BIOINFORMATICS DISCOVERY NOTE Vol. 26 no. 9 2010, pages 1135–1139
doi:10.1093/bioinformatics/btq111

Sequence analysis

Advance Access publication March 17, 2010

Mutual information is critically dependent on prior assumptions: would the *correct* estimate of mutual information please identify itself?

Andrew D. Fernandes[1,2,*] and Gregory B. Gloor[1]

[1]Department of Biochemistry, The University of Western Ontario, London, ON N6A 5C1 and [2]Department of Applied Mathematics, The University of Western Ontario, London, ON N6A 5B7, Canada

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Mutual information (MI) is a quantity that measures the dependence between two arbitrary random variables and has been repeatedly used to solve a wide variety of bioinformatic problems. Recently, when attempting to quantify the effects of sampling variance on computed values of MI in proteins, we encountered striking differences among various novel estimates of MI. These differences revealed that estimating the 'true' value of MI is not a straightforward procedure, and minor variations of assumptions yielded remarkably different estimates.

Results: We describe four formally equivalent estimates of MI, three of which explicitly account for sampling variance, that yield non-equal values of MI given *exact* frequencies. These MI estimates are essentially non-predictive of each other, converging only in the limit of implausibly large datasets. Lastly, we show that all four estimates are biologically reasonable estimates of MI, despite their disparity, since each is actually the Kullback–Leibler divergence between random variables conditioned on equally plausible hypotheses.

Conclusions: For sparse contingency tables of the type universally observed in protein coevolution studies, our results show that estimates of MI, and hence inferences about physical phenomena such as coevolution, are critically dependent on at least three prior assumptions. These assumptions are: (i) how observation counts relate to expected frequencies; (ii) the relationship between joint and marginal frequencies; and (iii) how non-observed categories are interpreted. In *any* biologically relevant data, these assumptions will affect the MI estimate as much or more-so than observed data, and are *independent* of uncertainty in frequency parameters.

Contact: andrew@fernandes.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 8, 2009; revised on February 21, 2010; accepted on March 10, 2010

1 BACKGROUND

Mutual information (MI) is an information-theoretic quantity measuring the dependence between two arbitrary random variables. It is used in a vast array of bioinformatic disciplines such as phylogenetics (Atchley *et al.*, 2000; Korber *et al.*, 1993), RNA secondary structure prediction (Bindewald and Shapiro, 2006), transcription factor binding site analysis (Tomovic and Oakeley, 2007) and protein coevolution (Dunn *et al.*, 2008; Wollenberg and Atchley, 2000), among many other fields.

Given two discrete random variables X and Y, the MI $\mathcal{I}(X,Y)$ shared by them is given by

$$\mathcal{I}(X,Y)=\sum_{y\in Y}\sum_{x\in X}\Pr(x,y)\log\left(\frac{\Pr(x,y)}{\Pr(x)\cdot\Pr(y)}\right). \quad (1)$$

Note that the summations in (1) can be replaced by abstract integration, allowing MI to be defined for arbitrary probability measures. However, for the remainder of this article we will implicitly assume that X and Y are discrete and finite. Virtually all bioinformatic applications of (1) involve multinomial likelihoods for $m\times n$, 2D contingency tables. A d-dimensional contingency table is a representation of the number of observed occurrences of d categorical variables, and when $d=2$ Kullback (1978) gives the total estimated MI as

$$\mathcal{I}(X,Y)=\sum_{i}\sum_{j}p_{ij}\log\left(\frac{p_{ij}}{r_i\cdot s_j}\right), \quad (2)$$

where p_{ij} is the joint probability that $\Pr(X=x_i,Y=y_j)$, r_i the marginal probability $\Pr(X=x_i)$ and s_i the marginal probability $\Pr(Y=y_j)$. Of course, the probability parameters p_{ij}, r_i and s_j are not observed. Instead, they must be inferred from the entries of a contingency table of observed events, where n_{ij} represents the number of observed events of joint class (i,j), $n_{i\star}$ the marginal counts $n_{i\star}=\sum_j n_{ij}$ and $n_{\star j}$ the marginal counts $n_{\star j}=\sum_i n_{ij}$.

