# Information-theoretic dissection of pairwise contact potentials


Melissa S. Cline *

Center for Biomolecular Science and Engineering

Baskin School of Engineering

University of California

Santa Cruz, CA

cline@soe.ucsc.edu

Kevin Karplus

Computer Engineering

Baskin School of Engineering

University of California

Santa Cruz, CA

karplus@soe.ucsc.edu

Richard H. Lathrop

Information and Computer Science

University of California

Irvine, CA

rickl@uci.edu

Temple F. Smith

Biomedical Engineering

Boston University

Boston, MA

Modular Genetics Inc.

Cambridge, MA

tsmith@darwin.bu.edu

Robert G. Rogers Jr.

Biomedical Engineering

Boston University

Boston, MA

rogers@darwin.bu.edu

David Haussler

Center for Biomolecular Science and Engineering

Baskin School of Engineering

University of California

Santa Cruz, CA

haussler@soe.ucsc.edu

**Abstract**

Pairwise contact potentials have a long, successful history in protein structure prediction. They provide an easily-estimated representation of many attributes of protein structures, such as the hydrophobic effect. In order to improve on existing potentials, one should develop a clear understanding of precisely what information they convey. Here, using mutual information, we quantified the information in amino acid potentials, and the importance of hydropathy, charge, disulfide bonding, and burial. Sampling error in mutual information was controlled for by estimating how much information cannot be attributed to sampling bias. We found the information in amino acid contacts to be modest: 0.04 bits per contact. Of that, only 0.01 bits of information could not be attributed to hydropathy, charge, disulfide bonding, or burial.

# 1 Introduction

Contact potentials, first invented in 1975 by Levitt and Warshel [1], are an essential component of many methods for protein structure design, protein structure prediction, and docking prediction [2, 3, 4]. Contact potentials are generally estimated according to the frequencies of residue contacts observed in a reference database of protein structures. While there are many different types of potentials, this paper focuses on amino acid potentials which estimate the probability of contact according to the statistical likelihood of finding two amino acids within a certain pairwise distance. These potentials are sometimes called Miyazawa-Jernigan potentials [5].

What makes these contact potentials effective? The general view is that they provide an easily-calculated representation of many aspects of protein structure, such as hydropathy, electrostatics, disulfide bonding, and stacking of aromatic side-chains. Hydropathy is the largest component of pairwise potentials [6, 7], and pairwise potentials show a strong statistical relation to hydropathic indices [8, 9]. Simple Hydrophobic-Polar (HP) potentials, which reflect only the hydropathy of the amino acids, perform respectably compared to potentials which reflect amino acid identity [10, 11, 12].

*To whom correspondence should be addressed. Mailing address: Center for Biomolecular Science and Engineering, Jack Baskin School of Engineering, University of California, Santa Cruz, CA 95064 USA. Phone: 1-831-459-4250, Fax: 1-831-459-4829

Unfortunately, pairwise potentials also reflect data collection artifacts. Thomas and Dill [12] observed that the likelihood of H-P and P-P contacts are underestimated due to the dominance of H-H contacts and limitations imposed by chain connectivity and excluded volume. Pairwise potentials also reflect the size [13], secondary structure composition [13], and stability [14] of the proteins in the reference database. Further, there appears to be more to protein structure than is captured in amino-acid pairing frequencies. There are proteins for which no simple pairwise potential could predict all actual contacts successfully [15, 16]. Sunyaev et. al. [17] tested statistical pairwise preferences in the context of backbone conformation, accessibility, and pairwise distances, and showed that most of the signal comes from roughly half of the amino acids. The other half appear to be "average", with no marked preference for any contact or environment.

While pairwise amino acid potentials have limitations, they have also been the basis of many successful methods. Thus, rather than dismiss pairwise potentials, we should study their behavior for the purpose of devising better potentials. Exactly what information do pairwise potentials represent? To what extent do they encode hydropathy, burial, electrostatics, and disulfide bonding, and to what extent do they encode further information?

We addressed these questions using *mutual information*, an information-theoretic measure that quantifies how much knowing one quantity reveals about another [18]. With mutual information, we analyzed amino acid contacts according to specific amino acid characteristics, and quantified the amount of the pairwise signal that could be derived from these characteristics alone.

One pitfall of mutual information is that it is prone to sampling error [19]. To avoid this pitfall, we estimated the mutual information that we would expect if sampling error were the only source of observed mutual information. This allowed us to test and reject the hypothesis that interacting amino acids were distributed independently, and to measure how much information was present between interacting amino acids beyond that which could be attributed to sampling error.

