Title
Statistical evaluation of coevolution methods for predicting inter-protein contacts

Permalink
https://escholarship.org/uc/item/0s7405kz

Author
Avila-Herrera, Aram

Publication Date
2015

Peer reviewed|Thesis/dissertation

Statistical evaluation of coevolution methods for predicting inter-protein contacts

by

Aram Avila-Herrera

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Statistical evaluation of coevolution methods for predicting inter-protein contacts

Aram Avila-Herrera

Abstract

In this work I rigorously benchmark thirteen methods for detecting residue coevolution between proteins and illustrate their strengths and weaknesses under a variety of practical data scenarios, including cross-species protein interactions such as those involved in the host-pathogen "arms race" between lentiviruses and the mammals they infect. I investigate the effects of a variety of null distributions on the false positive rate of predictions. Additionally, I provide computational tools that facilitate a standardized coevolutionary benchmark analysis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction to coevolution

1.1 A little biology and motivation

The interactions between proteins—the building blocks of cellular machinery—drive most biological processes in a cell. Understanding the basic connections and dependencies between these building blocks is invaluable in learning how cells function, adapt, and how they can be manipulated into performing new tasks or correcting harmful behaviors, as in disease for example.

Through the lens of evolution, analyses of protein sequences can reveal the successes and hint at the failures of nature's experiments. By looking at related proteins across multiple species, we can identify the regions within those proteins that have not tolerated changes—are conserved—and infer that they pinpoint an important structural or otherwise functional region.

However, proteins seldom work in isolation. To carry out their functions, they typically require assembly into complexes. Thus, mildly deleterious or neutral mutations in one region may favor compensatory mutations in another region (within the same protein or in another member of the complex). For example, in signaling complexes, a sensory receptor and its ligand may evolve as a pair while maintaining similar three-dimensional shapes, charge, or polarity in the regions (residues) involved in maintaining the specific interactions. These pairs of mutations appear correlated in a multiple sequence alignment (MSA) and are a signature for an important relationship between a pair of proteins.

These correlated mutations can happen over various timescales; obligate symbionts and pathogens typically adapt quickly to changing conditions and new hosts. Viruses exist solely by hijacking their hosts' cellular machinery to reproduce, engaging in an evolutionary arms race with their hosts' immune systems—each optimizing the interactions that benefit their own survival, including restriction factors, degradation pathways, and antibodies (in the case of a vertebrate immune system). On the other hand, compensating mutations in an essential cellular pathway tend to be rarer and are found only after comparing sufficiently diverged species.

Structural and systems biology have had great success in identifying and characterizing many of these important interactions (e.g. Nucleosome [3], Proteasome [44], regulation in protein networks [37], [74]). However, resolving very large complexes and unstructured proteins remains technically difficult, a daunting task as the number of proteins is ever increasing (Uniprot currently catalogs approximately 47 million [72]).

Coevolutionary information extracted from MSAs may provide a complementary means to assign importance to the complicated molecular networks of the cell. However, before predicting contacting residues or antagonist viral proteins, we must first define and learn to reliably measure coevolution.

1.2 Measuring coevolution

Coevolution—"the change of a biological object triggered by the change of a related object" [79]—is a powerful concept when applied to molecular sequence analysis because it reveals positional relationships that are worth preserving across evolutionary time scales. Sequence evolution is constrained by essential molecular interactions, such as contacts within a protein or RNA structure, as well as inter-molecular interactions within protein complexes and signaling pathways. These constraints define an epistasis (i.e. genetic interaction) between sites (residues or base-pairs) where the probability of a substitution depends on the states of other sites involved in an interaction [21]. For example, a mildly deleterious or neutral mutation may change the fitness landscape such that compensatory or advantageous mutations at another site become more likely. Because epistasis can induce correlation between substitution patterns among columns in multiple sequence alignments, many methods have been developed that use evidence of coevolving alignment columns to detect physical interactions within and between biomolecules. These methods draw inspiration from diverse techniques in molecular phylogenetics, inverse statistical mechanics, Bayesian graphical modeling, information theory, sparse inference, and spectral theory (reviewed in [20], [40]).

Despite good rationale for coevolutionary approaches, physically interacting alignment columns have been notoriously difficult to identify from correlated patterns of sequence evolution for several reasons. First, shared evolutionary history creates a background of correlated substitution patterns against which it can be difficult to distinguish additional constraints derived from physical interactions. Common phylogeny is particularly strong within a gene family (e.g. predicting intra-molecular contacts). But it is also present across gene families within a species or even between species (e.g. predicting host-virus protein interactions), especially at shorter evolutionary distances where gene trees mirror species trees more closely. Coevolution methods have used a variety of approaches to counter the dependence induced by shared phylogeny, including removing closely related sequences from alignments to reduce non-independence [8], [26], differential weighting of sequences when computing statistics [17], [19], [56], and null distributions that directly model or indirectly account for phylogeny [22], [60], [9], [75].

A second challenge arises when trying to distinguish correlated evolution that arises from direct versus indirect interactions. Alignment columns that are implicated in an interaction only by transitivity can be strongly correlated, and most columns are involved in multiple, partially overlapping interactions. For these reasons, close physical interactions may not produce patterns of substitution that are significantly more highly correlated than the background present in structures. This problem has been the focus of a recent class of coevolutionary methods that focuses on reducing the number of incorrect predictions by disentangling direct from indirect correlations [56], [39], [7], [18], [12]. An alternative point of view considers these networks of indirectly correlated residues as protein sectors that can easily, through cooperative substitutions, respond to fluctuating evolutionary pressures [53].

Finally, due to low statistical power—resulting in part from the previous two challenges—physically interacting sites can typically only be detected in multiple sequence alignments that span large evolutionary divergences and contain many hundreds to thousands of sequences. Recent evaluations of a number of coevolution methods concluded that accurate contact predictions require alignments with one to five times as many sequences (with <90% sequence redundancy) as positions [42], [35].

To date, coevolutionary prediction of physically interacting alignment columns has been applied with success to intra-molecular contacts [17], [52], [36], [51] and well-characterized inter-molecular interactions [58], such as bacterial two-component signaling systems [41], enzyme complexes [28], and fertilization proteins [13].

Due in part to the rapidly changing computational environment, few comprehensive benchmarks exist, especially those that evaluate inter-protein contact prediction. Dutheil [20] rigorously benchmark many mutual information based methods along with a suite of phylogeny based statistics on intra-molecular contacts in rRNA. Clark et al. [12] compares newer direct coupling based methods against multi-dimensional mutual information in predicting intra-protein contacts. Mao et al. [50] evaluate mutual information based methods and newer direct coupling based methods on inter-protein contact prediction in non-interacting proteins (i.e. under the null hypothesis), but do not evaluate power and precision in true interactions.

In this work I rigorously benchmark thirteen methods for detecting residue coevolution between proteins and illustrate their strengths and weaknesses under a variety of practical data scenarios, including cross-species protein interactions such as those involved in the host-pathogen “arms race” between lentiviruses and the mammals they infect. I investigate the effects of a variety of null distributions on the false positive rate of predictions. Additionally, I provide computational tools that facilitate a standardized coevolutionary benchmark analysis.

I am optimistic that this contribution will guide structural and systems biologists to use coevolutionary tools to resolve large protein complexes, identify novel drug targets, predict drug resistance mutations, rationally design vaccines, and for other uses related to protein interactions that are unknown at this time. Additionally, this work should serve to motivate comprehensive sequencing to generate the data needed to identify coevolution in interactions with under-represented taxa. Finally, I hope that the statistically rigorous evaluation framework developed will serve as proving grounds for more powerful methods development.

Chapter 2

Benchmarking coevolution methods

In this chapter I address whether existing coevolutionary methods are specific and sensitive enough to predict structural contacts between interacting proteins from their alignments. I confirm that such analyses benefit from deep diverse alignments and from employing the empirical distribution of coevolution scores in estimating their P-values. I also report on additional features of the alignments that are important for predicting contacts.

2.1 Background

2.1.1 Classes of coevolution methods

The thirteen coevolutionary methods benchmarked in this analysis fall into three general groups (Table 2.1). *Information-based* methods are various flavors of mutual information between pairs of alignment columns, each pair considered independently. *Direct* methods are those that consider pairs of sites in the context of a sparse global statistical model for contacts in the multiple sequence alignment, i.e. they attempt to find a small number of couplings that best explain the observed correlations between alignment columns. In effect, *Direct* methods aim to remove indirect correlations that arise from transitivity. *Phylogenetic* methods explicitly use a substitution rate matrix and phylogenetic tree in their calculation of a coevolution statistic that can be configured to take into account the biochemical and physical properties of amino acid residues. The *Phylogenetic* methods implemented in the CoMap package additionally report a P-value estimated from an internal simulation of independently evolving sites. In this benchmark I use the CoMap P-value as a statistic for comparison with other coevolution methods.

Other differences among the coevolution methods include the incorporation of two additional techniques that have been shown to improve performance, re-weighting sequences such that similar sequences contribute less to the final score [8] and applying an Average Product Correction (APC) to remove background noise and phylogenetic signal from the raw coevolution statistics

[19].

2.1.2 Alignments and structures

A gold standard data set for which to benchmark the performance of coevolution detecting methods does not yet exist. Residues may coevolve due to structural constraints or longer range functional constraints, however the latter is more difficult to validate and to retrieve experimental results for. In comparison, structural information is readily available for many protein complexes through the Protein Data Bank (PDB) [4], and can be used to validate predictions of coevolving residues.

Therefore, I benchmarked the coevolution methods on 33 within-species pairs of bacterial protein families with an ortholog in *E. coli* that also have an associated representative co-crystal structure deposited in the protein data bank (PDB). These include a set of paired alignments compiled by Ovchinnikov, Kamisetty, and Baker [58] (Ovch32), plus the histidine kinase-response regulator (HisKA-RR) bacterial two-component system from Procaccini et al. [62], provided by the authors. I included HisKA-RR, because it is a well-characterized interaction with a deep, diverse multiple sequence alignment (8998 sequences for each gene) and genetic evidence supporting several interactions. For these reasons, HisKA-RR has also been used previously in coevolutionary analyses [65].

Because the HisKA-RR alignment is so deep, it enabled us to quantify the effects of alignment size and diversity by uniformly down-sampling the full alignment to produce a wide range of smaller pairs of HisKA and RR multiple sequence alignments. These sub-sampled alignments have six different numbers of sequences (5, 50, 250, 500, 1000, 5000) with phylogenies also sub-sampled from the original tree (Figure 2.26). The 32 alignment pairs in Ovch32 naturally varied in size (range 216–6732 sequences) (Figure 2.27).

In addition to the number of sequences in the alignments (N), I consider the phylogenetic diversity (PD [25]) of the alignments—also captured in the effective number of sequences (N_{eff}) as calculated by PSICOV [39], the diversity within individual alignment columns measured by entropy, the alignment length (L) (i.e. the number of alignment columns), the proportion of contacting residues in the alignment.

2.1.3 Measuring performance

The performance of each method to distinguish contacting pairs of residues (positives) from other residue pairs (negatives) was measured as previously described [39], [24]. Briefly, for each pair of multiple sequence alignments from two interacting proteins, I compared every site in the first protein to every site in the second protein and scored these pairs of alignment columns for coevolution using each of the methods in Table 2.1. I then used these coevolution scores to predict inter-protein contacts—pairs of amino acid residues that are less than 8 angstroms (Å) apart from each other (measured between C_βs)—in a representative co-crystal structure.

I evaluated performance using power (also called recall and true positive rate

(TPR)) (Equation (2.1)) and precision (also called positive predictive value (PPV)) (Equation (2.3)) at a range of low false positive rates (FPR)—the proportion of negatives falsely predicted as positives (Equation (2.2)). Power and precision are complementary performance measures that quantify the percentage of interacting residue pairs that are found and the percentage of identified residue pairs that are interacting, respectively. Precision is a useful measure of performance in cases where positives (contacting pairs of residues) are overwhelmed by negatives (non-contacting residues). A method with high precision is helpful for generating lists of high confidence pairs of residues for expensive follow-up studies, even if it misses a number of truly interacting sites and therefore has relatively low power. I additionally examined four threshold-independent performance measures, area under Receiver-Operator Curve (auROC), area under precision-recall curve (auPR), maximum F_1-score (f_{max}) (Equation (2.4)), maximum ϕ (ϕ_{max}) (Equation (2.5)). See Table 2.5.

$$TPR = \frac{TP}{TP + FN} \tag{2.1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.2}$$

$$PPV = \frac{TP}{TP + FP} \tag{2.3}$$

$$F_1 = \frac{2 \cdot PPV \cdot TPR}{TPR + PPV} \tag{2.4}$$

$$\phi = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \tag{2.5}$$

An alternate definition of contacts I explored defines contacts as residue-pairs with less than 6Å between their closest non-hydrogen atoms. I also evaluated performance in the HisKA-RR sub-alignments using a stricter definition of contacts that, in addition to spatial proximity ($C_\beta < 8$Å), requires biochemical evidence for the role of the contacting residues in determining the specificity of the protein interaction. A list of such residues in representative sequences is found in [10], [47], [32], [70], [45]. Trends in the results were generally similar across these choices of definition for true interactions, but I observed some differences in performance between definitions when the false positive rate (FPR) is controlled (Figures 2.1 and 2.2).

2.1.4 Common null distributions

In practice, when applying the methods in this study the structure usually is not known. In fact, one of the main goals of extracting coevolutionary information from sequence alignments is to use it to predict structural contacts. One therefore uses a null distribution to control false predictions. Specifically, an upper quantile of the distribution of coevolutionary statistics in the absence of coevolutionary constraint is used as a threshold; one declares any pair of sites with a statistic exceeding the threshold a predicted contact.

The goal is to minimize false predictions by predicting contacts only when statistics are much larger than expected under the null distribution. A variety of null distributions are commonly used, including theoretical limiting distributions [19], [2], [27], the empirical distribution of observed scores (under the assumption that most pairs of sites are not coevolving) [30], and parametric, semi-parametric, and non-parametric bootstrap distributions [22], [76]. Theoretical and empirical nulls are computationally inexpensive compared to bootstrap methods, which require accurately simulating thousands of large data sets (See Chapter 3)

The empirical distribution of observed coevolution scores in an alignment is commonly used as the null distribution. A P-value ($P_{empirical}$) for a score S is simply the proportion of scores that are more extreme than S. This straightforward method can be easily applied with any statistic. However, it assumes that no pairs of sites are coevolving and should therefore produce thresholds that are too strict when there are some coevolving sites in the data set (i.e. making it harder to predict real contacts).

Another choice of null distribution is a theoretical distribution with parameters that are pre-determined or estimated from the observed scores, also assuming that virtually no pairs of sites are coevolving. Often a standard normal distribution is used. Under this assumption, I standardized the coevolution scores to Z-scores and compared these to upper quantiles of the standard normal distribution (mean = 0, variance = 1). I then used the resulting upper-tail P-values (P_{normal}) to predict contacting residue pairs.

Two properties of these distributions are that there is a one-to-one mapping of scores to P-values within each alignment pair, but not necessarily across alignment pairs. Therefore, P-values in short protein families (or in analyses with extensive filtering) may be overly conservative and perhaps not comparable to P-values in other alignment pairs.

2.2 Benchmark results

My primary finding is that many coevolutionary methods are able to detect inter-molecular contacts at low FPRs in alignments with hundreds of diverse sequences from each protein, consistent with previous studies of intra-molecular contacts [12], [20], specifically when the alignments are deeper than they are long [42], [35]. I capture this rectangular quality in the statistic $\mathrm{N_{eff}}/\mathrm{L}$, where $\mathrm{N_{eff}}$ is the effective number of sequences as calculated by PSICOV [39] and L is the total number of columns in both alignments. I observe similar trends when using the number of sequences (N) or their phylogenetic diversity (PD) [25], rather than $\mathrm{N_{eff}}/\mathrm{L}$, to compare performance. The relationship between N, PD, and $\mathrm{N_{eff}}$ is explored further in Section 2.2.3.1: Diversity of sequences. The diversity of residues within the individual alignment columns that make up each pair is another important factor to consider, and is explored in Section 2.2.3.2: Variability in alignment column-pairs.

2.2.1 Power and precision

Both power and precision improve with increasing N_{eff}/L for nearly all coevolutionary methods (Figures 2.4 and 2.5), in the HisKA-RR data set. However, for alignments with $N_{\mathrm{eff}}/L < 1.0$, power at $FPR < 5\%$ remains relatively low ($< 50\%$), and even lower ($< 10\%$) when controlling the false positive rate more strictly ($FPR < 0.1\%$). Precision is expectedly higher at $FPR < 0.1\%$ than at $FPR < 5\%$, but also remains below 50% for "square" alignments. Additionally, the performance metrics f_{max} and ϕ_{max} show that there are no score thresholds (i.e. the strictness of predictions) that achieve both high precision and power in alignments with $N_{\mathrm{eff}}/L \lesssim 3.0$ (Figures 2.6 to 2.8). Despite the smaller range in N_{eff}/L values, these performance trends are also observed across the 32 alignments in Ovch32 (Figures 2.9 to 2.11).

However, in the HisKA-RR alignment, I observed two exceptions to this trend when using the strictest definition for contacting pairs (i.e. requiring residue $C_\beta < 8$Å coupled with biochemical evidence for specificity determination). First, the standard MI statistic is the most precise method for detecting contacting sites in alignments with $N_{\mathrm{eff}}/L > 1.6$ and $FPR < 0.1\%$ (Figure 2.2). Second, mutual information normalized by the joint entropy (MI_j) has relatively high power compared to the *Information-based* methods and is the most powerful method for detecting contacting sites that are supported by experimental evidence at $FPR < 5\%$ (Figure 2.1). However, MI_j has drastically lower power at $FPR < 0.1\%$ (Figure 2.3). These findings suggest MI_j may be useful for detecting as many contacts as possible if a moderate FPR can be tolerated. *Information-based* methods are straightforward to compute, adding to their utility in these settings.

2.2.2 False positive rate

I used the sampled sub-alignments of HisKA-RR and the 32 alignments in Ovch32 to compare the performance of two commonly used null distributions and to evaluate the sensitivity of each approach to alignment size. For each null distribution and coevolutionary statistic, I first employed the non-contact pairs of residues to assess if the FPR was truly controlled or not at given target FPRs of 5% and 0.1%.

2.2.2.1 Empirically distributed null ($P_{empirical}$)

Contrary to my expectations, I found that the empirical null distribution produces nominal FPRs that exceed target FPRs (Figures 2.12, 2.13, 2.16 and 2.17). However, it is the *Direct* methods that best control the nominal FPR in both sets of alignments, marginally exceeding the target FPR in only a couple of cases. The *Information-based* methods fared well in the alignments in [58], however the HisKA-RR sub-alignments reveal that at $\mathrm{N_{eff}}/\mathrm{L} < 0.3$, control of the FPR is lost, especially in $\mathrm{MI_{Hmin}}$. The *Phylogenetic* method that consistently exceeded the target FPR was the CoMap correlation analysis ($\mathrm{CMP_{cor}}$) which makes no assumptions regarding the biochemical properties of the amino acids. These results suggest that the empirical null distribution is not as conservative of an approach as one might expect from including contacting residue pairs in the null distribution. Although, it may suffer from some of the same effects that make the normal null distribution anti-conservative, such as shared phylogeny or structural constraints, alignments with very few sequences (e.g. 5–50) have a limited number of possible scores which leads to ties in P-values between contacting and non-contacting residues. If contacts and non-contacts have roughly the same $P_{empirical}$ values, the target and nominal FPRs should be similar. But with large ammounts of ties, predictions are made in blocks, possibly forcing discontinuous jumps in the nominal FPR with respect to the target FPR. This could compound or balance the anti-conservativeness of $P_{empirical}$.

2.2.2.2 Normally distributed null (P_{normal})

Using a standard normal as the null distribution, we found that nominal FPRs consistently exceed the target FPR across the range of $\mathrm{N_{eff}/L}$ values in both the HisKA-RR sub-alignments and the alignments in [58] (Figures 2.14, 2.15, 2.18 and 2.19). In general, as $\mathrm{N_{eff}/L}$ increases, the nominal FPR for *Direct* methods increases while it decreases in *Information-based* methods. Nominal FPRs were observed to be as great as twice to 20 times the target FPR for target FPRs 5% and 0.1% respectively. This suggests that either non-contacting residue pairs carry signals of coevolution (e.g. due to phylogeny, structural, or other evolutionary constraints) and/or that Z-scores of coevolution statistics have variance greater than one across non-contacting residues (e.g. due to an underestimated standard deviation across residue pairs resulting from within protein constraints or residues appearing in many pairs). Three of the four phylogeny aware CoMap methods controlled the nominal FPR below the target in all cases suggesting that the charge compensation analysis is predicting long-range residue interactions as well as contacts.

Thus, while the normal distribution applied to standardized coevolution statistics can practically be used as a null distribution, we conclude that this approach results in elevated rates of false positive predictions, likely due to shared phylogeny or structural constraints affecting non-contacting residue pairs. A theoretical null (e.g. non-central gamma [29]) that is parameterized for individual column pairs may therefore be more appropriate and is explored in Section 3.3.

2.2.3 Other alignment features important for performance

2.2.3.1 Diversity of sequences

To investigate whether higher power in larger alignments results primarily from the number sequences per se or depends upon the diversity of the sequences, I compared the performance across alignments with different diversity values but the same number of sequences. I quantified diversity using phylogenetic diversity (PD) [25] and the effective number of sequences as calculated by PSICOV (N_{eff}) [39] (Figures 2.20 to 2.22).

For HisKA-RR sub-alignments, I found weak positive and negative relationships between the nominal false positive rate and PD for some methods in alignments with 5000 sequences at given target false positive rates. For each group of equally sized alignments for each method (and for each null distribution and significance threshold), I tested whether the false positive rate correlates with PD using Spearman's rho. Few methods had uncorrected P-values < 0.05 and none when controlling for the 336 comparisons (smallest uncorrected P: 1.73e-3; ρ: 0.85 for $\mathrm{MI_j}$ at $\mathrm{N} = 5000$, $P_{empirical} < 0.001$). Testing for a bulk correlation (ignoring method; normalizing PD by alignment size) reveals a weak positive correlation ($\rho = 0.27$, $P <$ 1.9e-29) at P_{normal} and $P_{empirical} < 0.05$ but not < 0.001. Overall this suggests that the false positive rate may increase with more diverse sequences at loose significance thresholds. Alternatively, the PD ranges were too small to detect a relationship with false positive rate.

While the range in diversity for alignments with 5 sequences is small (PD: 7.5-11, $\mathrm{N_{eff}}$: 5), under the normal distribution, the false positive rate is better controlled in diverse alignments. However, under the empirical null, the *Information-based* methods do not control the FPR for these alignments and have larger false positive rates as diversity increases in these alignments.

One caveat of the HisKA-RR analysis is that (for computational reasons) I generated sub-alignments by random sampling and therefore only explored a range of phylogenies close to the typical diversity for each alignment size. I observe fairly strong correlations between cutoff-independent performance metrics and $\mathrm{N_{eff}}$ (and also $\mathrm{N_{eff}/L}$ as L is constant in HisKA-RR). The alignments in Ovch32 provide a broader range of phylogenetic scenarios. Across these 32 interactions, N_{eff} is at best weakly correlated with the same performance metrics (Table 2.2). However, accounting for alignment length (with N_{eff}/L) reveals that there is a relationship between alignment depth and performance. (Table 2.3, Figures 2.21 and 2.22) show that high N_{eff} alone does not guarantee good performance. For example, taking the best performing method at each alignment pair, the alignment pair with the highest N_{eff} had at best the fourth poorest $\phi_{\max}$. Conversely, the third smallest N_{eff} corresponds to the third best $\phi_{\max}$; and at FPR < 0.001, it had the highest precision (PPV $= 63\%$). Interestingly, it also has the shortest length ($L = 168$ columns), suggesting that perhaps taking into account the proportion of possible contacts may play an important role in estimating expected performance.

2.2.3.2 Variability in alignment column-pairs

To explore the effect of substitution rate variation across sites in HisKA and RR, I parsed the performance results according to the entropy of the two alignment columns (one from each gene) in every pair of evaluated sites. For each alignment size (6 sizes; 10 alignments of each size), I split columns into below- versus above-median entropy separately for each gene, and then classified pairs of sites into the resulting four groups.

Then, for a subset of methods, I computed power and precision separately for each rate category group. This analysis showed that faster evolving (i.e. above-median-HisKA paired with above-median-RR) contacts are generally the easiest to detect with coevolutionary methods. Dually conserved residues (i.e. low-HisKA paired with low-RR) are the next easiest to detect (Figure 2.23). I conclude that $\mathrm{MI_w}$'s drop in performance at 5000 sequences may be due to dually-variable columns being improperly reweighted. These analyses show that sequence variation quantitatively affects the accuracy of coevolution analyses, with most methods performing best when coevolving residue pairs have similar substitution rates.

2.2.3.3 Proportion of contacts in the alignment

Finally, I looked at the relationship between performance and the proportion of residue pairs that are contacts. Comparing across the structures in the Ovch32 data set, we observed the proportion of contacts is correlated with precision at FPR $< 0.1\%$ (Figure 2.24, Table 2.4). This means that most strongly coevolving residues in a protein pair are more likely to be physically interacting in co-crystal structures with a larger fraction interface residues. Power is also correlated with the proportion of contacts, though not as strongly as precision (Figure 2.25).

2.3 Conclusions

In general, I confirm that coevolutionary methods that adjust for background phylogenetic signal through sequence re-weighting and/or average product correction (APC) (e.g. DI, $\mathrm{DI_{plm}}$, and PSICOV) perform better than the phylogeny unaware mutual information (MI) based methods and the phylogeny aware approaches that explicitly use evolutionary models.

CoMap performance is an interesting case because, in contrast to DI, $\mathrm{DI}_{\mathrm{plm}}$, and PSICOV, it was not designed to find contacting residues. In the smallest alignments (5 sequences) I tested, it can have slightly better performance than the other methods. However, its poor performance in other alignments may indicate that it is identifying a set of coevolving residue pairs that partially overlap with contacting residues. It remains to explore whether CoMap can be used to prioritize residue pairs predicted by the other methods for functional assays.

We are still left with the challenge of how to choose an appropriate P-value cutoff in a real analysis when the structure is unknown. Since my findings indicate that nominal FPRs exceed target FPRs using P_{normal} and $P_{empirical}$ for nearly all methods, stricter P-value cutoffs than the target false positive rate seem warranted. But it is not clear how much stricter will be needed in any given alignment pair without additional information to guide such modifications (e.g. incorporating alignment properties such as $\mathrm{N}_{\mathrm{eff}}/\mathrm{L}$ into a model for each coevolution method). Hence, in most applications one must simply aim to control a target FPR, knowing that the true error rate is likely to be larger. For this reason, the empirical null distribution may be the best choice to use as it controls error rates across the majority of alignment sizes, target FPRs, and coevolution methods tested (Figures 2.12, 2.13, 2.16 and 2.17). As a rule of thumb, the empirical null overall controls the FPR for the Direct methods, however in small alignments (5 sequences or $\mathrm{N}_{\mathrm{eff}}/\mathrm{L}$

< 0.3) it can be up to 1.5 times the target FPR. I therefore recommend sequencing deeply enough to attain $\mathrm{N_{eff}}/\mathrm{L} > 1.0$ to control FPR and > 2.0 to ensure modest TPR and PPV.

Null distributions that directly use an evolutionary model under the null hypothesis (proteins are evolving independently) have the potential to provide realistic thresholds for individual column pairs. Controlling the FPR using *P*-values obtained through simulation is explored in Chapter 3.

