Classification:
Biophysics and Computational Biology

Title:
Coevolutive, Evolutive and Stochastic Information in Protein-Protein Interactions

Author Affiliation:
Miguel Andrade‡, Camila Pontes‡ and Werner Treptow*
Laboratório de Biologia Teórica e Computacional (LBTC), Universidade de Brasília DF, Brasil
Supporting Information Placeholder

Corresponding Author:
Werner Treptow
Laboratório de Biologia Teórica e Computacional (LBTC), Universidade de Brasília DF, Brazil
+55 61 3107-3098
treptow@unb.br

Keywords:
Coevolution; Mutual Information; Protein-Protein Interaction

Author contributions. WT designed research; MA and CP performed research; MA, CP and WT analyzed data; WT wrote original draft; WT, MA and CP reviewed and edited. MA and CP contributed equally to this work.
ABSTRACT
Here, we investigate the contributions of coevolutive, evolutive and stochastic information in determining protein-protein interactions (PPIs) based on primary sequences of two interacting protein families \(A\) and \(B\). Specifically, under the assumption that coevolutive information is imprinted on the interacting amino acids of two proteins in contrast to other (evolutive and stochastic) sources spread over their sequences, we dissect those contributions in terms of compensatory mutations at physically-coupled and uncoupled amino acids of \(A\) and \(B\). We find that physically-coupled amino-acids at short range distances store the largest per-contact mutual information content, with a significant fraction of that content resulting from coevolutive sources alone. The information stored in coupled amino acids is shown further to discriminate multi-sequence alignments (MSAs) with the largest expectation fraction of PPI matches – a conclusion that holds against various definitions of intermolecular contacts and binding modes. When compared to the informational content resulting from evolution at long-range interactions, the mutual information in physically-coupled amino-acids is the strongest signal to distinguish PPIs derived from cospeciation and likely, the unique indication in case of molecular coevolution in independent genomes as the evolutive information must vanish for uncorrelated proteins.

SIGNIFICANCE
The problem of predicting protein-protein interactions (PPIs) based on multi-sequence alignments (MSAs) appears not completely resolved to date. In previous studies, one or more sources of information were taken into account not clarifying the isolated contributions of coevolutive, evolutive and stochastic information in resolving the problem. By benefiting from data sets made available in the sequence- and structure-rich era, we revisit the field to show that physically-coupled amino-acids of proteins store the largest (per contact) information content to discriminate MSAs with the largest expectation fraction of PPI matches – a result that should guide new developments in the field, aiming at characterizing protein interactions in general.
INTRODUCTION

While being selected to be thermodynamically stable and kinetically accessible in a particular fold (1, 2), interacting proteins A and B coevolve to maintain their bound free-energy stability against a vast repertoire of non-specific partners and interaction modes. Protein coevolution, in the form of a time-dependent molecular process, then translates itself into a series of primary-sequence variants of A and B encoding coordinated compensatory mutations (3) and, therefore, specific protein-protein interactions (PPIs) derived from this stability-driven process (4). As a ubiquitous process in molecular biology, coevolution thus apply to protein interologs, either paralogous or orthologous, under cospeciation or in independent genomes.

Thanks to extensive investigations in the past following ingenious approaches based on the correlation of phylogenetic trees (5–7) and profiles (8), gene colocalization (9) and fusions (10), maximum coevolutionary interdependencies (11) and correlated mutations (12, 13), the problem of predicting PPIs based on multi-sequence alignments (MSAs) appears to date resolvable, at least for small sets of paralogous sequences - recent improvements (14–18) resulting from PPI prediction allied to modern coevolutionary approaches (19–23) to identify interacting amino acids across protein interfaces.

In these previous studies, however, the information was taken into account as a whole, and it was not clarified, as discussed in recent reviews (4, 24), the isolated contributions of coevolutionary, evolutive and stochastic information in resolving the problem.

Here, by benefiting from much larger datasets made available in the sequence- and structure-rich era, we revisit the field by quantifying the amount of information that protein A stores about protein B stemming from each of these sources and, more importantly, their effective contributions in determining PPIs based on MSAs (Scheme 1). Specifically, under the assumption that the coevolutionary information is imprinted on the interacting amino acids of protein interologs in contrast to other (evolutive and stochastic) sources spread over their sequences, we want that information to be dissected in terms of compensatory mutations at physically-coupled and uncoupled amino acids of A and B. Anticipating our findings, we show that physically-coupled amino-acids store the largest per-contact mutual information (MI) content to discriminate concatenated MSAs with the largest expectation fraction of PPI matches – a conclusion that holds against various definitions of intermolecular protein contacts and binding modes. A significant fraction of that information results from coevolutive sources alone. It is important to acknowledge that, even though our findings are derived for a set of interologs under cospeciation, they can be generalized to cases of non-cospeciating interologs given that the underlying thermodynamical principles must be the same for all cases.

Scheme 1. Structural contacts mapped into M-long multi-sequence alignment (MSA) of protein interologs A and B. A set of pairwise protein-protein interactions is defined by associating each sequence j in MSA B to a sequence k in MSA A in one unique arrangement, \( \{ l(k) \} \), determined by the coevolution process \( \zeta \) to which these protein families were subjected.
THEORY AND METHODS
Decomposition of Mutual Information. In detail, consider two proteins $A$ and $B$ that interact via formation of $i = 1, \ldots, N$ amino-acid contacts at the molecular level. Proteins $A$ and $B$ are assumed to coevolve throughout $M!$ distinct processes $z$ described by the stochastic variable $Z$ with an uniform probability mass function $\rho(z)$, $\forall z \in \{1, \ldots, M!\}$. Given any specific process $z$, their interacting amino-acid sequences are respectively described by two $N$-length blocks of discrete stochastic variables $X^n = (x_1, \ldots, x_n)$ and $Y^n = (y_1, \ldots, y_n)$ with probability mass functions $(\rho(x^n), \rho(y^n), \rho(x^n, y^n|z))$ such that,

$$\begin{align*}
\rho(x^n) &= \sum_z \rho(x^n, y^n|z) \\
\rho(y^n) &= \sum_z \rho(x^n, y^n|z)
\end{align*}$$

(1)

and

$$\sum_{x^n, y^n} \rho(x^n, y^n|z) = 1$$

(2)

for every joint sequence $(x^n, y^n)_{x^n}$ defined in the alphabet $\chi$ of size $|\chi|$. Under these considerations, the amount of information that protein $A$ stores about protein $B$ is given by the mutual information $I(X^n; Y^n|z)$ between $X^n$ and $Y^n$ conditional to process $z$ (25). As made explicit in eq. [1], we are particularly interested in quantifying $I(X^n; Y^n|z)$ for the situation in which marginals of the $N$-block variables $(\rho(x^n), \rho(y^n))$ are assumed to be independent of process $z$ meaning that, for a fixed sequence composition of proteins $A$ and $B$ only their joint distribution depends on the process. Furthermore, by assuming $N$-independent contacts, we want that information to be quantified for the least-constrained model $\rho^*(x^n, y^n|z)$ that maximizes the conditional joint entropy between $A$ and $B$ - that condition ensures the mutual information to be written exactly, in terms of the individual contributions of contacts $i$.

