Sequence analysis

Using information theory to search for co-evolving residues in proteins

L. C. Martin1, G. B. Gloor2, S. D. Dunn2 and L. M. Wahl1,•

1Department of Applied Mathematics and 2Department of Biochemistry, University of Western Ontario, London, Ontario, Canada N6A 5B7

Received on July 20, 2005; revised on September 2, 2005; accepted on September 8, 2005
Advance Access publication September 13, 2005

ABSTRACT
Motivation: Some functionally important protein residues are easily detected since they correspond to conserved columns in a multiple sequence alignment (MSA). However important residues may also mutate, with compensatory mutations occurring elsewhere in the protein, which serve to preserve or restore functionality. It is difficult to distinguish these co-evolving sites from other non-conserved sites.

Results: We used Mutual Information (MI) to identify co-evolving positions. Using in silico evolved MSAs, we examined the effects of the number of sequences, the size of amino acid alphabet and the mutation rate on two sources of background MI: finite sample size effects and phylogenetic influence. We then assessed the performance of various normalizations of MI in enhancing detection of co-evolving positions and found that normalization by the pair entropy was optimal. Real protein alignments were analyzed and co-evolving isolated pairs were often found to be in contact with each other.

Availability: All data and program files can be found at http://www.biochem.uwo.ca/cgi-bin/CDD/index.cgi

Contact: lwahl@uwo.ca

Supplementary information: http://www.biochem.uwo.ca/cgi-bin/CDD/index.cgi

1 INTRODUCTION

Multiple sequence alignments (MSAs) of homologous proteins carry evolutionary information. Changes due to the mutation of amino acids at various positions in a protein are passed on to the next generation and may therefore reappear, through speciation, in the proteins of closely related organisms. In many instances functionality imposes constraints on these mutations, and important amino acid positions are often conserved. However mutation studies have shown that many non-conserved positions may also be functionally important. In particular, there may be compensatory mutations such that a mutation in one column of the MSA is typically matched by a corresponding mutation in another column. These co-evolving mutations are of key interest as they may identify residues that interact within the protein, residues which are functionally or structurally important, or possibly key sites of interaction between the protein and its substrate.

Mutual information (MI) is a measure of reduction of uncertainty (Ash, 1965; Cover and Thomas, 1991). In an MSA, the MI between two positions (columns) reflects the extent to which knowledge of the amino acid at one position allows us to predict the identity of the amino acid at the other position. MI values are high if substitutions at the two positions are correlated. MI is thus a natural measure with which to identify sites of correlated and compensatory mutations in homologous proteins.

Many investigators have reasoned that co-evolving positions in sequence alignments must be correlated and thus identifiable. Such identification is difficult because the signal that occurs due to compensatory mutations must be separated from the background signals that occur because of phylogenetic influences and small sample size effects (Atchley, 2000). Oliviera et al. (2002) developed a qualitative method, Correlated Mutation Analysis (CMA), to identify co-evolving residues based on the variability, entropy and correlation of the residues. However, their method determines important sites using a moving cutoff that differs with every protein family. Singer et al. (2002), in a variation of CMA using empirically derived likelihood scores, predicted residues to be in contact with a 16% accuracy for a variety of protein families. Pritchard et al. (2001) used the number of branching events for each residue to identify co-evolving sites. In previous studies that used MI, Korber et al. (1993) compared MI from an MSA to that from multiple randomizations of the original MSA and identified strongly co-varying residue pairs. Clarke (1995) corrected MI values by a metric corresponding to the number of observed amino acid pairs at each position. Wollenberg and Atchley (2000) used a model of phylogenetic substitution to attempt to correct for this background. However, most of these studies focused on those residue pairs with the highest MI values and assigned only qualitative significance to their scores. Tillier and Liu (2003) used mutual information and mutual interdependence to identify positions that co-evolved with each other, but were not co-evolving significantly with other positions in the alignment. Their method was highly selective, and did not rely on knowing the underlying relationships between the sequences, but could potentially miss positions that co-evolved
with several others. As yet, no single method has demonstrated general utility or achieved widespread acceptance.

Here we review the mathematical underpinnings of MI and its application to MSAs. Using simulated MSAs we quantify the background levels of MI due to the limited numbers of sequences and the underlying mutation rates; we then use in silico evolution to examine the contribution of shared evolutionary history. To simulate co-evolution due to structure or function, some positions are constrained to mutate in concert; the effects of these constraints are analyzed by comparing the MI of such pairs with the distribution of the MI of all position pairs in the alignment. In this comparison, the unconstrained positions provide a distribution of background signals arising from the phylogenetic relationships among the sequences and their limited number, sources common to all position pairs. A number of normalizations were tested for enhancement of the co-evolutionary signals above the background in the simulated sequences; division of MI by the pair entropy, a mathematically justifiable normalization, was optimal. This normalization was then applied to real protein MSAs, allowing us to identify position pairs that usually make contact in 3D protein structures.

