Avila-Herrera and Pollard *BMC Bioinformatics* (2015) 16:268
DOI 10.1186/s12859-015-0677-y

BMC Bioinformatics

RESEARCH Open Access

Coevolutionary analyses require phylogenetically deep alignments and better null models to accurately detect inter-protein contacts within and between species

Aram Avila-Herrera[1,2] and Katherine S. Pollard[1,2,3,4*]

Abstract

Background: When biomolecules physically interact, natural selection operates on them jointly. Contacting positions in protein and RNA structures exhibit correlated patterns of sequence evolution due to constraints imposed by the interaction, and molecular arms races can develop between interacting proteins in pathogens and their hosts. To evaluate how well methods developed to detect coevolving residues within proteins can be adapted for cross-species, inter-protein analysis, we used statistical criteria to quantify the performance of these methods in detecting inter-protein residues within 8 angstroms of each other in the co-crystal structures of 33 bacterial protein interactions. We also evaluated their performance for detecting known residues at the interface of a host-virus protein complex with a partially solved structure.

Results: Our quantitative benchmarking showed that all coevolutionary methods clearly benefit from alignments with many sequences. Methods that aim to detect direct correlations generally outperform other approaches. However, faster mutual information based methods are occasionally competitive in small alignments and with relaxed false positive rates. Two commonly used null distributions are anti-conservative and have high false positive rates in some scenarios, although the empirical distribution of scores performs reasonably well with deep alignments.

Conclusions: We conclude that coevolutionary analysis of cross-species protein interactions holds great promise but requires sequencing many more species pairs.

Keywords: Coevolution, Methods comparison, Inter-protein, Cross-species, Host-virus, Contact prediction, Protein interaction

Background

Coevolution—"the change of a biological object triggered by the change of a related object" [1]—is a powerful concept when applied to molecular sequence analysis because it reveals positional relationships that are preserved 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 [2]. 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. Understanding the basic connections and dependencies between these molecular machines 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.

Because epistasis can induce correlation between substitution patterns among columns in multiple sequence alignments, many methods have been developed that

*Correspondence: kpollard@gladstone.ucsf.edu
[1] Bioinformatics Graduate Program, University of California, San Francisco, USA
[2] Gladstone Institute of Cardiovascular Disease, University of California, San Francisco, USA
Full list of author information is available at the end of the article

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 [3, 4]).

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 [5, 6], differential weighting of sequences when computing statistics [7–9], and null distributions that directly model or indirectly account for phylogeny [10–13].

A second challenge arises when trying to distinguish correlated evolution that arises from direct versus indirect interactions. Alignment columns that are indirectly implicated in an interaction 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 coevolution methods that focuses on reducing the number of incorrect predictions by disentangling direct from indirect correlations [9, 14–17]. 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 [18]. Proteins are in fact quite dynamic, and many *unstructured* proteins are known to have important interactions [19, 20]. Coevolution methods have the exciting potential to reveal these hard to identify interactions, however distinguishing spurious correlations from true non-structural interactions remains a challenge.

The main barrier to overcoming this challenge is the impressively difficult task of compiling “gold standard” data sets in which true coevolving sites are clearly defined. Structural and systems biology have had great success in identifying and characterizing many important interactions (e.g. Nucleosome [21], Proteasome [22], regulation in protein networks [23, 24]). However, resolving large complexes and unstructured proteins remains technically difficult, a daunting task as the number of proteins is ever increasing.

Finally, due to low 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 [25, 26]. Even in the current data rich era of computational biology, such deep alignments are difficult to obtain, especially for cross-species protein interactions (e.g. host and pathogen interactions) because both members of the interaction must be equally deeply sequenced. Additionally, resolving orthologs and paralogs is not trivial.

Despite these challenges, coevolutionary prediction of physically interacting alignment columns has been applied with success to intra-molecular contacts [7, 27–29] and well-characterized inter-molecular interactions [30], such as bacterial two-component signaling systems [31], enzyme complexes [32], and fertilization proteins [33]. Although the signal-to-noise ratio is too low and the search space too large to use sequence evolution 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 [34].

However, the ability to now measure physical interactions between biomolecules with high-throughput technologies, such as affinity purification followed by mass spectrometry (APMS) [35], two-hybrid methods [36, 37], and protein complementation assays [38], 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.

One particularly exciting application of this approach is to characterize and potentially manipulate interacting residues in host-virus and host-parasite protein interactomes [23, 39]. Newly emerging data on antibody and antigen sequences within a host [40] 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 inter-molecular data sets of modest size.

To this end, we designed data processing scripts, statistical evaluation and visualization tools, and simulation pipelines that allowed us to easily extend a suite of coevolution methods designed for intra-protein interaction prediction (Table 1) so that they can be used to test for patterns of correlated sequence evolution at pairs of sites in two different proteins, potentially from different sets of organisms in different parts of the tree of life (e.g. human-bacteria, bacteria-phage interactions). We then applied this integrated framework for coevolutionary analysis to refine and annotate a recently derived human-HIV1 protein-protein interaction network [23] 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, we hypothesized that power would be low in these settings.

To quantify the limitations of coevolutionary methods when only a handful of sequences are available, we used a data set of 33 within-species bacterial protein-protein interactions. To systematically determine the parameters that affect performance, we focused on the well-characterized interaction between bacterial histidine kinase A (HisKA) and its response regulator (RR), for which a co-crystal structure and thousands of sequences are available. By sub-sampling HisKA-RR sequence pairs, we show that most methods have appreciable precision or power at low false positive rates for alignments with ∼500 or more sequences. However, the best performing method for a particular analysis will depend on whether power or precision is more important, the number of non-redundant sequences in the alignment, and whether the goal is to find structurally or functionally linked residues (i.e. long range interactions). By expanding this analysis to 32 additional bacterial interactions [30], we showed that these trends generalize beyond the specific example of HiskA and RR. We conclude that coevolution methods are able to identify some residues important for cross-species protein-protein interactions, but this approach will benefit greatly from additional sequence data.

Results

Performance benchmarking of coevolution methods

The coevolutionary methods benchmarked in our analyses fall into three general groups (Table 1). Information-based methods are various flavors of mutual information (MI) between pairs of sites, each considered independently. Direct methods are those that consider pairs of

Table 1 List of methods benchmarked

	Method	APC	Re-weighting	Reference	Software package
Information-based	MI	No	None	[8, 71]	infCalc
	VI			[65]	
	MI_j			[8]	
	MI_{Hmin}				
	MI_w		seq %id	[9]	DCA
Direct	DI	Yes	seq %id, pseudocount		
	DI_{256}			[68]	Code S1 in [68]
	DI_{32}				
	DI_{plm}		seq %id	[72]	plmDCA
	PSICOV		Blosum, pseudocount	[14]	PSICOV
Phylogenetic	CMP_{cor}	No	Downsampling	[10]	CoMap
	CMP_{chg}			[2]	
	CMP_{vol}				
	CMP_{pol}				

Coevolution methods benchmarked fall into three categories. Information-based methods: MI: mutual information [71], VI: variation of information [65], MI_j: MI divided by alignment column-pair entropy, MI_{Hmin}: MI divided by minimum column entropy [8], MI_w: MI with adjusted amino acid probabilities. Direct methods: DI: direct information—MI with re-estimated joint probabilities [9], DI_{256}, DI_{32}: DI using Hopfield-Potts for dimensional reduction (256 and 32 patterns respectively) [68], DI_{plm}: Frobenius norm of coupling matrices in 21-state Potts model using pseudolikelihood maximization [72], PSICOV: sparse inverse covariance estimation [14]. Phylogenetic methods: CoMap *P*-values for four analyses CMP_{cor}: substitution correlation analysis [10], CMP_{pol} for polarity compensation, CMP_{chg} for charge compensation, CMP_{vol} for volume compensation [2] sites in the context of a sparse global statistical model for contacts in the multiple sequence alignment. Phylogenetic methods explicitly use a substitution rate matrix and phylogenetic tree in their calculation of a coevolution statistic. The phylogenetic tree is used to account for the relatedness of the sequences—the observed sequences are themselves correlated due to their shared evolutionary histories. The substitution rate matrix may take into account the biochemical and physical properties of amino acid residues. The main phylogenetic method we report on, CoMap, reports a *P*-value based on internal simulation of independently evolving sites. In this benchmark we use this *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 [5] and applying an Average Product Correction (APC) to remove background noise and phylogenetic signal from “raw” coevolution statistics [8].

To benchmark coevolution methods, we used 33 within-species pairs of proteins with co-crystal structures determined from *E. coli* proteins. These include a set of paired alignments compiled by [30] (Ovch32), plus the histidine kinase-response regulator (HisKA-RR) bacterial two-component system from Procaccini et al. [41], provided by the authors. We included HisKA-RR, because it is a well-characterized interaction with a very large, diverse multiple sequence alignment (8998 sequences for each gene pair) and genetic evidence supporting several interactions. For these reasons, HisKA-RR has also been used previously in coevolutionary analyses [42].