2.4 Figures

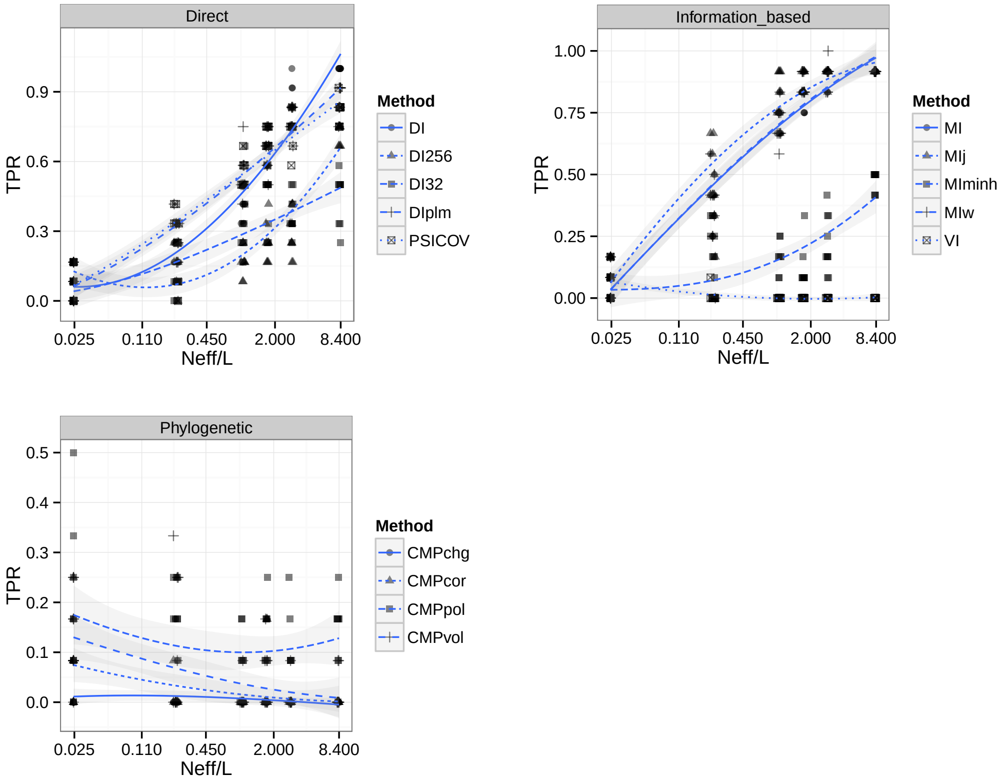

Figure 2.1: HisKA-RR with contacts defined using biochemical evidence for specificity determination. Power (TPR) at FPR < 5%. 95% confidence intervals for loess smooths are shown in gray

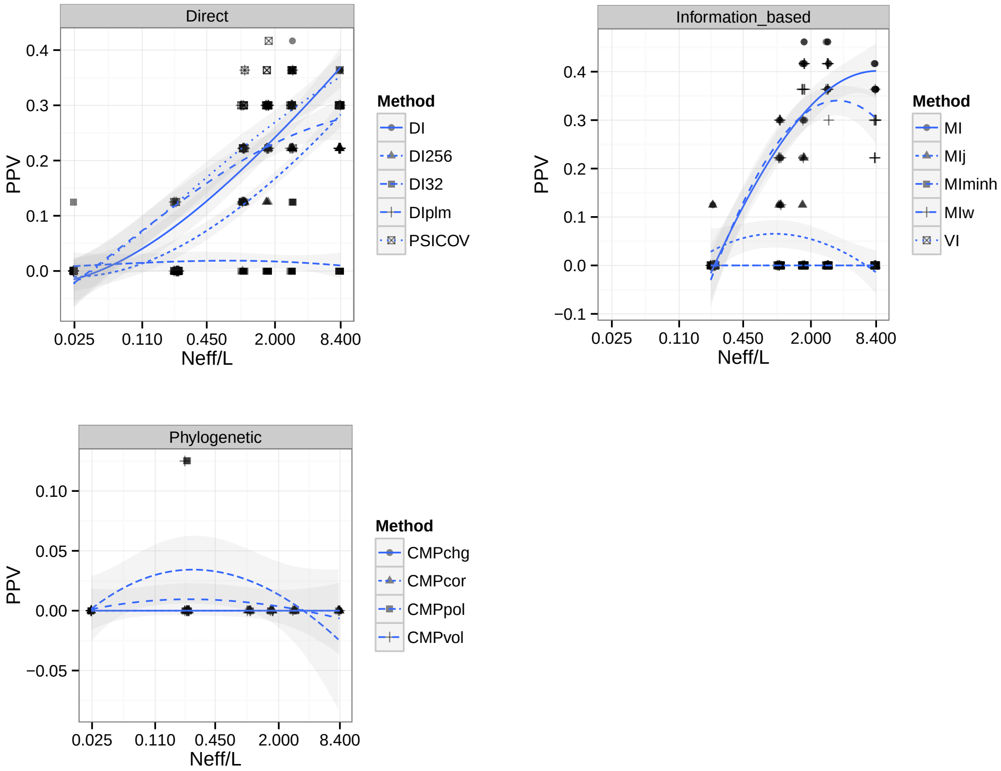

Figure 2.2: HisKA-RR with contacts defined using biochemical evidence for specificity determination. Precision (PPV) at FPR $< 0.1\%$. 95% confidence intervals for loess smooths are shown in gray

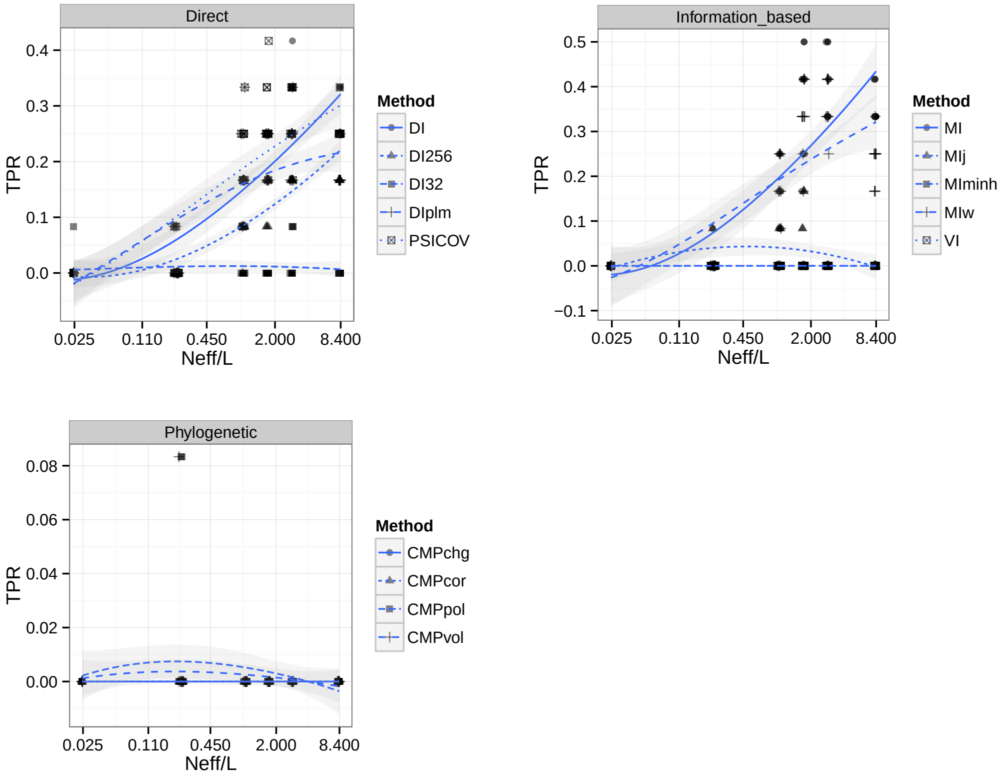

Figure 2.3: HisKA-RR with contacts defined using biochemical evidence for specificity determination. Power (TPR) at FPR $<$ 0.1%. 95% confidence intervals for loess smooths are shown in gray

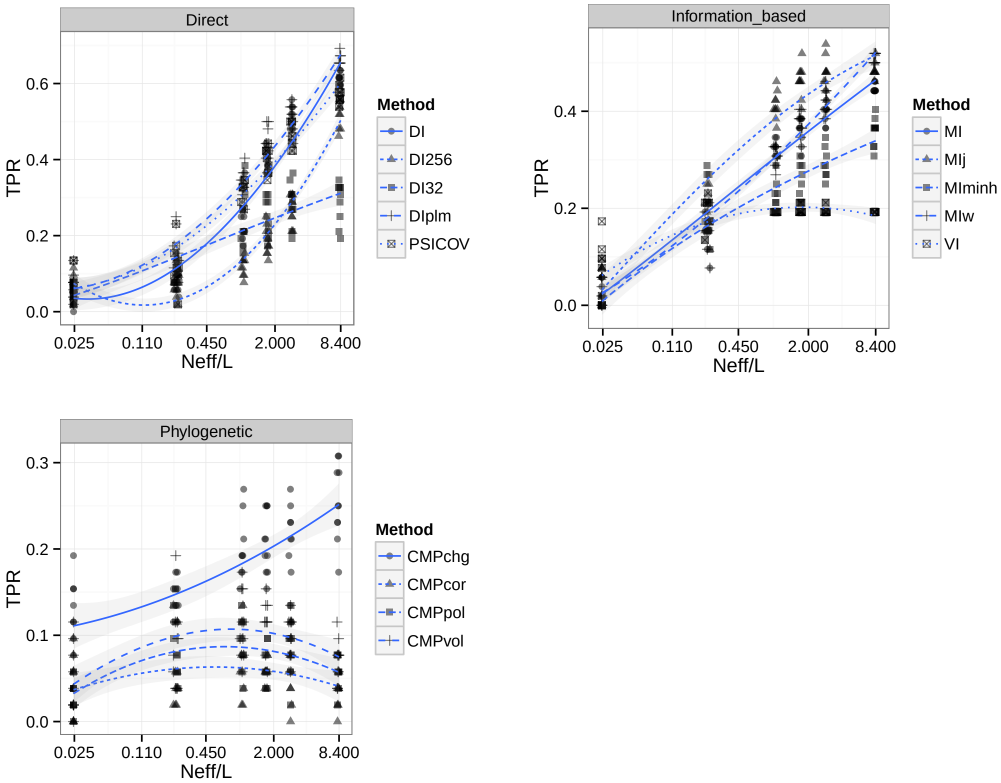

Figure 2.4: HisKA-RR with contacts defined using spatial proximity ($C_\beta <$ 8Å). Power (TPR) at FPR $< 5\%$. 95% confidence intervals for loess smooths are shown in gray

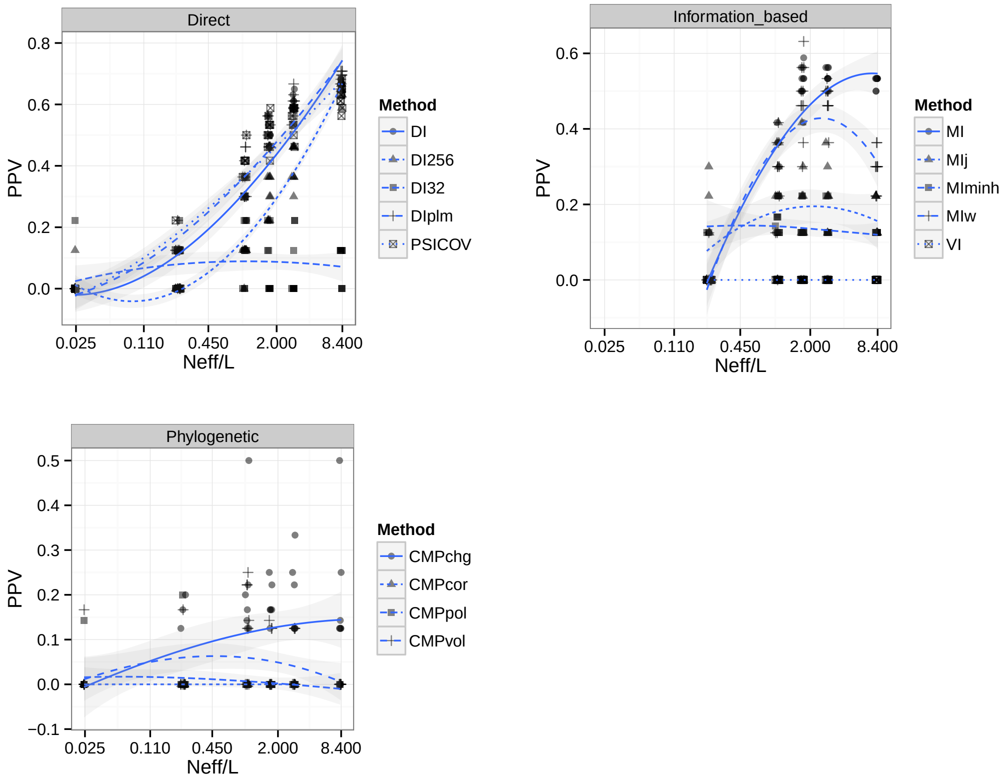

Figure 2.5: HisKA-RR with contacts defined using spatial proximity ($C_\beta < 8$Å). Precision (PPV) at FPR < 0.1%. 95% confidence intervals for loess smooths are shown in gray

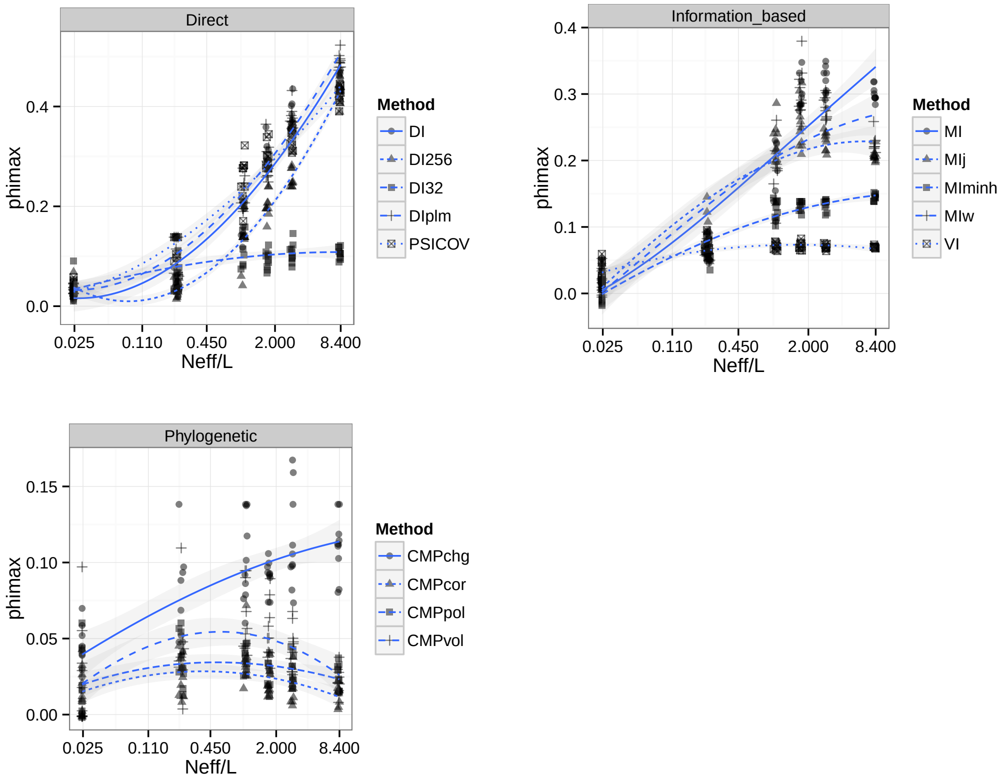

Figure 2.6: HisKA-RR with contacts defined using spatial proximity ($C_\beta <$ 8Å). Phi_{max}. 95% confidence intervals for loess smooths are shown in gray

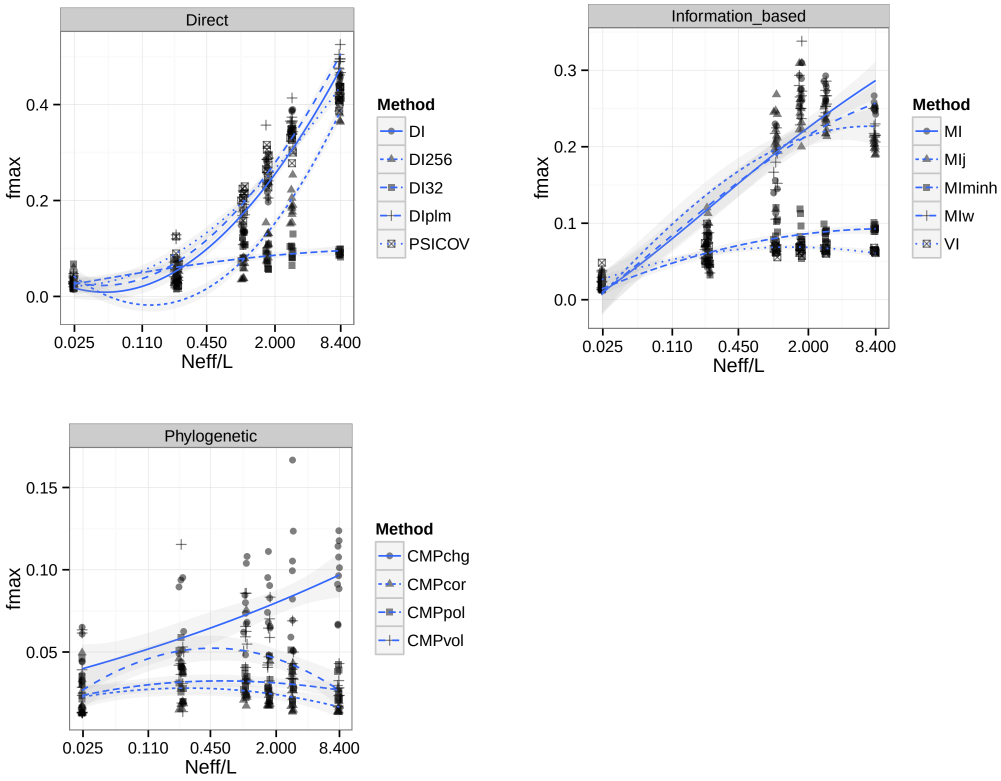

Figure 2.7: HisKA-RR with contacts defined using spatial proximity ($C_\beta <$ 8Å). $F_{\max}$. 95% confidence intervals for loess smooths are shown in gray

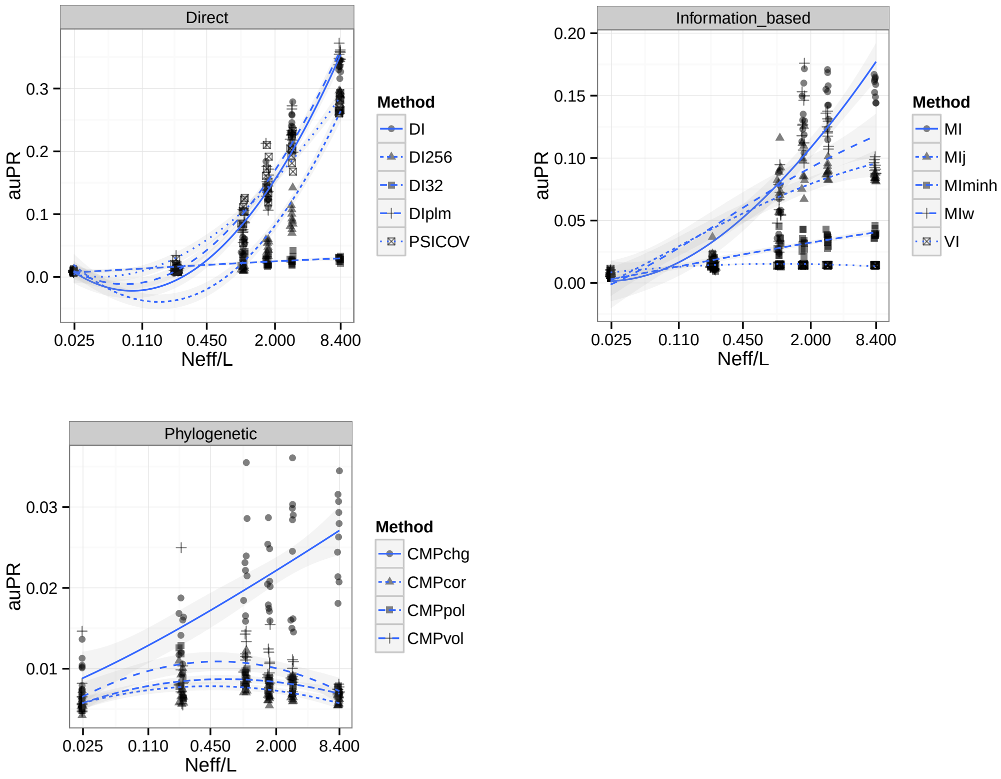

Figure 2.8: HisKA-RR with contacts defined using spatial proximity ($C_\beta <$ 8Å). AuPR. 95% confidence intervals for loess smooths are shown in gray

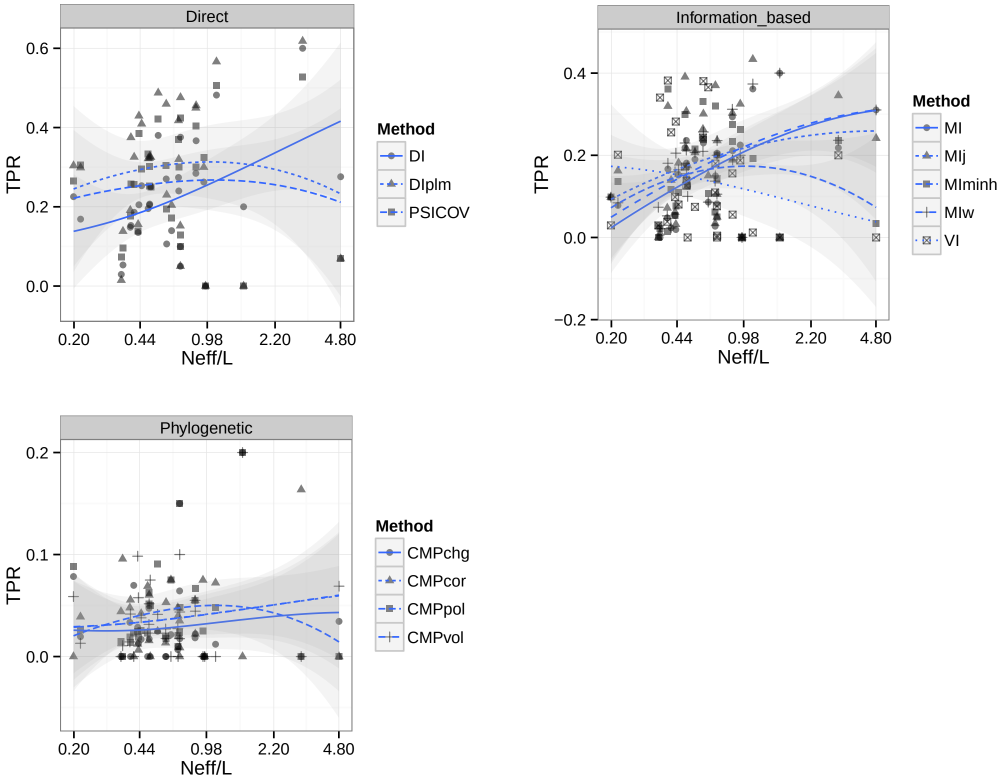

Figure 2.9: Ovch32 alignments [58] with contacts defined using spatial proximity ($C_\beta < 8$Å). Power (TPR) at FPR $< 5\%$. 95% confidence intervals for loess smooths are shown in gray

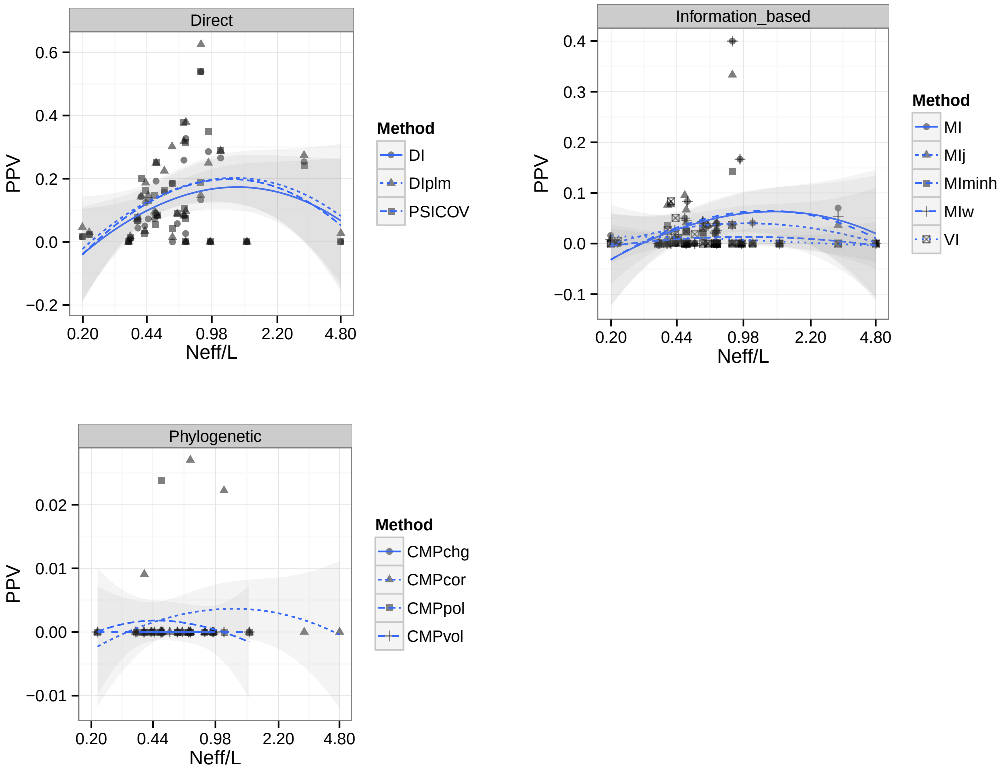

Figure 2.10: Ovch32 alignments [58] with contacts defined using spatial proximity ($C_\beta < 8$Å). Precision (PPV) at FPR $< 0.1\%$. 95% confidence intervals for loess smooths are shown in gray

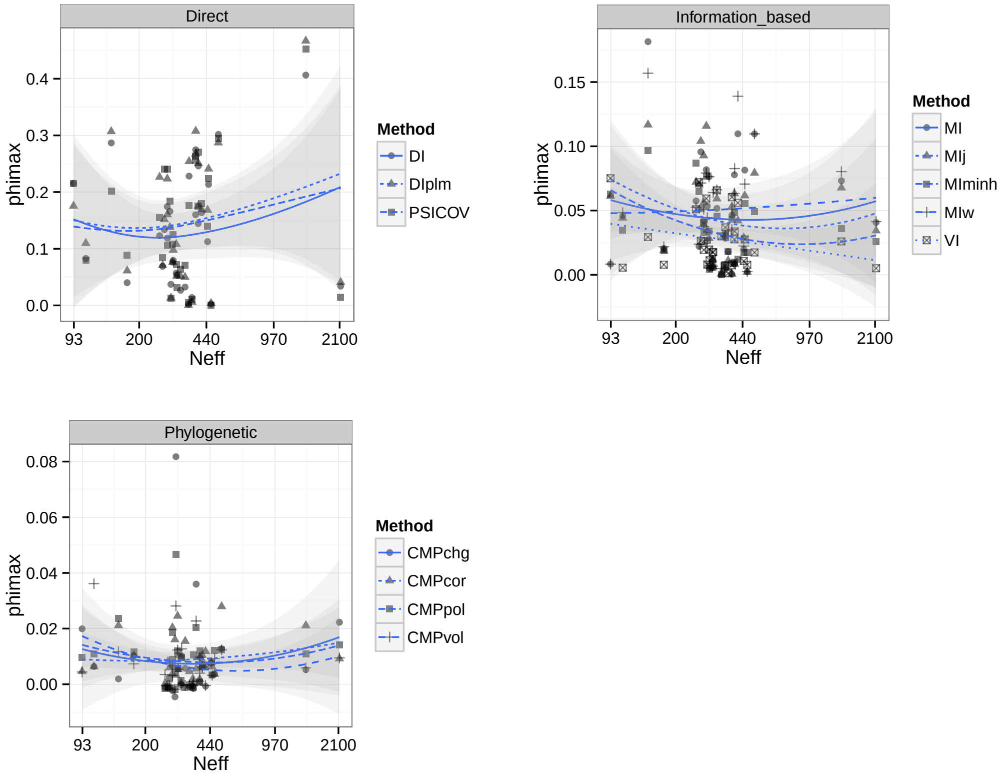

Figure 2.11: Ovch32 alignments [58] with contacts defined using spatial proximity ($C_\beta < 8$Å). Phi$_{\text{max}}$. 95% confidence intervals for loess smooths are shown in gray

