Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction

Stanley D Dunn
Lindi M Wahl
Gregory B Gloor

Available at: https://works.bepress.com/stanley_dunn/5/
Sequence analysis

Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction

S. D. Dunn\textsuperscript{1}, L. M. Wahl\textsuperscript{2} and G. B. Gloor\textsuperscript{1,*}

\textsuperscript{1}Department of Biochemistry and \textsuperscript{2}Department of Applied Mathematics, University of Western Ontario, London, Ontario, Canada, N6A 5C1

Received on October 9, 2007; revised on November 15, 2007; accepted on December 2, 2007

Advance Access publication December 5, 2007

Associate Editor: Alfonso Valencia

\textbf{ABSTRACT}

\textbf{Motivation:} Compensating alterations during the evolution of protein families give rise to coevolving positions that contain important structural and functional information. However, a high background composed of random noise and phylogenetic components interferes with the identification of coevolving positions.

\textbf{Results:} We have developed a rapid, simple and general method based on information theory that accurately estimates the level of background mutual information for each pair of positions in a given protein family. Removal of this background results in a metric, MI\textsubscript{p}, that correctly identifies substantially more coevolving positions in protein families than any existing method. A significant fraction of these positions coevolve strongly with one or only a few positions. The vast majority of such position pairs are in contact in representative structures. The identification of strongly coevolving position pairs can be used to impose significant structural limitations and should be an important additional constraint for \textit{ab initio} protein folding.

\textbf{Availability:} Alignments and program files can be found in the Supplementary Information.

\textbf{Contact:} ggloor@uwo.ca

\textbf{Supplementary information:} Supplementary data are available at Bioinformatics online.

\section{INTRODUCTION}

In the course of evolution, the amino acid sequences of a family of orthologous proteins slowly change while the fold of the native structure is maintained. Some sequence positions change little, implying that mutations at these sites are not tolerated, and the residues conserved in such positions are inferred to be important for the structure or function of the protein. Sequence positions that vary are expected to be less critical, yet in many cases mutations at these non-conserved positions also cause disruption of structure or loss of function.

How, then, do non-conserved positions change during evolution? It is believed that mutations in these positions can occur because they are either accompanied or preceded by compensatory changes in other variable positions (Fitch et al., 1970; Yanofsky et al., 1964). Such compensation would result in a coupling between changes in the two positions, or coevolution (Fitch et al., 1970). Since compensation often involves residues that are proximal in the folded structure (Fitch and Markowitz, 1970; Poon and Chao, 2005; Yanofsky et al., 1964), identification of coevolving positions is expected to be useful in the \textit{ab initio} prediction of protein structure from multiple sequence alignments (Fariselli et al., 2001; Gobel et al., 1994; Vendruscolo et al., 1997; Vendruscolo and Domay, 2000).

A number of different approaches to identifying coevolving positions have been developed including methods to detect the differences between observed versus expected frequencies of residue pairs (OMES) (Kass and Horovitz, 2002; Larson et al., 2000), the McLachlan Based Substitution correlation (McBASC) (Gobel et al., 1994; Olmea et al., 1999), Statistical Coupling Analysis (Lockless and Ranganathan, 1999), Mutual Interdependency (Tillier and Lui, 2003), Coevolution Analysis using Protein Sequence (Fares and Travers, 2006) and Mutual Information (MI) based methods (Chiu and Kolodziejczak, 1991; Cover and Thomas, 1991; Korber et al., 1993; Wollenberg and Atchley, 2000). Among these, only the MI, McBASC and OMES methods do not use any structural or phylogenetic information. These methods also do not depend on estimating the significance of coevolution by computationally expensive simulations. A recent study showed that McBASC and OMES were able to identify contacting pairs better than the MI or statistical-coupling methods (Fodor and Aldrich, 2004a).

In the context of multiple sequence alignments, MI is an attractive metric because it explicitly measures the dependence of one position on another, but its usefulness has been limited by three factors. First, positions with higher variability, or entropy, will tend to have higher levels of both random and nonrandom MI than positions of lower entropy (Fodor and Aldrich, 2004a; Martin et al., 2005), even though the latter are more constrained and would seem more likely to depend on neighboring positions. Second, random MI arises because the alignments do not contain enough sequences for background noise to be negligible; our previous modeling studies showed that alignments should contain at least 125 sequences before the random signal begins to subside relative to non-random MI (Martin et al., 2005). A third complicating factor is that all position pairs have MI due to the phylogenetic relationships of the organisms represented in the alignment.

*To whom correspondence should be addressed.