Virtually all applications of (2) make two fundamental assumptions when estimating the frequency parameters from the contingency counts. These assumptions are that: (i) $p_{ij}\simeq n_{ij}/n$, where $n=\sum_i\sum_j n_{ij}$; and (ii) $p_{i\star}=\sum_j p_{ij}$ and $p_{\star j}=\sum_i p_{ij}$, as per the marginal counts. When assumed to hold, these conditions can be used to map a given table of counts n_{ij} to a unique point-estimate of MI, which we term $\mathcal{I}_p$. A graphical depiction of the overall procedure for estimating $\mathcal{I}_p$ from n_{ij} is shown in Supplementary Figure A-1.

This point-estimate $\mathcal{I}_p$ is the most commonly used estimate of MI reported in the literature. However, to mitigate the effect of categories n_{ij} with small counts, pseudocounts (Durbin *et al.*, 1998) have been used for estimating motifs for both protein domains (Buslje *et al.*, 2009; Henikoff and Henikoff, 1996) and transcription factor binding sites (Nishida *et al.*, 2009). In essence, pseudocounts estimate p_{ij} as proportional to $n_{ij}+c_{ij}$, where all $c_{ij}>0$ denote small constants, resulting in a different point-estimate

*To whom correspondence should be addressed.

of MI. Since our work focuses on describing variability that is intrinsic to the computation of MI from frequencies, rather than the inference of frequencies from counts, we constrain c_{ij} to be zero as this is the standard pseudocount value in comparable studies that utilize MI or a variant to study substitution covariance in proteins.

A notable exception to the implicit estimate of $p_{ij} \simeq n_{ij}/n$ is the work of Meyer *et al.* (2008), who detail several methods by which frequencies can be inferred from counts for computing MI. As with previous literature and in contrast to this work, however, Meyer *et al.* (2008) only infer frequency point-estimates from the count data, effectively discarding information regarding sampling variance, and implicitly but incorrectly assume that $\mathcal{I}_a$, defined below, is the only correct estimate of MI.

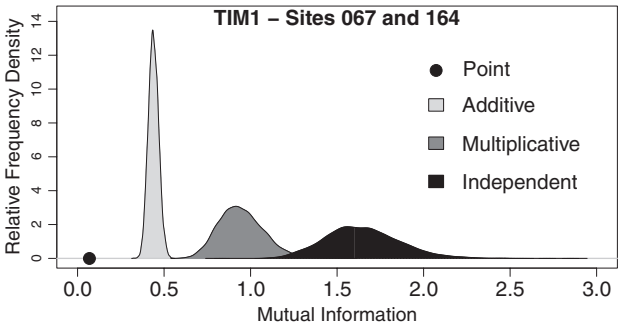

Fig. 1. Representative magnitudes and distributions for four estimates of MI. The figure is described in detail in Section 2. All logarithms use the natural base.

2 METHODS

2.1 Estimate of sampling variance

Estimates of sampling variance for $\mathcal{I}_a$, $\mathcal{I}_m$ and $\mathcal{I}_u$ were computed using standard Bayesian methods. Posterior estimates of p_{ij} were inferred from observed n_{ij} via Dirichlet priors and multinomial likelihoods, resulting in Dirichlet posteriors (Hutter and Zaffalon, 2005). All Dirichlet hyperparmeter components were set to $1/2$ following Berger and Bernardo (1992) as this value formally minimizes the influence of the prior on the posterior and is formally equivalent to using Jeffreys' reparameterization-invariant prior (Berger *et al.*, 2009).

2.2 Multiplicatively constrained MI

The decomposition of $\log(p_{ij}) = s_{ij} + d_{ij}$ is based on the observation that the p_{ij} parameters themselves comprise a probability density, and this density can be viewed as a proportional composition. The analysis of proportional compositions is best done via their logarithms (Aitchison, 1986) because such log-compositions are isomorphic to standard Euclidean vector spaces (Egozcue *et al.*, 2003). In these spaces, vector addition and scalar multiplication correspond to the physical amalgamation of compositional mixtures. Therefore, rather than being a theoretical artifice, the use of $\log(p_{ij})$ has a simple, direct and physical interpretation. The mathematical theory and a concrete example of computing $\mathcal{I}_m$ from p_{ij} is fully detailed in Supplementary Material B.

3 RESULTS

Our main results are 2-fold in that: (i) the two common assumptions that $p_{ij} \simeq n_{ij}/n$ and that the joint frequencies sum to the marginal frequencies are *not* required to estimate MI; and (ii) not using or modifying these assumptions results in dramatically varying estimates of MI especially for protein covariation data. The implication is that these two assumptions provide far more *a priori* information, and hence bias, to the estimate of MI than previously believed.