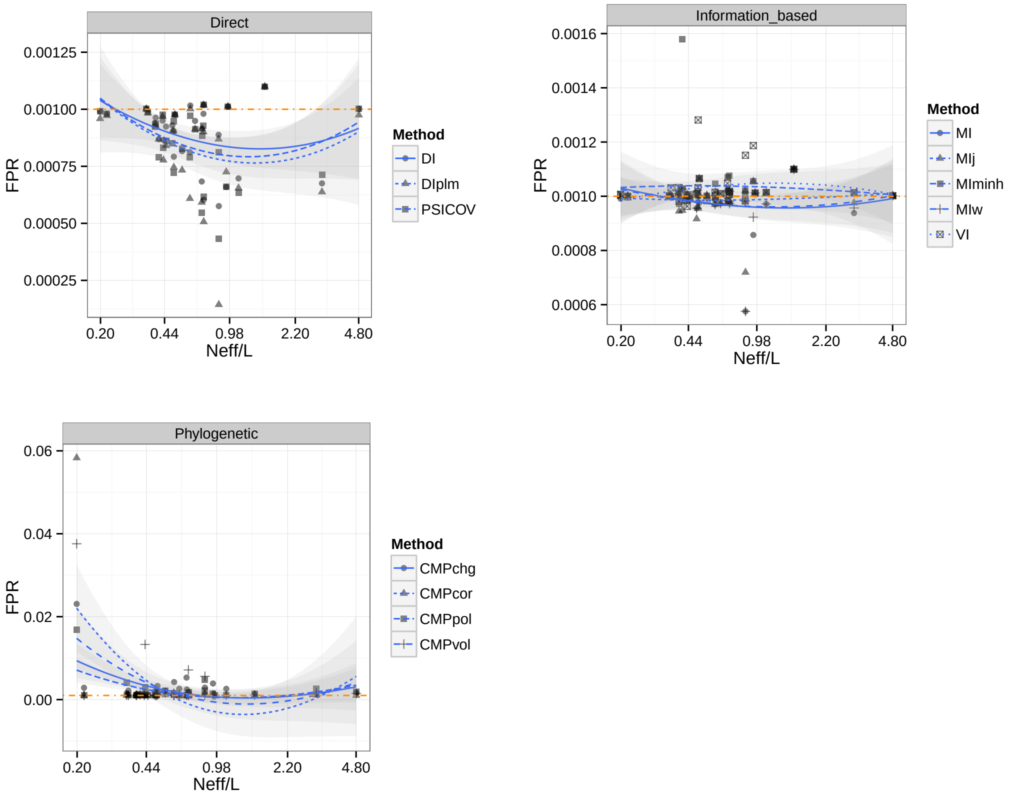

Figure 2.12: Ovch32 alignments [58] with contacts defined using spatial proximity (C_β < 8Å). False positive rate (FPR) at $P_{empirical} < 0.001$. 95% confidence intervals for loess smooths are shown in gray

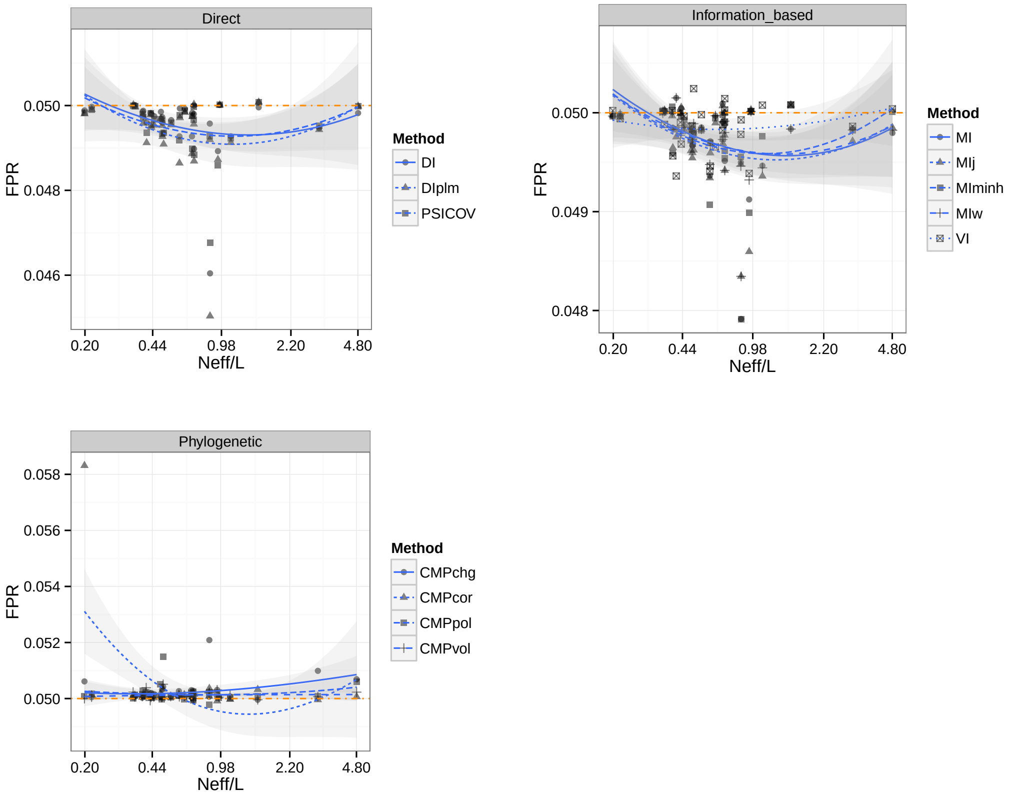

Figure 2.13: Ovch32 alignments [58] with contacts defined using spatial proximity ($C_\beta < 8$Å). False positive rate (FPR) at $P_{empirical} < 0.05$. 95% confidence intervals for loess smooths are shown in gray

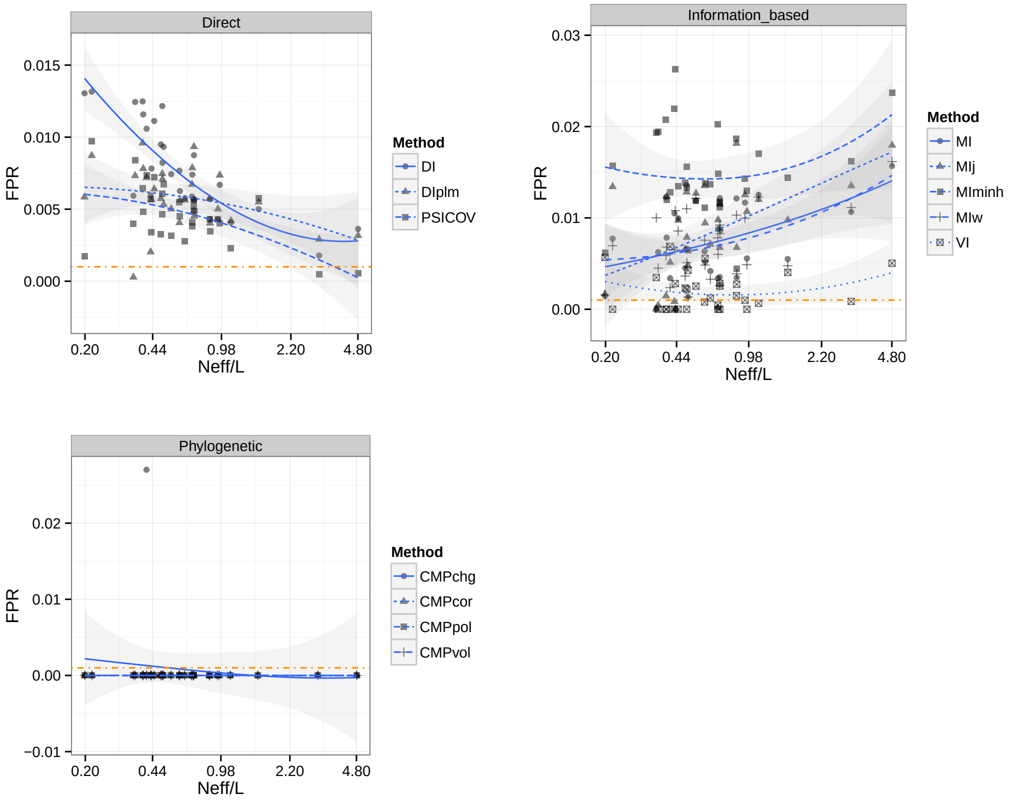

Figure 2.14: Ovch32 alignments [58] with contacts defined using spatial proximity ($C_\beta < 8$Å). False positive rate (FPR) at $P_{normal} < 0.001$. 95% confidence intervals for loess smooths are shown in gray

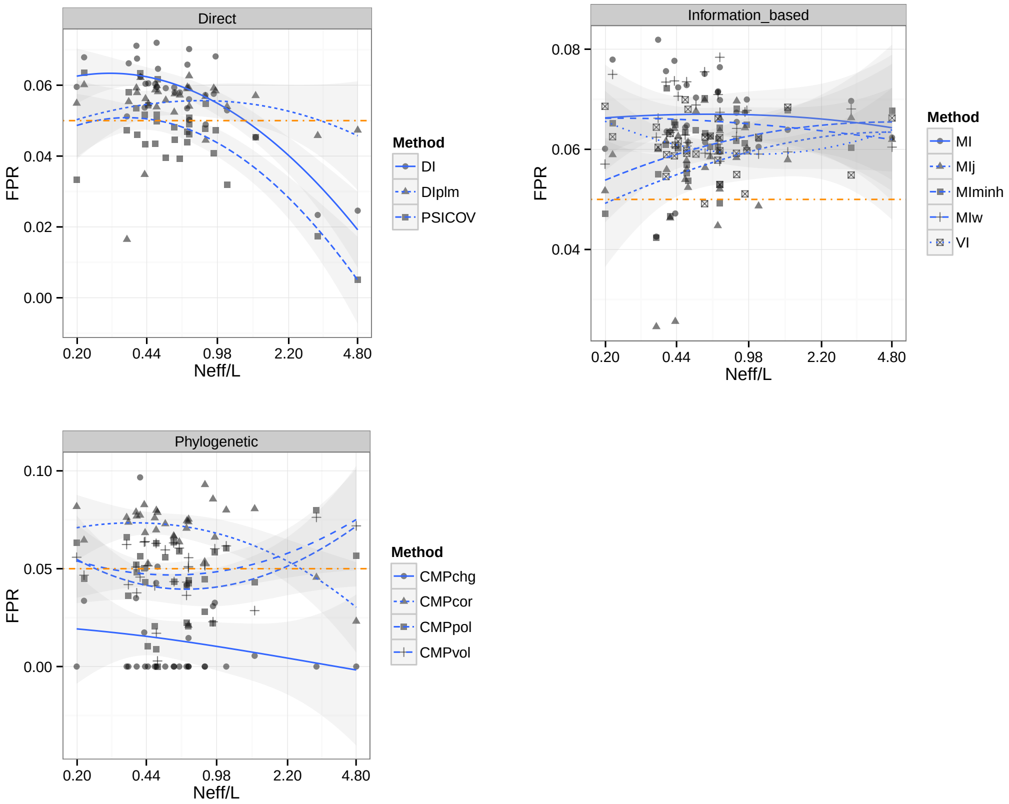

Figure 2.15: Ovch32 alignments [58] with contacts defined using spatial proximity ($C_\beta < 8$Å). False positive rate (FPR) at $P_{normal} < 0.05$. 95% confidence intervals for loess smooths are shown in gray

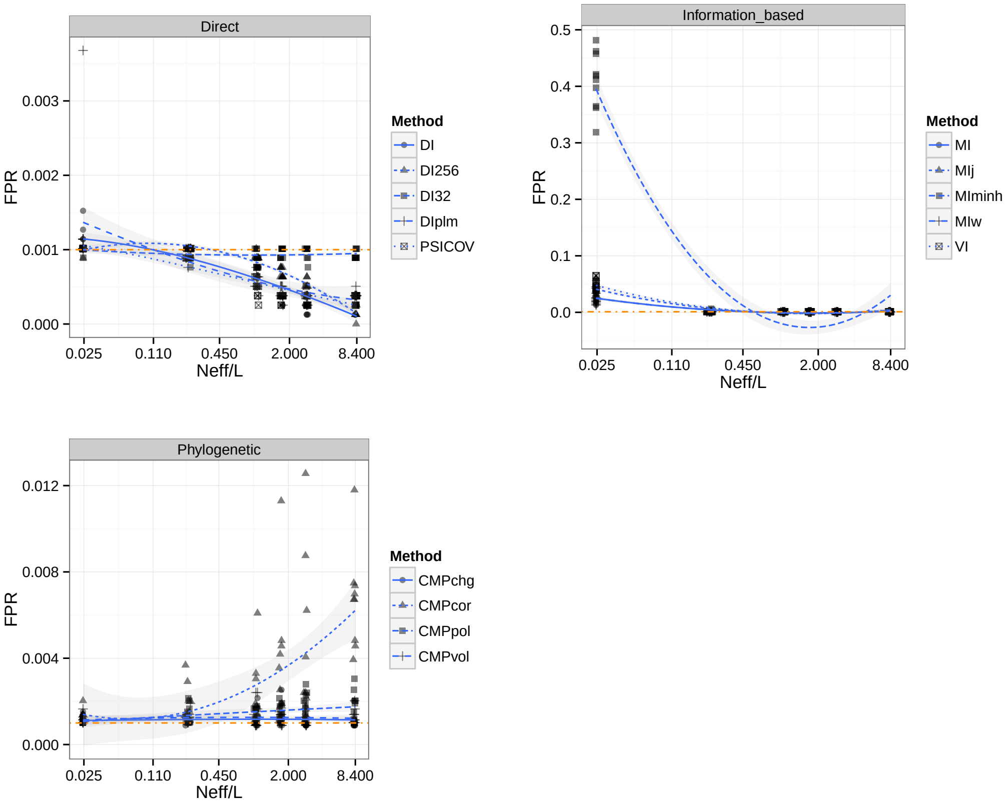

Figure 2.16: HisKA-RR with contacts defined using spatial proximity ($C_\beta <$ 8Å). False positive rate (FPR) at $P_{empirical} < 0.001$. 95% confidence intervals for loess smooths are shown in gray

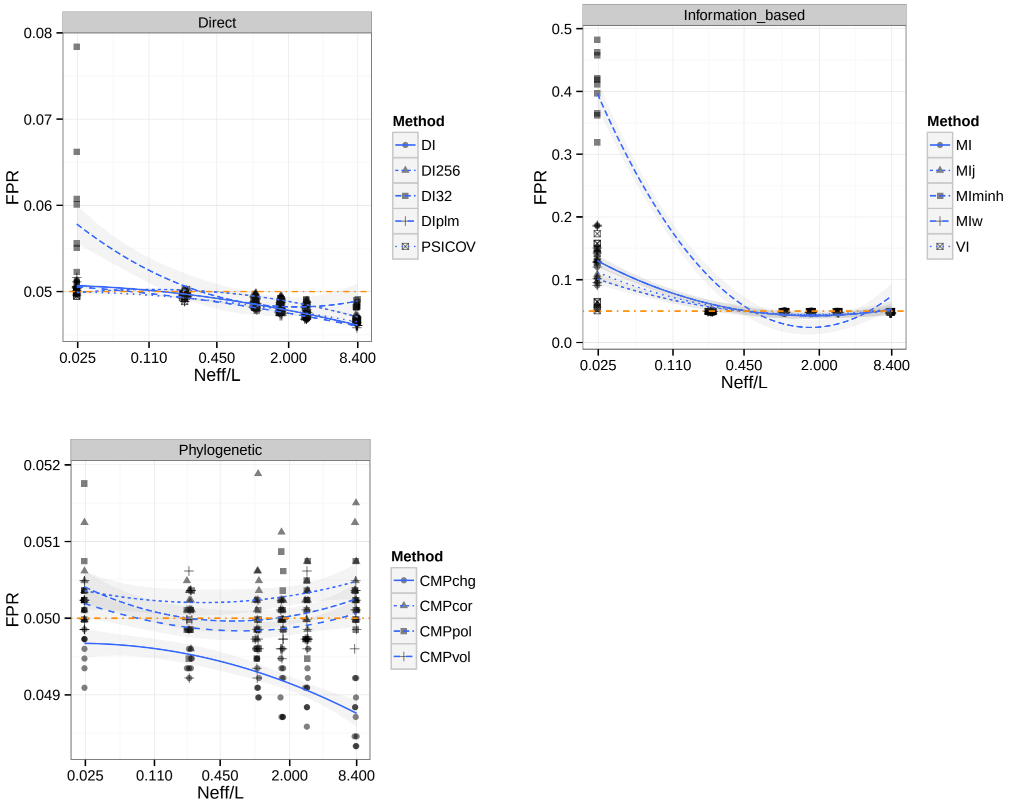

Figure 2.17: HisKA-RR with contacts defined using spatial proximity ($C_\beta <$ 8Å). False positive rate (FPR) at $P_{empirical} < 0.05$. 95% confidence intervals for loess smooths are shown in gray

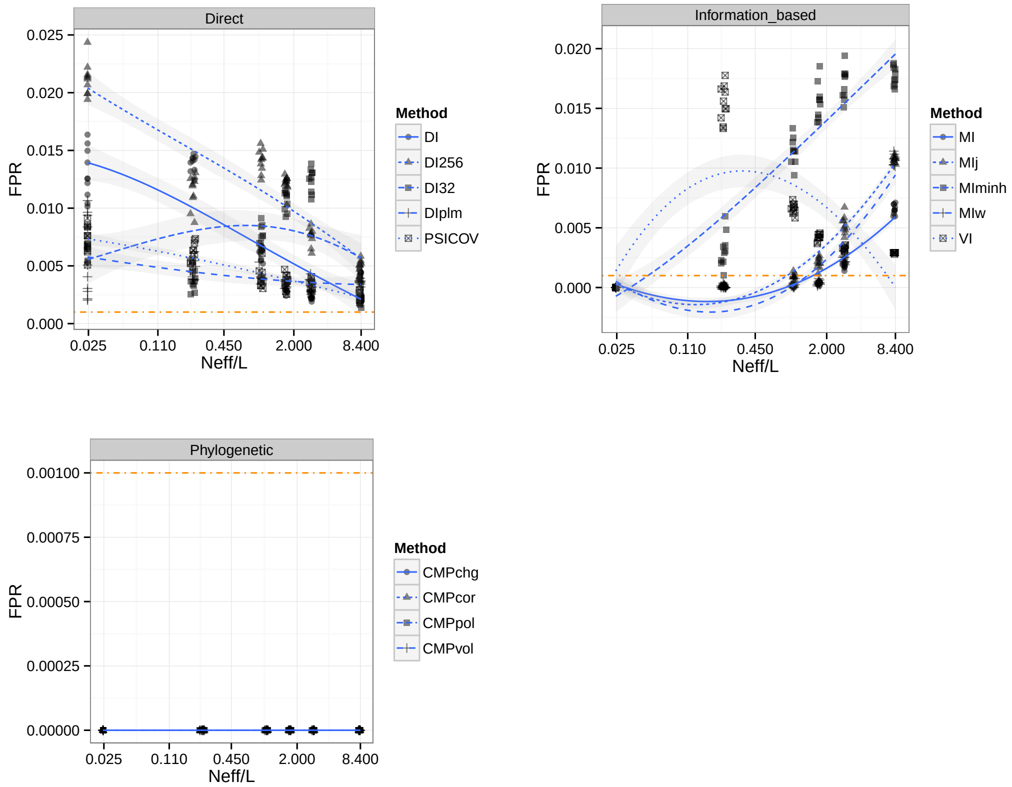

Figure 2.18: HisKA-RR with contacts defined using spatial proximity ($C_\beta < 8$Å). False positive rate (FPR) at $P_{normal} < 0.001$. 95% confidence intervals for loess smooths are shown in gray

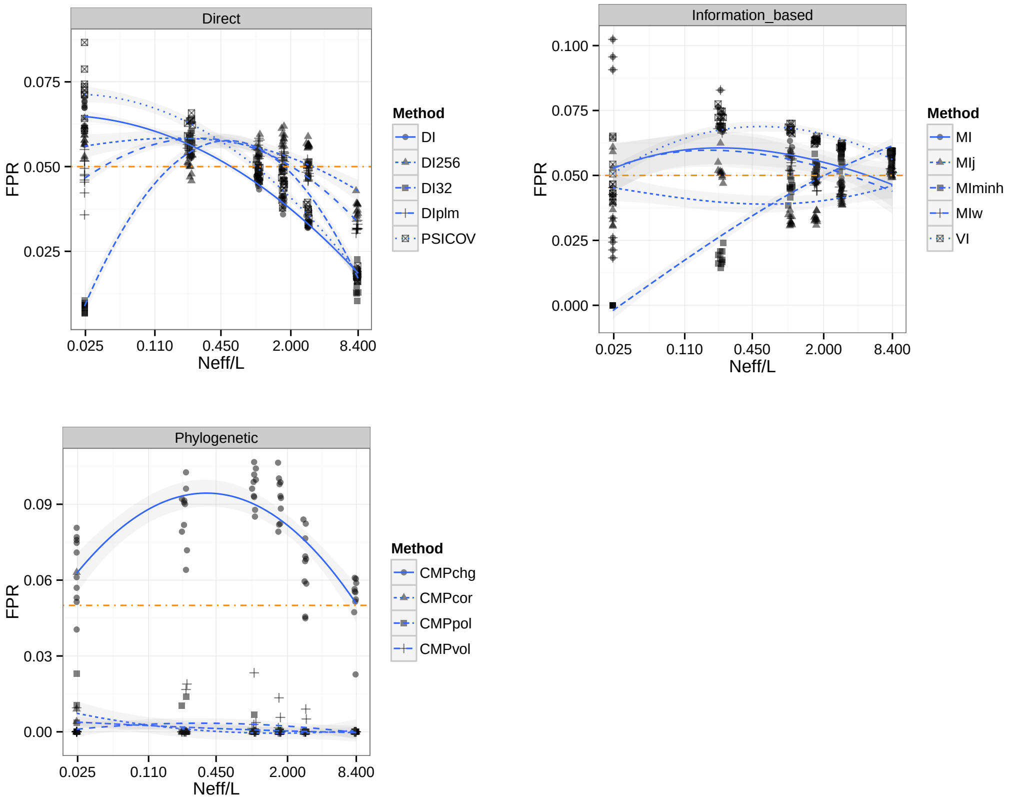

Figure 2.19: HisKA-RR with contacts defined using spatial proximity ($C_\beta <$ 8Å). False positive rate (FPR) at $P_{normal} < 0.05$. 95% confidence intervals for loess smooths are shown in gray

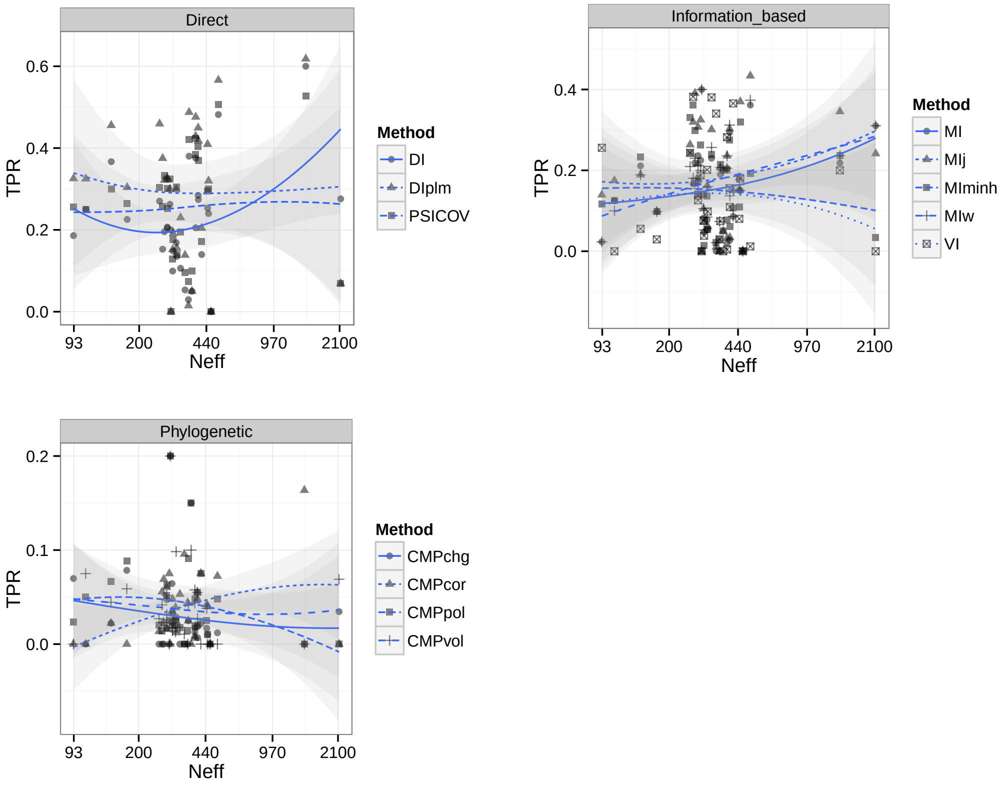

Figure 2.20: Ovch32 alignments [58] with contacts defined using spatial proximity ($C_\beta < 8$Å). Power (TPR) at FPR $< 5\%$. 95% confidence intervals for loess smooths are shown in gray

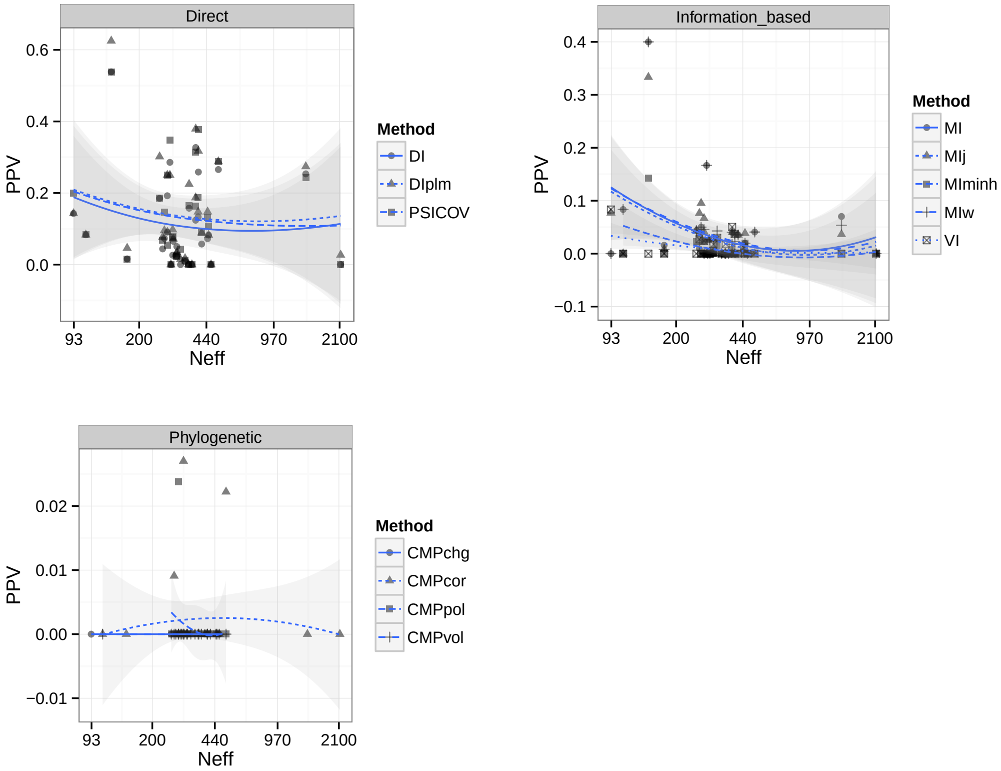

Figure 2.21: Ovch32 alignments [58] with contacts defined using spatial proximity ($C_\beta < 8$Å). Precision (PPV) at FPR $< 0.1\%$. 95% confidence intervals for loess smooths are shown in gray