2 MUTUAL INFORMATION

MI is a well known measure in Information Theory. Motivations for its properties and an axiomatic derivation are given by Ash (1965) and complementary material is found in Cover and Thomas (1991). We first give an overview of MI and then describe how it is applied to an MSA.

2.1 Definition and formulation of MI

MI is based on Shannon’s entropy, a measure of uncertainty for a random variable. The entropy $H(X)$ for a discrete random variable $X$, with possible values defined by an alphabet $\mathcal{X} = \{x_1, x_2, \ldots, x_K\}$ and associated probability distribution $p(X) = \{p(x_1), p(x_2), \ldots, p(x_K)\}$, where $\sum_{i=1}^{K} p(x_i) = 1$, is given by

$$H(X) = - \sum_{i=1}^{K} p(x_i) \log_b p(x_i).$$  \hspace{1cm} (1)

It follows that when $X$ can only take on one value, i.e. $X = x$, with probability $p(x) = 1$, then $H(X) = 0$, i.e. there is no uncertainty. When the probability is equally distributed over all members of the alphabet, i.e. $p(x_i) = 1/K$ for all $i$, then $H(X)$ is maximized. The choice of logarithm base $b$ serves to scale the entropy; for example, setting $b = K$ gives $\max[H(X)] = 1$. We use $b = 20$ throughout this work unless otherwise noted. Note that we formally set $\log 0 = 0$, the continuous limit, for elements whose probability is 0.

The concept of entropy can be easily extended to the case of two random variables where we have ordered pairs $(x_i, y_j)$. In this instance, it is helpful to think of the pairs as elements of an extended alphabet, $\mathcal{X} \times \mathcal{Y}$, whose elements are all possible distinct pairs. If we have a pair of random variables, then joint or pair entropy is defined as follows:

$$H(X, Y) = - \sum_{i=1}^{K} \sum_{j=1}^{L} p(x_i, y_j) \log_b p(x_i, y_j).$$  \hspace{1cm} (2)

Conditional entropy is given by $H(X|Y) = H(X, Y) - H(Y)$. Note that $H(X, Y) = H(Y, X)$, but that generally $H(X|Y) \neq H(Y|X)$ [equality is obtained if and only if $H(X) = H(Y)$].

Mutual Information, $MI(X, Y)$, is the reduction of uncertainty (as measured by entropy) of random variable $X$ given random variable $Y$ (as well as $X \rightleftharpoons Y$). We may thus write

$$MI(X, Y) = H(X) - H(X|Y)$$  \hspace{1cm} (3)

$$= H(Y) - H(Y|X)$$  \hspace{1cm} (4)

$$= H(X) + H(Y) - H(X, Y).$$  \hspace{1cm} (5)

MI is symmetric, i.e. $MI(X, Y) = MI(Y, X)$. Also note that $H(X, Y) = H(X)$, giving $MI(X, Y) = H(X)$.

Some important bounds on $H(X, Y)$ and $MI(X, Y)$ may be proven but are most readily observed from the Venn diagrammatic representation of entropy space shown in Figure 1. The joint entropy, $H(X, Y)$ is the area bounded by $H(X)$ and $H(Y)$, counting the area of intersection only once. Thus, one may show that $H(X, Y) \leq H(X) + H(Y)$ and $H(X, Y) \geq \max[H(X), H(Y)]$, with the equality occurring when one of $H(X)$ or $H(Y)$ is subsumed in the other. We see that $MI(X, Y)$, the area of intersection, is obtained at the expense of $H(X, Y)$. It follows that

$$0 \leq MI(X, Y) \leq \min[H(X), H(Y)].$$  \hspace{1cm} (6)

2.2 Mutual Information applied to MSAs

In an MSA, the amino acids in a given column can be considered as a set of observations of a random variable $x$. An estimate of the entropy $H(x)$ is obtained by using the observed amino acid frequencies, $f(x_i)$, in place of the underlying probabilities, $p(x_i)$, in Equation (1). A similar estimate for $H(X, Y)$ can be made using a second column in the MSA and the frequencies of all ordered pairs $(x_i, y_j)$ occurring in the rows of the two columns. It is then straightforward to obtain an estimate of $MI(X, Y)$ for that pair of columns by applying Equation (5).

