University of Brasilia - UnB
Institute of Biology - IB
Department of Cell Biology - CEL
Computational and Theoretical Biology Lab - LBTC

Assessing the Utility of Mutual Information Stored in Protein-Protein Interfaces to Infer Specific Protein Partners

Camila Ferreira Thé Pontes

Supervisor: Prof. Werner Treptow

A Thesis presented to the Molecular Biology Graduate Program at the University of Brasilia in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

March 16, 2021

Acknowledgements

I would first like to thank my supervisor, prof. Werner Treptow, for the time and patience dedicated to this work. He encouraged me to learn many things, and was always quick to respond and provide feedback when I needed.

I would like to thank my colleagues from the Computational and Theoretical Biology Lab (LBTC) at the University of Brasilia for all the support and shared experiences. I really enjoyed the time I spent working there.

I would like to acknowledge the support of Alfonso Valencia and his group at the Barcelona Supercomputing Center (BSC) for receiving me there are also providing important feedback on my work. I am very grateful for the opportunity they gave me to be part of the group.

I would like to thank prof. Marcelo Marotta, prof. João Gondim and their students from the Computer Science Department at the University of Brasilia for collaborating with me in parallel projects, giving me the opportunity to apply my knowledge in another research field.

I would also like to acknowledge the support of my family and friends, especially during covid times. Without their emotional support it would have being significantly tougher to finish this work.

Finally, I would like to acknowledge the financial support received from CAPES (PhD fellowship), FAPDF (financial support to attend conferences) and EuroLab4HPC (financial support for a 3 months internship at the BSC).

Abstract

Proteins are essential for several cellular processes. Hence, one of the central objectives in Biology is to understand the relationships between sequence, structure and function of these macromolecules. In this context, marks left by the coevolutionary process in interacting protein sequences are an important source of structural information. In fact, statistical correlations between amino acid sites in protein sequences are at the basis of state-of-the-art methods for prediction of inter- and intra-protein contacts, template-free structure prediction, identification of functional sites and specificity determining residues, inference of interacting paralogs, among other applications. In line with that, the present work conveys a set of theoretical results on how specific protein partners can be recovered based on sequence information alone. In the first chapter, a decomposition of the mutual information (MI) present in protein-protein complexes is carried out, considering the hypothesis that MI in proteins is originated from a combination of coevolutive, evolutive and stochastic sources. It was observed that the interface contains on average, by contact, more information than the rest of the protein complex, a result that holds when considering both Shannon and Tsallis MI as a measure of information. This observation led to the conclusion that the interface contains the strongest information signal for distinguishing the correct set of protein partners in interacting protein families. Building on that, the utility of using MI encoded on protein-protein interfaces to recover the correct set of protein partners is assessed in the second chapter. A genetic algorithm (GA) was developed to explore the space of possible concatenations between a pair of interacting protein families using the interface MI as objective function. Using the GA, interface MI maximization was performed for 26 different pairs of interacting protein families and it was observed that optimized concatenations corresponded to degenerate solutions with two distinct er-

ror sources, arising from mismatches among (i) similar and (ii) non-similar sequences. When mistakes made among similar sequences were disregarded, type-(i) solutions were found to resolve correct pairings at best true positive (TP) rates of 70% - far above the very same estimates in type-(ii) solutions. These results hold when the optimizations are made based on Tsallis MI. These findings raise further questions about the mechanisms behind protein partners coevolution and help rationalize literature data showing a sharp deterioration of TP rates with increasing sequence number in MI-based approaches.