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

doi:10.1093/bioinformatics/btm604
positions. This can be thought of as the degree of correlation between two correlates strongly with the entropy of the positions (Fodor and to 1, which occurs when all 20 residues are equally distributed. The calculated value varies from 0, in the case of complete conservation, and all other positions in the set, divided by the average \( M_{lb} \) of all positions in the set. We call this term the average product correction, (APC), as shown in Equation (5).

In contrast we expected that the \( M_I \) related to structure or function, \( M_{lp} \), would be highly specific, and for a given position would be found only with a limited number of other positions. Therefore, the difference between the APC and total \( M_I \) for a given pair of positions should isolate \( M_{lp} \). We use \( M_{lp} \) to denote this difference, i.e., \( M_{lp}(a,b) = M(a,b) - APC(a,b) \). As illustrated in the Results to follow, a remarkable feature of \( M_{lp} \) is that this correction also removes the influence of entropy.

3 METHODS

3.1 The average product correction

To determine the conditions under which the APC might closely approximate \( M_{lp} \), we begin by defining \( M(a,x) \) as the mean mutual information of column \( a \), that is, \( M(a,x) = \frac{1}{n} \sum M(a,x) \), where \( n \) is the number of columns in the alignment, \( m = n - 1 \) for convenience, and the summation is over \( x=1 \) to \( n, x \neq a \).

Similarly, \( M(x,y) \) denotes the overall mean mutual information, \( M(x,y) = \frac{1}{mn} \sum M(x,y) \), where the indices run \( x=1 \) to \( m, y = x+1 \) to \( n \).

We also define the mean joint entropy, \( H(x,y) \), and the mean joint entropy between a specific column \( a \) and all other columns, \( H(a,x) \), in the analogous way. Finally, we use \( \bar{H}(x,y) \) to denote the mean entropy of all residues in the alignment.

Expanding the product \( M(a,x)MB(x,y) \), we find that it equals:

\[
\left( \sum_{x,y} M(a,x) \right) \left( \sum_{x,y} M(b,x) \right) = \left( mH(a) + \sum_{x,y} H(x) - \sum_{x,y} H(a,x) \right) \cdot \left( mH(b) + \sum_{x,y} H(x) - \sum_{x,y} H(b,x) \right)
\]

\[
\approx \bar{H}(H(a) + H(b) - H(a,x) - H(b,x) + \bar{H}) + H(a)H(b) - H(a)bH(x) = H(a,x) + H(a,x)H(b,x).
\]

The approximation enters because we have used \( \bar{H} \) to approximate both \( \sum_{x,y} H(x)/m \) and \( \sum_{x,y} H(b,x)/m \), which holds for \( m \) large.

The critical assumption which allows us to proceed further is the following. We assume that the joint entropy contains an additive component from each residue, i.e., that we can approximate \( H(a,b) \) as \( H(a,b) \approx H(x,y) + \delta a + \delta b \). Clearly, some \( \delta a \) will be positive and some negative, and by the definition of \( H(x,y) \) we find that \( \sum_{x,y} \delta x = 0 \). Thus \( \sum_{x,y} \delta x = -\delta a \), and we find:

\[
H(a,x) = H(x,y) + \delta a - \delta a/m,
\]

and since \( m \) is large.

(Wollenburg and Atchley, 2000). This latter source may be limited to some degree by excluding highly similar sequences from closely related species from the alignment, but cannot be eliminated (Martin et al., 2005; Tillier and Lüi, 2003). Each of these sources of \( M_I \) will tend to obscure the desired signal based on the structural or functional relationships of positions.

Here we develop a simple method to estimate the expected levels of background \( M_I \) arising from the random and phylogenetic sources. We begin by calculating the \( M_I \) between positions of two different protein families in a joint alignment of sequences from the same set of organisms. We find that in the absence of any structural or functional relationships, the \( M_I \) between positions may be estimated with surprising accuracy from the average levels of \( M_I \) observed for those positions in comparison to the average of all joint pairs. Correction of the \( M_I \) values obtained between positions within a protein family by this factor significantly enhances the signal between positions that are close together in the folded protein structure. We further show that the same method can be applied using average \( M_I \) values derived from positions within a single protein family. Since large joint alignments are difficult to produce, this approach is much more accessible. Application of this correction provides a substantial improvement compared to previously published methods for using sequence analysis to find positions that are proximal in the protein structures.