The MI between two columns reflects the degree to which the pattern in the two columns is correlated. If amino acids occur independently at the two sites, the theoretical value for MI is zero. However, the estimate of the MI will only be zero if the observed pair frequencies reflect all possible pairings for the observed amino acid frequencies. Such a case is illustrated in Figure 10e, where $X = \{A, C; f(A) = 1/3, f(C) = 2/3\}$ and $Y = \{E, M; f(E) = 1/2, f(M) = 1/2\}$. Adding or deleting a row would result in different pair frequencies and yield a non-zero value of $MI(X, Y)$, as would interchanging dissimilar amino acids within a column.

When considering independent columns within finite MSAs, those alignments with fortuitous pairings resulting in vanishing MI values are the exception, as any deviation from the ‘all possible
pairs’ representation will yield a positive MI value. Also, as the number of amino acids occurring increases, the minimum sample size for which zero MI is possible increases quickly. If all 20 amino acids are present in each column and are equally probable, then zero MI is only possible if the frequency of each pair of amino acids is \(1/20^2\). This condition cannot be met in an MSA with <400 homologous proteins, and is unlikely to be met in columns of finite length. A statistical interpretation of MI is discussed in Cline et al. (2002), along with some cautionary remarks on the sample size.

3 SIMULATED MSAs

To investigate the properties of MI when applied to MSAs, we generated and analyzed simulated data. Although we recognize that in silico evolved sequences are unable to mimic many of the salient features of real MSAs, this process allowed us to isolate the effects of a number of key parameters on background MI. In particular, we examined the variation of MI with the number of sequences in the alignment, the size of the amino acid alphabet and the mutation rate used to generate the MSA. All parameters impact the mean entropy of the aligned columns, which is the measurable characteristic for MSAs obtained from real proteins. We first simulated MSAs in which mutations at each position in the protein were completely independent. This allowed us to quantify background MI. We then introduced co-evolving sites and evaluated the performance of MI in finding those sites. We expect that any correlated substitutions will have an observed MI above background values.

3.1 Background MI

There are two sources of background MI: finite sample size effects and phylogenetic influence.

Finite sampling effects. To quantify the effects of finite sample size on MI, we produced simulated MSAs in two ways. We produced matrices of \(n_s\) rows (sequences) and \(n_c\) columns (residues). Initially, all the entries were identical. Each position in the matrix was then allowed to ‘mutate’ with probability \(u\). If a mutation occurred, one of \(n_a\) (alphabet size) amino acids was randomly selected to replace the existing one. In the first case all members of the alphabet were equally probable. We then repeated this process, but selected the amino acid introduced by the mutation by using frequencies of amino acids observed in MSAs from real proteins. In particular, each column was randomly assigned to have the amino acid frequencies from one of the 314 ungapped columns of the MSAs for ATP synthase epsilon and gamma subunits (see the website for MSAs). Figure 2 shows the rapid decrease in mean MI as the number of sequences in these randomly generated alignments increases for either scenario. We note that alignments with amino acid frequencies from real proteins carry significantly less background MI on average. This is presumably because a number of columns are somewhat conserved and so have low entropies. Recall that the minimum entropy of either column is the maximum value that the MI can attain. Figure 2 suggests that for realistic amino acid frequencies, the contribution of background MI from finite column lengths is not too severe (when compared with values typically found in real protein alignments, see for example Fig. 9), particularly when the number of sequences in the alignment is greater than \(\sim 150\).

Phylogenetic influence. The second source of background MI is the evolutionary past inherent in every MSA of homologous proteins. In biologically meaningful alignments, no two positions can be independent, since two residues in the same protein are generally inherited as a unit. Thus, even when substitutions at the two sites occur with complete independence, the shared evolutionary history contributes a large background to the measured MI between those positions.

We quantified this effect by creating MSAs through in silico evolution. Binary trees with \(n_s\) tips were evolved using an initial sequence of identical amino acids, a branching probability \(p_b\), a site mutation probability \(\mu\) and equiprobable amino acid substitution. Sequences at the tips of the branches were then collected to form the simulated MSA.

Note that the ‘timestep’ used in this simulation is arbitrary, since both \(\mu\) and \(p_b\) are probabilities per timestep. The important non-dimensional parameter is the ratio of \(\mu\) to \(p_b\), which gives the average number of substitutions between branch nodes. Thus arbitrary values of \(\mu\) and \(p_b\) will give equivalent results, as long as this ratio is maintained. A value of \(p_b = 0.01\) was used in all the simulations.

This simple evolutionary model gives higher background levels of MI than other more complex models of evolution, thus providing a stringent test of our ability to detect co-evolution. MSAs evolved using a variety of more complex models from the PAML package (http://abacus.gene.ucl.ac.uk/software/paml.html) exhibited background MI with the same behavior though slightly reduced in magnitude (Supplementary data are shown at http://www.biochem.uwo.ca/cgi-bin/CDD/index.cgi).