Resumo

Proteínas são essenciais para diversos processos celulares. Assim, um dos objetivos centrais da Biologia é entender as relações entre sequência, estrutura e função dessas macromoléculas. Nesse contexto, as marcas deixadas pelo processo coevolutivo em sequências de proteínas parceiras são uma importante fonte de informação estrutural. De fato, as correlações estatísticas entre sítios de aminoácidos em sequências de proteínas são a base dos métodos mais modernos para a previsão de contatos inter- e intra-proteínas, predição de estrutura tridimensional, identificação de sítios funcionais e resíduos determinantes de especificidade, inferência de interações entre parálogos, entre outras aplicações. Em consonância com isso, o presente trabalho apresenta um conjunto de resultados teóricos sobre como proteínas parceiras específicas podem ser recuperadas com base apenas nas informações da sequência. No primeiro capítulo, é realizada uma decomposição da informação mútua (MI) presente nos complexos proteína-proteína, considerando a hipótese de que a MI em proteínas se origina de uma combinação de diferentes fontes: coevolutiva, evolutiva e estocástica. Foi observado que a interface contém, em média por contato, mais informações do que o restante do complexo protéico, resultado que se mantém quando se considera tanto a MI de Shannon quanto a de Tsallis como medida de informação. Essa observação levou à conclusão de que a interface contém o sinal de informação mais forte para distinguir o conjunto correto de proteínas parceiras em famílias de proteínas que interagem. Com base nisso, a utilidade de usar a MI armazenada em interfaces proteína-proteína para recuperar o conjunto correto de proteínas parceiras é avaliada no segundo capítulo. Um algoritmo genético (GA) foi desenvolvido para explorar o espaço de possíveis concatenações entre um par de famílias de proteínas que interagem usando a MI da interface como função objetivo. Usando o GA, a maximização da MI da interface foi realizada para 26 pares de famílias de proteínas que interagem e foi observado que concatenações otimizadas correspondem a soluções degeneradas com duas fontes de erro distintas, decorrentes de pareamentos errados entre (i) sequências similares e (ii) não similares. Quando os erros cometidos com sequências semelhantes foram desconsiderados, as soluções do tipo (i) apresentaram taxas de verdadeiros positivos (TP) de 70 % - muito acima das mesmas estimativas para soluções do tipo (ii). Esses resultados se mantêm quando as otimizações são feitas com base na MI de Tsallis. Essas descobertas levantam questões sobre os mecanismos por trás da coevolução de proteínas parceiras e ajudam a racionalizar os dados da literatura que mostram uma forte deterioração das taxas de TP com o aumento do número de sequência em abordagens baseadas em MI.

Contents

List of Figures

List of Tables

List of Abbreviations

CASP: Critical Assessment of protein Structure Prediction

DCA: Direct Coupling Analysis

GA: Genetic Algorithm

HK: Histidine Kinase

IPA: Iterative Pairing Algorithm

MI: Mutual Information

MSA: Multiple Sequence Alignment

PDB: Protein Data Bank

RR: Response Regulator

TP: True Positive

Glossary

Amino acids (or residues): fundamental units that constitute proteins.

Amino acid site: structural position in a family of homologous proteins.

Coevolution model: set of preferential affinities that associate a set of proteins from family A to a set of proteins from family B.

Gene: functional unit of genetic material that can be transcribed and translated, giving rise to a protein.

Genome: set of genes (or genetic material) of a given organism.

Homologous proteins: proteins that have a common ancestor and, in general, perform the same function in different species.

Interaction interface: structural region where the contacts between the proteins of a protein complex are formed.

Multiple sequence alignment (MSA): alignment of a set of primary sequences of homologous proteins, so that each column corresponds to a specific structural site.

Natural selection: process that determines greater reproductive success of individuals better adapted to the environment.

Paralogous proteins: proteins that have been duplicated within the genome of the same species.

Phylogenetic tree: dendrogram that represents the evolutionary relationships of a group of organisms.

Primary structure (or primary sequence): linear sequence of amino acids that compose a protein.

Protein (peptide or polypeptide): biological macromolecule consisting of a chain of amino acids folded into a specific three-dimensional structure.

Protein complex: system composed of two or more interacting proteins.

Selective pressure: environmental factors that favor the occurrence of certain characteristics in a population.

Chapter 1

Introduction

Biological systems, as conceived under Van Valen's Red Queen hypothesis [4], exist in a dynamic equilibrium. In the book *Alice through the looking glass*, by Lewis Carroll, the Red Queen says "it takes all the running you can do, to keep in the same place" - this is how functional interactions remain stable over the years. When two species interact in nature, several factors contribute to coupled evolutionary changes. Modifications in one species generate selective pressures that lead to modifications in the other species, resulting in maintained affinity. In this way, it is possible for an interaction to continue over thousands of years.