Our conclusions are based on the behavior of four equally valid, yet often dramatically different estimates of MI, described below. Differences among the estimates are particularly acute for the sparsely populated contingency tables commonly observed in studies of covariant substitution between two protein alignment sites. Note that although the natural base is used throughout, our results are invariant to choice of logarithm base.

3.1 Four estimates of MI

A representative sample of our four different estimates of MI is shown in Figure 1, with each being described below. Numerous other example plots with their associated contingency tables are given as Supplementary Material E. Our dataset consists of a representative structural alignment of 454 sequences of triose-phosphate isomerase (TIM1) gathered from GENBANK, restricted to a maximum of 90% identity. Site numbering follows the *Escherichia coli* Protein Data Bank entry **1TRE** and only utilizes ungapped sites exhibiting clear homology (Dunn *et al.*, 2008). The validity of using only ungapped sites can be substantiated by the extensive analysis of TIM1 functional covariance detailed by Merlo *et al.* (2007), with the additional benefit that sample-size effects do not complicate our analysis.

3.1.1 Point-estimate MI The first estimate of MI considered is the point-estimate $\mathcal{I}_p$ described above, without pseudocounts. This is the most common estimate of MI reported in the literature.

3.1.2 Additively constrained MI Rather than using pseudocounts, we instead use Bayesian techniques to estimate the posterior distribution of each component p_{ij} (see Section 4). We emphasize that essentially *any* technique could be used to estimate the posterior distribution of the p_{ij} *without* qualitative change in our findings. The most important property of what we deem the additively constrained estimate of MI, or $\mathcal{I}_a$, is that estimates of p_{ij} given n_{ij} are given a *range* rather than a single value. In this article, the range of p_{ij} is given as a parameter posterior distribution. However, even simple uniform-type error-estimates yield qualitatively similar findings.

Specifically, given a table of counts n_{ij}, a specific set of joint probabilities p_{ij} is drawn from the posterior density of $\Pr(p_{ij}|n_{ij})$. The marginal probability parameters $p_{i\star}$ and $p_{\star j}$ are computed through the previously discussed additive constraints $p_{i\star} = \sum_j p_{ij}$ and $p_{\star j} = \sum_i p_{ij}$. The procedure is repeated to build a Monte Carlo estimate of the posterior distribution of MI. This estimate of MI, deemed $\mathcal{I}_a$, is the one most commonly used whenever sampling variance is accounted for (Hutter and Zaffalon, 2005).

3.1.3 Multiplicatively constrained MI The constraint that the joint frequencies precisely sum to the marginal frequencies is not the only mathematically or physically reasonable assumption, however. For example, if $\log(p_{ij})$ is used rather than p_{ij} for parameterization, standard techniques from linear algebra can be used to partition $\log(p_{ij}) = s_{ij} + d_{ij}$ as further elucidated in Section 4. The components s_{ij} specifically quantify row/column

independence and the components d_{ij} specifically quantify row/column *dependence*. We deem the estimate of MI comparing p_{ij} and s_{ij} to be $\mathcal{I}_m$.

Although the use of the log-frequencies may be unfamiliar, we emphasize that their use is quite common in standard statistical theory. Log-frequencies are the 'natural' parameter space for the multinomial distribution on which (2) is based. As the multinomial is an exponential-family distribution, many of its properties have been elucidated and are best understood in the logarithmic natural parameter space.

3.1.4 Independent, unconstrained MI Although perhaps counter-intuitive, there is no *a priori* necessity to condition on the joint and marginal frequencies being consistent. To understand why, consider the following thought experiment. The table of joint counts n_{ij} is given to Alice from which she is told to estimate the p_{ij} parameters without assuming any particular structure in the data. Independently, the table of counts is summed to the marginals $n_{i\star}$ and $n_{\star j}$ which are given to Bob in order to estimate $p_{i\star}$ and $p_{\star j}$, again without further assumptions. Alice and Bob may each choose *different*, yet equally valid, methods of inferring frequencies from counts, some of which have been cataloged by Meyer *et al.* (2008). Alice's frequencies are inferred under the hypothesis of dependence whereas Bob's are inferred under the hypothesis of independence. Only *after* inferring frequencies from counts do Alice and Bob need to compare their respective joint frequencies using MI. Since Alice and Bob do not share information before comparing frequencies, there is no *a priori* requirement that joint frequencies sum to equal the marginal frequencies, and in general $p_{i\star} \neq \sum_j p_{ij}$ and $p_{\star j} \neq \sum_i p_{ij}$.