An MSA of real proteins typically represents a small sample of branch tips from a much larger evolutionary tree. We might therefore simulate a very large binary tree and then randomly eliminate branches until \(n_s\) ‘species’ remain in the simulated MSA. This strategy is, however, formally equivalent to the technique described...
then find the value of an MSA. These values are size was adjusted to investigate various possible interpretations of data point is the average of values from 50 MSAs.

above, which simulates only the species that are ultimately part of the MSA. It should be noted that real MSAs do not meet this ideal of random sampling. In general, there are many examples of sequences from closely related sequences (e.g. pathogens). Finally, although it may be extremely difficult to estimate \( \mu/p_b \) from a real MSA of homologous proteins, we are fortunate in that the mean entropy of the columns in the MSA grows monotonically with mutation rate (Fig. 3). Thus to simulate data that are analogous to a given real MSA of length \( n_s \), we can choose an arbitrary branching ratio \( p_b \) and then find the value of \( \mu \) that gives an entropy similar to that of the real data.

Combined effects. This framework allows us to quantify the combined contribution of finite sequence lengths and shared evolutionary history to background MI. Figure 4 plots the mean background MI versus the number of sequences in an alignment; the figure is analogous to Figure 2, but for sequences evolved in silico as described above. Once again, we have not included any correlated substitutions in the simulated MSA, and therefore we might expect that the MI should be close to zero. However Figure 4 shows significant background MI due to the inherent correlations in sequences that share an evolutionary history. Figure 4a shows the decline in background MI as we add more sequences to the alignment; to generate these data \( \mu/p_b \) was held fixed at 0.1, and the alphabet size was adjusted to investigate various possible interpretations of an MSA. These values are \( n_a = 2 \) (of theoretical interest), 4 bases, 20 amino acids or 64 codons. In the second panel, we held \( n_a \) fixed at 20 and varied \( \mu \). This dataset demonstrates that for very low values of \( \mu/p_b \), the background MI contributed by shared evolutionary history is small, but is still increasing even when we have 1000 sequences in the alignment.

Figure 5 offers two additional views of the multi-dimensional relationship between background MI and the parameters \( n_a \) and \( \mu/p_b \). In Figure 5a background MI is plotted against \( \mu \) for various MSA lengths. We can see that MI generally increases with higher mutation rate, rising rapidly to a plateau value. With large numbers of sequences in the alignment, MI may exhibit a peak before plateauing at a lower level. Some intuition can be gained by noting that the worst case for finite sampling effects (i.e. the longest alphabet and smallest number of sequences) in this dataset, \( n_a = 20 \), has the highest plateau. As stated previously, the ideal case of zero background MI cannot be realized with short columns, but is approached more closely as \( n_a \) increases. Thus plateau levels are lower for alignments with more sequences.

Finally, in Figure 5b we plot the mean background MI versus the mean column entropy since this value is accessible when working with real MSAs, whereas the mutation rate is unknown. From this

**Fig. 3.** Mean column entropy versus mutation rate (a), and number of sequences (b), for in silico evolved MSAs. Both plots are for amino acids alphabet size of 20 \((n_a)\) and for sequences 100 residues in length \((n_s)\). A binary tree branching probability of 0.01 \((p_b)\) was used to evolve the sequences. Each data point is the average of values from 50 MSAs.

**Fig. 4.** Mean background MI versus number of sequences, \( n_s \), for various ‘amino acid’ alphabet sizes, \( n_a \) (mutation rate held constant at \( \mu = 0.01 \)) (a), and for various mutation rates, \( \mu \) (amino alphabet size held constant at \( n_a = 20 \)) (b) for in silico evolved MSAs \((p_b = 0.01)\). The logarithm base is set to the alphabet size (i.e. \( b = n_a \)) in the calculation of MI values. Both plots are for sequences 100 residues in length \((n_s)\). Each data point is the average of values from 50 MSAs.