The concept of coevolution, formulated by Ehrlich and Raven in the 1960s [5], has been subject of discussion over the years. In a simplified way, coevolution is a process characterized by reciprocal evolutionary changes between interacting species, guided by natural selection. This process, however, can not be observed, as it occurs slowly over thousands of years. Thus, Ehrlich and Raven ask themselves "without recourse to long-term experimentation on single systems, what can be learned about the coevolutionary responses of ecologically intimate organisms?". In fact, there is no way to directly uncover the evolutionary history of species, but the coevolutionary process leaves marks on biological systems, from which the history can be reconstructed.

Coevolution is observed not only in macroscopic systems, defined by prey-predator or host-parasite interactions, but also in microscopic systems, defined by interactions between macromolecules. Just as a parasite undergoes changes in its morphological characteristics in order to maintain affinity for its host over the years, a biological molecule undergoes changes in its composition [6, 7], in order to maintain affinity for its molecular partners [8, 9]. Among these microscopic systems are interacting protein partners, which will be investigated in the present work.

Interactions between proteins are at the basis of their molecular function. The multiple amino acid substitutions that can occur over time at interaction interfaces have the potential to modify binding specificity and are key to the evolutionary changes that occur in networks of protein-protein interactions [10–15]. When two sites are in direct contact, a mutation at one site can modify the selective pressures acting on the other site, and vice versa. Then, what exactly determines the affinity between two proteins?

The establishment of a protein complex is a spontaneous process that configures a change of state in a system composed by two or more components. This change in state occurs only when associated with favorable (negative) Gibbs free energy, which, in the case of protein interaction, is called binding free energy. In the bound state, protein complexes are characterized by contacts between amino acids at the interface [8, 9] (Figure 1.1A). Therefore, to verify if a given interaction is favorable, it is necessary to analyze which sites are in contact and what is the quality of these contacts, *i.e.* the physicochemical compatibility of the residues. Several studies show that amino acid sites in the three-dimensional structure of proteins tend to restrict possible changes in neighboring sites, *i.e.* nearby amino acid sites coevolve [16–20].

To perform a coevolutionary analysis on amino acid sites, however, it is necessary to sample a sufficiently large and diverse number of interacting homologs of proteins A and B (Figure 1.1C). Since it is not possible to extract this information from databases of 3D structures, as many proteins do not yet have a determined structure, analyses should be performed based solely on primary sequences (Figure 1.1B). But is it really possible to obtain useful and reliable structural information from primary sequences alone? In principle, yes, but the study of protein-protein interactions based solely on primary sequences is a complex problem that requires sophisticated theoretical approaches.

In the decade of 1990, Valencia *et al.* pioneer work established the utility of cor-

