Augmenting information theory with opposite polarity of amino acids for protein contact prediction

Yancy Liao[1], Jeremy Selengut[1]

[1]Department of Computer Science, University of Maryland - College Park

The covariation of protein residue pairs is often measured by mutual information scores. Residue pairs with high mutual information scores, however, are not guaranteed to be functionally or structurally significant, since their covariation may be due to a number of indirect effects. We propose filtering for residue pairs whose covariational patterns suggest a direct interaction based on their chemical properties. Specifically, we took three protein families (groL, gudD, and BADH) and filtered them for pairs that covary by opposite polarity of charge ([K or R] with [E or D]). The resulting sets of pairs had moderate improvements in physical proximity compared to using mutual information alone. This suggests a potentially new approach of applying mutual information in tandem with chemical properties to improve contact prediction.

Introduction

Residue pairs which share a high degree of covariation can offer insights concerning the structure and function of a protein family. When changes in one residue is mirrored by compensatory changes in the other residue, this suggests that the residues have co-evolved.

Covariation can be used to infer residue pairs which are in physical contact or which have functional relationships. This can then be used to construct models of protein folding or mutational processes (Juan et al., 2013). While these inferences have become more widely accessible with the proliferation of sequence data, they are imperfect. Apart from random background noise, there are a variety of confounding factors by which residue pairs can exhibit a spurious degree of covariation. Such spurious relationships can come about, e.g., through a transitive chain of direct interactions, or because of the presence of a common lineage within the sequence data (Marks et al., 2012).

Covariational methods are split broadly into two groups: local and global (Marks et al., 2012). While global methods are typically more complex and avoid the problem of transitivity by modeling sequences as a whole, local methods are relatively simpler computationally and allow us to easily isolate for the contributions of each amino acid pairing, and thereby to augment the method to more heavily favor those pairings which correspond to meaningful chemical interactions.

Corrected Mutual Information

In this paper, we focus on one commonly used local method called mutual information. In particular, we use a variant which "corrects" for a portion of the background noise including the impact of common lineage.

Mutual information is a way to quantify the dependence of two random variables, or how much knowing one variable can tell you about the other. Concretely, suppose X and Y are positions in a multiple sequence alignment (MSA). Then the mutual information of X and Y is:

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Like many other local methods, mutual information revolves around a comparison of the observed distribution of pairs ($p(x, y)$) with the expected distribution under an assumption of independence ($p(x)p(y)$). The contribution of each amino acid pairing (or combination of x and y) is additive, so as we make use of later, we can meaningfully distinguish particular pairings which have an out-sized effect on the overall mutual information score.

Introduced by Dunn et al. (2008), corrected mutual information (cMI) is a minor variation which subtracts away a quantity called the average product correction (APC). The APC of positions X and Y is defined as:

$$APC(X, Y) = \frac{MI(X, \bar{*})MI(\bar{*}, Y)}{\overline{MI}}$$

where, if we allow n to be the length of the MSA, $\overline{MI} = \frac{2}{n(n-1)} \sum_{A=1}^{n} \sum_{B \neq A}^{n} MI(A, B)$ is the average of MI scores across all position pairs, and $MI(X, \bar{*}) = \frac{1}{n-1} \sum_{B \neq X}^{n} MI(X, B)$ is the average of MI scores among position pairs featuring X.

And finally, we have:

$$cMI(X, Y) = MI(X, Y) - APC(X, Y)$$

By incorporating a positional averaging, the APC under certain theoretical assumptions approximates the amount of background noise or shared ancestry inherent in its input positions, which is then subtracted away from the MI to produce the cMI, which has been empirically shown to be a more accurate measure of the covariation from structural or functional constraints (Dunn et al., 2008).

Methods

The selection of protein families on which to test our hypothesis was constrained by finding proteins with: ample sequence length, a representative hidden markov model (HMM), a PDB

crystal structure with multiple chains, and a sufficient number of representatives in our sequence database. With these criteria in mind, we selected and tested on three protein families: groL, gudD, and BADH. In this section we focus on giving a detailed breakdown of the process for groL.

Chaperonin GroEL (groL) is a protein found in bacteria which is involved with protein folding. It has an HMM of length 526, denoted TIGR02348 in the TIGRFAMs database. It has a PDB crystal structure of length 525, denoted 2EU1 in the Protein Data Bank. The crystal structure consists of 14 different chains.