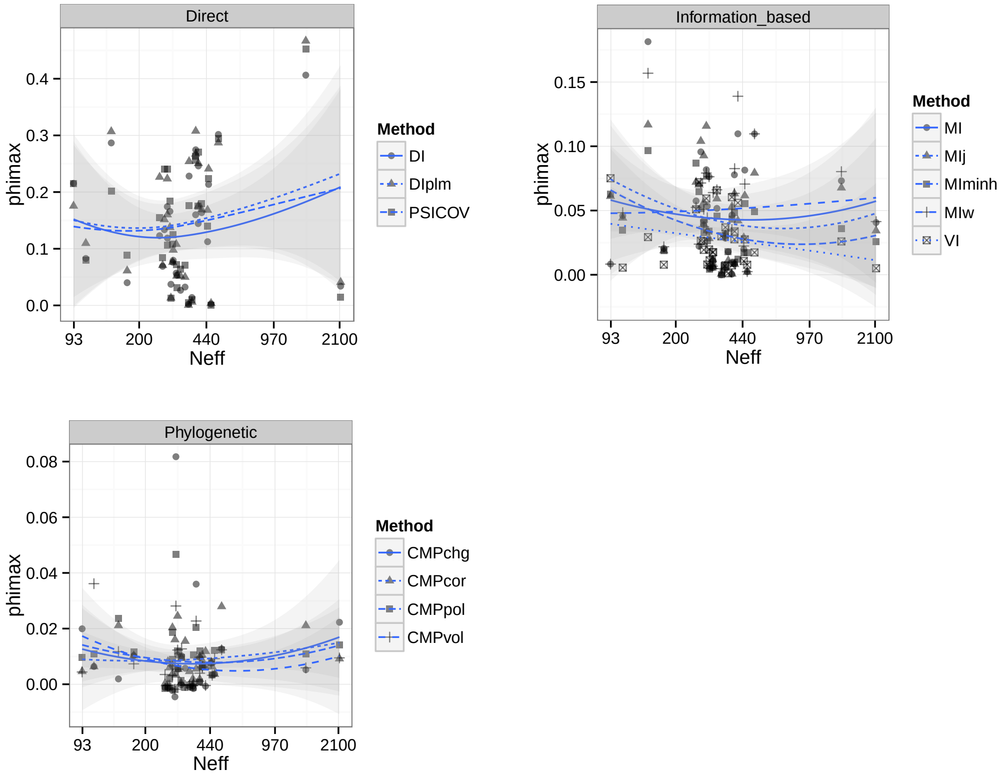

Figure 2.22: Ovch32 alignments [58] with contacts defined using spatial proximity ($C_\beta < 8$Å). Phi$_{\text{max}}$. 95% confidence intervals for loess smooths are shown in gray

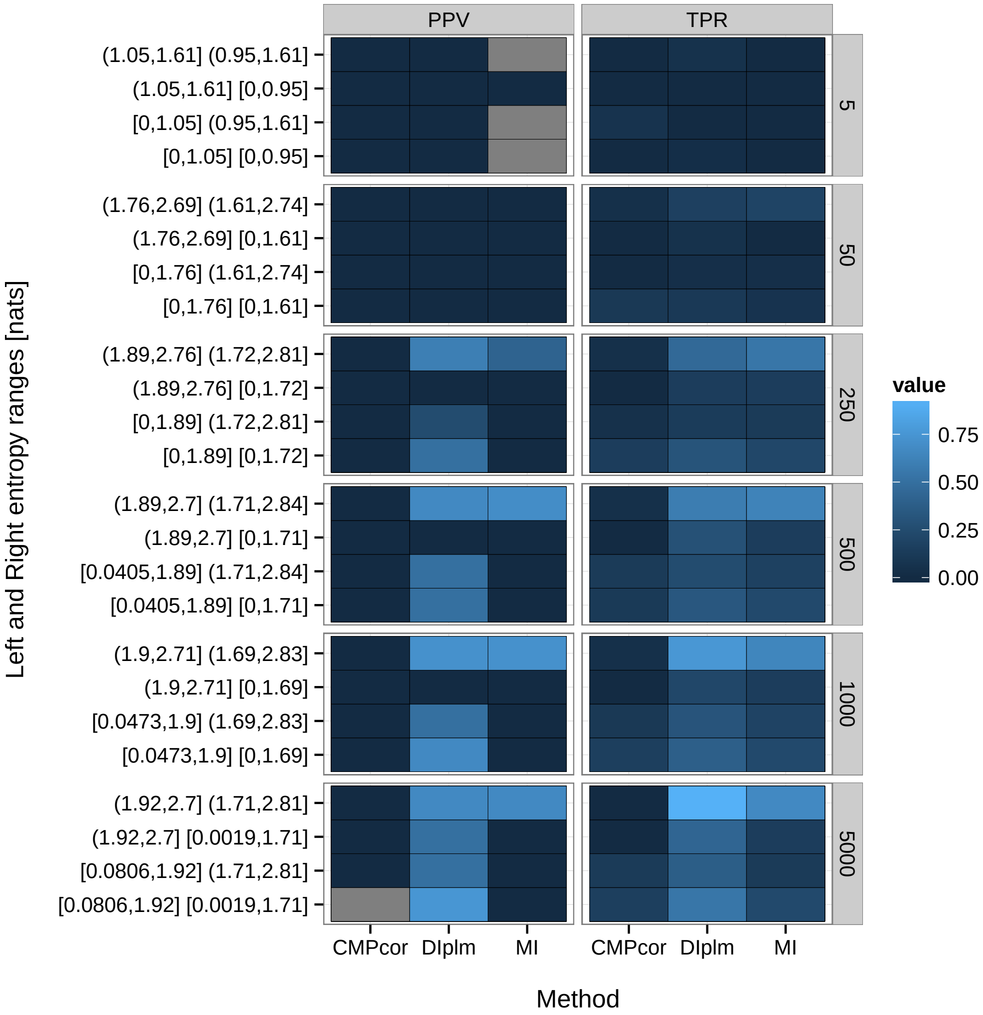

Figure 2.23: HisKA-RR with contacts defined using spatial proximity (C_β < 8Å). Performance is evaluated on subsets of alignment pairs with above and below median entropy for each HisKA and RR column. Power (TPR) at FPR $< 5\%$ and Precision (PPV) at FPR $< 0.1\%$ are shown for CMP_{cor}, DI_{plm}, and MI.

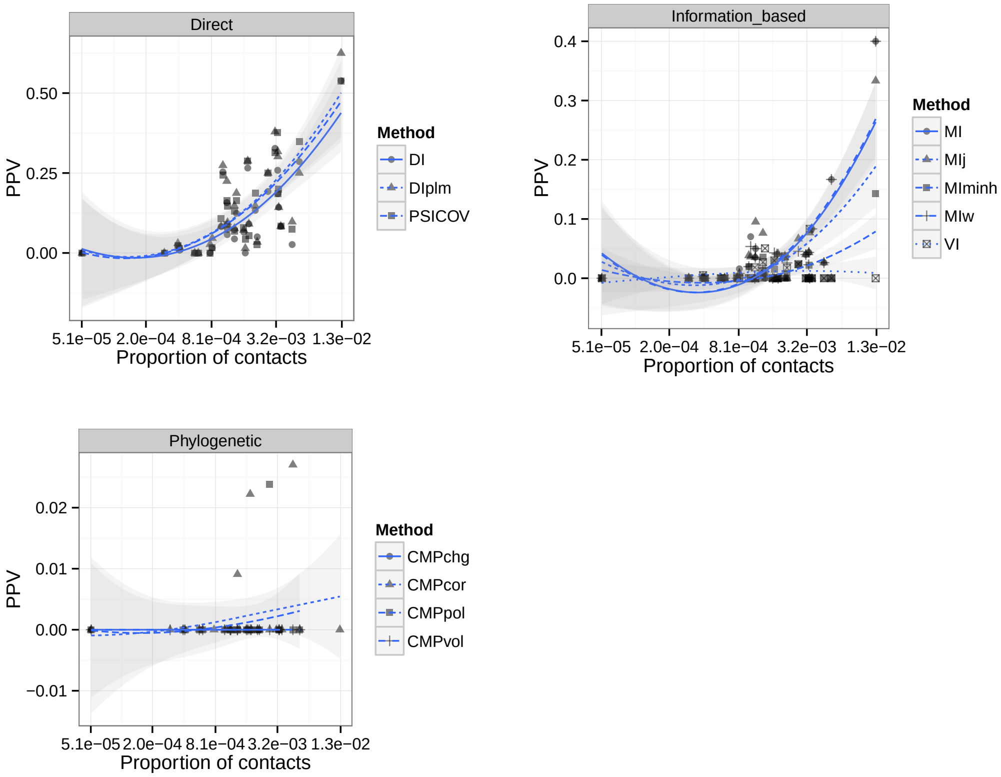

Figure 2.24: HisKA-RR with contacts defined using spatial proximity (C_β $<$ 8Å). Precision (PPV) at FPR $<$ 0.1% 95% confidence intervals for loess smooths are shown in gray

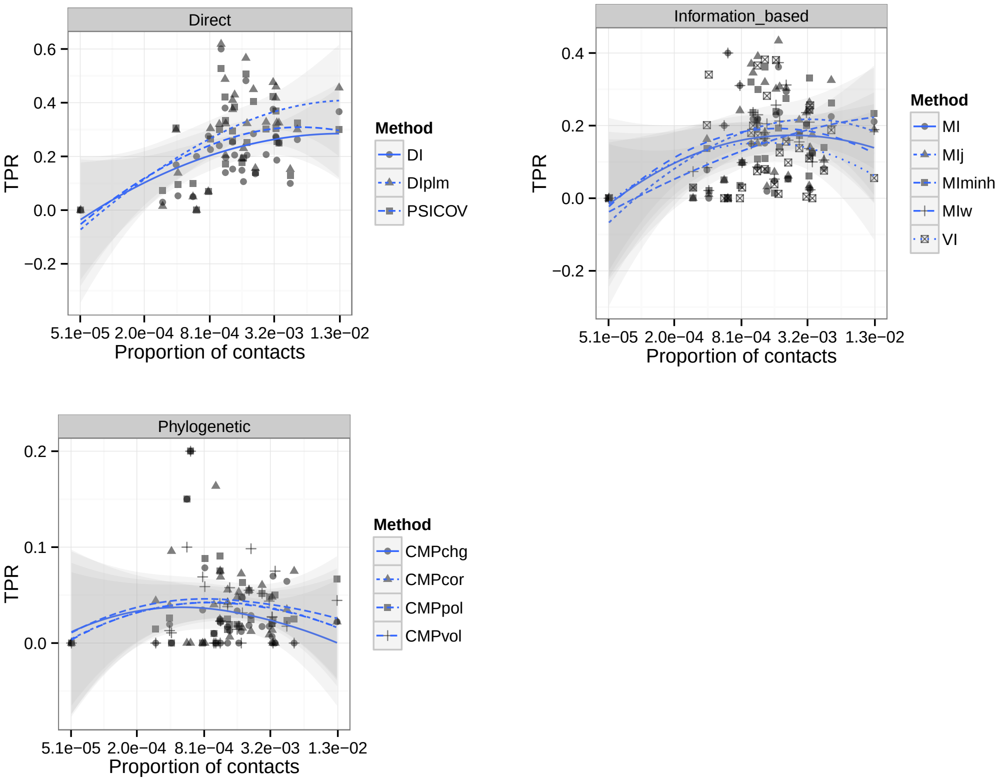

Figure 2.25: HisKA-RR with contacts defined using spatial proximity ($C_\beta <$ 8Å). Power (TPR) at FPR $<$ 5% 95% confidence intervals for loess smooths are shown in gray

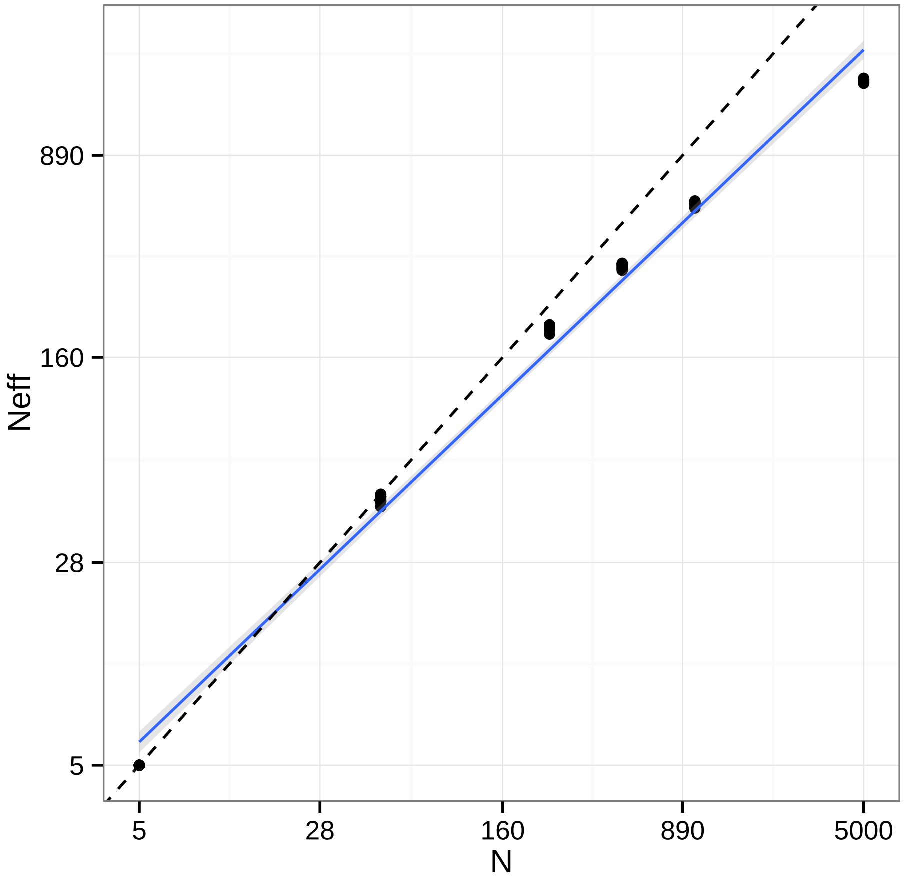

Figure 2.26: HisKA-RR. N_{eff} versus N for 60 sub-sampled alignments. 95% confidence intervals for loess smooths are shown in gray

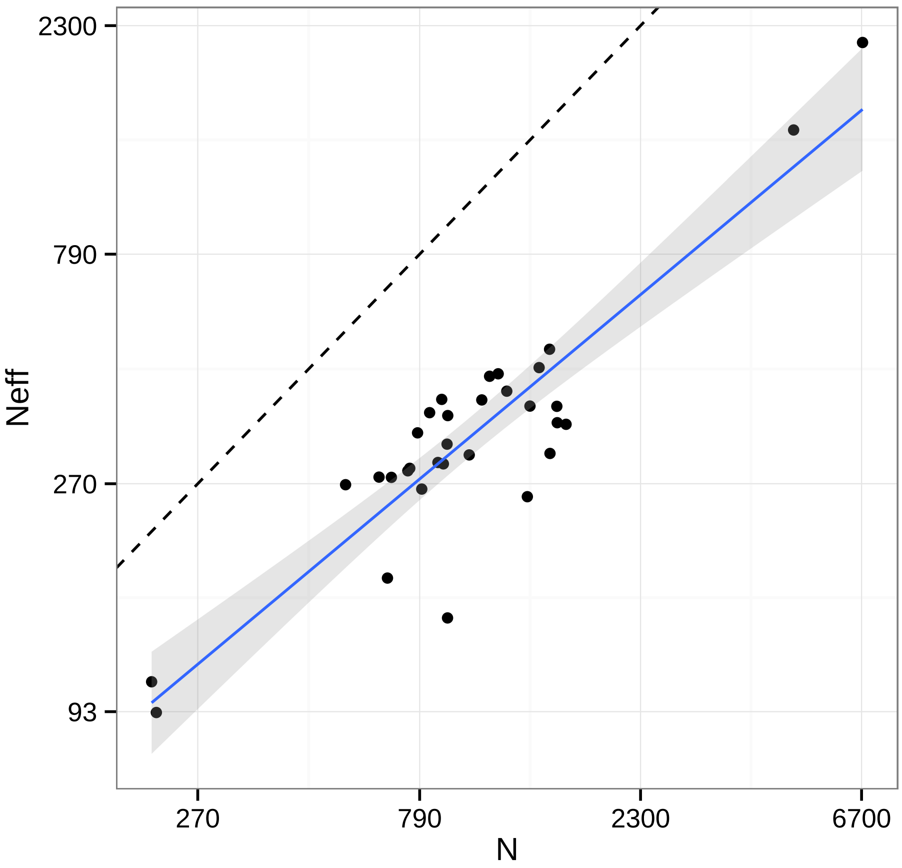

Figure 2.27: Ovch32. N_{eff} versus N for 32 alignments in [58]. 95% confidence intervals for loess smooths are shown in gray

2.5 Tables

	Method	APC	Re-weighting	Reference	Software package
Information-based	MI	No	None	[66, 19]	infCalc
	VI			[54]	
	MI_j			[19]	
	$\mathrm{MI}_\mathrm{Hmin}$				
	MI_w		seq %id	[56]	DCA
Direct	DI	Yes	seq %id, pseudocount		
	DI_{256}			[14]	Code S1 in [14]
	DI_{32}				
	DI_plm		seq %id	[24]	plmDCA
	PSICOV		Blosum, pseudocount	[39]	PSICOV
Phylogenetic	$\mathrm{CMP}_\mathrm{cor}$	No	Downsampling	[22]	CoMap
	$\mathrm{CMP}_\mathrm{chg}$			[21]	
	$\mathrm{CMP}_\mathrm{vol}$				
	$\mathrm{CMP}_\mathrm{pol}$				

Table 2.1: Coevolution methods benchmarked fall into three categories. **Information-based methods:** MI: mutual information [66], VI: variation of information [54], MI_j: MI divided by alignment column-pair entropy, $\mathrm{MI}_\mathrm{Hmin}$: MI divided by minimum column entropy [19], MI_w: MI with adjusted amino acid probabilities. **Direct methods:** DI: direct information—MI with re-estimated joint probabilities [56], DI_{256}, DI_{32}: DI using Hopfield-Potts for dimensional reduction (256 and 32 patterns respectively) [14], DI_plm: Frobenius norm of coupling matrices in 21-state Potts model using pseudolikelihood maximization [24], PSICOV: sparse inverse covariance estimation [39]. **Phylogenetic methods:** CoMap *P*-values for four analyses $\mathrm{CMP}_\mathrm{cor}$: substitution correlation analysis [22], $\mathrm{CMP}_\mathrm{pol}$ for polarity compensation, $\mathrm{CMP}_\mathrm{chg}$ for charge compensation, $\mathrm{CMP}_\mathrm{vol}$ for volume compensation [21].

	HisKA-RR		Ovch32	
Metric	ρ	P	ρ	P
auPR	0.554336	7.70407e-69	-0.2055977	6.47389e-05
auROC	0.519496	3.06530e-59	-0.0128938	8.04243e-01
$\mathrm{f}_{\max}$	0.537467	4.73041e-64	-0.1890605	2.45104e-04
$\phi_{\max}$	0.525671	7.33669e-61	-0.0556136	2.84684e-01

Table 2.2: Spearman correlations of threshold-independent metrics with $\mathrm{N}_{\mathrm{eff}}$

	HisKA-RR		Ovch32	
Metric	ρ	P	ρ	P
auPR	0.554336	7.70407e-69	0.162724	1.63855e-03
auROC	0.519496	3.06530e-59	0.170879	9.36027e-04
$\mathrm{f}_{\max}$	0.537467	4.73041e-64	0.127323	1.39922e-02
$\phi_{\max}$	0.525671	7.33669e-61	0.202785	8.18168e-05

Table 2.3: Spearman correlations of threshold-independent metrics with $\mathrm{N}_{\mathrm{eff}}/\mathrm{L}$

Metric	FPR	ρ	P
PPV	0.001	0.337364	3.80652e-10
	0.05	0.585178	1.81358e-35
TPR	0.001	0.202065	8.68295e-05
	0.05	0.209154	4.79314e-05

Table 2.4: Spearman correlations of Power (TPR) and Precision (PPV) with the proportion of contacts in an interaction.

	Prediction	
C_β distance	Coevolving	Not coevolving
< 8Å	TP	FN
≥ 8Å	FP	TN

Table 2.5: Confusion matrix.

	Method	Software package	Version	URL
Information-based	MI	infCalc	v0.1.2	https://github.com/aavilahe/infcalc
	VI			
	$\mathrm{MI_j}$			
	$\mathrm{MI_{Hmin}}$			
	$\mathrm{MI_w}$	DCA	”2011/12”	http://dca.ucsd.edu/DCA/DCA.html
Direct	DI			
	$\mathrm{DI_{256}}$	Code S1 in [59]	”2013”	http://doi.org/10.1371/journal.pcbi.1003176.s002
	$\mathrm{DI_{32}}$			
	$\mathrm{DI_{plm}}$	plmDCA	symmetric_v2	http://plmdca.csc.kth.se/
	PSICOV	PSICOV	V1.09	http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/
Phylogenetic	$\mathrm{CMP_{cor}}$	CoMap	1.5.1b5	http://home.gna.org/comap/doc/html/index.html
	$\mathrm{CMP_{chg}}$			
	$\mathrm{CMP_{vol}}$			
	$\mathrm{CMP_{pol}}$			

Table 2.6: Versions and sources of coevolution methods benchmarked.

Chapter 3

Null distributions

Commonly used null distributions that are parameterized by the observed scores can vary in their ability to control the false positive rate (Section 2.2.2). An evolutionary model under the null hypothesis (proteins are evolving independently) can potentially be used to generate realistic null distributions. In this chapter I present a semi-parametric bootstrap method for deriving null distributions of coevolution scores for individual residue-pairs in an inter-protein coevolution analysis. I find that this method, while well motivated, fails to control the false positive rate when fast or slow evolving alignment columns are inaccurately simulated.

An analytic method based on G-tests and an empirical method based on non-interactors are proposed for estimating P-values.

3.1 Background

In hypothesis testing, a null distribution of a test statistic is necessary in judging whether there is enough evidence to reject a null hypothesis. In the case of predicting interprotein contacts from coevolution scores, we are concerned with rejecting the hypothesis that a pair of residues are not contacts.

To do so I compared the coevolution score to a (null) distribution of scores we expect to measure in non-contacts and quantify it in a P-value—the probability that the null distribution would generate a score as or more extreme than the observed value. However, assumptions on what shape and scale that distribution should have affect the performance in predicting contacts. For example, an overly conservative null distribution may assume that non-contacts are likely to have scores as high as contacting residues, while an anti-conservative null would assume non-contacts have lower scores than they truly do.

An ideal null should account for shared phylogeny of the sequences in an alignment, finite sample effects, and evolutionary rates of individual alignment columns. However, these can be difficult and computationally expensive to implement and simpler, faster alternatives are common in practice.

In Section 2.1.4 I employed two common null distributions in the estimation of coevolution P-values. While these are computationally inexpensive, I showed that they suffer from elevated rates of false positives, likely due to shared phylogeny or structural constraints affecting non-contacting residue pairs.

Bootstrap methods (e.g. in [22], [76]), though more computationally intensive, have been used to generate null distributions of coevolution scores by simulating independent evolution. Though these methods theoretically account for shared phylogeny (by maintaining the phylogenetic relationships between the sequences) and give a null distribution for each alignment-column pair or class of column-pairs, they have typically been employed in *intra*molecular analyses.

A non-parametric bootstrap method—recently used in an interprotein analysis in [50]—shuffles individual alignment columns tens of thousands of times. This shuffling method is faster than semi-parametric simulations, conserves the entropy and composition of the observed columns, however it does not preserve the relationships between the sequences, potentially leading to null distributions that are overly anti-conservative.

As a middle ground, theoretical distributions that account for sequence variation in individual column-pairs could be used to generate null distributions of scores for each column-pair without the need for expensive simulations.

3.2 P-values from simulated alignments ($P_{bootstrap}$)

This approach aims to independently simulate alignments for each protein family (group of orthologous proteins) in an inter-protein analysis such that the simulated alignments resemble the observed alignments in terms of substitution rates, amino acid composition, and phylogenetic relationships. However, substitutions are generated independently for each protein family and are therefore not correlated beyond any correlation induced by similarities in the phylogenies of the two gene families. This should directly account for phylogenetic effects in the null distribution and therefore have the potential to more accurately control FPRs. Simulations are computationally intensive and not suitable for all methods as they can greatly increase computational time. To explore the possibility of using this approach with current coevolution methods, I implemented a semi-parametric bootstrap null distribution for a subset of phylogeny-unaware methods using the HisKA-RR sub-alignments used in Chapter 2.

3.2.1 Simulating independently evolving pairs of alignments

In order to classify pairs of sites as coevolving or not coevolving using a semi-parametric bootstrapped null distribution, I calculated a P-value ($P_{bootstrap}$) for the score at every pair of positions by comparing the observed score to the distribution of scores simulated for that pair under the null hypothesis (independent evolution).

To simulate alignments, I used FastTree (version 2.1.7 SSE3) [61] to build maximum likelihood phylogenetic trees for the HisKA and RR protein families. I used `hmmbuild` from the HMMER3 package [23] (version 3.0 March 2010) to build a profile hidden Markov model (pHMM) for each family. I sampled amino acid residues from a first order Markov chain to generate an initial sequence for each family. Finally, I used Revolver (version 1.0) [43] to simulate 1000 8998-sequence alignments for each family independently. Revolver can simulate the evolution of a given root sequence that adheres to the domain constraints imposed by a pHMM, and preserves a similar phylogenetic history to the observed alignment. Revolver used the WAG substitution matrix and indel probabilities were set to zero in order to simulate constant length alignments. Gaps from the observed alignment were then overlaid on the simulated alignment. I automated this process in a pipeline available at https://github.com/aavilahe/simulate_tools.

3.2.1.1 Example simulation of Response Regulator (RR) alignment with `simulate_tools`:

First a phylogenetic tree for the RR alignment was built using FastTree (version 2.1.7 SSE3) with options `-gamma -nosupport -wag` and saved in `RR.tree`.

The following steps were then automated in the `simulate_tools` pipeline (`runSimAli`):

1. Build profile HMM
2. Sample starting "root sequence" for simulation using first order Markov chain
3. Generate xml control file for Revolver
    - No tree scaling
    - Heterogeneous rates (alpha = 1, ncats = 9)
    - No indels

4. Run Revolver

```
runSimAli --tree RR.tree      \
          --outdir /path/to/output \
          --num_sims 1000 JobNameRR RR.phy
```

From these simulated master alignments, sequences corresponding to the observed sub-alignments were extracted to create a total of 60,000 sub-alignments, each corresponding to one of the original 60 observed sub-alignments.