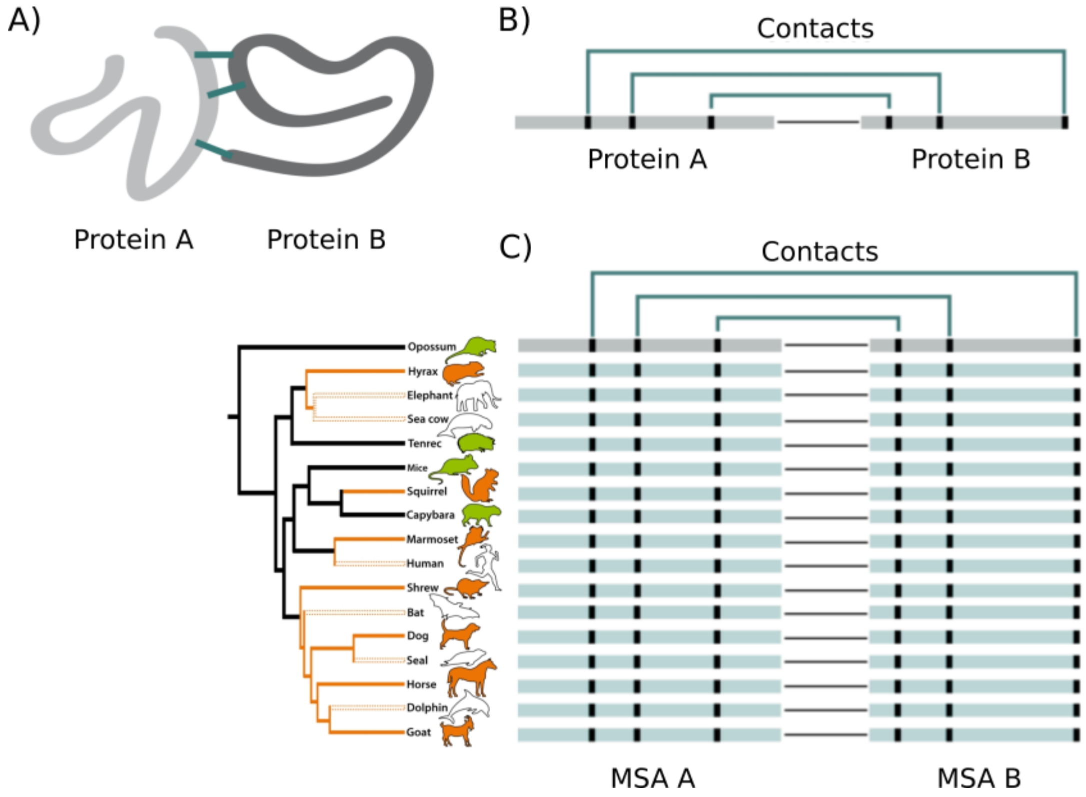


Figure 1.1: A) Protein complex formed by protein A and protein B. Contacts at the interaction interface are being shown in turquoise. Proteins are folded into their tertiary (three-dimensional) structure. B) Primary (linear) structure of proteins A and B. Interface contacts are being shown in turquoise. C) Multiple sequence alignment (MSA) of proteins A and B homologs, which are concatenated to form a set of pairwise interactions. On the left side, the phylogenetic tree of the species in whose genomes are the genes encoding the aligned proteins is shown.

related mutations to the inference of both intra- and inter-protein contacts [21, 22]. The basic idea was that statistically correlated positions in a MSA correspond to coevolving sites that should be close in the 3D protein structure. In this same decade, the Critical Assessment of protein Structure Prediction (CASP) competition [23] was inaugurated to promote further advances in the field of protein structure prediction in the following years. Afterwards, statistical methods were developed to disentangle direct from indirect correlations, contributing to great advances in contact prediction [24, 25]. Finally, the most recent methods combine global statistical models with deep learning and attention mechanisms to obtain structural models of proteins extremely close to the experimentally determined structure [26]. In summary, structural predictions based on primary protein sequences are not only possible, but have reached a very advanced level of accuracy over the last 30 years.

Besides intra- and inter-protein contact prediction, a set of additional problems can be addressed using coevolutionary approaches [27, 28]. For instance, the problem of determining whether two protein families interact or not can be approached by assessing the similarity between their phylogenetic trees [2, 29]. Also, by looking at phylogeny and correlated mutations at the same time, it is possible to infer sets of specificity determining sites, which explain functional differences in protein subfamilies [30, 31]. More recently, the problem of paralog matching has being thoroughly investigated by Bitbol *et al.* and Gueudré *et al.*, who developed coevolutionary methods to infer the correct set of interacting paralogs within bacterial genomes [3, 32, 33].

In this work, a theoretical characterization of protein interactions will be carried out using Information Theory concepts, statistical models and optimization algorithms. In Chapter 2, a decomposition of the mutual information (MI) present in protein-protein complexes is carried out, considering the hypothesis that MI in proteins is originated from a combination of coevolutive, evolutive and stochastic sources. It was observed that the interface contains on average, by contact, more information than the rest of the protein complex, a result that holds when considering both Shannon and Tsallis MI as a measure of information. This observation led to the conclusion that the interface contains the strongest information signal for distinguishing the correct set of protein partners in interacting protein families.