2 APPROACH

Shannon’s entropy \( (H) \) for column \( a \) in a multiple sequence alignment is a measure of the randomness of the residues in the column (Cover and Thomas, 1991). It is calculated as the sum, over all residues in the column, of the frequency of occurrence of each residue in that column, \( p(x,a) \), multiplied by the \( \log_{2} \) of that frequency: \( H(a) = - \sum x p(x,a) \log_{20} p(x,a) \). The resulting value varies from 0, in the case of complete conservation, to 1, which occurs when all 20 residues are equally distributed. The observed joint entropy of each pair of positions, \( H(a,b) \), is calculated similarly, except di-residue frequencies are used and the sum extends over the possible combinations. The joint entropy values can range from 0 to 2.

\( M_I \) measures the reduction of uncertainty about one position given information about the other (Cover and Thomas, 1991). This can be thought of as the degree of correlation between two positions \( a \) and \( b \) in a multiple sequence alignment. \( M_I \) is calculated \( M_I(a,b) = H(a) + H(b) - H(a,b) \). \( M_I \) can vary between 0 and 1, with larger values reflecting more interdependence between the positions.

As mentioned above, three factors confound the use of \( M_I \) in identifying covarying positions in protein families. First, \( M_I \) correlates strongly with the entropy of the positions (Fodor and Aldrich, 2004a; Martin et al., 2005). We have previously shown that the influence of entropy can be partially removed by the \( M_{IR} \) correction (Martin et al., 2005):

\[
M_{IR}(a,b) = M_I(a,b)/H(a,b).
\]

In addition, the \( M_I \) between a pair of positions in a protein family is composed of \( M_I \) due to structural-interactions, functional constraints, random noise and shared ancestry, as proposed by Wollenburg and Atchley (2000). Thus the challenge is to separate the signal caused by structural and functional constraints, \( M_{IR} \), from the background, \( M_{lo} \) which is the sum of contributions from random noise and shared ancestry.

To address this challenge, we postulated that each position in a multiple sequence alignment may have a particular propensity toward \( M_{lo} \), that is related to its entropy and phylogenetic history, and that the \( M_{lo} \) between any two positions is the product of their propensities. It then follows that \( M_{lo} \) for positions \( a \) and \( b \) may be expressed as the product of the average \( M_{lo} \) values of positions \( a \) and \( b \) with all other positions in the set, divided by the average \( M_{lo} \) of all positions in the set. We call this term the average product correction, (APC), as shown in Equation (5).

```latex
\[
M_{IR}(a,b) = M_I(a,b)/H(a,b).
\]
```
3.3 Sequence alignments and structural information

Protein families containing at least 125 independent sequences and at least one solved structure were the same as those used previously (Martin et al., 2005), but were filtered to remove duplicate families and families with only low-resolution structures. A list of the resulting 93 protein families and their alignments are given in Supplementary Table 1. Only ungapped positions in the alignments were considered in our analysis. Residues separated by \( \leq 6 \) \( \AA \) were defined to be in contact.

3.4 Other covariance measures

An independent method of calculating covariance is to use the sum-of-squares method of Kass and Horovitz, 2002 using the formula:

\[
OMES = \frac{\sum (N_0 - N_e)^2}{N_e}
\]

(7)

Here, \( N_e \) is the observed number of di-amins in a pair of columns, \( N_e \) is the expected number, \( N \) is the total number of possible di-amino acid pairs, and \( N_i \) is the total number sequences in the alignment. This equation returns the difference between the expected and observed di-amino acid frequencies for a pair of columns. This formula returns 0 if one column or both columns are absolutely conserved, or if the residues in both columns are assorted randomly.

The calculation of MC\(BASC\) is more involved, and was performed as described (Fodor and Aldrich, 2004a) using the software provided by the authors (http://www.afodor.net/). Unlike MI or OMES, MC\(BASC\) returns a high covariance value both for strongly covarying, non-conserved positions, and for highly conserved pairs of positions in a multiple sequence alignment.

3.5 Conversion of MI values to Z-scores

We determined how different a given covariance value was relative to all other values in the data set. The mean and SD of the values determined by each of the algorithms were calculated for all pairs of positions. The number of SD from the mean, i.e. the Z-score, was determined for each value or for each corrected value in a given data set. Initially, we ensured that the entropy of both positions exceeded 0.3 since previous work showed that positions with an entropy below this value often display MI because of insufficient variability (Fodor and Aldrich, 2004a; Martin et al., 2005; Tillier and Lui, 2003). Later analyses were conducted without an entropy cutoff as MI\(p\) provides an entropy-independent measure. The entropy cutoffs used for each analysis are indicated.