This thought experiment shows the plausibility of $\Pr(p_{ij}|n_{ij})$ being wholly unconstrained by either $\Pr(p_{i\star}|n_{i\star})$ or $\Pr(p_{\star j}|n_{\star j})$. We deem such an unconstrained estimate of MI to be $\mathcal{I}_u$. Of course, given any reasonable dataset n_{ij} it is likely but not strictly necessary that the inferred parameters be reasonably consistent with $p_{i\star} \sim \sum_j p_{ij}$ and $p_{\star j} \sim \sum_i p_{ij}$. Far from being a statistical artifice, this thought-experiment informally yet precisely describes the information-independence between the hypotheses of 'dependence' versus 'independence' in contingency table analysis: small 'errors' in the frequencies inferred by Alice are necessarily independent of small 'errors' in the frequencies inferred by Bob.

3.2 Properties of the estimates

In examining the ∼24 000 pairs of alignment sites for TIM1, it appears that both the central tendency and dispersion of the estimates robustly follow the ordering $\mathcal{I}_p < \mathcal{I}_a < \mathcal{I}_m < \mathcal{I}_u$ as displayed in Figure 1 and the Supplementary Material. In general, all three distributions were distinct except when both alignment sites were highly conserved (see sites 010 and 077 in Supplementary Material E) in which case $\mathcal{I}_p \sim 0$ and $\mathcal{I}_m \sim \mathcal{I}_u$ were their largest.

This latter case was among the most intriguing since site-pair conservation, which implies that one entry of the n_{ij} receives the majority of observations, minimizes $\mathcal{I}_p$ while maximizing $\mathcal{I}_m$ and $\mathcal{I}_u$. This discrepancy highlights how much 'information' is implicitly added by the assumption that $n_{ij} = 0 \Rightarrow p_{ij} = 0$.

3.2.1 Point-estimate comparison The inability of $\mathcal{I}_p$ to meaningfully predict any of the three distributional estimates is shown by the left-hand panels of Figure 2. Although the relative ordering of the estimates was almost always as shown, this ordering

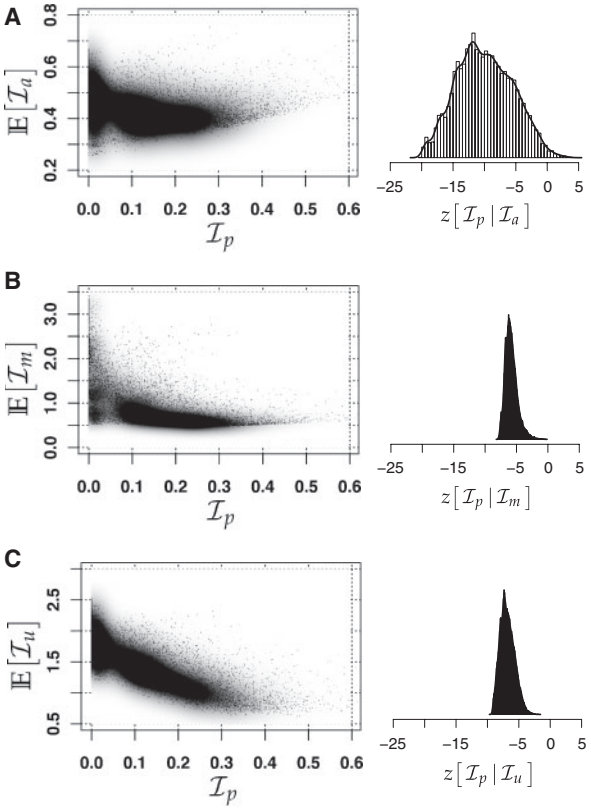

Fig. 2. Comparisons of $\mathcal{I}_p$ versus the expected value, denoted by '$\mathbb{E}[\cdot]$', of (**A**) $\mathcal{I}_a$, (**B**) $\mathcal{I}_m$ and (**C**) $\mathcal{I}_u$. Left-hand panels show scatter plots of raw values for all ∼24 000 ungapped site-pairs of TIM1. The slight negative correlation for (**C**) is discussed in the text. Right-hand panels show the *relative* difference between $\mathcal{I}_p$ and the distributional estimate given as a standardized z-score. Note that distributional estimates such as shown in Figure 1 are approximately Gaussian, implying that the z-score is a reasonable measure of correspondence between point- and distributional-estimate. Although mildly correlated, the large dispersion of each scatterplot show that $\mathrm{MI_p}$ is not a meaningful predictor of $\mathcal{I}_a$, $\mathcal{I}_m$ or $\mathcal{I}_u$.