For the least-constrained distribution $(\rho^*(x^n, y^n|z))$, the conditional mutual information

$$I(X^n; Y^n|z) = H(X^n) + H(Y^n) - H(X^n, Y^n|z)$$

(3)

writes in terms of the Shannon’s information entropies

$$\begin{align*}
H(X^n) &= -\sum_{x^n} \rho^*(x^n) \ln \rho^*(x^n) \\
H(Y^n) &= -\sum_{y^n} \rho^*(y^n) \ln \rho^*(y^n) \\
H(X^n, Y^n|z) &= -\sum_{x^n, y^n} \rho^*(x^n, y^n|z) \ln \rho^*(x^n, y^n|z)
\end{align*}$$

(4)

associated with the conditional joint distribution $(\rho^*(x^n, y^n|z))$ and the derived marginals $(\rho^*(x^n), \rho^*(y^n))$ of the $N$-block variables. From its entropy-maximization property, the critical distribution $\rho^*(x^n, y^n|z)$ factorizes into the conditional two-site marginal of every contact $i$

$$\rho^*(x^n, y^n|z) = \prod_{i=1}^N \rho^*(x_i, y_i|z)$$

(5)

then allowing eq.[4] to be written extensively, in terms of the individual entropic contributions

$$\begin{align*}
H(X^n) &= \sum_{x^n} H(X^n|x^n) \\
H(Y^n) &= \sum_{y^n} H(Y^n|y^n) \\
H(X^n, Y^n|z) &= \sum_{x^n, y^n} H(X^n, Y^n|x^n, y^n, z)
\end{align*}$$

(6)

such that,

$$I(X^n; Y^n|z) = \sum_{i=1}^N I(X_i; Y_i|z)$$

(7)

(cf. SI for details). In eq. [7], the conditional mutual information achieves its lower bound of zero if $X^n$ and $Y^n$ are conditionally independent given $z$ i.e., $\rho(x^n, y^n|z) = \rho^*(x^n) \times \rho^*(y^n)$. For the case of per-
fectly correlated variables \( \rho'(x^a,y^a)\), the conditional mutual information is bound to a maximum which cannot exceed the entropy of either block variables \( H(x^a) \) and \( H(y^a) \).

Given a known set of protein amino-acid contacts and their underlying primary sequence distributions defining the stochastic variables \( x^a \) and \( y^a \), eq. [7] thus establishes the formal dependence of their mutual information with any given process \( z \). Because “contacts” can be defined for a variety of cutoff distances \( r_c \), eq. [7] is particularly useful to dissect mutual information in terms of physically-coupled and uncoupled protein amino acids. In the following, we explore eq. [7] in that purpose by obtaining the two-site probabilities in eq. [5]

\[
p'(x_i,y|z) = \sum_{x_{i+1},...,x_n,y_{i+1},...,y_n} \delta_{y_{i+1},...,y_n} \delta_{x_{i+1},...,x_n} p'(x_{i+1},...,x_n,y_{i+1},...,y_n|z) \equiv f_{x_i,y_i}
\]

from the observed frequencies \( f = \{ f_{x_i,y_i}\} \) in the multiple-sequence alignment \( \{ x_i^a, y_i^a|z \}_M \) where the \( N \)-length amino-acid block \( i \) of protein \( B \) is joint to block \( k \) of protein \( A \) in one unique arrangement \( \{ l(k|z)\}_M \) for \( 1 \leq k \leq M \) (cf. Scheme 1 and Computational Methods).

**Computational Methods.** Table-S1 details the interacting protein systems considered in the study. For each system under investigation, amino-acid contacts defining the discrete stochastic variables \( x^a \) and \( y^a \) including physically coupled amino acids at short-range cut-off distances (\( r_c \leq 8.0 \AA \)) and physically uncoupled amino-acids at long-range cut-off distances (\( r_c > 8.0 \AA \)) were identified from the x-ray crystal structure of the bound state of proteins \( A \) and \( B \). The reference (native) multi-sequence alignment \( \{ x_i^a, y_i^a|z \}_M \) of the joint amino-acid blocks associated to \( x^a \) and \( y^a \) was reconstructed from annotated primary-sequence alignments published by Baker and coworkers (22), containing \( M \) paired sequences with known PPI interactions and defined in the alphabet of 20 amino acids plus the gap symbol \( ( |z|=21 \). Scrambled MSA models were generated by randomizing the pattern \( \{ l(k|z)\}_M \) in which block \( l \) is joint to block \( k \) in the reference alignment.

For any given MSA model, two-site probabilities \( \rho'(x_i,y|z) = f_{x_i,y_i} \) were derived from the observable frequencies \( f_{x_i,y_i} \) regularized by a pseudocount effective fraction \( \lambda' \) in case of insufficient data availability as devised by Morcos and coauthors (19). More specifically, two-site frequencies were calculated according to

\[
f_{x_i,y_i} = \frac{\lambda'_i}{|z|} + (1-\lambda') \frac{1}{M'_i} \sum_{m=1}^{M_i} \frac{1}{n'_i} \delta_{x_i^a,y_i^a,m^a} \equiv f_{x_i,y_i}
\]

where, \( n'_i = |\{ m^a | 1 \leq m^a \leq M_i, Hamming Distance(m,m^a) \geq \delta h \}| \) is the number of similar sequences \( m^a \) within a certain Hamming distance \( \delta h \) of sequence \( m \) and \( M'_i = \sum_{n=1}^{M_i} (n'_i)^{-1} \) is the effective number of distinguishable primary sequences at that distance threshold - the Kronecker delta \( \delta_{x_i^a,y_i^a,m^a} \) ensures counting of \( (x_i,y_i) \) occurrences only. In eq. [9], two-site frequencies converge to raw occurrences in the sequence alignment for \( \lambda'=0 \) or approach the uniform distribution \( \frac{1}{|z|} \) for \( \lambda'=1 \); eq. [9] is identical to the equation devised by Morcos and coauthors (19) by rewriting \( \lambda' = \frac{\lambda}{\lambda + M'_i} \). Here, two-site probabilities \( \rho'(x_i,y|z) = f_{x_i,y_i} \) were computed from eq. [9] after unbiasing the reference MSA by weighting down primary sequences with amino-acid identity equal to 100%. An effective number of primary sequences \( M'_i = M \) (cf. Table-S1) was retained for analysis and a pseudocount fraction of \( \lambda'=0.001 \) was used to regularize data without largely impacting observable frequencies. Single-site probabilities \( \{ \rho(x^a), \rho(y^a) \} \) were derived from \( \rho'(x_i,y|z) \) by marginalization via eq. [1].