Building on that, the utility of using MI encoded on protein-protein interfaces to recover the correct set of protein partners is assessed in Chapter 3. A genetic algorithm (GA) was developed to explore the space of possible concatenations between a pair of interacting protein families using the interface MI as objective function. Using the GA, interface MI maximization was performed for 26 different pairs of interacting protein families and it was observed that optimized concatenations corresponded to degenerate solutions with two distinct error sources, arising from mismatches among (i) similar and (ii) non-similar sequences. When mistakes made among similar sequences were disregarded, type-(i) solutions were found to resolve correct pairings at best true positive (TP) rates of 70% - far above the very same estimates in type-(ii) solutions. These results hold when the optimizations are made based on Tsallis MI. These findings raise further questions about the mechanisms behind protein partners coevolution and help rationalize literature data showing a sharp deterioration of TP rates with increasing sequence number in MI-based approaches.

One of the central objectives of Biology is to understand the relationship between sequence, structure and function of proteins and how the evolution of these macromolecules takes place in the space defined by these three elements [34]. The present investigation helps clarifying to which extent it is possible to extract structural, functional and evolutionary information from primary protein sequences. The fact that this type of information can be extracted from primary sequences has had important implications for several areas in Biology, e.g. Synthetic Biology (engineering of protein compounds), and Systems Biology (study of protein-protein interaction networks). I hope the present work will bring further advances to this relevant field.

Chapter 2

Decomposition of Mutual Information in Protein Complexes

2.1 Introduction

Proteins are essential for virtually all cellular processes. Therefore, these macromolecules are selected to be thermodynamically stable [6, 7] and kinetically accessible [35–37] in a given conformation. Additionally, protein partners coevolve to maintain the stability of their bound state [8, 9]. The process of protein coevolution translates into a series of primary sequence variants containing coordinated compensatory substitutions, which can be extracted from a multiple sequence alignment (MSA).

In fact, correlated mutations have been used in several contexts for inference of structural information [3, 16, 21, 24, 26, 38–53]. In this regard, a relevant fact has been pointed out a few years ago: the signal extracted from correlated mutations in MSAs is likely originated from a couple of different sources [54]. In past mutual information (MI)-based studies, however, the observed signal was never decomposed in a clear manner, which means that a characterization of the isolated contributions of coevolutive, evolutive and stochastic information to the total MI is still missing.

In this chapter, the MI signal stored in protein-protein interactions originating from each of the aforementioned sources will be quantified. More specifically, assuming that the coevolutive information is likely stored in amino acids that are in physical contact at the interface (*i.e.*, with less than 8Å distance between carbon betas), the MI will be dissected in terms of compensatory substitutions in physically coupled and non-coupled amino acids. This will be done considering two different statistical frameworks: Shannon statistics, and Tsallis statistics. The latter has not yet been explored in past investigations, but has potential to be even better than the former for modelling protein coevolution.

2.1.1 Objectives

General objective

The general objective of this chapter is to characterize the role of different sources of mutual information (coevolutive, evolutive and stochastic) in protein-protein interactions.

Specific objectives

The specific objectives of this chapter are the following:

1. characterize interacting positions in the interface of protein complexes in terms of information content, comparing them to more structurally distant pairs of positions;
2. decompose the mutual information between interacting proteins into coevolutive, evolutive and stochastic information;
3. compare the results obtained using Shannon and Tsallis statistics.

2.2 Theory and Methods

In this section, some basic theoretical concepts and formulations relevant for this chapter will be introduced. First, an overview of how Information Theory is applied in the theoretical field of protein coevolution is given, followed by an introduction of Shannon and Tsallis statistics. Then, a novel framework for mutual information decomposition is presented, and, finally, the protein systems investigated in this chapter are shown. To ease the reading and understanding of the theoretical formulation, a unified notation is used, which may differ from the notations used in the reference texts. The notation pattern used will be briefly described in the following.