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 (Additional file 13: Figure S1). The 32 alignment pairs in Ovch32 naturally varied in size (range 216–6732 sequences) (Additional file 13: Figure S2).

In addition to the number of sequences in the alignments (N), we consider the phylogenetic diversity (PD [43]) of the alignments—also captured in the effective number of sequences (N_{eff}) as calculated by PSICOV [14], 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.

For each pair of multiple sequence alignments from two interacting proteins, we 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 1. We then used coevolution scores to predict inter-domain pairs of amino acid residues that are less than 8 angstroms (Å) to each other, measured between C_βs, in the representative co-crystal structure (See Methods and Table 2).

We evaluated performance using power (also called recall, sensitivity, and true positive rate (TPR)) (Eq. 1) and precision (also called positive predictive value (PPV)) (Eq. 3) at a range of low false positive rates (FPR)—the proportion of negatives falsely predicted as positives (Eq. 2). The false positive rate is equivalent to 1 - specificity. 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. We 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}) (Eq. 4), maximum ϕ (ϕ_{max}) (Eq. 5).

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$PPV = \frac{TP}{TP + FP} \tag{3}$$

$$F_1 = \frac{2 \cdot PPV \cdot TPR}{TPR + PPV} \tag{4}$$

$$\phi = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \tag{5}$$

We also evaluated performance using two stricter definitions of contacts. First, we defined contacts as residue-pairs with less than 6Å between their closest non-hydrogen atoms. We then evaluated performance in

Table 2 TP: True positive, FP: False positive, TN: True negative, FN: False negative

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

the HisKA-RR sub-alignments using a definition of contacts that, in addition to spatial proximity ($C_\beta < 8$Å), requires biochemical evidence for the role of the contacting residues in determing ortholog- and paralog- specificity of the interaction (i.e. reducing cross-talk between orthologous and paralagous interacting proteins). A list of such residues in representative sequences is found in Casino et al. [44], Li et al. [45], Haldimann et al. [46], Skerker et al. [47], and Laub and Goulian [48]. Trends in the results were generally similar across these choices of definition for true interactions, but we observed some differences in performance between definitions when the false positive rate (FPR) is controlled (Additional file 13: Figure S8 and S10).

Physically interacting sites can be accurately detected in large sequence alignments

Our 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 [3, 17], specifically when the alignments are deeper than they are long [25, 26]. We capture this rectangular quality in the statistic N_{eff}/L, where N_{eff} is the effective number of sequences as calculated by PSICOV [14] and L is the total number of columns in both alignments. We observe similar trends when using the number of sequences (N) or their phylogenetic diversity (PD) [43], rather than N_{eff}/L, to compare performance.

Both power and precision improve with increasing N_{eff}/L for nearly all coevolutionary methods in the HisKA-RR data set (Fig. 1). 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" ($N_{\mathrm{eff}}/L = 1.0$) alignments. Additionally, the performance metrics $f_{\max}$ and $\phi_{\max}$ show that there are no score thresholds (i.e. the strictness of predictions) that achieve both high precision and power in alignments with $N_{\mathrm{eff}}/L \lesssim 3.0$ (Additional file 13: Figure S15-S17). Despite the smaller range in N_{eff}/L values, these performance trends are also observed across the Ovch32 alignments (Additional file 13: Figure S11 and S19).

However, in the HisKA-RR alignment, we 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 % (Additional file 13: Figure S10, Additional file 11). 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 % (Additional file 13: Figure S8, Additional file 12). However, MI_j has drastically lower power at FPR < 0.1 % (Additional file 13: Figure S9). 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.

CoMap performance is an interesting case because, in contrast to DI, DI_{plm}, and PSICOV, it was not initially designed to find contacting residues, rather a mix of both short and long-range interactions. In the smallest alignments (5 sequences) we tested, we occasionally observe CMP_{chg} has higher power than the Direct methods (Mann-Whitney U $P = 0.003$). However, its lower performance in other alignments may indicate that it is identifying a set of coevolving residue pairs that partially overlap with contacting residues. Additionally, a filtering step necessary to run CoMap on large alignments may be limiting its performance (See Methods). It remains to explore whether CoMap can be used to prioritize residue pairs predicted by the other methods for functional assays.

Finally, we 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 % (Additional file 13: Figure S24, Additional file 10). 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 (Additional file 13: Figure S25).

Diversity of sequences is important for accurately detecting contacts

The diversity of residues within the individual alignment columns that make up each pair is another important factor to consider. To explore this, we assessed performance among column pairs with respect to their marginal entropies. We computed power and precision separately for each rate category group (See Additional file 13: Supplemental Methods). 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 (Fig. 2). We conclude that 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

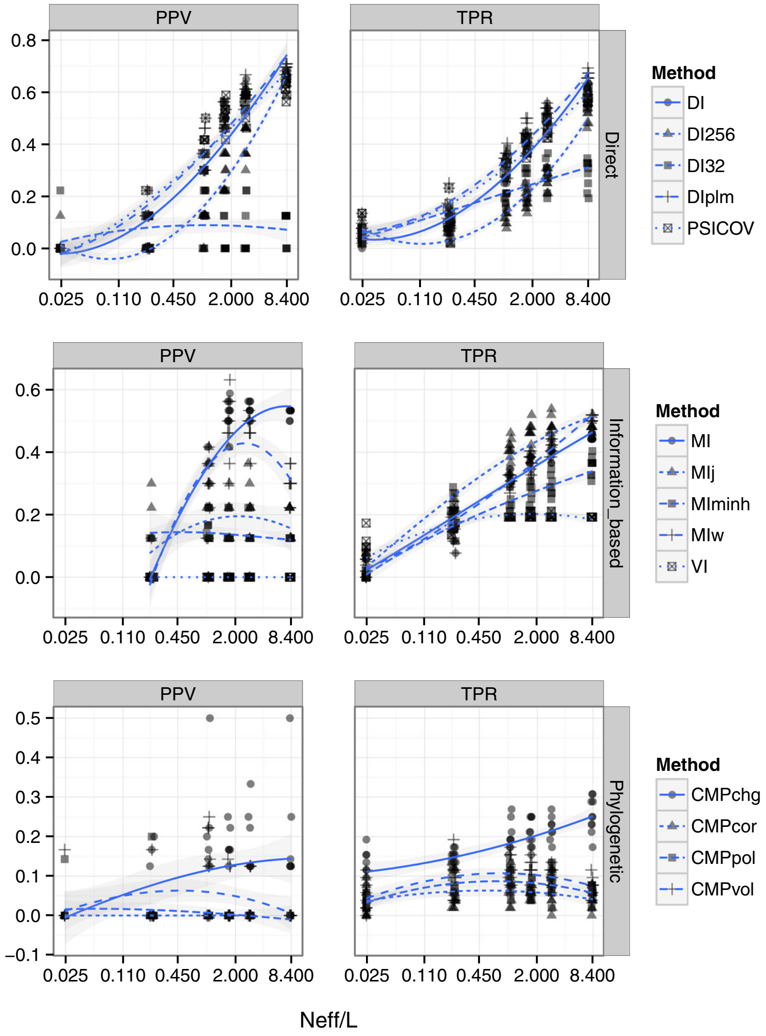

Fig. 1 Coevolution statistics differ in their ability to detect residue contacts in HisKA-RR sub-alignments. Direct methods benefit from larger, more diverse alignments. Left: Precision (PPV) at false positive rate (FPR) < 0.1 %. Right: Power (TPR) at false positive rate (FPR) < 5 %. Blue lines indicate a loess fit to each method, 95 % confidence intervals are shown in gray. See Abbreviations and Table 1 for abbreviations

accuracy of coevolution analyses, with most methods performing best when coevolving residue pairs have similar substitution rates.

To investigate whether higher power in larger alignments results primarily from the number sequences per se or depends upon the diversity of the sequences, we compared the performance across alignments with different diversity values but the same number of sequences. We quantified diversity using phylogenetic diversity (PD) [43] and the effective number of sequences as calculated by PSICOV (N_{eff}) [14] (Additional file 13: Figure S5 and S6).

