Supporting Information

Coevolutive, Evolutive and Stochastic Information in Protein-Protein Interactions

Miguel Andrade, Camila Pontes and Werner Treptow

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 $i=1,\dots,N$ 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 $\rho(z)$, $\forall z\in\{1,\dots,M\}$. Given any specific process z, their interacting amino-acid sequences are respectively described by two N-length blocks of discrete stochastic variables $(X_1,\dots,X_N)$ and $(Y_1,\dots,Y_N)$ with probability mass functions $\{\rho(x_1,\dots,x_N),\rho(y_1,\dots,y_N),\rho(x_1,\dots,x_N,y_1,\dots,y_N|z)\}$ such that

$$\begin{cases}\rho(x_1,\dots,x_N)=\sum\limits_{y'_1,\dots,y'_N}\rho(x_1,\dots,x_N,y'_1,\dots,y'_N|z)\\ \rho(y_1,\dots,y_N)=\sum\limits_{x'_1,\dots,x'_N}\rho(x'_1,\dots,x'_N,y_1,\dots,y_N|z)\end{cases}\tag{S1}$$

and

$$\sum_{x'_1,\dots,x'_N,y'_1,\dots,y'_N}\rho(x'_1,\dots,x'_N,y'_1,\dots,y'_N|z)=1\tag{S2}$$

over every joint sequence $\{x_1,\dots,x_N,y_1,\dots,y_N\}_{|\chi|^{2N}}$ defined in the alphabet χ 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 $\{\rho(x_1,\dots,x_N),\rho(y_1,\dots,y_N)\}$ 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 $\rho^*(x_1,\dots,x_N,y_1,\dots,y_N|z)$ 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], $\rho^*(x_1,\dots,x_N,y_1,\dots,y_N|z)$ factorizes into the conditional joint distributions of individual contacts i

$$\rho^*(x_1,\dots,x_N,y_1,\dots,y_N|z)=\prod_{i=1}^{N}\rho^*(x_i,y_i|z)\tag{S3}$$

such that

$$\begin{cases}\rho^*(x_1,\dots,x_N|z)=\sum\limits_{y'_1,\dots,y'_N}\rho^*(x_1,\dots,x_N,y'_1,\dots,y'_N|z)=\prod\limits_{i=1}^{N}\sum\limits_{y'_i}\rho(x_i,y'_i|z)=\prod\limits_{i=1}^{N}\rho(x_i|z)\\ \rho^*(y_1,\dots,y_N|z)=\sum\limits_{x'_1,\dots,x'_N}\rho^*(x'_1,\dots,x'_N,y_1,\dots,y_N|z)=\prod\limits_{i=1}^{N}\sum\limits_{x'_i}\rho(x'_i,y_i|z)=\prod\limits_{i=1}^{N}\rho(y_i|z)\end{cases}\tag{S4}$$

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

$$H(X_1,...,X_N,Y_1,...,Y_N|z) = -\sum_{x_1^{'},...,x_N^{'},y_1^{'},...,y_N^{'}} \rho(x_1^{'},...,x_N^{'},y_1^{'},...,y_N^{'}|z)\ln\rho(x_1^{'},...,x_N^{'},y_1^{'},...,y_N^{'}|z)$$

$$= -\sum_{x_1^{'},y_1^{'}} \rho^*(x_1^{'},y_1^{'}|z)\ln\rho^*(x_1^{'},y_1^{'}|z)\times[\overbrace{\sum_{x_2^{'},...,x_N^{'},y_2^{'},...,y_N^{'}} \rho^*(x_2^{'},...,x_N^{'},y_2^{'},...,y_N^{'}|z)}^{=1}]...$$

$$-[\overbrace{\sum_{x_1^{'},...,x_{N-1}^{'},y_1^{'},...,y_{N-1}^{'}} \rho^*(x_1^{'},...,x_{N-1}^{'},y_1^{'},...,y_{N-1}^{'}|z)}^{=1}]\times\sum_{x_N^{'},y_N^{'}} \rho^*(x_N^{'},y_N^{'}|z)\ln\rho^*(x_N^{'},y_N^{'}|z) \quad \text{(S5)}$$

$$= \sum_i -\sum_{x_i^{'},y_i^{'}} \rho^*(x_i^{'},y_i^{'}|z)\ln\rho^*(x_i^{'},y_i^{'}|z)$$