Stochastic or random variables will be represented in capital letters (*e.g.* X and Y), and blocks of these variables will be represented by capital letters with the size superscripted (*e.g.*, $X^N = (X_1, .., X_N)$ and $Y^N = (Y_1, ..., Y_N)$). The realization of a stochastic variable will be represented by a lowercase letter (*e.g.*, x is the realization of X and x^N is the realization block of X^N). In sums, when only one subscript variable (or set) is indicated, it means that the sum runs over all possible values of that variable (or all values contained in that set), *e.g.*, $\sum_i x_i$ is the sum of x_i for all values from i and $\sum_A a$ is the sum of all values $a \in A$. The set containing all possible values of a variable x will be denoted by $\{x\}$.

Probability distributions will be denoted by ρ, with $\rho(x)$ being the probability of a stochastic variable X assuming the value x, $\rho(x_i)$ the probability of X_i assuming the value value x_i, $\rho(x^N)$ the probability of the block X^N assuming the values x^N and $\rho(x, y)$ the probability of X assuming the value x and Y assuming the value of y simultaneously (joint probability). The set containing the probabilities of all the values that X can assume will be denoted by $\rho(X) = \{\rho(x)\}$. The frequency of a symbol a in the i-th column of a symbol array will be denoted by $f_i(a)$ and the joint frequency of a and b in the i-th and j-th columns of an array of will be denoted by $f_{ij}(a, b)$. The frequency of a given pair (x, y) in a pair of columns index by i will be denoted by f_{x_i,y_i}.

2.2.1 Information Theory and its applications in protein coevolution

In his classic work published in 1949 [55], Shannon states that the fundamental problem of communication is to reproduce, at one point, exactly or in an approximate way, a message emitted at another point. Ignoring the semantic aspects, it is interesting to note that a given message is selected from a set of possible messages. Therefore, a communication system must be able to operate considering each possible messages, and not just the one that will actually be transmitted, as this information is unknown at the time the system is being designed.

If the number of messages in the set is finite, then that number, or any monotonic function of that number, can be considered a measure of the information generated when a message is chosen, with all possibilities being equally likely. For informational analysis, the most natural choice of function is the logarithmic function which, in addition to being more intuitive, is more convenient for mathematical reasons. The choice of the logarithmic basis corresponds to the choice of the information measurement unit. If the base used is base 2, for example, the resulting unit is *bit*. If the natural logarithm is used, the resulting unit is called *nat*.

In the context of evolutionary protein analysis, it can be said that within a given family each amino acid site in the primary sequence represents a source of evolutionary information. The analysis of multiple sequence alignments (MSA) is usually done by extracting the site-specific frequencies of amino acids, *i.e.*, the frequency distribution of each MSA column (Figure 2.1). The nature of this distribution is, in general, informative about the evolutionary pressure that is acting on a certain site [56]. For example, if all members of a family of globular proteins present a conserved Cysteine in a certain site, there is strong evidence that this residue plays an important functional role or that it is necessary for the maintenance of the globular structure [35–37].

The conservation of a residue at a given site in the sequence is related to the entropy of that site [57–60], *i.e.*, the amount of information of that source. A very conserved site has low entropy value, while a variable site has high entropy value. The entropy H of a given source i, represented by the stochastic variable X_i, in a family of homologous