The conditional mutual information in eq. [7] was computed from single- and joint-entropies according to eq. [3]. Given the fact that the maximum value of \( I(X_i;Y|z) \) is bound to the conditional joint entropy, eq. [7] was computed in practice as a per-contact entropy-weighted conditional mutual information (26), \( H(X_i,Y|z) I(X_i;Y|z) \), to avoid that contributions of contacts between highly variable sites are overestimated. Because \( H(X_i,Y|z) \) and \( I(X_i;Y|z) \) have units of nats, eq. [7] is dimensionless in the present form.
RESULTS AND DISCUSSION

Details of all protein systems under investigation are presented in Table-S1. Each system involves two families of protein interologs \( A \) and \( B \) with known PPIs derived from cospeciation in the same genome (27). We denote by \( \{x^*_i, y^*_i \}_{i=1}^N \) their reference concatenated MSA associated to the native process \( z^* \). For convenience, in the following, we present and discuss results obtained for a representative system \( A \) and \( B \) – the protein complex TusBCD (chains B and C of 2DIP) which is crucial for tRNA modification in \( Escherichia coli \). Similar results and conclusions hold for all other systems in Table-S1 as presented in supplementary figures S1 through S4 (cf. SI).

![Figure 1. Informational analysis of protein complex TusBCD, chains B and C.](https://example.com/figure1.png)

**Figure 1.** Informational analysis of protein complex TusBCD, chains \( B \) and \( C \). (A) Three-dimensional representation of stochastic variables \( X^* \) and \( Y^* \) as defined from physically coupled amino acids at short-range cutoff distances \( r_c \leq 8.0 \text{ Å} \) (turquoise) and physically uncoupled amino-acids at long-range cutoff distances \( r_c > 8.0 \text{ Å} \) (gray). (B) Conditional mutual information \( \langle I(X^*; Y^*|z^*) \rangle_{M-n} \) as a function of the number \( M-n \) of randomly paired proteins in the reference (native) MSA. \( \langle I(X^*; Y^*|z^*) \rangle_{M-n} \) are expectation values estimated from a generated ensemble of 500 MSA models. (C) Mutual information gap \( N^{-1} \Delta I_{M-n} \) between reference and 100 random models featuring \( M \) randomly paired sequences. (D) Per-contact mutual information gap \( N^{-1} \Delta I_{M-n} \). (E) Mutual information decomposition \( N^{-1} \Delta I_{M-n} \) according to eq. [11] and comparison with functional mutual information \( |MI_{p,t,s=8\text{ Å}}| \) and direct information \( DI_{r_c \leq 8\text{ Å}} \). In B, C, D and E error bars correspond to standard deviations.

**Decomposition of Mutual Information.** Fig. 1A shows the three-dimensional representation of variables \( X^* \) and \( Y^* \) embodying every possible amino-acid pairs along proteins \( A \) and \( B \) and their decomposition in terms of physically coupled amino acids at short-range cutoff distances \( r_c \leq 8.0 \text{ Å} \) and physically uncoupled amino-acids at long-range cutoff distances \( r_c > 8.0 \text{ Å} \). In Fig. 1B, the total mutual information \( I(X^*; Y^*|z^*) \) across every possible amino-acid pairs of \( A \) and \( B \) (coupled + uncoupled) amounts to 987.88 in the reference (native) MSA (process \( z^* \)). As estimated from a generated ensemble of scrambled MSA models, expectation values for the mutual information \( \langle I(X^*; Y^*|z^*) \rangle_{M-n} \) decreases significantly as decorrelation or the number \( M-n \) of mismatched proteins in the reference MSA increases, for \( 0 \leq n \leq M \). The result also holds at the level of individual protein contacts \( i \) as \( I(x_i; y_i|z^*) \) for the reference alignment is systematically larger than \( \langle I(x_i; y_i|z^*) \rangle_{M-n} \) for scrambled MSA models with \( M \) sequence mismatches (Fig. S1).
As a measure of correlation, it is not surprising that mutual information in the reference MSA is larger than that of scrambled alignments. Not expected, however, is the fact that correlation does not vanish at scrambled models meaning that part of the calculated mutual information results at random. Supporting that notion, low mutual information between amino acid positions of multiple sequence alignments are typically indicative of independently evolved proteins. Subtraction of that stochastic source from \( I(X^n;Y^n|z^n) \) as computed in the form of a mutual-information gap

\[
\Delta I_{M,n} \equiv I(X^n;Y^n|z^n) - I(X^n;Y^n|z^n)_{M \to n}
\]

between the reference model and models with \( M \) mismatched sequences then reveals the isolated nonstochastic contributions to the total correlation between proteins \( A \) and \( B \). Here, the information gap \( \Delta I_{M,n} \) amounts to \( \sim 440 \) for every possible amino-acid pairs along proteins \( A \) and \( B \).

Fig. 1C shows the individual contributions of physically coupled and uncoupled amino acids to the total mutual information gap, \( \Delta I_{M} = \Delta I_{M,c,\leq 8\Delta} + \Delta I_{M,r,\leq 8\Delta} \). As a direct consequence of the extensive property of eq. [7], \( \Delta I_{M,c,\leq 8\Delta} \) increases with \( r_c \) and, consequently, with the block length \( N \) of the corresponding stochastic variables. As such, the information imprinted at physically uncoupled amino acids accounts for most of the total mutual information gap (438.8132 \( \pm 4.5159 \)). When normalized by the block length or the number of contacts (Fig. 1D), \( N^{-1}\Delta I_{M,c,\leq 8\Delta} \) reveals a distinct dependence being larger for physically coupled amino acids than uncoupled ones (0.0653 \( \pm 0.0015 \) versus 0.0340 \( \pm 0.0004 \)). The information gap profile as a function of amino-acid pair distances shown in Fig. S2 makes sense of the result by showing few larger values of \( \Delta I_{M} \) at short amino-acid pair distances in contrast to many smaller ones at long distances.