$$= \sum_i H(X_i,Y_i|z)$$

given that

$$\begin{cases} \sum_{x_2^{'},...,x_N^{'},y_2^{'},...,y_N^{'}} \rho^*(x_2^{'},...,x_N^{'},y_2^{'},...,y_N^{'}|z) = \prod_{i=2}^{N} \sum_{x_i^{'},y_i^{'}} \rho^*(x_i^{'},y_i^{'}|z) = 1 \\ ... \\ \sum_{x_1^{'},...,x_{N-1}^{'},y_1^{'},...,y_{N-1}^{'}} \rho^*(x_1^{'},...,x_{N-1}^{'},y_1^{'},...,y_{N-1}^{'}|z) = \prod_{i=1}^{N-1} \sum_{x_i^{'},y_i^{'}} \rho^*(x_i^{'},y_i^{'}|z) = 1 \end{cases} \quad \text{(S6)}$$

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

$$\begin{cases} H(X_1,...,X_N|z) = -\sum_{x_1^{'},...,x_N^{'}} \rho(x_1^{'},...,x_N^{'}|z)\ln\rho(x_1^{'},...,x_N^{'}|z) \\ \quad = -\sum_{x_1^{'}} \rho^*(x_1^{'}|z)\ln\rho^*(x_1^{'}|z)\times[\overbrace{\sum_{x_2^{'},...,x_N^{'}} \rho^*(x_2^{'},...,x_N^{'}|z)}^{=1}]... \\ \quad -[\overbrace{\sum_{x_1^{'},...,x_{N-1}^{'}} \rho^*(x_1^{'},...,x_{N-1}^{'}|z)}^{=1}]\times\sum_{x_N^{'}} \rho^*(x_N^{'}|z)\ln\rho^*(x_N^{'}|z) \\ \quad = \sum_i -\sum_{x_i^{'}} \rho^*(x_i^{'}|z)\ln\rho^*(x_i^{'}|z) \\ \quad = \sum_i H(X_i|z) \\ H(Y_1,...,Y_N|z) = -\sum_{y_1^{'},...,y_N^{'}} \rho(y_1^{'},...,y_N^{'}|z)\ln\rho(y_1^{'},...,y_N^{'}|z) \\ \quad = -\sum_{y_1^{'}} \rho^*(y_1^{'}|z)\ln\rho^*(y_1^{'}|z)\times[\overbrace{\sum_{y_2^{'},...,y_N^{'}} \rho^*(y_2^{'},...,y_N^{'}|z)}^{=1}]... \\ \quad -[\overbrace{\sum_{y_1^{'},...,y_{N-1}^{'}} \rho^*(y_1^{'},...,y_{N-1}^{'}|z)}^{=1}]\times\sum_{y_N^{'}} \rho^*(y_N^{'}|z)\ln\rho^*(y_N^{'}|z) \\ \quad = \sum_i -\sum_{y_i^{'}} \rho^*(y_i^{'}|z)\ln\rho^*(y_i^{'}|z) \\ \quad = \sum_i H(Y_i|z) \end{cases} \quad \text{(S7)}$$

where

$$\begin{cases} \sum_{x_2^{'},...,x_N^{'}} \rho^*(x_2^{'},...,x_N^{'}|z) = \prod_{i=2}^{N} \sum_{x_i^{'}} \rho^*(x_i^{'}|z) = 1,..., \sum_{x_1^{'},...,x_{N-1}^{'}} \rho^*(x_1^{'},...,x_{N-1}^{'}|z) = \prod_{i=1}^{N-1} \sum_{x_i^{'}} \rho^*(x_i^{'}|z) = 1 \\ \sum_{y_2^{'},...,y_N^{'}} \rho^*(y_2^{'},...,y_N^{'}|z) = \prod_{i=2}^{N} \sum_{y_i^{'}} \rho^*(y_i^{'}|z) = 1,..., \sum_{y_1^{'},...,y_{N-1}^{'}} \rho^*(y_1^{'},...,y_{N-1}^{'}|z) = \prod_{i=1}^{N-1} \sum_{y_i^{'}} \rho^*(y_i^{'}|z) = 1 \end{cases} \quad \text{(S8)}$$

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 $I(X^N;Y^N|z)$ across the entire distribution of $M!$ distinct coevolution processes reads as

$$I(X_1,...,X_N;Y_1,...,Y_N|Z) = \sum_{z} \rho(z')I(X_1,...,X_N;Y_1,...,Y_N|z') \quad \text{(S9)}$$