Physical distance within the crystal structure was defined as the minimum distance between atoms in the r-groups. For glycine, which has no r-group, we used the location of its c-alpha atom. We further refined the notion of distance between two positions to be the minimum distance for those positions across combinations of the 14 approximately identical chains.

Using HMMER 3.0, we performed an HMM search against the Uniref full database which found over 5000 sequences above the trusted cut-off, a number of sequences well surpassing the minimum for MI methods (Gloor et al., 2005). We aligned the crystal structure sequence against the HMM as well, in order to translate residue positions from HMM to crystal structure, and vice versa.

With the 5000 sequences as our data set, we computed cMI for each pair of positions from the 526. We selected those pairs above the 99.5th percentile in cMI, or the top 687 pairs of positions. For each pair, we translated them to the crystal structure and computed their physical distance.

We then generated 196 pairs of positions taken at random, and computed their physical distance. We also generated 99 pairs of ”adjacent” positions, meaning they are within 5 positions apart on the sequence. Intuitively, we would expect the high cMI positions, since they are covarying, to be closer in distance than the random positions. We would also expect the

adjacent positions to be fairly close in distance since they are near in terms of sequence order.

Finally, with regards our hypothesis, we wished to filter the high cMI positions for co-variations from amino acids having opposite polarity. We selected high cMI pairs for which [K or R] with [E or D] interactions accounted for at least 15% of their MI score. Since opposite polarities have a physical explanation for covarying, we expected those positions to be closer.

Results

We first plotted the three sets of position pairs: high cMI, random, and adjacent, according to their physical distance (in angstroms) versus their cMI score (note that while regular MI is always non-negative, it is possible for cMI to be negative).

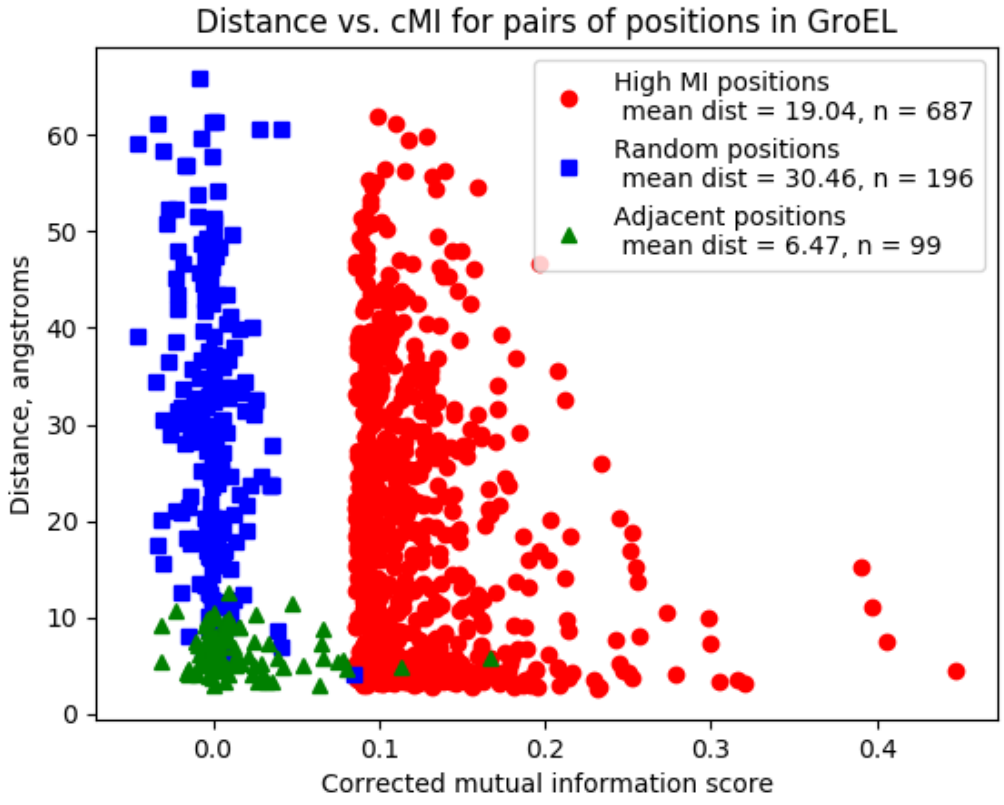


Figure 1:

We see in Figure 1 that the random positions are spread fairly evenly in terms of distance, from 0 angstroms apart to over 60 angstroms apart, with a mean distance of 30.46. Meanwhile, they are clustered tightly around 0 in cMI score. The high cMI positions, naturally, only occur on the right side of the graph, with a clear separation from the random positions. While still fairly spread out in terms of distance, the high cMI positions are weighted more heavily towards the lower side of the graph, indicating their closer distances while having a mean of 19.04. The adjacent positions are salient in the uniformity of their closer distances with a mean of 6.47.