does not necessarily imply that $\mathcal{I}_p$ is a 'closer' estimate to $\mathcal{I}_a$ than either $\mathcal{I}_m$ or $\mathcal{I}_u$. For example, if the 'distance' between the point and distribution-mean estimates is taken to be the standardized score z-score, then $\mathcal{I}_a$ is considerably *less predictive* of $\mathcal{I}_p$ than either $\mathcal{I}_m$ or $\mathcal{I}_u$ are.

The negative correlation seen in Figure 2C further highlights the information inherent in the assumption that zero counts imply a zero frequency. When both sites are highly conserved, implying that the majority of counts occur in one n_{ij} category, $\mathcal{I}_p$ tends to zero while $\mathcal{I}_m$ and $\mathcal{I}_u$ tend to overlap and attain their largest values (see sites 010 and 077 in Supplementary Material E). The point estimate $\mathcal{I}_p$ assumes that non-observed categories contain precisely zero information, whereas the distributional estimates assume that each of the ∼400 categories contributes a 'small amount' of MI to the total, as shown in a plot of $\sum_i p_i \log(p_i)$ in Supplementary Figure A-2.

Among the distributional estimates, scatter plots of mean values for $\mathcal{I}_a$, $\mathcal{I}_m$ and $\mathcal{I}_u$ (Supplementary Fig. A-3) show that values are only mildly predictive of each other. The strongest agreement was

seen between $\mathcal{I}_m$ and $\mathcal{I}_u$, with the weakest between $\mathcal{I}_a$ and $\mathcal{I}_u$. This observation can be interpreted to imply that joint-to-marginal consistency is a less-stringent constraint in the multiplicative $\log(p_{ij})$ parameter space than the additive p_{ij} parameter space.

3.2.2 Convergence of the estimates Although the four estimates differ substantially when the contingency table n_{ij} is sparse, standard arguments show that a sufficient condition for all four estimates to converge is asymptotic normality. Such normality is guaranteed only if a sufficient number of observations (usually 5–10) are seen in *every* category n_{ij}. In this case, the multinomial likelihood is well-approximated by the multivariate normal and all estimates converge in distribution.

In actual protein data, site-specific residues have well-known constraints due to participation in catalytic function or secondary/tertiary structural elements. These biological constraints imply that asymptotic normality *cannot* be achieved in real data. Therefore, we expect that non-commensurate estimates of MI as shown in Figure 1 are not exceptional and are in fact inescapable.

A more detailed examination of the convergence of $\mathcal{I}_p$, $\mathcal{I}_a$, $\mathcal{I}_m$ and $\mathcal{I}_u$ with respect to sample-size and dependence-to-independence ratio is presented in Supplementary Materials D and F. These examinations show that under biologically realistic amino acid frequencies displaying maximal variance as given by the WAG-model equilibrium frequencies (Whelan and Goldman, 2001) and *no* phylogenetic correlation, between 1000 and 10 000 sequences are needed for convergence, with greater dependency requiring a *larger* number of sequences. Since phylogeny and amino acid preference (smaller variance) can be seen by inspection to reduce the effective sample size, the *actual* number of biological replicates required would be much greater.

Although the example presented in Supplementary Material D might seem to imply that $\mathcal{I}_p$ is the 'better' estimate of the four when sample sizes are small and there is relatively little dependence, such a conclusion is incorrect because it is not known *a priori* that dependency is in fact small. In fact, it is arguable that since values of $\mathcal{I}_p$ appear to be relatively insensitive to both the number of sequences and the amount of modeled dependence that $\mathcal{I}_p$ is less able to discern true covariation from 'noise' than the other measures. An examination of convergence issues for different groups of protein orthologs is detailed in Supplementary Material F. Although it is *theoretically* impossible that one of the four estimates is, in any global sense, a 'better' estimator of MI than the others, it is certainly possible that *in practice* one or more estimators may outperform the others. However, we emphasize that this study provides *no* evidence of such superiority even in light of the single example illustrated by Supplement Material D.