3.2.2 False positive rate

Unfortunately, I found that *P*-values calculated using the bootstrap null distribution were heavily influenced by the error in simulating alignment columns with appropriate amino acid variation. Residue pairs for which the bootstrap simulated alignment columns have too much sequence variation (compared to the corresponding observed alignment column) tend to have small *P*-values, regardless of whether or not they are contacting residues. Simulation error increased with alignment size, as did nominal FPRs. Consequently, at a target FPR of 5%, the nominal FPR was not adequately controlled for alignments with more than 5 sequences ($\mathrm{N}_{\text{eff}}/\mathrm{L} = 0.02$) for any method except PSICOV. Interestingly, PSICOV is the method least affected by the simulation error.

Recalculating the nominal FPR using only alignment column pairs that were moderately well simulated (such that no more than 25% simulations were over or under conserved) showed much lower FPR for all methods except $\mathrm{MI}_{\mathrm{Hmin}}$ (Figure 3.1). MI and VI are controlled below a target FPR $< 5\%$. At a stricter target FPR $< 0.1\%$, PSICOV, MI, and VI are the only methods with completely controlled FPR at all alignment sizes. MI_{w}, DI, and MI_{j} are controlled in alignments with fewer than 1000, 500, 250 sequences respectively. Together these results suggest that the DI, MI_{w}, MI_{j}, and VI are sensitive to the amount of variation in alignments, while PSICOV and $\mathrm{MI}_{\mathrm{Hmin}}$ are more robust to that variation. This is important in directly comparing fast or slowly evolving column-pairs and in comparing observed scores to simulated null scores.

$\mathrm{MI}_{\mathrm{Hmin}}$'s higher FPR suggests it is identifying coevolving residues that are not structurally close. Perhaps some of them may be part of an alternate network of evolutionarily important residues, for example "protein sectors [53]" that span more than one protein. Or perhaps its null scores are simply distributed unhelpfully.

3.2.2.1 Comparison with CoMap

CoMap internally estimates P-values using a similar simulation approach. Nominal FPRs for CoMap methods, using their P-values directly, resemble those of the *Information-based* approaches using the normal distribution as a null (twice to 20 times the target FPR) (Figures 3.2 and 3.3). I conclude that it is very important for the evolutionary conservation of alignment columns in the bootstrap null distribution to closely match conservation levels in the observed data. Despite using currently accepted techniques for generating bootstrap distributions, I found that matching conservation levels this closely is challenging. This is an important problem for future research in the coevolution field.

3.3 Non-central gamma distributed null (P_{gamma})

In addition to having a single null distribution for all pairs of sites, theoretical nulls can be estimated for each pair of sites, taking into account properties of the alignment columns involved. However, the theoretical null may be unique to the coevolution statistic used and may not even have a closed-form expression. The non-central gamma distribution as derived by [29] can be used for estimating P-values (P_{gamma}) for MI and related methods due to their relationship with G-tests. Here the shape and scaling parameters depend on the number of observations (e.g. number of sequences in alignment) and number of realizations of the two categorical variables (e.g. number of different residues and gaps), and the non-centrality parameter is used to specify MI under the null hypothesis. This is an interesting area for future methods research.

3.4 Empirical P-values revisited

An alternative to independently simulating alignments is to permute the connections in an interaction network and estimate a null distribution using scores from non-interacting proteins. The permutation approach is similar to $P_{empirical}$ in Section 2.1.4 except that by generating pairs of non-interactions, it ensures that all of the null scores are generated according to the null hypothesis—an assumption that is generally not true in $P_{empirical}$.

Using an empirical distribution from non-interacting proteins could prove useful in detecting pairs of interacting proteins from non-interactors in the case where contacting proteins share many weak coevolution signals as opposed to a few strong signals.

An implementation is demonstrated in Chapter 4.

3.5 Figures

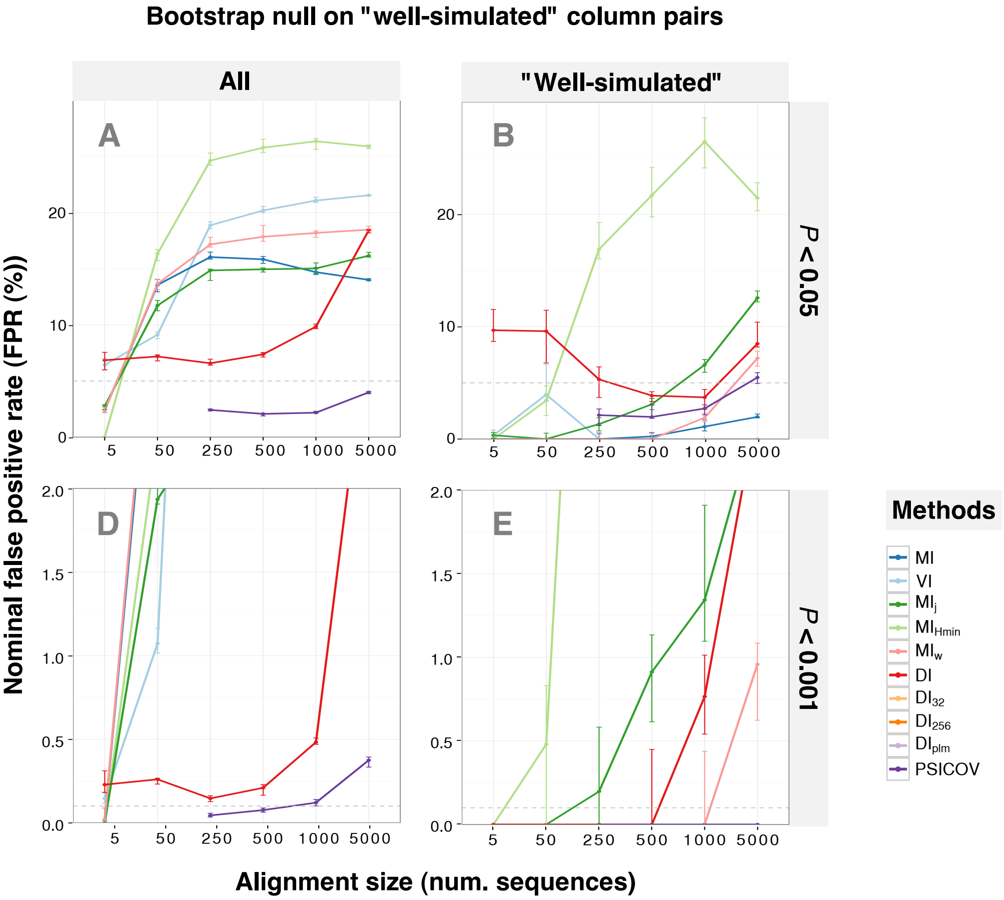

Figure 3.1: HisKA-RR with contacts defined using spatial proximity (C_β < 8Å). $P_{boostrap}$ fails to control the FPR except for in PSICOV at target FPR $< 5\%$ in HisKA-RR alignments. Eliminating residue pairs with large simulation errors shows PSICOV and $\mathrm{MI_{Hmin}}$ are most robust to sequence variation differences across sites.

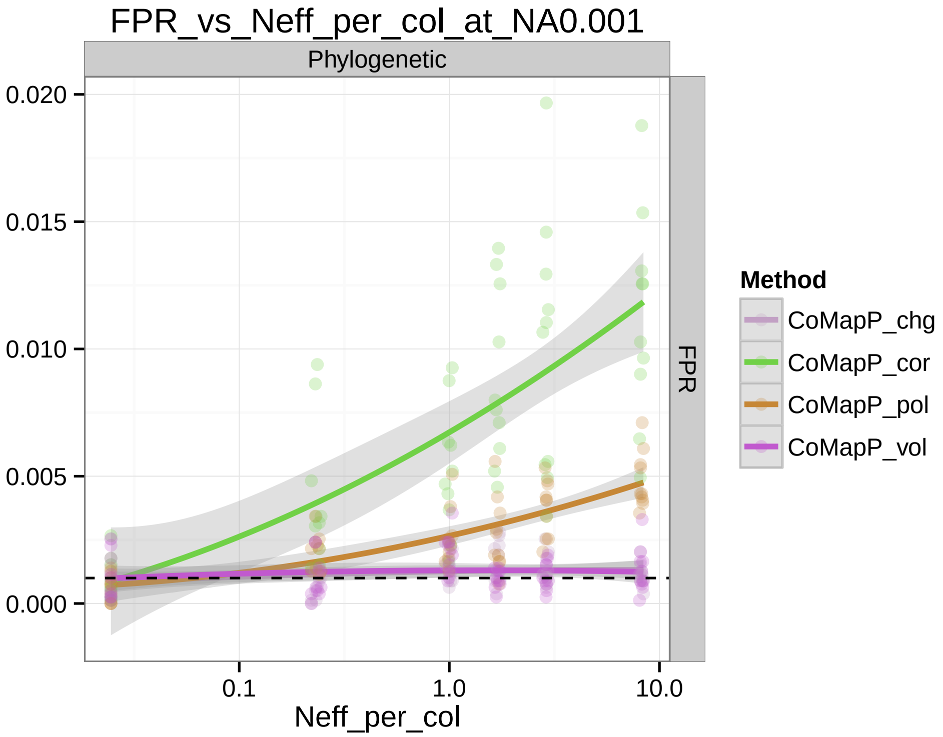

Figure 3.2: HisKA-RR with contacts define using spatial proximity. Nominal false positive rate (FPR) at $P < 0.05$ using CoMap's internal P-values

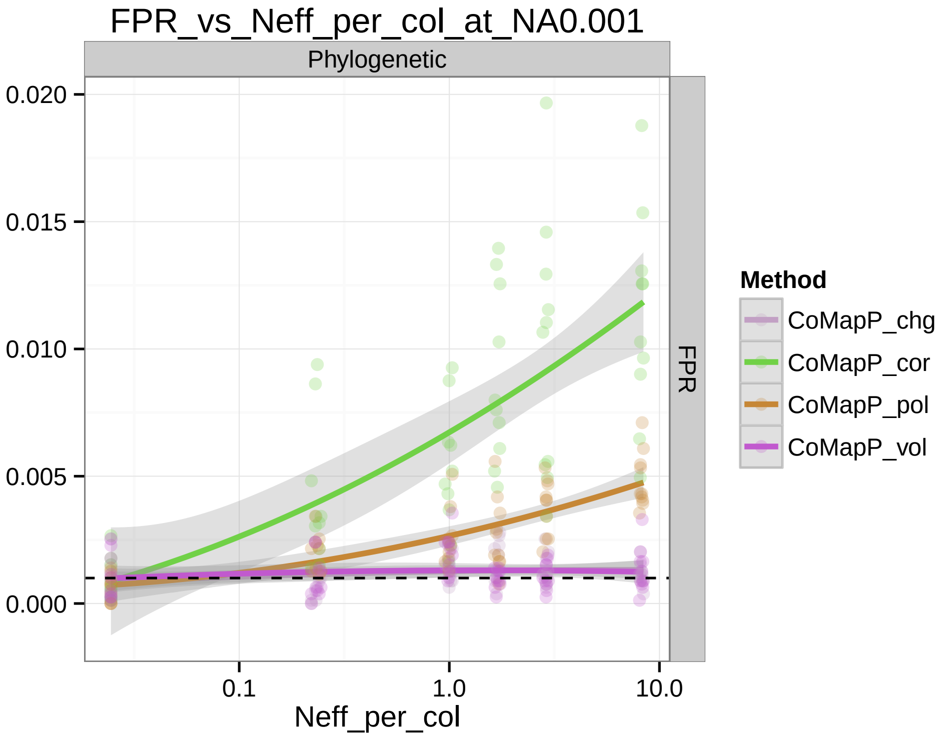

Figure 3.3: HisKA-RR with contacts define using spatial proximity. Nominal false positive rate (FPR) at $P < 0.001$ using CoMap's internal P-values

Chapter 4

Coevolution analyses of networks and cross-species interactions

Coevolution analyses of cross-species protein interactions potentially reveals how tightly linked organisms adapt to each other at the molecular level. Microbes interact with the human gut, plant roots cooperate with mycorrhizae, and viruses rewire host cells for their reproduction. In this chapter I apply coevolution methods benchmarked in Chapter 2 to measure the strength of coevolution in HIV1-human interactions. Then I focus on a particular arms-race between HIV1 protein Vif and its antagonist human A3G.

4.1 Background

The ability to now measure physical interactions between biomolecules with high-throughput technologies, such as affinity purification followed by mass spectrometry (APMS) [57], two-hybrid methods [6], [73], and protein complementation assays [55], raises the possibility of using sequence coevolution to refine predicted interactions in an experimentally reduced search space. For example, correlated substitution patterns in pairs of proteins could help determine if an experimentally measured interaction is likely to represent direct physical contact versus an indirect interaction in a complex or a false positive. Coevolutionary analysis could also be informative regarding which of the sites in a pair of interacting molecules are most likely to be in physical contact. Although, many coevolution studies have been conducted on HIV-human interactomes [1], [46], [82], this is the first to my knowledge that directly aims to measure coevolution between the lentiviral and mammal interactions.

Although the signal-to-noise ratio is too low and the search space too large to use sequence coevolution to effectively identify pairs of physically interacting protein residues across entire proteomes—most pairs of sites with correlated substitution patterns are not in direct contact, and most physically interacting sites do not have statistically correlated substitution patterns [78]—I applied my integrated framework for coevolutionary analysis (See Chapter 5) to refine and annotate a recently derived human-HIV1 protein-protein interaction network [37] and to test for coevolution in the well studied arms-race interaction between the mammalian cytidine deaminase APOBEC3G (A3G) and its HIV1 antagonist, Vif. Because fewer than ten orthologous mammal-lentivirus proteome pairs have been sequenced and mammalian divergence is low, I hypothesized that power would be low in these settings.

4.2 The interaction network of HIV and human proteins shows only weak evidence of coevolution across mammals

I sought to use inter-protein residue coevolution to refine a recently derived APMS protein-protein interaction network of the HIV-human interactome [37]. This study detected human proteins that interact with each HIV protein, either via direct physical contact or as members of complexes. Specifically, I hoped to use evidence of sequence coevolution to resolve direct versus indirect protein interactions amongst all human proteins measured to interact with each HIV protein. Secondly, I wanted to know if coevolutionary signals are strong enough to pinpoint key residues involved in the interfaces of any direct interactions.

For each protein in the HIV genome (nine polyproteins and an additional nine protease products), I computed a multiple sequence alignment with as many sequenced immunodeficiency viruses that infect mammals with sequenced genomes. I downloaded viral proteomes from Uniprot [72] and computationally processed the polyproteins for unannotated viruses.

Similarly, I leveraged a set multiple alignments of each human protein (22,947 CCDS records) with the sequences of its orthologs from any mammal with a sequenced immunodeficiency virus [49].

4.2.1 Affinity purification mass spectrometry (APMS) interactors

Together, this produced 425 pairs of host-virus protein alignments (out of the original 497 detected interactions) with up to eight immunodeficiency viruses and their primate, feline, and bovidae hosts ($<\mathrm{N}> = 7.35$, $<\mathrm{N}_{\mathrm{eff}}> = 5.39$). In an attempt to reduce the search space for high scoring interactions, I filtered out alignment columns with little variation (keeping no more than the top 300 most variable columns in the joined alignment). This step is especially necessary for the *Direct* methods, such as DI, to fit in RAM as they invert a $20 \cdot \mathrm{L}$ by $20 \cdot \mathrm{L}$ matrix as part of their algorithm. However, even with filtering, $\mathrm{N}_{\mathrm{eff}}/\mathrm{L}$ for the interactors is especially low. More than half of the interactors have lower $\mathrm{N}_{\mathrm{eff}}/\mathrm{L}$ than observed in either the HisKA-RR data set or Ovch32 in Chapter 2 Figure 4.1.

4.2.2 Empirical null distributions

I created two sets of null interaction networks. The first network is constructed by permuting the interactions such that HIV and human proteins in the original interactome are mispaired. The other is by randomly pairing human proteins not observed in the APMS interactome with HIV proteins.

Furthermore, these null networks can be used to generate a single null distribution using all null interactions, or separate null distributions per each virus protein. Coevolution methods are applied to the nulls exactly as to the APMS interactors to generate null score distributions. P-values are named $P^{n}{}_{empirical}$ for the non-interactor null and $P^{p}{}_{empirical}$ for the permuted null.

The permuted network consists of 382 paired alignments. And the non-interactor null consists of 400 alignments (alignments with fewer than 5 pairing sequences were discarded). The nulls had slightly lower $\mathrm{N_{eff}/L}$ than the median in the APMS interactors Table 4.1, Figure 4.1.

Since $\mathrm{N_{eff}/L}$ can have an effect on the distribution of scores (e.g. it may indirectly set an upper bound and induce discreteness in MI when $\mathrm{N} \ll 400$), it is important to either regularize the scores or ensure that the null and observed alignments are matched in terms of conservation, depth, and diversity.

4.2.3 Simulated null distributions

In my first attempt at understanding coevolutionary relationships between lentiviral and mammal proteins, I quantified coevolution using $\mathrm{MI_j}$ and used a $P_{bootstrap} < 0.001$ cutoff to predict coevolving residue pairs. For each protein pair, I varied the significance threshold and computed the count of significantly coevolving residue-pairs. I then compared this statistic for interacting protein pairs from the APMS network versus a control set of 100 randomly chosen lentivirus-mammal protein pairs not included in the APMS network.

I found that APMS detected interactions have only marginally more counts of significant signals of coevolution compared to non-interactions (best auROC = 0.541 at $P_{bootstrap} < 0.001$), and therefore counts of coevolving residues are not sensitive enough to distinguish direct interactions or the residues involved in them for this set of virus and host proteins.

Initially, it was unclear if this was due to low power from lack of sequence depth, a poor choice of coeovlution statistic, or that there was no detectable coevolutionary signals in the substitution patterns between lentivirusal and mammal interactors.

Based on my subsequent benchmarking in Chapter 2, I concluded that this lack of signal may result from low power in this setting. However, the results from Section 3.2.1: Simulating independently evolving pairs of alignments suggest that false positives in the null could be drowning a coevolution signal and motivated the empirical approach with non-interacting and permuted nulls.

I ran five *Information-based* methods (MI, $\mathrm{MI_w}$, $\mathrm{MI_{Hmin}}$, $\mathrm{MI_j}$, and VI) and a *Direct* method DI. The results with the empirical null distributions corroborate the earlier simulation results, establishing that without more sophisticated models and null distributions, coevolution problems with small $\mathrm{N_{eff}}/\mathrm{L}$ will remain out of reach (Figure 4.2).

4.3 Coevolution methods identify some residues known to affect host-virus interactions in Vif-A3G

Viral infectivity factor (Vif) is a lentiviral accessory protein whose primary function is to target the antiviral cytidine deaminase APOBEC3G (A3G) of its mammalian hosts through ubiquitination. Because the two protein families are in an evolutionary arms race [16], [15], I hypothesized that they would be an informative example for exploring the utility of coevolution methods in host-virus protein pairs (i.e. inter-protein, inter-species interactions). This is a novel application of coevolution analysis, which has primarily been applied to residues within a protein or between pairs of proteins in the same genome.

A major challenge in performing coevolutionary analysis on cross-species protein pairs is acquiring appropriate data, including paired alignments and protein structures for validation. For Vif-A3G, I was able to identify 16 pairs of sequences ($\mathrm{N_{eff}} = 10.0$) from different primates (A3G orthologs) and their lentiviruses (Vif orthologs) in public databases (Table 4.2). My benchmarking results on HisKA-RR indicate that such small protein families push the useful limits of the coevolution statistics I tested ($\mathrm{N_{eff}/L} = 0.014$). The low sequence diversity of A3G ($\mathrm{N_{eff}} = 3.04$) within primates compared to Vif ($\mathrm{N_{eff}} = 11.3$) within primate lentiviruses also presents challenges. Hence, I expect coevolutionary analysis to potentially have limited power in this scenario. To quantitatively evaluate performance requires validated Vif-A3G interactions.

The structure of Vif in complex with A3G has not been solved. However, biochemical assays have solidly identified regions important for binding and ubiquitination along the individual reference sequences of HIV1 Vif [11], [63], [80], [34] and human A3G [81], [64] (Table 4.3). For this analysis, I therefore take the residues in biochemically-validated regions to be positives even though they might not be contacts (i.e. C_β distance $\geq$ 8Å), and assume that all remaining residues are negatives, even though other sites (including sites deleted in these reference sequences) are possibly involved in the interaction. While further experimentation is needed to understand the relationship between functionally important sites and the structure of the protein interaction, as well as the effects of mutations in these sites on the fitness of lentiviruses, I explore whether any clues can be identified in the limited data that describes the coevolutionary history of the Vif-A3G residues.

First, I computed coevolutionary statistics for all Vif-A3G residue pairs and evaluated how well the statistics pinpoint the positive functionally important residues compared to negatives. For this evaluation, I used the empirical distribution of scores as a null distribution to determine statistical significance (i.e., $P_{empirical}$) because they have lower false positive rates across N_{eff}/L values at strict significance thresholds. Because the positives and negatives are single residues in each sequence instead of inter-protein residue pairs, I summarized $P_{empirical}$ for each residue by assigning it the most significant $P_{empirical}$ across all inter-protein pairs to which it belongs, and then explored the Vif and A3G results individually. From my benchmarking on the bacterial data sets, we know that significance thresholds that control the FPR vary by method and N_{eff}/L, and that strict thresholds that yield very low ($\sim$ 2–3%) power are typically needed to control FPR in small alignments. I therefore chose to identify a significance threshold for each method that maximizes precision on the known functional sites in each protein. Then, I estimated power and FPR at these thresholds.

On Vif, with the exception of CMP_{cor} and DI_{32}, the maximum precisions for each method ranged from 9 to 20% (i.e. only one or two residues out of ten predicted to be positives are truly positives) (Figure 4.20). At these precision-optimized thresholds, MI_{j} and MI_{minh} predict almost every Vif residue to be coevolving; a stricter threshold would not result in a lower proportion of incorrect predictions. In contrast, the precisions for CMP_{cor}, CMP_{pol}, and DI_{32} are the highest (20%, 40%, 100% respectively). However, this comes at the cost of making the fewest number of predictions with the latter only making a single prediction. For these methods, less strict thresholds are needed to identify a greater proportion of positives at the cost of increasing the proportion of false discoveries. Across all methods, low f_{max} and ϕ_{max} values (0.26 and below) suggest there are no significance thresholds that balance power and precision for this data set.

I observed similarly low performance on A3G (Figure 4.21). Encouragingly, I note that positions 128-130 are correctly identified by multiple methods (Figure 4.19). Residues at position 130 (e.g., D vs A) are highly likely to be adaptations that conferred species-specific resistance to Vif-induced degradation in Old World Monkeys 5-6MYA [16], [15]. Position 128, that also provides species-specific resistance, is thought to be more recent [16], [15],

[77]. While these coevolution methods alone may not yet be accurate enough to identify functional residues, they potentially enhance other evolutionary analyses. For example, of the many Apobec sites under positive selection [15], it is reasonable that lentiviruses are more likely shaping the evolution of those sites that coevolve with Vif than sites that coevolve with other viral or virus-like agents.

Secondly, I visualized the localization of Vif residues predicted to be coevolving with A3G on a partial structure of Vif in complex with cofactors utilized for protein ubiquitination [31] (Figure 4.18). In [31], the authors are able to see that a critical subset of the Vif positives is solvent-exposed. I re-evaluated performance with only these residues as the positives (Table 4.3). There is poor precision to identify the putative solvent-exposed interface among the methods; CMP_{cor} at 50% and CMP_{vol} at 10% are the only methods with precision $> 6\%$ (Figure 4.22).

My analysis of the Vif-A3G interaction confirms that power to detect functionally important residues in each protein family is also low in inter-protein analyses between species, even though it is plausible that an arms race between lentivirus and mammal would give rise to stronger signals of coevolution compared to background. It is important to consider that perhaps the positions I considered positives may not all be of equal evolutionary importance across primates. Interfaces may be gained or lost and the rapid evolution of the two proteins likely produces many alternative solutions to maintaining an antagonistic interaction. There were many predicted positions that were not in the positives and further systematic validation and more comprehensive sequencing of lentiviruses and primates is needed to determine which pairs of residues are actually in close proximity or functionally required for other reasons. Additionally, there appears to be some level of complementarity in the predictions made by VI and $\mathrm{MI}_{\mathrm{minh}}$ and the CMP methods, which measure different biochemical trade offs between coevolving residues. This strengthens the rationale for integrating methods to better predict interface residues experiencing potentially different evolutionary constraints (e.g. structural, catalytic activity, specificity). Coevolutionary analysis can help to generate and prioritize candidates for these experiments.

4.4 Conclusions

In two example analyses involving HIV-human protein interactions, I further demonstrate that coevolutionary analyses of cross-species protein-protein interactions are largely hindered by a lack of phylogenetically deep protein alignments.

Notably missing in the HIV-human interactome are capsid and p6 protein interactions. If performance can be improved to the levels seen in the bacterial data sets, either by methods development or further sequencing, coevolution methods would become an important asset in discovering interactions missed by experimental methods

4.5 Figures

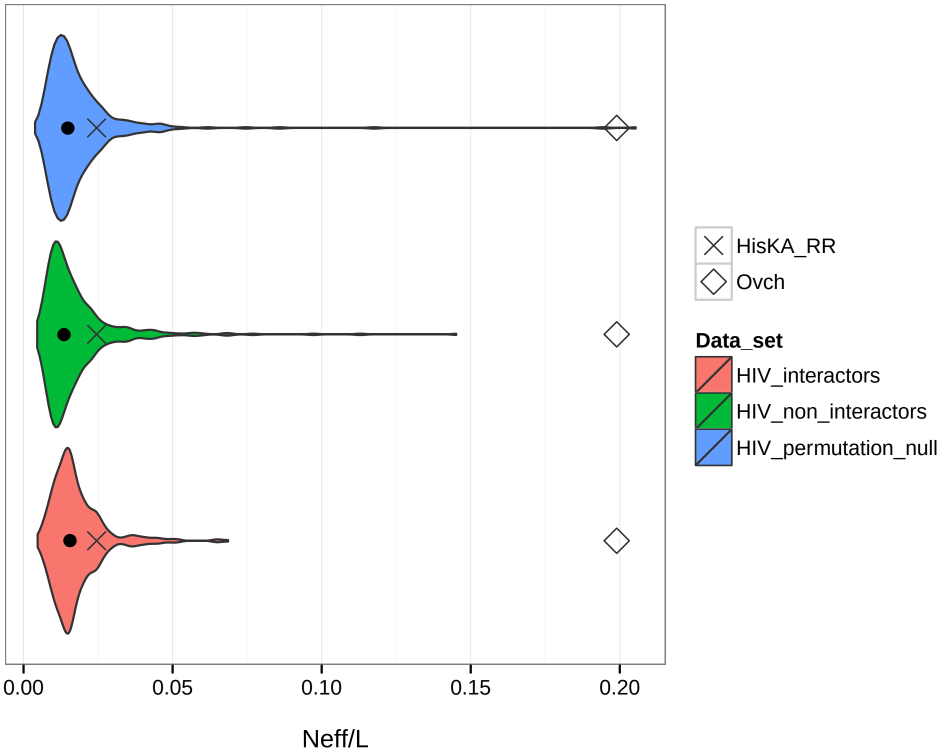

Figure 4.1: N_{eff}/L densities for APMS-derived HIV-human interactors, permuted nulls, and non-interactors. Median N_{eff}/L for the HIV alignments (black dot) is much lower than the minimum N_{eff}/L values seen in the bacterial data sets (HisKA-RR: cross, Ovch32: diamond)

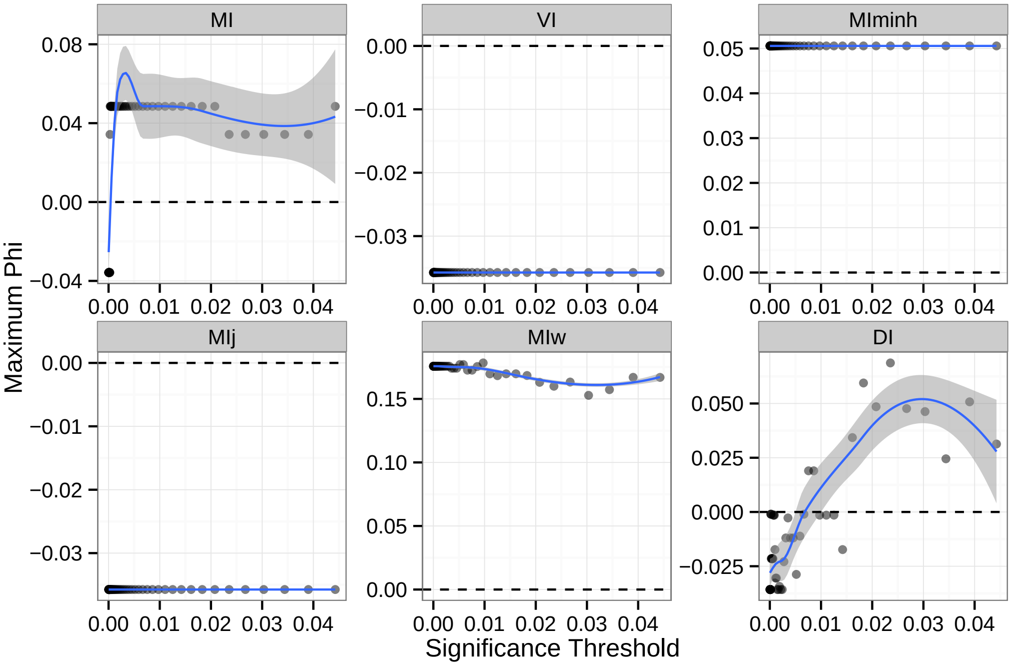

Figure 4.2: Predicting HIV-human interactors is difficult with such small $\mathrm{N_{eff}/L}$. $\mathrm{Phi_{max}}$ at varying FDR-corrected $P^n{}_{empirical} < \alpha$.