We then plotted within the high cMI group, highlighting those with covariation coming from opposite polarity amino acids (KR-ED correlation).

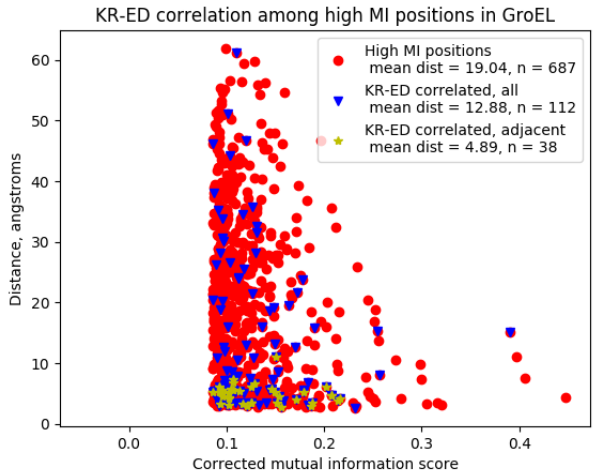


Figure 2:

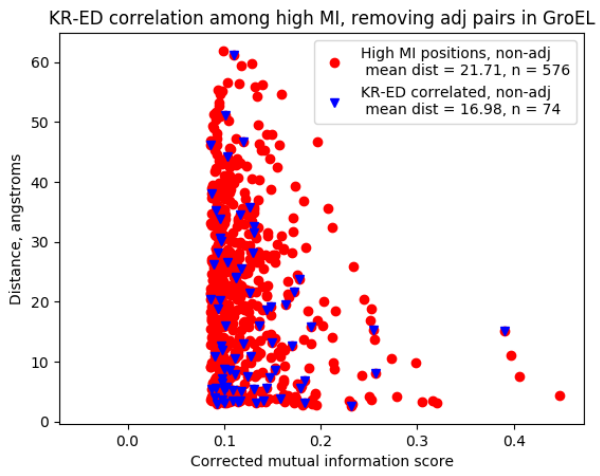


Figure 3:

We see in Figure 2 that of the 687 high cMI positions, 112 of them were found to be KR-ED correlated. The KR-ED correlated pairs do indeed have a closer mean distance (12.88) compared to the overall high cMI (19.04). However, of the 112 KR-ED correlated pairs, we found that 38 of them were sequentially adjacent. This is a fairly high percentage. And as we know from the previous Figure 1, adjacent pairs are almost always close in distance. So, what portion of the improved distance in KR-ED correlated pairs could be explained by the sequen-

tially adjacent? In Figure 3, we excluded the sequentially adjacent and plotted 576 non-adjacent high cMI positions alongside 74 non-adjacent KR-ED correlated positions. The former had a mean of 21.71, while the latter had a mean of 16.98, meaning the difference was diminished slightly, but still presents a moderate improvement.

The overall results we presented thus far for groL hold true for the other two protein families, gudD and BADH, as well, and their figures are included in the supplemental section.

Discussion and Future Work

We have seen that filtering for opposite polarity of covarying amino acid pairs improves upon a mere reliance of high cMI score, even when removing the effects of selecting for sequentially adjacent pairs. The improvement effect is moderate, on an averaged basis, and by no means gives a guaranteed set of contacting pairs.

In this study we focused exclusively on opposite polarity. In the course of the study we did attempt to filter by other interactions, such as those for hydrophobic affiliation, but obtained weak or mixed results. We also ran through the most common specific pairings found for ”close” positions versus ”far” positions but did not detect a discernibly generalizable pattern. Nevertheless, this line of investigation may hold untapped potential in leading to additional, more fine-grained filters based on the chemical properties of covariational patterns, and thereby augmenting local methods like mutual information in producing more accurate contact predictions.