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

$$I(X_1,...,X_N;Y_1,...,Y_N|Z) = I(X_1,...,X_N;Y_1,...,Y_N) + I(X_1,...,X_N,Y_1,...,Y_N|Z) - I(X_1,...,X_N|Z) - I(Y_1,...,Y_N|Z) \quad \text{(S10)}$$

in terms of the information entropies

$$\begin{cases} I(X_1,...,X_N|Z)=H(X_1,...,X_N)-H(X_1,...,X_N|Z) \\ I(Y_1,...,Y_N|Z)=H(Y_1,...,Y_N)-H(Y_1,...,Y_N|Z) \\ I(X_1,...,X_N,Y_1,...,Y_N|Z)=H(X_1,...,X_N,Y_1,...,Y_N)-H(X_1,...,X_N,Y_1,...,Y_N|Z) \\ I(X_1,...,X_N;Y_1,...,Y_N)=H(X_1,...,X_N)-H(Y_1,...,Y_N)-H(X_1,...,X_N,Y_1,...,Y_N) \end{cases} \tag{S11}$$

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

$$\begin{cases} \rho^*(x_1,...,x_N|z)=\rho^*(x_1,...,x_N) \\ \rho^*(y_1,...,y_N|z)=\rho^*(y_1,...,y_N) \end{cases}, \tag{S12}$$

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

$$\begin{cases} H(X_1,...,X_N|Z)=H(X_1,...,X_N) \\ H(Y_1,...,Y_N|Z)=H(Y_1,...,Y_N) \end{cases} \tag{S13}$$

thus simplifying eq. [S10]

$$I(X_1,...,X_N;Y_1,...,Y_N|Z)=H(X_1,...,X_N)+H(Y_1,...,Y_N)-H(X_1,...,X_N,Y_1,...,Y_N|Z) \tag{S14}$$

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

$$\begin{aligned} I(X_1,...,X_N;Y_1,...,Y_N|Z) &= \sum_{i=1}^{N} H(X_i|Z)+H(Y_i|Z)-H(X_i,Y_i|Z) \\ &= \sum_{z'} \rho(z') \sum_{i=1}^{N} H(X_i|z')+H(Y_i|z')-H(X_i,Y_i|z') \\ &= \sum_{z'} \rho(z') \sum_{i=1}^{N} I(X_i;Y_i|z') \end{aligned} \tag{S15}$$

implying

$$I(X_1,...,X_N;Y_1,...,Y_N|z)=\sum_{i=1}^{N} I(X_i;Y_i|z) \tag{S16}$$

as a direct consequence of eq. [S9].

REFERENCES

(1) Cover, T. M.; Thomas, J. A. *Elements of Information Theory 2nd Edition*, 2 edition.; Wiley-Interscience: Hoboken, N.J., 2006.

(2) Ovchinnikov, S.; Kamisetty, H.; Baker, D. Robust and Accurate Prediction of Residue–Residue Interactions across Protein Interfaces Using Evolutionary Information. *eLife* **2014**, *3*, e02030. https://doi.org/10.7554/eLife.02030.