For HisKA-RR sub-alignments, we 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), we tested whether the false positive rate correlates with PD using Spearman's rho. Few methods had uncorrected P-values < 0.05 and none did when controlling for the 336 comparisons (smallest uncorrected P: 1.73e-3; ρ: 0.85 for MI_j at 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, N_{eff}: 5), under the normal distribution, the false positive rate is better controlled in diverse alignments. However, under the empirical null,

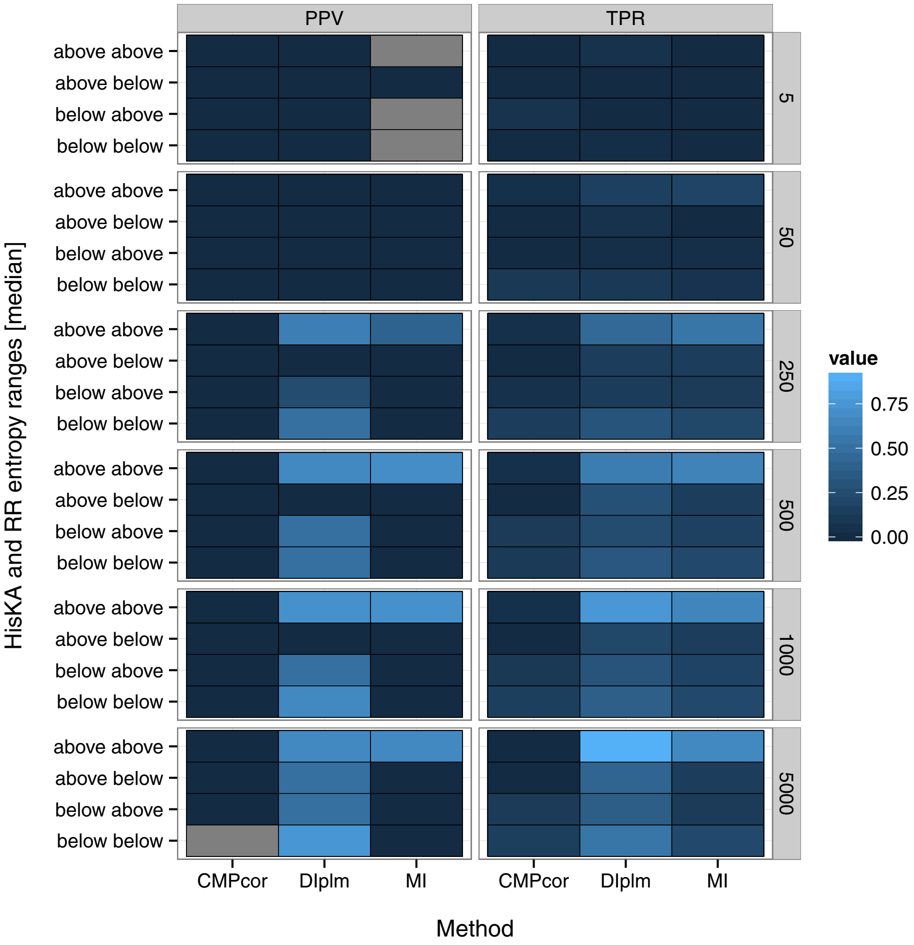

Fig. 2 Detecting coevolving alignment columns is easier when individual alignment columns have similar levels of variation. Column pairs in the HisKA-RR sub-alignments are parsed according to above- or below- median entropy for each size alignment size (number of sequences: N). Left: Median precision (PPV) at FPR < 0.1 %. Right: Median power (TPR) at FPR < 5 %. See Abbreviations and Table 1 for abbreviations

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) we generated sub-alignments by random sampling and therefore only explored a range of phylogenies close to the typical diversity for each alignment size. We observe fairly strong correlations between cutoff-independent performance metrics and N_{eff} (and also 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 weakly negatively correlated with the same performance metrics (Additional file 8). However, accounting for alignment length (with N_{eff}/L) reveals that there is a positive relationship between alignment depth and performance. (Additional file 9, Additional file 13: Figure S5 and S7) 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 ϕ_{max}. Conversely, the third smallest N_{eff} corresponds to the third best ϕ_{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.

Choice of null distribution affects performance

The previous results show performance based on the known HisKA-RR structure. In practice, when applying the methods in our study the structure usually is not known. 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 by chance under the null distribution. A variety of null distributions are commonly used, including theoretical limiting distributions [8, 49, 50], the empirical distribution of observed scores (under the assumption that most pairs of sites are not coevolving) [51], and parametric, semiparametric, and non-parametric bootstrap distributions [10, 52]. Theoretical and empirical nulls are computationally inexpensive compared to bootstrap methods, which require accurately simulating thousands of large data sets (See Additional file 13: Supplemental Text).

We used our sampled sub-alignments of HisKA-RR and the Ovch32 alignments [30] 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, we first employed the non-contact pairs of residues to assess if the FPR was truly controlled or not at given target FPRs (α) of 5 % and 0.1 %.

The normal distribution can be used as theoretical null for mutual information and its normalized variants. Under this assumption, coevolution scores are standardized to Z-scores and compared to upper quantiles of the standard normal distribution (mean = 0, variance = 1). We then used the resulting upper-tail *P*-values (P_{normal}) to predict contacting residue pairs. We found that nominal FPRs using this approach consistently exceed the target FPR across the range of N_{eff}/L values in both the HisKA-RR sub-alignments and the Ovch32 alignments [30] (Fig. 3; Additional file 13: Figure S12-S14). In general, as N_{eff}/L increases, the nominal FPR for Direct methods decreases while it increases in Information-based methods, suggesting that Direct methods truly benefit from deeper alignments. Nominal FPRs were observed to be as great as twice to 24 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

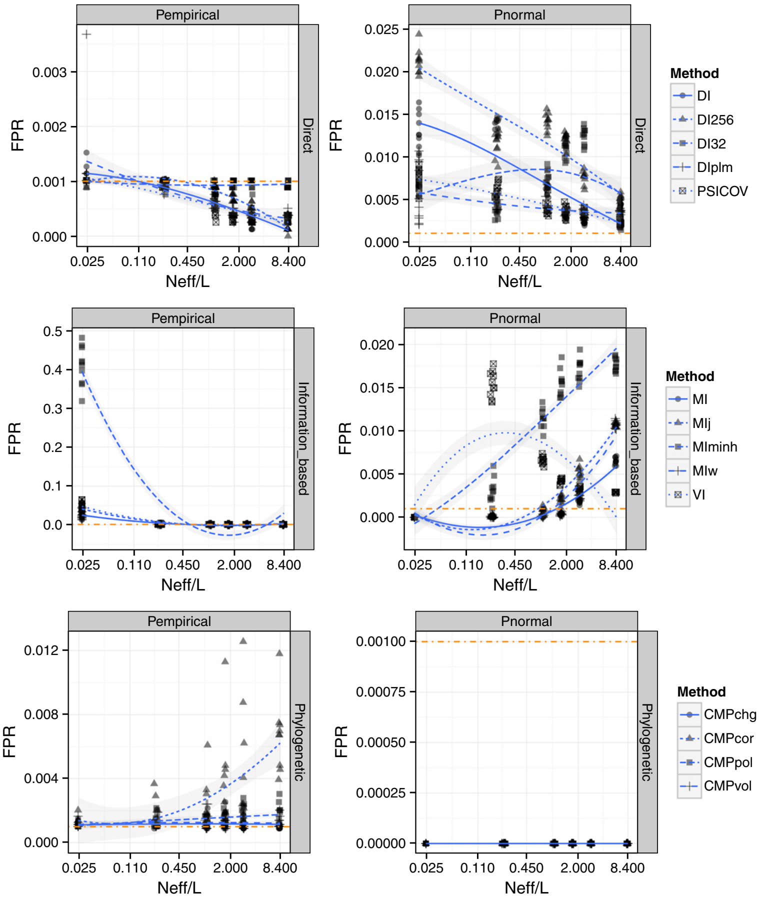

Fig. 3 Null distributions for coevolution statistics differ in their control of the false positive rate (FPR). Nominal FPRs for a given target FPR 0.1 % (Dashed orange line) are shown for the HisKA-RR sub-alignments. Left: Nominal FPRs using the empirical distribution of score ranks as the null distribution (i.e. using $P_{empirical}$). Right: Nominal FPRs assuming standardized scores have a standard normal null distribution (i.e. using P_{normal}). Blue lines indicate a loess fit for each method, 95 % confidence intervals are shown in gray. See Abbreviations and Table 1 for abbreviations

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, structural constraints affecting non-contacting residue pairs, or coevolution scores not being normally distributed (Additional file 13: Figure S30-S32). A theoretical null (e.g. non-central gamma [53]) that is parameterized for individual column pairs may therefore be more appropriate (See Additional file 13: Supplemental Text) and warrants future investigation.