Acknowledgments

I am extremely grateful to Dr. Jeremy Selengut for his generous guidance and expertise while working on this project; to members of the CBCB including Dr. Mihai Pop, Jay Ghurye, and Nidhi Shah for their helpful insights during conversations; and to IARPA for providing the

funding for this project to occur.

References and Notes

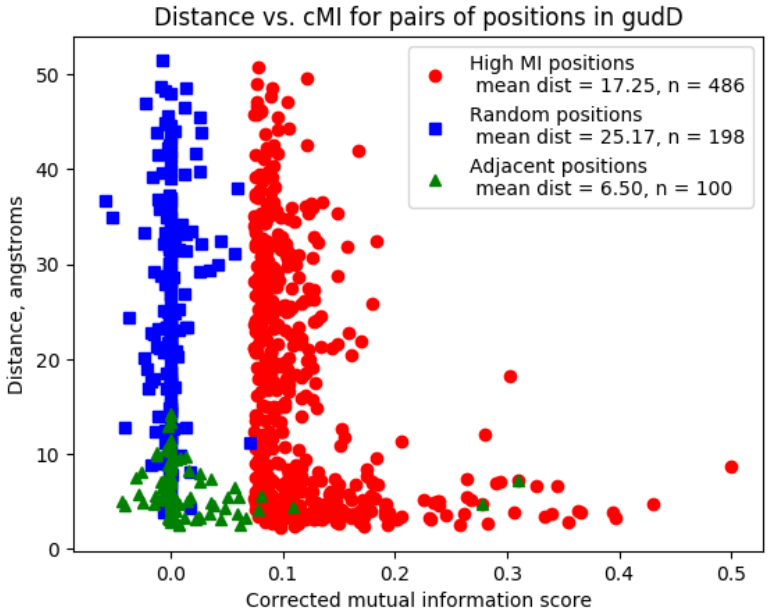
1. Dunn, S.D., Wahl, L.M., Gloor, G.B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, Volume 24, Issue 3, Pages 333-340. https://doi.org/10.1093/bioinformatics/btm604

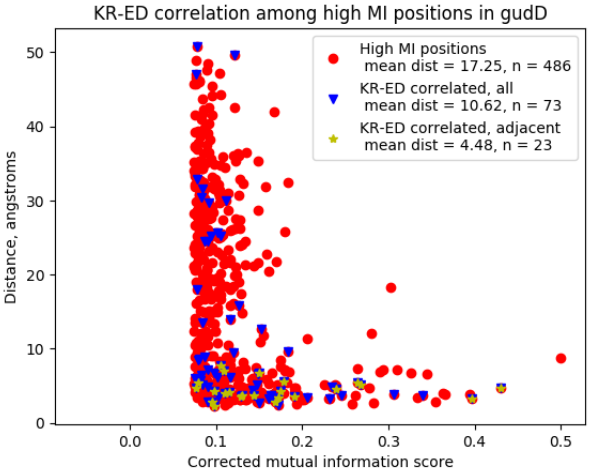
2. Gloor, Gregory B., Martin, Louise C., Wahl, Lindi M., Dunn, Stanley D. (2005) Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions. *Biochemistry* 44 (19), Pages 7156-7165. https://doi.org/10.1021/bi050293e

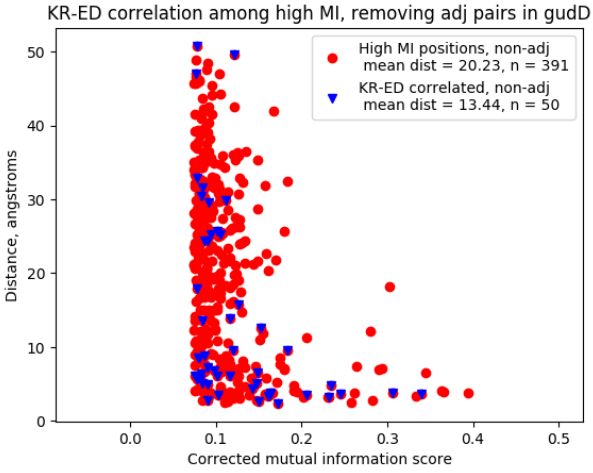
3. Juan, David de, Pazos, Florencio, Valencia, Alfonso. (2013) Emerging methods in protein co-evolution. *Nature Review Genetics* 14, Pages 249-261. https://doi.org/10.1038/nrg3414