4 RESULTS

4.1 The use of joint alignments of protein families

We made joint alignments of unrelated protein families from the same set of organisms to obtain a data set lacking MI\(p\). We reasoned that, if most of the proteins are orthologs, the two proteins of a joint alignment should share a similar phylogenetic history since they come from the same organisms. We ignored the possibility of horizontal gene transfer for this analysis. Joint pairs of positions between the two families should have only phylogenetic and random MI, i.e. MI\(p\), which is strongly related to the entropy of the positions (Martin et al., 2005). Since the proteins are unrelated and their positions share no MI\(p\), our starting assumption may be tested by examining the relationship between the MI and APC values for positions between the protein families.
Fig. 2. Scatter plots of the relationships between the MI for each pair of positions in a multiple sequence alignment, the APC and entropy. Panel A shows the relationship between MI and APC for the joint pairs of positions in a concatenated GAPDH and NDK multiple sequence alignment. These are labeled MI and APC. In panel B, MI is derived from positions within GAPDH only (MIGAPDH) and is plotted vs. APC. In panel C, both MI and APC were calculated between GAPDH positions. Panels D-F show scatter plots of MI vs. the sum of the entropy for position pairs derived from the same data sets as A-C. These plots show that MI values derived using the joint alignment (Fig. 2B) are widely distributed for both GAPDH and NDK in a concatenated GAPDH and NDK multiple sequence alignment. Panel B shows the distribution of values for GAPDH MI calculated using MI and APC values derived from the joint alignment. The distribution of positions in contact is strongly shifted to the right, and accounts for a substantial fraction of the shoulder and tail in the overall distribution.

Table 1. The effect of APC on MI

<table>
<thead>
<tr>
<th>MIa</th>
<th>Joint</th>
<th>Joint</th>
<th>GAPDH</th>
<th>GAPDH</th>
<th>GAPDH</th>
</tr>
</thead>
<tbody>
<tr>
<td>APCb</td>
<td>Joint</td>
<td>Joint</td>
<td>GAPDH</td>
<td>GAPDH</td>
<td>GAPDH</td>
</tr>
<tr>
<td>Mean</td>
<td>0.085</td>
<td>0.000</td>
<td>0.100</td>
<td>0.015</td>
<td>0.000</td>
</tr>
<tr>
<td>SD</td>
<td>0.038</td>
<td>0.013</td>
<td>0.046</td>
<td>0.029</td>
<td>0.022</td>
</tr>
<tr>
<td>Mean Z (D&lt;6 Å)</td>
<td>na</td>
<td>na</td>
<td>0.525</td>
<td>0.801</td>
<td>1.082</td>
</tr>
<tr>
<td>Median Z (D&lt;6 Å)</td>
<td>na</td>
<td>na</td>
<td>0.370</td>
<td>0.448</td>
<td>0.841</td>
</tr>
<tr>
<td>Median Z (D&gt;12 Å)</td>
<td>na</td>
<td>na</td>
<td>−0.0316</td>
<td>−0.065</td>
<td>−0.094</td>
</tr>
<tr>
<td>Median Z (D&gt;12 Å)</td>
<td>na</td>
<td>na</td>
<td>−0.178</td>
<td>−0.241</td>
<td>−0.147</td>
</tr>
</tbody>
</table>

aAlignment that is the source of MI or APC values.
bThe minimal distance between non-hydrogen atoms of the residues.

cValues of MIp for the joint pairs is similar to a normal distribution. We infer that the underlying distribution of scores is near normal in the absence of structural or functional relationships between positions.

The average value of MIp is 0, and it is notable that the SD of MIp for the joint pairs is markedly reduced relative to that for MI (Table 1). Since the joint data contain only background MI, we interpret this reduction in SD to confirm that the average product correction removes a significant proportion of the background MI.

4.1.1 Application of the joint values to GAPDH We next compared the MI values between pairs of positions in the GAPDH alignment with APC values calculated from the joint alignment (Fig. 2B). The distribution differs from that seen in Figure 2A in that more points are scattered upwards, suggesting that they contain additional MI due to structural or functional relationships. This interpretation is supported by the distribution of points representing pairs of positions with minimal inter-residue distances of ≤6 Å in the GAPDH structure, which are highlighted as open symbols. The SD of MIp determined in this way is substantially less than the SD.
of MI (Table 1). Furthermore, we observe that the mean or the median MIp values for spatially proximal residues are significantly greater than the mean or median MI values. These values are expressed as Z-scores (the number of SDs from the mean) in Table 1. Thus MIp calculated by subtraction of the joint APC values enhances the signals of many pairs that are expected to have MIp due to proximity.