**Fig. 5.** Mean background MI versus mutation rate (a), and mean entropy (b). Both plots are for \( n_a = 20 \), \( n_s = 100 \) and \( p_b = 0.01 \). Each data point is the average of values from 50 MSAs.
We first compare the MI for every possible pair of columns in our in silico evolved data, as shown in Figure 6 [because of symmetry, only $M(X,Y)$ for $Y > X$ is reported], thus determining whether the MI of partnered columns (shown as filled circles in this figure) is significantly higher than average. We see that this is not the case for a significant number of the partnered sites. The problem with this approach stems from the fact that the upper bound on MI for a given pair of columns is the lesser of the entropies of either column. Even though the partnered columns have an MI closer to saturation (45 degree line in Fig. 7a) for a given minimum column entropy than other independent columns with similar minimum entropies, the wide range of entropies present in the MSA is such that the level of background MI values due to the more entropic columns (portion of the cloud to the right) often surpasses the MI of the less entropic partnered columns. We observe that the background MI is highly correlated with the minimum entropy suggesting that normalizing $M(X,Y)$ by $\min\{H(X), H(Y)\}$ would ‘level the playing field’ and improve detection of those partnered columns with at least one low column entropy. This is true, as can be seen in Figure 7b, which shows MI/$\min\{H(X), H(Y)\}$ values; here the signal from the partnered sites is clearly stronger than the background. We can show, however, that detection ability may be further enhanced if we normalize by the upper bound on the MI, $H(X,Y)$. Even though the background MI is less correlated to this quantity (Fig. 7c), the separation of the MI for partnered columns from that of background $H(X,Y)$ peers is greater. This is demonstrated in Figure 7d, where we see that a greater fraction of the partnered sites have normalized MI values surpassing 4 standard deviations.

For a complete analysis, we normalized MI by all possible terms (or reasonable combinations of terms) appearing in Equations (3)–(5) and tested each normalization. Each set of scores was then converted to Z scores (i.e. number of standard deviations from the mean) and column pairs that scored above an assigned $Z$ cutoff were identified. Sensitivity (ratio of true positives to true positives plus false negatives) and specificity (ratio of true positives to true positives plus false positives) were then calculated.

We tested normalizing $M(X,Y)$ by each of the following: $H(X,Y), (H(X) + H(Y)), \min\{H(X), H(Y)\}, \max\{H(X), H(Y)\}, \min\{H(XY), H(IXY)\}$ and $\max\{H(XY), H(IXY)\}$. Normalizations $M(X,Y)/H(X,Y)$ and $M(X,Y)/(H(X) + H(Y))$ are not independent; the distributions for each differ but the rank ordering of all MI scores remains invariant. However, since a $Z$-score cutoff, and not a
percentile cutoff, is a more practical guideline for detecting significant sites in real proteins, we tested both.

The results of these tests for a $Z = 4$ cutoff are shown in Figure 8. Though most normalizations perform nearly equally well in terms of specificity, $MI/H(X, Y)$ is demonstrably better with respect to sensitivity. Both measures are optimized for values of $\mu > 10^{-3}$, which in turn corresponds (viz. Fig. 3a) to a mean entropy of $\sim 0.3$ for alignments whose size is on the order of $n_s = 100$. Therefore these tests suggest a natural entropy cutoff when analyzing real protein MSAs. Note that in both instances, raw MI performed significantly less well than any normalization in its ability to detect partners and filter out false positives.

4 ANALYSIS OF REAL PROTEIN FAMILIES

In this section, we apply normalized MI to real protein MSAs. We begin by describing a representative example, and then outline two different analyses that were performed to assess the predictive power of MI and its various normalizations with respect to residue contact.

4.1 Isolated pairs in triosephosphate isomerase

A multiple sequence alignment for triosephosphate isomerase was generated as described (Cover and Thomas, 1991). Mutual information and $Z$ scores were calculated as for the simulated alignments. Residues having an entropy $>0.3$ were then selected for the analysis that follows. This entropy cutoff restricted the analysis to those positions that contained sufficient variability that they were unlikely to show significant MI due to genetic linkage. One could also argue that those residues with entropy $<0.3$ are readily identifiable as (nearly) conserved columns. Residues that scored high in normalized MI relative to their peers ($Z > 4$) were identified and found to fall into two groups. The first was composed of residues that scored highly with one, and only one, other residue. We will call these ‘isolated pairs’. Isolated pairs often rank amongst the highest in the normalized MI distributions. The second group is composed of residues that score highly with more than one other residue and so form a network. Outliers belonging to this group have been determined to often be located at or near functional sites [they are discussed in detail in Gloor et al. (2005) and will not be further discussed here]. We now give an example illustrating the effect of normalization on those residues that are later identified as isolated pairs.

In Figure 9 we show MI and normalized MI for all pairs above the 0.3 entropy cutoff in the TIM alignment. In each plot, the data points for the isolated pairs are shown as open circles. One pair is nearly an outlier with respect to the MI distribution, and has the second-highest MI score. The other two pairs rank much lower and are at MI levels met by a significant portion of the cloud. After normalization, these three pairs have the highest $Z$ scores in the entire alignment (5.9, 4.4, 4.1). The three column pairs, (184,227), (68,70) and (135,176), were also the only outliers for this protein and were each singly connected, or isolated, pairs. The minimum separation between non-hydrogen atoms of these residues in the X-ray structure 1IIH.PDB (Noble et al., 1991) are 3.0, 3.5 and 7.6 Å.

