Gamma(n+21\lambda)} \prod_\alpha \frac{\Gamma(n_\alpha^i + \lambda)}{\Gamma(\lambda)}, \tag{4.2}$$

where n is the total number of amino acids in column i and λ is the pseudo-count of the Dirichlet prior. Notice that we treat gap symbols in the alignment simply as a 21st amino acid so that our alphabet size is 21.

Similarly, the probability $P(D_{ij}|w)$ of a pair of columns given a weight matrix for the pair of columns is

$$P(D_{ij}|w) = \prod_{\alpha,\beta} (w_{\alpha\beta})^{n_{\alpha\beta}^{ij}}, \tag{4.3}$$

where $w_{\alpha\beta}$ is the joint probability to see α at position i and β at position j, and $n_{\alpha\beta}^{ij}$ is the number of times the pair of amino acids $(\alpha\beta)$ occurs (on the same row) in

columns (ij) of the alignment. Using again a Dirichlet prior $P(w) \propto \prod_{\alpha\beta} w_{\alpha\beta}^{\lambda'-1}$ and integrating out the unknown weight matrix w, we have

$$P(D_{ij}) = \int_{\sum_{\alpha\beta} w_{\alpha\beta}=1} P(D_{ij}|w)P(w)dw = \frac{\Gamma(21^2\lambda')}{\Gamma(n+21^2\lambda')} \prod_{\alpha\beta} \frac{\Gamma(n_{\alpha\beta}^{ij}+\lambda')}{\Gamma(\lambda')}. \tag{4.4}$$

The conditional probability of column i given column j is given by $P(D_i|D_j) = P(D_{ij})/P(D_j)$. As the supplementary material shows (section 4.7), consistency requires that $\lambda = 21\lambda'$, and we use the Jeffreys' or information geometry prior $\lambda' = 1/2$ (i.e. uniform in the determinant of the Fisher information matrix). As a measure of dependence between two columns i and j we use the ratio of likelihoods of the joint and independent models for the columns

$$R_{ij} = \frac{P(D_{ij})}{P(D_i)P(D_j)} \tag{4.5}$$

For large counts $n_{\alpha\beta}^{ij}$ the logarithm of R is approximately proportional to the mutual information of the amino acid distributions in columns i and j. For small counts the ratio R_{ij} takes into account finite-size corrections. It also takes into account that the dependent model has more free parameters than the independent models. As a result, values of $R_{ij} > 1$ can be interpreted as indicating positive evidence of dependence between positions i and j.

Let T denote a spanning tree in which each node is one of the positions i in the joint alignment. We (arbitrarily) pick one node r to be the root of the tree and direct all edges in the tree toward the root. In this directed 'dependence tree' T each node i (except for the root) will have a single outgoing edge pointing to its 'parent' $\pi(i)$ (see Fig. 4.1). Given an assignment a and dependence tree T we calculate the probability $P(D|a,T)$ of the joint alignment by letting each column i depend on the parent column $\pi(i)$. That is, we have

$$P(D|a,T) = P(D_r) \prod_{i \neq r} P(D_i|D_{\pi(i)},a,T), \tag{4.6}$$

with r the root node and the product is over all nodes except for the root. Using (4.5) we can rewrite this as

$$P(D|a,T) = \left[\prod_i P(D_i)\right] \left[\prod_{i \neq r} R_{i\pi(i)}\right], \tag{4.7}$$

where the first product is over all positions (including the root) and the second product is over all edges in the tree T. Note that only the second product depends on the assignment a and tree T, and that (4.7) is independent of the choice of the root and

orientation of the edges in the tree. Note also that the position $\pi(i)$ that position i depends on may lie either within the same protein or in the other protein.

To calculate the probability of the alignment independent of a particular dependence tree we sum over all $|T|$ possible spanning trees of the L positions:

$$P(D|a) = \frac{1}{|T|} \sum_{T} P(D|a, T). \tag{4.8}$$

As shown in [28], this sum can be calculated efficiently as a matrix determinant. Let M denote the Laplacian of the matrix R

$$M_{ij} = \delta_{ij} \sum_{k} R_{ik} - R_{ij} \tag{4.9}$$

from which one row and column have been removed. We then simply have

$$P(D|a) = \frac{\prod_i P(D_i)}{|T|} \det(M). \tag{4.10}$$

Given a uniform prior, $P(a) = \text{constant}$, over assignments, the posterior probability becomes proportional to the determinant, i.e. $P(a|D) \propto \det(M)$.

4.6.2 Generalization: Orphan predictions

The general model just presented can easily be generalized in various ways. Here we discuss the generalizations that we use to predict orphan interactions. Since genomes have typically different numbers of orphan kinases and orphan regulators we have to relax the assumption that each protein has precisely one interaction partner. Although there are other possibilities, in our implementation we only consider assignments in which each protein is connected to *at most* one other protein at a time. For each genome we assign a number of interactions that is equal to the minimum of the number of orphan kinases and the number of orphan regulators. This typically leaves some proteins without an interaction partner. In addition, since there are 7 kinase classes, with a separate multiple alignment for each, a full orphan assignment consists of 7 joint alignments in parallel.

The probability $P(D|a)$ of the data given an orphan assignment is the product of the probabilities for each of the 7 joint alignments of interacting pairs, the 7 alignments of unassigned kinases, and 7 alignments of the receiver domains of unassigned regulators. That is, we also divide unassigned receivers into 7 classes. Let us focus on a single kinase class. We let J denote the joint alignment of the interacting pairs, with J^k the kinases in the joint alignment and J^r the receivers in the joint alignment. In addition, let K denote the alignment of unassigned kinases, and R the alignment

of unassigned receivers for this class. We now assume that we can factorize the joint probability of this data as follows

$$P(J, K, R) = P(K|J^k)P(R|J^r)P(J). \tag{4.11}$$

In particular, we will assume that the kinases in K were drawn from the same distribution as the kinases in J, and that the receivers in R were drawn from the same distribution as the receivers in J. We again write the conditional probabilities of unassigned kinases and receivers in terms of dependence trees T^k and T^r for the kinase and receiver positions. We then have

$$P(K|J^k, T^k) = \frac{P(K, J^k|T^k)}{P(J^k|T^k)} \tag{4.12}$$

and

$$P(R|J^r, T^r) = \frac{P(R, J^r|T^r)}{P(J^r|T^r)}. \tag{4.13}$$

Note, however, that in both these expressions the numerator and denominator are entirely equivalent to expression (4.7). That is, these conditional probabilities can be calculated, using equations (4.2), (4.4), (4.5), and (4.7), in terms of the counts of the number of times different combinations of amino acids occurs in pairs of positions in the kinases K, the kinases J^k, the receivers R, and the receivers J^r.

We would in principle calculate the probabilities $P(K|J^k)$ and $P(R|J^r)$ by summing over all possible spanning trees T^k and T^r, which involves calculating determinants precisely as in equation (4.10). However, as described in the supplementary material (section 4.7), numerical stability issues with the calculation of these determinants (see [46]) force us to use an approximation when we run multiple kinases/receiver classes in parallel. Instead of calculating determinants we thus approximate $P(K|J^k) \approx P(K|J^k, T^k)$ using the dependence tree T^k that maximizes the joint probability $P(J^k|T^k)$ of all cognate kinases in the class, and approximate $P(R|J^r) \approx P(R|J^r, T^r)$ by using the dependence tree T^r that maximizes the probability $P(J^r|T^r)$ of all cognate receivers in the class. Similarly, for the joint probability $P(J)$ we also approximate $P(J) \approx P(J|T^*)$ where T^* is the dependence tree that maximizes the probability of cognate kinase/receiver pairs in the class.

Finally, it is trivial to incorporate 'training' examples of known interacting proteins in our Bayesian network model. We simply add the known interacting pairs to the alignments and keep these pairs fixed, i.e. they are not sampled over. In our case we added all cognate pairs for each of the 7 classes to the corresponding joint alignments J. In this way the 'frozen' cognate pairs in the alignment act as 'seeds' that are used in sampling orphan assignments.

4.6.3 Gibbs sampling

To calculate the posterior probabilities $P(x, y|D)$ that members x and y interact we sample the distribution $P(a|D)$ using a Markov chain Monte-Carlo method known as Gibbs sampling. Let r_g denote the maximum of the number of orphan kinases and the number of orphan regulators in genome g. We first sample a genome g with probability $P(g) \propto \binom{r_g}{2}$. If the sampled genome has more kinases than regulators we pick two kinases (k_1, k_2) at random and sample over the current assignment and the assignment with the interaction partners of these kinases exchanged. Note that if one kinase is currently unassigned the exchange would cause the other kinase to become unassigned. If both kinases are unassigned the move will leave the current assignment unchanged. If the sampled genome has more regulators than kinases we pick a pair of regulators (r_1, r_2) at random and again sample over the current interaction assignment and the assignment with the interaction partners swapped. If one or both of the regulators are unassigned we also sample over the kinase class that each unassigned regulator is assigned to. That is, if both regulators are assigned we sample over 2 assignments, if one is unassigned we sample over $2*7 = 14$ assignments, and if both are unassigned over $7 * 7 = 49$ assignments. For the cognate predictions of Fig. 4.2 the move-set simplifies since each protein is guaranteed to be assigned to precisely one interaction partner.

For each kinase/receiver pair (x, y) we then determine the fraction $f(x, y)$ of sampled assignments that have x and y assigned as interaction partners. Note that, since we cannot assume that each orphan has only one interaction partner, these fractions cannot be directly interpreted as posterior probabilities of interaction. That is, if a certain kinase interacts 1/4 of the time with each of four different receivers this might simply indicate that this orphan kinase can interact with all four receivers. The results in Fig. 4.2 and in table 4.1 were obtained by performing 10 independent sampling runs in each case, and averaging the observed frequencies $f(x, y)$ from each of the runs.

4.6.4 Phylogenetic permutation test

To assess whether the high correlations seen between amino acid pairs of kinases and receivers in the HisKA class could be explained by phylogeny alone, we constructed a null model that conserves all evolutionary relationships, but associates kinases with non-cognate regulators. We first map orthology relations between all cognate kinase/regulator pairs. Two cognate pairs are considered orthologs when they are best reciprocal hits and align over more than 80% of their lengths with at least 80% amino acid identity. Next we filter out orthologous cliques; sets of orthologous cognate pairs that are all orthologous to each other. The result is a collection of n orthologs groups of cognate pairs. We define the *overlap* of a pair of orthologous groups as the number

of genomes in which the representatives of both groups exist and produce a list of all pairs of orthologous groups sorted by overlap. Starting from the pair with highest overlap we then create multiple alignments of 'true' and 'false' kinase/regulator pairs by applying the following rule for each entry in the list: We first check that both groups of cognate pairs have not yet been used. If not, we extract the sequences from the genomes in which both cognate pairs occur. These cognate pair sequences are added directly to the alignment of 'true' pairs. The same kinase and receiver domain sequences are added also to the alignment of 'false' pairs but now with, in each genome of the group, the kinase of the first cognate pair assigned to the regulator of the second pair and vice versa. In this way the alignments of 'true' and 'false' pairs will consist of the same set of proteins with the precise same phylogenetic relationships between interacting pairs. We then determine R_{ij} for all pairs of positions from both 'true' and 'false' alignments.

4.6.5 Network structure analysis

Owing to the different overall number of TCS genes in the different kinase classes, both the sensitivity and specificity of the predictions will likely vary from class to class. As different genomes have different numbers of TCSs in different classes, combining predictions from all classes might introduce biases in our TCS network analysis. We therefore focus on the by far most common class of HisKA kinases and their receivers for the TCS networks prediction and comparison. We first extracted all HisKA kinases from all genomes and all regulators that interact with HisKA kinases. For the latter we took all regulators in cognate pairs with HisKA kinases as well as all orphan regulators that were classified as HisKA receivers during most of the Monte-Carlo sampling for the orphan predictions.

Whereas the Monte-Carlo sampling is most suited for predicting the most likely interaction partners of each kinase and regulator, it is not well suited for an unbiased inference of the entire signaling network in each genome since the total number of interactions is fixed in each genome to at most one per protein per time point during the sampling. In addition, in the Monte-Carlo sampling only orphan interactions were sampled and cognate interactions where kept fixed. Therefore, to predict genome-wide TCS signaling interactions allowing for an arbitrary number of connections, and treating cognates and orphans in the same way, we use the following procedure.

During the Monte-Carlo sampling runs that were used to predict orphan interaction partners, we also kept track of the numbers $n_{\alpha\beta}^{ij}$ of interacting HisKA kinase/receiver pairs that have the combination of amino acids $(\alpha\beta)$ at positions (ij). By averaging these over the sampling runs we obtain the average counts $\langle n_{\alpha\beta}^{ij} \rangle$ that summarize the amino-acid composition at all pairs of position in predicted interacting HisKA pairs. Using the average counts $\langle n_{\alpha\beta}^{ij} \rangle$ we determined the position dependency statistics R_{ij} and determined three dependence trees T^*, T^k and T^r that each maxi-

mize the sum of $\log(R_{ij})$ along their edges. Whereas T^* takes into account all kinase and receiver positions, T^k only takes into account kinase positions and T^r only the receiver positions, respectively. Finally we estimated the joint probabilities for amino acid combination $(\alpha\beta)$ to occur at positions (ij) as

$$p^{ij}_{\alpha\beta} = \frac{\langle n^{ij}_{\alpha\beta}\rangle + \lambda}{\sum_{\alpha\beta}(\langle n^{ij}_{\alpha\beta}\rangle + \lambda)}. \tag{4.14}$$

The marginal probabilities p^i_α for amino acid α to occur at position i are given by summing the joint probabilities, e.g. $p^i_\alpha = \sum_\beta p^{ij}_{\alpha\beta}$.

Using these joint and marginal probabilities we can then calculate, for any kinase-receiver pair with sequences S_k and S_r, respectively, the log-ratio of the probabilities of their sequences (S_k, S_r) under the dependent model, that describes the probability distribution of all kinase and receiver positions in terms of the optimal tree T^*, and two independent models, that describe the dependencies of the kinase and receiver positions separately, using the optimal trees T^k and T^r, respectively. This ratio $X(S_k, S_r)$ is given by the expression

$$X(S_k, S_r) = F(S_k, S_r|T^*) - F(S_k|T^k) - F(S_r|T^r) \tag{4.15}$$

with

$$F(S|T) = \sum_{(ij)\in T} \log[p^{ij}_{S_i S_j}] - \log[p^i_{S_i}] - \log[p^j_{S_j}] \tag{4.16}$$

where S_i is the amino acid that occurs at position i in the sequence S, and the sum is over all edges in the tree T. For each genome, we calculate the log-ratio $X(S)$ for all kinase-receiver pairs, including both orphans and cognates, and predict an interaction to occur between any pair for which $X(S) \geq 1$. At the chosen (conservative) cut-off of 1, about half of all the predictions between cognate kinases and cognate receivers correspond to cognate pairs (see section 4.7). To calculate the connectivity distribution we counted the number of predicted interaction partners for each TCS gene and obtained reverse cumulative distributions separately for cognate kinases, cognate receivers, orphan kinases, and orphan receivers.

In order to determine the orthology relationships between cognates and orphans, we first extracted the sequences of all kinase domains belonging to HisKA kinases as well as the sequences of all receiver domains of HisKA response regulators. For each kinase or receiver domain we then identified orthologous domains in the 398 other genomes. A domain $\tilde{d}$ is considered an ortholog of domain d when

1. d and $\tilde{d}$ are each other's reciprocal best match.

2. d and $\tilde{d}$ align over 80% of their lengths.

3. d and $\tilde{d}$ are at least 60% identical at the amino acid level.

Under these relatively stringent constraints, we typically find orthologous domains in between 4 and 10 other genomes. We then counted how often the orthologs of cognate pairs are themselves cognates pairs, how often only one of the members of a cognate pair has an ortholog, how often this single ortholog is itself part of a cognate pair and how often it is an orphan, etcetera. These ortholog statistics are shown in section 4.7.

For each genome we determined the number of cognate pairs C, the number of orphan kinases K, and the number of orphan receivers R and determined

1. The fraction f_{cc} of all $T_{cc} = C^2$ possible interactions between cognate kinases and cognate receivers that are predicted.

2. The fraction f_{co} of all $T_{co} = CR$ possible interactions between cognate kinases and orphan receivers that are predicted.

3. The fraction f_{oc} of all $T_{oc} = KC$ possible interactions between orphan kinases and cognate receivers that are predicted.

4. The fraction f_{oo} of all $T_{oo} = KR$ possible interactions between orphan kinases and orphan receivers that are predicted.

For each category of interactions, we ordered all genomes with respect to the total number of possible interactions T. We then calculated running averages of both the f values and T values over windows of 25 consecutive genomes to determine the average dependence of f on T. Standard errors s_e of the running averages of f were also calculated by determining the variance $\mathrm{var}(f)$ of f across the 25 genomes in each window, and are given by $s_e = \sqrt{\mathrm{var}(f)/25}$.

Finally, for each category we fitted f to a power-law function of T as follows. For a genome with T possible interactions of which n are predicted to exist we estimate f as $f = (n+1)/(T+2)$ and logarithmically transform (T, f) to a data-point $(x, y) = (\log(T), \log(f))$. We then fit a linear function $y = ax + b$ to the set of data points (x, y) by finding the line that minimizes the average distance of the data points to the line (which is also the first principal component axis).

4.7 Supplementary material

4.7.1 Classifying receiver domains

Similar to previous work, i.e. [49,74], we found that cognate response regulators that interact with different types of kinases show distinct amino acid compositions in their receiver domains and that these differences can be used to predict, for each receiver domain, what kind of kinase it will interact with.

We divided the multiple alignment of all cognate receiver domains into 8 sub-alignments corresponding to sets of regulators that interact with kinases of each particular kinase class. For each of the 8 alignments we then constructed a position specific weight matrix

$$w^c_{i\alpha} = \frac{n^c_{i\alpha} + \lambda}{\sum_\alpha (n^c_{i\alpha} + \lambda)}. \tag{4.17}$$

Here $n^c_{i\alpha}$ is the total number receivers of class c that have an amino acid α in column i of the alignment (gaps are treated as a 21st amino acid) and λ is the pseudo-count resulting from the Dirichlet prior (we used the Jeffreys' prior $\lambda = 1/2$). $w^c_{i\alpha}$ is thus the estimated probability of seeing amino acid α in position i of a receiver of class c.

Given a receiver with sequence S we can now determine the posterior probability $P(c|S)$ that it belongs to class c. We have

$$P(c|S) = \frac{P(S|c)P(c)}{\sum_{c'} P(S|c')P(c')} \quad \text{with} \quad P(S|c) = \prod_i w^c_{S_i i}, \tag{4.18}$$

where S_i is the amino acid in the ith position of receiver sequence S and the product runs over all positions in the receiver. We assumed a uniform prior $P(c) = 1/8$.

We tested to what extent this simple model is capable of correctly classifying receiver sequences. For each cognate receiver we calculated the posterior probability $P(c|S)$ of the class c given the receiver sequence S, using the WMs $w^c_{i\alpha}$ constructed from all receiver sequences. We then assigned the receiver to the class c that maximizes $P(c|S)$. The results in Fig. 4.8 show that for the three most abundant types of kinases (HisKA, H3, and HisKin), and for the Hwe kinases as well, the classifier predicts almost perfectly which kinase type the respective receivers interact with. For the other classes the classification is still correct in the majority of the cases.

The types of mis-classifications match what is to be expected based on the domain architectures. Both long and short hybrids contain an HisKA domain and their receivers are sometimes mistaken for a receiver that interacts with a single HisKA domain kinase. Both long hybrid kinases and Hpt kinases contain an Hpt domain and the most common misclassification is between receivers that interact with a kinase with a single Hpt domain and receivers that interact with long hybrids. Because of this, and because the number of cognate pairs of the Hpt class is very small, we

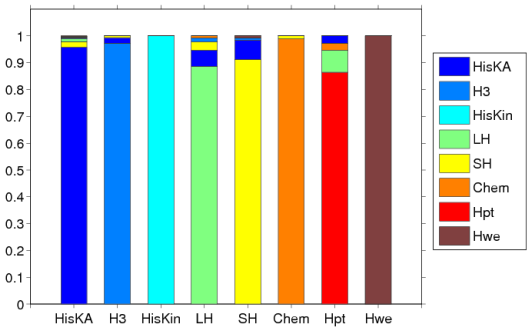


Figure 4.8: Predicted classification of receivers. Each bar represents the set of all receivers that are member of a cognate pair with kinases of a particular type (indicated below the bar). In each bar the colors indicate what fraction of the cognate receivers of this type is classified with each type of kinase. The legend on the right shows the correspondence between color and kinase type. SH and LH stand for short and long hybrid, respectively, and Chem stands for chemotaxis.

have treated the Hpt and long hybrid classes as one class in our analysis (leaving 7 classes in total). Although they also contain an Hpt domain, cognate receivers of chemotaxis kinases are very rarely mistaken with receivers of Hpt and long hybrid kinases, probably due to the fact that cognate regulators of chemotaxis kinases are mainly CheB and CheY regulators which have very specific functions in chemotaxis and correspondingly a specific amino acid composition. Overall, the WM model predicts the correct type of kinase for 96% of the cognate receiver domains.

4.7.2 Details of the Bayesian network model

We first derive why consistency requires that the pseudo-count λ of the Dirichlet prior for the marginal probabilities w_α is related to the pseudo-count λ' of the joint probabilities $w_{\alpha\beta}$ through

$$\lambda = 21\lambda' \tag{4.19}$$

In section 4.6 we calculated expressions for $P(D_i)$ and $P(D_{ij})$ in terms of λ, λ', the joint counts $n^{ij}_{\alpha\beta}$ and the marginal counts n^i_α and n^j_β. The conditional probability is then given by $P(D_i|D_j) = P(D_{ij})/P(D_j)$. However, we could have also calculated the conditional probability by introducing the *conditional* probabilities $w_{\alpha|\beta}$ which give the probability that α occurs at position i given that β occurred at position j. In terms of this parametrization we obtain

$$P(D_i|w, D_j) = \prod_{\alpha,\beta} (w_{\alpha|\beta})^{n^{ij}_{\alpha\beta}}. \tag{4.20}$$

Using again a Dirichlet prior with pseudo-count λ', the integral over possible conditional probabilities $w_{\alpha|\beta}$ then gives

$$P(D_i|D_j) = \prod_\beta \left[\int P(D_i|w, D_j) P(w) dw_{\alpha|\beta} \right] = \prod_\beta \left[\frac{\Gamma(21\lambda')}{\Gamma(n^j_\beta + 21\lambda')} \prod_\alpha \frac{\Gamma(n^{ij}_{\alpha\beta} + \lambda')}{\Gamma(\lambda')} \right]. \tag{4.21}$$

It is easy to see that this will only match the conditional probability we calculated through $P(D_i|D_j) = \frac{P(D_{ij})}{P(D_j)}$ if $\lambda = 21\lambda'$. In addition, in section 4.6 we also noted that equation 4.7 is independent of the choice of the root. However, this is also only true when $\lambda = 21\lambda'$.

4.7.2.1 Probabilities of unassigned kinases and receivers

The calculation of the joint probability $P(J, K, R)$, with J the alignments of assigned pairs, K the alignment of unassigned kinases, and R the alignment of unassigned receiver domains, is identical for each particular class of kinases. We thus focus on a single class. As described in the main paper we make the assumption that K depends only on the kinase sequences in J and R only on the receiver sequences in J. That is, for the probability of the kinases that are not assigned, only the amino acids in the *kinases* of the assigned pairs matter, not the amino acids of the *receivers* in the assigned pairs (and vice versa for the receivers). Formally, we thus assume that we can factorize $P(K, R, J)$ as follows

$$P(K, R, J) = P(K|J^k) P(R|J^r) P(J) \tag{4.22}$$

with J^k the sequences of the assigned kinases and J^r the sequences of the assigned receivers.

Since the calculation of $P(K|J^k)$ and $P(R|J^r)$ is identical we focus on the calculation of the kinase probabilities $P(K|J^k)$. We first calculate this conditional probability for a specific dependence tree T, i.e. we calculate $P(K|J^k, T)$. Note that, in contrast to the dependence tree for the joint alignment J, this tree includes only positions within the kinase. We now use the general identity

$$P(K|J^k, T) = \frac{P(K, J^k|T)}{P(J^k|T)} \tag{4.23}$$

and use equations 4.2, 4.4, 4.5, and 4.7 to calculate the factors in numerator and denominator. In particular, let K^{ij} denote the set of counts in the ith and jth columns of the unassigned kinases K, with $K^{ij}_{\alpha\beta}$ the number of times the combination $(\alpha\beta)$ occurs at positions (ij). Similarly let k^{ij} denote the counts in columns i and j of the kinases in J^k with $k^{ij}_{\alpha\beta}$ the number of times combination $(\alpha\beta)$ occurs in columns

(ij). We also have the marginal counts K^i_α and k^i_α in columns K^i and k^i. Using equation 4.7

$$P(D|T) = \left[\prod_i P(D^i)\right]\left[\prod_{i\neq r} R_{i\pi(i)}\right], \tag{4.24}$$

we have

$$P(K|J^k,T) = \left[\prod_i \frac{P(k^i+K^i)}{P(k^i)}\right]\prod_{i\neq r}\frac{R_{i\pi(i)}(k^{ij}+K^{ij})}{R_{i\pi(i)}(k^{ij})}, \tag{4.25}$$

where the function $P(n^i)$ of the set of marginal counts n^i is given by expression 4.2

$$P(n^i) = \frac{\Gamma(21\lambda)}{\Gamma(n+21\lambda)}\prod_\alpha \frac{\Gamma(n^i_\alpha+\lambda)}{\Gamma(\lambda)}, \tag{4.26}$$

and the function $R_{ij}(n^{ij})$ of the set of counts n^{ij} in a pair of columns (ij) is given by combining equation 4.4:

$$P(n^{ij}) = \frac{\Gamma(21^2\lambda')}{\Gamma(n+21^2\lambda')}\prod_{\alpha\beta}\frac{\Gamma(n^{ij}_{\alpha\beta}+\lambda')}{\Gamma(\lambda')}, \tag{4.27}$$

with equation 4.5: $R_{ij}(n^{ij}) = P(n^{ij})/[P(n^i)P(n^j)]$. In summary, the conditional probability $P(K|J^k,T)$ given a dependence tree T can be determined by using the exact same expressions as used for calculating $P(J|T)$ in the main paper, only now we calculate the ratio of the probabilities of the alignment containing both counts K and J^k and the alignment containing only counts J^k.

Finally, if we define the ratio of R_{ij} values:

$$\tilde{R}_{ij} = \frac{R_{ij}(k^{ij}+K^{ij})}{R_{ij}(k^{ij})}, \tag{4.28}$$

we could again calculate the sum over spanning trees by defining the Laplacian matrix

$$\tilde{M}_{ij} = \delta_{ij}\left(\sum_k \tilde{R}_{ik}\right) - \tilde{R}_{ij}, \tag{4.29}$$

from which one row and column have been removed, and using

$$P(K|J^k) = \left[\prod_i \frac{P(k^i+K^i)}{P(k^i)}\right]\frac{1}{|T|}\det(\tilde{M}). \tag{4.30}$$

However, as detailed below, calculating this determinant accurately is a challenging numerical problem which has currently not been satisfactorily solved, see e.g. [46],

and we instead use the approximation of only using the dependence tree T^* with maximal probability, i.e.

$$P(K|J^k) \approx P(K|J^k, T^*). \tag{4.31}$$

We choose the dependence tree T^* that maximizes the probability of all the cognate kinases and keep this tree fixed throughout the sampling runs. In addition, to reduce numerical error due to small spurious correlations, we score positions that show no evidence of dependence with any other position according to a WM model. In particular, all positions i for which $\log(R_{ij}) < 10$ for all positions j are excluded from the dependence tree T^* and are scored with a WM model.

4.7.2.2 Approximation of the determinant

The matrix components R_{ij} and $\tilde{R}_{ij}$ correspond to the ratios of probabilities of all observed data in columns (i, j) under a general dependent model and under the assumption that i and j are independent. These in turn involve the ratios of products of gamma functions whose arguments, i.e. the number of occurrences of certain combinations of letters in certain columns, can become quite large. As a result, some of the matrix components are extremely large numbers, and others are extremely small numbers. In principle this is no numerical problem because we can easily calculate the logarithms of the matrix entries instead of the matrix entries themselves. However, when we calculate the determinant we need to calculate combinations of products, sums, and differences of these matrix entries and this is numerically very challenging.

In order to approximate the determinant we used the same approach as in [46]. We rescaled all matrix entries as follows

$$R_{ij} \to 10^{C(\frac{\log(\frac{R_{ij}}{R_{\min}})}{\log(\frac{R_{\max}}{R_{\min}})}-1)} \tag{4.32}$$

where $R_{\max}$ ($R_{\min}$) is the maximal (minimal) entry of the matrix R_{ij}. This function essentially rescales and shifts all the $\log(R_{ij})$ values such that they now map to the interval $\left[10^{-C}, 1\right]$. These scaled R values can be considered a more conservative estimate of dependence, as they diminish the relative difference in dependence between different pairs of positions [46].

For our predictions of cognate two-component interactions as well as polyketide synthase interactions, we set $C = 5$, calculated $\log(R_{\max})$ as well as $\log(R_{\min})$ at the beginning of the simulation and kept it fixed during the simulation (the highest $\log(R_{ij})$ values correspond to pairs of residues (ij) that lie in the same protein and thus do not depend on the current assignment). In order to keep the absolute log-probability differences of different assignments approximately the same the resulting determinants need to be rescaled by an appropriate factor in order to counteract the reduction of log-probability differences due to the rescaling of the R-matrix entries.

We chose this factor by demanding that the model reduces to the maximum-likelihood tree model in the case of one dominating tree. Let $\det(M')$ be the minor of the Laplacian with scaled R values and $\det(M)$ the minor of the Laplacian with the actual R values. We then approximate $\det(M)$ as

$$\det(M) \approx \left[\frac{\det(M')}{10^{-C(n-1)(1+\frac{\log(R_{min})}{\alpha})}} \right]^{\frac{\alpha}{C\log(10)}} \tag{4.33}$$

where $\alpha = \log(\frac{R_{max}}{R_{min}})$ and n is the dimension of the matrix R_{ij}. Note that this approximation is also very accurate in the case of a set of dominating dependence trees with similar likelihoods.

In an attempt to reduce numerical error due to positions that show no dependence on other positions to start with, we do not score all columns according to the general model, but filter out a subset of positions that show either very low variability, or that show no dependence on any of the other positions. In particular all positions with entropy less than 10% of the maximum possible entropy log(21), and all positions with more than 50% gaps are filtered out. These positions are scored using a simple WM model, i.e. with the probability of the letter independent of other columns. Again, this complication is to reduce numerical errors and would not be necessary if we had a better numerical procedure for calculating the determinant.

4.7.2.3 Sampling scheme for the sum-over-trees model

Without any prior knowledge about the connectivity nor about the dependence tree structure, our search space is vast and there is a great danger of getting stuck in local optima during the sampling procedure. In order to deal with this problem we used simulated annealing starting from a relatively high 'temperature'. We sample from the distribution $P(D)^{1/T}$ setting $T = 100$ at the start and decreasing T linearly with time until $T = 1$ is reached. Due to the 'heating', the probability distribution over the space of assignments is effectively flattened and it is easier to move out of local maxima in this initial phase. After $T = 1$ is reached we continue sampling at $T = 1$ and allow the system to reach equilibrium. In a final phase of sampling (still at $T = 1$) we record interaction partners to estimate the posterior distribution of interaction for any kinase/regulator pairs. The simulated annealing resulted in a significant improvement in performance compared to simulations where $T = 1$ is used throughout (data not shown).