3.2.3 Boundedness of the estimates Using the method of Lagrange multipliers, it is straightforward to show that both $\mathcal{I}_p$ and $\mathcal{I}_a$ have a finite range (Hutter and Zaffalon, 2005). Specifically, for a d-dimensional contingency table with m_k marginal classes possessing $m_1 \times m_2 \times \cdots \times m_d$ joint classes, it can be shown that $0 \leq \mathcal{I}_p, \mathcal{I}_a \leq \log\left(\max_k\{m_k\}\right)$.

In contrast, both $\mathcal{I}_m$ and $\mathcal{I}_u$ are *unbounded* and can, with non-zero probability, take any non-negative value. Such unboundedness does not necessarily make these $\mathcal{I}_m$ and $\mathcal{I}_u$ incongruous with $\mathcal{I}_p$ and $\mathcal{I}_a$ since large values of the former only occur if the joint and marginal frequencies are asymptotically inconsistent. Examples of such unlikely parameter estimates are discussed in Supplementary Material C.

4 CONCLUSIONS

MI is a tool commonly used to define and study a type of generalized covariance between two random variables. In this work, we show that for sparsely populated contingency tables there exist at least four different estimates of MI. These estimates are poor predictors of one other and often differ greatly when compared on both relative and absolute scales. Unlike previous work, where differences between estimates of MI are universally attributed to differences in how frequency parameters are estimated from counts, we show that the definition of MI *itself* may be responsible for uncertainty in its value. For example, given *precise* frequency parameters $p_{ij} = \frac{1}{16} \cdot \left[\begin{smallmatrix} 3 & 7 \\ 1 & 5 \end{smallmatrix}\right]$, the additive $\mathcal{I}_a = -5 \cdot \log(2) + 9/8 \cdot \log(3) + 5/8 \cdot \log(5)$ whereas the multiplicative $\mathcal{I}_m = -6 \cdot \log(2) + 3/16 \cdot \log(3) + 1/2 \cdot \log(\sqrt{105} - 7) + 1/2 \cdot \log(15 - \sqrt{105}) + 5/16 \cdot \log(5) + 7/16 \cdot \log(7)$. Thus, even for *exact* frequencies p_{ij} we have $\mathcal{I}_a \neq \mathcal{I}_m$. To our knowledge, this is the first demonstration that definition of MI *itself* possesses intrinsic uncertainty. Such uncertainty may explain why MI, either in raw or corrected form, is in often an unreliable estimate of categorical dependency (Dunn *et al.*, 2008; Martin *et al.*, 2005).

Minimizing such differences requires that the estimates mutually converge, and such convergence requires implausibly large biological datasets. Thus, the assumed relationship between the joint and marginal frequencies and the interpretation of unobserved categories will inevitably have non-negligible consequences. The breadth of these consequences may explain the broad range of results observed by Buslje *et al.* (2009) with respect to the MI-based detection of protein coevolution. Moreover, the example convergence discussed in Supplementary Material D taken in the context of the Central Limit Theorem implies that a *necessary* although insufficient condition for an MI estimate to be principally dependent on the data rather than assumptions is that the expected value of each of the four estimates must be equal. As shown in Supplement Material F, an analysis of 10 different alignments across 7 functionally and structurally diverse ortholog families reveals that even with alignments with $\sim$1000 sequences, estimates of MI are primarily dependent on prior assumptions, not observed data.

By carefully tracking all conditional assumptions inherent in our four MI estimates, we see that both $\mathcal{I}_p$ and $\mathcal{I}_a$ are inferred through $\Pr(p_{ij}|n_{ij})$ and $\Pr(p_{i\star}, p_{\star j}|n_{ij}, p_{ij}, A)$, where A denotes the hypothesis of an additive constraint. The $\mathcal{I}_p$ estimate builds on $\mathcal{I}_a$ with the additional conditional restriction that $p_{ij} = n_{ij}/n$, a strong assumption that collapses the distribution of $\mathcal{I}_a$ to the single-point $\mathcal{I}_p$, itself a poor predictor of $\mathcal{I}_a$. Similarly to $\mathcal{I}_a$, the $\mathcal{I}_m$ estimate is inferred through $\Pr(p_{ij}|n_{ij})$ and $\Pr(p_{i\star}, p_{\star j}|n_{ij}, p_{ij}, M)$, where M denotes the hypothesis of multiplicative subspace decomposition. Lastly, $\mathcal{I}_u$ is inferred via $\Pr(p_{ij}|n_{ij})$, $\Pr(p_{i\star}|n_{i\star})$ and $\Pr(p_{\star j}|n_{\star j})$ with the only conditional assumptions coming from observed data. All four of the resultant MI estimates have the form $\mathcal{I}(p,q) = \sum_{ij} p_{ij} \log(p_{ij}/q_{ij})$ and can, therefore, be equivalently viewed as Kullback–Leibler divergences between distributions conditioned on different hypotheses and parameter estimates (Kullback, 1978).