# 2 Methods

## 2.1 Dataset of Pairwise Contacts

Our dataset consisted of the pairwise contacts observed in the structures of 208 proteins, most of which were cytosolic. From these structures, we extracted a set of 22,707 amino acid contacts. In the tradition of threading algorithms, we used only the contacts between residues in secondary structural elements: $\alpha$-helices or $\beta$-strands. All contacts had a $C_\beta$ distance of eight angstroms or less, and were separated in the polypeptide chain by four or more residues, precluding contacts between adjacent turns in $\alpha$-helices. While we also studied contacts between adjacent turns of $\alpha$-helices, they appeared to be less informative (data not shown). These 22,707 contacts were made by 18,198 residues, for an average of approximately 1.25 contacts per residue.

For each residue in each pairwise contact, the dataset described the residue's amino acid type and its *solvent exposure*. Solvent exposure is an accessibility measure designed for threading algorithms. Compared to other measures of accessibility and burial, it retains less indirect information about the protein sequence. First, the side chains of all amino acids were replaced with that of Alanine. Then, to avoid spurious holes in the interior of the protein, the radii of $C_\beta$ and the water sphere were both increased, to 2.1 angstroms and 2.4 angstroms respectively. Finally, Eisenberg's algorithm [20] was used to calculate the accessible area [21].

## 2.2 Mutual information

The *mutual information* $I(X,Y)$ of two discrete random variables $X$ and $Y$ quantifies the amount of information that either variable reveals about the other[22, 18]. It as calculated as follows:

$$\begin{aligned} I(X,Y) &= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= E\left[\log_2 \frac{p(x,y)}{p(x)p(y)}\right] \end{aligned}$$

where the sum is over all possible pairs of values $x$ and $y$ of the variables $X$ and $Y$, and E[.] is the expected value. If $X$ and $Y$ are independent, their mutual information is zero. Otherwise, mutual information is positive, with larger values indicating greater dependency. When calculated with base 2 logarithms, mutual information is expressed in units of *bits*.

Let $x$ and $y$ represent the amino acid types of two residues in contact, and let $P(x, y)$ represent the probability of their contact. If pairwise interactions were independent of amino acid type, then $P(x, y)$ would be $P(x)P(y)$. By making this substitution in the formula for mutual information, we see that the mutual information $I(X, Y)$ would be zero. On the other hand, suppose pairwise interactions depended entirely on amino acid type, with amino acid type $x$ interacting only with amino acid type $y$. Then, the chance of seeing any $(x, y)$ pair would be $P(x, y) = P(x)$. By making this substitution in mutual information formula, we see that $I(X, Y)$ would be $\sum_{x,y} P(x) \log_2 \frac{1}{P(y)}$, which works out to approximately 3.9 bits.

## 2.3 Residue alphabets

Measuring mutual information generally requires assigning each residue into some category. The most obvious set of categories is the twenty amino acid types. Mutual information measured according to amino acid type tells us the extent to which pairwise contacts are determined by the many factors related to amino acid properties: hydropathy, aromaticity, electrostatics, and so forth.

There are other categories associated with other amino acid attributes. Measuring mutual information according to these categories quantifies the importance of the corresponding attributes. For example, we might wish to ask whether charge explains most of the information in pairwise contacts, or other attributes are also important. To analyze the importance of charge, one could group residues according to charge: positive, negative, and neutral. Then, by measuring mutual information with this three-letter alphabet and comparing it to that measured with the twenty-letter amino acid alphabet, one can estimate the importance of charge to pairwise contacts. If pairwise contacts contained a large amount of information besides charge, then mutual information measured according to charge, $I_c(X, Y)$, would be small compared to that measured by amino acid type, $I_A(X, Y)$. This is because mutual information is calculated from the likelihood of pairwise contacts, $P(X, Y)$. If contact likelihood depended on more than charge, then assessing contact likelihood

| Alphabet | Contents |
|---|---|
| I | Cys and Other |
| II | Positive, Negative, Other |
| III | Hydrophobic and Polar |
| IV | Cys, Positive, Negative, Other Polar, Other Hydrophobic |
| V | The standard alphabet of amino acid types |

| Alphabet | Category | Contents |
|---|---|---|
| I | Cys | Cys |
| | Other | All Others |
| II | Positive | Arg, His, Lys |
| | Negative | Asp, Glu |
| | Neutral | All Others |
| III | Hydrophobic | Ala, Cys, Gly, Ile, Met, Phe, Pro, Trp, Tyr, Val |
| | Polar | Arg, Asn, Asp, Glu, Gln, His, Lys, Ser, Thr |
| IV | Cys | Cys |
| | Positive | Arg, His, Lys |
| | Negative | Asp, Glu |
| | Other Polar | Asn, Gln, Ser, Thr |
| | Other Hydrophobic | Ala, Gly, Ile, Leu, Met, Phe, Pro, Tyr, Val |

Table 1: This table describes the five alphabets used to relate amino acid contact patterns to biochemical forces, and details the amino acids assigned to each alphabet.

using charge alone would account for only part of the total signal encoded in pairwise contacts. On the other hand, if contact likelihood depended mostly on charge, then little information would be lost by estimating contact likelihood from charge, and $I_C(X,Y)$ would be close compared to $I_A(X,Y)$.