4.7.3 Reconstruction of cognate pairs

4.7.3.1 Results for the small classes

The results of the reconstruction of cognate pairs for the smaller kinase classes are shown in figure 4.9. The smaller kinase classes, particularly the chemotaxis and HWE

classes, have only very few kinases and regulators per genome and therefore random scoring, i.e. where every possible kinase/receiver pair inside the same genome is assigned the same probability of interaction, already produces a reasonable number of correct predictions. Additionally, the sizes of the corresponding alignments are very small and there is only little co-evolutionary information. Nonetheless, it is apparent in figure 4.9 that the method produces highly accurate predictions on these smaller classes as well.

4.7.3.2 Performance of the extended model on all cognate pairs

The prediction of orphan interactions requires two extensions to our model. Response regulators must be allowed to interact with kinases of any class and, due to unequal numbers of kinases and regulators, our way of assigning kinases and regulators demands that in every assignment a number of kinases and regulators do not have any interaction partner (see section 4.6).

A simple way of testing the performance of the former extension is to run our MCMC simulation with all cognate pairs of all 7 classes at the same time. The results are shown in figure 4.10. Due to the fact that the search space is now much bigger as every kinase can interact with any response regulator of the 7 classes, i.e. every regulatory can switch between the 7 classes of kinases, the quality of our predictions, though still quite accurate, generally decreases. It is important to note that although the chemotaxis and HWE families are very small and thus contain very little co-evolutionary information, the algorithm predicts the interaction partners of kinases of these classes with very high accuracy. This is due to the fact that regulators of the chemotaxis and HWE families form clearly distinct subfamilies (see figure 4.8) and thus, since they come in very small numbers per genome, a correct classification of their class membership is sufficient for determining their right interaction partners.

4.7.4 Network structure predictions

As described in the main text, for the prediction of the two-component signaling network structure, we assign a log-ratio score to any kinase/regulator pair of the HisKA class. In figure 4.11, we show the PPV/sensitivity curve for this log-ratio score. The used cut-off of 1 corresponds to a sensitivity of 0.56 and a PPV-value of 0.48. Note that although, at this cut-off, every second prediction corresponds to a non-cognate pair, the false positive rate is very low (0.04). Also note that for figure 4.11 we consider all predicted interactions between proteins belonging to different cognate pairs as false positives, which is very conservative since cross-talk between cognates is likely to exist. If we use the log-ratio score to predict HisKA orphan interactions, we get a p-value of 10^{-7} for the set of known *Caulobacter* interactions (and 10^{-3} when in addition the putative cognate pairs are excluded from our dataset

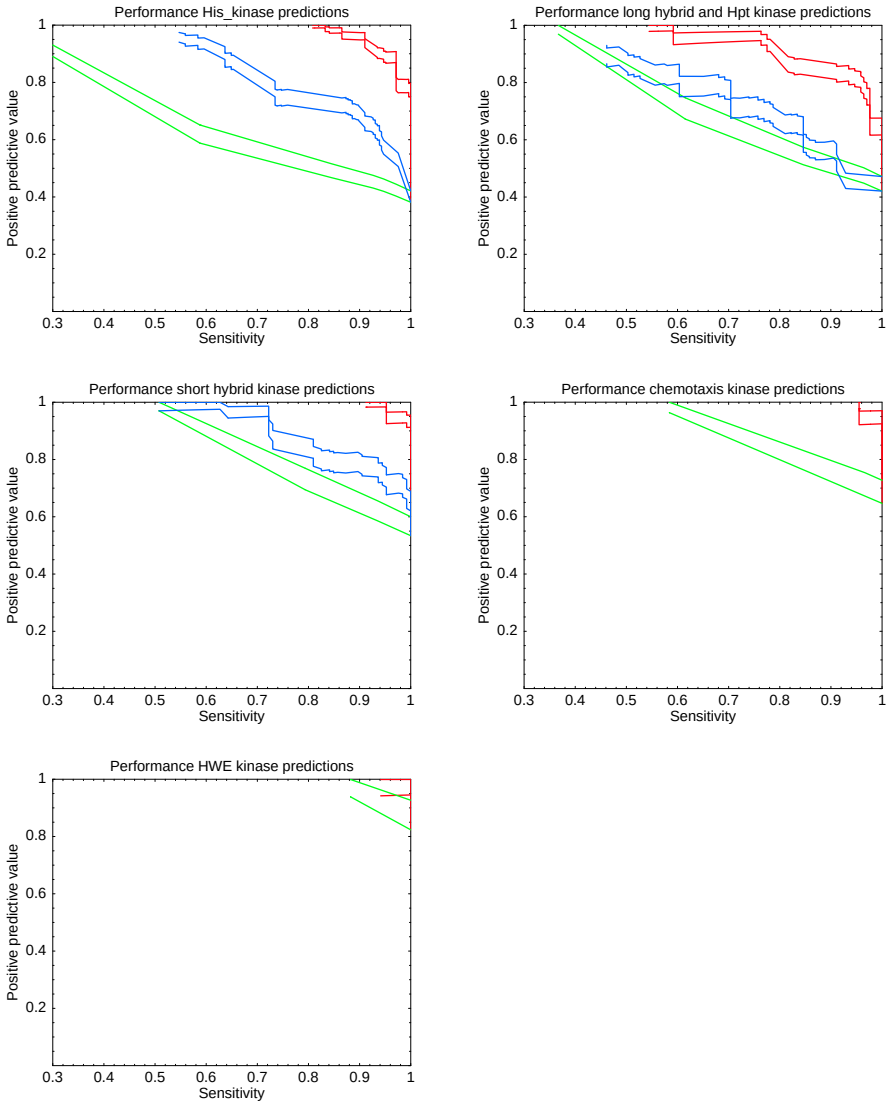


Figure 4.9: Analysis of the predictions for cognate pairs for the His_kinase (top left), long hybrid/Hpt (top right), short hybrid (middle left), chemotaxis (middle right) and HWE classes (bottom left). In all figures, the red curves show the performance of the model in which $P(D|a,T)$ is averaged over all dependence trees, the blue curve shows the performance of the model $P(D|a,T^*)$ that uses only the best dependence tree, and the green line shows the performance of random predictions. For the chemotaxis and HWE predictions, the blue curve is not shown as it is identical to the green curve due to the fact that there are no pairs of positions with a $\log(R)$ value higher than our threshold of 10 and the sequences are thus scored with a simple position-specific weight matrix model. All pairs of curves show estimated PPV plus and minus one standard error.

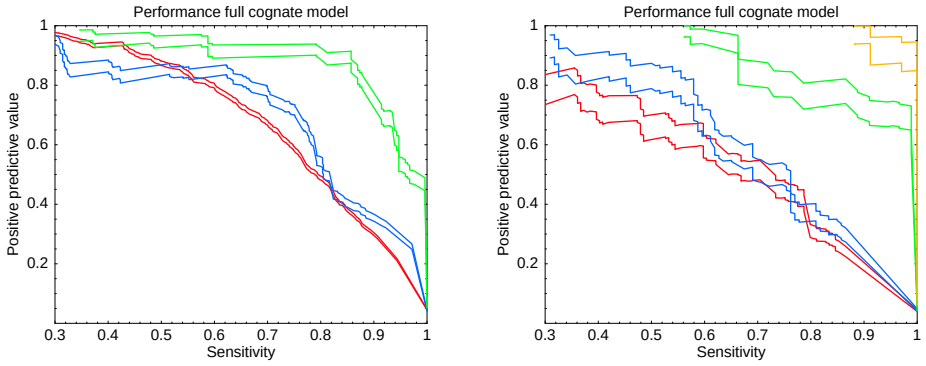


Figure 4.10: Reconstruction of cognate pairs when response regulators are allowed to interact with kinases of any of the 7 classes.**Left panel:** Quality of predictions for kinases of class HisKA (red line), H3 (blue line) and HisKin (green line). **Right panel:** Quality of predictions for kinases of class long hybrid/Hpt (red line), short hybrid (blue line), chemotaxis (green line) and HWE (orange line). All pairs of curves show estimated PPV plus and minus one standard error.

(see below)).

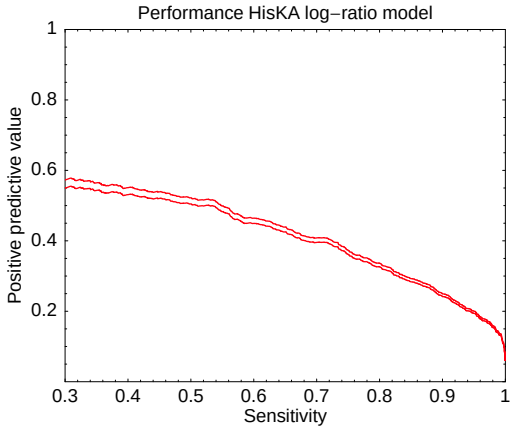


Figure 4.11: Reconstruction of cognate pairs with the log-ratio model that is used to predict network structure. The curves show estimated PPV plus and minus one standard error.

4.7.5 P-value calculation

In order to test the significance of our predictions in *Caulobacter Crescentus*, we calculated a p-value as follows. For each of the HisKA kinases with known interactions, we collected the posterior probabilities of interaction for all orphan regulators. We then sorted the entire list of all predictions by posterior and ranked each prediction, starting at rank 0 for the prediction with highest posterior. We then summed the ranks of all known interactions, obtaining the rank-sum r_{tot}, and calculated the probability

$P(r \leq r_{\text{tot}})$, of getting a rank-sum r not larger than r_{tot} with random predictions (i.e. a randomly ordered list).

When the total number of predictions, n, is larger than r_{tot}, the probability $P(r \leq r_{\text{tot}})$ can be very well approximated analytically as follows. Let X_i be the rank of the known interaction i ($X_i \in \{0, .., n-1\}$) and l the total number of known interactions.Then, for $m < n$,

$$P(\sum_{i=1}^{l} X_i = m) = \frac{1}{n^l} \binom{m+l-1}{l-1} \tag{4.34}$$

where $\frac{1}{n^l}$ is the probability for the variables X_i to take on any value between 0 and $n-1$ and $\binom{m+l-1}{l-1}$ is the total number of possible combinations of l numbers that sum up to m. In our problem, two known interactions cannot have the same rank, but this effect should be small as the number of known interactions is small compared to the number of possible interactions. From equation (4.34), we calculate the p-value,

$$P(r \leq r_{\text{tot}}) = \sum_{f=0}^{r_{\text{tot}}} \frac{1}{n^l} \binom{f+l-1}{l-1} \tag{4.35}$$

For the orphan predictions in *Caulobacter* we obtain a p-value of $7.5 \cdot 10^{-18}$. Some of the predicted pairs are found to actually lie near each other on the genome (although they were not predicted to be in the same operon, and where thus not classified as orphan pairs). If we exclude these putative cognate pairs the p-value becomes $1.1 \cdot 10^{-9}$.

4.7.6 Comparison with orphan interactions

4.7.6.1 Orphans in *Caulobacter crescentus*

The orphan kinase ChpT of *Caulobacter crescentus* only has a HisKA domain and does thus not fall into the HisKA class as defined in table 2 in the main text (ChpT does not have an ATP-binding domain). However, to increase the number of experimentally determined interactions that we could use to benchmark our predictions, we added the ChpT kinase as well as its orthologs as defined by COG [75] to our set of orphan HisKA kinases (for our predictions, we only use the HisKA domain (see above), so the absence of the ATP-binding domain does not cause any difficulties).

4.7.6.2 Additional orphan interactions

Besides *Caulobacter crescentus* that accounts for the largest part of known orphan interactions, there are three more species with experimentally determined orphan

kinase	regulator	posterior	se	exp evidence
HP0244	HP0703	0.9427	0.0485	[76]
HP0244	HP1043	0.05336	0.0487	[76]
HP0244	HP1021	0.0039	0.0022	
HP0244	HP1067	0	0	
HP0244	HP0616	0	0	
HP0244	HP0393	0	0	
HP0244	HP0019	0	0	

Table 4.3: Predictions for the one orphan HisKA kinase in *Helicobacter pylori* for which an interaction is known. There are 7 orphan regulators in *H. pylori* and we show the posterior probabilities for all of them. Posterior probabilities and their standard errors were calculated over 20 sampling runs.

interactions involving HisKA kinases, namely *Helicobacter pylori*, *Bacillus subtilis* and *Ehrlichia chaffeensis*. Our predictions for these species are shown in tables 4.3, 4.4 and 4.5. As in table 4.1, the list of predictions is shown ordered by posterior, up to and including all known interactions. Correct predictions are shown in green, incorrect predictions (at odds with the experimental results) are shown in red. All other predictions are shown in black. Posterior probability and standard error of the posterior probability over 20 sampling runs are shown for each prediction.

In *H. pylori* the known interaction matches the top prediction of the algorithm which is assigned a 94% posterior probability.

In *B. subtilis* it is known that the regulator Spo0F interacts with all Kin kinases, i.e. KinA, KinB, KinC, KinD, and KinE. Indeed we predict that Spo0F interacts with all these kinases with nonzero probability. The interaction probabilities of Spo0F with all other kinases is zero (data not shown). Table 4.4 shows, however, that the fraction of time Spo0F is associated with each of these kinases varies significantly across the different Kin kinases, with Spo0F associating with KinC more than 65% of the time, with KinD 28% of the time and only roughly 4% with the other Kin kinases. Note also that some of the Kin kinases are predicted to interact with other regulators as well.

For kinase ECH_0299 of *E. chaffeensis* (an ortholog of NtrY), we correctly predict that it interacts only with ECH_0339 (an ortholog of NtrX). Kinase ECH_0885 is the only example where our predictions clearly disagree with the experimental evidence. Whereas the experimental evidence suggests that ECH_0885 interacts only with ECH_0773, we assign 100% posterior probability to ECH_0885 interacting with ECH_1012.

kinase	regulator	posterior	se	exp evidence
KinA	Spo0F	0.0361	0.0060	[77]
KinB	CheV	0.7929	0.0348	
KinB	Spo0A	0.1649	0.0256	
KinB	YneI	0.0412	0.0294	
KinB	Spo0F	0.0006	0.0004	[77]
KinC	Spo0F	0.6765	0.0731	[77]
KinD	YneI	0.5215	0.0975	
KinD	Spo0F	0.2840	0.0692	[77]
KinE	Spo0A	0.4516	0.0768	
KinE	YneI	0.3751	0.0972	
KinE	CheV	0.1649	0.0366	
KinE	Spo0F	0.0028	0.0008	[77]

Table 4.4: Predictions for orphan HisKA kinases with known interactions in *B. subtilis*. There are 6 orphan regulators in total in *B. subtilis*. For every known interaction shown there are several kinds of evidence, see [77]. Posterior probabilities and their standard errors were calculated over 20 sampling runs.

kinase	regulator	posterior	se	exp evidence
ECH_0299(NtrY)	ECH_0339(NtrX)	1	0	[78]
ECH_0885(PleC)	ECH_1012(CtrA)	1	0	[78]
ECH_0885(PleC)	ECH_0773(PleD)	0	0.2236	[78]

Table 4.5: Predictions for the two orphan HisKA kinases with known interactions in *E. chaffeensis*. There are 3 orphan regulators in total in *E. chaffeensis*. Posteriors and their standard errors were calculated over 20 sampling runs.

	CK	OK	-
CR	9.184	0.009	3.59
OR	0.015	0.055	1.02
-	1.326	0.346	382.5

Table 4.6: Ortholog statistics for cognate pairs. For each cognate kinase/receiver pair and each of the 398 other genomes, there can be either: no orthologs for both (-,-), two orthologs that form a cognate pair (CK,CR), no ortholog for the kinase and an ortholog for the receiver which is an orphan receiver (-,OR), etcetera. The table shows the average number of times each of the 9 possible combinations occurs for cognate kinase/receiver pairs.

4.7.7 Ortholog statistics

Our predictions suggest that orphan kinases interact predominantly with orphan regulators, that cognate kinases interact predominantly with cognate regulators, and that there is relatively little interaction between orphan kinases and cognate regulators or between cognate kinases and orphan regulators. Since orphans and cognates almost certainly share a common phylogenetic ancestry, we decided to investigate to what extent cognates and orphans change class on relatively short evolutionary time scales. To this end we determined orthologous genes for each cognate kinase/regulator pair, for each orphan kinase, and for each orphan regulator.

Table 4.6 shows the ortholog statistics for cognate pairs. For each cognate kinase/regulator pair there are 9 possibilities for its orthologs in each of the 398 other genomes varying from the cognate pair mapping to another cognate pair in the other genome, to absence of orthologs for both genes in the pair. The table shows the average number of occurrences of each of the 9 possibilities. The table shows that in on average over 380 genomes there are no orthologs for either gene. The next most common occurrence is that the cognate pair maps to a cognate pair (in on average 9.184 genomes). After that it is by far most likely that only one of the two genes has an ortholog. In all cases cognates are significantly more likely to map to cognates than to orphans.

Similarly, for each orphan kinase we counted the number of times that it has no ortholog in each of the 398 other genomes, the number of times the ortholog is itself an orphan, and the number of times the ortholog is part of a cognate pair. Finally, for each orphan receiver we counted the number of times it has no ortholog in each of the other genomes, the number of times its ortholog is an orphan, and the number of times its ortholog is part of a cognate pair. These orphan ortholog statistics are shown in table 4.7.

The table shows that for both orphan kinases and for orphan receivers there are on average a handful of genomes with orthologs. In both cases, if there is an ortholog, it is much more likely to be an orphan as well.

	Orphan	Cognate	-
Orphan kinase	3.78	0.61	393.6
Orphan receiver	4.595	1.153	392.25

Table 4.7: Ortholog statistics for orphans. For both orphan kinases and orphan receivers, the table shows how many of 398 other genomes on average have: an ortholog that is also an orphan, an ortholog that is part of a cognate pair, or no ortholog.

4.7.8 Prediction of polyketide synthase interactions: classification model

In order to compare the quality of our predictions to the simple classification scheme proposed in [15], we calculated posterior probabilities of interaction using only the information about the class membership of the head and tail sequences as follows. We used the annotation of [15] to label every head (tail) as H1 (T1), H2 (T2), H3 (T3) or as 'unclustered'. For a given head sequence of class H_i of a genome g, we assign an interaction probability of 0 to all tails of classes T_j with $j \neq i$ and a probability of $1/n_g^i$, where n_g^i is the number of tails of class i of genome g, to all tails of class i of genome g. If the head belongs to the class of unclustered heads, it is assigned a probability of $1/n_g$ to interact with any of the n_g tails of genome g. In other words, for the H1, H2 and H3 classes, each head sequence can only interact with tail sequences of the correct corresponding tail class, but within the corresponding class, every tail is equally likely to be an interaction partner. Heads that are unclustered can interact with any tail of the same genome with equal probability.

Chapter 5

Discussion and outlook

It is becoming clear that pairs of co-evolving residues in proteins are not isolated from each other, but form chains or networks which connect residues that can be distant in space ([21, 22, 35] and chapter 2). We have shown that dependence tree models are well suited to describe such networks in a statistically sound and computationally tractable way. A crucial ingredient of our model is that it inherently distinguishes between dependencies that are direct and others that can be explained indirectly. In the context of domain structures, we could show that this distinction leads to a significant improvement in the prediction of interacting residues (chapter 2). The successful application of our Bayesian network model to the prediction of protein-protein interactions strongly suggest that our model should also be well-suited for the inference of interacting residues between interacting proteins or protein domains. Indeed, albeit based on a different mathematical model, recent work on the family of bacterial two-component systems has shown that the distinction between direct and indirect dependencies in joint alignments of interacting kinase/regulator pairs leads to a very strong improvement in prediction accuracy of inter-protein contacts [16]. An obvious and interesting direction of future research would be to use our Bayesian model for the inference of inter-protein/domain contacts on a large scale and to investigate whether our results on protein domains can be generalized to interacting protein families.

Another possible road of future research would be to extend and generalize our method for the prediction of protein-protein interactions. At this point, it is instructive to consider the general problem of predicting protein-protein interactions in two different limits - in a scenario where there are families of paralogous proteins and in a second scenario, where there are several different protein families, which each only consist of orthologous proteins (figure 5.2). So far our work has focused on a special case of the former problem, namely on a scenario where there are two large families of paralogous proteins that are known to interact specifically. For this scenario, as illustrated in figure 5.1, we considered assignments of putative interaction

partners that assign to each member of the first family exactly one member of the second family. A possible extension of our model would be to consider more general assignments, where proteins are either allowed to have no interaction partners at all and/or where proteins are allowed to have several interaction partners, to deal with cases where the interactions are less specific. The latter extension may be helpful to for example infer interactions between growth factors and their receptors, which can be quite unspecific [79, 80], or to improve our predictions for orphan two-component proteins, which we have shown to be more promiscuous (cf. chapter 4). Assignments that allow for several interaction partners require an extension of our mathematical framework because the probability of the data can no longer be written simply in terms of the joint probability of interacting pairs of sequences. It is not clear whether such an extension would allow for an exact analytical expression for the probability of the data and/or an efficient sampling scheme.

Assignments that allow proteins not to have any interaction partners are feasible within our current mathematical framework. Such assignments would be particularly useful in cases where some interaction partners are missing, for example because of difficulties in detecting all members of the two protein families due to low sequence similarity (as in the case of the kinases of the Hpt class, cf. chapter 4) or because certain proteins in the dataset have evolved a different function and do not interact any more. This can be very problematic in situations where the number of proteins per genome is small and an assignment that forces each protein to have an interaction partner leads to an incorrect pairing of proteins.

Preliminary results with assignments that allow each protein to have either one or no interaction partner have shown that even in the case of non-interacting protein families, assignments that include interacting proteins are favoured because of the background signal that is due to the phylogenetic relatedness of the sequences (cf. top left panel in figure 4.2 that shows that even non-interacting pairs of proteins have many highly dependent pairs of residues). Presumably, there would be a similar effect if several interaction partners per protein were allowed. Thus, an extension of our model to more general assignments would most likely require the introduction of a phylogenetic correction, be it in the form of a simple correction as used for the prediction of contacts (chapter 2) or in the form of an evolutionary model for pairs of amino acid residues.

The problem of inferring interaction partners in the second limit, namely in the case of families that consist only of orthologous proteins (figure 5.2), could in principle also be tackled within our Bayesian framework. Under the assumption that if two proteins interact, all their orthologs from other genomes also interact (as is generally assumed [7, 62, 81]), joint alignments of putatively interacting pairs of proteins could be constructed by fusing the protein sequences of the interacting pair and all their orthologs. Whereas in the case of two families of paralogous proteins, only one joint alignment needs to be constructed (figure 5.1), here there would be as many joint

alignments as putatively interacting protein pairs (figure 5.2). A suitable score of interaction for a given pair could be the probability of their joint alignment divided by the probability of the two single alignments, where each probability would be calculated by summing over all possible spanning trees. Such a score would quantify the amount of dependence that is seen between inter-protein residues. A difficulty with this approach is that since there is only one protein of each family per genome and, to calculate the interaction score of two proteins, only sequences from genomes where orthologs of both proteins are present can be used, the joint alignments tend to be rather small. However, with the growing number of fully sequence genomes, it should soon become feasible to construct alignments of reasonable size for many pairs of protein families.

Ultimately, the Bayesian framework for both scenarios may be merged into one general model. In such a model, assignments would include both paralogous and orthologous protein sequences so as to make optimal use of all the information available for the estimation of dependencies between inter-protein residues.

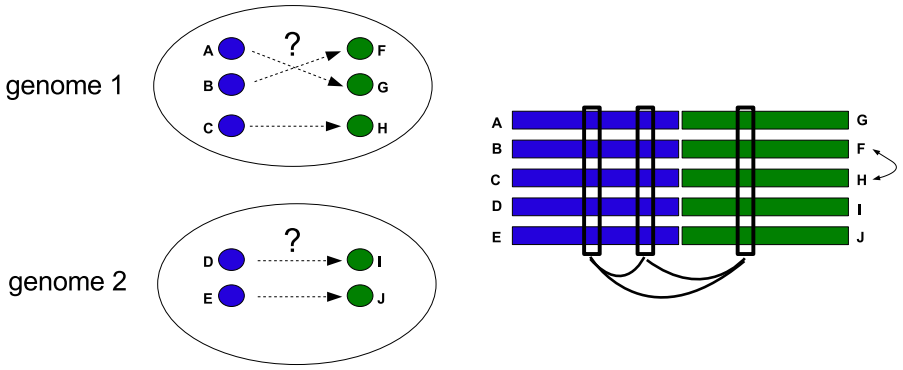


Figure 5.1: Predicting protein-protein interactions in the case of two families of paralogous proteins (blue and green) that interact specifically. Here, a possible solution to the problem is given by an assignment of each protein of the first family to exactly one protein of the second family (left). This induces one joint alignment of interacting sequence pairs (right). The true interactions are inferred by permuting the sequences of the same genome of one family in the alignment (here the green one, indicated by the arrow) and searching for the assignment that maximizes the dependencies between columns in the joint alignment.

Finally, dependence tree models could be generally used for the refinement and characterization of multiple sequence alignments of protein families or protein domains. Although a generalization of profile hidden Markov models to models that take into account the dependence structure of the underlying sequences may not be feasible, our Bayesian model could be used to refine existing multiple alignments, for example by rearranging gaps in such a way as to maximize the probability of the alignment, i.e. the total amount of dependencies seen between pairs of columns. In this context, it would be interesting to investigate how in such a refined alignment,

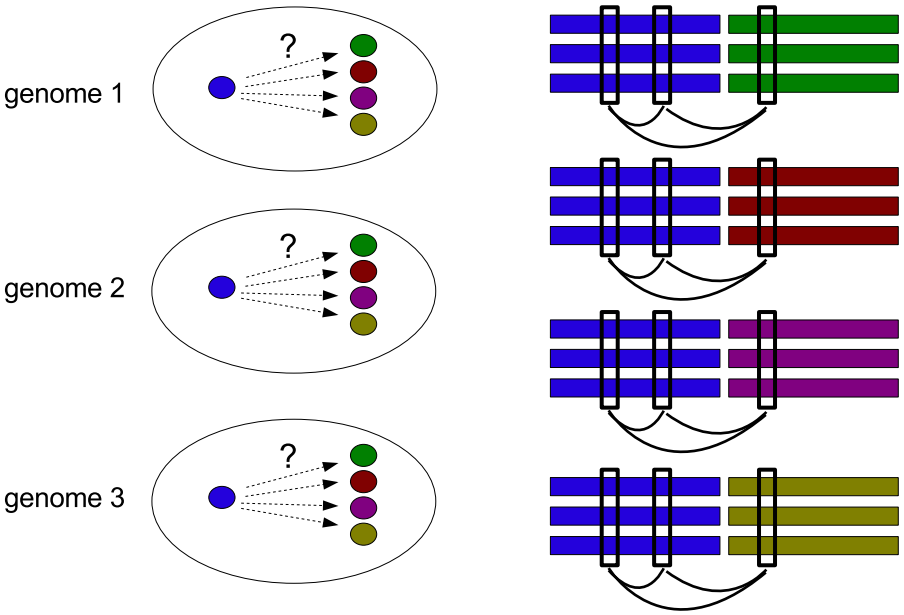


Figure 5.2: Predicting protein-protein interactions in the case of several protein families that consist only of orthologs, illustrated as circles with different colours. It is assumed that all orthologs of an interacting protein pair also interact. For each putatively interacting pair, all orthologous pairs are collected, fused and used to construct a joint alignment (right). Interactions are inferred for those pairs that show the strongest dependencies in their joint alignment, which can be quantified, using dependence tree models, by the ratio of the probability of the joint alignment versus the probability of the two single alignments.

or generally in any alignment, local correlations relate to structural properties of the proteins. Previous work [39] and a preliminary analysis on our dataset of domain alignments (chapter 2) has shown that within alpha helices, there is a clear enrichment of dependencies between pairs of residues that are 3 to 4 residues apart, corresponding to the periodicity of the helix. It may thus also be possible to improve current methods for the prediction of protein secondary structure (see e.g. [82]) with the help of co-evolutionary information. Lastly, our model may be used for the discovery of protein subfamilies. This could be done by clustering the sequences of a given protein alignment based on differences in the dependence structure or based on differences in the distribution of pairs of amino acids in particular correlated pairs of columns.

Part II

Inference of Protein-RNA Interactions from High-Throughput Sequencing Data

Chapter 6

Introduction

Eukaryotic cells contain tens of thousands of mRNAs, whose location, activity and fate must be highly coordinated and regulated [83]. In most of these regulatory processes, such as nuclear export, splicing, localization and stability, RNA-binding proteins (RBPs) are of crucial importance [84]. However, only a small number of RBPs have been characterized experimentally so far, and for most of the several hundred RBPs that are encoded in eukaryotic genomes, little is known about the regulatory 'logic' by which they determine the fate of RNAs. Insights gained from a number of experimental studies in recent years have led to the proposition of the so-called RNA regulon model [84,85] (figure 6.1). This model hypothesizes that, in analogy to transcription factors that co-regulate a set of related genes, each RBP targets a distinct set of functionally related mRNAs, which as such form a RNA regulon. Additionally, the model predicts that since many eukaryotic proteins have several different functions, their corresponding transcripts can be part of several regulons. Thus, the fate of each mRNA is not only determined by one RBP, but is a result of the interplay of many RBPs (figure 6.1).

A nice example of a set of RBPs whose regulatory functions appear to fit well into the RNA regulon paradigm are the Puf1-5 RBPs in *S.cerevisiae* [86]. Whereas Puf1 and Puf2 targets strongly overlap, Puf3, Puf4 and Puf5 each target distinct, barely overlapping sets of mRNAs. Puf3 associates almost exclusively with a set of roughly 150 mRNAs that encode mitochondrial proteins, and it is thought that Puf3 is involved in the transport of these mRNAs to the mitochondrium [86]. Puf1 and Puf2 preferentially target mRNAs that encode membrane-associated proteins, Puf4 mostly binds mRNAs that encode nucleolar ribosomal RNA-processing factors and Puf5 preferentially associates with mRNAs that encode chromatin modifiers and components of the spindle body [86].