Under the assumption that the coevolutive information is imprinted on the interacting amino acids of interologs in contrast to other (evolutive and stochastic) sources spread over their primary sequences, the difference between short- and long-range contributions provides us with per-contact estimates for the information content resulting from coevolution alone that is,

\[
N^{-1}\Delta I_{M,\leq 8\Delta} \equiv N^{-1}\Delta I_{M,c,\leq 8\Delta} - N^{-1}\Delta I_{M,r,\leq 8\Delta}
\]

where, \( N^{-1}\Delta I_{M,c,\leq 8\Delta} \) represents the per-contact mutual information resulting from evolution. As shown in Fig. 1E, \( N^{-1}\Delta I_{M,c,\leq 8\Delta} \) amounts to 0.0264 \( \pm 0.0014 \) which compares well to independent measures of coevolutionary information ie., functional mutual information (MI\(_{p,r,\leq 8\Delta} \)) (26) and direct information (DI\(_{s,\leq 8\Delta} \)) (19), 0.0340 \( \pm 0.0037 \) and 0.0202 \( \pm 0.0019 \). More specifically, MI\(_{p,r} \) is a metric MI\(\equiv MI−APC \) formulated by Dunn and coworkers (26) in which mutual information is subtracted from structural or functional relationships (APC) whereas, DI is based on the direct coupling analysis that removes all kinds of indirect correlations by following a global statistical approach (19). According to definition in eq. [11], we then conclude that \( \sim 40 \% \) of the information content stored in physically coupled amino acids of the protein complex TusBCD results from coevolutive sources alone.

**Degeneracy and Error Analysis of Short and Long-Range Correlations.** The present analysis reveals quantitative differences between short- and long-range correlations of proteins \( A \) and \( B \). Because \( N^{-1}\Delta I_{M,c,\leq 8\Delta} \) provides us with an unbiased (intensive) estimate for proper comparison of the information contents between coupled and uncoupled amino acids, in the following, we focus our attention on \( N^{-1}\Delta I_{M,c,\leq 8\Delta} \) to dissect their effective contributions in determining PPIs based on sequence alignments. Accordingly, let us define the total number \( \omega_{M,n} \) of native-like MSA models generated by randomization of \( M−n \) sequence pairs in the reference alignment

\[
\omega_{M,n} \equiv \sum_{n \subseteq X(r)} \omega_{M,n}
\]

in terms of rencontres numbers \( \omega_{M,n} \)

\[
\omega_{M,n} = \frac{M!}{m!} \sum_{q=0}^{M-n} \frac{(-1)^q}{q!}
\]
or permutations of the reference sequence set \( \{ l(k)|z^* \}_M \) with \( n \) fixed positions satisfying
\[
\sum_{k=0}^{M} \omega_{k,n} = M!
\]
(in combinatorics language). Here, \( S(r) \) denotes the set of fixed positions \( n \)
\[
S(r) = \{ n|0 \leq n \leq M, N^{-1} \Delta I_{M-n,r} \leq \delta I \}
\]
(14)
for which the normalized mutual information gap \( N^{-1} \Delta I_{M-n,r} \) is smaller than a certain resolution \( \delta I \) independently from the corresponding block length \( N \) or the number of amino-acid contacts. In simple terms, \( \omega \) informs us on the degeneracy or the number of scrambled MSA models with a similar amount of mutual information of that in the reference multi-sequence alignment.

As shown in Table-S2, \( \omega_{M,n} \) is an astronomically increasing function of \( M-n \), identical for any definition of \( X^N \) and \( Y^N \) derived from the same number \( M \) of aligned sequences. For instance, there is 164548102752 alignments for the protein complex TusBCD with \( M-n=5 \) scrambled sequence pairs. In contrast, \( \omega \) depends on the stochastic variables at various resolutions \( \delta I \) (Fig. 2A). The number of native-like MSA models is substantially smaller for definitions of \( X^N \) and \( Y^N \) embodying physically-coupled amino acids in consequence of the smaller number \( M-n \) of randomly paired sequences required to perturb \( N^{-1} \Delta I_{M-n,r} \) of a fixed change \( \delta I \) such that \( \omega \) accumulates less over MSA models satisfying \( N^{-1} \Delta I_{M-n,r} \leq \delta I \) (Fig. 2B).

---

**Figure 2.** Degeneracy and error analysis for \( X^N \) and \( Y^N \) involving interacting amino acids at short-range distances \( r_c \leq 8.0 \text{Å} \) (turquoise) and long-range distances \( r_c > 8.0 \text{Å} \) (gray). (A) Total number \( \omega_{M} \) of native-like MSA models at various resolutions \( \delta I \). (B) Per-contact gaps of mutual information \( N^{-1} \Delta I_{M-n,r} \) as a function of the number \( M-n \) of scrambled sequence pairs in the reference alignment. (C) Expectation values \( \langle \epsilon \rangle_{S} \) (eq. [15]) for the fraction of sequence matches across native-like MSA models at various resolutions \( \delta I \). Dashed lines highlight differences at \( \delta I \) values of 0.01 and 0.02.
The degeneracy of $M$-long native-like MSA models at a given resolution $\delta I$ depends on the cutoff distance $r_c$ defining $x^n$ and $y^n$ (Fig. 2A). That condition imposes distinct boundaries for the amount of PPIs amenable of resolution across definitions of the stochastic variables in terms of coupled and uncoupled amino acids. Indeed, the expectation value

$$\langle \epsilon \rangle_S = \sum_{n \in S} (M \sum_{n \in S} \omega_{M,n})^{-1} \omega_{M,n} \times n$$

for the fraction $M^{-1}n$ of primary sequence matches among native-like MSA models mapped by $S$ in eq. [14] decreases substantially with $\omega_S$ meaning that $\langle \epsilon \rangle_S$ is systematically larger for physically-coupled amino-acids at various resolutions $\delta I$ (Fig. 2C). For instance, the fraction of matches at $\delta I = 0.02$ is $\sim 20\%$ larger for coupled amino-acids than the same estimate for amino acids at long-range distances ($0.8333$ versus $0.6991$). Linear extrapolation in Fig. 2C along increased values of $\delta I$ suggests even larger differences in the expectation fraction of PPI matches between short and long-range correlations of proteins $A$ and $B$.

Figure 3. Dependence with contact definition $r_c^*$ and docking decoys. (A) $N^{-1} \Delta I_{M,r_c}$ and $MI_{p,r_c}$ at various $r_c^*$. (B) $N^{-1} \Delta I_{M,r_c}$ (turquoise), $N^{-1} \Delta I_{M,r_c}$ (green), $MI_{p,r_c}$ (blue) at alternative interfaces generated by docking – only physically coupled amino acids as defined for $r_c \leq 8.0 \text{Å}$ were included in the calculations. Black bars represent the root-mean-square deviation (RMSD) between the native bound structure and docking decoys as generated by GRAMM-X (28). (C) Illustration of four docking decoys of chain B in the protein complex TusBCD (chain C is shown in gray).