Using mutual information, we assessed the importance of hydropathy, disulfide bonding, and charge. For hydropathy, the twenty amino acids were classified as either hydrophobic or polar. For disulfide bonding, they were classified as Cys or other. For charge, they were classified as positive, negative, or neutral. To measure the cumulative importance of these attributes, they were classified as Cys, positive, negative, other polar, or other hydrophobic. Table 1 details these classifications.

## 2.4 Mutual information with finite sample sizes

There is one problem with this approach. One cannot directly compare the mutual information obtained with the 20-letter alphabet to that obtained with a smaller alphabet because mutual information is over-estimated with small sample sizes [19]. With larger alphabets, this over-estimation becomes more pronounced.

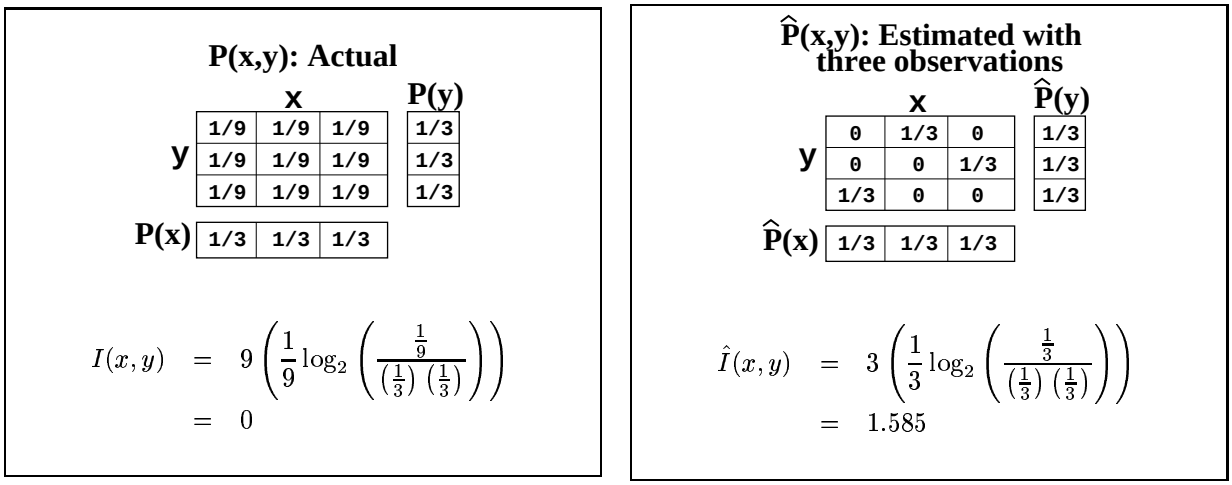


Figure I: This figure illustrates why mutual information is over-estimated with small sample sizes. At the left is the actual probability distribution of random variables $X$ and $Y$, with a mutual information of zero. At the right is the probability distribution computed with three observations, yielding an observed mutual information of 1.585 bits.

Consider the example in Figure I. Here, we assume we are given a random sample $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ of $n$ independent observations from the joint probability distribution of $X$ and $Y$. We let $\hat{P}(x, y)$ be the fraction of times the pair $(x, y)$ occurs in $D$, $\hat{P}(x)$ the fraction of times $x$ occurs in $x_1, ..., x_n$, and $\hat{P}(y)$ the fraction of times $y$ occurs in $y_1, ..., y_n$. At the left, the actual probability distribution for $X$ and $Y$ has mutual information of zero. At the right, a probability distribution estimated with $n = 3$ has observed mutual information of 1.585 bits. In the estimated joint probability distribution, sampling error will tend to suggest an artificial pairwise dependence. This artificial dependence will be more pronounced when the alphabet is larger and the probability distributions are more complex and the observations per parameter fewer in number. Therefore, the mutual information measured with the 20-letter alphabet will be more inflated than that measured with a 2-letter or 3-letter alphabet. This problem will occur to some extent with any finite-sized set of observations, or finite-sized database of reference protein structures.

Assume $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ are $n$ independent observations of two discrete random variables $X$ and $Y$. Assume that $\hat{P}(X)$ and $\hat{P}(Y)$ are their respective empirical probability distributions, and $\hat{P}(X, Y)$ is their empirical joint probability distribution. Consider two hypotheses:

Null Hypothesis $H_0$: $X$ and $Y$ are independent and distributed according to $\hat{P}(X,Y) = \hat{P}(X)\hat{P}(Y)$

Alternative Hypothesis $H_1$: $X$ and $Y$ are dependent and distributed according to $\hat{P}(X,Y) \neq \hat{P}(X)\hat{P}(Y)$