Although RBPs are involved in many diverse processes, they are built from only a few different RNA-binding modules [87]. The two most common of these modules, which both bind single-stranded RNA, are the RRM (RNA-recognition motif) and

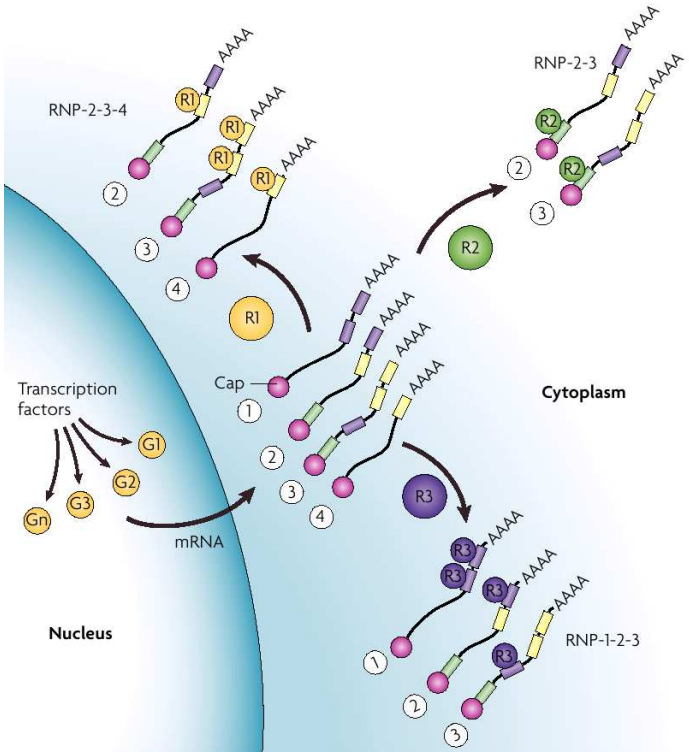


Figure 6.1: Illustration of the RNA regulon model (figure taken from [84]). RBPs are shown as coloured circles (R1, R2 and R3) and mRNAs are shown as black lines covered with coloured rectangles that denote the recognition elements of the corresponding RBPs. The groupings of mRNAs according to the RBP that they are bound to form the RNA regulons (RNP-1-2-3, RNP-2-3-4 and RNP-2-3 in the figure). There are three regulatory aspects that characterize the RNA regulon model. Firstly, a RNA regulon is a set of mRNAs that are bound by a particular RBP and that all have similar biological functions. Secondly, the transcripts of a single gene can participate in multiple regulons (e.g. transcripts 2 and 3 are part of all three regulons). As such, multifunctional proteins can participate in several regulons according to their separate functions. Thirdly, the fate of a single mRNA can be determined by the interplay of several RBPs (as illustrated by the presence of several different recognition elements on transcripts 2 and 3).

the KH (K-homology) domain. The RRM domain is by far the most abundant RNA-binding domain (RBD) in higher vertebrates and RRM domains can be found in 0.5 to 1 percent of all human genes [87–89]. In structural and biochemical studies, RRM domains have been shown to recognize anywhere between 4 and 8 nucleotides [89]. KH domains, which are ubiquitous in eukaryotes, eubacteria and archea, bind both single-stranded RNA and DNA and typically recognize a sequence stretch of 4 nucleotides [88].

A common theme in the architecture of RBPs is that they consist of several RBDs and it is thought that the specificity of RBP-RNA interaction is determined by the interplay of several of these domains, particularly in cases where the recognition motifs of single RBDs are very short [87, 90]. An extreme example in this respect are the proteins of the Pumilio family, whose RBDs each recognize only one particular nucleotide, but where the combination of several domains leads to a recognition sequence of up to eight nucleotides [86, 91, 92] .

The amino acids that connect the different RBDs are in many cases of great importance as they strongly influence the binding affinity and set the spacing of the RBD recognition sequences on the target RNA [87]. In cases where the linker is flexible, the spacing between the binding sites of consecutive RBDs is not fixed but can vary within a certain interval. A good example in this respect is the RBP Nova. Nova harbours three KH domains that each recognize a very similar 4-nucleotide motif [93] and target sites of Nova typically contain clusters of these sites spaced at distances of 2 to 6 nucleotides [94]. It may, indeed, be a general property of recognition elements of KH-domain containing RBPs that they consist of short motifs separated by spacers of variable length (see chapters 7 and 8).

In recent years, two important experimental techniques have been developed to map target RNAs of particular RBPs on a global scale, namely RIP-chip (ribonucleoprotein immunoprecipitation and microarray) [95] and CLIP (cross-linking and immunoprecipitation) [96]. In RIP-chip, endogenous RNA-protein complexes (RNPs) are isolated by immunoprecipitation or affinity-purification and the bound RNA is identified using DNA microarrays [95]. In the CLIP protocol, cells are first irradiated with UV light at 254 nm, which causes proteins to crosslink with their target RNAs. After nuclease treatment, which leads to the digestion of RNA that is unprotected by RBPs, the RBP of interest is immunoprecipitated and the target RNAs are isolated, reverse-transcribed and sequenced [96–98]. CLIP has the advantage that due to the digestion of unbound RNA and the use of sequencing technology the binding sites can be localized at a much higher resolution than in RIP-chip (roughly $50 - 100$ nucleotides, with the exact resolution depending on the sequencing technology, see e.g. [98]).

Nonetheless, there are several difficulties associated with the CLIP methodology. Firstly, the crosslinking efficiency with UV light at 254 nm is generally quite limited, resulting in low yields of target RNA. Secondly, although CLIP maps target sites at

fairly high resolution, it may in some cases still be very difficult to infer the exact binding site of the RBP, particularly in cases where the recognition sequence is very short and/or mostly structurally determined. Thirdly, among the typically thousands of genomic regions that CLIPped sequence tags map to, it is difficult to distinguish truly bound regions from background.

A novel protocol that can partially circumvent these difficulties is the so-called PAR-CLIP protocol (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation). In this experimental technique, cells are fed with modified uracils (4-thioU) prior to crosslinking and immunoprecipitation. The modified uracils are incorporated into nascent transcripts and crosslink with proteins at a 100 to 1000-fold higher efficiency. Importantly, crosslinking with modified nucleotides causes a very specific mutational bias that can be used as a criterion to distinguish truly bound sites from background as well as to infer the particular location of the binding sites. In the following chapter, we describe both the experimental and computational aspects of the PAR-CLIP method.

Chapter 7

Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP

Markus Hafner[1,4], Markus Landthaler[1,4,6], Lukas Burger[2], Mohsen Khorshid[2], Jean Hausser[2], Philipp Berninger[2], Andrea Rothballer[1], Manuel Ascano, Jr.[1], Anna-Carina Jungkamp[1,6], Mathias Munschauer[1], Alexander Ulrich[1], Greg S. Wardle[1], Scott Dewell[3], Mihaela Zavolan[2,5], and Thomas Tuschl[1,5]

1 Howard Hughes Medical Institute, Laboratory of RNA Molecular Biology, The Rockefeller University, New York, USA
2 Biozentrum der Universität Basel and Swiss Institute of Bioinformatics (SIB), Basel, Switzerland
3 Genomics Resource Center, The Rockefeller University, New York, USA
4 these authors contributed equally to this work
5 corresponding authors: Mihaela.Zavolan@unibas.ch, ttuschl@rockefeller.edu
6 present address: Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, Berlin, Germany

under review in Cell

RNA transcripts are subjected to post-transcriptional gene regulation by interacting with hundreds of RNA-binding proteins (RBPs) and microRNA-containing ribonucleoprotein complexes (miRNPs) expressed in a cell-type

dependent fashion. We developed a powerful cell-based crosslinking approach to determine at high resolution and transcriptome-wide the binding sites of cellular RBPs and miRNPs. The crosslinked sites are revealed by thymidine to cytidine transitions in the cDNAs prepared from immunopurified RNPs of 4-thiouridine-treated cells. We determined the binding sites and regulatory consequences for several intensely studied RBPs and miRNPs, including PUM2, QKI, IGF2BP1-3, AGO/EIF2C1-4 and TNRC6A-C. Our study revealed that these factors bind thousands to tens of thousands of sites containing defined sequence motifs and have distinct preferences for exonic versus intronic or coding versus untranslated regulatory transcript regions. The precise mapping of binding sites across the transcriptome will be critical to the interpretation of the rapidly emerging data on genetic variation between individuals and how these variations contribute to complex genetic diseases.

7.1 Introduction

Gene expression in eukaryotes is extensively controlled at the post-transcriptional level by RNA-binding proteins (RBPs) and ribonucleoprotein complexes (RNPs) modulating the maturation, stability, transport, editing and translation of RNA transcripts [99–101]. Vertebrate genomes encode several hundred RBPs [102], each containing one or more domains able to specifically recognize target transcripts. Furthermore, hundreds of microRNAs (miRNAs) bound by Argonaute (AGO/EIF2C) proteins mediate destabilization and/or inhibition of translation of partially complementary target mRNAs [103]. To understand how the interplay of these RNA-binding factors affects the regulation of individual transcripts, high resolution maps of *in vivo* protein-RNA interactions are necessary [84].

A combination of genetic, biochemical and computational approaches are typically applied to identify RNA-RBP or RNA-RNP interactions. Microarray profiling of RNAs associated with immunopurified RBPs (RIP-Chip) [104] defines targets at a transcriptome level, but its application is limited to the characterization of kinetically stable interactions and does not directly identify the RBP recognition element (RRE) within the long target RNA. Nevertheless, RREs with higher information content can be derived computationally from RIP-Chip data, e.g. for HuR [105] or for Pumilio [106].

More direct RBP target site information is obtained by combining *in vivo* UV crosslinking [107] with immunoprecipitation [108, 109] followed by the isolation of crosslinked RNA segments and cDNA sequencing (CLIP) [97]. CLIP was used to identify targets of the splicing regulators NOVA1 [98], FOX2 [110] and SFRS1 [111] as well as U3 snoRNA and pre-rRNA [112], pri-miRNA targets for HNRNPA1 [113]

and EIF2C2/AGO2 protein binding sites [114]. CLIP is limited by the low efficiency of UV 254 nm RNA-protein crosslinking, and the location of the crosslink is not readily identifiable within the sequenced crosslinked fragments, raising the question of how to separate UV-crosslinked target RNA segments from background non-crosslinked RNA fragments also present in the sample.

Here we describe an improved method for isolation of segments of RNAs bound by RBPs or RNPs, referred to as PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation). To facilitate crosslinking, we incorporated 4-thiouridine (4SU) into transcripts of cultured cells and identified precisely the RBP binding sites by scoring for thymidine (T) to cytidine (C) transitions in the sequenced cDNA. We uncovered tens of thousands of binding sites for several important RBPs and RNPs and assessed the regulatory impact of binding on their targets. These findings underscore the complexity of post-transcriptional regulation of cellular systems.

7.2 Results

7.2.1 Photoactivatable nucleosides facilitate RNA-RBP crosslinking in cultured cells

Random or site-specific incorporation of photoactivatable nucleoside analogs into RNA *in vitro* has been used to probe RBP- and RNP-RNA interactions [115, 116]. Several of these photoactivatable nucleosides are readily taken up by cells without apparent toxicity and have been used for *in vivo* crosslinking [117]. We applied a subset of these nucleoside analogs (7.1A) to cultured cells expressing the FLAG/HA-tagged RBP IGF2BP1 followed by UV 365 nm irradiation. The crosslinked RNA-protein complexes were isolated by immunoprecipitation, and the covalently bound RNA was partially digested with RNase T1 and radiolabeled. Separation of the radiolabeled RNPs by SDS-PAGE indicated that 4SU-containing RNA crosslinked most efficiently to IGF2BP1. Compared to conventional UV 254 nm crosslinking, the photoactivatable nucleosides improved RNA recovery 100- to 1000-fold, using the same amount of radiation energy (7.1B). We refer to our method as PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation).

We evaluated the cytotoxic effects upon exposure of HEK293 cells to 100 μM and 1 mM of 4SU or 6SG in tissue culture medium over a period of 12 h by mRNA microarrays. The mRNA profiles of 4SU or 6SG treated cells were very similar to those of untreated cells (Supplementary Tables 7.1 and 7.2), suggesting that the conditions for endogenous labeling of transcripts were not toxic.

To guide the development of bioinformatic methods for the identification of binding sites, we first studied human Pumilio 2 (PUM2), a member of the Puf-protein family (7.2A) known for its highly sequence-specific RNA binding [92].

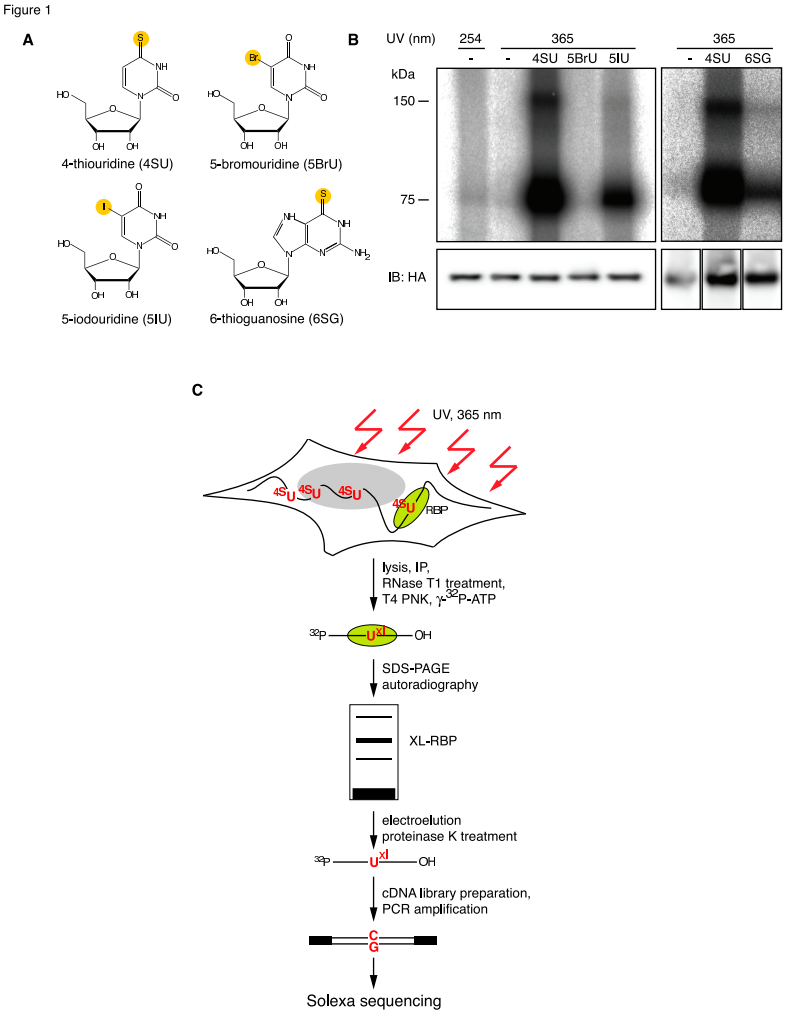


Figure 7.1: **PAR-CLIP methodology** (A) Structure of photoactivatable nucleosides. (B) Phosphorimages of SDS-gels that resolved 5'-^{32}P-labeled RNA-FLAG/HA-IGF2BP1 immunoprecipitates (IPs) prepared from lysates from cells that were cultured in media in the absence or presence of 100 µM photoactivatable nucleoside and crosslinked with UV 365 nm. For comparison, a sample prepared from cells crosslinked with UV 254 nm was included. Lower panels show immunoblots probed with an anti-HA antibody. (C) Illustration of PAR-CLIP. 4SU-labeled transcripts were crosslinked to RBPs and partially RNase-digested RNA-protein complexes were immunopurified and size-fractionated. RNA molecules were recovered and converted into a cDNA library and deep sequenced.

7.2.2 Identification of PUM2 mRNA targets and its RRE

PUM2 protein crosslinked well to 4SU-labeled cellular transcripts (7.2B). The crosslinked segments were converted into a cDNA library (7.1C) and Solexa sequenced [118]. The sequence reads were aligned against the human genome and EST databases. Reads mapping uniquely to the genome with up to one mismatch, insertion or deletion were used to build clusters of sequence reads (7.2C, Supplementary Methods). We obtained 7,523 clusters originating from about 3,000 unique transcripts, 93% of which were found within the 3' untranslated region (UTR) (Supplementary Figure 7.8) in agreement with previous studies [119]. All sequence clusters with mapping and annotation information are available online [1].

PhyloGibbs analysis [120] of the top 100 most abundantly sequenced clusters, as expected, yielded the PUM2 RRE, UGUANAUA [121] (Figure 7.2D). Unexpectedly, over 70% of all sequence reads that gave rise to clusters showed a T to C mutation compared to the genome (Supplementary Figure 7.8). Ranking of sequence read clusters according to the frequency of T to C mutation further enriched for the PUM2 RRE (Supplementary Figure 7.9) indicating that the T to C mutation is diagnostic of sequences interacting with the RBP. The T to C changes were not randomly distributed: the T corresponding to U7 of the RRE mutated at higher frequency compared to the Ts corresponding to U1 and U3 (Figure 7.2E). Our analyses indicate that the reverse transcriptase specifically misincorporated dG across from crosslinked 4SU residues and that local amino acid environment also affected crosslinking efficiency. Uridines proximal to the RRE also exhibited an increased T to C mutation frequency, indicating that crosslinks also form in close proximity to an RRE and that our method even captured PUM2 binding sites that did not have a U7 in its RRE.

7.2.3 Identification of QKI RNA targets and its RRE

To further validate our method, we applied it to the RBP Quaking (QKI), which contains a single heterogeneous nuclear ribonucleoprotein K homology (KH) domain (Figures 7.3A, B). The RRE ACUAAY was determined by SELEX [122], but *in vivo* targets are largely undefined. Mice with reduced expression of QKI show dysmyelination and develop rapid tremors or "quaking" 10 days after birth. Previous studies suggested that QKI participates in pre-mRNA splicing, mRNA export, mRNA stability and protein translation [123].

PhyloGibbs analysis of the 100 most abundantly sequenced clusters yielded the RRE AYUAAY (Figures 7.3C, D), similar to a motif identified by SELEX [122]. We found approx. 6,000 clusters mapping to 2,500 transcripts. Close to 75% of these clusters were derived from intronic sequences, supporting the hypothesis that QKI is

[1] *http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html*

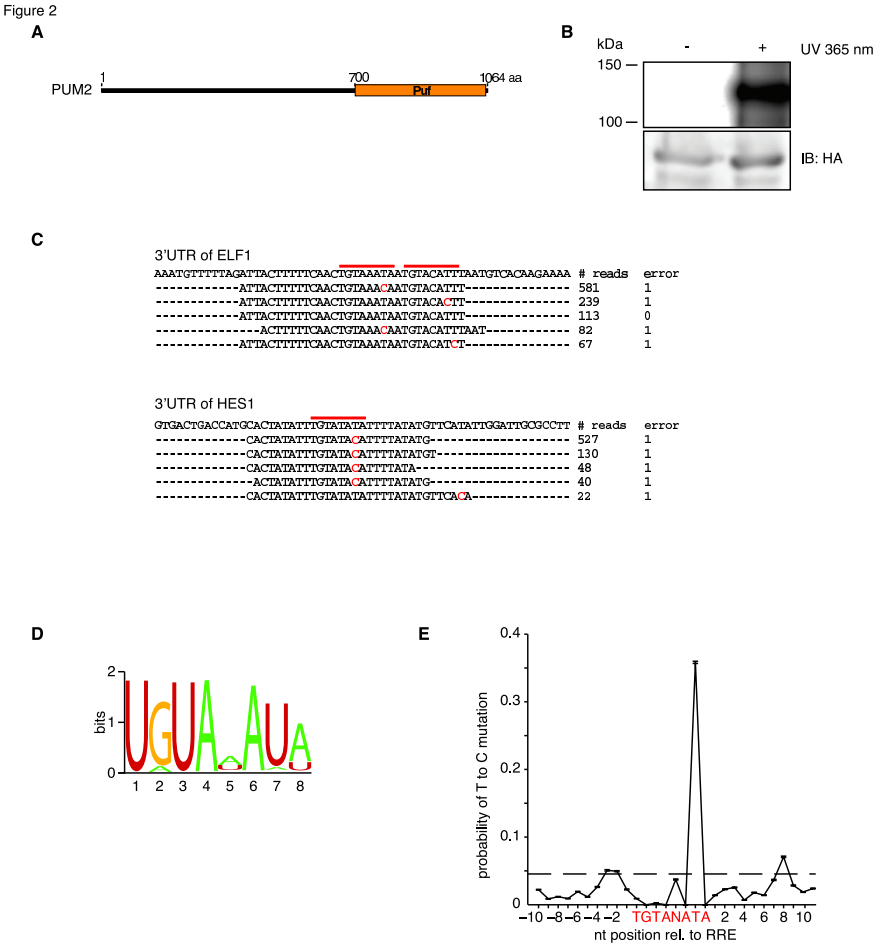


Figure 7.2: **RNA recognition by PUM2 protein**(A) Domain structure of PUM2 protein. (B) Phosphorimage of SDS-gel of radiolabeled FLAG/HA-PUM2-RNA complexes from non-irradiated or UV-irradiated 4SU-labeled cells. The lower panel shows an anti-HA immunoblot. (C) Alignments of PAR-CLIP cDNA sequence reads to corresponding regions in the 3'UTR of ELF1 and HES1 Refseq transcripts. The number of sequence reads (# reads) and mismatches (errors) are indicated. Red bars indicate the PUM2 recognition motif and red-letter nucleotides indicate T to C sequence changes. (D) Sequence logo of the PUM2 recognition motif generated by PhyloGibbs analysis of the top 100 sequence read clusters. (E) T to C positional mutation frequency for PAR-CLIP clusters anchored at the 8-nt recognition motif from all motif-containing clusters. The dashed line represents the average T to C mutation frequency within these clusters.

a splicing regulator [123] and 70% of the remaining exonic clusters fall into 3'UTRs (Supplementary Figure 7.8).

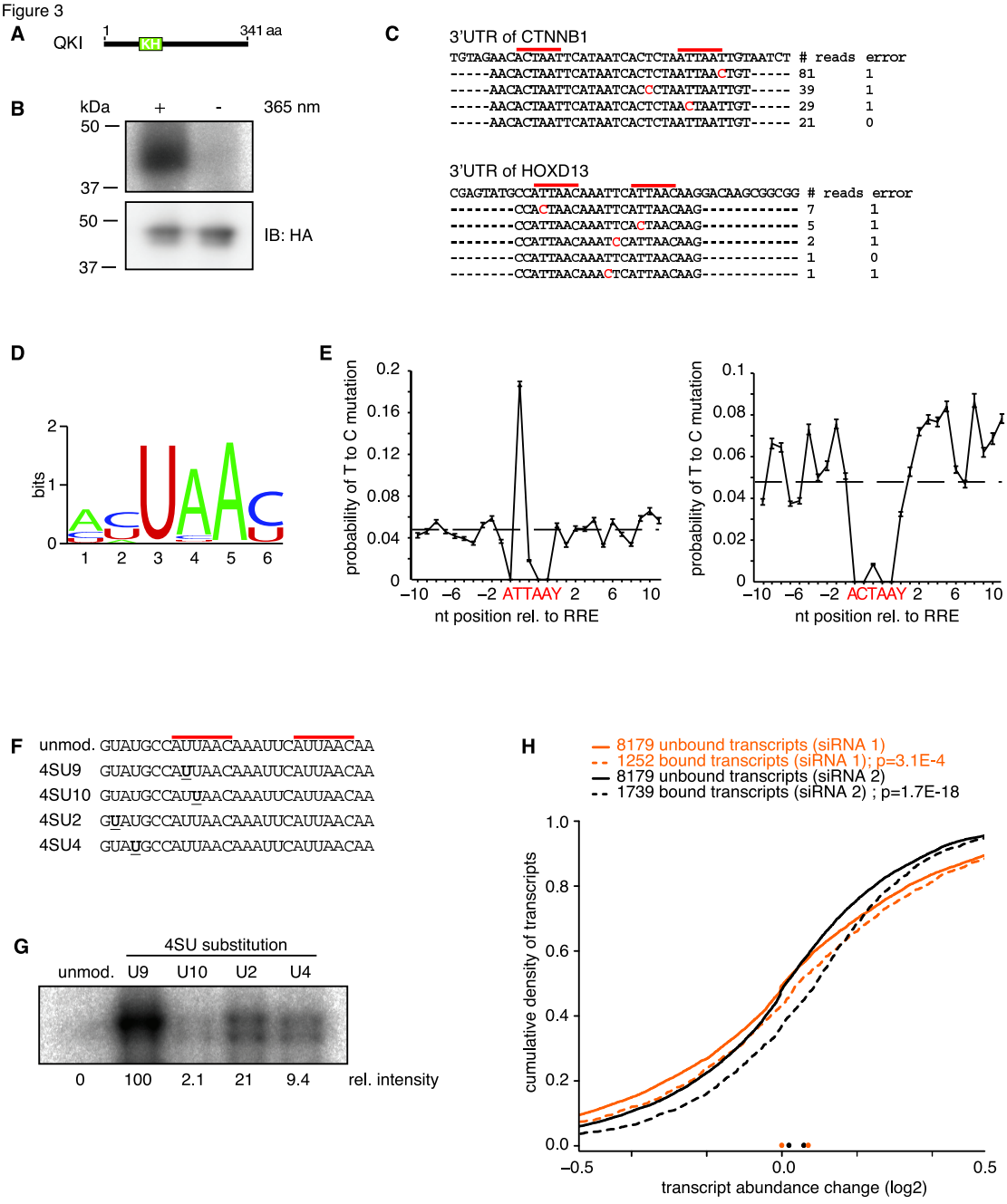
Mutation analysis of the clustered sequence reads showed that T corresponding to U2 in AUUAAY was frequently altered to C whereas the T corresponding to U3 in AUUAAY or ACUAAY remained unaltered (Figure 7.3E). Crosslinking of 4SU residues located in the immediate vicinity to the RRE was mostly responsible for exposing the motif with C2, showing that crosslinking inside the recognition element is not a precondition for its identification. Hence, the discovery of RREs is unlikely to be prevented by sequence-dependent crosslinking biases as long as deep enough sequencing captures these interaction sites at and nearby the RRE.

7.2.4 T to C mutations occur at the crosslinking sites

To better characterize the T to C transition observed in crosslinked RNA segments, we UV 365 nm crosslinked oligoribonucleotides containing single 4SU substitutions to recombinant QKI (Figures 7.3F, G). The crosslinking efficiency varied 50-fold and mirrored the results of the mutational analysis (Figure 7.3G). The least effective crosslinking was observed for placement of 4SU at position 3 of the QKI RRE (4SU9), and the most effective crosslinking was found at position 2 of the QKI RRE (4SU10); the crosslinking efficiency for two positions outside of the RRE (4SU2 and 4SU4) was intermediate. Neither of these substitutions affected RNA-binding to recombinant QKI protein as determined by gel-shift analysis, whereas mutations of the recognition element weakened the binding between 2.5- and 9-fold (Supplementary Table 7.5).

Next, we sequenced libraries prepared from non-crosslinked as well as QKI-protein-crosslinked oligoribonucleotides containing 4SU at indicated positions (Figure 7.3F). The fraction of sequence reads with T to C changes obtained from non-irradiated 4SU-containing oligoribonucleotides varied between 10 and 20%, and increased to 50 to 80% upon crosslinking (Supplementary Table 7.6). The variation of the degree of T to C changes in the crosslinked samples is most likely determined by background of non-crosslinked oligoribonucleotides. Presumably, the T to C transition frequency is increased upon crosslinking as a direct consequence of a chemical structure change of the 4SU nucleobase upon crosslinking to protein amino acid side chains, resulting in altered stacking or hydrogen bond donor/acceptor properties directing the preferential incorporation of dG rather than dA during reverse transcription (Supplementary Figure 7.8). At the doses of 4SU applied to cultured cells, about 1 out of 40 uridines was substituted by 4SU as determined by HPLC analysis of the nucleoside composition of total RNA. Assuming a 20% T to C conversion rate for a non-crosslinked 4SU-labeled site, we estimated that the average T to C conversion rate of 40-nt sequence reads derived from background non-crosslinked sequences will be near 5%. Clusters of sequence reads with average T to C conversion above this threshold, irrespective of the number of sequence reads, most certainly represent crosslinking sites.

Figure 3


A
QKI
1
341 aa
KH
B
kDa
+
-
365 nm
50
37
50
37
IB: HA
C
3'UTR of CTNNB1
TGTAGAACACTAATTCATAATCACTCTAATTAATTGTAATCT # reads error
-----AACACTAATTCATAATCACTCTAATTAACTGT----- 81 1
-----AACACTAATTCATAATCACCTAATTAATTGT----- 39 1
-----AACACTAATTCATAATCACTCTAACTAATTGT----- 29 1
-----AACACTAATTCATAATCACTCTAATTAATTGT----- 21 0
3'UTR of HOXD13
CGAGTATGCCATTAACAAATTCATTAACAAGGACAAGCGGCGG # reads error
--------CCACTAACAAATTCATTAACAAG---------- 7 1
--------CCATTAACAAATTCACTAACAAG---------- 5 1
--------CCATTAACAAATCCATTAACAAG---------- 2 1
--------CCATTAACAAATTCATTAACAAG---------- 1 0
--------CCATTAACAAACTCATTAACAAG---------- 1 1
D
bits
E
probability of T to C mutation
ATTAAY
nt position rel. to RRE
probability of T to C mutation
ACTAAY
nt position rel. to RRE
F
unmod. GUAUGCCAUUAACAAAUUCAUUAACAA
4SU9 GUAUGCCAUUAACAAAUUCAUUAACAA
4SU10 GUAUGCCAUUAACAAAUUCAUUAACAA
4SU2 GUAUGCCAUUAACAAAUUCAUUAACAA
4SU4 GUAUGCCAUUAACAAAUUCAUUAACAA
G
4SU substitution
unmod.
U9
U10
U2
U4
0
100
2.1
21
9.4
rel. intensity
H
8179 unbound transcripts (siRNA 1)
1252 bound transcripts (siRNA 1); p=3.1E-4
8179 unbound transcripts (siRNA 2)
1739 bound transcripts (siRNA 2) ; p=1.7E-18
cumulative density of transcripts
transcript abundance change (log2)

The ability to separate signal from noise by focusing on clusters with a high frequency of T to C mutations rather than clusters with the largest number of reads, represents a major enhancement of our method over UV 254 nm crosslinking methods.

To assess whether the transcripts identified by PAR-CLIP are regulated by QKI, we analyzed the mRNA levels of mock-transfected and QKI-specific siRNA-transfected cells with microarrays. Transcripts crosslinked to QKI were significantly upregulated upon siRNA transfection, indicating that QKI negatively regulates bound mRNAs (Figure 7.3H), consistent with previous reports of QKI being a repressor [123].

7.2.5 Identification of IGF2BP family RNA targets and its RRE

We then applied PAR-CLIP to the FLAG/HA-tagged insulin-like growth factor 2 mRNA-binding proteins 1, 2, and 3 (IGF2BP1-3) (Figures 7.4A, B), a family of highly conserved proteins that play a role in cell polarity and cell proliferation [124]. These proteins are predominantly expressed in the embryo and regulate mRNA stability, transport and translation. They are re-expressed in various cancers [125, 126] and IGF2BP2 has been associated with type-2 diabetes [127]. The IGF2BPs are highly similar and contain six canonical RNA-binding domains, two RNA recognition motifs (RRMs) and four KH domains (Figure 7.4A). Therefore, target recognition for this protein family appears complex, with only a small number of coding and non-coding RNA targets being known so far. A precise definition of the RREs is missing [124].