Another choice of null distribution is the observed empirical distribution of the coevolution statistics. 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 also 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). Although, we found that the empirical null distribution does produce nominal FPRs that exceed target FPRs (Fig. 3; Additional file 13: Figure S13). As the proportion of contacts increases, the $P_{empirical}$-values become more conservative (Additional file 13: Figure S26 and S27). The Direct methods best control the nominal FPR in both sets of alignments, marginally exceeding the target FPR in only a couple of cases (maximum FPR/α = 3.68). The Information-based methods controlled the FPR below 1.58 times α in the Ovch32 alignments [30], however the HisKA-RR sub-alignments reveal that at $N_{eff}/L < 0.3$, control of the FPR is lost, especially in MI_{Hmin} (FPR/$\alpha > 400$). The Phylogenetic method that consistently exceeded the target FPR was the CoMap correlation analysis (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. In some methods like MI_{minh}, 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 diminish the anti-conservativeness of $P_{empirical}$.

Cross-species case study: applying coevolution methods to Vif-A3G identifies some residues known to affect host-virus interactions

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 [54, 55], we 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, we were able to identify 16 pairs of sequences (N_{eff} = 10.0) from different primates (A3G orthologs) and their lentiviruses (Vif orthologs) in public databases (Additional file 5). Our benchmarking results on HisKA-RR indicate that such small protein families push the useful limits of the coevolution statistics we tested ($N_{eff}/L = 0.014$). The low sequence diversity of A3G (N_{eff} = 3.04) within primates compared to Vif (N_{eff} = 11.3) within primate lentiviruses also presents challenges. Hence, we 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 [56–59] and human A3G [60, 61] (Table 3). For this analysis, we therefore take the residues in biochemically-validated regions to be positives even though they might

Table 3 Important residues for the Vif-A3G interaction

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

HIV1 Vif [56–59]. Human A3G [60, 61]

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, we explore whether any clues can be identified in the limited data that describes the coevolutionary history of the Vif-A3G residues.

First, we computed a subset of 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, we 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, we 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 (Additional file 7). From our 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. we therefore chose to identify a significance threshold for each method that maximizes precision on the known functional sites in each protein. Then, we 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) (Additional file 13: Figure S34). 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_{pol}, CMP_{cor}, 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.

We observed similarly low performance on A3G (Fig. 4). Encouragingly, we note that positions 128-130 are correctly identified by multiple methods (Fig. 5). 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 [54, 55]. Position 128, that also provides species-specific resistance, is thought to be more recent [54, 55, 62]. 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 [55], 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, we 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 [63] (Additional file 7, Additional file 13: Figure S36). In [63], the authors are able to see that a critical subset of the Vif positives is solvent-exposed. We re-evaluated performance with only these residues

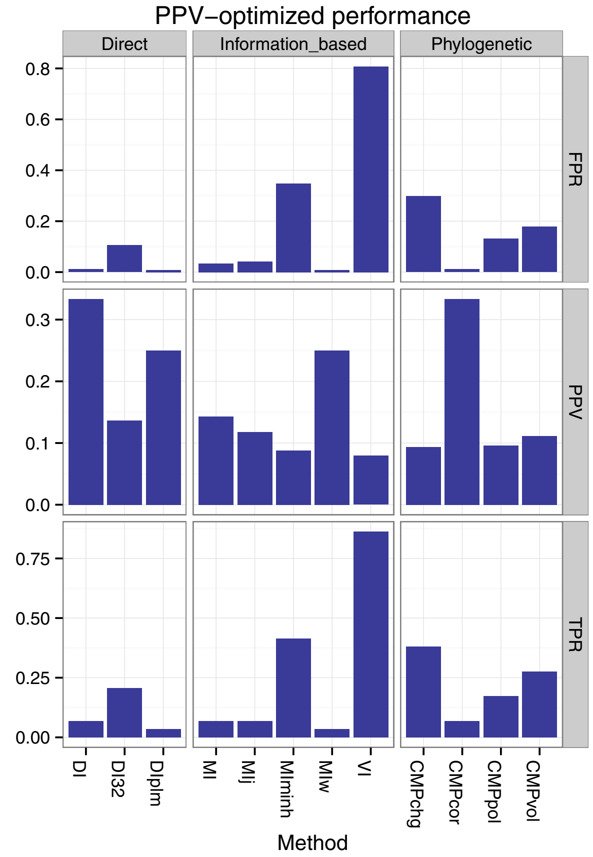

Fig. 4 Power (TPR), precision (PPV), and false positive rate (FPR) for predicting antiviral protein A3G residues (not pairs) essential for interacting with its viral antagonist Vif at $P_{empirical} < \alpha$ thresholds that maximize PPV for each coevolution method. Residues defined as positive are taken from previous functional mutation studies in Table 3. See Abbreviations and Table 1 for abbreviations

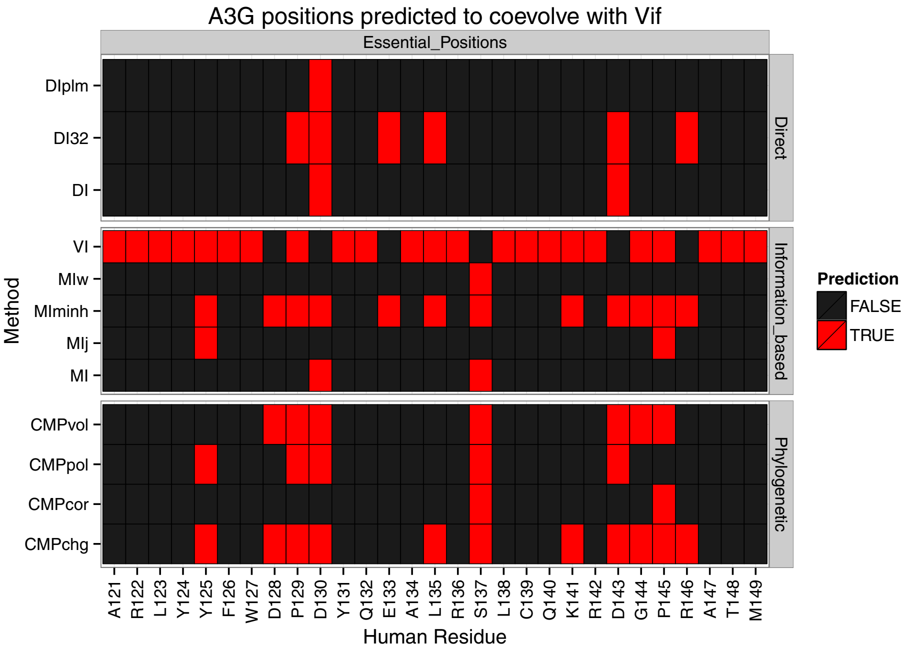

Fig. 5 A3G residues currently known to be essential for binding its viral antagonist Vif. Predictions of residues that coevolve with Vif (red) made at a threshold that maximizes precision (PPV) using currently known essential residues identify position D130 which was previously implicated in species specific resistance

as the positives (Table 3). There is poor precision to identify the putative solvent-exposed interface among the methods; CMP_{cor} at 40 % and CMP_{vol} at 10 % are the only methods with precision $>$ 6 % (Additional file 13: Figure S35).

Our 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 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 MI_{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.

A toolkit for inter-molecular coevolution analysis

Due to the diversity of coevolution methods and the time spanned during which they were developed, it is no surprise that they vary widely in the input and output formats they tolerate. Additionally, many of the coevolution methods we tested are computationally expensive, so we prepared our workflow to take advantage of multiprocessing workstations and high performance computing clusters. We outline a few utilities we developed to aid in processing sequences, structures, and coevolution results for benchmarking and making predictions and visualizations.

Our toolkit consists of three parts, (1) a collection of wrappers for running the coevolution programs from the command line and where possible in a Sun Grid Engine super computing environment (https://github.com/aavilahe/coevo_tools), (2) an R package for evaluating performance and calculating $P_{empirical}$ and P_{normal} (https://github.com/aavilahe/coevo_analysis_Rpackage), and (3) pre- and post-processing utilities to facilitate input and output format management, mapping alignments to structural models, and visualizing coevolving residues on protein structures (https://github.com/aavilahe/coevo_analysis_pypackage).

We also implemented the canonical mutual information statistic, the normalizations of mutual information in Martin et al. [64], and VI, the information theoretic distance described by Meila [65] (https://github.com/aavilahe/infcalc).

File formats

The coevolution methods we tested accept three different file formats and alignments as two separate files or one horizontally concatenated file. The different formats, (fasta, phylip, raw reads) store more or less meta-data and have limits on the length of sequence identifiers.

Our `coevo` package at https://github.com/aavilahe/coevo_analysis_pypackage depends on the Biopython library and contains many auxiliary functions and executable python scripts for input file preparation.