From a statistical viewpoint, none of these estimates are 'more correct' than any other since each merely conditions on different assumptions. What we have shown is that estimates of MI are necessarily *as* or *more* affected by these assumptions than the actual

observed data. The 'correct' conditional assumptions to use, we believe, are those that more closely match known or hypothesized biological constraints. For example, a pair of protein sites forming a putative salt bridge may assume that $p_{ij}=0$ for all joint classes possessing uncharged amino acids. Similarly, sites involved in putative α-helices may *a priori* disallow proline and proportionally weight the prior probability of other amino acids by their known helical-forming propensities. Unlike the filtering methodology of Codoñer *et al.* (2008) that is applied only to $\mathcal{I}_a$, corrections to MI as per Dunn *et al.* (2008), or different ways of inferring frequencies as per Meyer *et al.* (2008), this information needs to assist in the construction of the MI computation *itself*.

A consequence of requiring that mathematical and biological assumptions be concordant is the impossibility of meaningfully inferring coevolution using only pair-count data. Like the examples above that conditioned on possible salt-bridges or α-helices, we posit that additional biological knowledge is a necessary prerequisite for meaningful inferences about coevolution. The difficult problem of how to best express such often-incomplete biological information as conditional prior-probabilities remains an open question for future research.

ACKNOWLEDGEMENTS

Chris DeHaan provided helpful discussions and ideas. Lindi Wahl provided mentorship, manuscript review and funding. All computations were done using the **R** language (2009).

Funding: Natural Sciences and Engineering Research Council of Canada.

Conflict of Interest: none declared.

REFERENCES

Aitchison,J. (1986) *The statistical analysis of compositional data. Monographs on statistics and applied probability*. Chapman and Hall, New York, London.

Atchley,W.R. *et al.* (2000) Correlations among amino acid sites in bhlh protein domains: an information theoretic analysis. *Mol. Biol. Evol.*, **17**, 164–178.

Berger,J.O. and Bernardo,J.M. (1992) Ordered group reference priors with application to the multinomial problem. *Biometrika*, **79**, 25–37.

Berger,J.O. *et al.* (2009) The formal definition of reference priors. *Ann. Stat.*, **37**, 905–938.

Bindewald,E. and Shapiro,B.A. (2006) RNA secondary structure prediction from sequence alignments using a network of *k*-nearest neighbor classifiers. *RNA*, **12**, 342–352.

Buslje,C.M. *et al.* (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, **25**, 1125–1131.

Codoñer,F.M. *et al.* (2008) Reducing the false positive rate in the non-parametric analysis of molecular coevolution. *BMC Evol. Biol.*, **8**, 106.

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Durbin,R. *et al.* (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.

Egozcue,J.J. *et al.* (2003) Isometric logratio transformations for compositional data analysis. *Math. Geol.*, **35**, 279–300.

Henikoff,J.G. and Henikoff,S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.*, **12**, 135–143.

Hutter,M. and Zaffalon,M. (2005) Distribution of mutual information from complete and incomplete data. *Comput. Stat. Data Anal.*, **48**, 633–657.

Korber,B.T. *et al.* (1993) Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl Acad. Sci. USA*, **90**, 7176–7180.

Kullback,S. (1978) *Information theory and statistics*. Dover, Mineola, New York.

Martin,L.C. *et al.* (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.

Merlo,L.M.F. *et al.* (2007) An empirical test of the concomitantly variable codon hypothesis. *Proc. Natl Acad. Sci. USA*, **104**, 10938–10943.

Meyer,P. *et al.* (2008) *minet:* A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Nishida,K. *et al.* (2009) Pseudocounts for transcription factor binding sites. *Nucleic Acids Res.*, **37**, 939–944.

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at `http://www.R-project.org` (last accessed date February 21, 2010).

Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941.

Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.

Wollenberg,K.R. and Atchley,W.R. (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc. Natl Acad. Sci. USA*, **97**, 3288–3291.