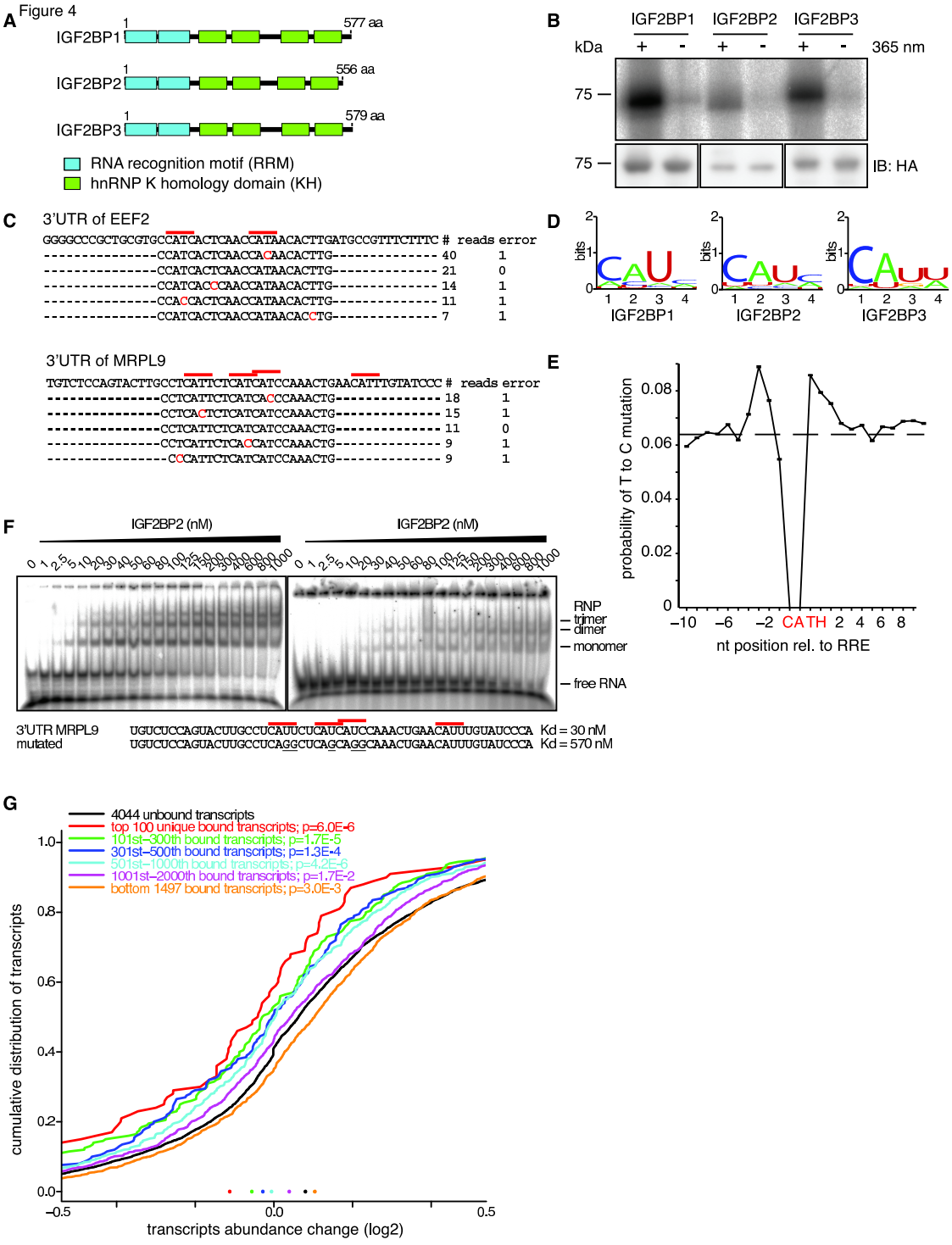
The three IGF2BPs recognized a highly similar set of target transcripts (Supple-

Figure 7.3 *(facing page)*: **RNA recognition by QKI protein**. (A) Domain structure of QKI protein. (B) Phosphorimage of SDS-gel resolving radiolabeled RNA crosslinked to FLAG/HA-QKI IPs from non-irradiated or UV-irradiated 4SU-labeled cells. The lower panel shows the anti-HA immunoblot. (C) Alignments of PAR-CLIP cDNA sequence reads to the corresponding regions in the 3'UTRs of the CTNNB1 and HOXD13 transcripts. Red bars indicate the QKI recognition motif and red-letter nucleotides indicate T to C sequence changes. (D) Sequence logo of the QKI recognition motif generated by PhyloGibbs analysis of the top 100 sequence read clusters. (E) T to C positional mutation frequency for PAR-CLIP clusters anchored at the AUUAAY (left panel) and ACUAAY (right panel) RRE; Y = U or C. The dashed line represents the average T to C mutation frequency within these clusters. (F) Sequences of synthetic 4SU-labeled oligoribonucleotides with QKI recognition motifs, derived from a sequence read cluster aligning to the 3'UTR of HOXD13 shown in (C) 4SU-modified residues are underlined. (G) Phosphorimage of SDS-gel resolving recombinant QKI protein after crosslinking to radiolabeled synthetic oligoribonucleotides shown in (F). (H) Stabilization of QKI-bound transcripts upon siRNA knockdown. Two distinct siRNA duplexes (1, orange traces and 2, black traces) were used for QKI knockdown and changes in transcript stability relative to mock transfection were inferred from microarray analysis. Shown are the distributions of changes upon siRNA transfection for transcripts that did (dashed lines) or did not (solid lines) contain QKI PAR-CLIP clusters. The p-values obtained in the Wilcoxon rank-sum test comparing the changes in targeted and non-targeted transcripts are indicated.

mentary Table 7.3), suggesting similar and redundant functions. PhyloGibbs analysis of the clusters derived from mRNAs (Figure 7.4C) yielded the sequence CAUH (H=A, U, or C) as the only consensus recognition element (Figure 7.4D), contained in more than 75% of the top 1000 clusters for IGF2BP1, 2 or 3 (Supplementary Figure 7.10). In total, we identified over 100,000 sequence clusters recognized by the IGF2BP family that map to about 8,400 protein-coding transcripts. The annotation of the clusters was predominantly exonic (ca. 90%) with a slight preference for 3'UTR relative to coding sequence (CDS) (Supplementary Figure 7.8). The mutation frequency of all sequence tags containing the element CAUH (H = A, C, or U) showed that the crosslinked residue was positioned inside the motif, or in the immediate vicinity (Figure 7.4E). The consensus motif CAUH was found in more than 75% of the top 1000 targeted transcripts, followed in more than 30% by a second motif, predominantly within a distance of three to five nucleotides (Supplementary Figure 7.10). *In vitro* binding assays showed that nucleotide changes of the CAUH motif decreased, but did not abolish the binding affinity (Figure 7.4F and Supplementary Table 7.5).

To test the influence of IGF2BPs on the stability of their interacting mRNAs, as reported previously for some targets [124], we simultaneously depleted all three IGF2BP family members using siRNAs and compared the cellular RNA from knockdown and mock-transfected cells on microarrays. The levels of transcripts identified by PAR-CLIP decreased in IGF2BP-depleted cells, indicating that IGF2BP proteins stabilize their target mRNAs. Moreover, transcripts that yielded clusters with the highest T to C mutation frequency were most destabilized (Figure 7.4G), indicating that the ranking criterion that we derived based on the analysis of PUM2 and QKI

Figure 7.4 *(facing page)*: **RNA recognition by the IGF2BP protein family**. (A) Domain structure of IGF2BP1-3 proteins. (B) Phosphorimage of an SDS-gel resolving radiolabeled RNA crosslinked to FLAG/HA-IGF2BP1-3 IPs. The lower panel shows anti-HA immunoblots. (C) Alignments of IGF2BP1 PAR-CLIP cDNA sequence reads to the corresponding regions of the 3'UTRs of EEF2 and MRPL9 transcripts. Red bars indicate the 4-nt IGF2BP1 recognition motif and nucleotides marked in red indicate T to C sequence changes. (D) Sequence logo of the IGF2BP1-3 RRE generated by PhyloGibbs analysis of the top 100 sequence read clusters. (E) T to C positional mutation frequency for PAR-CLIP clusters anchored at the 4-nt recognition motif from all motif-containing clusters. The dashed line represents the average T to C mutation frequency within these clusters. (F) Phosphorimage of native PAGE resolving complexes of recombinant IGF2BP2 protein with wild-type (left panel) and mutated target oligoribonucleotide (right panel). Sequences and dissociation constants (K_d) are indicated. (G) Destabilization of IGF2BP-bound transcripts upon siRNA knockdown. A cocktail of three siRNA duplexes targeting IGF2BP1, 2, and 3 was used, as well as a mock transfection and changes in transcript stability were monitored by microarray analysis. Distributions of transcript level changes for IGF2BP1-3 PAR-CLIP target transcripts versus non-targeted transcripts are shown. IGF2BP1-3 target sequences were ranked and divided into bins. The p-values indicate the significance of the difference between the changes of target versus non-target transcripts, as given by the Wilcoxon rank-sum test and are corrected for multiple testing.


A
Figure 4
IGF2BP1
577 aa
IGF2BP2
556 aa
IGF2BP3
579 aa
RNA recognition motif (RRM)
hnRNP K homology domain (KH)
B
IGF2BP1
IGF2BP2
IGF2BP3
kDa
365 nm
75
IB: HA
C
3'UTR of EEF2
reads
error
3'UTR of MRPL9
D
bits
IGF2BP1
IGF2BP2
IGF2BP3
E
probability of T to C mutation
nt position rel. to RRE
CATH
F
IGF2BP2 (nM)
RNP
trimer
dimer
monomer
free RNA
3'UTR MRPL9
mutated
Kd = 30 nM
Kd = 570 nM
G
4044 unbound transcripts
top 100 unique bound transcripts; p=6.0E-6
101st–300th bound transcripts; p=1.7E-5
301st–500th bound transcripts; p=1.3E-4
501st–1000th bound transcripts; p=4.2E-6
1001st–2000th bound transcripts; p=1.7E-2
bottom 1497 bound transcripts; p=3.0E-3
cumulative distribution of transcripts
transcripts abundance change (log2)

data generalizes to other RBPs.

For comparison to conventional and high-throughput sequencing CLIP [97, 98], we also sequenced cDNA libraries prepared from UV 254 nm crosslinking. Of the 8,226 clusters identified by UV 254 nm crosslinking of IGF2BP1, 4,795 were found in the PAR-CLIP dataset. Although UV 254 nm crosslinking identified the identical segments of a target RNA as PAR-CLIP, the position of the crosslink could not be readily deduced, because no abundant diagnostic mutation was observed (Supplementary Figure 7.11).

7.2.6 Identification of miRNA targets by AGO and TNRC6 family PAR-CLIP

To test our approach on RNP complexes, we selected the protein components mediating miRNA-guided target RNA recognition. In animal cells, miRNAs recognize their target mRNAs through base-pairing interactions involving mostly 6-8 nucleotides at the 5' end of the miRNA (the so called "seed") [103]. Target sites were thought to be predominantly located in the 3'UTRs of mRNAs, and computational miRNA target prediction methods frequently resort to identification of evolutionarily conserved sites that are located in 3'UTRs and are complementary to miRNA seed regions [103, 128].

We isolated mRNA fragments bound by miRNPs from HEK293 cell lines stably expressing FLAG/HA-tagged AGO or TNRC6 family proteins [129]. The AGO IPs revealed two prominent RNA-crosslinked bands of 100 and 200 kDa, representing AGO, and likely TNRC6 and/or DICER1 protein. The TNRC6 IPs showed one prominent RNA-crosslinked protein of 200 kDa (Figure 7.5A).

From clusters (Figure 7.5B) formed by at least 5 PAR-CLIP sequence reads and containing more than 20% T to C transitions, we extracted 41 nt long regions centered over the predominant T to C transition or crosslinking site. The length of the crosslink-centered regions (CCRs) was selected to include all possible registers of miRNA/target-RNA pairing interactions relative to the crosslinking site.

PAR-CLIP of individual AGO proteins yielded on average about 4,000 clusters that overlapped, supporting our earlier observation that AGO1-4 bound similar sets of transcripts [129]. We therefore combined the sequence reads obtained from all AGO experiments, which yielded 17,319 clusters of sequence reads at a cut-off of 5 reads. These clusters distributed across 4,647 transcripts with defined GeneIDs, corresponding to 21% of the 22,466 unique HEK293 transcripts that we identified by digital gene expression (DGE).

PAR-CLIP of individual TNRC6 proteins yielded on average about 600 clusters that also overlapped substantially, again consistent with our observation that TNRC6 family proteins bind similar transcripts [129]. We therefore combined all sequence reads from all TNRC6 experiments, yielding 1,865 clusters and CCRs. More than

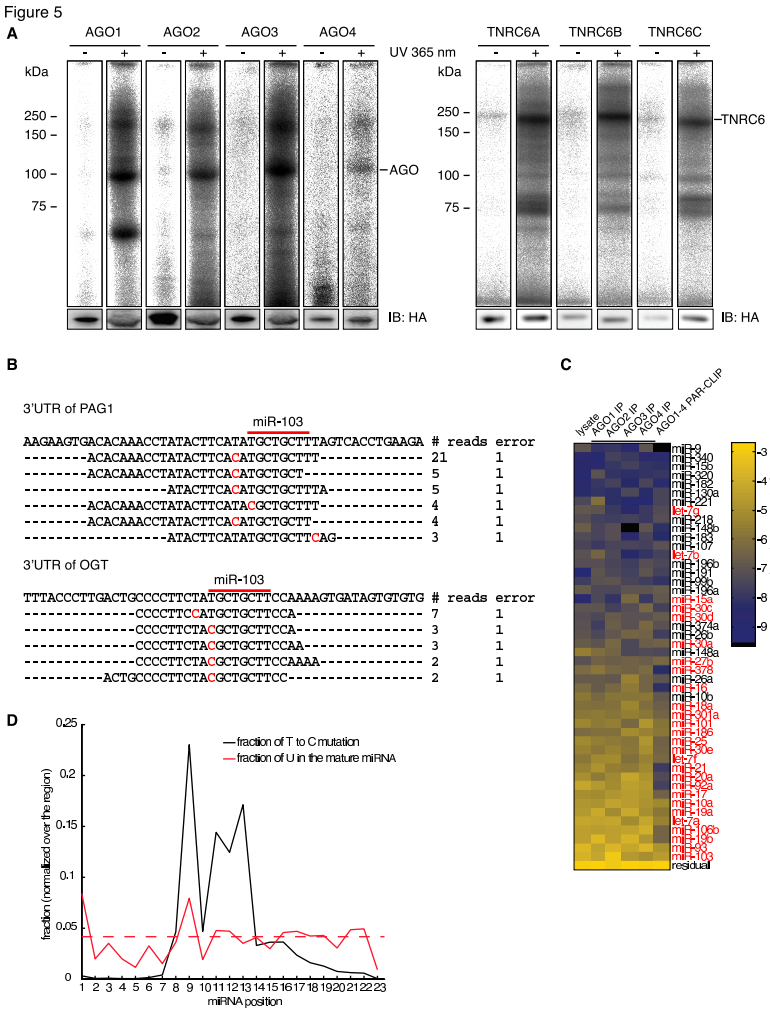


Figure 7.5: **AGO protein family and TNRC6 family PAR-CLIP** (A) Phosphorimage of SDS-gels resolving radiolabeled RNA crosslinked to the FLAG/HA-AGO1-4 and FLAG/HA-TNRC6A-C IPs. The lower panel shows the immunoblot with an anti-HA antibody. (B) Alignment of AGO PAR-CLIP cDNA sequence reads to the corresponding regions of the 3'UTRs of PAG1 and OGT. Red bars indicate the 8-nt miR-103 seed complementary sequence and nucleotides marked in red indicate T to C mutations. (C) miRNA profiles from RNA isolated from untreated HEK293 cells, non-crosslinked FLAG/HA-AGO1-4 IPs, and combined AGO1-4 PAR-CLIP libraries. The color code represents relative frequencies determined by sequencing. miRNAs indicated in red were inhibited by antisense oligonucleotides for the transcriptome-wide characterization of the destabilization effect of miRNA binding. (D) T to C positional mutation frequency for miRNA sequence reads is shown in black, and the normalized frequency of occurrence of uridines within miRNAs is shown in red. The dashed red line represents the normalized mean U frequency in miRNAs.

50% of these TNRC6 CCRs fell within 25 nt of an AGO CCR, and 26% overlapped by at least 75%, indicating that AGO and TNRC6 members bind to the same sites (Supplementary Figure 7.12).

7.2.7 Comparison of miRNA profiles from AGO PAR-CLIP to non-crosslinked miRNA profiles

To relate the potential miRNA-target-site-containing CCRs to the endogenously expressed miRNAs, we determined the miRNA profiles from total RNA isolated from HEK293 cells, and miRNAs isolated from non-crosslinked AGO1-4 IPs by Solexa sequencing [118], and compared them to the profile from the miRNAs present in the combined AGO1-4 PAR-CLIP library. miRNA profiles obtained from total RNA and IP of the four AGO proteins in non-crosslinked cells correlated well (Figure 7.5C) supporting our observation that AGO1-4 bind the same targets [129]. The most abundant among the 557 identified miRNAs and miRNAs* were miR-103 (7% of miRNA sequence reads), miR-93 (6.5%), and miR-19b (5.5%). The 25 and 100 most abundant miRNAs accounted for 72% and 95% of the total of miRNA sequence reads, respectively. Comparison of the miRNA profile derived from the combined AGO PAR-CLIP library with the combined non-crosslinked libraries showed a good correlation (Spearman correlation coefficient of 0.56, Figure 7.5C and Supplementary Figure 7.12).

Importantly, in the AGO PAR-CLIP library, the majority of miRNA sequence reads derived from prototypical miRNAs [130] displayed T to C conversion near or above 50%. The T to C conversion was predominantly concentrated within positions 8 to 13 (Figure 7.5D), residing in the unpaired regions of the AGO protein ternary complex [131]. Five of the 100 most abundant miRNAs in HEK293 cells lack uridines at position 8-13, yet only 2 of those miRNAs, miR-374a and b, showed no crosslinking, because uridines at residues 14 and higher can still be crosslinked (data not shown). This frequency of crosslinks was substantially lower in the miRNAs whose expression did not correlate between AGO-IP and AGO PAR-CLIP samples compared to the miRNAs whose expression correlated well (Supplementary Figure 7.12).

7.2.8 mRNAs interacting with AGOs contain miRNA seed complementary sequences

Independent of any pairing models for miRNAs and their targets, we first determined the enrichment of all 16,384 possible 7-mers within the 17,319 AGO CCRs, relative to random sequences with the same dinucleotide composition. The most significantly enriched 7-mers, except for a run of uridines, corresponded to the reverse complement of the seed region (position 2-8) of the most abundant HEK293 miRNAs, and they

were most frequently positioned 1-2 nt downstream of the predominant crosslinking site within the CCRs (Figure 7.6A). This places the crosslinking site near the centre of the AGO-miRNA-target-RNA ternary complex, where the target RNA is proximal to the Piwi/RNase H domain of the AGO protein [131]. The polyuridine motif lies within the region of target RNA that may be able to basepair with the 3' half of miRNA loaded into AGO proteins [131, 132]. Therefore, these stretches of uridine may contribute directly to miRNA-target RNA hybridization or, as has been suggested previously, they may represent an independent determinant of miRNA targeting specificity [133, 134].

To further examine the positional dependence of target RNA crosslinking, we aligned the CCRs containing 7-mer seed complements to the 100 most abundant miRNAs and plotted the position-dependent frequency of finding a crosslinked position (Figure 7.6B). This identified two additional crosslinking regions, which correspond to the unpaired 5' and 3' ends of the target RNA exiting from the AGO ternary complex, indicating that the window size of 41 nt centered on the predominant crosslink position always included the miRNA-complementary sites.

We then computed the number of occurrences of miRNA-complementary sequences of various lengths in the CCRs and calculated their enrichment. The most significant enrichment was generally obtained with 8-mers that were complementary to miRNA seed regions (pos. 1-8). Inspection of the region between 3 nt upstream and 9 nt downstream of the predominant crosslinking site reveals that approximately 50% of the CCRs contain 6-mers corresponding to one of the top 100 expressed miRNAs, with a 1.5-fold enrichment over random 6-mers (Supplementary Figure 7.12). Given that 6-mers still showed some degree of excess conservation in comparative genomics studies [1, 2] and that our analysis was focused on a narrow window directly downstream of the crosslinking site, our results suggest that the majority of the CCRs represent *bona fide* miRNA binding sites. Furthermore, the number of miRNA seed complements for all known miRNAs correlated well with the expression levels of miRNAs found in HEK293 cells, and less well with miRNA profiles of other tissue samples (Supplementary Figure 7.13).

The nucleotide composition of CCRs that contained at least one 7-mer seed complementary to one of the top 100 expressed miRNA showed a slightly elevated U-content (approx. 30% U) compared to those CCRs not containing seed matches (Supplementary Figure 7.13), which was expected from previous bioinformatic analyses of functional miRNA-binding sites.

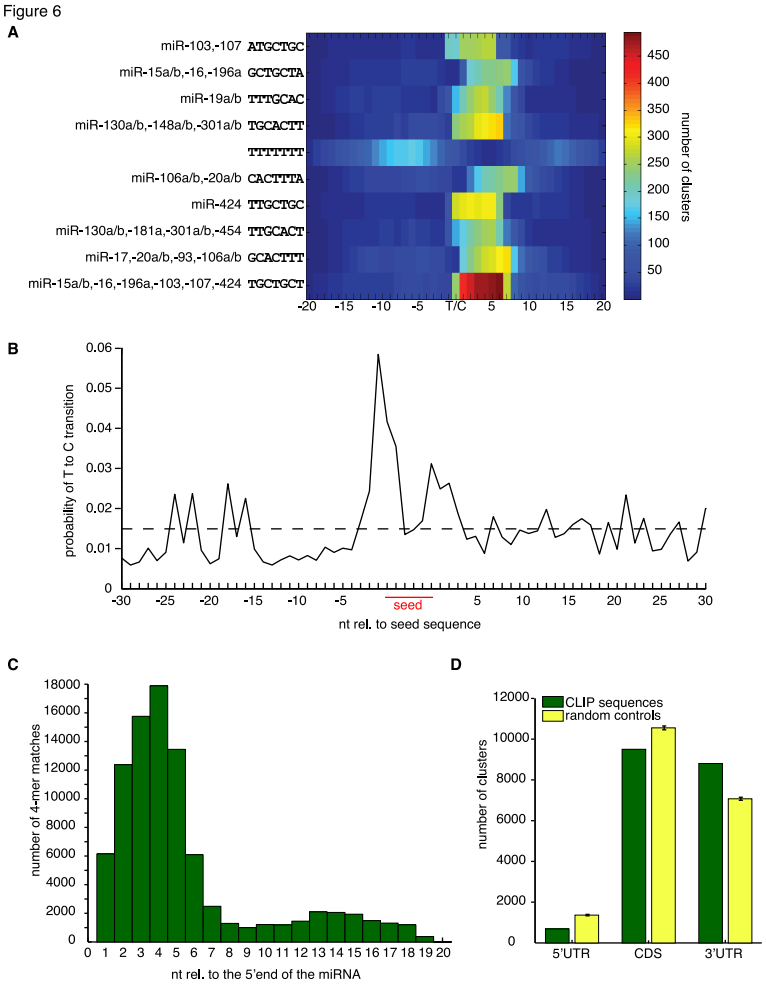


Figure 7.6: **AGO PAR-CLIP identifies miRNA seed-complementary sequences in HEK293 cells.** (A) Representation of the 10 most significantly enriched 7-mer sequences within PAR-CLIP CCRs. T/C indicates the predominant T to C transition within clusters of sequence reads. (B) T to C positional mutation frequency for clusters of sequence reads anchored at the 7-mer seed complementary sequence (pos. 2-8 of the miRNA) from all clusters containing seed-complementary sequences to any of the top 100 expressed miRNAs in HEK293 cells. The dashed line represents the average T to C mutation frequency within the clusters. (C) Identification of 4-nt base-pairing regions contributing to miRNA target recognition. CCRs with at least one 7-mer seed complementary region to one of the top 100 expressed miRNAs were selected. The number of 4-nt contiguous matches in the CCRs relative to the 5'end of the matching miRNA was counted. (D) Analysis of the positional distribution of CCRs. The number of clusters annotated as derived from the 5'UTR, CDS or 3'UTR of target transcripts is shown (green bars). Yellow bars show the expected location distribution of the crosslinked regions if the AGO proteins bound without regional preference to the target transcript.

7.2.9 Non-canonical and 3'end pairing of miRNAs to their mRNA targets is limited

Structural and biochemical studies of the ternary complex of *T. thermophilus* Ago, guide and target indicated that small bulges and mismatches could be accommodated in the seed pairing region within the target RNA strand [131]. We therefore searched for putative target RNA binding sites that did not conform to the model of perfect miRNA seed pairing, but rather contained a discontinuous segment of sequence complementarity to either target or miRNA with a minimum of 6 base pairs. We only considered pairing patterns if they were significantly enriched in CCRs compared to dinucleotide randomized sequences, and if the CCRs containing them did not at the same time contain perfectly pairing seed-type sites. We identified 891 CCRs with mismatches and 256 with bulges in the seed region. Mismatches occurred most frequently across from position 5 of the miRNA as G-U or U-G wobbles, U-U mismatches and A-G mismatches (A residing in the miRNA). Therefore, it appears that only a small fraction of the miRNA target sites that we isolated (less than 6.6%), contained bulges or loops in the seed region.

To assess the role of auxiliary base pairing outside of the seed region, we selected CCRs that contained a 7-mer seed match to one of the 100 most abundant miRNAs. Supporting earlier computational results [133], we also detected a weak signal for contiguous 4-nt long matches to positions 13-18 of the miRNA (Figure 7.6C).

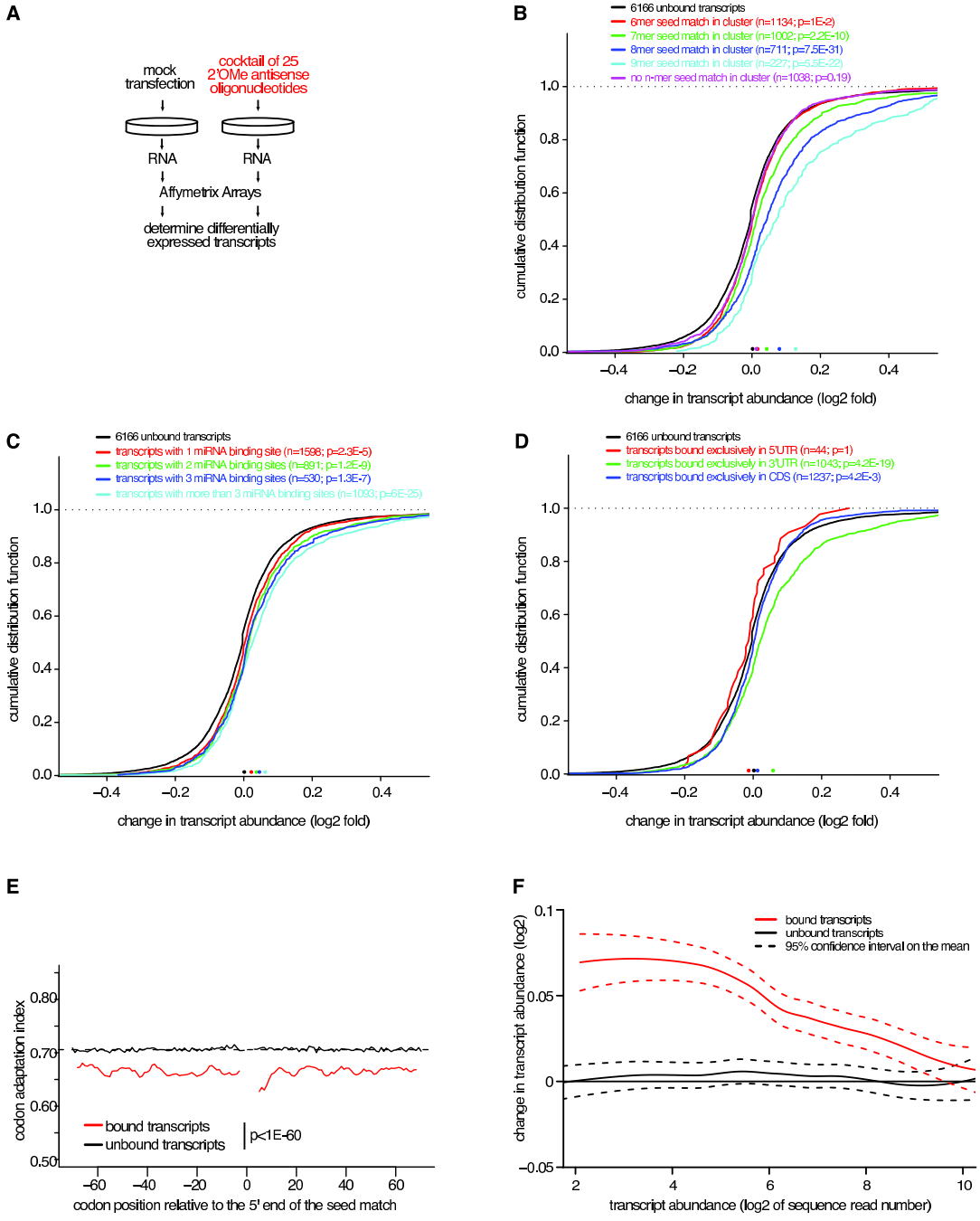
7.2.10 miRNA binding sites in CDS and 3'UTR destabilize target mRNAs to different degrees

The majority (84%) of AGO CCRs originated in exonic regions, with only 14% from intronic, and 2% from undefined regions. Of the exonic CCRs, 4% corresponded to 5'UTRs, 50% to CDS, and 46% to 3'UTRs (Figure 7.6D).

Evidence of widespread binding of miRNAs to the CDS was reported before [1, 135]. However, miRNAs are believed to predominantly act on 3'UTRs [103], with relatively few reports providing experimental evidence for miRNA-binding to individual 5'UTRs or CDS [135–139].

To obtain evidence that AGO CCRs indeed contain functional miRNA-binding sites, we blocked 25 of the most abundant miRNAs in HEK293 cells (Figure 7.5C) by transfection of a cocktail of 2'-O-methyl-modified antisense oligoribonucleotides and monitored the changes in mRNA stability by microarrays (Figure 7.7A). Consistent with previous studies of individual miRNAs [133], the magnitude of the destabilization effects of transcripts containing at least one CCR depended on the length of the seed-complementary region and dropped from 9-mer to 8-mer to 7-mer to 6-mer matches (Figure 7.7B). We did not find evidence for significant destabilization of transcripts that only contained imperfectly paired seed regions.

Figure 7


A
mock transfection
cocktail of 25 2'OMe antisense oligonucleotides
RNA
RNA
Affymetrix Arrays
determine differentially expressed transcripts
B
6166 unbound transcripts
6mer seed match in cluster (n=1134; p=1E-2)
7mer seed match in cluster (n=1002; p=2.2E-10)
8mer seed match in cluster (n=711; p=7.5E-31)
9mer seed match in cluster (n=227; p=5.5E-22)
no n-mer seed match in cluster (n=1038; p=0.19)
cumulative distribution function
change in transcript abundance (log2 fold)
C
6166 unbound transcripts
transcripts with 1 miRNA binding site (n=1598; p=2.3E-5)
transcripts with 2 miRNA binding sites (n=891; p=1.2E-9)
transcripts with 3 miRNA binding sites (n=530; p=1.3E-7)
transcripts with more than 3 miRNA binding sites (n=1093; p=6E-25)
cumulative distribution function
change in transcript abundance (log2 fold)
D
6166 unbound transcripts
transcripts bound exclusively in 5'UTR (n=44; p=1)
transcripts bound exclusively in 3'UTR (n=1043; p=4.2E-19)
transcripts bound exclusively in CDS (n=1237; p=4.2E-3)
cumulative distribution function
change in transcript abundance (log2 fold)
E
codon adaptation index
bound transcripts
unbound transcripts
p<1E-60
codon position relative to the 5' end of the seed match
F
bound transcripts
unbound transcripts
95% confidence interval on the mean
change in transcript abundance (log2)
transcript abundance (log2 of sequence read number)

Next, we examined whether the change in stability of CCR-containing transcripts correlated with the number of binding sites. We found that multiple sites were more destabilizing compared to single sites (Figure 7.7C), and that multiple binding sites may also reside within a single 41-nt CCR (Supplementary Figure 7.13). Both of these findings are in agreement with previous observations [133].