SUPPLEMENTARY FIGURES AND TABLES

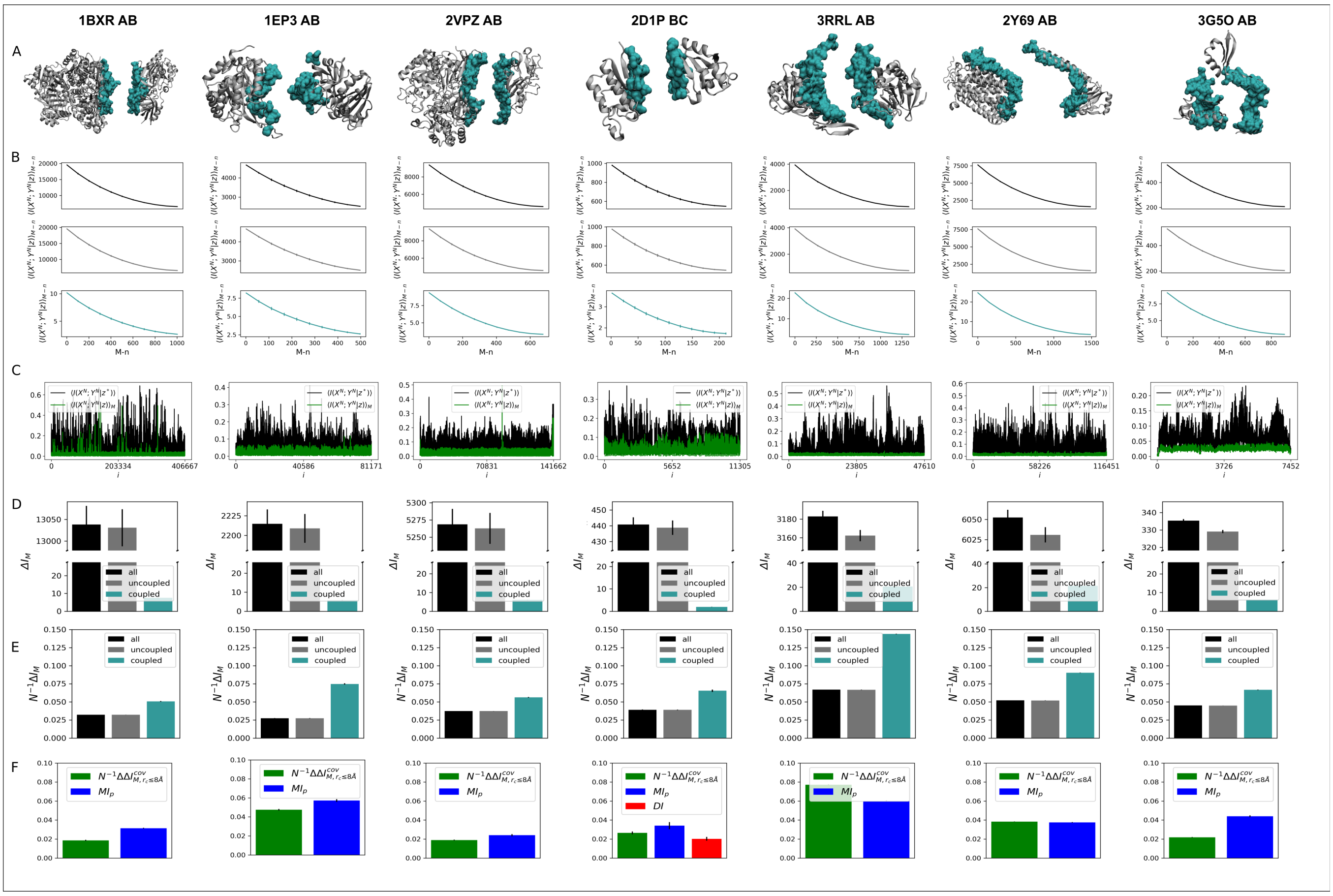

Fig. S1. Informational analysis of protein-protein complexes used in Baker and coworkers.[2] All protein complexes but 2G5O are obligate dimers. (A) Three-dimensional representation of stochastic variables X^N and Y^N as defined from physically coupled amino acids at short-range cutoff distances $r_c \leq 8.0\,\mathring{A}$ (turquoise) and physically uncoupled amino-acids at long-range cutoff distances $r_c > 8.0\,\mathring{A}$ (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 ~ 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 ΔI_M between reference and 100 random models featuring M randomly paired sequences. (E) Per-contact mutual information gap $N^{-1}\Delta I_{M,r_c}$. (F) Mutual information decomposition ($N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\,\mathring{A}}$) and comparison with functional mutual information ($MI_{p,r_c\leq 8\,\mathring{A}}$) and direct information ($DI_{r_c\leq 8\,\mathring{A}}$). In C, D, E and F error bars correspond to standard deviations.