Dependence with Contact Definition $r_c^*$ and Docking Decoys. So far, our results were determined by defining physically coupled amino acids at short-range cutoff distances ($r_c \leq r_c^*$) and physically uncoupled amino-acids at long-range cutoff distances ($r_c > r_c^*$), for $r_c^* = 8.0 \text{Å}$. Further analysis shows however a clear dependence of the per-contact mutual information gap $N^{-1} \Delta I_{M,r_c}$ of coupled amino acids with $r_c^*$ – which is not the case for uncoupled ones. As shown in Fig. 3A that distinction is due the coevolutive information stored at short-range distances which reaches a maximum at $r_c^* \approx 8.0 \text{Å}$ in contrast to evolutive sources uniformly spread over an entire range of $r_c^*$ values. Particularly interesting the result strongly support the assumption that coevolutive information is imprinted preferentially on physically-coupled amino acids of interologs in contrast to other (evolutive and stochastic) sources spread over their primary sequences - a conclusion further supported by calculations of the mutual information $MI_p$ subtracted from structural or functional relationships as a function of $r_c^*$.

At $r_c^* = 8.0 \text{Å}$, there is still a clear dependence of the per-contact mutual information gap $N^{-1} \Delta I_{M,r_c}$ with binding modes or docking decoys of proteins $A$ and $B$. Shown in Fig. 3B, the per-contact mutual information gap reaches a maximum at the bound configuration of $A$ and $B$ (RMSD = 0.0 Å), meaning that $N^{-1} \Delta I_{M,r_c}$ embodies coevolutive pressures in the native amino acids contacts beyond their accessibility at the molecular surface of proteins. The conclusion is further supported by
noticing that the isolated coevolutive content $N^{-1}\Delta I_{M,r,\sim \lambda}$ for the bound configuration of $A$ and $B$ or the associated mutual information $MI_p$ subtracted from structural-functional relationships are larger than the very same estimates for any docking decoys.

CONCLUDING REMARKS. Overall, molecular coevolution as the maintenance of the binding free-energy of interacting proteins leads their physically coupled amino-acids to store the largest per-contact mutual information $N^{-1}\Delta I_{M,r,\sim \lambda}$, at $r_c \approx 8.0 \text{ Å}$, with a significant fraction of the information resulting from coevolutive sources alone. In the present formulation, coupled amino acids are related to the smallest degeneracy $o_\lambda$ of native-like MSA models and, therefore, to the largest expectation fraction of PPI matches $\langle \epsilon \rangle_\lambda$ across such models. These findings hold against any other definition of protein contacts, either across a variety of limitrophe distances $r_c$ discriminating coupled and uncoupled amino acids or alternative binding interfaces in docking decoys. Although presented for the protein complex Tus-BCD, results and discussion also extent to other protein systems as shown in supplementary figures S1 through S4 (cf. SI).

Advances in PPI prediction (14–18) are highly welcome in the contexts of paralog matching, host-pathogen PPI network prediction and interacting protein families prediction. Recent studies suggest strategies like maximizing the interfamily coevolutionary signal (14), iterative paralog matching based on sequence “energies” (15) and expectation–maximization (18), which have been capable of accurately matching paralogs for some study cases. Despite these advances, the problem of PPI prediction remains unsolved for sequence ensembles in general, especially for proteins that coevolve in independent genomes though likely resulting from the same free-energy constraints - examples are phage proteins and bacterial receptors, pathogen and host-cell proteins, neurotoxins and ion channels, to mention a few. Far from being trivial, PPI prediction from primary sequences alone is highly complex only solvable within a certain accuracy to date. In this regard, our findings reveal the largest amount of non-stochastic information $N^{-1}\Delta I_{M,r,\sim \lambda}$ available in primary sequences under coevolution to discriminate MSA models with the largest expectation fraction of PPIs matches $\langle \epsilon \rangle_\lambda$. Optimality of $N^{-1}\Delta I_{M,r,\sim \lambda}$ and $\langle \epsilon \rangle_\lambda$ might be therefore of practical relevance in search of more effective heuristics to resolve PPIs from arbitrary multi-sequence alignments. When compared to evolutive sources, that information is the strongest signal to characterize protein interactions derived from cospeciation and likely, the unique indication in case of coevolution without cospeciation as non-stochastic information of uncoupled amino acids must vanish in independent proteins.

We believe the results are of broad interest as the stability principles of protein systems under coevolution must be universal, either under cospeciation or in independent genomes. We therefore anticipate that decomposition of evolutive and coevolutive information imprinted in physically-coupled and uncoupled amino acids and evaluation of their potential utility in resolving MSA models in terms of degeneracy and fraction of PPI matches should guide new developments in the field, aiming at characterizing protein interactions in general.

ACKNOWLEDGEMENTS

Comments of Antônio Francisco P. de Araújo. Fernando Melo and Michael Klein on the manuscript are gratefully acknowledged. The research was supported in part by the Brazilian Agencies CNPq, CAPES and FAPDF under Grants 305008/2015-3, 23038.010052/2013-95 and 193.001.202/2016. WT thanks CAPES for doctoral fellowship to MA and CP.

REFERENCES
1. Garcia LG, Treptow WL, Pereira de Araújo AF (2001) Folding simulations of a threedimensional protein model with a nonspecific hydrophobic energy function. *Phys Rev E* 64(1):011912.

2. Treptow WL, Barbosa MAA, Garcia LG, Araújo AFP de (2002) Non-native interactions, effective contact order, and protein folding: A mutational investigation with the energetically frustrated hydrophobic model. *Proteins Struct Funct Bioinforma* 49(2):167–180.

3. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18(4):309–317.

4. de Juan D, Pazos F, Valencia A (2013) Emerging methods in protein co-evolution. *Nat Rev Genet* 14(4):249–261.

5. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) Co-evolution of proteins with their interaction partners. *J Mol Biol* 299(2):283–293.

6. Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* 14(9):609–614.

7. Gertz J, et al. (2003) Inferring protein interactions from phylogenetic distance matrices. *Bioinforma Oxf Engl* 19(16):2039–2045.

8. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci* 96(8):4285–4288.

9. Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23(9):324–328.

10. Marcotte CJV, Marcotte EM (2002) Predicting functional linkages from gene fusions with confidence. *Appl Bioinformatics* 1(2):93–100.

11. Tillier ERM, Biro L, Li G, Tillo D (2006) Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. *Proteins* 63(4):822–831.

12. Pazos F, Valencia A (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* 47(2):219–227.

13. Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* 4:165.

14. Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A (2016) Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc Natl Acad Sci* 113(43):12186–12191.

15. Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS (2016) Inferring interaction partners from protein sequences. *Proc Natl Acad Sci* 113(43):12180–12185.