Then we analyzed the impact on stability for transcripts with CCRs exclusively present either in the CDS or the 3'UTR; there were not enough transcripts to assess the impact of CCRs derived from the 5'UTR. CDS-localized sites only marginally reduced mRNA stability (Figure 7.7D), independent of the extent of seed pairing. To gain more insights into miRNA binding in the CDS, we examined the codon adaptation index (CAI) [140] around crosslinked seed matches, and found that the sequence environment of crosslinked seed matches differed from that of non-crosslinked seed matches in the CAI. The bias in codon usage extended for at least 70 codons up- as well as downstream of the crosslinked seed matches (Figure 7.7E), which also correlates well with the marked increase in the A/U content around the binding sites that would lead to a codon usage bias. It was recently reported that miRNA regulation in the CDS was enhanced by inserting rare codons upstream of the miRNA-binding site, presumably due to increased lifetime of miRNA-target-RNA interactions as ribosomes are stalled [141]. These observations suggest that transcripts with reduced translational efficiency form at least transient miRNP complexes amenable to UV crosslinking.

The abundance of mRNAs expressed in HEK293 cells varied over 5 orders of mag-

Figure 7.7 *(facing page)*: **Relationship between various features of miRNA/target RNA interactions and mRNA stability.** (A) FLAG/HA-AGO2-tagged HEK293 cells were transfected with a cocktail of 25 2'-O-methyl modified antisense oligoribonucleotides, inhibiting miRNAs marked in red in Figure 7.5C, or mock transfected, followed by microarray analysis of the change of mRNA expression levels. (B) Transcripts containing CCRs were categorized according to the presence of n-mer seed complementary matches and the distributions of stability changes upon miRNA inhibition are shown for these categories. The stability change for transcripts harboring CCRs without identifiable miRNA seed-complementary regions is also shown. The p-values indicate the significance of the difference between the transcript level changes of transcripts containing CCRs versus transcripts without CCRs, as given by the Wilcoxon rank-sum test and are corrected for multiple testing. (C) Transcripts were categorized according to the number of CCRs they contained. (D) Transcripts were categorized according to the positional distribution of CCRs. Only transcripts containing CCRs exclusively in the indicated region are used. (E) Codon adaptation index (CAI) for transcripts containing 7-mer seed complementary regions (pos. 2-8) in the CDS for the miR-15, miR-19, miR-20, and let-7 miRNA families. The red and the black lines indicate the CAI for seed-complementary sequence containing transcripts bound and not bound by AGO proteins determined by AGO PAR-CLIP. (F) LOESS regression of total transcript abundance in HEK 293 cells (log2 of sequence counts determined by digital gene expression (DGE)) against fold change of transcript abundance (log2) determined by microarrays after transfection of the miRNA antagonist cocktail versus mock transfection of AGO-bound and unbound transcripts.

nitude as shown by DGE profiling. When we related the expression level of CCR-containing transcripts with the magnitude of transcript stabilization after miRNA inhibition, we found that miRNAs preferentially act on transcripts with low and medium expression levels (Figure 7.7F). Highly expressed mRNAs appear to avoid miRNA regulation [142], at least for those miRNAs expressed in HEK293 cells. However, we cannot fully rule out that the weaker response of highly abundant targets may be due to lower affinity and reduced occupancy of miRNA binding sites in highly abundant transcripts.

Earlier studies defining miRNA target regulation were carried out by transfection of miRNAs into cellular systems originally devoid of these miRNAs [143–145]. We transfected miRNA duplexes corresponding to the deeply conserved miR-7 and miR-124 into FLAG/HA-AGO2 cells, performed PAR-CLIP, and also recorded the effect on mRNA stability upon miR-7 and miR-124 transfection by microarray analysis. Transcripts containing miR-7- or miR-124-specific CCRs were destabilized, especially when CCRs were located in the 3'UTR (Supplementary figure 7.14).

7.2.11 Context-dependence of miRNA binding

Not every seed-complementary sequence in the HEK293 transcriptome yielded a CCR, thereby providing an opportunity to identify sequence context features specifically contributing to miRNA target binding and crosslinking. For seed-complementary sites that were crosslinked and those that were not crosslinked, we computed the evolutionary selection pressure by the ElMMo method [2], the mRNA stability scores by TargetScan context score [133], and sequence composition and structure measures for the regions around the miRNA seed complementary sites. The feature that distinguished most crosslinked from non-crosslinked seed matches was a 25% lower free energy required to resolve local secondary structure involving the miRNA-binding region (Supplementary Figure 7.14), associated with a 6% increase in the A/U content within 100 nt around the seed-pairing site. These differences were similar for sites located in the CDS and 3'UTRs. Compared to non-crosslinked sites, crosslinked sites are under stronger evolutionary selection (ElMMo) and in sequence contexts facilitating miRNA-dependent mRNA degradation (TargetScan context score).

The location of AGO CCRs within transcript regions was non-random and 7-mer or 8-mer sites within the 3'UTR were preferentially located near the stop codon or the polyA tail in transcripts with relatively long 3'UTRs (more than 3 kb) (Supplementary Figure 7.14). The location of CCRs in the CDS was biased towards the stop codon for the transfected miR-7 and 124, but not for the endogenous miRNAs (Supplementary Figure 7.14).

Finally, we wanted to examine how miRNA targets defined by PAR-CLIP compared in regulation of target mRNA stability to those predicted by ElMMo [2], TargetScan context score [133], TargetScan Pct [146] and PicTar [147]. In each case, we

selected the same number of highest-scoring sites containing a 7-mer seed-complement to the top 5 expressed miRNAs (let-7a, miR-103, miR-15a, miR-19a and miR-20a). The analysis was limited to 3'UTR sites due to restriction by the prediction methods. The effect on mRNA stability, as assessed by miRNA antisense inhibition, was overall equivalent for transcripts harboring CCRs compared to transcripts predicted by ElMMo, TargetScan context score, TargetScan Pct and PicTar [147] (Supplementary Figure 7.14).

7.3 Discussion

Maturation, localization, decay and translational regulation of mRNAs involve formation of complexes of RBPs and RNPs with their RNA targets [99, 100]. Several hundred RBPs are encoded in the human genome, many of them containing combinations of RNA-binding domains which are drawn from a relatively small repertoire, resulting in diverse structural arrangements and different specificities of target RNA recognition [87]. Furthermore hundreds of miRNAs function together with AGO and TNRC6 proteins to destabilize target mRNAs and/or repress their translation [103]. Collectively, these factors and their presumably combinatorial action constitute the code for post-transcriptional gene regulation. Here we describe an approach to directly identify transcriptome-wide mRNA-binding sites of regulatory RBPs and RNPs in live cells.

7.3.1 PAR-CLIP allows high-resolution mapping of RBP and miRNA target sites

We showed that application of photoactivatable nucleoside analogs to live cells facilitates RNA-protein crosslinking and transcriptome-wide identification of RBP and RNP binding sites. We concentrated on 4SU after it became apparent that the crosslinking sites in isolated RNAs were revealed upon sequencing by a prominent transition from T to C in the cDNA prepared from the isolated RNA segments. Compared to regular UV 254 nm crosslinking in the absence of photoactivatable nucleosides, our method has two distinct advantages. We obtain higher yields of crosslinked RNAs using similar radiation intensities, and more importantly, we can identify crosslinked regions by mutational analysis. Studies using conventional UV 254 nm CLIP have not reported the incidence of deletions and substitutions [97, 98, 114], except for recent work by Grannemann et al. on the U3 snoRNA that showed an increase of deletions at the RBP binding site [112]. Our own analysis indicates that mutations in sequence reads derived from UV 254 nm CLIP were at least one order of magnitude less frequent than T to C transitions observed in PAR-CLIP (Supplementary Figure 7.8).

From an experimental perspective, it is important to note that crosslinked RNA segments, irrespective of the methods of isolation, are always contaminated with non-crosslinked RNAs, as shown by consistent identification of rRNAs, tRNAs, and miRNAs. Compared to crosslinked RNA fragments, these unmodified RNA molecules are more readily reverse transcribed, which underscores the need for separation of crosslinked signal from non-crosslinked noise. We now provide a method that accomplishes this critical task.

7.3.2 Context dependence of 4SU crosslink sites

It is conceivable that binding sites located in peculiar sequence environments, e.g. those completely devoid of U, may exist and cannot be captured using 4SU-based crosslinking. However, such sites are extremely rare. Only about 0.4% of 32-nt long sequence segments, representative of the length of our Solexa sequence reads, are U-less, corresponding to an occurrence of one such segment in every 8 kb of a transcript.

Nonetheless, to provide a means to resolve such unlikely situations, we explored the use of other photoactivatable nucleosides, such as 6SG to identify IGF2BP1 binding sites. We found a good correlation between the sequence reads obtained from a given gene with 4SU and 6SG (Pearson correlation coefficient 0.65, Supplementary Table 7.4). Moreover, the sequence read clusters, representing individual binding sites, overlapped strongly: 59% out of the 47,050 6SG clusters were also identified with 4SU, despite of the fact that the environment of IGF2BP1 binding sites was strongly depleted for guanosine. Interestingly, the sequence reads obtained after 6SG crosslinking were enriched for G to A transitions, pointing to a structural change in 6SG analogous to the situation in PAR-CLIP with 4SU. Because 6SG appears to have lower crosslinking efficiency compared to 4SU, we recommend to first use 4SU and then resort to 6SG when the data indicates that the sites of interest are located in sequence contexts devoid of uridines. It is important to point out that neither of these photoactivatable nucleotides appears to be toxic under our recommended conditions.

7.3.3 miRNA target identification

When applying PAR-CLIP to isolate miRNA-binding sites, we were surprised to find nearly 50% of the binding sites located in the CDS. However, miRNA inhibition experiments showed that miRNA binding at these sites only caused small, yet significant mRNA destabilization. In spite of the difference in their efficiency of triggering mRNA degradation, CDS and 3'UTR sites appear to have similar sequence and structure features. The sequence bias around CDS sites is associated with an increased incidence of rare codon usage, which could in principle reduce translational rate, thereby providing an opportunity for transient miRNP binding and regulation. Similar observations were made previously using artificially designed reporter systems [141].

The use of the knowledge of the crosslinking site allowed us to narrowly define the miRNA-binding regions for matching the site with the most likely miRNA endogenously co-expressed with its targets, and to assess non-canonical miRNA-binding modes. We were able to explain the majority of PAR-CLIP binding sites by conventional miRNA-mRNA seed-pairing interactions [133], yet found that about 6% of miRNA target sites might best be explained by accepting bulges or mismatches in the seed pairing region, similar to the interaction between let-7 and its target lin-41 [148] and those recently observed in biochemical and structural studies of *T. thermophilus* Ago protein [131, 132].

7.3.4 The mRNA ribonucleoprotein (mRNP) code and its impact on gene regulation

We were able to identify all of the crosslinkable RNA-binding sites present in about 9,000 of the top-expressed mRNA in HEK293 cells representing approximately 95% of the total mRNA molecules of a cell. One of the surprising outcomes of our study was that each of the examined RBPs or miRNPs bound and presumably controlled between 5 and 30% of the more than 20,000 transcripts detectable in HEK293 cells. These results demonstrate that a transcript will generally be bound and regulated by multiple RBPs, the combination of which will determine the final gene-specific regulatory outcome. Exhaustive high-resolution mapping of RBP- and RNP-target-RNA interactions is critical, because it may lead to the discovery of specific combination of sites (or modules) that may control distinct cellular processes and pathways. To gain further insights into the dynamics of mRNPs it will be important to also map the sites of RNA-binding factors, such as helicases, nucleases or polymerases, where the specificity determinants are poorly understood. The precise identification of RNA interaction sites will be extremely useful for interrogating the rapidly emerging data on genetic variation between individuals and whether some of these variations possibly contribute to complex genetic diseases by affecting post-transcriptional gene regulation.

7.4 Methods

7.4.1 PAR-CLIP

Human embryonic kidney (HEK) 293 cells stably expressing FLAG/HA-tagged IGF2BP1-3, QKI, PUM2, AGO1-4, and TNRC6A-C [129] were grown overnight in medium supplemented with 100 μM 4SU. Living cells were irradiated with 365 nm UV light. Cells were harvested and lysed in NP40 lysis buffer. The cleared cell lysates were treated with RNase T1. FLAG/HA-tagged proteins were immunoprecipitated with

anti-FLAG antibodies bound to Protein G Dynabeads. RNase T1 was added to the immunoprecipitate. Beads were washed and resuspended in dephosphorylation buffer. Calf intestinal alkaline phosphatase was added to dephosphorylate the RNA. Beads were washed and incubated with polynucleotide kinase and radioactive ATP to label the crosslinked RNA. The protein-RNA complexes were separated by SDS-PAGE and electroeluted. The electroeluate was proteinase K digested. The RNA was recovered by acidic phenol/chloroform extraction and ethanol precipitation. The recovered RNA was turned into a cDNA library as described [118] and Solexa sequenced. The extracted sequence reads were mapped to the human genome (hg18), human mRNAs and miRNA precursor regions. Transfection of siRNAs and mRNA profiling by array analysis were described previously [129]. For a more detailed description of the methods, see the Supplementary Material.

7.4.2 Supplementary Information: Bioinformatic Analysis

7.4.2.1 Adapter removal and sequence annotation

The basic method for removing adaptors and assigning a functional annotation to the sequence reads was described in [149]. Briefly, we used an in-house ends-free local alignment algorithm (score parameters: 2 for match, -3 for mismatch, -2 for gap opening, -3 for gap extension) to align the Solexa adapter to the 3' end of each sequence read, allowing for the possibility that the adapter was not completely sequenced [1]. We removed from the reads the fragments that aligned to the adaptor as long as the number of matches exceeded that of mismatches by at least 3. Sequences that were either too short (less than 20 nt) or too repetitive (using a cut-off of 0.7 and 1.5 in the entropy of the mono- and dinucleotide distributions, respectively, of individual sequence reads [149] were discarded because they would probably map to multiple genomic locations. The remaining sequences were mapped to the hg18 version of the human genome assembly that was downloaded from the University of California at Santa Cruz (*http://genome.cse.ucsc.edu*) and to a database of sequences whose function (rRNA, tRNA, sn/snoRNA, miRNA, mRNA, etc.) is already known. These were obtained from the sources specified in [149]. The Oligomap algorithm [149] was used for this purpose, and all the perfect and 1-error (mismatch or insertion or deletion (indel) mappings were obtained. Based on the GMAP [150] genome mapping of human mRNA transcripts from NCBI downloaded on November 4, 2008, we determined whether the sequence reads mapped to intronic or exonic regions of genes. Based on the coding region annotation of transcripts in GenBank, we determined whether the exonic sequence reads originated from the 5'UTR, CDS or 3'UTR.

7.4.2.2 Generation of clusters of mapped sequence reads

For subsequent analyses only sequence reads that were at least 20 nucleotides long and mapped uniquely to the genome with at most one error were used. A single-linkage clustering of the sequence reads was performed, with two reads being placed in the same cluster if they overlapped by at least one nucleotide in their genomic mappings. Each cluster was then annotated based on the functional annotation of sequence reads that covered most of the cluster length. We then considered all the mRNA-annotated clusters containing at least 5 mRNA-annotated sequence reads, and we defined a scoring scheme to identify the clusters that had the highest probability of being real crosslinking sites (see below: Identification of high confidence clusters).

[1] Software can be downloaded from *http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html*.

7.4.2.3 Analysis of the mutational spectra

From the clusters defined above, all sequence reads were used that mapped uniquely and with one error (mismatch or indel) to the genome to infer the mutational bias of the method. For each library, we calculated the proportion of mutations involving each of the four nucleotides as well as the proportion of each of the four nucleotides in the crosslinked sequence reads (see Figure 7.81B and C).

7.4.2.4 Identification of high-confidence clusters

We used the crosslinked clusters of PUM2 and QKI, to define criteria for selecting high-confidence binding sites. The criteria that we tested reflected the mechanistic aspects of generating the sequence reads. Our preliminary analysis revealed that T to C mutations are by far the most frequently observed mutations in these data sets, and that they are most frequent inside or in the immediate vicinity of the binding motifs as opposed to the rest of the sequence (see Figs. 7.2E, 7.3E, and 7.4E). This suggested that the observed mutational bias is directly linked to the crosslinking event and should thus be a good criterion for separating true crosslinked sites from background sequence reads. The preliminary analysis also indicated a strong bias for having G nucleotides at the last position of a sequence read and also at the genomic position immediately upstream of a sequence read. This bias reflects the sequence specificity of the RNase T1, and may again help in the identification of sequence reads that map to multiple sites or for discriminating random RNA turnover products unrelated to RNase T1 treatment. Finally, we observed that many clusters with abundantly sequenced reads contained more than one position with a T to C mutation. The results of testing these criteria for their ability to select clusters that contained the known binding motif for QKI and PUM2 are shown in Figure 7.9. For QKI, binding motifs were defined as occurrences of ACUAA or AUUAA, which we identified from a very small number of clusters. The first of these motifs was also identified previously through SELEX experiments [122]. For PUM2, in order to account for additional motif variants besides the consensus UGUANAUA, binding motifs were identified as matches to the weight matrix (as inferred by MotEvo [3] that resulted from the motif search (see below). We found that ranking of the clusters by the number of T to C mutations in all reads in the clusters of sequence reads leads to the strongest enrichment in clusters with a binding site (Figure 7.9). The figures show the fraction of the crosslinked clusters that contain at least one occurrence of the known binding motif as a function of the number of clusters that passed a given cut-off in the selection criterion (e.g. total number of sequence reads, total number of T to C mutations, total number of sequence reads with a G at position -1 relative to their genomic locus). The cut-off decreases from the left to the right of the x-axis. It is clear that, particularly for PUM2, the number of T to C mutations strongly correlates with the

presence/absence of the motif in the cluster. For comparison, we also show the same plots when using as the ranking criterion not the total number of T to C mutations in the cluster, but just the total number of sequence reads per cluster. For QKI, this leads to a significantly lower enrichment of clusters with recognition elements. We also investigated how the fraction of clusters with the known binding motif depends on the number of distinct crosslinking positions (i.e. positions with at least one T to C mutation) inside the cluster (Figure 7.9). The fraction of clusters with a binding site increases steadily from 0 to 5 crosslinking positions for both proteins, with the strongest increase from 0 to 1 for PUM2 and between 0 and 2 crosslinking positions for QKI. When requiring that at least two positions with T to C mutations are present in the cluster, the fraction of clusters with a binding site increases roughly by 20 % for PUM2, and by more than 40 % for QKI. These considerations led us to the following procedure for defining high confidence clusters for any given RBP. We first selected all the clusters with at least two crosslinking positions and, secondly, within this subset, we ranked all clusters by the total number of T to C mutations in all sequence reads in the cluster.

7.4.2.5 Extraction of peaks and crosslink-centered regions (CCRs) from sequence read clusters

From each ranked, mRNA-annotated cluster, a peak region, defined as a 32-nt long region with the highest average sequence read density, was extracted. Because the T to C mutation was diagnostic for the site of crosslinking, we focused our motif analysis on regions anchored at the position in a cluster with the most T to C mutations. We then investigated the mutational profile around this position and we found that this profile approaches the background profile after about 20 nt to the left and right of the main site of T to C mutations. Thus, these 41-nt long regions centered on the main site of T to C mutations are most likely to contain the binding sites and we focused our motif search on these regions.

7.4.2.6 RNA recognition element search

For the motif search defining the core of a RNA recognition site we selected, for each RBP, the top 100 high confidence clusters, defined as described above. We selected the 41-nt region centered on the main T to C mutation site and searched for over-represented sequence motifs using PhyloGibbs [120]. We used a first-order Markov model as the background model and searched each set of sequences for three motifs of lengths varying between 4 and 8 nt, demanding an expected total number of 50 motifs. For each parameter setting, we performed five replicate runs. This generally resulted for each RBP in various shifted versions of the same motif. Therefore we hierarchically clustered all the weight matrices that we obtained from these runs, allowing

for partial overlap of at least 4 nucleotides between pairs of weight matrices. In the clustering procedure, two weight matrices were fused if the posterior probability of their stemming from the same as opposed to two different probability distribution was larger than 0.2 (for a description of the Bayesian calculation, see [149], section 4.1). Replicating this procedure multiple times yielded very similar results (not shown). For each protein, we selected the largest cluster of weight matrices, i.e. the cluster that contained most of the weight matrices that we obtained in replicate runs, and created the final weight matrix by summing up the counts for each nucleotide of the weight matrices belonging to this cluster. Since the clustering procedure also allows the fusion of only partially overlapping weight matrices, the resulting weight matrices are typically longer (roughly 10 nucleotides) than the motif length that we imposed in individual runs, and can contain stretches of low information content. We therefore selected for each RBP, the window with highest information content. For PUM2 and QKI, the length of this window was 8 and 6 nt, respectively, in accordance with the known or expected consensus motifs [106, 122], respectively. For the IGF2BPs, we chose a window length of 4 nt, which is believed to be the size of binding motifs of KH-domains [90]. To identify binding sites in PUM2 clusters of aligned sequence reads using the inferred weight matrix, we used the MotEvo algorithm [3], which is based on a hidden Markov model that models the input sequences as contiguous stretches of nucleotides drawn from a background or a weight matrix model. We chose for the background a first order Markov model (which makes every nucleotide dependent on the preceding nucleotide in the sequence). The background model parameters (dinucleotide frequencies) were estimated from the set of input sequences. MotEvo was run in the prior-update mode, meaning that we attempted to find the prior probabilities for sites and background that maximize the likelihood of the sequence data. MotEvo generates as an output a list of sites for the given input weight matrix as well as their corresponding posterior probabilities. Note that not all matches to the weight matrix are reported, but only the subset of matches whose corresponding sequence is more likely under the weight matrix model than the background model. We chose a cut-off of 0.4 on the posterior probability to define the set of binding sites.

7.4.2.7 Determination of the location of sequence read clusters within functional mRNA regions

For each RBP, we investigated whether clusters of mapped sequence reads preferentially originated in 5'UTR, CDS or 3'UTR (Figure 7.8A). As a result of our annotation pipeline, we could assign probabilities to each cluster to belong to either 5'UTR, CDS and 3'UTR based on the annotation of individual sequence reads within the cluster (see above). Taking together these probabilities for all clusters, we obtained estimates of the numbers of clusters originating in each of these three regions. We compare these numbers to those that we would expect if clusters were sampled uniformly from any-

where along the transcripts. This would for instance result in many more clusters from 3' compared to 5'UTR regions simply because 3'UTRs tend to be longer than the 5'UTRs. We determined all the transcripts to which a cluster mapped, and based on the GenBank annotation of the CDS of these transcripts, we calculated the fraction of the cluster nucleotides that fell in the 5'UTR (f_5), CDS (f_{CDS}), and 3'UTR (f_3). In the cases in which the cluster mapped to several transcripts belonging to the same gene, these fractions were averaged over all transcripts. The expected proportion of nucleotides sequenced from each region was calculated by summing these fractions for all clusters. The variance was determined by noting that the probability that a nucleotide was sampled from a particular region, e.g. 5'UTR, is Bernoulli distributed with parameter f_5, which has a variance of $f_5(1 - f_5)$. The total variance was then computed as the sum of all the variances.

7.4.2.8 Distance distribution between consecutive CAU-motifs in the IGF2BP RNA binding sites

Since each of the IGF2BPs has 4 KH domains and we found only one clear motif, we hypothesized that all KH domains have the same or a very similar binding specificity. In analogy to what has been observed for the neuronal RBP involved in splicing, Nova [94], we propose that the binding specificity of the IGF2BPs arises from the concerted action of several KH-domains that each recognize the same 4 letter sequence (CAUH), which should be apparent by a preferred spacing between subsequent occurrences of the motif as determined by the distance of corresponding KH-domains in the structure of the IGF2BPs. We calculated, for each IGF2BP separately, the distribution of distances between subsequent occurrences of the CAU-motif in clusters unambiguously derived from the 3'UTR of protein coding genes. We restricted ourselves to these clusters since 3'UTR regions are overrepresented in clusters of the IGF2BPs and each region, 5'UTR, CDS and 3'UTR, has different sequence biases that need to be taken into account when modeling background distributions. In order to reduce boundary effects due to the finite length of the clusters, we extended each cluster region 32 nt to the right and left (the genomic regions are shown on the website: *http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html*) . We then compared this distance distribution to the distance distribution of consecutive occurrences of the CAU motif in randomly chosen 3'UTR regions of the same length distribution as the clusters of mapped sequence reads. To estimate the mean and standard deviation of the relative frequency of each inter-motif distance in the background dataset, we repeated the random selection of 3'UTR regions 1000 times.

7.4.2.9 Enrichment of identified binding motifs in all clusters

We defined the binding motifs for PUM2, QKI and IGF2BPs using a subset of high-confidence clusters for each protein. To determine to what extent these motifs were indeed representing the binding sites of the proteins in the complete data sets, we collected, for each protein and for each cluster, all the respective crosslink-centered regions (CCRs) and ranked them by the number of T to C mutations. We then calculated for varying cut-offs on the number of T to C mutations the fraction of clusters above the given cut-off that contain at least one binding site (Figure 7.10, blue traces). The binding site was defined to be UGUANAUA for PUM2, ACUAA or AUUAA for QKI and CAU or two CAUs separated by no more than 10 nucleotides for the IGF2BPs. To estimate the number of sites expected by chance, we generated 1000 sets of random sequences with the same nucleotide frequencies as the CCRs (dinucleotide shuffling for PUM2 as well as QKI and mononucleotide shuffling for the IGF2BPs, due to the small length of the binding motif). For all proteins, the CCRs are clearly enriched in the respective binding motifs. The enrichment is strongest for PUM2, which has the longest recognition motif. For the IGF2BPs, the enrichment for the CAU-spacer-CAU motif is much stronger than for the CAU motif due to the clustering of the CAU motif (see previous section).For PUM2, we additionally determined the enrichment only for the first half of motif UGUA. This short motif is clearly enriched and is contained in more than 72 percent of all CCRs, suggesting the presence of other variants of the PUM2 motif besides the consensus UGUANAUA.

7.4.2.10 Analysis of siRNA knockdown experiments

We imported the CEL files into the R software[1] using the BioConductor affy package (Gentleman et al., 2004). The transcript probe set intensities were background-corrected, adjusted for non-specific binding and quantile normalized with the GCRMA algorithm (Wu, 2006). Probe sets with more than 6 of the 11 probes mapping ambiguously to the genome were discarded, as were probe sets that mapped to multiple genes. We then collected all probe sets matching a given gene, and we selected for further analysis the RefSeq transcript with median 3'UTR length corresponding to that gene. In total 16,063 transcripts were identified. The log-intensity of probe sets mapping to the gene were then averaged to obtain the expression level per RefSeq transcript. The changes of transcript abundances were computed as the logarithm of the ratio of transcript expression in the cocktails of siRNA treated samples and mock-transfected cells.

To study the effect of individual proteins on the mRNA stability of their targets, we performed the following analysis. We first made the links between clusters of mapped Solexa sequence reads and expression data based on the NCBI Gene ID.

[1] *http://www.R-project.org*

That is, both the transcripts that were crosslinked and those whose expression was measured on microarrays have associated Gene IDs in the Gene database of NCBI. We mapped both the mapped sequence read clusters as well as the transcripts on microarrays to their corresponding genes, and thus identified which genes that were represented on microarrays have been crosslinked. From this set of genes we removed those that are likely off-targets of the transfected siRNAs. As previous studies showed, complementarity to the first 8 nucleotides of the miRNA is a good indicator that the transcript will be downregulated by a miRNA or siRNA, so we defined as putative off-targets those genes whose representative RefSeq transcripts carried such complementary sites in their 3'UTR. We divided the list of genes sorted by the maximum score of any cluster associated with a given gene.In order to improve the target identification and the assessment of the target response, we used some specific information that was available for individual data sets. For instance, for the IGF2BPs we only considered clusters with at least 2 positions of T to C changes, because we previously observed that this criterion improves the accuracy of target identification for the PUM2 and QKI. Thus, for the IGF2BPs we divided the bound transcripts into the following bins, top 100 genes, 101-300 genes, 301-500 genes and 501-1000 genes and calculated the log2fold change of transcript abundance. To determine whether the siRNA knockdown has an effect on mRNA stability, we compared these distributions with the distribution of log-fold changes of genes that did not have any associated clusters from CLIP analysis. For QKI, we performed the same analysis starting from clusters with a single T to C mutation site, but that additionally contained the QKI motif.

7.4.2.11 Generation and ranking of clusters of mapped sequence reads for AGO and TNRC6 family PAR-CLIP

To generate sequence read clusters for the cDNA libraries from the AGO and TNRC6 PAR-CLIP we used sequence reads of at least 20 nt in length and with unique, perfect or 1-error mapping to the genome. We clustered the reads with single-linkage criterion, meaning that we placed two reads in the same cluster if they overlapped by at least one nucleotide in their genomic mappings. We then selected the clusters that contained at least 5 mRNA-annotated reads and at least 2 positions at which T to C mutations occurred in the sequence reads relative to the genomic sequence, and we ranked them by the total number of T to C mutations which, as we described above, is indicative of the number of crosslinks.

7.4.2.12 Definition of CCRs for sequence read clusters of AGO and TNRC6 PAR-CLIP

In each ranked, mRNA-annotated cluster we identified the position with the largest number of T to C mutations, and we constructed the mutation frequency profile around this position. We found that this profile approaches the background after about 20 nucleotides to the left and right of the position with the maximum number of T to C changes, and we therefore extracted a genomic region of 41 nucleotides centered on this position for further analyses.

7.4.2.13 Filtering to remove unspecific "background" clusters for AGO and TNRC6

Because it is still possible that a substantial number of the clusters we obtained contain degradation products of abundantly expressed mRNAs and because a number of proteins that associate with the RISC complex have a molecular weight that is similar to that of AGO proteins and may be responsible for some of the sequence reads/clusters that we obtained in the experiment with FLAG-tagged AGO we have collected PAR-CLIP data for a number of proteins and identified the AGO-specific clusters as follows. We built similar clusters for all the proteins that we investigated (PUM2, QKI, IGF2BP1-3, AGO1-4, TNRC6A-C), we compared the clusters, and when two clusters bound by two different proteins overlapped by more than 75% of their total length we considered that the two proteins shared a cluster. Finally, we discarded the following AGO clusters: clusters in which no position had a T to C mutation rate greater than 0.2, the experimentally determined T to C mutation rate at non-crosslinked sites; clusters that were shared between AGO libraries and libraries of other RBPs, with the number of sequence reads in the AGO libraries being less than 1/10 of the number of sequence reads in the other library. After applying these filters we obtained 17,319 AGO1-4 binding regions. We applied the same procedure to the clusters that we obtained from miR-124 and miR-7 transfection experiments.