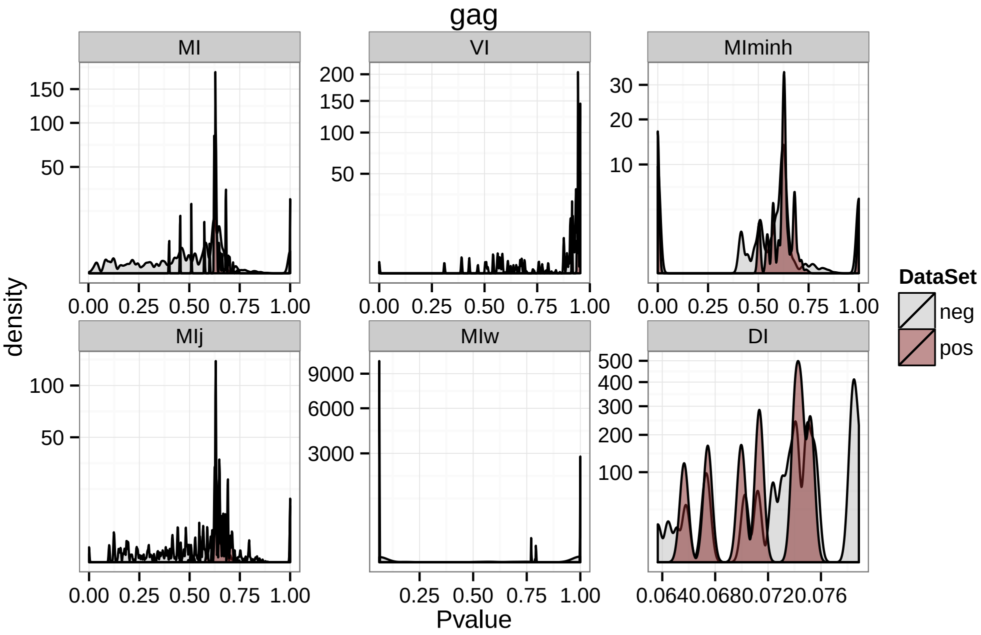

Figure 4.3: $P^n{}_{empirical}$ distributions for group-specific antigen (Gag) APMS-interactors (pos: red) and non-interactors (neg: grey).

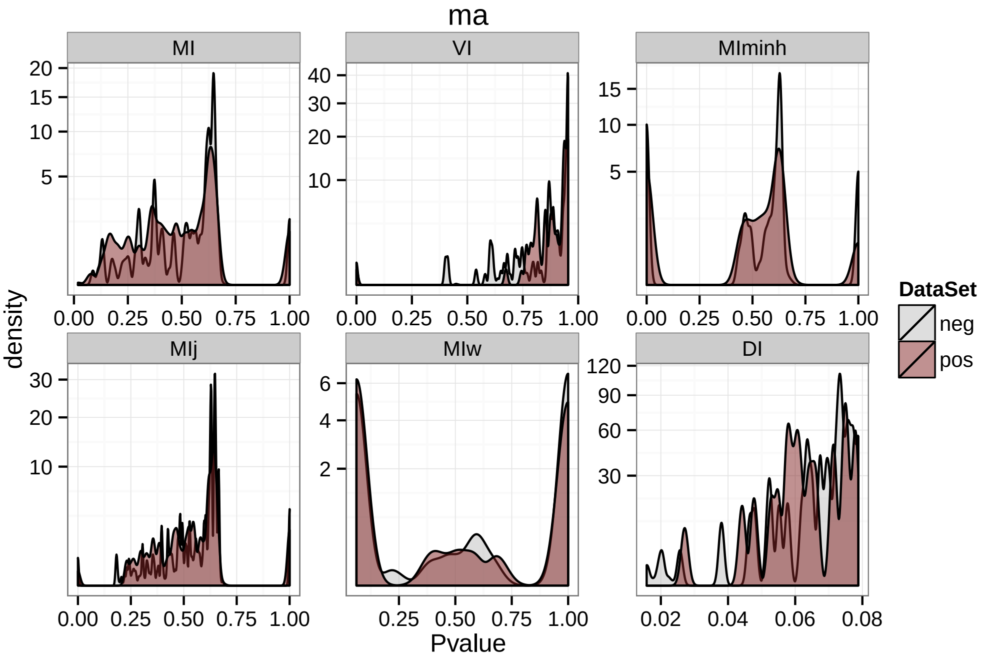

Figure 4.4: $P^n{}_{empirical}$ distributions for matrix (Ma) APMS-interactors (pos: red) and non-interactors (neg: grey).

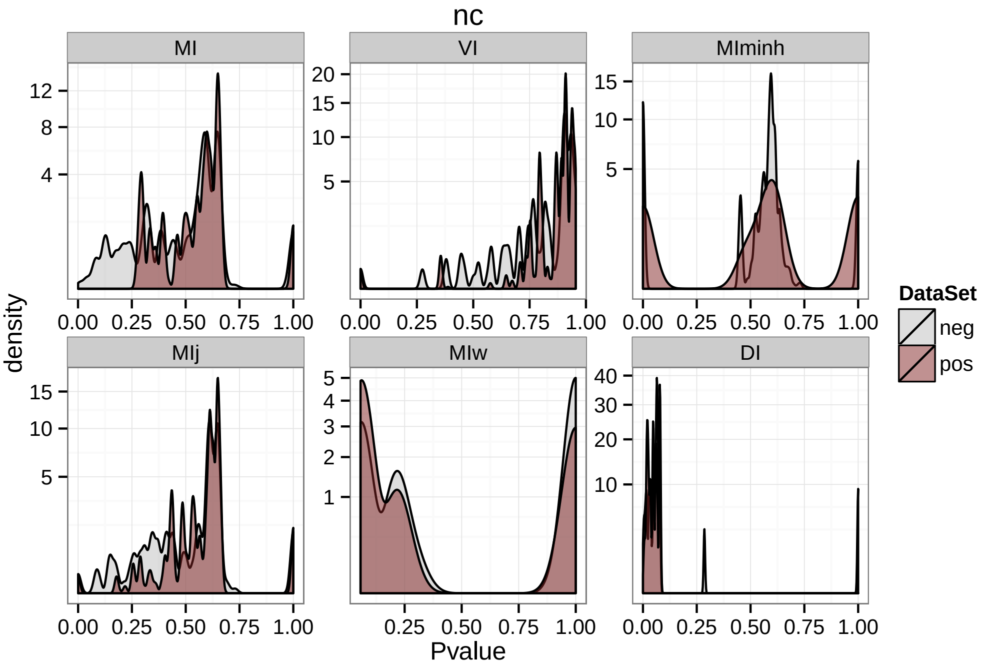

Figure 4.5: $P^{n}{}_{empirical}$ distributions for nucleocapsid (Nc) APMS-interactors (pos: red) and non-interactors (neg: grey).

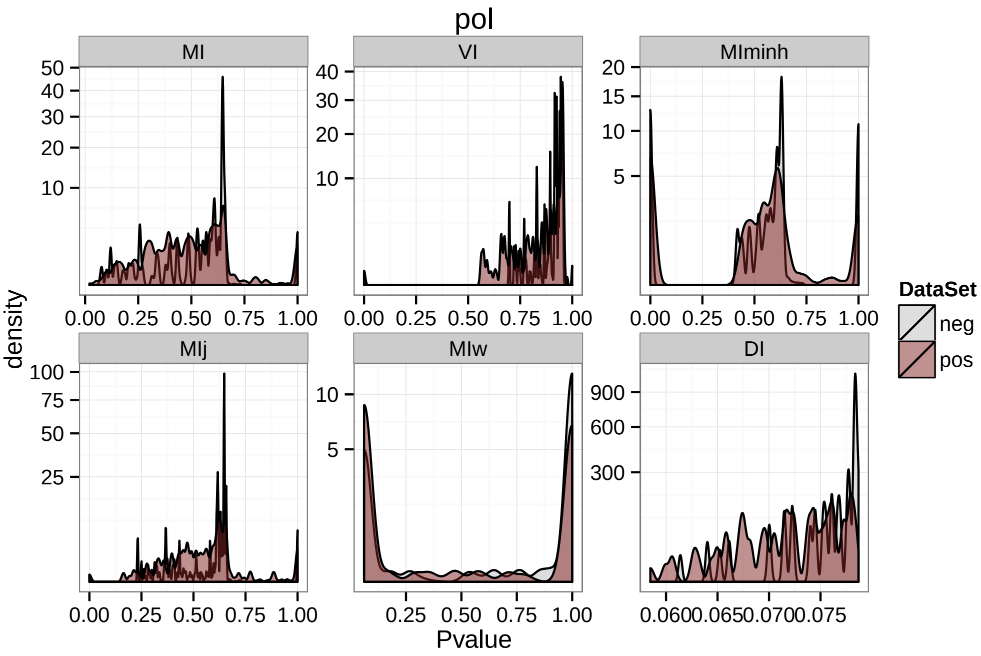

Figure 4.6: $P^n{}_{empirical}$ distributions for polymerase (Pol) APMS-interactors (pos: red) and non-interactors (neg: grey).

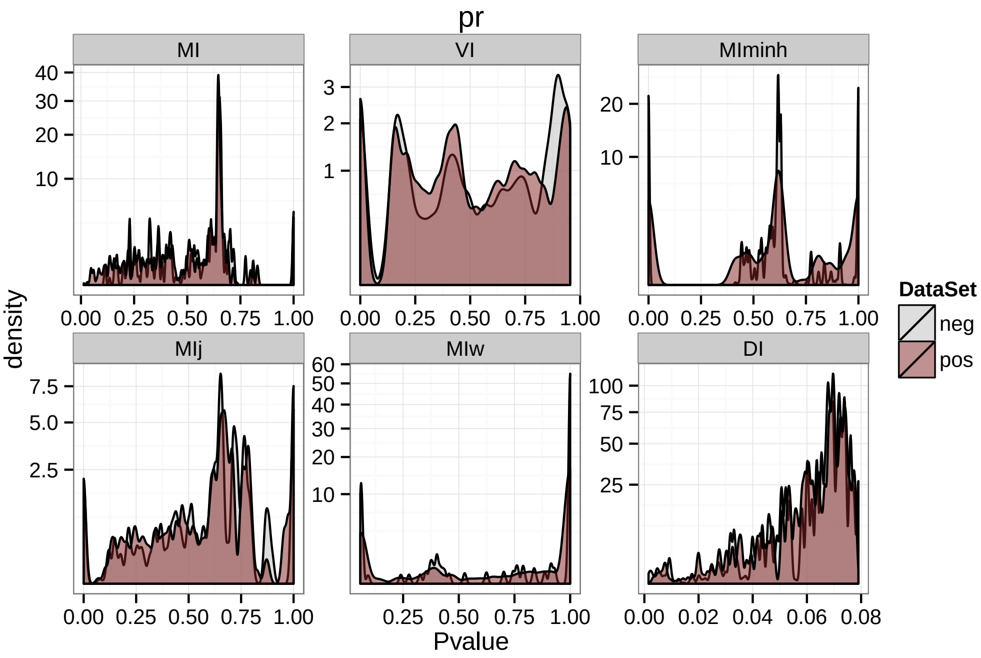

Figure 4.7: $P^n{}_{empirical}$ distributions for protease (Pr) APMS-interactors (pos: red) and non-interactors (neg: grey).

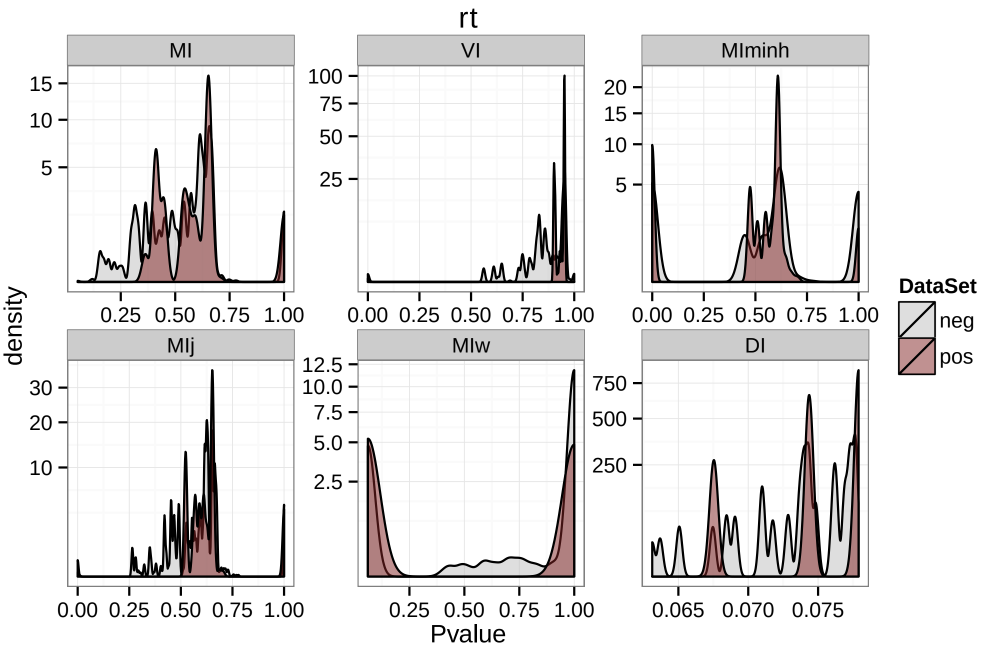

Figure 4.8: $P^{n}{}_{empirical}$ distributions for reverse transcriptase (Rt) APMS-interactors (pos: red) and non-interactors (neg: grey).

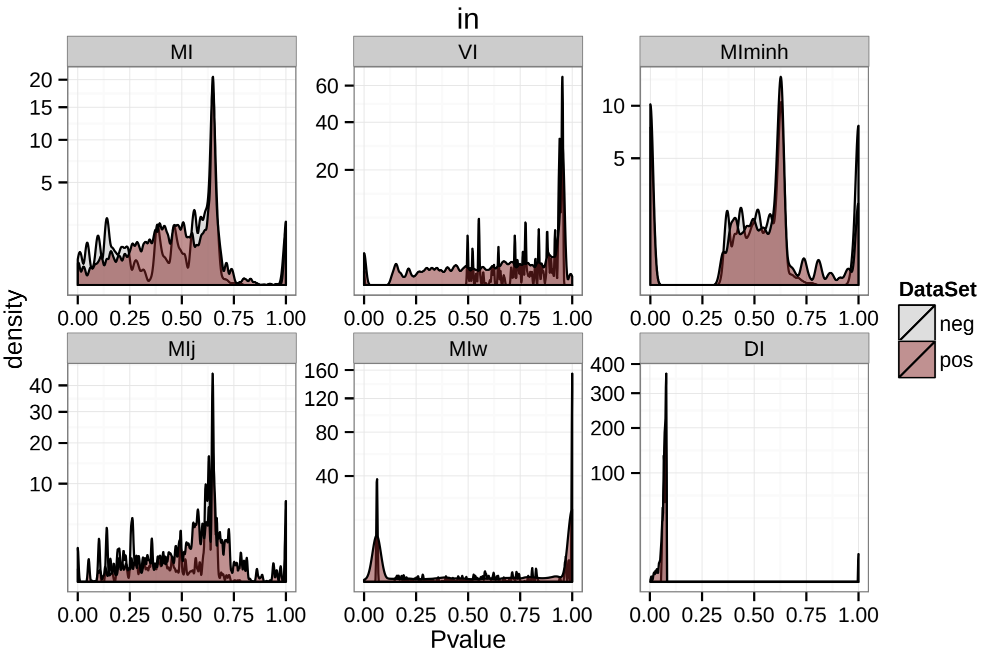

Figure 4.9: $P^n{}_{empirical}$ distributions for integrase (In) APMS-interactors (pos: red) and non-interactors (neg: grey).

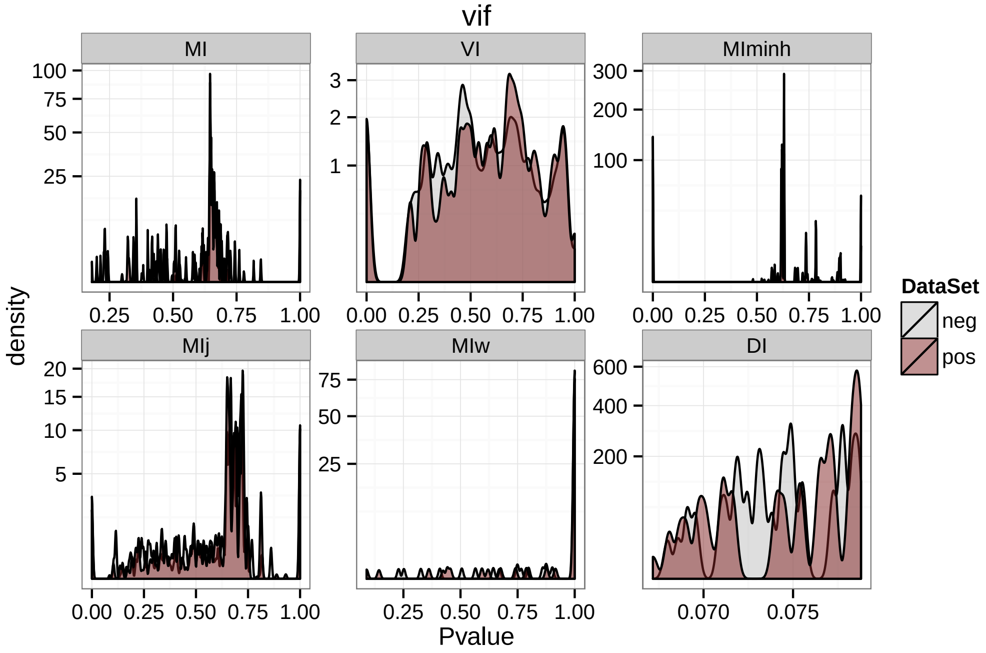

Figure 4.10: $P^n{}_{empirical}$ distributions for viral infectivity factor (Vif) APMS-interactors (pos: red) and non-interactors (neg: grey).

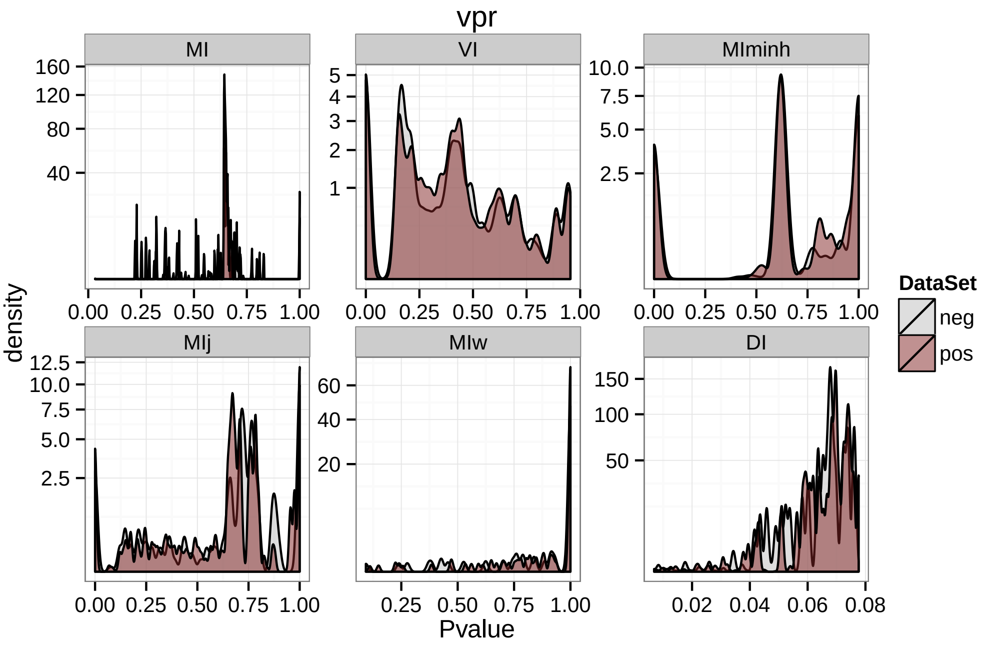

Figure 4.11: $P^{n}{}_{empirical}$ distributions for viral protein r (Vpr) APMS-interactors (pos: red) and non-interactors (neg: grey).

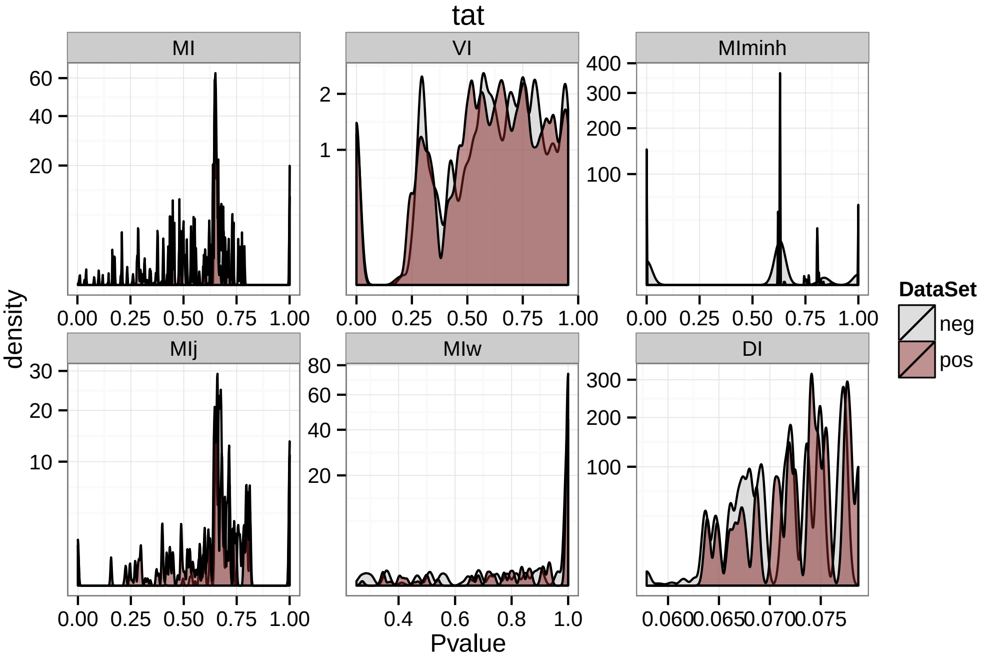

Figure 4.12: $P^{n}{}_{empirical}$ distributions for transactivator of transcription (Tat) APMS-interactors (pos: red) and non-interactors (neg: grey).

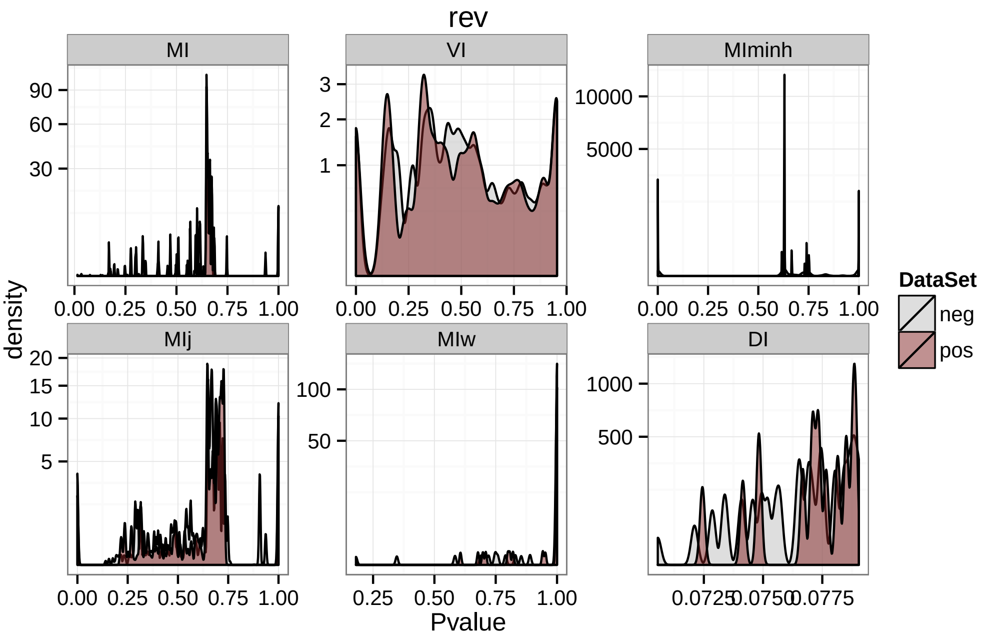

Figure 4.13: $P^n{}_{empirical}$ distributions for regulator of expression of virion proteins (Rev) APMS-interactors (pos: red) and non-interactors (neg: grey).