Let $E$ denote expectation over the distribution $\hat{P}(X,Y)$ in hypothesis $H_1$. A general test of the Hypothesis $H_1$ is the log ratio test, using the statistic $\hat{I}(X,Y) = \hat{I}(X,Y|D)$

$$\begin{aligned}
\hat{I}(X,Y) &= \frac{1}{n} E\left[\log \frac{P(D|H_1)}{P(D|H_0)}\right] \\
&= \frac{1}{n} E\left[\log \frac{\prod_{i=1}^{n} \hat{P}(x_i,y_i)}{\prod_{i=1}^{n} \hat{P}(x_i)\hat{P}(y_i)}\right] \\
&= \frac{1}{n} E\left[\sum_{i=1}^{n} \log \frac{\hat{P}(x_i,y_i)}{\hat{P}(x_i)\hat{P}(y_i)}\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} E\left[\log \frac{\hat{P}(x_i,y_i)}{\hat{P}(x_i)\hat{P}(y_i)}\right] \\
&= E\left[\log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)}\right] \\
&= \sum_{x,y} \hat{P}(x,y) \log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)}
\end{aligned}$$

where, as before, the last sum is over all possible pairs of values $x$ and $y$ of the variables $X$ and $Y$. Thus, the log ratio test statistic $\hat{I}(X,Y)$ is the observed mutual information between $X$ and $Y$, estimated according to the empirical joint probability distribution of $X$ and $Y$.

$n\hat{I}(X,Y)$ is known as the *g-statistic*, a statistic used in tests of independence [23] and similar in purpose to the more familiar though less effective [24] chi-square statistic $T(X,Y)$:

$$T(X,Y) = \sum_{x,y} \frac{(n\hat{P}(x,y) - n\hat{P}(x)\hat{P}(y))^2}{n\hat{P}(x)\hat{P}(y)} = n \sum_{x,y} \frac{(\hat{P}(x,y) - \hat{P}(x)\hat{P}(y))^2}{\hat{P}(x)\hat{P}(y)}$$

Theory shows that for a sufficiently large sample size, both statistics $n\hat{I}(X,Y)$ and $T(X,Y)$ have approximately the same distribution: approximately chi-square with $(M-1)^2$ degrees of freedom, where $M$ is the size of the alphabet [24, 25]. For large sample sizes, if $n\hat{I}(X,Y)$ (or respectively $T(X,Y)$) exceeds the 99th percentile of the chi-square distribution with $(M-1)^2$ degrees of freedom, one can reject the null hypothesis of independence between $X$ and $Y$ with 99% confidence. In this case, the difference between $n\hat{I}(X,Y)$ (or

$T(X,Y)$) and the chi-square distribution yields an estimate of the dependence between $X$ and $Y$.

Both $n\hat{I}(X,Y)$ and $T(X,Y)$ are ineffective for smaller sample sizes, because they do not yield a good approximation of the chi-squared distribution. More specifically, both statistics are biased in that their expected values, when estimated from a finite sample size, can differ substantially from their actual values. This can be demonstrated for $T(X,Y)$ as it was for $\hat{I}(X,Y)$ in Figure I. A rule of thumb is that there should be five or more observations of every possible pair of values $(x,y)$ for the approximation to be valid [26]. Unfortunately, we did not have this much data in all cases.

To solve this problem, we used an alternate approach aimed at smaller sample sizes. Rather than making a chi-squared approximation to $n\hat{I}(X,Y)$, we estimated the distribution of $\hat{I}(X,Y)$ empirically, using random independent permutations of $X$ and $Y$, as follows. Suppose the observed data is $D = \{(x_1,y_1),...,(x_n,y_n)\}$, and consider a permutation $\sigma$ of the numbers $1,...,n$, which reorders the observations of $Y$ with respect to $X$. Under the null hypothesis that $X$ and $Y$ are independent, for any permutation $\sigma$, the permuted data $\sigma(D) = \{(x_1,y_{\sigma(1)}),...,(x_n,y_{\sigma(n)})\}$ would have the same probability as $D$. We calculated an empirical distribution of $\hat{I}(X,Y)$ using the observed data set $D$ as follows. Let $\sigma_1,...,\sigma_n$ be $N$ random permutations of the numbers $1,...,n$. Then, for any value $v$, we estimated

$$\hat{P}(\hat{I}(X,Y) \geq v) \approx \frac{\left|\{i : 1 \leq i \leq N \text{ and } \hat{I}(X,Y|\sigma_i(D)) > v\}\right|}{N}$$

In other words, the probability that $\hat{I}(X,Y)$ exceeds $v$ is estimated as the fraction of N permutations for which this happens. We used this empirical distribution in place of the chi-squared distribution in our analysis. In particular, for a given dataset $D$, if $v = \hat{I}(X,Y|D)$ and $\hat{P}(\hat{I}(X,Y) \geq v) \leq 0.01$, then we rejected the null hypothesis that $X$ and $Y$ are independent with a confidence of 99%. In other words, if $\hat{I}(X,Y|D) \geq \hat{I}(X,Y|\sigma_i(D))$ for at least 99% of the random permutations $\sigma_i, 1 \leq i \leq n$, then we rejected the null hypothesis with a confidence of 99%. Similar random permutation analysis is seen in genotyping, where it is used to analyze haplotype frequencies [27].