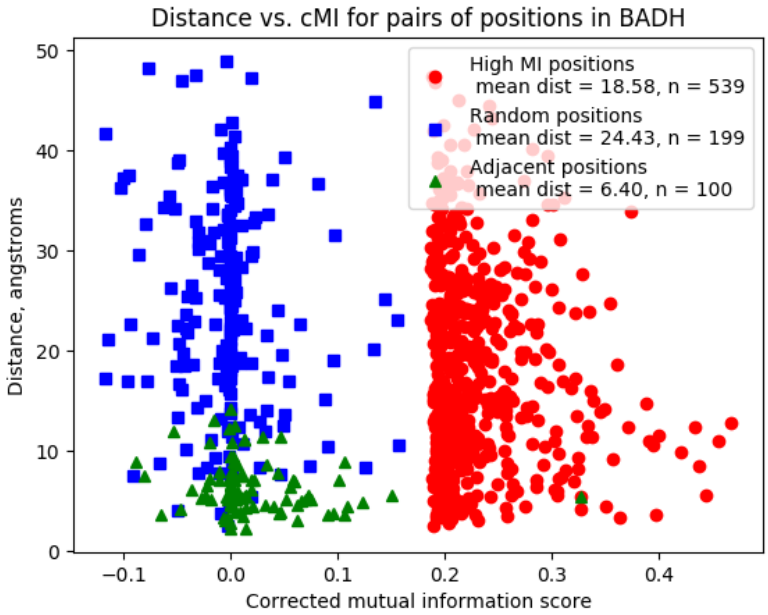
4. Marks, Debora S., Hopf, Thomas A., Sander, Chris. (2012) Protein structure prediction from sequence variation. *Nature Biotechnology* 30, Pages 1072-1080. https://doi.org/10.1038/nbt.2419

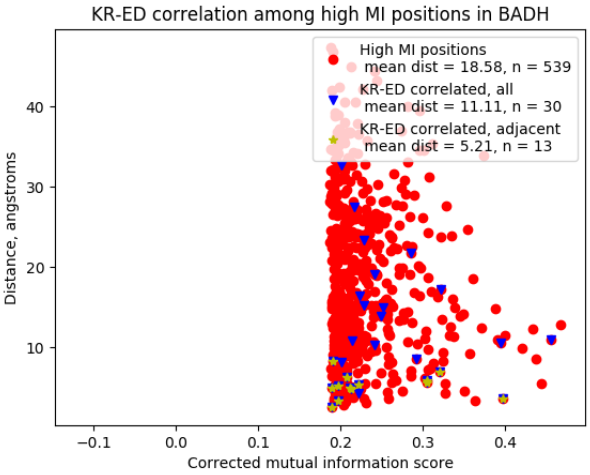
Supplemental

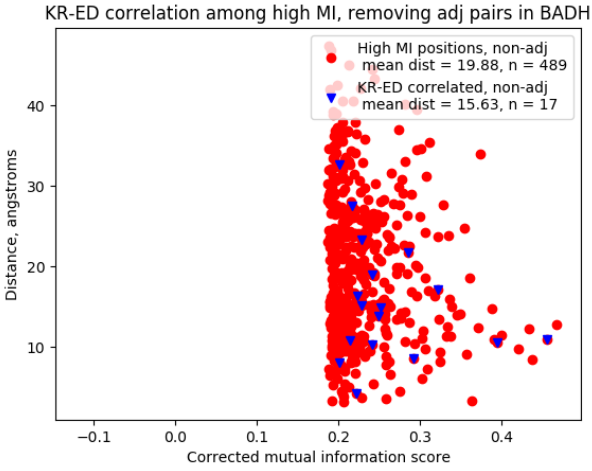







Distance vs. cMI for pairs of positions in BADH
High MI positions
mean dist = 18.58, n = 539
Random positions
mean dist = 24.43, n = 199
Adjacent positions
mean dist = 6.40, n = 100
Distance, angstroms
Corrected mutual information score


KR-ED correlation among high MI positions in BADH
High MI positions
mean dist = 18.58, n = 539
KR-ED correlated, all
mean dist = 11.11, n = 30
KR-ED correlated, adjacent
mean dist = 5.21, n = 13
Distance, angstroms
Corrected mutual information score


KR-ED correlation among high MI, removing adj pairs in BADH
High MI positions, non-adj
mean dist = 19.88, n = 489
KR-ED correlated, non-adj
mean dist = 15.63, n = 17
Distance, angstroms
Corrected mutual information score