16. Várnai C, Burkoff NS, Wild DL (2017) Improving protein-protein interaction prediction using evolutionary information from low-quality MSAs. *PLOS ONE* 12(2):e0169356.

17. Bitbol A-F (2018) Inferring interaction partners from protein sequences using mutual information. *PLOS Comput Biol* 14(11):e1006401.
18. Correa Marrero M, Immink RGH, de Ridder D, van Dijk ADJ. Improved inference of intermolecular contacts through protein–protein interaction prediction using coevolutionary analysis. *Bioinformatics*. doi:10.1093/bioinformatics/bty924.

19. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci* 108(49):E1293–E1301.

20. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.

21. Jeong C-S, Kim D (2012) Reliable and robust detection of coevolving protein residues. *Protein Eng Des Sel* 25(11):705–713.

22. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci*:201314045.

23. Burger L, Nimwegen E van (2010) Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. *PLOS Comput Biol* 6(1):e1000633.

24. Juan D, Pazos F, Valencia A (2008) Co-evolution and co-adaptation in protein networks. *FEBS Lett* 582(8):1225–1230.

25. MacKay DJC (2003) *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, UK; New York). 1 edition.

26. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24(3):333–340.

27. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* 3:e02030.

28. Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein–protein docking. *Nucleic Acids Res* 34(suppl_2):W310–W314.
Coevolutive, Evolutive and Stochastic Information in Protein-Protein Interactions
Miguel Andrade, Camila Pontes and Werner Tretow
Laboratório de Biologia Teórica e Computacional (LBTC), Universidade de Brasília DF, Brasil

Derivation of Main Text Equations
Consider two proteins \( A \) and \( B \) that interact via formation of \( i=1,\ldots,N \) independent amino-acid contacts at the molecular level. Proteins \( A \) and \( B \) are assumed to evolve throughout \( M \) distinct coevolution processes \( Z \) described by the stochastic variable \( Z \) with probability mass function \( \rho(z) \), \( \forall z \in \{1,\ldots,M\} \). Given any specific process \( z \), their interacting amino-acid sequences are respectively described by two \( N \)-length blocks of discrete stochastic variables \( \{X_i,...,X_N\} \) and \( \{Y_1,...,Y_N\} \) with probability mass functions \( \{\rho(x_1,...,x_N),\rho(y_1,...,y_N),\rho(x_1,...,x_N,y_1,...,y_N|z)\} \) such that
\[
\rho(x_1,...,x_N) = \sum_{y_1,...,y_N} \rho(x_1,...,x_N,y_1,...,y_N|z)
\]
\[
\rho(y_1,...,y_N) = \sum_{x_1,...,x_N} \rho(x_1,...,x_N,y_1,...,y_N|z)
\]
and
\[
\sum_{x_1,...,x_N,y_1,...,y_N} \rho(x_1,...,x_N,y_1,...,y_N|z) = 1
\]
over every joint sequence \( \{x_1,...,x_N,y_1,...,y_N\}_i \), defined in the alphabet \( \chi \) of size \(|\chi|\).

Under these considerations, we are interested in quantifying the amount of information that protein \( A \) stores about the interacting amino-acids of protein \( B \) conditional to any given coevolution process. As made explicit in eq. [1], we are particularly interested in the situation in which marginals of the \( N \)-block variables \( \{\rho(x_1,...,x_N),\rho(y_1,...,y_N)\} \) are independent of process \( z \) meaning that, for a fixed sequence composition of proteins \( A \) and \( B \) only their joint distribution depends on coevolution. Furthermore, by assuming \( N \)-independent contacts, we want that information to be quantified for the least-constrained model \( \rho^*(x_1,...,x_N,y_1,...,y_N|z) \) that maximizes the conditional joint entropy between \( A \) and \( B \) - that condition ensures the mutual information to be written exactly, in terms of the individual contributions of contacts \( i \). Given its entropy-maximization property (1), \( \rho^*(x_1,...,x_N,y_1,...,y_N|z) \) factorizes into the conditional joint distributions of individual contacts \( i \)
\[
\rho^*(x_1,...,x_N,y_1,...,y_N|z) = \prod_{i=1}^{N} \rho^*(x_i,y_i|z)
\]
such that
\[
\rho^*(x_1,...,x_N|z) = \sum_{y_1,...,y_N} \rho^*(x_1,...,x_N,y_1,...,y_N|z) = \prod_{i=1}^{N} \sum_{y_i} \rho(x_i,y_i|z) = \prod_{i=1}^{N} \rho(x_i|z)
\]
\[
\rho^*(y_1,...,y_N|z) = \sum_{x_1,...,x_N} \rho^*(x_1,...,x_N,y_1,...,y_N|z) = \prod_{i=1}^{N} \sum_{x_i} \rho(x_i,y_i|z) = \prod_{i=1}^{N} \rho(y_i|z)
\]
are marginals for any specific \( N \)-block sequence of proteins \( A \) and \( B \). Eq. [S3] ensures the conditional joint entropy to be written extensively in terms of entropic contributions of contact \( i \).
given that

\[
\sum_{x_{s},y_{s} \in \mathcal{S}} \rho^{*}(x_{s},x_{l},y_{l},z) = \prod_{i=2}^{N} \sum_{x_{s}} \rho^{*}(x_{s},y_{l}|z) = 1 \\
\sum_{x_{s},y_{s} \in \mathcal{S}} \rho^{*}(x_{s},x_{l},y_{l},z) = \prod_{i=1}^{N-1} \sum_{x_{s}} \rho^{*}(x_{s},y_{l}|z) = 1
\]

are normalized conditional joint probabilities of \(2(N-1)\)-block sequences. The consequence for the conditional entropy of the individual block variables is then clear

\[
H(X_{1},...,X_{N}|z) = -\sum_{x_{1},...,x_{N}} \rho(x_{1},...,x_{N}) \ln \rho(x_{1},...,x_{N}|z) \\
= -\sum_{x_{1}} \rho^{*}(x_{1}|z) \ln \rho^{*}(x_{1}|z) \times \prod_{i=2}^{N} \sum_{x_{s}} \rho^{*}(x_{s},x_{l},y_{l}|z) ... \\
= \left[ \sum_{x_{1}} \rho^{*}(x_{1}|z) \ln \rho^{*}(x_{1}|z) \times \sum_{x_{s}} \rho^{*}(x_{s},x_{l},y_{l}|z) \ln \rho^{*}(x_{s},y_{l}|z) \right] \\
= \sum \left[ \sum \rho^{*}(x_{s},y_{l}|z) \ln \rho^{*}(x_{s},y_{l}|z) \right] \\
= \sum H(x_{1},...,x_{N}|z) \\