The number of random permutations $N$ needed for this analysis was determined empirically, as follows. In informal terms, there are enough samples to yield a statistically-viable result when the observed results are

stable, and adding more observations does not change the results. We observed that 1,000 samples yielded a stable estimate of $\hat{I}(X,Y)$, so we chose a slightly larger $N$ of 5,000 in a spirit of caution.

Define the *independent information* $\hat{I}_I(X,Y)$ as the expected value of $\hat{I}(X,Y|\sigma(D))$, or the average mutual information observed in the permuted observation sets:

$$\hat{I}_I(X,Y) = \hat{I}_I(X,Y|\sigma(D)) = \frac{1}{N}\sum_{i=1}^{N}\hat{I}(X,Y|\sigma_i(D))$$

Here, $\hat{I}_I(X,Y)$ represents the bias in the empirical estimation of $\hat{I}(X,Y)$.

In cases when we rejected the null hypothesis, we measured the dependence of $X$ and $Y$ according to the *excess information* $\hat{I}_E(X,Y)$:

$$\hat{I}_E(X,Y) = \hat{I}_E(X,Y|D) = \hat{I}(X,Y|D) - \hat{I}_I(X,Y|D)$$

$\hat{I}_E(X,Y)$ describes the amount of information between $X$ and $Y$ in excess of the sampling bias.

## 2.5 Conditioning mutual information on burial

Conventional wisdom states that much of the information in amino-acid contact potentials stems from hydropathy, and the burial or exposure of hydrophobic residues. To quantify this effect, we factored out the burial/exposure signal by conditioning mutual information on burial, as follows:

1. According to a specific solvent exposure threshold, we classified each pair as either buried ($B$) or exposed ($E$) according to the more-buried residue. We empirically determined the probability of burial and exposure, $P(\mathrm{B})$ and $P(\mathrm{E})$, according to the proportion of pairs classified as buried.

2. For the buried pairs only, we estimated two quantities: the excess mutual information for the buried set, $I_E(A_1,A_2|\mathrm{B})$; and our confidence in rejecting the null hypothesis $H_0$, that that $A_1$ and $A_2$ are independent in buried contacts.

3. We repeated Step 2 for the exposed pairs, estimating $I_E(A_1,A_2|\mathrm{E})$ and an analogous confidence level.

4. We estimated the excess information conditioned on burial according to the results of the previous steps, yielding $I_E(A_1, A_2|\mathrm{C}) = I_E(A_1, A_2|\mathrm{B})P(\mathrm{B}) + I_E(A_1, A_2|\mathrm{E})P(\mathrm{E})$.

The next question was what solvent exposure value to use as a threshold. We tested all possible threshold values, graphing conditioned excess mutual information as a function of solvent exposure threshold. The lowest point on this graph represents the exposure/burial threshold at which hydropathy explains the greatest amount of the pairwise information. This amount is the difference between the unconditioned and conditioned excess information.

# 3 Results

## 3.1 Information content of amino acid contacts

Our first question was how much information is contained in pairwise amino acid contact potentials. To assess this, we grouped the amino acid contacts according to amino acid type and measured their observed mutual information $\hat{I}(A_1, A_2)$. As described in Section 2 we estimated their independent mutual information $I_I(A_1, A_2)$, the amount of mutual information we would expect from sampling error if that was the only source of observed mutual information, and their excess mutual information $I_E(A_1, A_2)$, the amount of mutual information observed beyond independent information.

Table 2 shows this mutual information. While the observed mutual information was approximately 0.06 bits, about one-fifth of that was probably an artifact of sampling error. The excess mutual information was approximately 0.04 bits. This is modest. At approximately 1.25 contacts per residue, an average of 0.04 bits of information per contact corresponds to 0.05 bits per residue. For contrast, one sequence of close homology yields approximately three bits of information per residue [28]. However, the amino acid contact signal is almost certainly real. Following the steps in Section 2, we observed that in at least 4995 of 5000 trials, the observed mutual information $\hat{I}(X, Y|D)$ in the original dataset $D$ was greater than that in the permuted dataset $\sigma(D)$, $\hat{I}(X, Y|\sigma(D))$. Thus, we rejected the null hypothesis $H_0$ that the two amino acids are independent with a confidence of at least 99%.

| Alphabet | Mutual Information | | |
|---|---|---|---|
| | Observed | Independent | Excess |
| Amino Acid Type (V) | 0.0605 | 0.0205 | 0.0400 |

Table 2: Shown is the amount of information found in pairwise contacts, estimated according to amino acid type. While observed mutual information is the simplest measure, it is affected by sampling error; Independent information describes the amount likely to result from sampling error. Excess information describes how much information was observed beyond what could be attributed to sampling error, and is the more reliable indicator of how much information is truly present.