4.2 Larger datasets—distant residue contact

Multiple sequence alignments for the Conserved Domain Database (CDD) (Marchler-Bauer et al., 2005) were obtained from the NCBI (January 2005) and used for the analysis. Alignments for protein families from PFAM (Alex et al., 2004), COG (Tatusov et al., 2003) and CDD were extracted from this dataset if they included one or more structures and contained at least 30 sequences. The 941 resulting alignments were examined for the presence of poorly aligned members. Sequences were removed from the alignments if they had 20% or more gaps than the reference structure’s sequence. Applying this rule resulted in 616 suitable sequence families. The MI and $Z$ score for each pair of residues in the 616 protein families above the 0.3 entropy cutoff was determined as for the modeled
alignments. Seventy-five percent of these families (464 of 616) had one or more residue pairs with a Z score >4 (‘high-scoring pairs’). This yield is substantially below that seen with alignments generated manually, where 22 of 23 alignments yielded one or more high-scoring residue pairs (Gloor et al., 2005). Thus, the alignments from the domain databases are generally suitable for this analysis, but increased sensitivity can be achieved by careful editing of the alignments.

We used these alignments to test the same hypothesis that was tested in Section 4.1 (that isolated pairs should be close in 3D space), but we added the constraint that they be separated by at least 10 residues in the reference sequence. Poon and Chao (2005) have shown experimentally that co-evolving residues are often in physical proximity, and this measure was also used by Tillier and Liu (2003) to test their ability to find co-evolving positions. For this, we measured the distance between positions in the 361 alignments (142 COG, 95 PFAM and 124 CDD alignments) that contained at least one isolated pair separated by 10 or more residues in sequence. The data in the first row of Table 1 show that the mean distance between the 435 isolated pairs in the complete set of 361 protein families is 17.4 Å compared with the average inter-residue distance of 23.0 Å in these protein families.

The significance of this difference was tested by randomly selecting with replacement, for each isolated pair, amongst those pairs from the data with the same sequence separation (bootstrap sampling). For example, for the complete dataset, 435 random inter-residue distances were selected such that the randomly selected set had the same sequence separations as the real dataset. This sampling was repeated for $10^6$ trials. The proportion of the trials that had a mean distance of 17.4 Å or less was then reported. We observed that none of the $10^6$ random samples had a mean distance of 17.3 Å or less (i.e. $P < 10^{-6}$), and thus conclude that it was extremely unlikely to choose a set of 435 inter-residue distances with this mean distance from the set of residue pairs with the same sequence separations by chance. This was done for datasets of MSAs with varying number of sequences and the results are shown in Table 1. All but the smallest MSA sets (8–24 sequences) have isolated pairs whose average distance is significantly less than that expected by chance.

Our modeled data also indicated that there should be a strong relationship between the specificity of the method and the number of sequences in the alignment. We therefore stratified the data and found that the high-scoring residue pairs from the 36 alignments containing at least 125 sequences were separated by <7 Å on average, but had a mean sequence separation of over 50 residues. Almost 60% of these residues were in contact (28/47, defined as a side-chain separation of 5 Å or less in the representative structure). In contrast, the mean distance for residue pairs from the alignments containing <125 sequences was at least 11 Å, and generally <30% of these were in contact. Therefore, the number of sequences in the alignments is a critical parameter for the detection of such positions, and we conclude that the modeled data accurately represent the minimum number of sequences required to extract useful MI values from real alignments.

### 4.3 Maximum residue MI

As described above, the various normalizations changed both the absolute value of MI and the rank order of pairs. To focus on the effects of the latter in our ability to detect pair contact, we examined the relationship between the maximum score achieved by a residue under the various normalizations, and the distance between the residue and its MI partner in the 83 alignments containing at least 125 sequences. Note that we are no longer restricting our analysis to isolated pairs. As before, this analysis was restricted to residues with entropies of 0.3 or greater and residue pairs separated by 10 or more other residues in the reference sequence.

This set of MSAs provided 1,097,964 residue pairs of which 23,612 (or 2.2%) were in contact (a side chain separation of 5 Å or less). The number of residue positions that met the entropy cutoff constraint was 8,556. The number of contacts identified by each normalization is given in Table 2. We observe that maximum MI without any normalization identifies the least number of contacts, whereas there is a significant improvement in accuracy when maximum conditional entropy, maximum entropy, sum of entropies and pair entropy are each used to normalize MI. Pair entropy and sum of entropies identified the greatest number of contacting positions.