A typical processing step may involve truncating sequence identifiers when converting between sequence formats, taking care that they remain informative and unique. For example:

```
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')
```

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
```

The coevolution methods return tab, space, or comma delimited output with and without headers. The scores returned are often indexed by column numbers of the concatenated alignment and not the original two alignments of interest, and can be numbered starting from 0 or 1.

The `scores` module in our `coevo` package includes definitions for the various formats we encountered, extracts the relevant indices and scores, optionally merges results from different methods, and processes them to a standard tab delimited format with appropriate headers and indices that correspond to the alignments of interest. For example:

```
import coevo
import coevo.scores as cs

def merge_tabs(tab_list):
    '''Outer join tab_list on indices'''

    return cs.merge_dfs(tab_list, left_index = True,
                                  right_index = True,
                                  how = 'outer')

# length of left protein
HK_length = 87

# define two formats, optionally assign a suffix
mi_fmt = cs.Format(prog = 'infCalc', suff = '')
dca_fmt = cs.Format(prog = 'mfDCA', suff = '')
hp32_fmt = cs.Format(prog = 'hpDCA', suff = 'p_32')

# load and standardize coevo results
mi_tab = mi_fmt.load("path/to/HK_RR.out")
dca_tab = dca_fmt.load("path/to/HK_RR.mfDCA")
hp32_tab = hp32_fmt.load("path/to/HK_RR.hpDCA")

# drop intraprotein column-pairs
mi_tab = mi_tab.drop_intraprotein(mi_tab, HK_length)
dca_tab = dca_tab.drop_intraprotein(dca_tab, HK_length)
hp32_tab = hp32_tab.drop_intraprotein(hp32_tab, HK_length)