4.2 Application of intraprotein average MI values to correct for MIb
Because of the difficulty of constructing large joint data sets for all proteins of interest, we explored the use of intrafamily values to calculate the APC corrections. Most pairs should share only MIb since they are neither proximal nor linked functionally; thus pairs sharing significant MIb should have a small effect on average MI values. To test this hypothesis, we calculated the correlation between the position average MI values from the joint data sets with position average MI values calculated using intraprotein data sets for all protein families in the complete set of 289 joint alignments, using an entropy cutoff of 0.3. The correlation coefficient (r) ranged between 0.71 and 0.99 with 256 (88%) of the r values being ≥0.9 (Supplementary Fig. 1). Interestingly, for 259 of the 289 joint alignments the r value was greater than that seen in the GAPDH-NDK joint alignment, which was 0.897. The enzyme SAICAR, which catalyzes a step in purine biosynthesis, was an outlier in this analysis with 15 of the 16 lowest r values derived from joint alignments involving this enzyme; indeed, all r values below 0.83 were from SAICAR-containing joint alignments. No single protein family predominated at the other end of the distribution. We concluded that the average intraprotein and the average joint MI values were strongly related, and inferred that the corresponding joint and intrafamily APC corrections were largely equivalent.

In GAPDH, subtraction of intrafamily APC to obtain MIp reduced the average value to 0, and again resulted in a major decrease in the standard deviation of the resulting intrafamily MIp values (Table 1). Furthermore, the mean or median Z-scores for spatially proximal pairs of positions were greater than those obtained with MIp derived from the joint alignment. The mean or median Z-scores for residue pairs separated by ≥12 Å was ≤0 (Table 1). The effect of this correction on pairs in contact can be seen clearly by comparing panels A and C in Figure 2. In Figure 3B note that the distribution of MIp scores for pairs in contact is shifted strongly to the right compared to the bulk of the pairs. We conclude that most pairs of positions in contact exhibit coevolution levels ranging from modest to strong as measured by MIp.

4.2.1 Application of intraprotein MIp to multiple protein families
We next compared the ability of intrafamily MIp to identify contacting pairs in a data set of 83 high quality multiple sequence alignments (Methods) to other published methods for finding coevolving positions. First, we measured the fraction of pairs in contact (non-hydrogen atom separations ≤6 Å) as a function of rank order using a variety of methods: MI, MIr (Martin et al., 2005), MIp, MIa, OMES (Fodor and Aldrich, 2004a) or McBASC (Fodor and Aldrich, 2004a). The statistical coupling method was not tested as it is less effective than both McBASC and OMES (Fodor and Aldrich, 2004a, b). Neither did we test any of the methods that rely on information other than the sequence alignment itself, or any methods that relied on bootstrapping to estimate significance.

In our data set of 83 protein families, each containing at least 125 orthologous members and a representative structure, ~8% of randomly chosen pairs are in contact. Figure 4 shows the proportion of pairs, ranked n or higher in each data set that were in contact for each method using either a 0 or 0.3 entropy cutoff as indicated. Among published methods, MIr was best, but none of the methods performed nearly as well as MIp at either entropy cutoff. The MIa method was almost equivalent to MIp in this analysis when an entropy cutoff of 0.3 was used. Interestingly, all of the methods except MIp, MIa and McBASC showed a rapid decay versus rank in identifying contacting pairs. In contrast the top five pairs identified by any of these methods were nearly equally likely to be in contact as the top ranked pair of the method.

Finally, we compared each method to find the one that identified the greatest number of contacting pairs in representative structures of the 83 protein family alignments. The pairwise scores from each method were converted to Z-scores and sorted from the highest to lowest score. Single-paired positions are defined as those pairs where each partner had a Z-score above the threshold with only the other residue. In previous work we showed that the residues corresponding to

![Fig. 4. Plots of the likelihood of contact between pairs of residues in representative structures versus the coevolution ranks for different algorithms. Coevolution between pairs of positions was calculated using either MI, MIr, MIa, MIp, OMES or McBASC across all 83 protein families. Two entropy cutoff values were used in this analysis: 0.3 for the MI-based methods, and 0 for McBASC and OMES. MIp was tested at both entropy cutoffs. For each protein family, the scores from each method were sorted from highest to lowest, the fraction of pairs at each rank or higher that were in contact in a representative structure was tabulated, and averages were plotted. All the methods were significantly better than random chance in identifying contacting pairs. The random line shows the results of a single randomization, the value of which converges to the random contact frequency of 8%.](image-url)
The multiple sequence alignment for the MIp4.2.2 Application of intraprotein strongly enhances the coevolution score of residues in contact. methods (Supplementary Table 2). We conclude that MIp entropy cutoff of 0. Data from all 83 protein families were included. The solid lines indicate methods that were evaluated with an entropy cutoff of 0.3, the dashed lines indicate methods evaluated with an entropy cutoff of 0. Strikingly, three to four times as many pairs as any other method (other than MIa) were identified by each method at each Z-score threshold ranging from 2–7. The bottom panel shows the fraction of single pairs in contact that are found by each method at a Z-score cutoff of 0. Strikingly, Z75% and 80% of pairs were in contact across a broad range of Z-score cutoffs in the MIr, OMES, MIa and MIp methods. As expected from the results shown in Figure 4, MIp performed best with an entropy cutoff of 0. Strikingly, MIp uncovered three to four times as many pairs as any other method (other than MIa) while maintaining the same accuracy as the other methods (Supplementary Table 2). We conclude that MIp strongly enhances the coevolution score of residues in contact.