7.4.2.14 Analysis of crosslinked position with respect to miRNA seed-complementary sequence

We identified all the target regions (T to C anchored regions of 41 nucleotides) that have an 8-mer (A opposite miRNA position 1 and perfect match at miRNA positions 2-8) seed match and we extended symmetrically the seed-complementary region by 20 nt to the left and right. We then computed the positional T to C mutation frequency in these regions and normalized it over the length of the target region.

7.4.2.15 Identification of pairing regions of miRNAs within CCRs

To determine whether positions other than the seed region may be involved in base-pairing interaction with targets, we first took the T to C anchored target regions and identified those that had at least a 6-mer (2-6 and A opposite miRNA position 1, 2-7 or 3-8) seed complementarity to at least one of the top 100 most expressed miRNAs in HEK293 cells. For each of these T to C anchored regions and each miRNA that matched to it, we identified all the occurrences of complementarities of at least 4 nucleotides between the miRNA and the putative target region. Each of these was counted with a weight 1/n towards the positional profile of miRNA-target site matches, with n being the number of miRNAs that matched the putative target region.

7.4.2.16 Analysis of transcript stabilization as a function of the type of miRNA binding sites

We constructed the distribution of log-fold-changes of transcripts with various types of PAR-CLIP clusters, and we compared them with the distribution of log-fold-changes of transcripts that did not yield PAR-CLIP clusters, although they were expressed, as determined by the microarray measurements. The categories of transcripts were the following:

1. Transcripts with various types of miRNA seed matches
   At most 6-mer match: 1-6 (with A opposite miRNA position 1), 2-7, 3-8, 4-9 match to at least one of the top 100 most abundant miRNAs.
   At most 7-mer match: 1-7 (with A opposite miRNA position 1), 2-8, 3-9 match to at least one of the top 100 most abundant miRNAs.
   At most 8-mer match: 1-8 (with A opposite miRNA position 1), 2-9 match to at least one of the top 100 most abundant miRNAs.
   At most 9-mer match: 1-9 (with A opposite miRNA position 1) match to at least one of the top 100 most abundant miRNAs.

2. Transcripts with PAR-CLIP clusters originating exclusively in a particular transcript region (5'UTR, CDS, 3'UTR).

3. Transcripts with 1, 2, 3, 4 or more non-overlapping PAR-CLIP clusters.

7.4.2.17 Digital Gene Expression (DGE)

The sequence reads from the DGE (Illumina) experiments have been analyzed in a manner similar to that described above in the section "Adapter removal and sequence annotation". We only considered genomic and transcript matches containing the

GATC recognition sequence of the DpnII restriction enzyme directly upstream of the mapped sequence read. For our analyses we further used sequence reads that had a perfect match in the genome. The probability that a sequence read originates in a given locus was then computed as 1/n of loci to which the sequence read can be mapped. The sequence reads were also mapped to the mRNA sequences and then we computed an expression level per gene. This was defined as the sum of the weighted copies of all sequence reads that can be mapped to transcripts that originate in that gene. Finally, to assess the accuracy of the expression level measurements, we correlated the logarithm of the expression level measured Affymetrix GeneChip® microarray with the logarithm expression level measured using the DGE technology. The Spearman correlation coefficient was 0.68. We found a considerable number of transcripts that could be detected by sequencing (22,465) and that were undetectable on the microarrays (on which we measured 16,063 transcripts). Correlation between biological replicates of HEK293 cells was higher than 0.99.

7.4.2.18 Analysis of miRNA-induced destabilization of crosslinked and non-crosslinked miR-124 and miR-7 targets

We intersected the transcripts with the background-noise-filtered PAR-CLIP clusters obtained after miR-124 and miR-7 transfection (see “Filtering to remove unspecific “background” clusters for AGO and TNRC6” section above) with those for which we had destabilization and AGO-IP Affymetrix microarray measurements. We then constructed, for each miRNA, three non-overlapping sets of transcripts: those with PAR-CLIP clusters exclusively in the 3'UTR, with PAR-CLIP clusters exclusively in the CDS, and transcripts that did not yield any PAR-CLIP clusters. For each set, we computed the average log2 fold change upon miRNA transfection, and the average log2 fold enrichment in the AGO-IP. We compared these values between transcripts with and transcripts without PAR-CLIP clusters (Figure 7.14). The error bars on the bar plot represent 95% confidence intervals on the mean log2 fold changes. Finally, we performed Wilcoxon's rank sum test to assess the significance of the difference in the log2 fold changes of pairs of transcript sets. We also looked at various combinations of CLIP cluster locations (Figure 7.14) that occurred more than 25 times in a given data set. Finally, we compared the destabilization and AGO-binding of crosslinked and non-crosslinked single miR-124 and miR-7 seed matches (Figure 7.14). A seed match was defined as a match to nucleotides 1-7, 2-8 or 1-8 of the miRNA (both miRNAs start with U, so a 1-7 or 1-8 seed match also means having an A opposite nucleotide 1 of the miRNA). A seed match was considered "crosslinked" if it overlapped with a CLIP cluster from the corresponding transfection library.

7.4.2.19 Estimation of miRNA expression based on SOLEXA sequencing

The miRNA profile was generated from Solexa sequencing runs containing small RNAs from the following libraries: AGO1- IP and lysates of AGO1-4 IP, which were combined and denoted lysate in Figure 7.5C. The miRNA annotation was preformed as described in [130, 149].

7.4.2.20 Plots of motif frequency versus enrichment

We performed a 7-mer word enrichment analysis based on the T to C anchored target regions from the miRNA transfection experiments. We enumerated all words of length 7 and we determined their frequency in the real set as well as in a background set of shuffled sequences with the same dinucleotide content. For each 7-mer, we then calculated its enrichment as the ratio of the two frequencies. Additionally, we calculated for each 7-mer the posterior probability that the frequency of the 7-mer is different in foreground and background allowing for sampling noise [149]. To determine whether the enriched motifs may correspond to miRNAs, all significantly enriched motifs (with a posterior ≥ 0.99) were aligned with Needleman-Wunsch algorithm (penalties: gapopening -4, gapextension -4) to the reverse complement of the transfected and to the top 20 most expressed in HEK293 miRNAs. We only reported cases in which the enriched word mapped with 0 or 1 errors to the first 9 positions of one of these miRNAs.

7.4.2.21 Identification of significantly enriched types of miRNA binding sites

In order to identify individual miRNA binding sites in the sequence data we first defined a set of putative "binding models". These were either contiguous matches to at least 6 nucleotides of a miRNA, or matches that had a single structural defect. This was defined as either an internal loop or a bulge either in the miRNA or in the mRNA. For each of the 553 miRNAs we enumerated all these binding models, and we determined the enrichment of the T to C anchored regions in each of these models, relative to the average over 10 dinucleotide randomized sequence sets. Using a cutoff of 1.0e-20 in the probability that the real set had a lower frequency of occurrence compared to the randomized sets, which we used as a measure of the significance of the enrichment, we found all the T to C anchored regions that contained at least one significantly enriched binding model from one of the top 100 most expressed miRNAs within 10 nucleotides of the T to C mutation site. To obtain a comprehensive list of target sites we added to these the 7-mer nucleotide matches (within the same 10 nucleotides of the T to C mutation) to positions 1-7 or 2-8 of one of the top 100 most expressed miRNAs, irrespective of whether the T to C anchored regions were enriched in these 7-mers.

7.4.2.22 Correlation of miRNA seed family expression with frequencies of occurrence of seed-complementary motif

From all samples of smirnadb [130], all miRNAs that had at least 50 counts in total from all samples were used to build seed groups (defined by the motif found at positions 2-8). We added an additional sample, which was generated by pooling together the miRNA reads from deep sequencing of HEK293 small RNA as well as AGO1-4 IPs without crosslinking. For each sample, we computed the expression of a seed group as the sum of the sequence reads of all miRNAs that were part of the seed group. We correlated the seed expression with the frequency of the seed-complementary motif in the T to C anchored regions.

7.4.2.23 Co-occurrence of miRNA seed pairs within CCRs

To determine if the crosslinked regions are enriched in pairs of binding sites for highly expressed miRNAs. Assuming that not all of these sites may have been captured in our experiment, we used for this purpose the 17,319 cluster regions that we extended by 32 nucleotides on either side. We scanned these regions for non-overlapping 7-mers corresponding to the positions 2-8 of the top 20 most expressed miRNAs in HEK293 cells. We performed a similar procedure using 100 randomized variants of the extended clusters that preserved the dinucleotide composition. As additional controls we performed, first, the same procedure using 20 randomly selected miRNAs (Figure 7.13F) and secondly counting of the number of seed match pair occurrence in the extended clusters for 100 sets of 20 randomly selected miRNAs (Figure 7.13H). A visualization of seed match pair occurrence is shown in Figure 7.13G.

7.4.2.24 Properties of crosslinked and non-crosslinked miRNA seed matches

For the analyses whose results are presented in Figure 7.14 we needed to intersect the CLIP transcript sets with the transcript set measured by the Affymetrix microarray. In order to study the properties of crosslinked and predicted but non-crosslinked seed complementary matches we do not need to make this intersection, and we therefore considered the entire set of miRNA seed matches that are present in the representative RefSeq transcripts. We chose as the representative RefSeq transcript for a given gene that transcript that had the median 3'UTR length from all RefSeq transcripts corresponding to a gene. RefSeq transcripts that could not be detected in the DGE transcriptome profiling were discarded. For the analysis of the miR-124 and miR-7 transfection libraries, we scanned the 5'UTR, CDS and 3'UTRs of representative expressed RefSeq transcripts for 7-mer or 8-mer seed matches to miR-124 or miR-7, and intersected these with the background-noise-filtered miR-124 and miR-7 PAR-CLIP clusters to identify the crosslinked and non-crosslinked seed matches. In parallel, we scanned the 5'UTR, CDS and 3'UTRs of representative expressed RefSeq transcripts

for 7-mer and 8-mer seed matches to miR-15, miR-20, miR-103, miR-19, let-7 representing the top expressed miRNA families in HEK293 cells. These seed matches were then separated into crosslinked and non-crosslinked based on the intersection with the background-noise-filtered AGO1-4, PAR-CLIP clusters. Furthermore, because we wanted to analyze properties of the environment of the putative miRNA target sites, we only considered seed matches located at least 100 nucleotides away from either of the boundaries of the transcript. For each individual seed match, we computed the following quantities:

Selection pressure: is the posterior probability that a seed complementary region is under evolutionary selection pressure, as computed by the ElMMo algorithm described in [2].

Predicted destabilization score: is a score that characterizes the extent to which the environment of a seed match is favorable for its functionality in mRNA destabilization, as computed by the TargetScanS method [133]. For the analysis, we downloaded the TargetScan 5.1 from the *www.TargetScan.org* website.

Local AU content: is the proportion of A + U nucleotides within 50 nucleotides upstream and 50 nucleotides downstream of the miRNA binding site, defined as a 20 nt-long region, anchored at the 3'end by the seed-matching region.

Target site Eopen: is similarly defined in terms of the energy required to open the secondary structure of the target in a region of 20 nucleotides anchored at the 3'end by the seed-complementary region (opposite positions 1-8 of the miRNA). This was computed using the program RNAup of the Vienna package [151] with the following parameters: u=20 (length of the window required to be single-stranded), w=50 (maximal length of the interacting region). The rest of the parameters were left with their default values. The negative value of this energy can be viewed as a measure of accessibility.

We tested whether the four properties introduced above took significantly different values when comparing crosslinked to non-crosslinked seed matches using Wilcoxon's rank sum test.

7.4.2.25 Codon adaptation index around crosslinked and non-crosslinked seed matches

We compared the Codon Adaptation Index (CAI) [140] around crosslinked and non-crosslinked seed matches as follows. We estimated an optimal human codon usage by analyzing all the CDS from the 25% highest expressed genes among all the genes expressed in at least one of the two "whole brain" samples of the SymAtlas project [152]. This set of genes was determined by reanalyzing the two Affymetrix CEL files using the pipeline described above in the 'Analysis of miRNA knockdown and overexpression experiments' section. We then anchored all sequences at the codon covering the 5' end of seed match (1-7, 2-8, or 1-8 of miR-15, miR-20, miR-103,

miR-19, let-7 miRNAs) and computed the CAI for the 70 codons upstream and downstream of the anchor, i.e. a total of 141 codons. The 7-mer or 8-mer seed match is entirely covered by codons 0, 1 and 2, which highly constrains the codon usage at these positions, making it uninformative. The figure therefore does not show the CAI at these positions. For crosslinked seed matches, we smoothed the profile using a moving average of 5.

7.4.2.26 Analysis of positional bias of crosslinked and non-crosslinked regions

We set to determine whether crosslinked seed matches (1-7, 2-8, or 1-8 of miR-15, miR-20, miR-103, miR-19, let-7 miRNAs) have a positional bias relative to the STOP codon. Noting that at least in the 4 AGO PAR-CLIP libraries, crosslinked seed matches tended to be located in CDS of shorter lengths than their non-crosslinked counterparts, we performed local polynomial regression [153], fitting the distance between the seed matches and the STOP codon to the CDS length (Figures 7.14M, N). The loess fit and 95% confidence interval on the distance to the STOP codon given the CDS length were obtained using R's loess and predict loess functions with default parameters. The miRNA transfection and AGO PAR-CLIP libraries were separately analyzed, and loess fits were computed separately for crosslinked and non-crosslinked seed matches (Figs. 7.14K-N, shown in red and black, respectively). Finally, we represented the expected distance to the STOP codon as a function of the CDS length assuming that seed matches are distributed uniformly over the CDS (dashed blue curve). We used the same methodology to determine whether crosslinked sites are located preferentially towards a 3'UTR boundary (stop-codon or polyA-tail) instead of the stop-codon.

7.4.2.27 Comparison of the set of targets determined by the experimental assay (PAR-CLIP) and computational methods (ElMMo, TargetScan 5.1)

We computed the number of seed matches to each of the top 5 expressed miRNA families in the top 1000 CCRs from the AGO-PAR-CLIP. For each of these 5 miRNA families, we selected an equal number of target sites predicted by the ElMMo method, located on the mRNAs that could be detected in the DGE expression profiling (i.e. with at least one tag count), and starting from targets predicted with highest confidence. In addition, only genes that are expressed above the median on the arrays (i.e., the arrays in which the miRNAs are inhibited or not present) were considered in the analysis. We repeated the procedure using the TargetScan context scores, TargetScan PCT and Pictar. The ElMMo and TargetScan miRNA prediction methods only scan the mRNA 3'UTRs for target sites. Therefore, we determined a fourth set of miRNA

target sites through keeping only the CCRs harboring a seed match to at least one of the top 5 miRNA families, and located in the 3'UTR region of an mRNA. Finally, for each of these 6 sets of miRNA targets and each of the top 5 miRNA families, we determined the average log2 fold change in gene expression upon transfecting the antisense 2'-O-methyl oligonucleotide cocktail as well as the 95% confidence interval on the mean log2 fold change. We performed the same analysis on the miR-7 and miR-124 transfection data sets, this time analyzing only CCRs containing seed matches to miR-7 or miR-124.

7.4.2.28 Stability of transcripts containing CCRs with 6-mer seed complementary matches

For all mRNAs representative of genes detected through DGE profiling, we computed the number of 3'UTR-located 6-mer and 7-mer (or longer) seed matches to the top 5 expressed miRNA families. We then plotted the mean log2 fold change in gene expression following the transfection of the antisense 2'-O-methyl oligonucleotide cocktail as a function of the number of 6-mer and 7-mer (or better) seed matches, as well as the 95% confidence interval on the mean log2 fold change. Finally, we performed the same analysis on the miR-7 and miR-124 transfection data sets, this time analyzing only seed matches to miR-7 and miR-124.

Supplementary Figure S1

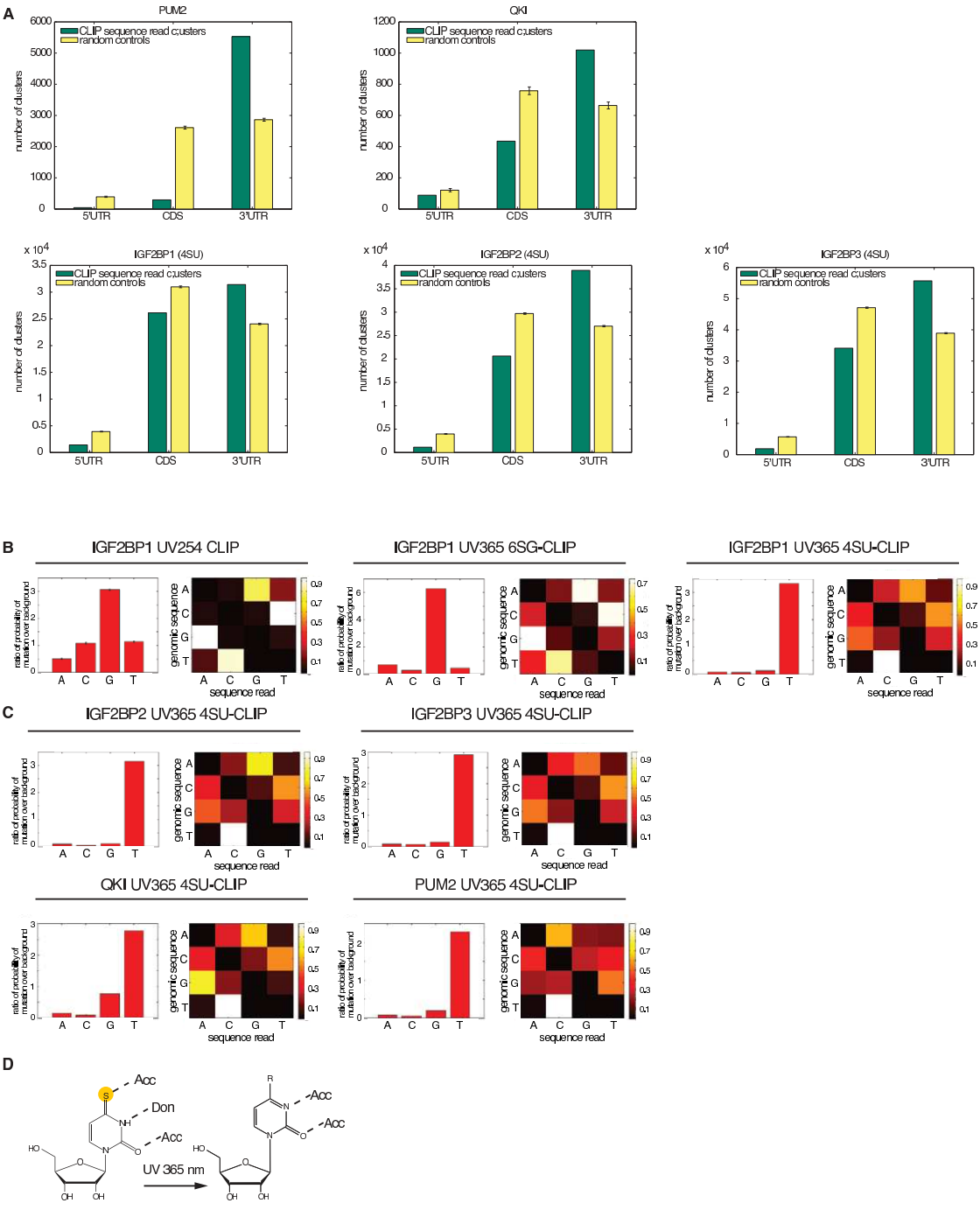


Figure 7.8 *(facing page)*: **Analysis of the transcript regional preferences and the mutational pattern of crosslinked sequences of PUM2, QKI, and IGF2BP1-3**. **(A)** For each 4SU-CLIP for the respective protein, the number of exonic sequence read clusters annotated as derived from the 5'UTR, CDS or 3'UTR of a target transcript is shown (green bars). Yellow bars show the expected location distribution of clusters if the RBPs bound without regional preference to the set of target transcripts. **(B)** Comparison of the mutational patterns observed with traditional UV 254 nm CLIP of HEK293 cells stably expressing FLAG/HA-tagged IGF2BP1 and that observed with UV 365 nm CLIP of cells grown in 6SG or 4SU containing medium. For each experimental condition two panels are shown: the left one indicates the mutation frequency of each of the four nucleotides relative to the frequency of occurrence of these nucleotides in all sequence reads; the right one shows, for each of the four nucleotides, the frequency of mutation towards each of the three others. In the right panels, white indicates high mutation frequency towards a particular nucleotide. In general, transitions are more frequent than other mutation types. Traditional 254 nm CLIP generates mutations preferably on Gs (left panel). Mutations after UV254 CLIP were twice as frequent at G compared to any other position (left panel) and predominantly identified as G to A transition (shown by the matrix in the right panel). Treatment of cells with 6SG (middle two panels, top row) resulted in a marked preference for mutations at G, about one order of magnitude compared to the other nucleotides with a preferred substitution of the G with an A. The preference for mutations at G is much more pronounced relative to that observed in the 254 nm crosslinked cells. 4SU-CLIP yields about a 30-fold increased mutation preference for T, nearly always to C. **(C)** Same analysis as in (B) for IGF2BP2 and 3, QKI, and PUM2. The mutational biases for these proteins are comparable. T is almost exclusively targeted for mutation, and is preferentially sequenced as C. **(D)** The increase in T to C transitions after 4SU-protein crosslinking can be rationalized by structural changes in donor/acceptor properties of 4SU after crosslinking to proximal amino acid side chains and subsequent incorporation of dG rather than dA in the reverse transcription; R representing a side chain.

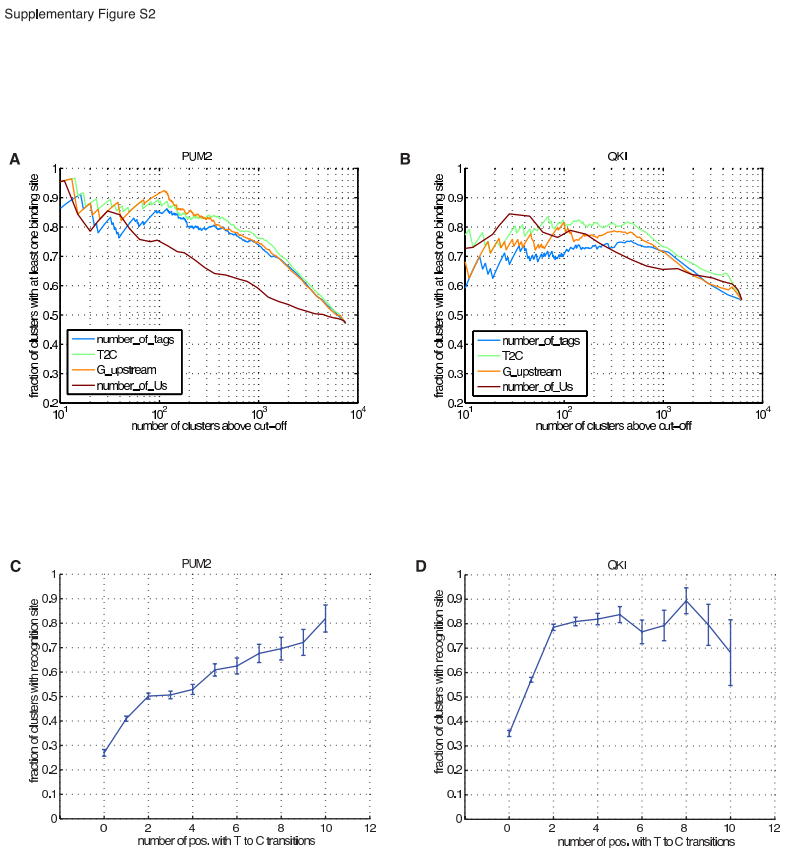


Figure 7.9: **Presence of the PUM2 and QKI recognition sequences in clusters generated by PAR-CLIP from cell lines stably expressing the respective epitope-tagged protein**. **(A)** Fraction of clusters containing the PUM2-recognition motif, versus the total number of clusters above a given cut-off on a particular property as indicated in each figure legend (G upstream: number of sequence reads with a G at position -1; T to C: number of sequence reads with a T to C mutation; number of sequences: total number of sequence sequence reads in the cluster, number_of_Us: number of uridines in the sequence read cluster. For each cut-off on a given property, the fraction of clusters with at least one binding site above the given cut-off is shown. Cut-off increases from right to left. The best signal is obtained by sorting according to the frequency of crosslinking events. **(B)** Same as in (A) for QKI. **(C-D)** Fraction of clusters with the recognition element (as indicated) for PUM2 **(C)** and QKI **(D)** versus the number of distinct crosslinking sites within a cluster indicated by a T to C change. The fraction of sites containing at least one recognition motif rises with the number of crosslinking sites.

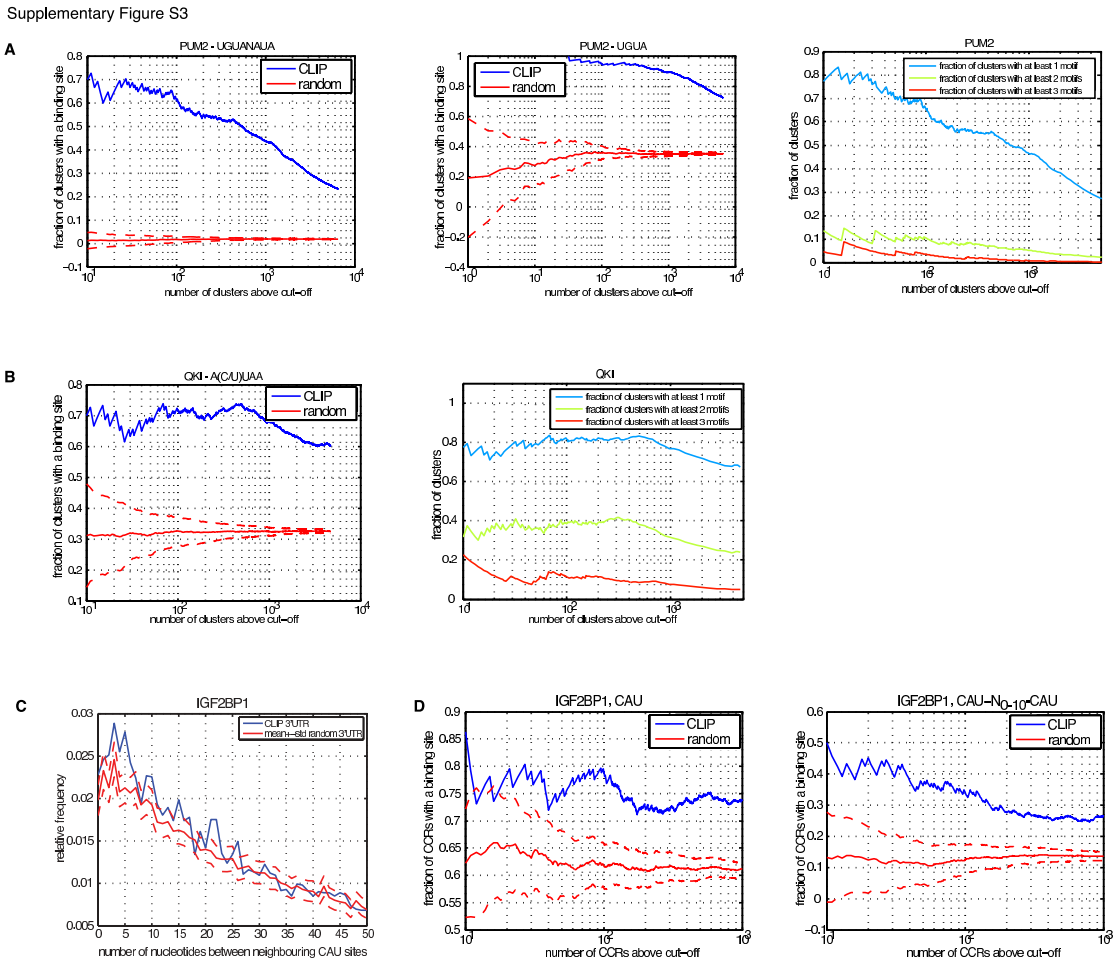


Figure 7.10: **(A)** Enrichment of binding motifs for PUM2 for the consensus motif UGUANAUA as well as the short variant UGUA compared to CCRs with randomized sequences (left and middle panels). The rightmost panel shows the fraction of clusters with at least one, two or three UGUANAUA motifs. Most clusters contain only one binding site. **(B)** Enrichment of the A(C/U)UAA binding motif in CCRs of QKI (left). The right panel shows the fraction of clusters with at least one, two or three motifs. A significant fraction of clusters contains two or more binding sites. **(C)** Distance between two neighboring CAU-motifs in crosslinked IGF2BP1 PAR-CLIP clusters (blue line) and in randomized transcripts (red line). CAU-motifs are enriched within 3-5 nt distance of each other in the crosslinked regions compared to randomized sequence sets. Only IGF2BP1 is shown because IGF2BP2 and 3 show the same results. **(D)** Enrichment of the CAU or CAU-N(0-10)-CAU binding motif for IGF2BP1 over randomized sequence sets of the same nucleotide composition. Equivalent analyses for IGF2BP2 and IGF2BP3 yield similar results (data not shown).

Supplementary Figure S4

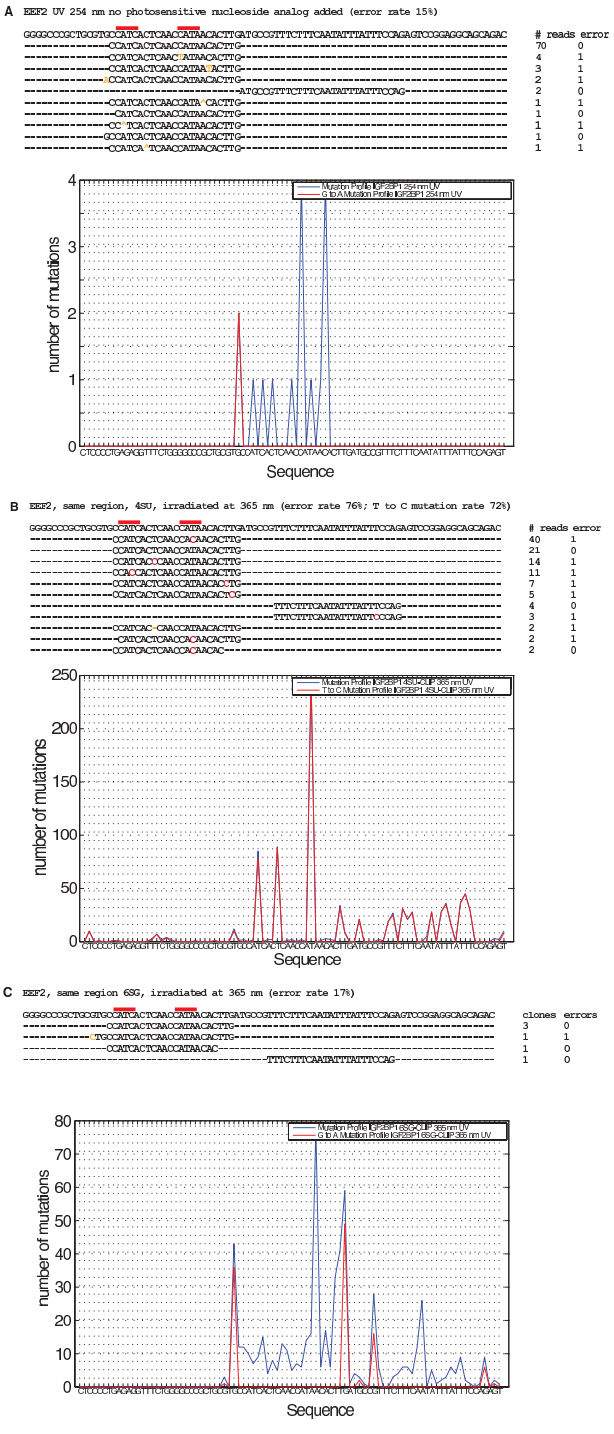


Figure 7.11 *(facing page)*: Alignment of sequences from CLIP experiments with IGF2BP1 against nucleotides 2784-2868 of the human EEF2 transcript (NM_001961). Nucleotides marked in red show the T to C changes, all other mismatches are marked in orange. Due to space limitations, not all reads that were sequenced are shown. **(A)** Alignment of sequences obtained from UV crosslinking at 254 nm. Lower panel: Profile for G to A mutations (red) and for any mutation (blue). **(B)** Alignment of sequences obtained after incorporation of 4SU into the transcript and crosslinking at 365 nm. Lower panel: mutational profile for T to C mutations (red) and for any mutation (blue). **(C)** Alignment of sequences obtained after incorporation of 6SG into the transcript and crosslinking at 365 nm. Lower panel: as in (A).