### Table 1. The mean distance between residue pairs that have Z scores of 4 or more is strongly dependent on the number of sequences in the alignments, and is independent of the separation of these residues in the primary sequence.

<table>
<thead>
<tr>
<th>Number of sequences in alignment</th>
<th>Mean distance (Å)</th>
<th>Num. of isolated pairs</th>
<th>Mean separation (residues)</th>
<th>Fraction of pairs in contact</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>8–1436</td>
<td>17.4</td>
<td>435</td>
<td>98.6</td>
<td>0.26</td>
<td>&lt;1E−6</td>
</tr>
<tr>
<td>&gt;150</td>
<td>6.9</td>
<td>32</td>
<td>54.5</td>
<td>0.66</td>
<td>&lt;1E−6</td>
</tr>
<tr>
<td>125–149</td>
<td>6.6</td>
<td>15</td>
<td>52.1</td>
<td>0.47</td>
<td>&lt;1E−6</td>
</tr>
<tr>
<td>100–124</td>
<td>12.5</td>
<td>15</td>
<td>75.1</td>
<td>0.33</td>
<td>4.6E−5</td>
</tr>
<tr>
<td>75–99</td>
<td>10.4</td>
<td>40</td>
<td>80.1</td>
<td>0.48</td>
<td>&lt;1E−6</td>
</tr>
<tr>
<td>50–74</td>
<td>18.8</td>
<td>101</td>
<td>120.0</td>
<td>0.24</td>
<td>&lt;1E−6</td>
</tr>
<tr>
<td>25–49</td>
<td>18.4</td>
<td>171</td>
<td>96.5</td>
<td>0.20</td>
<td>&lt;1E−6</td>
</tr>
<tr>
<td>8–24</td>
<td>26.3</td>
<td>61</td>
<td>121.7</td>
<td>0.07</td>
<td>0.45</td>
</tr>
</tbody>
</table>

The P-value was determined by the bootstrap analysis as described in the text, and represents the likelihood of selecting the number of residue pairs in the column from the total set of all inter-residue distances, with an average inter-residue distance equal to or less than that observed. Note that all the residue pairs used in this analysis are separated by 10 or more intervening amino acids. Details may be found at http://www.biochem.uwo.ca/cgi-bin/CDD/index.cgi.

### Table 2. Number of residue positions, X, whose maximum MI(X,Y) for all Y that are separated by 10 or more amino acids successfully predict side chain contact with Y:

<table>
<thead>
<tr>
<th>Normalization</th>
<th>Number of maxima in contact</th>
<th>Fraction of maxima in contact (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>min(H(XY), H(YX))</td>
<td>840</td>
<td>9.8</td>
</tr>
<tr>
<td>min(H(X), H(Y))</td>
<td>877</td>
<td>10.3</td>
</tr>
<tr>
<td>max(H(XY), H(YX))</td>
<td>1218</td>
<td>14.2</td>
</tr>
<tr>
<td>max(H(X), H(Y))</td>
<td>1218</td>
<td>14.2</td>
</tr>
<tr>
<td>H(X) + H(Y)</td>
<td>1228</td>
<td>14.4</td>
</tr>
<tr>
<td>H(X,Y)</td>
<td>1228</td>
<td>14.4</td>
</tr>
</tbody>
</table>

All normalizations are tested and results are expressed both as a number and as a fraction of eligible (X,Y) pairs.
residues, albeit by a small margin over the previous two, and identified 46% more than un-normalized MI.

Some of the table entries are identical. As noted earlier, normalization by \( H(X) + H(Y) \) is not independent from \( H(X, Y) \) normalization, and so we obtain the same rank ordering in the transformations induced by these normalizations. Normalization by minimum entropy or conditional entropy yields the same highest scoring pair since, upon rewriting Equation (3) as \( H(X|Y) = H(X) - H(Y|X) = H(X) - MI(X, Y) \), we can see that for a given \( MI(X, Y) \), finding the maximum entropy of \( X \) or \( Y \) is equivalent to finding the corresponding conditional entropy.

To reiterate, by exclusively considering the highest MI and normalized MI values for each residue position, we do not rely on the form of the underlying distributions, nor on a cutoff (i.e. 4 standard deviations from the distribution mean, the value of which requires guidance from the analysis of simulated evolutions).

We have shown that the rank ordering induced by pair normalization yields top scoring pairs that are more frequently in contact than those of other normalizations.