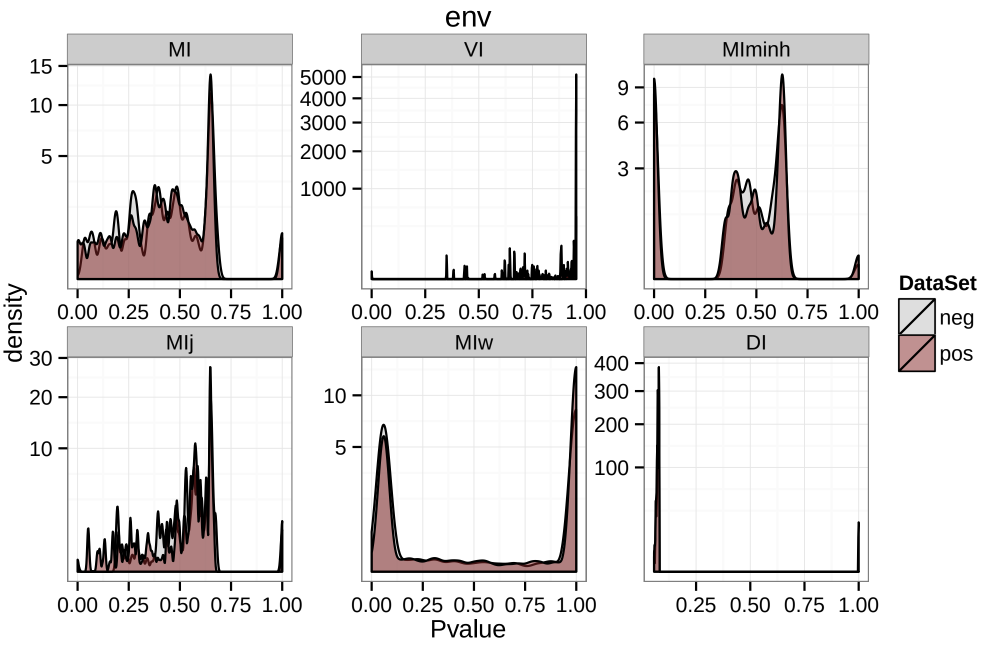

Figure 4.14: $P^n{}_{empirical}$ distributions for envelope (Env) APMS-interactors (pos: red) and non-interactors (neg: grey).

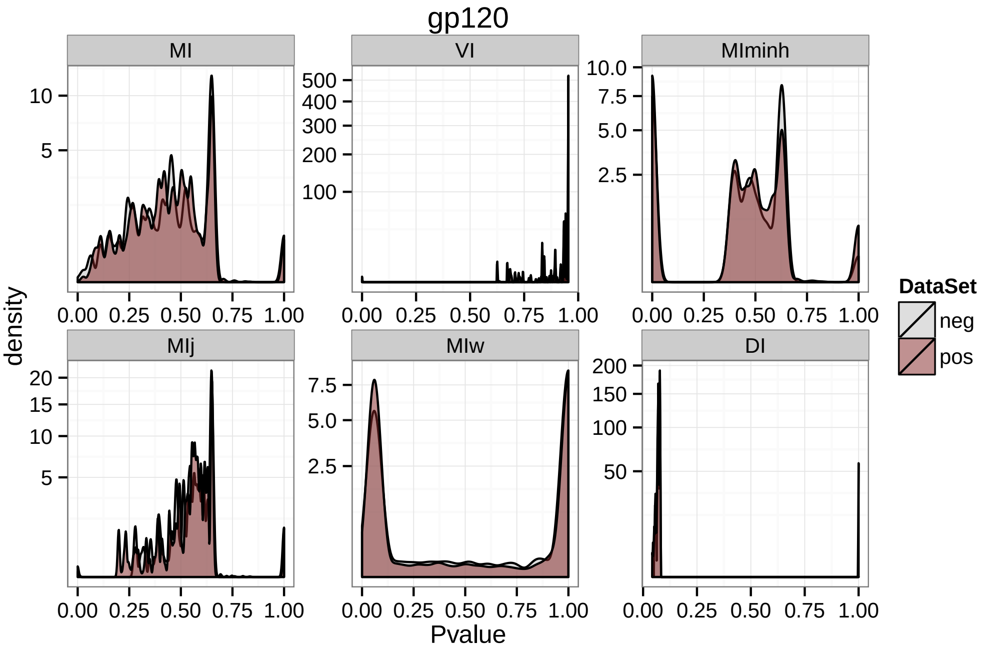

Figure 4.15: $P^{n}{}_{empirical}$ distributions for glycoprotein 120 (Gp120) APMS-interactors (pos: red) and non-interactors (neg: grey).

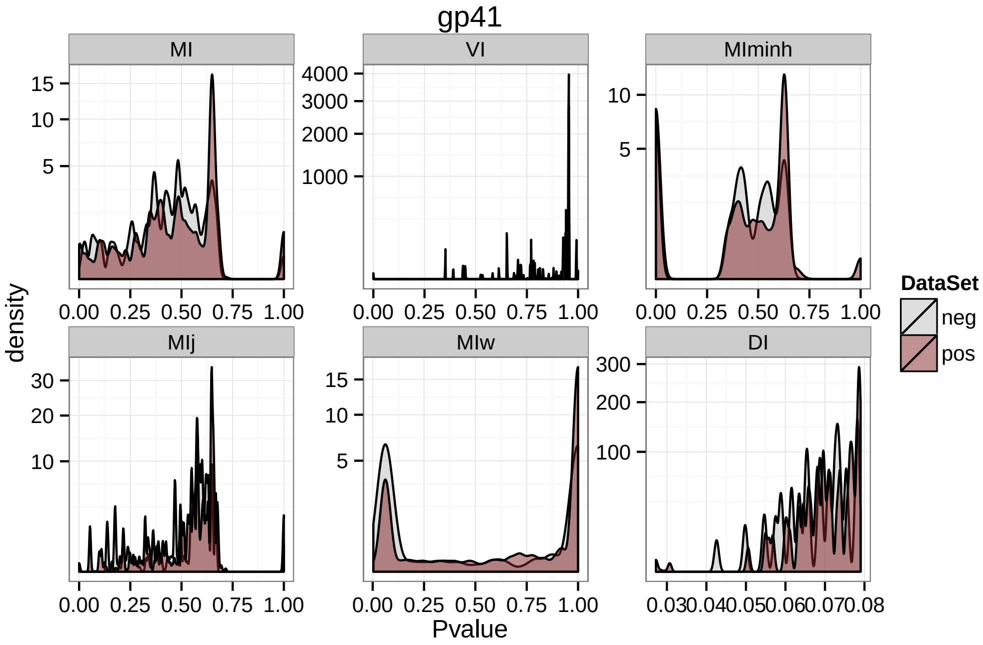

Figure 4.16: $P^n_{empirical}$ distributions for glycoprotein 41 (Gp41) APMS-interactors (pos: red) and non-interactors (neg: grey).

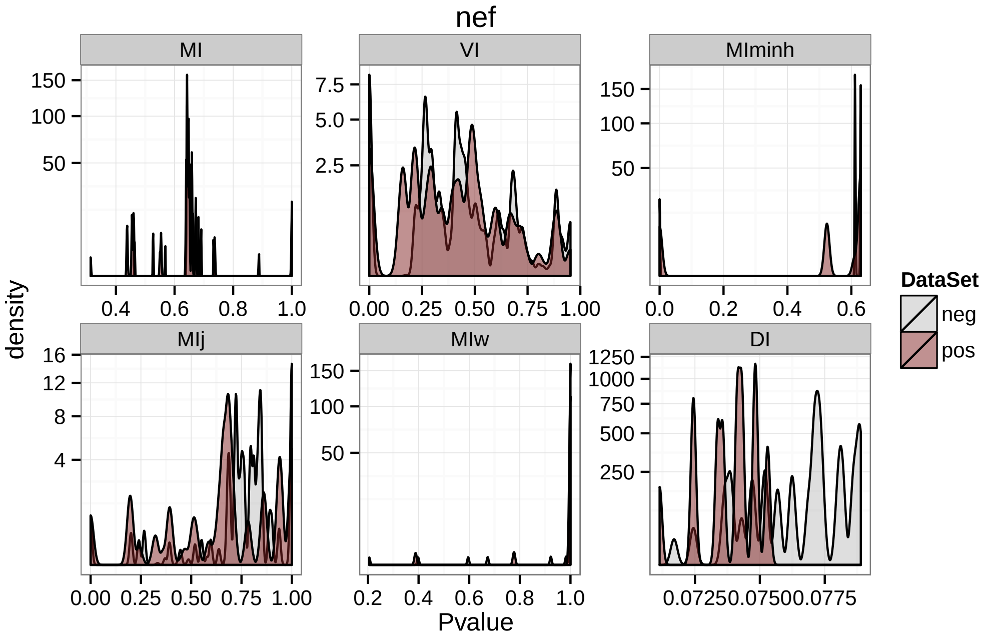

Figure 4.17: $P^n{}_{empirical}$ distributions for negative factor (Nef) APMS-interactors (pos: red) and non-interactors (neg: grey).

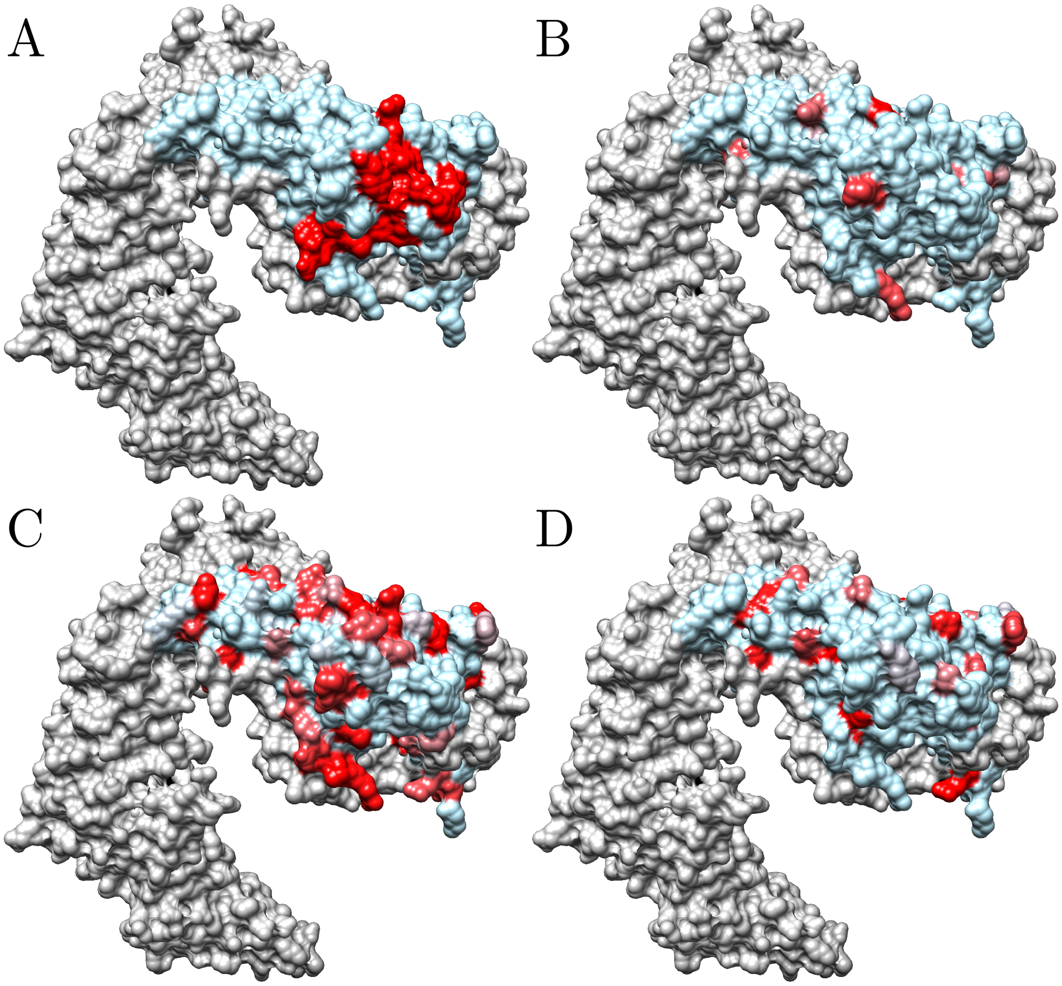

Figure 4.18: HIV1 Vif (light blue) in complex with co-factors (grey) without APOBEC3G (A3G) (PDB ID: 4N9F). Residues in red are predicted to be coevolving with A3G optimizing precision (PPV) using **A:** previously known essential residues, **B-D:** predictions using MI, DI, CMP$_{\text{vol}}$ respectively.

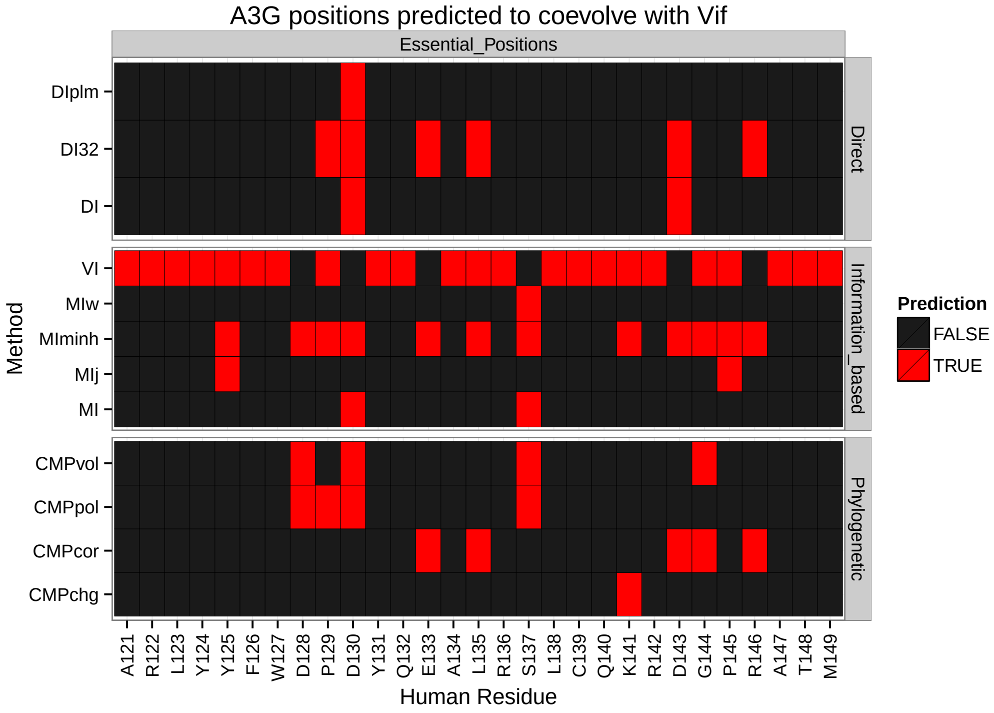

Figure 4.19: A3G positions predicted to coevolve with Vif (Human sequence shown). Mutations at positions 128 and 130 likely to be adaptations that conferred species-specific resistance to Vif-induced degradation in Old World Monkeys [16], [15], [77].

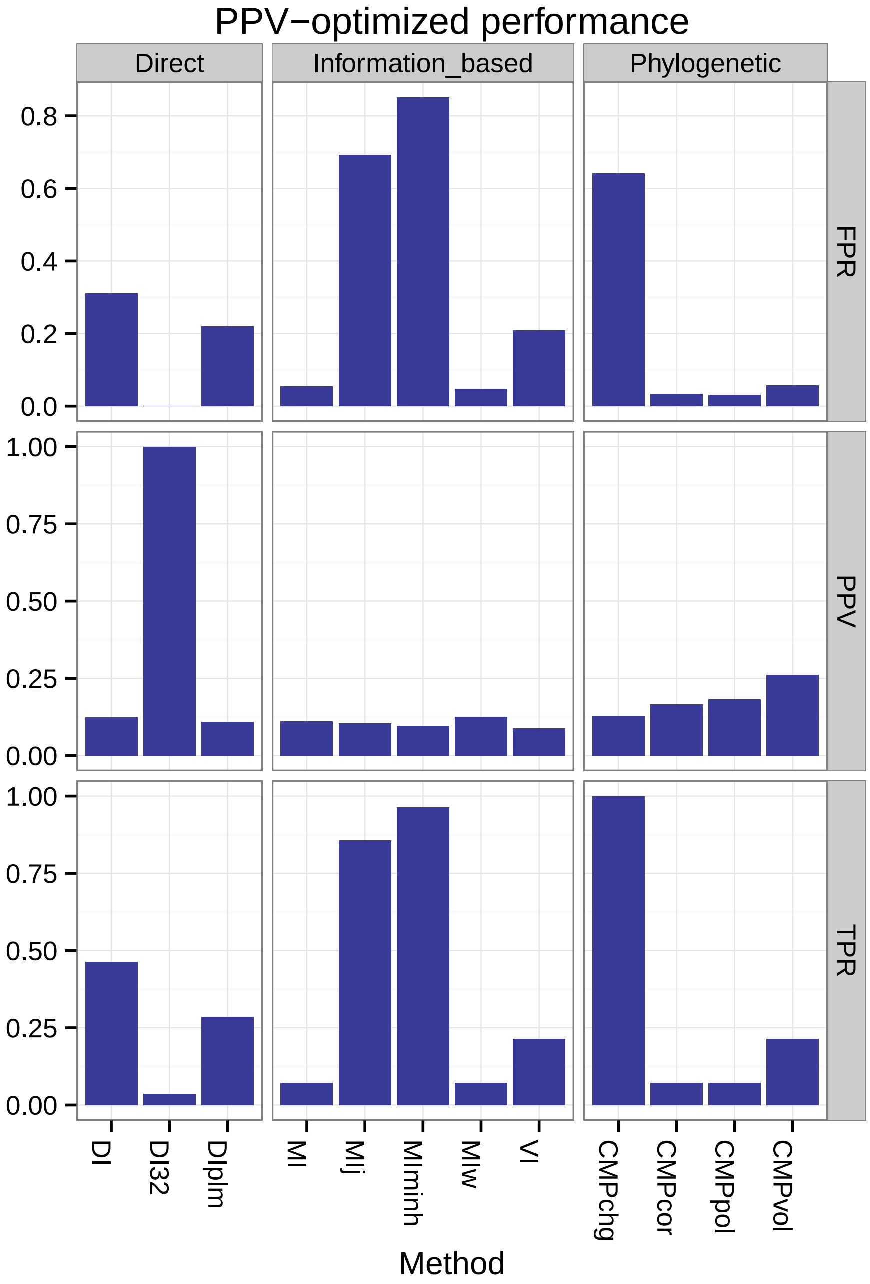

Figure 4.20: Performance at $P_{empirical} < \alpha$ that maximizes precision in Vif.

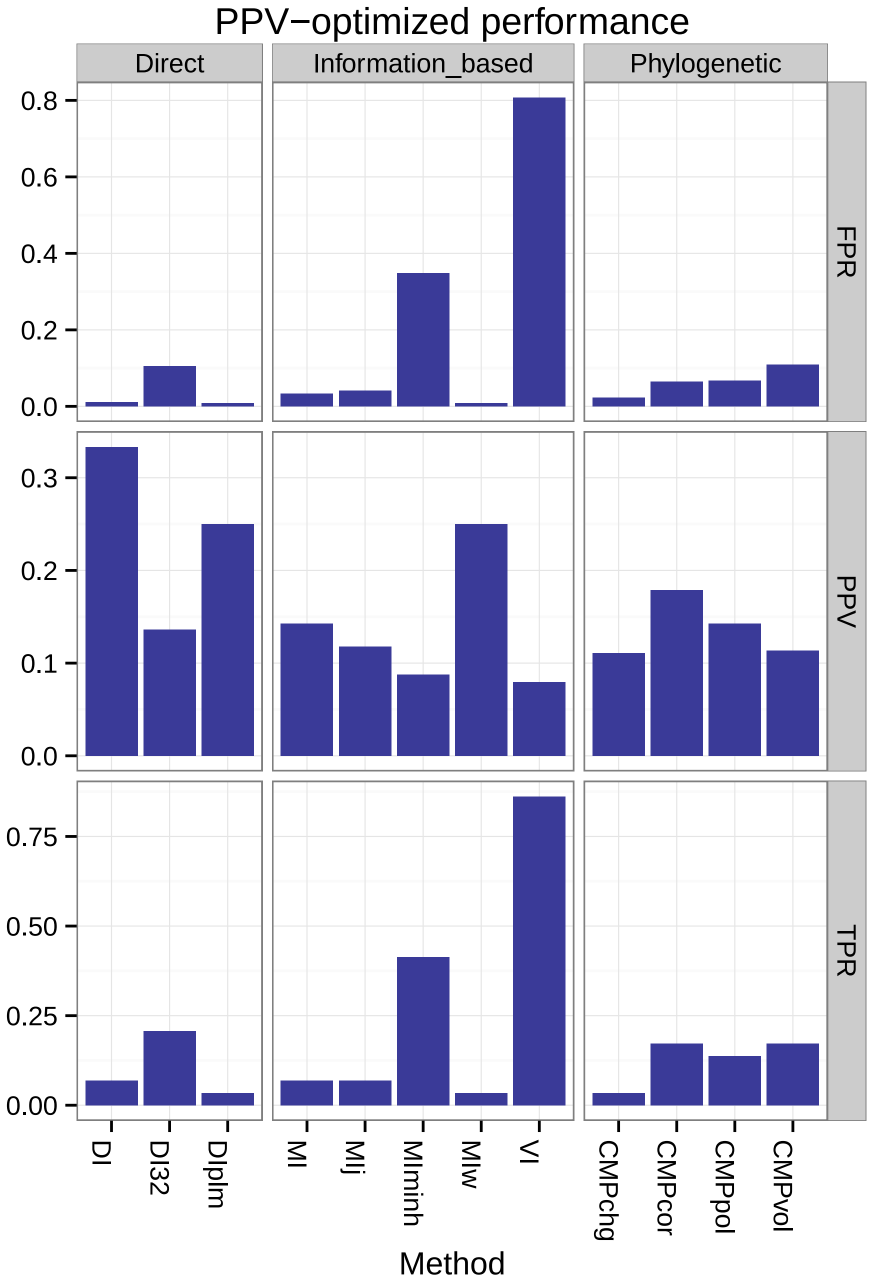

Figure 4.21: Performance at $P_{empirical} < \alpha$ that maximizes precision in A3G.

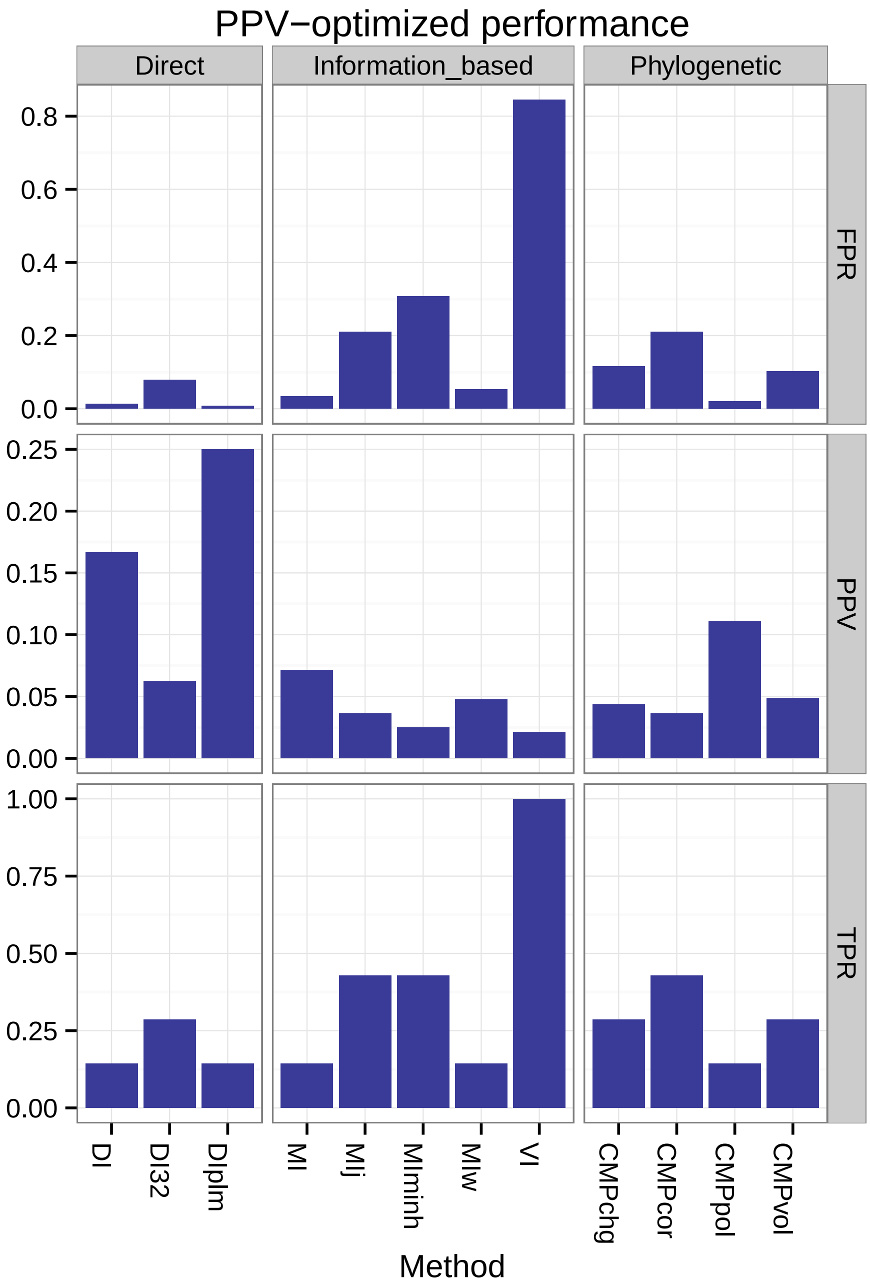

Figure 4.22: Performance at $P_{empirical} < \alpha$ that maximizes precision in Vif using critical residues.

4.6 Tables

			95% CI		
		μ	Lower	Upper	P-value
Interactors	Permutation null	0.0005	-0.0003	0.0014	0.2280
	Non-interactors	0.0016	0.0008	0.0024	0.0003
Non-interactors	Permutation null	-0.0011	-0.0019	-0.0002	0.0177

Table 4.1: Mann-Whitney test for location shift of $\mathrm{N_{eff}}$/L distributions. $\mathrm{H_0}$: location parameter $\mu = 0$.

Mammal	A3G accession	Lentivirus	Vif accession
Homo sapiens	NP_068594.1	*HIV1*	Q72499
		HIV2	Q74121
Pan troglodytes	NP_001009001.1	*SIVcpz*	Q1A266
Gorilla gorilla	AAT44394.1	*SIVgor*	ACM63194.1
Macaca mulatta	NP_001185622.1	*SIVmac*	P05903
Macaca nemestrina	ADU03765.1	*SIVmne*	AAA91932.1
Chlorocebus pygerythrus	AEY75955.1	*SIVver*	P27983
Chlorocebus tantalus	AEY75957.1	*SIVtan*	P89905
Chlorocebus sabaeus	AEY75959.1	*SIVsab*	AAA21506.1
Chlorocebus aethiops aethiops	AEY75961.1	*SIVgri*	AAA47589.1
Cercopithecus cephus	AGE34488.1	*SIVmus1*	ABO61046.1
		SIVmus2	ABO61055.1
Cercocebus torquatus	AGE34491.1	*SIVrcm*	AAK69675.1
Cercocebus atys	AGE34496.1	*SIVsmm*	P19506
Colobus guereza	AGE34499.1	*SIVcol*	AAK01034.1

Table 4.2: Accessions for primate A3G orthologs and lentiviral Vif orthologs, paired by host species.

	Position	Notes
Vif	21-23,26	A3G-specific
	30	
	40-44	
	55-72	A3G and A3F
A3G	121-149	essential for Vif-binding

Table 4.3: Important residues for the Vif-A3G interaction. HIV1 Vif [11, 63, 80, 34]. Human A3G [81, 64].

Chapter 5

Benchmarking software

I outline a few utilities I wrote to aid in processing sequences, structures, and coevolution results for benchmarking and making predictions and visualizations.

5.1 Background

Over the years many implementations for computing coevolution scores have been developed, either as standalone tools (See Table 2.1), web-servers [79], [68], or occasionally as part of libraries for evolutionary computation (http://biopp.univ-montp2.fr/wiki/index.php/Main_Page).

Unfortunately, computational platforms rapidly change and quickly become obsolete, and when coupled with the quick pace of computational biology, leaves many useful tools abandoned, or forgotten after publication.