# merge and save to file
merged = merge_tabs([mi_tab, dca_tab, hp32_tab])
cs.write_tab(merged, "path/to/HK_RR.tab")
```

Structure

Another important procedure is to map column numbers from a given alignment to a reference PDB structure. For example, we 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 [66] using `make_attributes.py` (e.g. Additional file 7, Additional file 13: Figure S36 shows Vif residues predicted to coevolve with A3G, each Vif residue is colored by most significant *P*-value out of all A3G residues).

Discussion

In this work we aimed to paint a picture of the performance of emerging methods to identify inter-protein contacts using coevolution and to identify properties of alignments where performance is expected to be best. As previously noted in intra-protein predictions [3, 9, 14], re-weighting of the sequences to account for the underlying phylogeny is important for inter-protein predictions as well, however as the comparison between MI_w and MI shows, it is important to tune the parameters controlling the re-weighting in cases where there are fast evolving alignment columns in an overall conserved protein family. Fortunately, methods that search for direct correlations—using a global statistical model for the sequence alignments—seem to be able to correct for the improper weighting (compare MI_w to DI). These methods are more precise at strict false positive rates than their counterparts especially when the alignments have $N_{eff}/L < 1.0$. However, it may be beneficial to use a faster, MI-based method if the use case allows for a relaxed FPR and is concerned with power versus precision.

We also investigated the use of three null models to control the false positive rate. Counter-intuitively, a null model that explicitly models evolution independently for each alignment fails to control the false positive rate. We believe that our simulated alignments are systematically scoring too low because they fail to capture the correct amount of variation in the observed alignments, resulting in artificially significant *P*-values, except for when the effects of having small alignment sizes results in overly conservative *P*-values. Using a standard normal or the empirical distribution of scores as null models also failed to control the false positive rate, likely due to the correlation structure imposed by the shared evolutionary history of the residues, the distribution of evolutionary rates of the residues, or because asymptotic assumptions do not hold at small sample sizes. Thus, choosing an appropriate *P*-value cutoff in a real analysis when the structure is unknown and alignment depth is shallow still remains a challenge. However, we show that in diverse enough alignments the empirical null successfully controls the false positive rate for Direct methods. As a future direction, we suggest exploring theoretical null distributions that can be parameterized for individual alignment column pairs such as [53] or further improving protein evolution simulators to generate distributions of scores where the evolutionary rates are more similar between the null and alternate hypothesis.

These results are encouraging, but still leave us with the challenge of how to choose an appropriate *P*-value cutoff in a real analysis when the structure is unknown. Since our 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. Additional information to consider when making such modifications should include incorporating alignment properties such as N_{eff}/L, and the expected proportion of contacts expected to exist (Additional file 13: Figure S27; Fig. 3). However, large data sets of many protein interactions are needed in order to be confident in parameters or prior probabilities to be used to correct the *P*-values. 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 (Fig. 3; Additional file 13: Figure S13). As a rule of thumb, the empirical null overall controls the FPR for the Direct methods, however in small alignments (5 sequences or $N_{eff}/L < 0.3$) it can be up to 1.5 times the target FPR. For the purposes of data collection and experimental design, we therefore recommend sequencing phylogenetically deeply enough to attain $N_{eff}/L > 1.0$ to control FPR and > 2.0 to ensure modest TPR and PPV.

A related problem to the one discussed here is to search a large set of protein pairs (within or between species) to determine which ones are interacting. In this setting, coevolution method performance is potentially more important than when predicting contacting residues for known interactions, because the search space will contain so many negatives (i.e., non-interacting pairs). A permissive *P*-value cutoff will lead to a large number of false positives and that may misinform investigators, while being too strict will lead to false negatives that keep potentially important findings hidden. It would be interesting to understand if thresholds and the methods for choosing them generalize to all protein-protein interactions. Different experimental techniques have strengths and weakness in identifying different types of interactions. Interactions may be transient, but highly critical, or tightly binding but too conserved to detect any sequence variation among the sequenced orthologs Mulberry Ideally, we would like to understand what a null model teaches us about phylogeny-induced correlations when structural inter- or intra-protein constraints are minimal, perhaps at an evolutionary stage where a protein interaction is acquired or lost. What can this reveal about the birth and death of protein interactions, regulatory networks, and neofunctionalization? Another challenge for predicting interacting protein pairs from coevolutionary tests is how to summarize statistics for individual pairs of residues to produce a single score for a pair of proteins. Although outside the scope of our work, such a strategy would likely involve comparing tails of score or *P*-value distributions. Deciding on how to define how much of the tail to consider will depend highly on having an estimate of the false positive rate. Based on some preliminary investigations of these questions, we conclude that it is unlikely that cross-species interacting protein pairs can be accurately distinguished from non-interacting pairs on a genome-wide scale.

The progress of high-throughput interaction mapping highlights the need for continued refinement of inter-protein coevolution detection methods. Given that improper re-weighting of sequences can negatively affect power and the false positive rate, perhaps expanding Direct methods to independently obtain sequence weights for each alignment or using an evolution-based probabilistic weight (such as in CoMap or using phylogenetic logistic regression) for unusual variation in each column is a logical next step forward. Another important contribution would be to develop a generalizable null model that can help differentiate contacts when there are very few sequences available for protein families. Furthermore, investigating the correlations among the coevolution statistics themselves in inter-protein data sets could potentially disentangle structural from non-structural coevolutionary forces as well as serving to construct an ensemble method. Comprehensively sequencing orthologous pairs of protein families is a straightforward way to test the usefulness of these future contributions while simultaneously enabling current methods to perform to their fullest.

Conclusion

We benchmarked 13 coevolution methods on 33 protein interactions with associated sequence alignments of varying depths. We conclude that coevolutionary analyses of cross-species protein-protein interactions is largely hindered by a lack of phylogenetically deep protein alignments for many proteins, and furthur demonstrate this in an example case involving an HIV1-human protein interaction. Additionally, we report that commonly used null distributions generally fail to control false positives in coevolutionary analyses, though errors are best controlled by the empirical null in large alignments.

Methods

Multiple sequence alignments

A master alignment of 8998 horizontally concatenated HisKA and RR sequences from Procaccini et al. [41] was graciously provided by the authors (Additional files 2 and 3). From this alignment, aligned sequences were sampled uniformly (each sequence had equal probability of being sampled) to create sub-alignments with 5, 50, 250, 500, 1000, and 5000 sequences. We sampled 10 sub-alignments of each alignment size (number of sequences in sub-alignment), resulting in 60 total alignment pairs (Additional file 4).

The Ovch32 alignments [30] were downloaded from complexes section of the Baker lab website (http://gremlin.bakerlab.org/complexes/PDB_benchmark_alignments.zip) on Aug 29, 2014 (Additional file 1). A stable link is located at the Dryad repository, doi:10.5061/dryad.s00vr/7 [67]. The corresponding structures were downloaded from PDB and processed to obtain contacts between representative protein chains. See Supplemental Files for accessions. Columns comprised of more than 75 % gaps were removed as in [30]. Additionally, only columns mapping to the representative structure were kept.

The CoMap implementation requires a preprocessing step to remove sequence redundancy (a data munging alternative to sequence weighting). This additional step was also necessary to prevent buffer underflow errors when evaluating likelihoods in very large input trees. Therefore, all alignments with more than 200 sequences were culled to contain the 200 most diverse sequences before being passed to CoMap. The sub-alignment used corresponds to the 200-leaf sub-tree that maximizes PD for each original input alignment and tree.

Measuring coevolution

The coevolution methods benchmarked are listed in Tables 1 and 4. Wrappers for the Direct methods are provided in our coevo_tools code repository to facilitate running from the command line (See Supplement for details). For methods in the plmDCA, mfDCA and hpDCA packages, MATLAB, or the MATLAB runtime executable is required as well as various MATLAB Toolbox dependencies and licenses. Default settings were used for all methods, including sequence re-weighting and APC. DI_{32} and DI_{256} are variations of DI in the hpDCA package with an additional parameter for tuning dimensionality reduction, “p”, set to 32 and 256 respectively as it had no default (a selection from a parameter search in [68]).

Evaluating coevolution performance

For each method, coevolution scores for pairs of amino acid positions were used to predict inter-domain pairs of amino acid residues that are close to each other in the representative co-crystal structure (PDB ID: 3DGE).

As previously described in Ezkurdia et al. [69], Monastyrskyy et al. [70], Jones et al. [14], and to be consistent with Morcos et al. [9], we define *positives* as pairs of alignment positions mapping to amino acid residues whose beta carbons (C_β) are less than 8 angstroms apart in 3DGE. All other pairs of alignment positions are considered *negatives*.

Table 4 Versions and sources of coevolution methods benchmarked

	Method	Software package	Version	URL
Information-based	MI	infCalc	v0.1.2	https://github.com/aavilahe/infcalc
	VI			
	MI_j			
	MI_{Hmin}			
	MI_w	DCA	"2011/12"	http://dca.rice.edu/portal/dca/download
Direct	DI			
	DI_{256}	Code S1 in [68]	"2013"	http://doi.org/10.1371/journal.pcbi.1003176.s002
	DI_{32}			
	DI_{plm}	plmDCA	symmetric_v2	http://plmdca.csc.kth.se/
	PSICOV	PSICOV	V1.09	http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/
Phylogenetic	CMP_{cor}	CoMap	1.5.1b5	http://home.gna.org/comap/doc/html/index.html