H(Y_{1},...,Y_{N}|z) = -\sum_{y_{1},...,y_{N}} \rho(y_{1},...,y_{N}) \ln \rho(y_{1},...,y_{N}|z) \\
= -\sum_{y_{1}} \rho^{*}(y_{1}|z) \ln \rho^{*}(y_{1}|z) \times \prod_{i=2}^{N} \sum_{y_{s}} \rho^{*}(y_{s},y_{l}|z) ... \\
= \left[ \sum_{y_{1}} \rho^{*}(y_{1}|z) \ln \rho^{*}(y_{1}|z) \times \sum_{y_{s}} \rho^{*}(y_{s},y_{l}|z) \ln \rho^{*}(y_{s}|z) \right] \\
= \sum \left[ \sum \rho^{*}(y_{s},y_{l}|z) \ln \rho^{*}(y_{s},y_{l}|z) \right] \\
= \sum H(y_{1},...,y_{N}|z)
\]

where

\[
\sum_{x_{s},x_{l}} \rho^{*}(x_{s},x_{l},z) = \prod_{i=2}^{N} \sum_{x_{s}} \rho^{*}(x_{s}|z) = 1, \ldots, \sum_{x_{s},x_{l}} \rho^{*}(x_{s},x_{l},z) = \prod_{i=1}^{N-1} \sum_{x_{s}} \rho^{*}(x_{s}|z) = 1
\\
\sum_{y_{s},y_{l}} \rho^{*}(y_{s},y_{l},z) = \prod_{i=2}^{N} \sum_{y_{s}} \rho^{*}(y_{s}|z) = 1, \ldots, \sum_{y_{s},y_{l}} \rho^{*}(y_{s},y_{l},z) = \prod_{i=1}^{N-1} \sum_{y_{s}} \rho^{*}(y_{s}|z) = 1
\]

are normalized probabilities of \((N-1)\)-block sequences. Throughout any specific coevolution process \(z\), the amount of information that protein \(A\) stores about the interacting amino-acids of protein \(B\) is given by the conditional mutual information \(I(X_{1},...,X_{N};Y_{1},...,Y_{N}|z)\) between the stochastic variables \((X_{1},...,X_{N})\) and \((Y_{1},...,Y_{N})\).

The expectation value of \(I(X_{1}^n;Y_{1}^m|z)\) across the entire distribution of \(M!\) distinct coevolution processes reads as

\[
I(X_{1},...,X_{N};Y_{1},...,Y_{N}|z) = \sum_{z} \rho(z) \left[I(X_{1},...,X_{N};Y_{1},...,Y_{N}|z) \right] \\
\]

the mutual information between the block variables conditionally to the discrete stochastic variable \(Z\). Eq. \([S9]\) can be rewritten

\[
I(X_{1},...,X_{N};Y_{1},...,Y_{N}|Z) = I(X_{1},...,X_{N};Y_{1},...,Y_{N}) + I(X_{1},...,X_{N};Y_{1},...,Y_{N}|Z) - I(X_{1},...,X_{N}|Z) - I(Y_{1},...,Y_{N}|Z)
\]

in terms of the information entropies
\[
\begin{align*}
I(X_1, \ldots, X_N | Z) &= H(X_1, \ldots, X_N) - H(X_1, \ldots, X_N | Z) \\
I(Y_1, \ldots, Y_N | Z) &= H(Y_1, \ldots, Y_N) - H(Y_1, \ldots, Y_N | Z) \\
I(X_1, \ldots, X_N, Y_1, \ldots, Y_N | Z) &= H(X_1, \ldots, X_N, Y_1, \ldots, Y_N) - H(X_1, \ldots, X_N, Y_1, \ldots, Y_N | Z) \\
I(X_1, \ldots, X_N, Y_1, \ldots, Y_N) &= H(X_1, \ldots, X_N) - H(Y_1, \ldots, Y_N) - H(X_1, \ldots, X_N, Y_1, \ldots, Y_N | Z)
\end{align*}
\] (S11)

associated with single and joint probability distributions \( \{ \rho^* (x_1, \ldots, x_N | z), \rho^* (y_1, \ldots, y_N | z), \rho^* (x_1, \ldots, x_N, y_1, \ldots, y_N | z) \} \) in eq. [S3 and S4]. For the condition in eq. [S1]

\[
\begin{align*}
\rho^* (x_1, \ldots, x_N | z) &= \rho^* (x_1, \ldots, x_N) \\
\rho^* (y_1, \ldots, y_N | z) &= \rho^* (y_1, \ldots, y_N)
\end{align*}
\] (S12)

the information entropy of either block variables \( H(X_1, \ldots, X_N | Z) \) and \( H(Y_1, \ldots, Y_N | Z) \) are independent of \( Z \)

\[
\begin{align*}
H(X_1, \ldots, X_N | Z) &= H(X_1, \ldots, X_N) \\
H(Y_1, \ldots, Y_N | Z) &= H(Y_1, \ldots, Y_N)
\end{align*}
\] (S13)

thus simplifying eq. [S10]

\[
I(X_1, \ldots, X_N; Y_1, \ldots, Y_N | Z) = H(X_1, \ldots, X_N) + H(Y_1, \ldots, Y_N) - H(X_1, \ldots, X_N, Y_1, \ldots, Y_N | Z)
\] (S14)

into the joint entropy differences between \( X_1, \ldots, X_N \) and \( Y_1, \ldots, Y_N \) when unconditionally and conditionally dependent on \( Z \). From eq. [S5, S7 and S13], the conditional mutual information then rewrites

\[
I(X_1, \ldots, X_N; Y_1, \ldots, Y_N | Z) = \sum_{i=1}^{N} H(X_i | Z) - H(X_i | Y_i | Z)
\]

\[
= \sum_{i} \rho(z) \sum_{i=1}^{N} H(X_i | z) + H(Y_i | z) - H(X_i | Y_i | z)
\]

\[
= \sum_{i} \rho(z) \sum_{i=1}^{N} I(X_i; Y_i | z)
\] (S15)

implying

\[
I(X_1, \ldots, X_N; Y_1, \ldots, Y_N | z) = \sum_{i=1}^{N} I(X_i; Y_i | z)
\] (S16)

as a direct consequence of eq. [S9].