4.2.2 Application of intraprotein MIp to Triosephosphate Isomerase

The multiple sequence alignment for the homodimeric enzyme triosephosphate isomerase, reported previously (Martin et al., 2005), was used to compare the single pairs identified by each method. Single paired positions at integer Z-score thresholds of 2 through 7 were identified, and positions adjacent in sequence were ignored. Table 2 shows that MIr identified 7 such pairs, five of which were in contact in the structure 1IIH (Noble et al., 1991). Four of these were separated by more than 10 residues in sequence. MIp identified 15 pairs, 11 in intra-subunit contact, and 9 of which were separated by more than 10 residues in sequence. MIp also identified one pair that were far apart within each single subunit, but in contact across the dimerization interface. All the pairs identified by MIr were also found by MIp, in each case with a strikingly higher Z-score. OMES identified six pairs, four in contact. McBASC found 8 pairs, three of which were in contact including one pair that made contact across the dimer interface. Interestingly, the OMES, McBASC and MIp methods each found several pairs that the other methods did not, likely because the underlying algorithms are different.

We examined the covariation of one pair corresponding to positions 140–189 in the 1IIH structure. This pair was chosen because it was highly variable in 8 structures derived from different organisms (see Supplementary Tables 3 and 4), and because it was identified only by MIp. As shown in Supplementary Table 3, interactions between these residues comprise a strong ionic interaction, an aromatic ring interaction and various aliphatic interactions. Similar interactions can