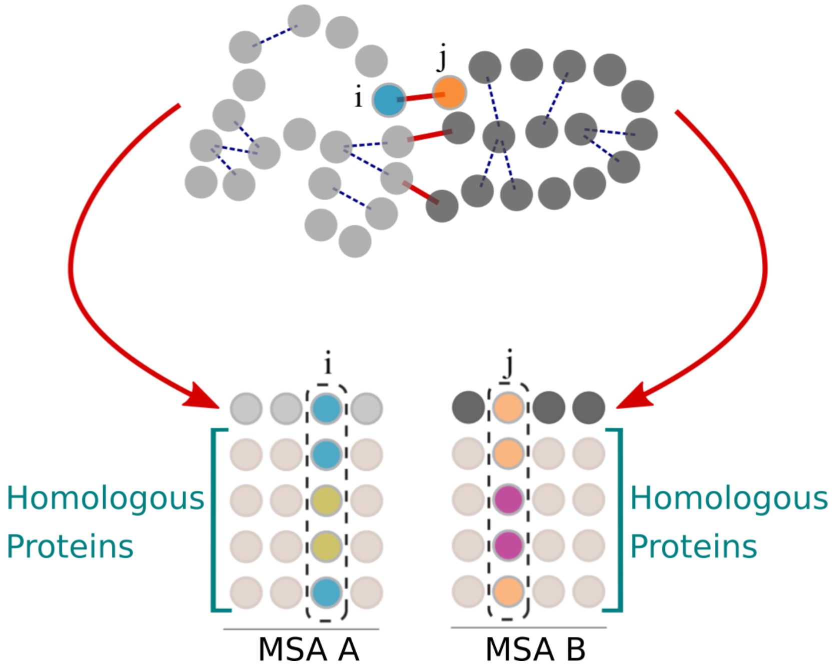


Figure 2.1: Information sources in globular proteins. The distribution of residues in sites i and j is being shown in the multiple sequence alignments (MSA). Above, proteins are represented in their tertiary structure, with indications of chemical interactions between residues within the same protein (in blue) and between proteins (in red).

proteins, is the negative sum of the individual frequencies of the residues that appear in that site multiplied by their logarithmic function, that is:

$$H(X_i) = -\sum_{A} f_i(a)\ln(f_i(a)) \tag{2.1}$$

where A represents the alphabet of all amino acids and $f_i(a)$ represents the frequency of amino acid $a \in A$ in source i, *i.e.*, the probability of the stochastic variable X_i assuming a value $x_i = a$. The joint entropy of the sources i and j, represented by the variables X_i and Y_j, is calculated from the joint probabilities of amino acid pairs:

$$H(X_i, Y_j) = -\sum_{A\times A} f_{ij}(a,b)\ln(f_{ij}(a,b)) \tag{2.2}$$

where $A \times A$ represents the set of all amino acid pairs and $f_{ij}(a,b)$ represents the frequency of the amino acid pair $(a,b) \in A \times A$ in sources i and j.

Further information can be extracted from the joint distribution of frequencies, considering pairs of sites instead of individual ones. In the native state, each residue in the polypeptide chain performs residue-residue interactions. The mutations that occur in a given site are, therefore, dependent on the physicochemical nature of the neighboring residues. As a result, the frequency distributions of amino acids at different sites must be dependent on each other. The detection of these correlations has the potential to clarify the physical interactions that define the native structure [21,24,25,49]. One way to detect correlation between pairs of amino acid sites is by calculating the mutual information, I, between two sources i and j, which can be defined as:

$$I(X_i;Y_j) = \sum_{A\times A} f_{ij}(a,b) \ln\left(\frac{f_{ij}(a,b)}{f_i(a)f_j(b)}\right) \tag{2.3}$$

In equation (2.3), the mutual information reaches its lower limit of zero if X_i and Y_j are independent. In the case where the variables are perfectly correlated, the mutual information has a maximum that cannot exceed the entropy of any of the sources, $H(X_i)$ and $H(Y_j)$ (Figure 2.2).

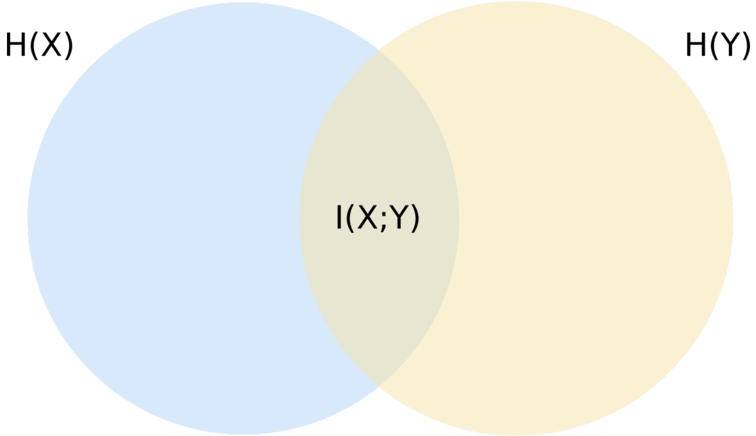


Figure 2.2: Venn diagram showing the relationship between the entropy of two stochastic variables, $H(X)$ and $H(Y)$, and the mutual information between these two variables, $I(X;Y)$.