REFERENCES

1. Cover TM, Thomas JA (2006) Elements of Information Theory 2nd Edition (Wiley-Interscience, Hoboken, N.J.). 2 edition.

2. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife 3:e02030.
SUPPLEMENTARY FIGURES AND TABLES

Fig. S1. Informational analysis of six protein-protein complexes used in Baker and coworkers. (A) Three-dimensional representation of stochastic variables $X_N$ and $Y_N$ as defined from physically coupled amino acids at short-range cutoff distances $r_c \leq 8.0 \text{Å}$ (turquoise) and physically uncoupled amino-acids at long-range cutoff distances $r_c > 8.0 \text{Å}$ (gray). (B) Conditional mutual information $\langle I(X_N;Y_N|z) \rangle_{M-n}$ as a function of the number $M-n$ of randomly paired proteins in the reference MSA. $\langle I(X_N;Y_N|z) \rangle_{M-n}$ are expectation values estimated from a generated ensemble of $\sim 100$ MSA models. (C) Conditional mutual information as a function of protein contact $i$. Mutual information $I(X_i;Y_i|z)$ for the reference alignment (black) is systematically larger than $\langle I(X_i;Y_i|z) \rangle_M$ for scrambled models (green) along every contact $i$. (D) Mutual information gap $\Delta I_M$ between reference and $100$ random models featuring $M$ randomly paired sequences. (E) Per-contact mutual information gap $\Delta I_{M-n}$, $r_c \leq 8 \text{Å}$ (F) Mutual information decomposition ( $N^{-1} \Delta \Delta I_{M-n}^{\text{Cov}}$, $r_c \leq 8 \text{Å}$) and comparison with functional mutual information ( $MI_{p.a}$, $r_c \leq 8 \text{Å}$) and direct information ( $DI_{r_c \leq 8 \text{Å}}$ ). In C, D, E and F error bars correspond to standard deviations.
Fig. S2. Information gap $\Delta I_m$ profile as a function of amino-acid (AA) pair distances. Shown are average values and the associated standard deviations (error bars) of $\Delta I_m$ at various pair distances. The profile shows few larger values of $\Delta I_m$ at short distances in contrast to many smaller ones at long distances.
Fig. S3. Degeneracy and error analysis for $X^N$ and $Y^N$ involving interacting amino acids at short-range distances $r_c \leq 8.0\,\text{Å}$ (blue), long-range distances $r_c > 8.0\,\text{Å}$ (red), or both (green). (A) Total number $\omega_S$ of native-like models at various resolutions $\delta I$. (B) Per-contact gaps of mutual information $N^{-1} \Delta I_{\mu \nu}$ as a function of the number $M-n$ of randomly paired sequences in the reference alignment. Error bars correspond to standard deviations. (C) Expectation values $\langle \epsilon \rangle_S$ for the fraction of sequence matches at various resolutions $\delta I$. 

\[ \text{Fig. S3. Degeneracy and error analysis for } X^N \text{ and } Y^N \text{ involving interacting amino acids at short-range distances } r_c \leq 8.0\,\text{Å} \text{ (blue), long-range distances } r_c > 8.0\,\text{Å} \text{ (red), or both (green). } (A) \text{ Total number } \omega_S \text{ of native-like models at various resolutions } \delta I \text{. } (B) \text{ Per-contact gaps of mutual information } N^{-1} \Delta I_{\mu \nu} \text{ as a function of the number } M-n \text{ of randomly paired sequences in the reference alignment. Error bars correspond to standard deviations. } (C) \text{ Expectation values } \langle \epsilon \rangle_S \text{ for the fraction of sequence matches at various resolutions } \delta I. \]
Fig. S4. Dependence with contact definition \( r_c^* \) and docking decoys. (A) \( N^{-1}\Delta I_{M, r} \) and \( MI_{r,r} \) at various \( r_c^* \). (B) \( N^{-1}\Delta I_{M, r} \) (turquoise), \( N^{-1}\Delta I_{M, r}^{cou} \) (green), \( MI_{r,r} \) (blue) at alternative interfaces generated by docking – only physically coupled amino acids as defined for \( r_c \leq 8.0 \) Å were included in the calculations. Black bars represent the root-mean-square deviation (RMSD) between the native bound structure and docking decoys.
Table-S1. Protein system A and B considered in the study.

| Complex description                      | PDB ID | Protein A                                      | Protein B                                      | M  | MSA length |
|------------------------------------------|--------|-----------------------------------------------|-----------------------------------------------|----|------------|
| Carbamoyl Phosphate Synthetase           | 1BXR   | Chain A: Carbamoyl-Phosphate Synthetase large subunit | Chain B: Carbamoyl-Phosphate Synthetase small subunit | 1004 | 1452       |
| Lactococcus Lactis Dihydroorotate Dehydrogenase B. | 1EP3   | Chain A: Dihydroorotate Dehydrogenase B (PYRD Subunit) | Chain B: Dihydroorotate Dehydrogenase B (pyrk Subunit) | 552  | 572        |
| Polysulfide reductase native structure   | 2VPZ   | Chain A: THIOSULFATE REDUCTASE                | Chain B: NRFC PROTEIN                          | 676 | 927        |
| heterohexameric TusBCD proteins          | 2D1P   | Chain B: Hypothetical UPF0116 protein yheM     | Chain C: Hypothetical protein yheL             | 216 | 214        |
| 3-oxoadipate coA-transferase             | 3RRL   | Chain A: Succinyl-CoA:3-ketoacid-coenzyme A transferase subunit A | Chain B: Succinyl-CoA:3-ketoacid-coenzyme A transferase subunit B | 1330 | 437      |
| Bovine heart cytochrome c oxidase        | 2Y69   | Chain A: Cytochrome C Oxidase Subunit 1        | Chain B: CYTOCHROME C OXIDASE SUBUNIT 2        | 1484 | 740        |
Table S2. Rencontres numbers \( \omega_{M-n} \) as a function of the number \( M-n \) of randomly paired sequences in the reference alignment \( \{(x^*_k, y^*_l | z^*_m)\}_{M} \).

| M-n | 1BXR AB | 1EP3 AB | 2VPZ AB | 2D1P BC | 3RRL AB | 2Y69 AB |
|-----|---------|---------|---------|---------|---------|---------|
| 0   | 1       | 1       | 1       | 1       | 1       | 1       |
| 1   | 0       | 0       | 0       | 0       | 0       | 0       |
| 2   | 503506  | 152076  | 228150  | 23220   | 883785  | 1100386 |
| 3   | 336342008 | 55761200 | 102515400 | 3312720 | 702444320 | 1087181368 |
| 4   | 378763143759 | 34439511150 | 77616972225 | 793810530 | 1160091564220 | 1811380056759 |
| 5   | 370346185008800 | 18453465396640 | 50999525216640 | 164548102752 | 151449649396704 | 2621268206581024 |
| ... | ...     | ...     | ...     | ...     | ...     | ...     |
| 40  | 1.963993579054E+119 | 4.115347994062E+108 | 1.788169031032E+112 | 1.862684992297E+91 | 1.830259053207E+124 | 1.558622316946E+126 |