### 3.2 Quantifying the importance of specific contact classes

Given the mutual information in amino acid contact potentials, our next question concerned where that information comes from. The likelihood that two residues interact is influenced by biochemical properties including hydropathy, charge, and disulfide bonding. To assess their importance, we grouped the amino acid contacts into corresponding contact alphabets as shown in Table 1 : Cys and Other (I); Positive, Negative, and Neutral (II); and Hydrophobic and Polar (III). We assessed their cumulative effect by grouping the residues according to their combination: Cys, Positive, Negative, Other Polar, and Other Hydrophobic (IV). For each alphabet, we followed the steps outlined in Section 2 to measure the observed, independent, and excess mutual information, and to see if we could reject the null hypothesis $H_0$. Table 3 shows the mutual information of each alphabet, and compares it to the 20-element alphabet of amino acid types (V).

In terms of observed mutual information, the amino acid alphabet (V) appeared to convey far more information than any other. However, this was the most complex alphabet, with twenty values for each row and column term, and the most affected by sampling error. Looking at excess mutual information, which was less sensitive to this error, the gap between the amino acid alphabet and the others shrinks considerably.

Half of the 0.04 bits of information in the 20-element amino acid alphabet (V) was conveyed in the 2-element HP alphabet (III). Cys and Other (I) accounted for 0.005 bits of information, or approximately one-eighth of the information of the amino acid alphabet. While this number might seem small, one should note that less than one-twentieth of the pairs in the dataset contained a Cys residue. So, while Cys-Cys pairs are infrequent, they represented a strong signal when they were present. Contacts between charged residues accounted for approximately one-fourth of the amino acid signal. For all alphabets, the observed information $I(X, Y|D)$ exceeded the permuted information $I(X, Y|\sigma(D))$ in more than 99% of all permutations $\sigma(D)$.

| Alphabet | Contents |
|---|---|
| I | Cys and Other |
| II | Positive, Negative, Other |
| III | Hydrophobic and Polar |
| IV | Cys, Positive, Negative, Other Polar, Other Hydrophobic |
| V | The standard alphabet of amino acid types |

| | | Mutual Information (bits) | | |
|---|---|---|---|---|
| Alphabet | Size | Observed | Expected | Excess |
| I | 2x2 | 0.0055 | 0.0001 | 0.0054 |
| II | 3x3 | 0.0105 | 0.0002 | 0.0103 |
| III | 2x2 | 0.0197 | $5.6\times10^{-5}$ | 0.0197 |
| IV | 5x5 | 0.0306 | 0.0010 | 0.0296 |
| V | 20x20 | 0.0605 | 0.0205 | 0.0400 |

Table 3: This table estimates the importance of specific biochemical properties that influence contact likelihood: hydropathy, charge, and disulfide bonding. Alphabet IV estimates their cumulative importance.

Thus, in all alphabets, we rejected the null hypothesis $H_0$ with a confidence of 99%.

Alphabet IV in Table 3 illustrated the cumulative effect of hydropathy, charge, and disulfide bonding on contact information. Altogether, three-quarters of the amino acid contact signal was described by the combined effect of hydropathy, charge, and disulfide bonding. Note that due to overlapping information, these contributions are not additive. For example, the physiochemical attraction between two polar residues of opposite charge is one component of both the charge signal and hydropathic signal.

### 3.3 Burial of hydrophobic residues

To estimate the importance of burial in pairwise contacts, we conditioned excess mutual information on burial as described in Section 2.5. Figure II depicts excess mutual information conditioned on burial for the five alphabets and for solvent exposure thresholds ranging from zero (fully buried) to 100 (fully exposed).

When conditioned excess mutual information on burial, most alphabets showed a drop in excess mutual information at almost every threshold. Only Cys and Other (I) appeared not to be affected, reflecting that disulfide bonding is not dependent on burial. For contrast, the amino acid alphabet (V) showed a substantial drop, reaching its lowest point at a solvent exposure threshold of 24. The HP alphabet (III) showed a drop of similar magnitude, reflecting the major importance of burial in the hydropathic signal. The charge alphabet