Supplementary Figure S5

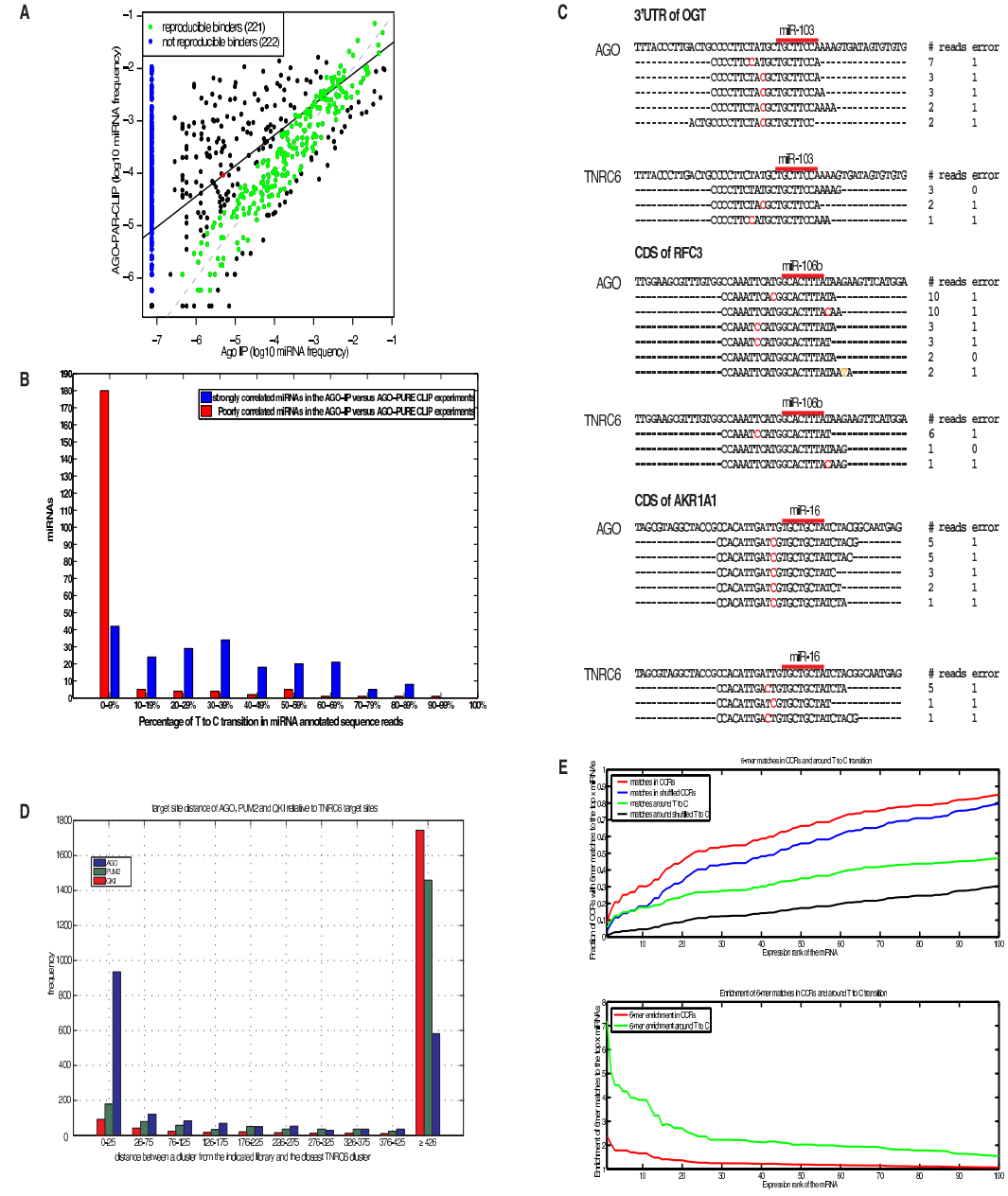
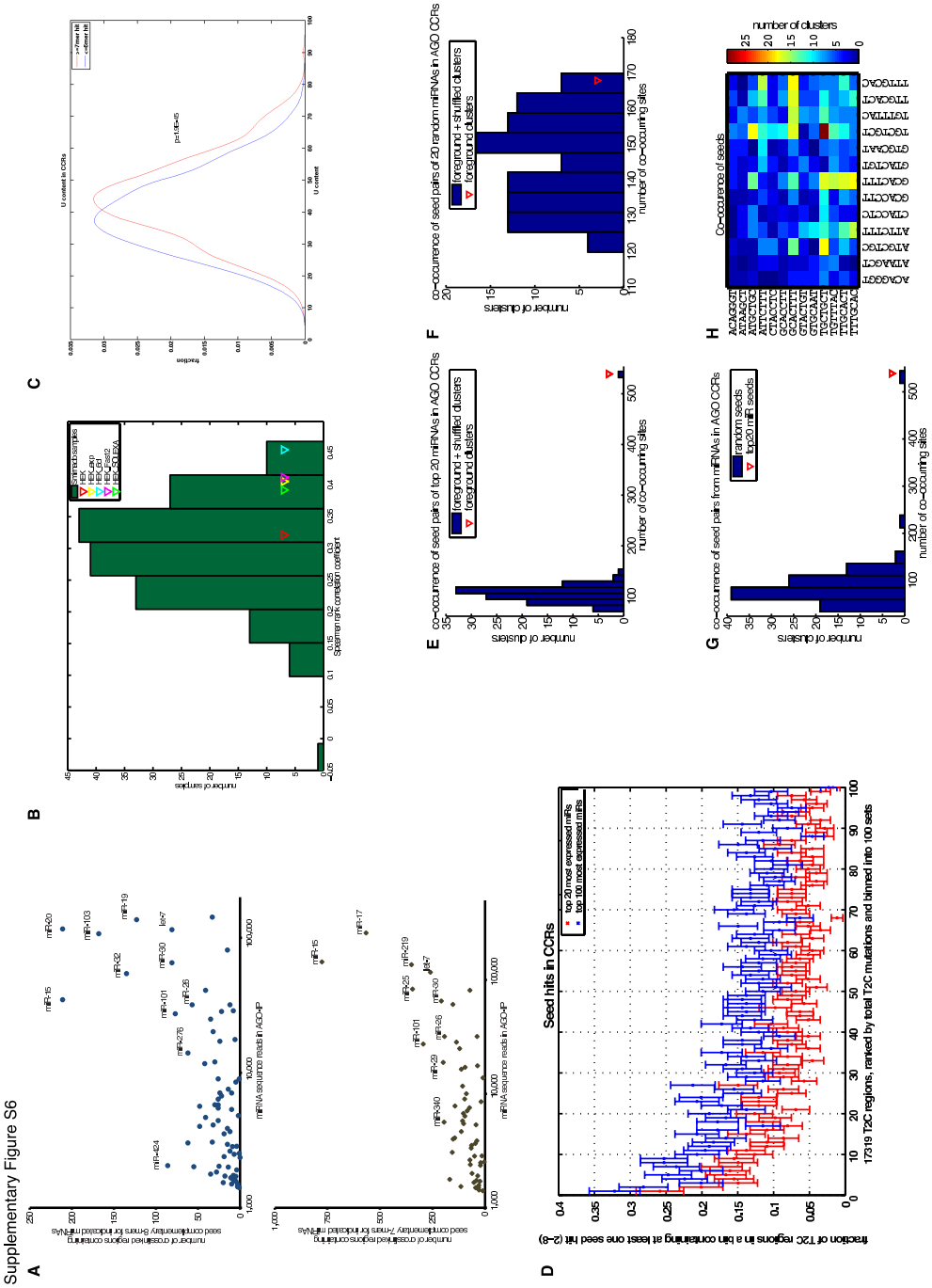


Figure 7.12 *(facing page)*: **CCRs from the AGO-PAR-CLIP contain miRNA seed complementary sites. (A)** Principal component analysis of the relative abundance of miRNAs derived from the combination of the AGO-PAR-CLIP libraries on one hand, and the non-crosslinked AGO-IPs on the other hand. The first principal component is projected onto the plane of $\log_{10}$-frequency in Ago-IP vs. $\log_{10}$-frequency in CLIP. The slope of the principal component was 0.58. Although for many miRNAs the expression levels measured by the two methods are quite comparable, there is a subset of miRNAs whose expression in the AGO-IP is systematically lower than the expression estimated based on the AGO-PAR-CLIP data (shown in blue). **(B)** The miRNAs that correlate well between the AGO-IP and the AGO-PAR-CLIP data (panel A: difference in $\log_{10}$ frequencies in Ago CLIP vs Ago IP smaller than 0.6, shown in green) are miRNAs with high frequency of T to C mutations in the AGO-PAR-CLIP, whereas miRNAs that were sequenced at least once in the Ago CLIP but were not detected in the Ago IP (blue) have a low frequency of T to C mutations. **(C)-(E)** AGO and TNRC6 proteins bind to the same regions on the target transcripts. **(C)** Alignments of AGO PAR-CLIP and TNRC6 PAR-CLIP cDNA sequence reads to regions in the 3'UTRs of OGT (NM_181672), the CDS of RFC3 (NM_002915) and the CDS of AKR1A1 (NM_006066). Red bars indicate 8 nt seed complementary sequences and nucleotides marked in red indicate T to C mutations diagnostic of the crosslinking position. **(D)** The distance between TNRC6 target sites and the nearest binding sites of QKI, PUM2, AGO have been computed. The histogram shows the number of TNRC6 target sites within a given nucleotide distance from the binding site of another RNA binding protein. Approximately 950 (i.e. ca. 50%) of the CCRs from the TNRC6 PAR-CLIP experiment fall within 25 nt of a CCR from the AGO-PAR-CLIP. **(E)** 6-mer enrichment in the full CCRs and the region ranging from 2 nt upstream to 10 nt downstream of the predominant crosslinking site. The upper panel shows the fraction of CCRs having a 6-mer hit for the top 100 expressed miRNAs. The background set consists of dinucleotide shuffled versions of either the full CCRs or the region around the crosslinking site. The lower panel shows the enrichment of 6-mers relative to the background set in the region indicated in previous panel (full CCRs, and 13 nt around the predominant crosslinking site)



Supplementary Figure S6

Figure 7.13 *(facing page)*: **CCRs from the AGO-PAR-CLIP are enriched for target sites for the most abundant miRNAs in HEK293 cells (A)** Correlation between occurrence of 8-mer (upper panel) and 7-mer (lower panel) seed matches in the CCRs and the abundance of the corresponding miRNA seed families. **(B)** Spearman correlation between the number of 7-mer (2-8) seed matches in the CCRs from AGO-PAR-CLIP and the experimentally determined counts of corresponding miRNA seeds in various miRNA samples from the smiRNAdb database *http://www.mirz.unibas.ch/smirnadb* and the HEK293 RNA analyzed in this study. Triangles indicate different HEK293 miRNA libraries. **(C)** Comparison of the U content of CCRs with at least a 7-mer seed match to the top 100 most abundant miRNAs versus CCRs with at most a 6-mer seed match to the top 100 most abundant miRNAs. The mean of the distributions was significantly different (ranksum test, p = 1.9E-45). **(D)** The number of crosslinking events correlates with the enrichment of the CCRs in the putative binding sites for the most abundantly expressed miRNAs. The frequency of the most strongly enriched miRNA seed motif (complementary to positions 2-8 of the miRNAs) was determined in the 17,319 AGO CCRs, which were sorted by the number of U-to-C changes and grouped into bins of 100. The frequency of miRNA seed-complementary motifs in the CCRs decreases with the number of U-to-C mutations in the clusters corresponding to these CCRs. **(E)** Number of pairs of non-overlapping seed (pos. 2-8) matches for the 20 most abundantly expressed miRNAs in HEK 293 cells in the crosslinked regions (red triangle) and in control regions (100 sets of dinucleotide shuffled crosslinked regions). Only the experimental set shows enrichment of miRNA pairs. **(F)** Number of co-occurring pairs of miRNA seed matches in the AGO crosslinked regions and the shuffled control regions for 20 randomly chosen miRNAs. **(G)** Number of co-occurring pairs of miRNA seed matches in the AGO crosslinked regions for 100 sets of 20 randomly chosen miRNAs. **(H)** Heat map representation of miRNA seed match co-occurrence. Only miRNA seed matches were counted that did not overlap and could therefore be bound simultaneously by two AGO-proteins. The scale indicates the absolute number of co-occurring pairs. Matches to the seed of miR-17 co-occur with matches to the seed of miR-19/miR-130/miR-301/miR-30/miR-15/miR-16. miR-16 seed matches have the tendency to co-occur with themselves.

Supplementary Figure S7

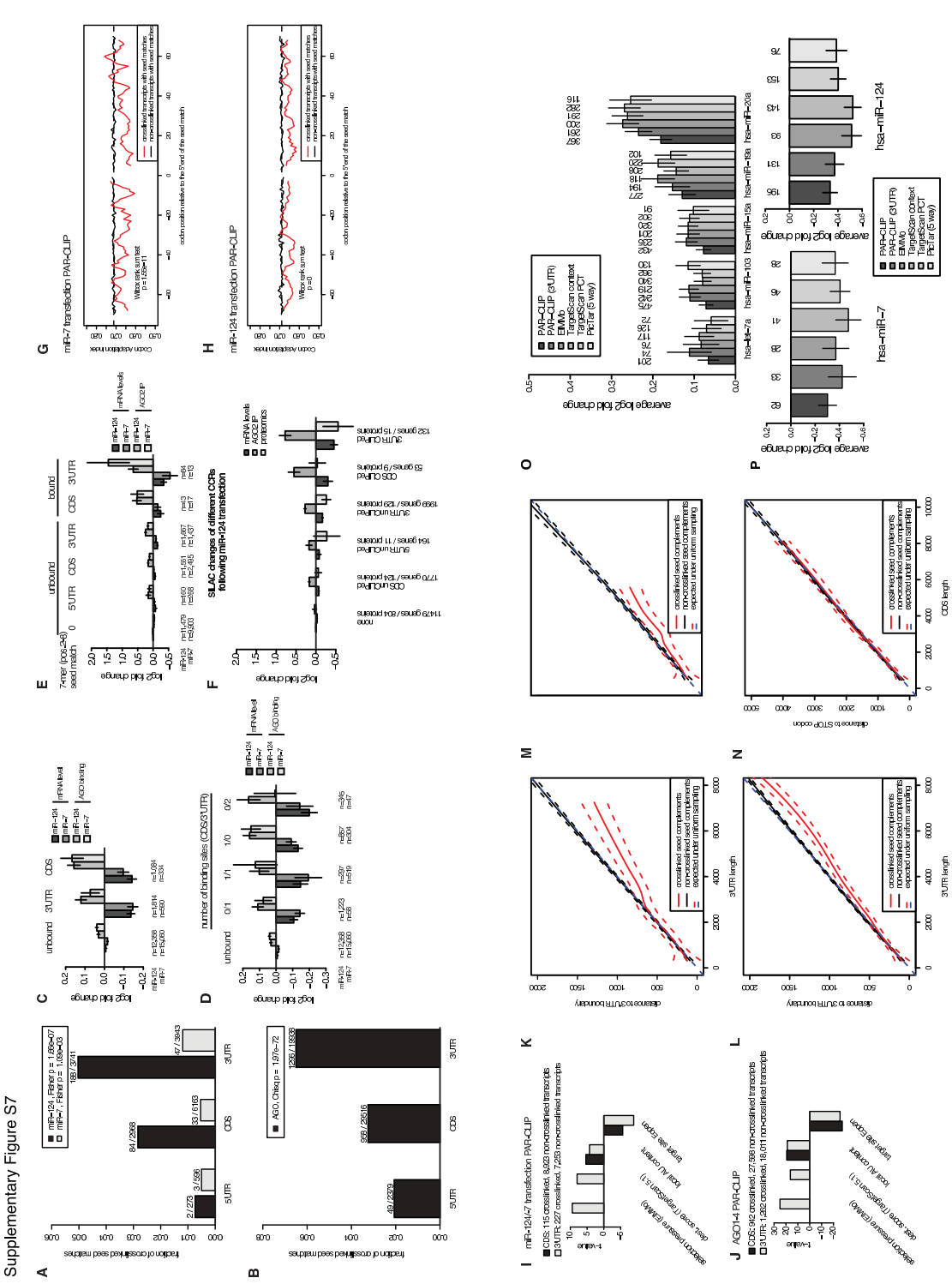


Figure 7.14 *(facing page)*: **Properties of CCRs containing miRNA seed complementary sites**. **(A)** Seed complementary sequences in the 3'UTR are more efficiently crosslinked than seed complementary regions in the CDS. Fraction of crosslinked seed matches (1-7 or 2-8) for the miR-124 (dark bars) and miR-7 (light bars) transfection experiments are shown; and in **(B)** the fraction of crosslinked seed matches for miR-15, miR-16, miR-19, and let-7 in the ALL_AGO dataset is shown. **(C)** Properties of AGO-PAR-CLIP sequence read clusters obtained after miR-124 and miR-7 transfection. Transcripts with PAR-CLIP sequence read clusters identified after miR-124 and miR-7 transfection (n indicates number of transcripts considered) are bound by AGO2 and destabilized. Transcript stability (dark grey bars) was determined as in Figure 3 by comparison of mRNA-abundance of mock-transfected and miR-124 and miR-7-transfected HEK293 cells. miR-7 and miR-124 mediated AGO2 binding (light grey bars) was determined by comparing transcripts enriched by AGO2-IPs of mock transfected and miR-124 and miR-7 transfected HEK293 cells [134]. Transcripts containing PAR-CLIP sequence read clusters were categorized according to the transcript region bound by AGO2 (CDS/3'UTR). **(D)** Same as in (C). Transcripts were categorized in more detail according to the number and region (CDS/3'UTR) of sequence read clusters identified. **(E)** Same as in (C). Transcripts containing a miR-124 and miR-7 seed complementary sequence but without PAR-CLIP sequence read clusters (unbound) were compared to transcripts with PAR-CLIP sequence read clusters with miR-124 and miR-7 seed complementary sequences (bound). The unbound and bound transcripts are categorized according to regions within the transcript (5' UTR, CDS, and 3'UTR). **(F)** In addition to the AGO2 binding and mRNA destabilization following miR-124 transfection shown in (G) for PAR-CLIP identified transcripts, changes in protein level following miR-124 transfection (as measured by SILAC in Hela cells [143]) are indicated. **(G-H)** Codon adaptation index (CAI) for regions upstream and downstream of CCRs (relative to 5' end of the seed match) found in the CDS for the **(G)** miR-7 and **(H)** miR-124 transfection experiments. The red and the black lines indicate the CAI for crosslinked and non-crosslinked transcripts, respectively. **(I)** The sequence context defines a functional miRNA binding site in the UTR as well as in the CDS. Four different criteria (selection pressure, destabilization score, local A/U content, target site openness) were compared for crosslinked transcripts containing 7-mer seed matches for a miR-124 and miR-7 and **(J)** the miR-15, miR-19, miR-20, and let-7 miRNA families in the AGO PAR-CLIP experiments compared to non-crosslinked transcripts containing the same 7-mer seed matches. **(K)** In 3'UTRs longer than 3,000 nts the crosslinked sites distribute preferentially near to the boundaries of the UTR. Distance from the region boundaries (stop codon and polyA signal, respectively) of CCRs with 7-mer seed complement regions falling in the 3'UTR to miR-124 and miR-7 in the transfection experiments (red line) and **(L)** 7-mer seed matches to the miR-15, miR-16, miR-19 and let-7 seed families from the AGO PAR-CLIP (red line) compared to non-crosslinked seed-matches (black lines). **(M)** Distance from the stop codon of CCRs falling in the CDS containing 7-mer seed matches of miR-124 and miR-7 (red line) or **(N)** 7-mer seed matches of the miR-15, miR-16, miR-19 and let-7 seed families (red line) compared to non-crosslinked seed-matches (black lines). Only for the miR-124 and miR-7 transfection experiments the crosslinked sites in the CDS distribute significantly closer to the stop-codon. **(O)** Comparison of PAR-CLIP with ElMMo, TargetScan S, TargetScan Pct, and PicTar miRNA target predictions. We determined the number of seed matches in the top 1000 CCRs for each of the indicated miRNAs. For each miRNA we selected the indicated number of sites (on mRNAs found by DGE and having a signal intensity above the median on the Affymetrix mRNA microarrays), starting from those with the best score, as given by the indicated prediction method. The figure shows average log2 fold changes of mRNA targets identified by the different methods upon miRNA inhibition (of miRNAs let-7a, miR-103, miR-15a, miR-19a, miR-20). **(P)** Average log2 fold changes of mRNA targets identified by various methods upon miR-7 and miR-124 transfection.

	0 nM 4SU	0 nM 4SU	100 nM 4SU	1000 nM 4SU
0 nM 4SU	1	0.94	0.96	0.95
0 nM 4SU	0.94	1	0.96	0.94
100 nM 4SU	0.96	0.96	1	0.99
1000 nM 4SU	0.95	0.94	0.99	1

Table 7.1: Toxicity of photoreactive 4SU nucleotides (correlation of mRNA abundance).

	0 nM 6SG	100 nM 6SG	1000 nM 6SG
0 nM 6SG	1	0.98	0.97
100 nM 6SG	0.98	1	1
1000 nM 6SG	0.97	1	1

Table 7.2: Toxicity of photoreactive 6SG nucleotides (correlation of mRNA abundance).

	IGF2BP1	IGF2BP2	IGF2BP3	QKI	PUM2
IGF2BP1	1	0.78	0.74	0.14	0.24
IGF2BP2	0.78	1	0.81	0.15	0.28
IGF2BP3	0.74	0.81	1	0.19	0.34
QKI	0.14	0.15	0.19	1	-0.31
PUM2	0.24	0.28	0.34	-0.31	1

Table 7.3: Correlation of sequence reads per transcript for the generated cDNA libraries.

	IGF2BP1 (UV254)	IGF2BP1 (4SU)	IGF2BP1 (6SG)
IGF2BP1 (UV254)	1	0.11	0.11
IGF2BP1 (4SU)	0.11	1	0.65
IGF2BP1 (6SG)	0.11	0.65	1

Table 7.4: Correlation of sequence reads per transcript comparing different crosslinking methods.

Protein	Modification	Sequence	Kd
QKI	unmod	GUAUGCCAUUAACAAAUUCAUUAACAA	93 nM
QKI	mutated1	GUAUGCCCACAUCAAAUUCAUUAACAA	264 nM
QKI	mutated2	GUAUGCCAUUAACAAAUUCCACAUCAA	386 nM
QKI	mutated1+2	GUAUGCCCACAUCAAAUUCCACAUCAA	871 nM
IGF2BP2	3'UTR C2orf43 (wt)	CATTGCCATACATTAACTCTCCATTTCTGCATTAACTTCATTT	7.6±3.5 nM
IGF2BP2	3'UTR C2orf43 (mutant 1)	CATTGCCATACAGGAACTCTCCAGGTCTGCAGGAACTTCATTT	15±2 nM
IGF2BP2	3'UTR C2orf43 (mutant 2)	CCTTGCCCTACCTTAACTCTCCCTTTCTGCCTTAACTTCCTTT	31±1 nM
IGF2BP2	3'UTR MRPL9 (wt)	TGTCTCCAGTACTTGCCTCATTCTCATCATCCAAACTGAA	29±2 nM
IGF2BP2	3'UTR MRPL9 (mutant 1)	TGTCTCCAGTACTTGCCTCAGGCTCAGCAGGCAAACTGAA	570±120 nM
IGF2BP2	3'UTR MRPL9 (mutant 2)	TGTCTCCAGTACTTGCCTCCTTCTCCTCCTCCCAACTGAA	76±13 nM
IGF2BP2	3'UTR MRP9 (wt)	CCTCATTCTCATCATCCAAACTG	n.d.
IGF2BP2	3'UTR C2orf43 (wt)	CCATACATTAACTCTCCATTTCTGCATTAACT	2100±500 nM

Table 7.5: Affinity of sequences identified by PAR-CLIP for QKI and IGF2BP2 (recombinant protein and synthetic RNA). Only modified sequence stretches are shown.

oligo	sequence	non-crosslinked, % mutated	crosslinked, % mutated
4SU9	GUAUGCCAUUAACAAAUUCAUUAACAAGUCCGUUCG	8.1	49.4
4SU10	GUAUGCCAUUAACAAAUUCAUUAACAAGUCCGUUCG	25.8	47.3
4SU2	GUAUGCCAUUAACAAAUUCAUUAACAAGUCCGUUCG	17.3	78.8
4SU4	GUAUGCCAUUAACAAAUUCAUUAACAAGUCCGUUCG	8.8	82.7

Table 7.6: In vitro PAR-CLIP experiment with synthetic oligoribonucleotides (shown in red) and recombinant QKI.

Chapter 8

Towards a recognition code of KH domain-containing RNA-binding proteins

Taken together with the work on Nova [94, 97], our analysis of the PAR-CLIP data of IGF2BP1-3 suggests a general mechanism by which KH domain-containing RNA-binding proteins (RBPs) achieve the specificity of interaction with their target RNAs. All four proteins contain several KH domains, most of which presumably recognize short 3-4 nucleotide long sequence stretches and for all proteins, the inter-motif spacings are not fixed, but appear to be constrained to a certain interval of preferred distances ([94] and figure 7.10C). Considering that many KH-domain containing RBPs harbour not only one, but several KH domains [87], and that they may form homo- and heterodimers [154, 155], these results suggest that, generally, sequences elements specifically bound by KH domain-containing RBPs may consist of short recognition motifs separated by spacers of variable length.

Most current state-of-the-art motif finding algorithms are not flexible enough to model variable spacer lengths, but typically search for one or many motifs that are not spatially constrained with respect to each other (see e.g. [3]). Additionally, they have been developed to discover transcription factor binding sites, which are typically much longer (8-12 nucleotides in eukaryotes and even longer in bacteria) than the sites of many RNA-binding domains, and it is well known that motif finding becomes very difficult when the motif is as short as 4 nucleotides (as for typical KH domains). A recently published motif finding tool that models insertions and deletions within sequence motifs and can thus deal with spacers of variable length is *Glam2* [156]. However, *Glam2* models the number of inserted positions between subsequent matches to a weight matrix with a geometric distribution, which is not in accordance with our observation for Nova and the IGF2BPs. Additionally, the model is not general enough for the description of configurations where certain motifs

re-occur several times (due to several RNA-binding domains having the same binding specificity or due to homodimerization) or possibly in a different order in the same input sequence (due to different modes of dimerization).

There is thus a need for motif finding algorithms that model, within a general framework, binding elements in terms of a variable number of motifs that are separated by spacers of varying lengths. Analogously to how the binding specificity is achieved by the combined binding of several RNA-binding domains, such an algorithm would score entire binding elements, consisting of a combination of motifs and spacer(s), which effectively lengthens the recognition sequence and should thus lead to a stronger enrichment signal. We have implemented a preliminary version of such an algorithm that uses Gibbs sampling to search for two (different or identical) motifs separated by a variable spacer whose minimal and maximal length can be preset. Preliminary results of an application of this algorithm to CLIP data of Nova [97] and the PAR-CLIP data of the IGF2BPs show that the method successfully infers both the binding motif and the spatial clustering of the motif for each protein.

We are planning to implement a general version of such a motif finder, based on a hidden Markov model, as illustrated in figure 8.1. In this model, there are three different types of states, a background state, motif states and spacer states. In the background state, nucleotides are emitted according to a background distribution that is independent of position, in the motif states, binding sites are emitted according to weight matrix models and in the spacer states, spacers of different lengths are emitted. The nucleotides of the spacer can either be modelled with the same distribution as in the background state or with their own distribution. The probability of the data for a particular configuration of binding sites and spacers is computed by integrating out the unknown parameters of the weight matrix models, the background distribution and the spacer distribution. The prior distribution over spacer lengths may either be specified by the user or integrated out, which would effectively favour polarized distributions. In the latter case, it may be useful to have separate spacer length distributions for each particular transition from one weight matrix to another one as this distribution reflects the spatial arrangement of subsequent domains in the RBP monomer or dimer. Finally, the space of configurations of our model is searched using a Gibbs sampling approach where at each time step the configuration of one input sequence is re-sampled. This can (at least approximately) be carried out using the forward-backward algorithms of hidden Markov model theory [3, 157].

In some applications, there is prior knowledge about the approximate location of the binding sites within the set of input sequences. For example, in PAR-CLIP data, there is an enrichment of T to C mutations in the immediate vicinity of binding sites (chapter 7) and a mutation profile along the input sequence thus already contains information about the possible location of binding sites. Prior knowledge may also come from previous insights into the binding preferences of the RBP. For example, for some RNA-binding proteins that bind single-stranded RNA it is known that binding

sites tend to lie in open structures such as stem loops [93,158]. Position-specific prior distributions, describing the probability for a binding site to start in a particular position based on properties such as mutation biases or accessibility [159], can also be easily incorporated into the proposed hidden Markov model. As such, the model can be optimally adjusted to the type of data and the properties of the RBP.

To our knowledge, the proposed motif finder would be the first motif finding tool designed specifically for the inference of binding sites of RBPs and the first algorithm to model, within a general framework, binding elements in terms of a variable number of motifs separated by spacers of varying lengths. We are planning to apply the algorithm to both the PAR-CLIP data of IGF2BP1-3, available CLIP data on Nova [97, 98] as well as RIP-Chip data [160]. In this way, we hope to gain insight into the recognition 'code' of KH domain-containing RBPs.