	CMP_{chg}			
	CMP_{vol}			
	CMP_{pol}			

We considered the following two alternative definitions of *positives*:

- Closest non-hydrogen atom-atom distance between residues is less than 6 angstroms [14]
- C_β distance is less than 8 angstroms *and* at least one residue is mentioned as important in determining specificity of the HisKA-RR interaction in [44–48].

Residue pairs are predicted as coevolving if their scores or *P*-values are above a given threshold (eg. top 1 %, $P < 0.05$) (Table 2).

Phylogenetic diversity

Phylogenetic diversity (PD) is calculated as the sum of the branch lengths in a tree built from the concatenated multiple sequence alignment of both proteins. Trees were built using FastTree (version2.1.7 SSE3) with options `-gamma -nosupport -wag`.

Additional files

Additional file 1: PDB IDs from [67]. (CSV 0.46 kb)

Additional file 2: Accessions for HisKA-RR alignment. (CSV 796 kb)

Additional file 3: HisKA-RR master alignment. (ZIP 1136 kb)

Additional file 4: Sequence identifiers for the 60 subsampled alignments from the HisKA-RR master alignment. (CSV 5611 kb)

Additional file 5: Accessions for Vif-A3G analysis. (CSV 0.73 kb)

Additional file 6: Uniprot identifiers for lentivirus proteins in HIV1-human interactome analysis. (CSV 0.803 kb)

Additional file 7: $P_{empirical}$ for Vif residues as an attributes file for use with UCSF Chimera [66]. (CSV 163 kb)

Additional file 8: Spearman test for correlations between cutoff independent metrics (ϕ_{max}, f_{max}, auPR, auROC) and N_{eff} for Ovch32 and HisKA-RR data sets. (CSV 0.24 kb)

Additional file 9: Spearman test for correlations between cutoff independent metrics (ϕ_{max}, f_{max}, auPR, auROC) and N_{eff}/L for Ovch32 and HisKA-RR data sets. (CSV 0.24 kb)

Additional file 10: Spearman test for correlations between precision (PPV) and the proportion of contacts for Ovch32 and HisKA-RR data sets. (CSV 0.13 kb)

Additional file 11: Mann-Whitney tests for difference in precision between MI and other methods for $N_{eff}/L > 1.6$ and FPR < 0.1 % in the HisKA-RR sub-alignments, with contacts defined from experimental studies. (CSV 0.48 kb)

Additional file 12: Mann-Whitney tests for difference in power between MI_J and other methods for $N_{eff}/L > 1.6$ and FPR < 5 % in the HisKA-RR sub-alignments, with contacts defined from experimental studies. (CSV 0.47 kb)

Additional file 13: Figure S1. HisKA-RR. Number of effective sequences (N_{eff}) versus number of sequence (N) in the 60 sub-sampled HisKA-RR alignments. Dashed line indicates the diagonal. Blue line indicates a linear fit with 95 % confidence intervals in gray. **Figure S2.** Ovch32. Number of effective sequences (N_{eff}) versus number of sequence (N) in the Ovch32 alignments. Dashed line indicates the diagonal. Blue line indicates a linear fit with 95 % confidence intervals in gray. **Figure S3.** Distribution of C_β distances in HisKA-RR interaction (PDB: 3DGE). **Figure S4.** Distribution of C_β distances in Ovch32 interactions [67] (See supplemental file for PDB accessions). **Figure S5.** Ovch32. Precision (PPV) versus N_{eff} at FPR < 0.1 %. Blue lines indicate a loess fit to each method, 95 % confidence intervals are shown in gray. **Figure S6.** Ovch32. Power (TPR) versus N_{eff} at FPR < 5 %. Blue lines indicate a loess fit to each method, 95 % confidence intervals are shown in gray. **Figure S7.** Ovch32. ϕ_{max} versus N_{eff}. Blue lines indicate a loess fit to each method, 95 % confidence intervals are shown in gray. **Figure S8.** HisKA-RR alt.. Power (TPR) vs N_{eff}/L at FPR < 5 %. A stricter definition of positives, defined experimentally in [46–48] is used. Blue lines indicate a loess fit to each method, 95 % confidence intervals are shown in gray. **Figure S9.** HisKA-RR alt.. Power (TPR) vs N_{eff}/L at FPR < 0.1 %. A stricter definition of positives, defined experimentally in [46–48] is used. Blue lines indicate a loess fit to each method, 95 % confidence intervals are shown in gray. **Figure S10.** HisKA-RR alt.. Precision (PPV) vs N_{eff}/L at FPR < 0.1 %. A stricter definition of positives, defined experimentally in [46–48] is used. Blue lines indicate a loess fit to each method, 95 % confidence intervals are shown in gray. **Figure S11.** Ovch32. Power (TPR) at FPR < 5 % and Precision (PPV) at FPR < 0.1 % versus N_{eff}/L. Blue lines indicate a loess fit to each method, 95 % confidence intervals are shown in gray.

Figure S12. HisKA-RR. Nominal false positive rate (FPR) for target FPR 5 %. **Figure S13.** Ovch32. Nominal false positive rate (FPR) for target FPR 0.1 %. **Figure S14.** Ovch32. Nominal false positive rate (FPR) for target 5 %. **Figure S15.** HisKA-RR. $\phi_{\max}$. **Figure S16.** HisKA-RR. $F_{\max}$. **Figure S17.** HisKA-RR. Area under precision-recall curve. **Figure S18.** HisKA-RR. Area under ROC curve. **Figure S19.** Ovch32. $\phi_{\max}$. **Figure S20.** Ovch32. $F_{\max}$. **Figure S21.** Ovch32. Area under precision-recall curve. **Figure S22.** Ovch32. Area under ROC curve. **Figure S23.** HisKA-RR. Median precision (PPV) at FPR <0.1 % and median power (TPR) at FPR < 5 % per rate categories of individual alignment columns. Rate categories are defined as above- and below- median entropy for the HisKA and RR columns in each set of 10 alignments of equal size (number of sequences (N)). **Figure S24.** Ovch32. Precision (PPV) versus the proportion of contacting pairs of residues in each interaction (i.e. contacting pairs divided by all pairs of residues) at FPR < 0.1 %. **Figure S25.** Ovch32. Precision (PPV) versus the proportion of contacting pairs of residues in each interaction (i.e. contacting pairs divided by all pairs of residues) at FPR < 5 %. **Figure S26.** Ovch32. False positive rate (FPR) versus the proportion of contacting pairs of residues in each interaction (i.e. contacting pairs divided by all pairs of residues) at $P < 0.05$ **Figure S27.** Ovch32. False positive rate (FPR) versus the proportion of contacting pairs of residues in each interaction (i.e. contacting pairs divided by all pairs of residues) at $P < 0.001$ **Figure S28.** HisKA-RR. The phylogenetic methods CTMP and Spidermonkey successfully ran on a subset of our alignments. Power (TPR) at FPR < 5 % and precision (PPV) at FPR < 0.1 %. Select methods are included for comparison. Blue line indicates a linear fit with 95 % confidence intervals in gray. **Figure S29.** HisKA-RR. The phylogenetic method CTMP and Spidermonkey successfully ran on a subset of our alignments. Nominal false positive rate (FPR) at target FPR 0.1 %. Select methods are included for comparison. Blue line indicates a linear fit with 95 % confidence intervals in gray. **Figure S30.** HisKA-RR. Quantile quantile plots of standardized coevolution scores are not always normally distributed. Scores are from 10 alignments with 5 sequences. **Figure S31.** HisKA-RR. Quantile quantile plots of standardized coevolution scores are not always normally distributed. Scores are from 10 alignments with 500 sequences. **Figure S32.** HisKA-RR. Quantile quantile plots of standardized coevolution scores are not always normally distributed. Scores are from 10 alignments with 5000 sequences. **Figure S33.** HisKA-RR. $P_{boostrap}$ fails to control the FPR except for PSICOV at target FPR < 5 % in HisKA-RR alignments. Eliminating residue pairs with large simulation errors shows PSICOV and MI$_{\text{Hmin}}$ are most robust to variation at individual sites. See Misc. Abbreviations and Table 1 for abbreviations. **Figure S34.** Vif. Power (TPR), precision (PPV), and false positive rate (FPR) for predicting viral protein Vif residues (not pairs) essential for interacting with its host target A3G at $P_{empirical} < \alpha$ thresholds that maximize PPV for each coevolution method. Residues defined as positive are taken from previous functional mutation studies in Table 3. See Abbreviations and Table 1 for abbreviations. **Figure S35.** Vif. Power (TPR), precision (PPV), and false positive rate (FPR) for predicting viral protein Vif residues (not pairs) essential for interacting with its host target A3G at $P_{empirical} < \alpha$ thresholds that maximize PPV for each coevolution method. Residues defined as positive are taken from previous functional mutation studies in Table 3. See Abbreviations and Table 1 for abbreviations.Vifcrit PPVoptbars **Figure S36.** Residues (red) on viral protein Vif (light blue) that are predicted to coevolve with it host target A3G (structure unknown). Cofactors are shown in gray. Predictions are made at a threshold that maximizes precision (PPV) using **A** known essential residues (Table 3) using **B-D** MI, **Figure S37.** HIV1-human. Distinguishing HIV1-human interactors from a protein pairs in a permuted network is difficult with small N$_{\text{eff}}$/L. $\phi_{\max}$ across a the number of predicted coevolving column-pairs per protein-pair versus $P\hat{\ }p_{empirical}$ threshold for making column-pair predictions. Blue line indicates a linear fit with 95 % confidence intervals in gray. **Figure S38.** HIV1-human. N$_{\text{eff}}$/L distribution of alignments in HIV1-human interactors The minimum N$_{\text{eff}}$/L seen in the HisKA-RR (red) and Ovch32 (orange) data sets is marked. (ZIP 28364 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AA carried out the analysis. AA and KSP designed the analysis and wrote the manuscript. All authors read and approved the final manuscript.

Authors' information

AA is a Bioinformatics graduate student at the University of California San Francisco in the laboratory of Dr. Katherine S. Pollard.
KSP is a Senior Investigator at the Gladstone Institutes and Professor of Epidemiology and Biostatistics at the University of California San Francisco.

Acknowledgements

We thank Martin Weigt for providing HisKA and RR alignments and for providing links to DCA source code. We also thank Julien Dutheil for help running CoMap correctly. We greatly appreciate Art Poon for his help in adapting Spidermonkey batch files for amino acid analyses. This work was supported by a National Institutes of Health bioinformatics training grant, a UCSF Graduate Research Mentorship Fellowship, institutional funding from Gladstone Institutes, and a gift from the San Simeon Fund.

Author details

[1]Bioinformatics Graduate Program, University of California, San Francisco, USA. [2]Gladstone Institute of Cardiovascular Disease, University of California, San Francisco, USA. [3]Department of Epidemiology and Biostatistics, University of California, San Francisco, USA. [4]Institute for Human Genetics, University of California, San Francisco, CA, 94158, USA.

Received: 20 February 2015 Accepted: 17 July 2015
Published online: 25 August 2015

References

1. Yip KY, Patel P, Kim PM, Engelman DM, McDermott D, Gerstein M. An integrated system for studying residue coevolution in proteins. Bioinformatics. 2008;24(2):290–2. doi:10.1093/bioinformatics/btm584.
2. Dutheil J, Galtier N. Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol Biol. 2007;7:242. doi:10.1186/1471-2148-7-242.
3. Dutheil JY. Detecting coevolving positions in a molecule: why and how to account for phylogeny. Brief Bioinform. 2012;13(2):228–43. doi:10.1093/bib/bbr048.
4. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet. 2013;14(4):249–61. doi:10.1038/nrg3414.
5. Buslje CM, Santos J, Delfino JM, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics. 2009;25(9):1125–31. doi:10.1093/bioinformatics/btp135.
6. Fares MA, Travers SA. A novel method for detecting intramolecular coevolution adding a further dimension to selective constraints analyses. Genetics. 2006;173(1):9–23. doi:10.1534/genetics.105.053249.
7. Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, Talsania S, et al. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc Natl Acad Sci USA. 1153;108(28):0–5. doi:10.1073/pnas.1105315108.
8. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008;24(3):333–40. doi:10.1093/bioinformatics/btm604.
9. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA. 2011;108(49):E1293–301. doi:10.1073/pnas.1111471108.
10. Dutheil J, Pupko T, Jean-Marie A, Galtier N. A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol. 2005;22(9):1919–28. doi:10.1093/molbev/msi183.
11. Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol. 1999;287(1):187–98. doi:10.1006/jmbi.1998.2601.
12. Caporaso JG, Smit S, Easton BC, Hunter L, Huttley GA, Knight R. Detecting coevolution without phylogenetic trees? tree-ignorant metrics of coevolution perform as well as tree-aware metrics. BMC Evol Biol. 2008;8(327). doi:10.1186/1471-2148-8-327.

13. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA. 2009;106(1):67–72. doi:10.1073/pnas.0805923106.
14. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012;28(2):184–90. doi:10.1093/bioinformatics/btr638.
15. Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol. 2010;6(1):e1000633. doi:10.1371/journal.pcbi.1000633.
16. Delaporte E, Wyler Lazarevic CA, Iten A, Sudre P. Large measles outbreak in geneva, switzerland, january to august 2011: descriptive epidemiology and demonstration of quarantine effectiveness. Euro Surveill Bull. 2013;18(6). http://www.ncbi.nlm.nih.gov/pubmed/23410259.
17. Clark GW, Ackerman SH, Tillier ER, Gatti DL. Multidimensional mutual information methods for the analysis of covariation in multiple sequence alignments. BMC Bioinformatics. 2014;15(1):157. doi:10.1186/1471-2105-15-157.
18. McLaughlin Jr RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature. 2012;491(7422):138–42. doi:10.1038/nature11500.
19. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: Introducing the d ^{2} concept. Ann Rev Biophys. 2008;37(1):215–46. doi:10.1146/annurev.biophys.37.032807.125924.
20. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005;6(3):197–208. doi:10.1038/nrm1589.
21. Ben-Shem A, Garreau de Loubresse N, Melnikov S, Jenner L, Yusupova G, Yusupov M. The structure of the eukaryotic ribosome at 3.0 å resolution. Science (New York, NY). 2011;334(6062):1524–9. doi:10.1126/science.1212642.
22. Lasker K, Forster F, Bohn S, Walzthoeni T, Villa E, Unverdorben P, et al. Molecular architecture of the 26s proteasome holocomplex determined by an integrative approach. Proc Natl Acad Sci USA. 2012;109(5):1380–7. doi:10.1073/pnas.1120559109.
23. Jager S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, et al. Global landscape of HIV-human protein complexes. Nature. 7381;481:365–70. doi:10.1038/nature10719.
24. Vinayagam A, Zirin J, Roesel C, Hu Y, Yilmazel B, Samsonova AA, et al. Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. Nat Methods. 2013;11(1):94–9. doi:10.1038/nmeth.2733.
25. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci USA. 1567;110(39):4–9. doi:10.1073/pnas.1314045110.
26. Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3d contacts and structures of protein complexes. Elife. 2014;3. doi:10.7554/eLife.03430.
27. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3d structure computed from evolutionary sequence variation. PloS ONE. 2011;6(12):e28766. doi:10.1371/journal.pone.0028766.
28. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell. 2012;149(7):1607–21. doi:10.1016/j.cell.2012.04.012.
29. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol. 2012;30(11):1072–80. doi:10.1038/nbt.2419.
30. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife. 2014;3:e02030. doi:10.7554/eLife.02030.
31. Juan D, Pazos F, Valencia A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Natl Acad Sci USA. 2008;105(3):934–9. doi:10.1073/pnas.0709671105.
32. Gershoni M, Fuchs A, Shani N, Fridman Y, Corral-Debrinski M, Aharoni A, et al. Coevolution predicts direct interactions between mtDNA-encoded and nDNA-encoded subunits of oxidative phosphorylation complex i. J Mol Biol. 2010;404(1):158–71. doi:10.1016/j.jmb.2010.09.029.
33. Clark NL, Gasper J, Sekino M, Springer SA, Aquadro CF, Swanson WJ. Coevolution of interacting fertilization proteins. PLoS Genet. 2009;5(7): e1000570. doi:10.1371/journal.pgen.1000570.
34. Yeang CH, Haussler D. Detecting coevolution in and among protein domains. PLoS Comput Biol. 2007;3(11):e211. doi:10.1371/journal.pcbi.0030211.
35. Morris JH, Knudsen GM, Verschueren E, Johnson JR, Cimermancic P, Greninger AL, et al. Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. Nat Protoc. 2014;9(11):2539–54. doi:10.1038/nprot.2014.164.
36. Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U. Yeast two-hybrid, a powerful tool for systems biology. Int J Mol Sci. 2009;10(6): 2763–88. doi:10.3390/ijms10062763.
37. Vidal M, Fields S. The yeast two-hybrid assay: still finding connections after 25 years. Nat Methods. 2014;11(12):1203–6. http://www.nature.com/articles/nmeth.3182.
38. Michnick SW, Ear PH, Landry C, Malleshaiah MK, Messier V. Protein-fragment complementation assays for large-scale analysis, functional dissection and dynamic studies of protein-protein interactions in living cells. Methods Mol Biol. 2011;756:395–425. doi:10.1007/978-1-61779-160-4_25.
39. Shapira SD, Gat-Viks I, Shum BO, Dricot A, de Grace MM, Wu L, et al. A physical and regulatory map of host-influenza interactions reveals pathways in h1n1 infection. Cell. 2009;139(7):1255–67. doi:10.1016/j.cell.2009.12.018.
40. Liao HX, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. 2013;496(7446): 469–76. doi:10.1038/nature12053.
41. Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: Orphans and crosstalks. PLoS ONE. 2011;6(5):e19729. doi:10.1371/journal.pone.0019729.
42. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. High-resolution protein complexes from integrating genomic information with molecular simulation. Proc Natl Acad Sci USA. 2212;106(52):4–9. doi:10.1073/pnas.0912100106.
43. Faith DP. Conservation evaluation and phylogenetic diversity. Biol Conserv. 1992;61(1):1–10. doi:10.1016/0006-3207(92)91201-3.
44. Casino P, Rubio V, Marina A. Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell. 2009;139(2):325–36. doi:10.1016/j.cell.2009.08.032.
45. Li L, Shakhnovich EI, Mirny LA. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. Proc Natl Acad Sci USA. 2003;100(8):4463–8. doi:10.1073/pnas.0737647100.
46. Haldimann A, Prahalad MK, Fisher SL, Kim SK, Walsh CT, Wanner BL. Altered recognition mutants of the response regulator PhoB: a new genetic strategy for studying protein-protein interactions. Proc Natl Acad Sci USA. 1436;93(25):1–6. http://www.ncbi.nlm.nih.gov/pubmed/8962056.
47. Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, et al. Rewiring the specificity of two-component signal transduction systems. Cell. 2008;133(6):1043–54. doi:10.1016/j.cell.2008.04.040.
48. Laub MT, Goulian M. Specificity in two-component signal transduction pathways. Annu Rev Genet. 2007;41:121–45. doi:10.1146/annurev.genet.41.042007.170548.
49. Tillier ERM, Lui TWH. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics. 2003;19(6):750–55. doi:10.1093/bioinformatics/btg072.
50. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins: Structure Function Bioinform. 2004;56(2):211–21. doi:10.1002/prot.20098.
51. Gouveia-Oliveira R, Roque FS, Wernersson R, Sicheritz-Ponten T, Sackett PW, Molgaard A, et al. InterMap3d: predicting and visualizing co-evolving protein residues. Bioinformatics. 2009;25(15):1963–5. doi:10.1093/bioinformatics/btp335.
52. Wollenberg KR, Atchley WR. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci USA. 2000;97(7):3288–91. doi:10.1073/pnas.070154797.
53. Goebel B, Dawy Z, Hagenauer J, Mueller JC. An approximation to the distribution of finite sample size mutual information estimates. In: Communications, 2005. ICC 2005. IEEE International Conference on, vol. 2. IEEE; 2005. p. 1102–6. https://ieeexplore.ieee.org/ielx5/9996/32110/01494518.pdf, doi:10.1109/ICC.2005.1494518.

54. Compton AA, Hirsch VM, Emerman M. The host restriction factor APOBEC3g and retroviral vif protein coevolve due to ongoing genetic conflict. Cell Host Microbe. 2012;11(1):91–8. doi:10.1016/j.chom.2011.11.010.
55. Compton AA, Emerman M. Convergence and divergence in the evolution of the APOBEC3g-vif interaction reveal ancient origins of simian immunodeficiency viruses. PLoS Pathog. 2013;9(1):e1003135. doi:10.1371/journal.ppat.1003135.
56. Chen G, He Z, Wang T, Xu R, Yu XF. A patch of positively charged amino acids surrounding the human immunodeficiency virus type 1 vif SLVx4yx9y motif influences its interaction with APOBEC3g. J Virol. 2009;83(17):8674–82. doi:10.1128/JVI.00653-09.
57. Russell RA, Pathak VK. Identification of two distinct human immunodeficiency virus type 1 vif determinants critical for interactions with human APOBEC3g and APOBEC3f. J Virol. 2007;81(15):8201–10. doi:10.1128/JVI.00395-07.
58. Zhang H, Pomerantz RJ, Dornadula G, Sun Y. 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. J Virol. 2000;74(18):8252–61. http://www.ncbi.nlm.nih.gov/pubmed/10954522.
59. He Z, Zhang W, Chen G, Xu R, Yu XF. Characterization of conserved motifs in HIV-1 vif required for APOBEC3g and APOBEC3f interaction. J Mol Biol. 2008;381(4):1000–11. doi:10.1016/j.jmb.2008.06.061.
60. Zhang L, Saadatmand J, Li X, Guo F, Niu M, Jiang J, et al. Function analysis of sequences in human APOBEC3g involved in vif-mediated degradation. Virology. 2008;370(1):113–21. doi:10.1016/j.virol.2007.08.027.
61. Russell RA, Smith J, Barr R, Bhattacharyya D, Pathak VK. Distinct domains within APOBEC3g and APOBEC3f interact with separate regions of human immunodeficiency virus type 1 vif. J Virol. 2009;83(4):1992–2003. doi:10.1128/JVI.01621-08.
62. Xu H, Svarovskaia ES, Barr R, Zhang Y, Khan MA, Strebel K, et al. A single amino acid substitution in human APOBEC3g antiretroviral enzyme confers resistance to HIV-1 virion infectivity factor-induced depletion. Proc Natl Acad Sci USA. 2004;101(15):5652–7. doi:10.1073/pnas.0400830101.
63. Guo Y, Dong L, Qiu X, Wang Y, Zhang B, Liu H, et al. Structural basis for hijacking CBF-beta and CUL5 e3 ligase complex by HIV-1 vif. Nature. 2014;505(7482):229–33. doi:10.1038/nature12884.
64. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics. 2005;21(22):4116–24. doi:10.1093/bioinformatics/bti671.
65. Meila M. Comparing clusterings–an information based distance. J Multivar Anal. 2007;98(5):873–95. doi:10.1016/j.jmva.2006.11.013.
66. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605–12. doi:10.1002/jcc.20084.
67. Ovchinnikov S, Kamisetty H, Baker D. Data from: Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. 2014. http://dx.doi.org/10.5061/dryad.s00vr.
68. Cocco S, Monasson R, Weigt M. From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. PLoS Comput Biol. 2013;9(8):e1003176. doi:10.1371/journal.pcbi.1003176.
69. Ezkurdia I, Graña O, Izarzugaza JMG, Tress ML. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins: Structure Function Bioinform. 2009;77(S9):196–209. doi:10.1002/prot.22554.
70. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue-residue contact prediction in CASP10: Contact prediction in CASP10. Proteins: Structure Function Bioinform. 2014;82:138–53. doi:10.1002/prot.24340.
71. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27:379–423. https://dx.doi.org/10.1002%2Fj.1538-7305.1948.tb01338.x.
72. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. Phys Rev E Stat Nonlinear Soft Matter Phys. 2013;87(1):012707. http://www.ncbi.nlm.nih.gov/pubmed/23410359.