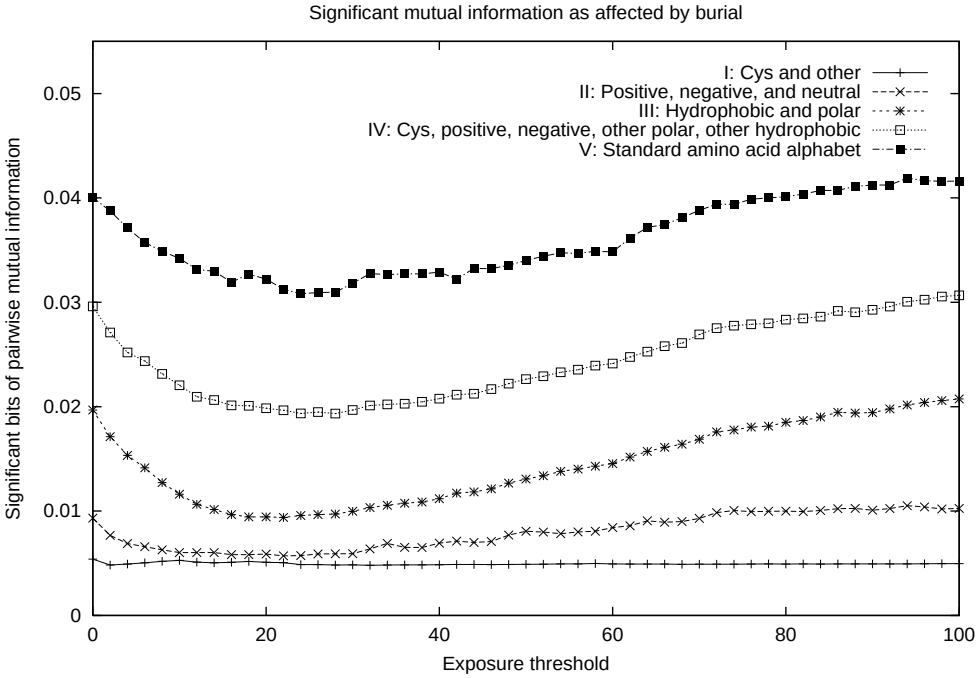


Figure II: Excess mutual information conditioned on burial and exposure, as a function of the solvent exposure threshold. Each point on this graph was obtained by dividing the contact pairs into buried and exposed subsets according to a solvent exposure threshold, estimating the excess mutual information for each subset, and combining the two measures according to the fraction or probability of exposure and burial, as defined in Section 2.5.

| Alphabet | Excess mutual information (bits) | | |
|---|---|---|---|
| | Unconditioned | Conditioned | Difference |
| I | 0.0054 | 0.0049 | 0.0006 |
| II | 0.0103 | 0.0057 | 0.0046 |
| III | 0.0197 | 0.0096 | 0.0101 |
| IV | 0.0296 | 0.0194 | 0.0102 |
| V | 0.0400 | 0.0308 | 0.0082 |

Table 4: This table compares the unconditioned excess mutual information to that conditioned on burial with a solvent exposure threshold of 24. At this threshold, conditioned mutual information was at its lowest, as shown in Figure II, and roughly 60% of all pairs were classified as buried.

(II) dropped less than most of the others. This reflects the lesser importance of charge in exposed positions, where the residues can interact with other molecules. In all cases, a solvent exposure threshold of 24 yielded the most marked decrease in excess mutual information. At this threshold, approximately 60% of the contacts were classified as buried. Table 4 shows the excess mutual information conditioned on this threshold.

Table 4 shows that the HP (III), cumulative (IV), and amino acid (V) alphabets all lost approximately 0.01 bits of information when conditioned on burial. The charge alphabet (II) lost approximately half of its signal, while the Cys-Other alphabet (I) showed little change.

One might ask why the HP alphabet (III) did not lose more information when conditioned on burial. There are two reasons. First, we used a simple binary classification of hydrophobic and polar, and of exposed and buried. A more detailed classification might show a greater change. Second, as observed previously [12], not all hydrophobic residues can be buried or polar residues exposed due to limitations from chain connectivity, excluded volume, and the dominance of the H-H contacts. Thus, the effect of burial on the HP signal is limited by structural constraints.

Finally, Table 4 shows a gap of 0.01 bits between the amino acid (V) and cumulative (IV) alphabets. This implies that pairwise amino acid contacts contained 0.01 bits of information that were not explained by hydropathy, charge, disulfide bonding, or burial. In summary, the familiar amino acid potentials convey 0.04 bits of information that could not be attributed to sampling bias. 0.01 bits can be attributed directly to burial. Of the remaining 0.03 bits, approximately 0.02 bits could be attributed to three amino acid properties: hydropathy, charge, and disulfide-bonding.

# 4 Conclusion

To improve contact potentials, and the many methods that use them, one must first understand how existing potentials work. We decomposed the signal present in pairwise amino acid contacts using mutual information. Mutual information is heavily influenced by sampling error; to avoid this problem, we have focused on *excess mutual information*, the amount of information beyond what we might expect from sampling error.

The amount of information found in amino acid contacts seems modest: 0.04 bits, less than one-twentieth of one bit. With an average of 1.25 contacts per residue, this corresponded to an average of 0.05 bits of information per residue. However, we observed this signal to be statistically significant. We rejected the null hypothesis $H_0$ that the two amino acids are independent with confidence of 99%. Of the 0.04 bits of information, three-fourths of this information can be explained by three amino acid properties: hydropathy, charge, and disulfide bonding.