Table 2. Single pairs in triosephosphate isomerase by each method

<table>
<thead>
<tr>
<th>pos_i</th>
<th>pos_j</th>
<th>Distance</th>
<th>MIr</th>
<th>OMES</th>
<th>McBASC</th>
<th>MIp</th>
</tr>
</thead>
<tbody>
<tr>
<td>184</td>
<td>227</td>
<td>2.8</td>
<td>7.5</td>
<td>9.1</td>
<td>10.5</td>
<td>184</td>
</tr>
<tr>
<td>142</td>
<td>145</td>
<td>3.1</td>
<td>7.5</td>
<td>9.1</td>
<td>10.5</td>
<td>142</td>
</tr>
<tr>
<td>179</td>
<td>181</td>
<td>3.1</td>
<td>7.5</td>
<td>9.1</td>
<td>10.5</td>
<td>179</td>
</tr>
<tr>
<td>216</td>
<td>241</td>
<td>3.1</td>
<td>7.5</td>
<td>9.1</td>
<td>10.5</td>
<td>216</td>
</tr>
<tr>
<td>221</td>
<td>224</td>
<td>3.1</td>
<td>7.5</td>
<td>9.1</td>
<td>10.5</td>
<td>221</td>
</tr>
<tr>
<td>68</td>
<td>70</td>
<td>3.4</td>
<td>5.3</td>
<td>7.9</td>
<td>9.5</td>
<td>68</td>
</tr>
<tr>
<td>138</td>
<td>141</td>
<td>3.3</td>
<td>5.3</td>
<td>7.9</td>
<td>9.5</td>
<td>138</td>
</tr>
<tr>
<td>39</td>
<td>246</td>
<td>3.9</td>
<td>5.3</td>
<td>7.9</td>
<td>9.5</td>
<td>39</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>7.3</td>
<td>4.2</td>
<td>5.7</td>
<td>7.1</td>
<td>6</td>
</tr>
<tr>
<td>140</td>
<td>189</td>
<td>4.5</td>
<td>5.0</td>
<td>7.1</td>
<td>9.5</td>
<td>140</td>
</tr>
<tr>
<td>48</td>
<td>64</td>
<td>4.6</td>
<td>6.2</td>
<td>7.7</td>
<td>9.5</td>
<td>48</td>
</tr>
<tr>
<td>135</td>
<td>176</td>
<td>5.6</td>
<td>5.6</td>
<td>6.7</td>
<td>7.7</td>
<td>135</td>
</tr>
<tr>
<td>209</td>
<td>228</td>
<td>5.6</td>
<td>7.8</td>
<td>7.7</td>
<td>9.5</td>
<td>209</td>
</tr>
<tr>
<td>139</td>
<td>186</td>
<td>5.7</td>
<td>7.7</td>
<td>9.5</td>
<td>9.5</td>
<td>139</td>
</tr>
<tr>
<td>10</td>
<td>235</td>
<td>8.0</td>
<td>9.5</td>
<td>9.5</td>
<td>9.5</td>
<td>10</td>
</tr>
<tr>
<td>76</td>
<td>98</td>
<td>8.0</td>
<td>10.4</td>
<td>9.5</td>
<td>9.5</td>
<td>76</td>
</tr>
<tr>
<td>214</td>
<td>221</td>
<td>8.0</td>
<td>3.2</td>
<td>9.5</td>
<td>9.5</td>
<td>214</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>7.3</td>
<td>4.2</td>
<td>5.7</td>
<td>9.5</td>
<td>9</td>
</tr>
<tr>
<td>187</td>
<td>219</td>
<td>10.7</td>
<td>4.5</td>
<td>7.5</td>
<td>9.5</td>
<td>187</td>
</tr>
<tr>
<td>93</td>
<td>169</td>
<td>11.3</td>
<td>6.1</td>
<td>9.5</td>
<td>9.5</td>
<td>93</td>
</tr>
<tr>
<td>187</td>
<td>245</td>
<td>12.7</td>
<td>6.3</td>
<td>9.5</td>
<td>9.5</td>
<td>187</td>
</tr>
<tr>
<td>193</td>
<td>231</td>
<td>19.4</td>
<td>5.3</td>
<td>9.5</td>
<td>9.5</td>
<td>193</td>
</tr>
<tr>
<td>15</td>
<td>71</td>
<td>23.8</td>
<td>7.5</td>
<td>9.5</td>
<td>9.5</td>
<td>15</td>
</tr>
<tr>
<td>112</td>
<td>178</td>
<td>26.28</td>
<td>9.5</td>
<td>9.5</td>
<td>9.5</td>
<td>112</td>
</tr>
</tbody>
</table>

1Z-scores reported with entropy cutoff of 0, except for MIr which had an entropy cutoff of 0.3.
2Pairs in contact across the dimerization interface, intra-subunit distance reported.
be seen among the many common pairs that are found at these positions. We conclude that these positions in the protein family are highly variable, yet must covary to maintain a side-chain interaction so that the correct secondary or tertiary structure is maintained.

5 DISCUSSION

A number of obstacles, including random noise, the influence of entropy, the phylogenetic history and the number of sequences required, complicate the identification of coevolving positions in multiple sequence alignments when using MI (Martin et al., 2005; Tillier and Lui, 2003). Several groups have used bootstrapping or other randomization methods to attempt to estimate the background coevolution signal (Fares and Travers, 2006; Wollenberg and Atchley, 2000). While bootstrapping can estimate the background due to random noise and the phylogenetic background, the algorithms are computationally expensive and have not been applied to large numbers of protein families. They are also potentially sensitive to our incomplete understanding of the substitution probabilities needed to model the ancestry of the protein family.

We have taken a different approach and developed a correction that rapidly and accurately estimates the background MI found in protein family multiple sequence alignments. Our method was initially based on the assumptions that the coevolution signal between pairs of unrelated positions is derived from random noise or from shared ancestry but not from structural or functional constraints; that these factors give each position in an alignment a particular propensity toward MIb; that MIb will be the product of the propensities; and that the gene encoding the protein is inherited as a unit. We have shown that the APC accurately estimates MI in the absence of structural or functional relationships. Furthermore, in real protein alignments the subtraction of the APC from MI results in a metric, MIp, that is independent of the entropy of the positions, and that provides a significant improvement over previously published methods in identifying co-evolving positions that are proximal in protein structure.

The MIp metric still requires a significant number of sequences in the protein family because it does not address the finite sample size effects inherent in MI (Martin et al., 2005). Supplementary Figure 4 shows that the correct identification of pairs in contact begins to approach the limit when the alignments contain at least 125 non-identical sequences. The exponential growth in the sequence databases, and the availability of sequences from a wide range of organisms enables this requirement to be met for a significant fraction of protein families. Interestingly, as seen in Supplementary Figure 5, the range of variability in the sequences has a smaller effect, so long as the sequences are not identical.