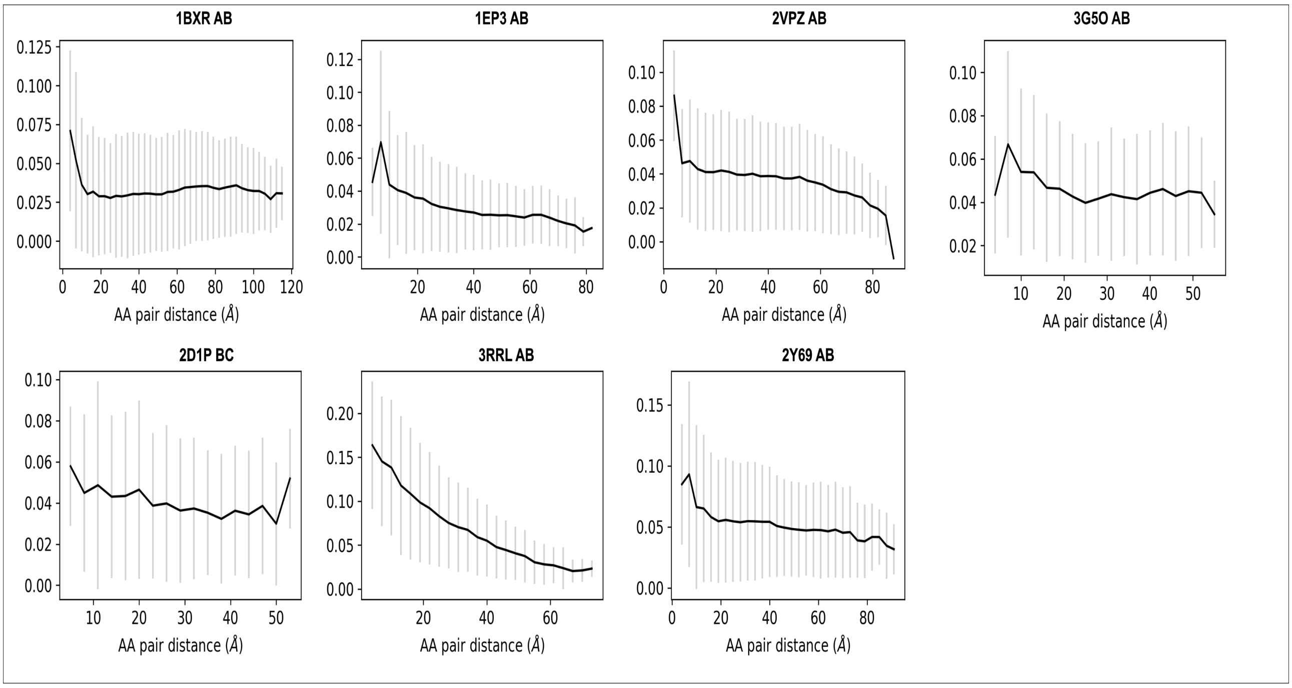

Fig. S2. Information gap ΔI_M profile as a function of amino-acid (AA) pair distances. Shown are average values and the associated standard deviations (error bars) of ΔI_M at various pair distances. The profile shows few larger values of ΔI_M at short distances in contrast to many smaller ones at long distances.

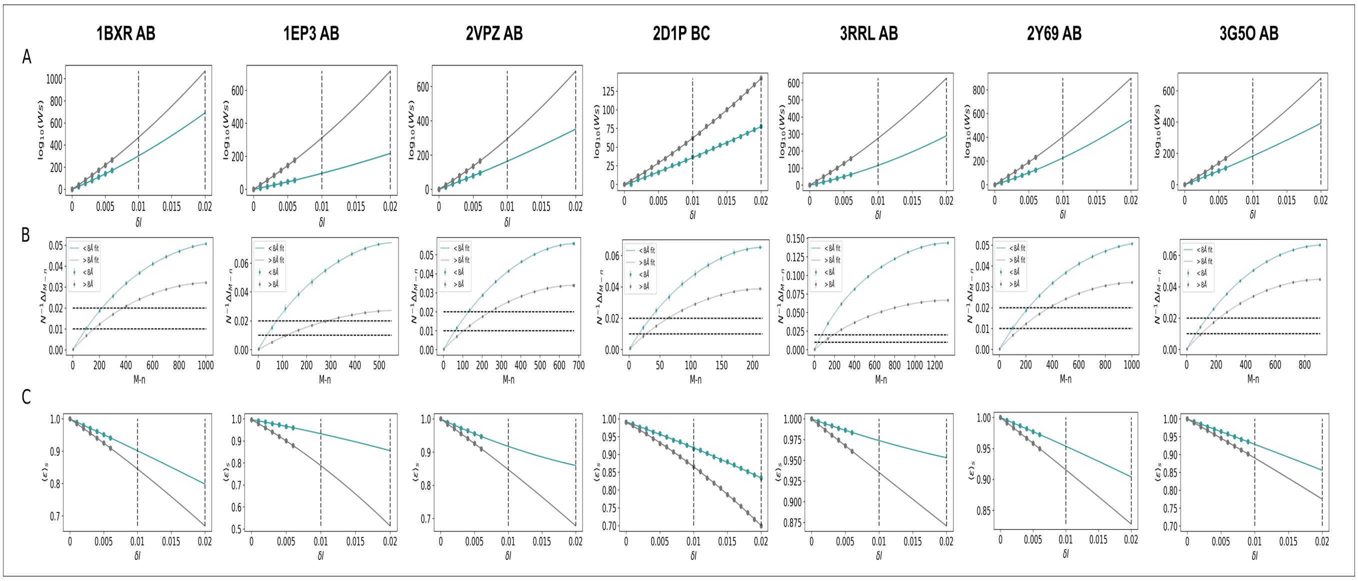

Fig. S3. Degeneracy and error analysis for X^N and Y^N involving interacting amino acids at short-range distances $r_c \leq 8.0\,\mathring{A}$ (blue), long-range distances $r_c > 8.0\,\mathring{A}$ (red), or both (green). (A) Total number ω_S of native-like models at various resolutions δI. (B) Per-contact gaps of mutual information $N^{-1} \Delta I_{M-n, r_c}$ 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 δI.