Amino acid potentials have been seen to carry artifacts of the size [13], secondary structure composition [13], and stability [14] of the proteins in the reference database. We cannot yet speak to the utility of the additional 0.01 bits of information conveyed in amino acid potentials: perhaps they reflect such artifacts, or perhaps they convey genuine and valuable information. The effects of such artifacts will likely diminish as the number of known protein structures grow. At that time, we shall be able to assess better the utility of this apparent extra information.

## Acknowledgments

This paper represents the contributions of many people. The dataset was assembled by Ljubomir Buturovic and Raman Nambudripad. Lydia Gregoret provided keen insights and comments on the manuscript. The solvent exposure measure was developed by David Eisenberg, and we thank him for sharing his algorithm with us.

This work was supported in part by NSF grants CDA-9115268, IRI-9123692, and BIR-9408579; DOE grant 94-12-048216; ONR grant N00014-91-J-1162; NIH grant GM17129; NFS and GAANN graduate fellowships; and the UCSC Divisions of Natural Sciences and Engineering.

# References


[1] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–8, 1975.

[2] W. A. Koppensteiner and M.J. Sippl. Knowledge-based potentials–back to the roots. *Biochemistry*, 63(3):247–52, 1998.

[3] A. Torda. Perspectives in protein-fold recognition. *Current Opinion in Structural Biology*, 7(2):200–5, 1997.

[4] S. Vajda, M. Sippl, and J. Novotny. Empirical potentials and functions for protein folding and binding. *Current Opinion in Structural Biology*, 7(2):222–8, 1997.

[5] S. Miyazawa and R. Jernigan. Residue-residue potentials with a favorable contact pair term and unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256(3):623–44, 1996.

[6] M. R. Betancourt and D. Thirumalai. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Science*, 8(2):361–9, 1999.

[7] S. H. Bryant and C. E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins: Structure, Function, and Genetics*, 16(1):92–112, May 1993.

[8] G. Casari and M. J. Sippl. Structure-derived hydrophobic potential, hydrophobic potential derived from x-ray structures of globular proteins is able to identify native folds. *Journal of Molecular Biology*, 224(3):725–32, 1992.

[9] H. Li, C. Tang, and N. S. Wingreen. Nature of driving force for protein folding: a result from analyzing the statistical potential. *Physical Review Letters*, 79(4):765–8, 1997.

[10] J. P. Kocher, M. J. Rooman, and S. J. Wodak. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *Journal of Molecular Biology*, 235(5):1598–613, 1994.

[11] B. H. Park, E. S. Huang, and M. Levitt. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *Journal of Molecular Biology*, 266(4):831–46, 1997.

[12] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: how accurate are they? *Journal of Molecular Biology*, 257(2):457–69, March 1996.

[13] E. Furuichi and P. Koehl. Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins: Structure, Function, and Genetics*, 31(2):139–49, 1998.

[14] L. Zhang and J. Skolnick. How do potentials derived from structural databases relate to "true" potentials? *Protein Science*, 7(1):112–22, 1998.

[15] L. A. Mirny and E. I. Shakhnovich. Protein structure prediction by threading. why it works and why it does not. *Journal of Molecular Biology*, 283(2):507–26, 1998.

[16] M. Vendruscolo and E. Domany. Pairwise contact potentials are unsuitable for protein folding. *Journal of Chemical Physics*, 109(24):11101–8, 1998.

[17] S. R. Sunyaev, F. Eisenhaber, P. Argos, E. N. Kuznetsov, and V. G. Tumanyan. Are knowledge-based potentials derived from protein structure sets discriminative with respect to amino acid types? *Proteins: Structure, Function, and Genetics*, 31(3):225–46, 1998.

[18] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley and Sons, first edition, 1991.

[19] D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841–54, December 1995.

[20] J. U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.

[21] R. Lathrop and T. Smith. Global optimum protein threading with gapped alignment and empirical pair score functions. *Journal of Molecular Biology*, 255(4):641–65, 1996.

[22] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(2):412–25, 2000.

[23] K. Williams. The failure of pearson's goodness of fit statistic. *Statistician*, 25:49, 1976.

[24] J. Rice. *Mathematical statistics and data analysis*, pages 310–12. Duxbury Press, second edition, 1995. Theorem A.

[25] M Schervish. *Theory of Statistics*, page 459. Springer-Verlag, 1995.

[26] J. Rice. *Mathematical statistics and data analysis*. Duxbury Press, second edition, 1995.

[27] R. M. Single, D. Meyer, J. A. Hollenbach, M. P. Nelson, J. A. Noble, H. A. Erlich, and G. Thomson. Haplotype frequency estimation in patient populations: The effect of departures from hardy-weinberg proportions and collapsing over a locus in the hla region. *Genetic Epidemiology*, 22(2):186–95, 2002.

[28] Kevin Karplus. Regularizers for estimating distributions of amino acids from small samples. In *Proceedings, 3rd International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA, July 1995. AAAI/MIT Press.