In the case of coevolution software, this has resulted in many re-implementations that are often hard to obtain and verify for correctness. In turn this has led to the use of many different file formats.

Striking a balance between a general purpose utilitie and disposable scripts for particular tasks is extremely difficult and is all but guaranteed to result in inefficiencies. Either the researcher rewrites the same tool multiple times or devotes entirely too much time to making the perfect software suite that may unfortunately only be used a handful of times.

My attempt to balance standardization and hacks resulted in a handful of utility functions I most frequently used in my analyses. They are organized into complementary `python` and `R` packages for use in a scientific computing Linux environment. The goal of the packages is to easily format and benchmark new coevolutionary tools as they become available.

5.2 Dealing with alignments and structures

A major aspect of coevolutionary analyses involves joining and splitting alignments, converting them to different input formats for various tools, and mapping column numbers to peptide chains in a structural model.

The `coevo` package at https://github.com/aavilahe/coevo_analysis_pypackage contains many auxiliary functions and executable python scripts for these purposes.

A typical processing step may involve mangling names:

```
import Bio.AlignIO
from coevo.aln_aux import formatting as fmt

aln = Bio.AlignIO.read('example.fasta', format = 'fasta')
seqids = (seq.id for seq in aln)
id_map = dict(fmt.make_strict_phylip_id_map(seqids))
aln = fmt.replace_ids(aln, id_map)
print aln.format('fasta')
```

Here is a trivial example for converting between sequence formats:

```
# The coevolution programs benchmarked span three sequence formats.
# Convert fasta to a strict sequential phylip format
fasta_to_phy.py < left.fa > left.phy
# Convert fasta to the format PSICOV requires
fasta_to_psicov.py < left.fa > left.psi
```

Some methods require one concatenated alignment, while others read in two separate ones:

```
# Join two alignments on sequence identifiers (horizontally concatenate)
join_fastas.py left.fa right.fa > left_right.fa

# Divide an alignment into two parts
# The left alignment will have 324 columns
split_faa_on_col.py left_right.fa 324 left.fa right.fa
```

Another important procedure is to map column numbers from a given alignment to a reference PDB structure. For example, I used `map_column_to_resnum.py`, and `get_dists.py` to map atomic distances to column-pairs in existing alignments in order to compare them to coevolution scores and *P*-values and to validate predictions. The HisKA-RR complex in (PDB: 3DGE) is actually an ABAB tetramer—two sets of identical chains form a structure such that a HisKA chain will make contact with two RR chains. One can use `min_dists.py` to get the minimum distances between residues from both interactions. For a detailed example, see https://github.com/aavilahe/coevo_analysis_pypackage/blob/dev/example/pdb_tests/example_3DGE_column_distances.sh.

Visualization of coevolution score summaries on individual residues can be accomplished by generating an attributes file for use with UCSF Chimera [59] using `make_attributes.py` (e.g. Figure 4.18 shows Vif residues predicted to coevolve with A3G, each Vif residue is colored by most significant P-value out of all A3G residues).

5.3 Benchmarking and P-values

The code repository at https://github.com/aavilahe/coevo_analysis_pypackage also contains functions for merging results from multiple coevolution tools, handling different offsets (e.g. counting from `0` or `1`), dropping intraprotein columns and renumbering concatenated alignments by their individual indices.

After preprocessing results, https://github.com/aavilahe/coevo_analysis_Rpackage packages nifty wrappers for measuring performance with ROCR [69]. It aims to transparently handle different types of scores (dissimilarities, distances, and similarities) and estimate P-values that depend on observed scores.

5.4 Coevolution methods themselves

The most important component of a coevolution analysis—the coevolution methods themselves—come in a variety of flavors and in different states of maturity. I have written light wrappers for running the methods in Table 2.1 (available in https://github.com/aavilahe/coevo_tools) that are called simply by `runWrapper aln.in aln.out`. Flexibility for running with specific parameters is maintained by optional positional arguments in some cases (e.g. specifying number of CPUs or regularization strength for PSICOV). The most useful wrappers abstract MATLAB away from the user in the case of the DI family of methods.

The mutual information methods used in this work live at http://github.com/aavilahe/infcalc. Originally, this package was intended to leverage `cython` and the `multiprocessing` library to parallelize calculations for simulated bootstrap alignments in Section 3.2.1. However, it is more convenient to take advantage of a Sun Grid Engine supercomputing environment if available. Example scripts for generating job submissions are included.

5.5 Simulating evolution

- See Section 3.2.1 for an overview of `simulate_tools`

Chapter 6

Discussion

6.1 Summary of conclusions

Measuring coevolution between two proteins is a fundamental step in broadening our understanding of the role that residue-residue interactions play in a global evolutionary landscape. Although recent breakthroughs in methodologies have reignited a growing interest in coevolutionary analysis—especially for predicting structural contacts in proteins—methods development has concentrated on intra-protein analyses and leaves unanswered whether current methods are effective in inter-protein analyses, especially when the two proteins are from different genomes. In this work I looked beyond intra-protein contacts, and focus on coevolution between two different proteins, using data sets that cover a variety of protein-protein interactions, and including cross-species interactions.

My benchmarks revealed that using Direct methods such as PSICOV when coupled with an empirical P-value typically result in the best performance for inter-protein analyses. However, it is critical to note that alignment depth is arguably the most important aspect of a coevolutionary analysis, and one should strive for obtaining hundreds to thousands of diverse sequences for analyses ($\mathrm{N_{eff}/L} \gtrsim 2.0$).

6.2 Future improvements on methodologies

6.2.1 Statistical learning

As larger data sets of paired protein families are compiled, applying statistical learning techniques to coevolution analyses is becoming feasible. [71] investigated combining $\mathrm{DI_{plm}}$ [24] and PSICOV [39] predictions with classifiers trained on intraprotein analyses. [50] builds on [71] by testing a similar classifier—trained on intraprotein analyses—on a null data set built from the Negatome, a non-interaction database [5]. At a cutoff equivalent to $P_{empirical} < 0.001$, their classifier is able to exclude virtually all false positives from intermolecular contacts and achieves near 90% precision in predicting intramolecular contacts. However, neither the FPR among all contacts, nor TPR is reported. Li et al. [46] uses a monumental set of 27 coevolution methods to predict intra- and inter-protein coevolution within HIV proteins, and identifies a near-optimal combination of methods to average into a single classifier for the given HIV data sets.

Coevolution statistics may have different distributions for each type of molecular interaction (e.g. RNA-protein), or it might depend on the evolutionary pressures involved (e.g. host-virus ubiquitination vs bacterial signaling). Therefore, a variety of data sets is needed in order to validate that a particular classifier generalizes.

Classifiers that combine many different coevolution statistics face the challenge of regularizing features. Regularization is important for comparing features on different scales to prevent, for example, features on large scales from dominating optimization of an objective function. Additionally, values of the same feature may cover different ranges depending on the underlying data set. Transforming the raw values through quantile normalization is one way to regularize features.

A logical next step is to apply the statistical learning techniques to moderately sized data sets of deep alignment-pairs of true interactions and non-interaction networks in a rigorous framework that incorporates prior knowledge of the alignments and predicted structural elements of the involved proteins. [38] accomplishes this in an intra-protein analysis, and although current studies suggest that a similar inter-protein analysis is reasonable, such a study has not yet been published. A multi-stage classifier such as in [38] may prove advantageous in predicting different classes of coevolving residues.

While statistical learning techniques aim to transform and partition the high dimensional space in which these multitudes of correlation statistics live, another approach to improving classification performance and perhaps our understanding of coevolution is to incrementally combine the best performing aspects of the various methodologies. For example, alignment column covariation is encoded in many different ways, as categorical variables representing amino acid identity or class, as dissimilarities from a reference, and expected number of substitutions on branches of a phylogenetic tree. Also, amino acid frequencies are currently estimated directly, or with an ad-hoc regularization or flat prior (e.g. pseudo-counts and L2 penalty in [24], [56]), while perhaps a statistically and computationally efficient estimator, such as a James-Stein type estimator discussed in [33] may be better suited for small-sample high-dimensional data.

6.3 Future application of coevolution

One particularly exciting application of this approach is to characterize and potentially manipulate interacting residues in host-virus and host-parasite protein interactomes [37], [67].

I showed a limited application in predicting important residues in the Vif-A3G interaction. As more sequence and structural data becomes available it will be interesting to see how coevolutionary data becomes incorporated into discovery and design of novel therapeutics for rapidly evolving pathogens and elusive drug targets.

Newly emerging data on antibody and antigen sequences within a host [48] offers an opportunity to harness coevolutionary signals to investigate the mechanisms of broadly neutralizing antibodies and immune evasion. The primary open question for these new applications is whether existing methods are sensitive and specific enough to detect coevolution with the levels of constraint and divergence that are present in sequences extracted from patient samples.

Combined with an atlas of fitness landscapes for mutations that confer susceptibility or resistance to treatments, these analyses may hopefully prove a powerful tool in understanding and overcoming many diseases.

References

[1] Millán Ortiz et al. "Evolutionary trajectories of primate genes involved in HIV pathogenesis." In: *Molecular biology and evolution* 26.12 (Dec. 2009), pp. 2865–2875. ISSN: 1537-1719. DOI: 10.1093/molbev/msp197.

[2] Elisabeth R. M. Tillier and Thomas W. H. Lui. "Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments". In: *Bioinformatics* 19.6 (Apr. 2003), pp. 750–755. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btg072.

[3] Adam Ben-Shem et al. "The structure of the eukaryotic ribosome at 3.0 Å resolution". eng. In: *Science (New York, N.Y.)* 334.6062 (Dec. 2011), pp. 1524–1529. ISSN: 1095-9203. DOI: 10.1126/science.1212642.

[4] H. M. Berman. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. ISSN: 13624962. DOI: 10.1093/nar/28.1.235. (Visited on 05/26/2015).

[5] P. Blohm et al. "Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis". en. In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D396–

D400. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkt1079. (Visited on 04/13/2015).

[6] Anna Brückner et al. "Yeast Two-Hybrid, a Powerful Tool for Systems Biology". en. In: *International Journal of Molecular Sciences* 10.6 (June 2009), pp. 2763–2788. ISSN: 1422-0067. DOI: 10.3390/ijms10062763. (Visited on 02/14/2015).

[7] L. Burger and E. van Nimwegen. "Disentangling direct from indirect co-evolution of residues in protein alignments". eng. In: *PLoS computational biology* 6.1 (2010), e1000633. DOI: 10.1371/journal.pcbi.1000633.

[8] C. M. Buslje et al. "Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information". eng. In: *Bioinformatics* 25.9 (2009), pp. 1125–31. DOI: 10.1093/bioinformatics/btp135.

[9] J. G. Caporaso et al. "Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics". eng. In: *BMC evolutionary biology* 8 (2008), p. 327. DOI: 10.1186/1471-2148-8-327.

[10] P. Casino, V. Rubio, and A. Marina. "Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction". eng. In: *Cell* 139.2 (2009), pp. 325–36. DOI: 10.1016/j.cell.2009.08.032.

[11] G. Chen et al. "A patch of positively charged amino acids surrounding the human immunodeficiency virus type 1 Vif SLVx4Yx9Y motif in-

fluences its interaction with APOBEC3G". eng. In: *Journal of virology* 83.17 (2009), pp. 8674–82. DOI: 10.1128/JVI.00653-09.

[12] Greg W. Clark et al. "Multidimensional mutual information methods for the analysis of covariation in multiple sequence alignments". en. In: *BMC Bioinformatics* 15.1 (May 2014), p. 157. ISSN: 1471-2105. DOI: 10.1186/1471-2105-15-157. (Visited on 02/12/2015).

[13] N. L. Clark et al. "Coevolution of interacting fertilization proteins". eng. In: *PLoS genetics* 5.7 (2009), e1000570. DOI: 10.1371/journal.pgen.1000570.

[14] S. Cocco, R. Monasson, and M. Weigt. "From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction". eng. In: *PLoS computational biology* 9.8 (2013), e1003176. DOI: 10.1371/journal.pcbi.1003176.

[15] A. A. Compton and M. Emerman. "Convergence and divergence in the evolution of the APOBEC3G-Vif interaction reveal ancient origins of simian immunodeficiency viruses". eng. In: *PLoS pathogens* 9.1 (2013), e1003135. DOI: 10.1371/journal.ppat.1003135.

[16] A. A. Compton, V. M. Hirsch, and M. Emerman. "The host restriction factor APOBEC3G and retroviral Vif protein coevolve due to ongoing genetic conflict". eng. In: *Cell host & microbe* 11.1 (2012), pp. 91–8. DOI: 10.1016/j.chom.2011.11.010.

[17] V. Dahirel et al. "Coordinate linkage of HIV evolution reveals regions of immunological vulnerability". eng. In: *Proceedings of the National*

Academy of Sciences of the United States of America 108.28 (2011), pp. 11530–5. DOI: 10.1073/pnas.1105315108.

[18] E. Delaporte et al. “Large measles outbreak in Geneva, Switzerland, January to August 2011: descriptive epidemiology and demonstration of quarantine effectiveness”. eng. In: *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 18.6 (2013). URL: http://www.ncbi.nlm.nih.gov/pubmed/23410259.

[19] S. D. Dunn, L. M. Wahl, and G. B. Gloor. “Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction”. eng. In: *Bioinformatics* 24.3 (2008), pp. 333–40. DOI: 10.1093/bioinformatics/btm604.

[20] J. Y. Dutheil. “Detecting coevolving positions in a molecule: why and how to account for phylogeny”. eng. In: *Briefings in bioinformatics* 13.2 (2012), pp. 228–43. DOI: 10.1093/bib/bbr048.

[21] J. Dutheil and N. Galtier. “Detecting groups of coevolving positions in a molecule: a clustering approach”. In: *BMC Evol Biol* 7 (2007), p. 242. DOI: 10.1186/1471-2148-7-242.

[22] J. Dutheil et al. “A model-based approach for detecting coevolving positions in a molecule”. eng. In: *Molecular biology and evolution* 22.9 (2005), pp. 1919–28. DOI: 10.1093/molbev/msi183.

[23] S. R. Eddy. "Accelerated Profile HMM Searches". eng. In: *PLoS computational biology* 7.10 (2011), e1002195. DOI: 10.1371/journal.pcbi.1002195.

[24] M. Ekeberg et al. "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models". eng. In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 87.1 (2013), p. 012707. URL: http://www.ncbi.nlm.nih.gov/pubmed/23410359.

[25] Daniel P. Faith. "Conservation evaluation and phylogenetic diversity". In: *Biological Conservation* 61.1 (1992), pp. 1–10. ISSN: 0006-3207. DOI: http://dx.doi.org/10.1016/0006-3207(92)91201-3.

[26] M. A. Fares and S. A. Travers. "A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses". eng. In: *Genetics* 173.1 (2006), pp. 9–23. DOI: 10.1534/genetics.105.053249.

[27] Anthony A. Fodor and Richard W. Aldrich. "Influence of conservation on calculations of amino acid covariance in multiple sequence alignments". en. In: *Proteins: Structure, Function, and Bioinformatics* 56.2 (May 2004), pp. 211–221. ISSN: 08873585. DOI: 10.1002/prot.20098. (Visited on 03/18/2015).

[28] M. Gershoni et al. "Coevolution predicts direct interactions between mtDNA-encoded and nDNA-encoded subunits of oxidative phosphorylation complex i". eng. In: *Journal of molecular biology* 404.1 (2010), pp. 158–71. DOI: 10.1016/j.jmb.2010.09.029.

[29] Bernhard Goebel et al. "An approximation to the distribution of finite sample size mutual information estimates". In: *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*. Vol. 2. IEEE, 2005, pp. 1102–1106. URL: https://ieeexplore.ieee.org/ielx5/9996/32110/01494518.pdf (visited on 02/12/2015).

[30] R. Gouveia-Oliveira et al. "InterMap3D: predicting and visualizing co-evolving protein residues". eng. In: *Bioinformatics* 25.15 (2009), pp. 1963–5. DOI: 10.1093/bioinformatics/btp335.

[31] Y. Guo et al. "Structural basis for hijacking CBF-beta and CUL5 E3 ligase complex by HIV-1 Vif". eng. In: *Nature* 505.7482 (2014), pp. 229–33. DOI: 10.1038/nature12884.

[32] A. Haldimann et al. "Altered recognition mutants of the response regulator PhoB: a new genetic strategy for studying protein-protein interactions". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 93.25 (1996), pp. 14361–6. URL: http://www.ncbi.nlm.nih.gov/pubmed/8962056.

[33] Jean Hausser and Korbinian Strimmer. "Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks". In: *The Journal of Machine Learning Research* 10 (2009), pp. 1469–1484. URL: http://dl.acm.org/citation.cfm?id=1755833 (visited on 04/14/2015).

[34] Z. He et al. "Characterization of conserved motifs in HIV-1 Vif required for APOBEC3G and APOBEC3F interaction". eng. In: *Journal*

of molecular biology 381.4 (2008), pp. 1000–11. DOI: 10.1016/j.jmb.2008.06.061.

[35] T. A. Hopf et al. "Sequence co-evolution gives 3D contacts and structures of protein complexes". In: *Elife* 3 (2014). DOI: 10.7554/eLife.03430.

[36] T. A. Hopf et al. "Three-dimensional structures of membrane proteins from genomic sequencing". eng. In: *Cell* 149.7 (2012), pp. 1607–21. DOI: 10.1016/j.cell.2012.04.012.

[37] S. Jager et al. "Global landscape of HIV-human protein complexes". eng. In: *Nature* 481.7381 (2012), pp. 365–70. DOI: 10.1038/nature10719.

[38] D. T. Jones et al. "MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins". en. In: *Bioinformatics* 31.7 (Apr. 2015), pp. 999–1006. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btu791. (Visited on 04/03/2015).

[39] D. T. Jones et al. "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments". eng. In: *Bioinformatics* 28.2 (2012), pp. 184–90. DOI: 10.1093/bioinformatics/btr638.

[40] D. de Juan, F. Pazos, and A. Valencia. "Emerging methods in protein co-evolution". eng. In: *Nature reviews. Genetics* 14.4 (2013), pp. 249–61. DOI: 10.1038/nrg3414.

[41] D. Juan, F. Pazos, and A. Valencia. "High-confidence prediction of global interactomes based on genome-wide coevolutionary networks". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.3 (2008), pp. 934–9. DOI: 10.1073/pnas.0709671105.

[42] H. Kamisetty, S. Ovchinnikov, and D. Baker. "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.39 (2013), pp. 15674–9. DOI: 10.1073/pnas.1314045110.

[43] T. Koestler, A. von Haeseler, and I. Ebersberger. "REvolver: modeling sequence evolution under domain constraints". In: *Mol Biol Evol* 29.9 (2012), pp. 2133–45. DOI: 10.1093/molbev/mss078.

[44] K. Lasker et al. "Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach". en. In: *Proceedings of the National Academy of Sciences* 109.5 (Jan. 2012), pp. 1380–1387. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1120559109. (Visited on 05/26/2015).

[45] M. T. Laub and M. Goulian. "Specificity in two-component signal transduction pathways". eng. In: *Annual review of genetics* 41 (2007), pp. 121–45. DOI: 10.1146/annurev.genet.41.042007.170548.

[46] Guangdi Li et al. “A new ensemble coevolution system for detecting HIV-1 protein coevolution”. en. In: *Biology Direct* 10.1 (Dec. 2015). ISSN: 1745-6150. DOI: 10.1186/s13062-014-0031-8. (Visited on 04/13/2015).

[47] L. Li, E. I. Shakhnovich, and L. A. Mirny. “Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.8 (2003), pp. 4463–8. DOI: 10.1073/pnas.0737647100.

[48] H. X. Liao et al. “Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus”. eng. In: *Nature* 496.7446 (2013), pp. 469–76. DOI: 10.1038/nature12053.

[49] M. C. Maher and R. D. Hernandez. “Rock, Paper, Scissors: Harnessing Complementarity in Ortholog Detection Methods Improves Comparative Genomic Inference”. en. In: *G3&amp;#58; Genes|Genomes|Genetics* 5.4 (Apr. 2015), pp. 629–638. ISSN: 2160-1836. DOI: 10.1534/g3.115.017095. (Visited on 06/07/2015).

[50] W. Mao et al. “Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution”. en. In: *Bioinformatics* (Feb. 2015). ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv103. (Visited on 04/03/2015).

[51] D. S. Marks, T. A. Hopf, and C. Sander. “Protein structure prediction from sequence variation”. eng. In: *Nature biotechnology* 30.11 (2012), pp. 1072–80. DOI: 10.1038/nbt.2419.

[52] D. S. Marks et al. “Protein 3D structure computed from evolutionary sequence variation”. eng. In: *PloS one* 6.12 (2011), e28766. DOI: 10.1371/journal.pone.0028766.

[53] R. N. McLaughlin Jr. et al. “The spatial architecture of protein function and adaptation”. eng. In: *Nature* 491.7422 (2012), pp. 138–42. DOI: 10.1038/nature11500.

[54] Marina Meila. “Comparing clusterings–an information based distance”. In: *Journal of Multivariate Analysis* 98.5 (2007), pp. 873–895. ISSN: 0047-259X. DOI: http://dx.doi.org/10.1016/j.jmva.2006.11.013.

[55] S. W. Michnick et al. “Protein-fragment complementation assays for large-scale analysis, functional dissection and dynamic studies of protein-protein interactions in living cells”. In: *Methods Mol Biol* 756 (2011), pp. 395–425. DOI: 10.1007/978-1-61779-160-4<sub>2</sub>5.

[56] F. Morcos et al. “Direct-coupling analysis of residue coevolution captures native contacts across many protein families”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.49 (2011), E1293–301. DOI: 10.1073/pnas.1111471108.

[57] J. H. Morris et al. “Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions”. In: *Nat Protoc* 9.11 (2014), pp. 2539–54. DOI: 10.1038/nprot.2014.164.

[58] S. Ovchinnikov, H. Kamisetty, and D. Baker. “Robust and accurate prediction of residue-residue interactions across protein interfaces using

evolutionary information”. In: *Elife* 3 (2014), e02030. DOI: 10.7554/eLife.02030.

[59] E. F. Pettersen et al. “UCSF Chimera–a visualization system for exploratory research and analysis”. eng. In: *Journal of computational chemistry* 25.13 (2004), pp. 1605–12. DOI: 10.1002/jcc.20084.

[60] D. D. Pollock, W. R. Taylor, and N. Goldman. “Coevolving protein residues: maximum likelihood identification and relationship to structure”. eng. In: *Journal of molecular biology* 287.1 (1999), pp. 187–98. DOI: 10.1006/jmbi.1998.2601.

[61] M. N. Price, P. S. Dehal, and A. P. Arkin. “FastTree 2–approximately maximum-likelihood trees for large alignments”. eng. In: *PloS one* 5.3 (2010), e9490. DOI: 10.1371/journal.pone.0009490.

[62] Andrea Procaccini et al. “Dissecting the Specificity of Protein-Protein Interaction in Bacterial Two-Component Signaling: Orphans and Crosstalks”. en. In: *PLoS ONE* 6.5 (May 2011). Ed. by Magnus Rattray, e19729. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0019729. (Visited on 02/18/2015).

[63] R. A. Russell and V. K. Pathak. “Identification of two distinct human immunodeficiency virus type 1 Vif determinants critical for interactions with human APOBEC3G and APOBEC3F”. eng. In: *Journal of virology* 81.15 (2007), pp. 8201–10. DOI: 10.1128/JVI.00395-07.

[64] R. A. Russell et al. “Distinct domains within APOBEC3G and APOBEC3F interact with separate regions of human immunodefi-

ciency virus type 1 Vif". eng. In: *Journal of virology* 83.4 (2009), pp. 1992–2003. DOI: 10.1128/JVI.01621-08.

[65] A. Schug et al. "High-resolution protein complexes from integrating genomic information with molecular simulation". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.52 (2009), pp. 22124–9. DOI: 10.1073/pnas.0912100106.

[66] Claude Elwood Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27 (Oct. 1948), pp. 379–423, 623–656. URL: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf (visited on 02/11/2015).

[67] S. D. Shapira et al. "A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection". eng. In: *Cell* 139.7 (2009), pp. 1255–67. DOI: 10.1016/j.cell.2009.12.018.

[68] F. L. Simonetti et al. "MISTIC: mutual information server to infer coevolution". en. In: *Nucleic Acids Research* 41.W1 (July 2013), W8–W14. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkt427. (Visited on 04/07/2015).

[69] T. Sing et al. "ROCR: visualizing classifier performance in R". en. In: *Bioinformatics* 21.20 (Oct. 2005), pp. 3940–3941. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bti623. (Visited on 06/08/2015).

[70] J. M. Skerker et al. "Rewiring the specificity of two-component signal transduction systems". eng. In: *Cell* 133.6 (2008), pp. 1043–54. DOI: 10.1016/j.cell.2008.04.040.

[71] M. J. Skwark, A. Abdel-Rehim, and A. Elofsson. “PconsC: combination of direct information methods and alignments improves contact prediction”. en. In: *Bioinformatics* 29.14 (July 2013), pp. 1815–1816. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btt259. (Visited on 04/13/2015).

[72] The UniProt Consortium. “UniProt: a hub for protein information”. en. In: *Nucleic Acids Research* 43.D1 (Jan. 2015), pp. D204–D212. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gku989. (Visited on 04/14/2015).

[73] Marc Vidal and Stanley Fields. “The yeast two-hybrid assay: still finding connections after 25 years”. In: *Nature methods* 11.12 (2014), pp. 1203–1206. URL: http://www.nature.com/articles/nmeth.3182 (visited on 02/14/2015).

[74] Arunachalam Vinayagam et al. “Integrating protein-protein interaction networks with phenotypes reveals signs of interactions”. In: *Nature Methods* 11.1 (Nov. 2013), pp. 94–99. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2733. (Visited on 05/26/2015).

[75] M. Weigt et al. “Identification of direct residue contacts in protein-protein interaction by message passing”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.1 (2009), pp. 67–72. DOI: 10.1073/pnas.0805923106.

[76] K. R. Wollenberg and W. R. Atchley. “Separation of phylogenetic and functional associations in biological sequences by using the parametric

bootstrap". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 97.7 (2000), pp. 3288–91. DOI: 10.1073/pnas.070154797.

[77] H. Xu et al. "A single amino acid substitution in human APOBEC3G antiretroviral enzyme confers resistance to HIV-1 virion infectivity factor-induced depletion". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.15 (2004), pp. 5652–7. DOI: 10.1073/pnas.0400830101.

[78] C. H. Yeang and D. Haussler. "Detecting coevolution in and among protein domains". eng. In: *PLoS computational biology* 3.11 (2007), e211. DOI: 10.1371/journal.pcbi.0030211.

[79] K. Y. Yip et al. "An integrated system for studying residue coevolution in proteins". eng. In: *Bioinformatics* 24.2 (2008), pp. 290–2. DOI: 10.1093/bioinformatics/btm584.

[80] H. Zhang et al. "Human immunodeficiency virus type 1 Vif protein is an integral component of an mRNP complex of viral RNA and could be involved in the viral RNA folding and packaging process". eng. In: *Journal of virology* 74.18 (2000), pp. 8252–61. URL: http://www.ncbi.nlm.nih.gov/pubmed/10954522.

[81] L. Zhang et al. "Function analysis of sequences in human APOBEC3G involved in Vif-mediated degradation". eng. In: *Virology* 370.1 (2008), pp. 113–21. DOI: 10.1016/j.virol.2007.08.027.

[82] Yuqi Zhao et al. “Integrated Analysis of Residue Coevolution and Protein Structures Capture Key Protein Sectors in HIV-1 Proteins”. en. In: *PLOS ONE* 10.2 (Feb. 2015). Ed. by Yuxian He, e0117506. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0117506. (Visited on 06/08/2015).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

June 11, 2015

Date