We have also mathematically demonstrated the validity of the APC correction. As derived above, the APC between two unrelated positions a and b correctly estimates MIb if the number of position pairs is large, and if an assumption is made that the joint entropy is composed of an additive component from each individual position. The MIp values deviate somewhat when the position entropy and the mean entropy are different, but this results in errors in the Z-scores of Z ≤ 0.2 when 50 or more positions are used for the calculation, and falls to Z ≤ 0.1 when the number of positions is Z ≥ 100 (Supplementary Fig. 2). These deviations are consistent with the algebraic modeling, and are much less than that observed for either MI or MIr (Supplementary Fig. 2). Furthermore, modeled coevolving positions in protein families shows that MIp is significantly more sensitive and more selective than other methods (Supplementary Fig. 3). Since the APC is an accurate estimate of the background MI, we conclude that the residual MIp is an accurate estimate of the coevolution signal caused by structural and functional correlations, MIb/c.

The ASC correction performed nearly as well as the APC correction in improving the coevolution signal between proximal residues. This result implies that the largest error in application of either correction is not related to whether the correction provides the most probable level of MIb for the data set, but rather whether the data set itself is large enough that observed levels of MI are likely to approach the most probable levels. Once again, it is beneficial to have as many sequences in the alignment as possible, which can be achieved using the intrafamily APC values, since joint alignments will usually contain fewer sequences.

Even when the intrafamily alignments were restricted to those in the joint alignments, however, we found that the intrafamily MIp gave lower SD and thus higher Z-scores for proximal positions in comparison to the joint MIp. Two factors may contribute to these effects. First, we and others have previously shown that positions with high MI can be divided into group coevolvers, which appear to coevolve with multiple positions and are often not in contact, and isolated pairs, which are in contact and coevolve strongly with each other only (Fares and Travers, 2006; Gloor et al., 2005; Martin et al., 2005). The intrafamily APC correction, but not the joint APC correction, may tend to suppress the signals from the group coevolvers, since their average MI will be elevated relative to other positions of similar entropy, and this could lead to a tightening of the residual intrafamily MIp values. The net effect of this would be to enhance the Z-scores of the isolated pairs. Second, the possible presence of entire genes acquired through lateral transfer will affect the validity of the joint APC correction but not the intrafamily APC correction, since the latter requires only that genes were inherited as units.

Interestingly, the right side of the intrafamily MIp distribution shown in Figure 3 has both a shoulder and a tail, and the positions in contact have a significantly different distribution than the bulk of the pairs. These observations suggest that, while a small number of important contacting pairs share high MIp, there are many more pairs with lower levels of MIp that may also contribute to protein structure or function. We have found that position pairs with lower amounts of MIp show strong correlations with local and long-distance secondary structure (in preparation), and are currently working to incorporate these constraints to further use MIp to aid in de novo protein folding.

We also noted that within the GAPDH tetrameric structure, positions in contact between subunits had substantially higher average and median MIp than those in contact within a subunit. Furthermore, MIp can be applied to membrane
proteins where it has been used to identify key coevolving residues that allowed the identification of structural and functional signals in the Major Intrinsic Protein family of membrane proteins (Arinaminpathy, Gloor, Gerstein and Engelman, submitted for publication). Together with the identification of the intersubunit contact of positions 15 and 71 in triosephosphate isomerase (Table 2), these preliminary observations imply that MIp can identify proximal positions across subunit interfaces, and may lead to methods to identify interacting protein partners.

6 CONCLUSION

We have found that the background MI of each pair of positions in a multiple sequence alignment can be accurately estimated from mean MI values of each position. We have shown that the resulting value, the APC, accurately estimates the background MI provided the number of positions in the protein family is equal to or greater than the size of a typical protein domain. This measure can be rapidly calculated from the MI values for a multiple sequence alignment. Subtraction of the APC from the MI produces a new covariance measure which we term MIp. The MIp is substantially more sensitive and selective than previous methods at finding pairs of positions in contact in real protein families.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the insightful comments of two anonymous reviewers. Work in the labs of S.D.D. and G.B.G. is supported by operating grants from the Canadian Institutes of Health Research. Work by L.M.W. is supported by the National Science and Engineering Research Council of Canada, and by the Canada Research Chairs program.

Conflict of Interest: none declared.

REFERENCES


Noble, M.E. et al. (1991) The adaptability of the active site of trypanosomal triosephosphate isomerase as observed in the crystal structures of three different complexes. Proteins, 10, 5069.