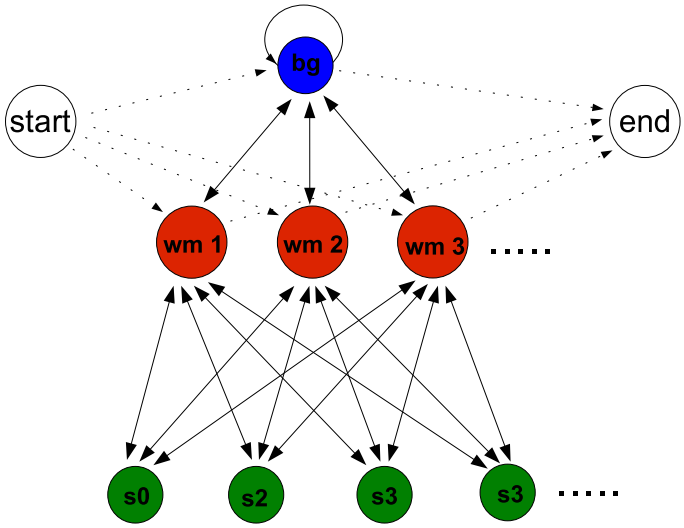


Figure 8.1: Illustration of the hidden Markov model for the inference of binding sites of RBPs. The blue circle corresponds to the background state (bg), red circles to motif states ($\mathrm{wm}_1 - \mathrm{wm}_3$), green circles correspond to spacer states ($s_0 - s_3$, where the index refers to the length of the spacer) and the white circles correspond to the start and end states of the model. The number of motif and spacer states is arbitrary and can be preset by the user, but for the sake of simplicity, we only show three motif states and four spacer states. An input sequence is modelled as a path through the hidden Markov model from the start state to the end state. In the background state (bg), nucleotides are emitted based on a zero or first order Markov model, in the motif states, binding sites are emitted according to position-specific weight matrices (wm) and in the spacer states, spacers of different lengths are emitted. The nucleotides of the spacer can be either modelled with the same model as in the background state or with a different zero or first order Markov model.

Acknowledgments

There are of course many people that helped and supported me along this stony, but rewarding road.

First of all, I would like to thank my supervisor Prof E. van Nimwegen and my co-supervisor Prof M. Zavolan for everything they have taught me, for all the critical and helpful discussions and for all the mathematical and biological insights. Without their dedication and support, none of this would have been possible.

Secondly, I would like to thank my friends Nacho, Ionas, Dimos, Jean, Mohsen and Phil for great discussions and late-night beers and, in particular, Mel for her continuous support during the final leg of the race.

And finally, and most importantly, I would like to thank my parents for all their patience, support and encouragement.

Bibliography

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

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

[3] E. van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, 8(Suppl 6):S4, 2007.

[4] S.R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1994.

[5] S. Lindgreen, P.P. Gardner, and A. Krogh. Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics*, 22(24):2988–2995, 2006.

[6] K.R. Wollenberg and W.R. Atchley. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *PNAS*, 97:3288–3291, 2000.

[7] F. Pazos and A. Valencia. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins: Structure, Function, and Genetics*, 47:219–227, 2002.

[8] E.R. Tillier and T.W. Liu. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, 19(6):750–755, 2003.

[9] A.A. Fodor and R.W. Aldrich. Influence of conservation on calculations of amino acids covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56:211–221, 2004.

[10] L.C. Martin, G.B. Gloor, S.D. Dunn, and L.M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22):4116–4124, 2005.

[11] M.A. Fares and S.A. Travers. A novel method for detecting intramolecular coevolution: Adding a further dimension to selective constraints analyses. *Genetics*, 173:9–23, 2006.

[12] R. Gouveia-Oliveira and A.G. Pedersen. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms for Molecular Biology*, 2:12, 2007.

[13] S.D. Dunn, L.M. Wahl, and G.B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 2008.

[14] C.H. Yeang and D. Haussler. Detecting coevolution in and among protein domains. *PLoS Computational Biology*, 3:e211, 2007.

[15] M. Thattai, Y. Burak, and B. I. Shraiman. The origins of specificity in polyketide synthase protein interactions. *PLoS Comp Biol*, 3(9):e186, 2007.

[16] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *PNAS*, 106:67–72, 2009.

[17] S. Maisnier-Patin and D.I. Andersson. Adaptation to the deleterious effect of antimicrobial drug resistance mutations by compensatory evolution. *Research in Microbiology*, 155:360–369, 2004.

[18] W.M. Fitch and E. Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.*, 4:579–593, 1970.

[19] C. Yanovsky, V. Horn, and D. Thorpe. Protein structure relationships revealed by mutational analysis. *Science*, 146:1593–1594, 1964.

[20] C.H. Yeang, J.F. Darot, H.F. Noller, and D. Haussler. Detecting the coevolution of biosequences-an example of RNA interaction prediction. *Biochemistry*, 44:7156–7165, 2005.

[21] S.W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 8(286):295–299, 1999.

[22] G.M. Süel, S.W. Lockless, M.A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1):59–69, 2003.

[23] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L.L. Sonnhammer, D.J. Studholme, C. Yeats, and S.R. Eddy. The Pfam protein families database. *Nucl. Acids Res.*, 32:D138–D141, 2004.

[24] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

[25] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley and Sons, 1991.

[26] W.R. Atchley, W. Terhalle, and A.W. Dress. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. *J.Mol.Evol.*, 48:501–516, 1999.

[27] J.B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956.

[28] M. Meilà and T. Jaakkola. Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, 16(1):77–92, 2006.

[29] S.R. Eddy. Profile hidden markov models. *Bioinformatics*, 14:755–763, 1998.

[30] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daughert, L. Duquenne, R. D. Finn, J. Goug, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natal, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. Interpro: the integrative protein signature database. *Nucleic Acids Res.*, 35:D224–228, 2009.

[31] F Pazos and A Valencia. Protein co-evolution, co-adaptation and interactions. *The EMBO Journal*, 27:2648–2655, 2008.

[32] D.K. Chiu and T. Kolodziejczak. Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosc.*, 7:347–52, 1991.

[33] G. Shackelford and K. Karplus. Contact prediction using mutual information and neural nets. *Proteins*, 69(Suppl 8):159–164, 2007.

[34] J.M. Izarzugaza, O. Graña, M.L. Tress, A. Valencia, and N.D. Clarke. Assessment of intramolecular contact predictions for CASP7. *Proteins*, 69(Suppl 8):152–158, 2007.

[35] G.B. Gloor, L.C. Martin, L.M. Wahl, and S.D. Dunn. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44:7156–7165, 2005.

[36] AA Fodor and RW Aldrich. On evolutionary conservation of thermodynamic coupling in proteins. *Journal of Biological Chemistry*, 279(18):19046–19050, 2004.

[37] L. Burger and E. van Nimwegen. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Molecular Systems Biology*, 4:165, 2008.

[38] O. Olmean, B. Rost, and A. Valencia. Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.*, 295:1221–1239, 1999.

[39] D.D. Pollock, W.R. Taylor, and N. Goldman. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.*, 287:187–198, 1999.

[40] T Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *PNAS*, 99(22):14116–14121, 2002.

[41] B Rost and C Sander. Conservation and prediction of solvent accessibility in protein families. *Proteins*, 20:216–226, 1994.

[42] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[43] C.S. Miller and D. Eisenberg. Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, 24(14), 2008.

[44] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, 8:113, 2007.

[45] R.D. Finn, M. Marshall, and A. Bateman. iPfam: Visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 21:410–412, 2005.

[46] J. Cerquides and R. López de Màntaras. Tractable Bayesian learning of tree augmented naive Bayes classifiers. *Proceedings of Twentieth International conference on Machine Learning*, 2003.

[47] B. Bollobás. *Modern Graph Theory*. Springer, Berlin, corr. 2nd printing edition, 1998.

[48] A. Stock, V. Robinson, and P. Goudreau. Two-component signal transduction. *Annu.Rev.Biochem.*, 69:183–215, 2000.

[49] T.W. Grebe and J.B. Stock. The histidine protein kinase superfamily. *Advances in Microbial Physiology*, 41:139–227, 1999.

[50] C. Fabret, V.A. Feher, and J.A. Hoch. Two-component signal transduction in *Bacillus subtilis*: How one organism sees its world. *Journal of Bacteriology*, 181(7):1975–1983, 1999.

[51] Z.-L. Tzeng and J.A. Hoch. Molecular recognition in signal transduction: the interaction surfaces of the SpoOF response regulator with its cognate phosphorelay proteins revealed by alanine scanning mutagenesis. *J.Mol.Biol.*, 272:200–212, 1997.

[52] N. Ausmees and C. Jacobs-Wagner. Spatial and temporal control of differentiation and cell cycle progression in *Caulobacter Crescentus*. *Annu.Rev.Microbiol.*, 57:225–247, 2003.

[53] L. Li, E.I. Shakhnovich, and L.A. Mirny. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proc Natl Acad Sci*, 100(8):4463–4468, 2003.

[54] ftp.ncbi.nlm.nih.gov/genomes/bacteria.

[55] http://hmmer.wustl.edu/.

[56] C.B. Do, M.S.P. Mahabhashyam, M. Brudno, and S. Batzoglou. Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15:330–340, 2005.

[57] E. van Nimwegen, M. Zavolan, N. Rajewsky, and E. D. Siggia. Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics. *Proc. Natl. Acad. Sci. USA*, 99:7323–7328, 2002.

[58] J.M. Skerker and M.T. Laub. Cell-cycle progression and the generation of asymmetry in *Caulobacter Crescentus*. *Nature Reviews Microbiology*, 3:325–337, 2004.

[59] N. Ohta and A. Newton. The core dimerization domains of histidine kinases contain specificity for the cognate response regulator. *Journal of Bacteriology*, 185(15):4424–4431, 2003.

[60] J.M. Skerker, M.S. Prasol, B.S. Perchuk, E.G. Biondi, and M.T. Laub. Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a systems-level analysis. *PLoS Biol*, 3(10):e334, 2005.

[61] P. Bork, L.J. Jensen, C. von Mering, A.K. Ramani, I. Lee, and E.M. Marcotte. Protein interaction networks from yeast to human. *Curr Opin Struct Biol.*, 14:292–299, 2004.

[62] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol.*, 12(3):368–373, 2002.

[63] B.A. Shoemaker and A.R. Panchenko. Deciphering protein-protein interactions. Part ii. Computational methods to predict protein and domain interaction partners. *PLoS*, 3(4):e43, 2007.

[64] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

[65] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK, 2003.

[66] J. Zapf, U. Sen, M. Madhusudan, J.A. Hoch, and K.I. Varughese. A transient interaction between two phosphorelay proteins trapped in a crystal lattice reveals the mechanism of molecular recognition and phosphotransfer in signal transduction. *Structure Fold.Des.*, 8:851–862, 2000.

[67] E. Biondi, S. Reisinger, J. Skerker, M. Arif, B. Perchuk, K. Ryan, and M. Laub. Regulation of the bacterial cell cycle by an integrated genetic circuit. *Nature*, 444(7121):899–904, 2006.

[68] J. Wu, J-L.Z. Ohta, and A. Newton. A novel bacterial tyrosine kinase essential for cell division and differentiation. *Pro Natl Acad Sci*, 96:13068–13073, 1999.

[69] K. Weissman and P. Leadlay. Combinatorial biosynthesis of reduced polyketides. *Nat Rev Microbiol*, 3:925–936, 2005.

[70] E. van Nimwegen. Scaling laws in the functional content of genomes. *Trends in Genet.*, 19(9):479–484, 2003.

[71] E. Alm, K. Huang, and A. Arkin. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comp Biol*, 2(11):e143, 2006.

[72] C. F. Higgins. ABC transporters: from microorganisms to man. *Annu Rev Cell Biol.*, 8:67–113, 1992.

[73] R.S. Kaczmarski and G. J. Mufti. The cytokine receptor superfamily. *Blood Rev*, 5(3):193–203, 1991.

[74] K. Koretke, A. Lupas, P. Warren, M. Rosenberg, and J. Brown. Evolution of two-component signal transduction. *Molecular Biology and Evolution*, 17:1956–1968, 2000.

[75] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278:631–637, 1997.

[76] D. Beier and R. Frank. Molecular characterization of two-component systems of *Helicobacter pylory*. *Journal of Bacteriology*, 182(8):2068–2076, 2000.

[77] P.J. Piggot and D.W. Hilbert. Sporulation of *Bacillus subtilis*. *Current Opinion in Microbiology*, 7:579–586, 2004.

[78] Y. Kumagai, Z. Cheng, M. Lin, and Y. Rikihisa. Biochemical activities of three pairs of *Ehrlichia chaffeensis* two-component regulatory system proteins involved in inhibition of lysosomal fusion. *Infect Immun*, 74:5014–5022, 2006.

[79] W. Rostène, P. Kitabgi, and S.M. Parsadaniantz. Chemokines: a new class of neuromodulator? *Nat Rev Neurosci*, 8(11):895–903, 2007.

[80] N. Itoh and D.M. Ornitz. Evolution of the Fgf and Fgfr gene families. *Trends Genet*, 20(11):563–569, 2004.

[81] D. Juan, F. Pazos, and A. Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *PNAS*, 105(3):934–939, 2008.

[82] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.

[83] R.E. Halbeisen, T.S. Galgano, T. Scherrer, and A.P. Gerber. Post-transcriptional gene regulation: From genome-wide studies to principles. *Cell. Mol. Life Sci.*, 65:798–813, 2008.

[84] J.D. Keene. RNA regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, 8:533–543, 2007.

[85] J.D. Keene and S.A. Tenenbaum. Eukaryotic mRNPs may represent posttranscriptional operons. *Molecular Cell*, 9:1161–1167, 2002.

[86] A.P. Gerber, D. Herschlag, and P.O. Brown. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS*, 2(3):342–354, 2004.

[87] B.M. Lunde, C. Moore, and G. Varani. RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology*, 8:479–490, 2007.

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

[89] A. Cléry, M. Blatter, and F.H. Allain. RNA recognition motifs: boring? not quite. *Curr Opin Struct Biol*, 18(3):290–298, 2008.

[90] R. Valverde, L. Edwards, and L. Regan. Structure and function of KH domains. *FEBS*, 275:2712–2726, 2008.

[91] X. Wang, P.D. Zamore, and T.M. Hall. Crystal structure of a Pumilio homology domain. *Mol. Cell*, 7:855–865, 2001.

[92] X. Wang, J. MacLachlan, P.D. Zamore, and T.M. Hall. Modular recognition of RNA by a human pumilio-homology domain. *Cell*, 110:501–512, 2002.

[93] K. Musunuru and R.B. Darnell. Determination and augmentation of RNA sequence specificity of the Nova K-homology domains. *Nucleic Acids Research*, 32(16):4852–4861, 2004.

[94] J. Ule, G. Stefani, A. Mele, M. Ruggiu, X. Wang, B. Taneri, T. Gaasterland, B. Blencowe, and R.B. Darnell. An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444:580–586, 2006.

[95] S.A. Tenenbaum, P.J. Lager, C.C. Carson, and J.D. Keene. Ribonomics: Identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays. *Methods*, 26:191–198, 2002.

[96] J. Ule, K. Jensen, A. Mele, and R.B. Darnell. CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods*, 37:376–386, 2005.

[97] J. Ule, K.B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R.B. Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302:1212–1215, 2003.

[98] D.D. Licatalosi, A. Mele, J.J. Fak, J. Ule, M. Kayikci, S.W. Chi, T.A. Clark, A.C. Schweitzer, J.E. Blume, X. Wang, J.C. Darnell, and R.B. Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456:464–469, 2008.

[99] K.C. Martin and A. Ephrussi. mRNA localization: Gene expression in the spatial dimension. *Cell*, 136:719–730, 2009.

[100] M.J. Moore and N.J. Proudfoot. Pre-mRNA processing reaches back totranscription and ahead to translation. *Cell*, 136:688–700, 2009.

[101] N. Sonenberg and A.G. Hinnebusch. Regulation of translation initiation in eukaryotes: Mechanisms and biological targets. *Cell*, 136:731–745, 2009.

[102] A.E. McKee, E. Minet, C. Stern, S. Riahi, C.D. Stiles, and P.A. Silver. A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. *BMC Dev. Biol.*, 5:14, 2005.

[103] D.P. Bartel. MicroRNAs: Target recognition and regulatory functions. *Cell*, 136:215–233, 2009.

[104] S.A. Tenenbaum, C.C. Carson, P.J. Lager, and J.D. Keene. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci USA*, 97(26):14085–14090, 2000.

[105] I. Lopez de Silanes, M. Zhan, A. Lal, X. Yang, and M. Gorospe. Identification of a target RNA motif for RNA-binding protein hur. *PNAS*, 101:2987–2992, 2004.

[106] A.P. Gerber, S. Luschnig, M.A. Krasnow, P.O. Brown, and D. Herschlag. Genome-wide identification of mRNAs associated with the translational regulator pumilio in Drosophila melanogaster. *Proc Natl Acad Sci USA*, 103(12):4487–4492, 2006.

[107] A.J. Wagenmakers, R.J. Reinders, and W.J. van Venrooij. Cross-linking of mRNA to proteins by irradiation of intact cells with ultraviolet light. *Eur. J. Biochem.*, 112:323–330, 1980.

[108] S.A. Adam and G. Dreyfuss. Adenovirus proteins associated with mRNA and hnRNA in infected HeLa cells. *J. Virol.*, 61:3276–3283, 1987.

[109] G. Dreyfuss, Y.D. Choi, and S.A. Adam. Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. *Mol. Cell. Biol.*, 4:1104–1114, 1984.

[110] G.W. Yeo, N.G. Coufal, T.Y. Liang, G.E. Peng, X.D. Fu, and F.H. Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol*, 16(2):130–137, 2009.

[111] J.R. Sanford, X. Wang, M. Mort, N. Vanduyn, D.N. Cooper, S.D. Mooney, H.J. Edenberg, and Y. Liu. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res*, 2009.

[112] S. Granneman, G. Kudla, E. Petfalski, and D. Tollervey. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *PNAS*, 106(24):9613–8, 2009.

[113] S. Guil and J.F. Cáceres. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nat Struct Mol Biol*, 14(7):591–596, 2007.

[114] S.W. Chi, J.B. Zang, A. Mele, and R.B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460:479–486, 2009.

[115] Y Kirino and Z Mourelatos. Site-specific crosslinking of human microRNPs to RNA targets. *RNA*, 14:2254–2259, 2008.

[116] K.M. Meisenheimer and T.H. Koch. Photocross-linking of nucleic acids to associated proteins. *Crit. Rev. Biochem. Mol. Biol.*, 32:101–140, 2005.

[117] A. Favre, G. Moreno, M.O. Blondel, J. Kliber, F. Vinzens, and C. Salet. 4-thiouridine photosensitized RNA-protein crosslinking in mammalian cells. *Biochem Biophys Res Commun*, 141(2):847–854, 1986.

[118] M. Hafner, P. Landgraf, J. Ludwig, A. Rice, T. Ojo, C. Lin, D. Holoch, C. Lim, and T. Tuschl. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, 44(1):3–12, 2008.

[119] M. Wickens, D.S. Bernstein, J. Kimble, and R. Parker. A PUF family portrait: 3'UTR regulation as a way of life. *Trends Genet*, 18(3):150–157, 2002.

[120] R. Siddharthan, E.D. Siggia, and E. Nimwegen. Phylogibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, 2005.

[121] A. Galgano, M. Forrer, L. Jaskiewicz, A. Kanitz, M. Zavolan, and A.P. Gerber. Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. *PLoS ONE*, 3:e3164, 2008.

[122] A. Galarneau and S. Richard. Target RNA motif and target mRNAs of the Quaking STAR protein. *Nat Struct Mol Biol*, 12(8):691–698, 2005.

[123] S. Chénard, C.A. Richard. New implications for the Quaking RNA binding protein in human disease. *J Neurosci Res*, 86(2):233–242, Feb 2008.

[124] J.K. Yisraeli. VICKZ proteins: a multi-talented family of regulatory RNA-binding proteins. *Biol Cell*, 97(1):87–96, 2005.

[125] B. Boyerinas, S.M. Park, N. Shomron, M.M. Hedegaard, J. Vinther, J.S. Andersen, C. Feig, J. Xu, C.B. Burge, and M.E. Peter. Identification of let-7-regulated oncofetal genes. *Cancer Res*, 68(8):2587–2591, 2008.

[126] E. Dimitriadis, T. Trangas, S. Milatos, P.G. Foukas, I. Gioulbasanis, N. Courtis, F.C. Nielsen, N. Pandis, U. Dafni, G. Bardi, and P. Ioannidis. Expression of oncofetal RNA-binding protein CRD-BP/IMP1 predicts clinical outcome in colon cancer. *Int J Cancer*, 121(3):486–494, 2007.

[127] Diabetes Genetics Initiative of Broad Institute of Harvard, MIT, Lund University, Novartis Institutes of BioMedical Research, Richa Saxena, Benjamin F Voight, Valeriya Lyssenko, Noël P Burtt, Paul I W de Bakker, Hong Chen, Jeffrey J Roix, Sekar Kathiresan, Joel N Hirschhorn, Mark J Daly, Thomas E Hughes, Leif Groop, David Altshuler, Peter Almgren, Jose C Florez, Joanne Meyer, Kristin Ardlie, Kristina Bengtsson Boström, Bo Isomaa, Guillaume Lettre, Ulf Lindblad, Helen N Lyon, Olle Melander, Christopher Newton-Cheh, Peter Nilsson, Marju Orho-Melander, Lennart Råstam, Elizabeth K Speliotes, Marja-Riitta Taskinen, Tiinamaija Tuomi, Candace Guiducci, Anna Berglund, Joyce Carlson, Lauren Gianniny, Rachel Hackett, Liselotte Hall, Johan Holmkvist, Esa Laurila, Marketa Sjögren, Maria Sterner, Aarti Surti, Margareta Svensson, Malin Svensson, Ryan Tewhey, Brendan Blumenstiel, Melissa Parkin, Matthew Defelice, Rachel Barry, Wendy Brodeur, Jody Camarata, Nancy Chia, Mary Fava, John Gibbons, Bob Handsaker, Claire Healy, Kieu Nguyen, Casey Gates, Carrie Sougnez, Diane Gage, Marcia Nizzari, Stacey B Gabriel, Gung-Wei Chirn, Qicheng Ma, Hemang Parikh, Delwood Richardson, Darrell Ricke, and Shaun Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–1336, Jun 2007.

[128] N. Rajewsky. microRNA target predictions in animals. *Nat.Genet.*, 38:S8–S13, 2006.

[129] M. Landthaler, D. Gaidatzis, A. Rothballer, P.Y. Chen, S.J. Soll, L. Dinic, T. Ojo, M. Hafner, M. Zavolan, and T. Tuschl. Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA*, 14(12):2580–2596, 2008.

[130] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A.O. Kamphorst, and et al. Landthaler, M. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129:1401–1414, 2007.

[131] Y. Wang, S. Juranek, H. Li, G. Sheng, T. Tuschl, and D.J. Patel. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature*, 456:921–926, 2008.

[132] Y. Wang, S. Juranek, H. Li, G. Sheng, T. Tuschl, and D.J. Patel. Nucleation, propagation and cleavage of target RNAs in ago silencing complexes. *Nature*, 461:754–761, 2009.

[133] K.K. Grimson, A.and Farh, W.K. Johnston, P. Garrett-Engele, L.P. Lim, and D.P. Bartel. Microrna targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, 27:91–105, 2007.

[134] J. Hausser, M. Landthaler, L. Jaskiewicz, D. Gaidatzis, and M. Zavolan. Relative contribution of sequence and structure features to the mRNA binding of argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Res.*, 19(11):2009–2020, 2009.

[135] A.A. Easow, G. Teleman and S.M. Cohen. Isolation of microrna targets by mirnp immunopurification. *RNA*, 13:1198–1204, 2007.

[136] J.J. Forman, A. Legesse-Miller, and H.A. Coller. A search for conserved sequences in coding regions reveals that the let-7 microrna targets dicer within its coding sequencene photosensitized. *PNAS*, 105:14879–14884, 2008.

[137] J.R. Lytle, T.A. Yario, and J.A. Steitz. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3'. *PNAS*, 104:9667–9672, 2007.

[138] F.C. Orom, U.A.and Nielsen, , and A.H. Lund. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Mol. Cell*, 30:460–471, 2008.

[139] Y. Tay, J. Zhang, A.M. Thomson, B. Lim, and I. Rigoutsos. MicroRNAs to nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455:1124–1128, 2008.

[140] P.M. Sharp and W.H. Li. The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acids Res.*, 15:1281–1295, 1987.

[141] S. Gu, L. Jin, F. Zhang, P. Sarnow, , and M.A. Kay. Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. *Nat. Struct. Mol. Biol.*, 16:144–150, 2009.

[142] A. Stark, J. Brennecke, N. Bushati, R.B. Russell, and S.M. Cohen. Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123:1133–1146, 2005.

[143] D. Baek, J. Villen, C. Shin, F.D. Camargo, S.P. Gygi, and D.P. Bartel. The impact of microRNAs on protein output. *Nature*, 455:64–71, 2008.

[144] L.P. Lim, P. Lau, N.C.and Garrett-Engele, A. Grimson, J.M. Schelter, J. Castle, D.P. Bartel, P.S. Linsley, and J.M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433:769–773, 2005.

[145] M. Selbach, B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455:58–63, 2008.

[146] R.C. Friedman, K.K.-H. Farh, C.B. Burge, and D. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19:92–105, 2009.

[147] S. Lall, D. Grun, A. Krek, K. Chen, Y.-L. Wang, C.N. Dewey, P. Sood, T. Colombo, N. Bray, and et al. MacMenamin, P. A genome-wide map of conserved microRNA targets in C. elegans. *Curr.Biol.*, 16:460–471, 2006.

[148] M.C. Vella, E.Y. Choi, S.Y. Lin, K. Reinert, and F.J. Slack. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes Dev.*, 18:132–137, 2004.

[149] P. Berninger, D. Gaidatzis, E. van Nimwegen, and M. Zavolan. Computational analysis of small RNA cloning data. *Methods*, 44(1):13–21, 2008.

[150] T.D. Wu and C.K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, 2005.

[151] I.L. Hofacker. Vienna RNA secondary structure server. *Nucl. Acids Res.*, 31:3429–3431, 2003.

[152] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, and G. et al. Kreiman. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, 101:6062–6067, 2004.

[153] W.S. Cleveland, E. Grosse, and W.M. Shyu. *Statistical Models in S*. Wadsworth & Brooks/Cole, 1992.

[154] A. Ramos, D. Hollingworth, S.A. Major, S. Adinolfi, G. Kelly, FW Muskett, and A. Pastore. Role of dimerization in KH/RNA complexes: the example of Nova KH3. *Biochemistry*, 41(13):4193–4201, 2002.

[155] J. Nielsen, M.A. Kristensen, M. Willemoës, F.C. Nielsen, and J. Christiansen. Sequential dimerization of human zipcode-binding protein IMP1 on RNA: a cooperative mechanism providing RNP stability. *Nucleic Acids Res*, 32(14):4368–4376, 2004.

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

[157] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Cambridge University Press, 2001.

[158] T. Aviv, Z. Lin, G. Ben-Ari, C.A. Smibert, and F. Sicheri. Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. *Nat Struct Mol Biol*, 13(2):168–176, 2006.

[159] U. Mückstein, H. Tafer, J. Hackermüller, S. Bernhart, P.F. Stadler, and I.L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–1182, 2006.

[160] D.J. Hogan, D.P. Riordan, A.P. Gerber, D. Herschlag, and P.O. Brown. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biology*, 6(10):e255, 2008.

Curriculum Vitae

Personal Data

Nationality:	Swiss
Date of Birth:	13/12/1979
Address:	Gerbergasse 16, 4001 Basel
Phone (work):	+41 61 267 15 75
Phone (private):	+41 77 470 79 23
Email:	lukas.burger@unibas.ch

Education and Research Positions

2005–2009	**PhD in Bioinformatics (summa cum laude)** from the University of Basel, Switzerland. Title: "*Inference of Biomolecular Interactions from Sequence Data*". Supervisor: Prof Erik van Nimwegen. Co-supervisor: Prof Mihaela Zavolan.
2004	**Diploma thesis in computational physics** Eidgenössische Technische Hochschule Zürich, ETHZ, Switzerland. Title: *"Emergence of Ring Structures in Neural Networks with Spike-Timing Dependent Plasticity*". Supervisor: Prof Richard Hahnloser.
2000-2004	**Diploma in physics** Eidgenössische Technische Hochschule Zürich, ETHZ, Switzerland.
1992-1999	**Matura** Realgymnasium der Kantonsschule Rämibühl, Zürich, Switzerland.

Awards

2009 Winner of the 2009 Swiss Institute of Bioinformatics Young Bioinformatician Award for the work on the "*Inference of Intra- and Interprotein Interactions from Genomic Data*".

Publications

- Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Jr., Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP*. under review in Cell.
- Lukas Burger and Erik van Nimwegen. *Disentangling Direct from Indirect Co-evolution of Residues in Protein Alignments*. PLoS Computational Biology, 6(1):e1000633, 2010.
- Lukas Burger and Erik van Nimwegen. *Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method*. Molecular Systems Biology 4:165, 2008.
- Lukas Burger and Erik van Nimwegen. *A Bayesian algorithm for reconstructing two-component signaling networks*. Lecture Notes in Bioinformatics, Proceedings of the 6th international Workshop on Algorithms in Bioinformatics (WABI), 2006.

Conference Presentations and Posters

- *Inference of intra- and inter-protein interactions from sequence alignments*. Talk at the BC2 Conference on Molecular Evolution, Basel, 2009.
- *Inference of binding specificity of RNA-binding proteins from 4-ThioU-CLIP data: Towards a recognition code of KH domain-containing RNA binding proteins*. Poster at the Pacific Symposium of Biocomputing, Hawaii, 2009.
- *Accurate prediction of protein-protein interactions from sequence alignments*. Invited talk for the seminar series of the soft condensed matter group, University of Munich, 2008.

- *Accurate prediction of protein-protein interactions from sequence alignments.* Poster at the Berlin Summer Meeting for Computational and Experimental Molecular Biology, Berlin, Germany, 2008.
- *Accurate prediction of protein-protein interactions from sequence alignments.* Presentation at the Biozentrum Symposium, Basel, Switzerland, 2008.
- *A Bayesian algorithm for reconstructing two-component signaling networks.* Presentation at the Bertinoro Computational Biology Meeting, Bertinoro, Italy, 2007.
- *A Bayesian algorithm for reconstructing two-component signaling networks.* Presentation at the 6th International Workshop on Algorithms in Bioinformatics (WABI), Zurich, Switzerland, 2006.
- *A computational method to predict interaction specificity in two-component systems.* Poster at the CTBP Summer School on Quantitative Approaches to Gene Regulatory Systems, San Diego, CA, USA, 2006.

Summer Schools and Special Courses

- Internship at the Institute of Quantum Optics, Quantum Nanophysics and Quantum Information, Universität Wien, 2004.
- Otto Warburg International Summer School and Workshop on Networks and Regulation. Max Plank Institute of Molecular Genetics, Berlin, 2005.
- Cold Spring Harbour Course in Advanced Bacterial Genetics, Cold Spring Harbour, USA, 2007.
- CTBP Summer School on Quantitative Approaches to Gene Regulatory Systems, San Diego, CA, USA, 2006.

I am grateful to the following academic teachers

- ETHZ: R. Hahnloser, R. Douglas, K. Martin, D. Pescia, E. Trubowitz, G. Blatter
- University of Basel: E. van Nimwegen, M. Zavolan, T. Schwede, U. Jenal, T. Vetter