5 DISCUSSION

5.1 The effect of normalizing MI by \( H(X, Y) \)

To clarify the effect of the normalization that we propose, it may be useful to consider several hypothetical MSAs and their corresponding Venn diagrams, as illustrated in Figure 10. In cases (a) and (b), the \( H(X) \) and \( H(Y) \) spaces overlap exactly (see Fig. 1 for labeling). This corresponds to two columns whose elements are uniquely paired, that is, an 'A' in column 1 occurs if and only if an 'E' occurs in column 2, etc. Although only unique pairs occur in both cases, in (b) greater column entropy results in greater MI. Normalization eliminates the influence of this disparity in entropy.

In case (c), the pairings are not unique; in particular an 'E' might correspond to either an 'A' or a 'C'. Here the uncorrected MI is similar to that in case (a), but when normalized there is a penalty associated with non-uniqueness. Likewise case (d) illustrates non- uniqueness in both directions. Because of the high column entropy the MI is large. Again, normalization reduces this effect. Case (e) demonstrates that when all possible ordered pairs occur at frequencies expected for two independent columns, both MI and normalized MI are zero.

When applied to MI from MSAs, the \( H(X, Y) \) normalization ranks the column pairs in the order of most unique mutations. For real protein evolution there may be subsets of the amino acids that are equally viable as compensatory mutation residues, several of which would appear in different homologous proteins. Those co-evolving sites would not score as well under this normalization.

Interestingly, maximizing \( MI(X, Y)/H(X, Y) \) corresponds to minimizing a normalized metric: \( D(X, Y) \equiv H(X, Y) - MI(X, Y) \) satisfies the three distance properties: (i) \( D(X, Y) = 0 \) iff \( X = Y \), (ii) \( D(X, Y) + D(Y, Z) \geq D(X, Z) \) and (iii) \( D(X, Y) = D(Y, X) \), thus

\[
\frac{D(X, Y)}{H(X, Y)} = 1 - \frac{MI(X, Y)}{H(X, Y)} \tag{7}
\]

is a normalized metric (Bardera et al., 2004).

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>AE</td>
<td>AE</td>
<td>AE</td>
<td>AE</td>
<td>AE</td>
</tr>
<tr>
<td>AE</td>
<td>AE</td>
<td>AE</td>
<td>AE</td>
<td>AM</td>
<td>AE</td>
</tr>
<tr>
<td>AE</td>
<td>CM</td>
<td>CE</td>
<td>CE</td>
<td>CE</td>
<td>CM</td>
</tr>
<tr>
<td>CM</td>
<td>CM</td>
<td>CE</td>
<td>DM</td>
<td>CM</td>
<td>CM</td>
</tr>
<tr>
<td>CM</td>
<td>DF</td>
<td>CE</td>
<td>DF</td>
<td>CF</td>
<td>CM</td>
</tr>
</tbody>
</table>

\[
\frac{MI}{H(X, Y)}
\]

\[
\begin{array}{cccccc}
1 & 1 & .58 & .52 & 0 \\
\end{array}
\]

Fig. 10. Examples illustrating the effect of normalizing MI by pair entropy for MSAs with varied column entropies (area of circles) and dependencies (area of overlap). Completely dependent columns do not gain by an increase in entropy (a versus b) with normalized MI. Similarly, columns with a similar amount of unnormalized MI rank lower in normalized MI (c versus d) if the overall entropy of the columns is greater.

6 CONCLUSIONS

We have investigated the properties of MI when applied to MSAs. In particular, we have quantified two sources of background MI for uncorrelated sites: finite sequence sampling effects and phylogenetic influences. Quantifying sampling effects suggests that background MI is a significant obstacle for alignments that have <125–150 sequences. A significant amount of background MI is due to the phylogenetic fingerprint on an MSA, even when mutations occur independently.

Evolution of MSAs via in silico evolution was carried out with a very weak set of rules for co-evolution, specifically no restrictions were placed either on the amino acid identity for mutations or on that of compensatory mutations when partnered sites mutated. We found that normalizing MI by the pair entropy optimized the ability to detect co-evolving sites over a large range of mutation rates. Importantly, our method requires no assumptions about either the underlying relationships between the proteins in the alignment or how rapidly or slowly the sequences may be evolving. The method only assumes that the MI caused by shared inheritance is relatively constant across each sequence, and that most pairs do not share significant MI due to structural or functional constraints. The significance scores thus reflect how much more MI a given residue has than its peers in that alignment.

We have applied this measure to MSAs obtained from real protein alignments and found that normalization notably improved the identification of important residue pairs that are in contact in the folded structures of the protein; most of the co-evolving, isolated pairs are in contact with their partner in representative structures. These residues may prove important to the structural stability of the folded protein.

ACKNOWLEDGEMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Canadian Institutes of Health Research.
and the Ontario Ministry of Science, Technology and Industry (Premier’s Research Excellence Awards). Funding to pay the Open Access publication charges for this article was provided by the University of Western Ontario.

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, 50–69.