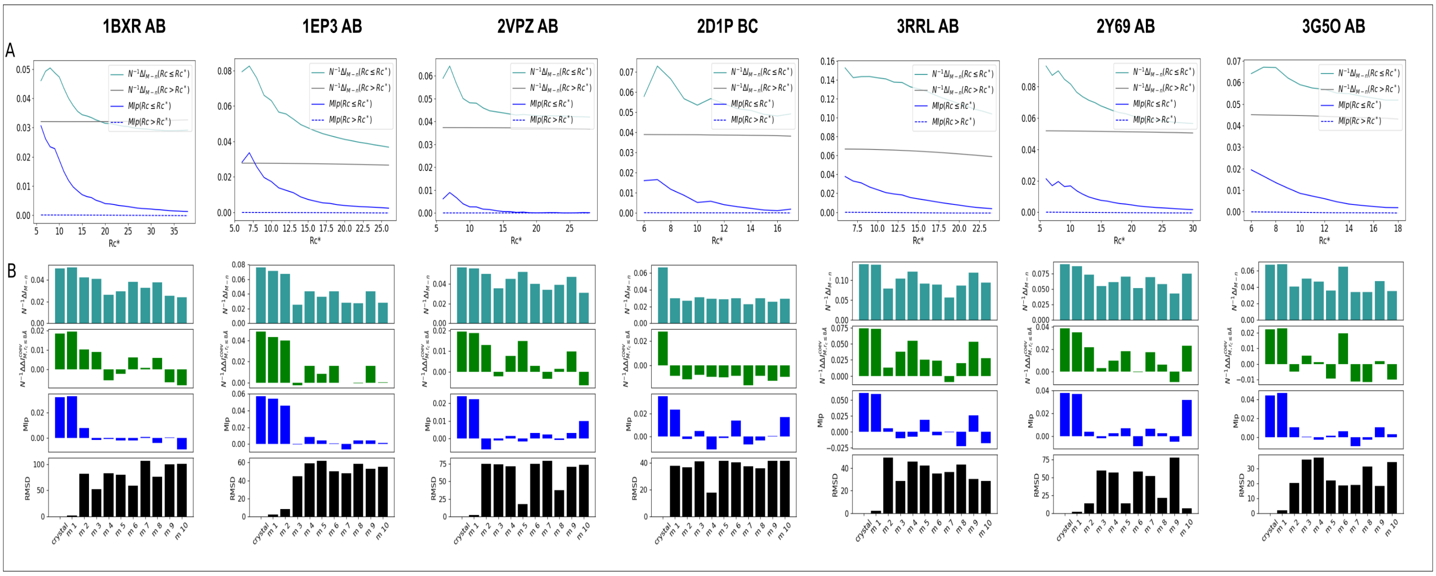

Fig. S4. Dependence with contact definition r_c^* and docking decoys. (A) $N^{-1}\Delta I_{M,r_c}$ and MI_{p,r_c} at various r_c^*. (B) $N^{-1}\Delta I_{M,r_c}$ (turquoise), $N^{-1}\Delta\Delta I_{M,r_c}^{Cov}$ (green), MI_{p,r_c} (blue) at alternative interfaces generated by docking – only physically coupled amino acids as defined for $r_c \leq 8.0\,\mathring{A}$ were included in the calculations. Black bars represent the root-mean-square deviation (RMSD) between the native bound structure and docking decoys.

Table-S1. *Rencontres* numbers $\omega_{M,n}$ as a function of the number $M-n$ of randomly paired sequences in the reference alignment $\{(x_k^N, y_l^N | z^*)\}_M$.

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	782444320	1087181368
4	378763143759	34439511150	77616972225	793810530	1168091564220	1811380056759
5	370346185008800	18453455396640	50999525216640	164548102752	1514469649396704	2621268206581024
…	…	…	…	…	…	…
40	1,963993579054E+119	4,115347994062E+108	1,788169031032E+112	1,8626849992297E+91	1,830259053207E+124	1,558622316946E+126