2.2.2 Shannon statistics and the Entropy Maximization Principle

Information Theory makes it possible to recover probability distributions based on partial knowledge, through a type of statistical inference called Maximum Entropy estimation. The Maximum Entropy estimate is the least biased estimate possible considering the information given. If we consider Statistical Mechanics as a form of statistical inference instead of a physical theory, it is possible to show that its classical computation rules, starting with the determination of the partition function, are an immediate consequence of the Entropy Maximization Principle [61].

Let X be a stochastic variable that assumes discrete values $x \in \Omega$. Suppose we don't know the probabilities $\{\rho(x)\}$ corresponding to each of the states, but we do know the expected value of a function $f(X)$

$$\langle f(X) \rangle = \sum_x \rho(x) f(x) \tag{2.4}$$

Based on this information, what would be the expected value of a function $g(X)$? At first glance, the problem does not seem to have a solution, as the information provided is insufficient to determine the probabilities $\{\rho(x)\}$. Equation (2.4) and the normalization condition

$$\sum_x \rho(x) = 1 \tag{2.5}$$

would have to be supplemented by additional $(n-2)$ conditions so that $\langle g(X) \rangle$ could be found.

The breakthrough provided by Information Theory is the discovery of a unique and unambiguous criterion to determine the "amount of uncertainty" represented by a discrete probability distribution, in line with our intuitive notion that a wider distribution represents a greater degree of uncertainty than one with a conspicuous peak. This quantity, which is positive, increases with the degree of uncertainty, and is additive for independent sources of uncertainty, is known as Shannon's entropy:

$$H(\rho(X)) = -\sum_x \rho(x) \ln(\rho(x)) \tag{2.6}$$

Now, to make inferences based on partial information, we need to find the probability distribution that has maximum entropy given the known information. This is the only unbiased estimate that can be made, *i.e.*, any other information that is incorporated into the model would be arbitrary, making it biased. To maximize equation (2.6) subject to the restrictions (2.4) and (2.5), we just have to introduce the Lagrange multipliers λ_0, λ_1 in the usual way. So, we have the Lagrangian

$$L(\rho(X), \lambda_0, \lambda_1) = -\sum_x \rho(x)\ln(\rho(x)) - \lambda_0(\sum_x \rho(x) - 1) - \lambda_1(\sum_x \rho(x)f(x) - \langle f(X)\rangle) \tag{2.7}$$

Making $\nabla L = 0$, we get

$$\rho^*(x) = e^{-\lambda_0 - \lambda_1 f(x)} \tag{2.8}$$

The constants λ_0, and λ_1 are determined by substituting in equations (2.4) and (2.5). The result can be written as

$$\langle f(X)\rangle = \frac{\partial}{\partial \lambda_1} \ln Z(\lambda_1) \tag{2.9}$$

$$\lambda_0 = \ln Z(\lambda_1) \tag{2.10}$$

where

$$Z(\lambda_1) = \sum_x e^{-\lambda_1 f(x)} \tag{2.11}$$

is called the partition function.

Now it is possible to rewrite $\rho^*(x)$ as

$$\rho^*(x) = \frac{e^{-\lambda_1 f(x)}}{Z(\lambda_1)} \tag{2.12}$$

which is a *Boltzmann-like* distribution. These results can be generalized for any number of $f(X)$ functions

$$\rho^*(x) = \frac{e^{-\lambda_1 f_1(x) - \ldots - \lambda_m f_m(x)}}{Z(\lambda_1, \ldots, \lambda_m)} \tag{2.13}$$

where

$$Z(\lambda_1, \ldots, \lambda_m) = \sum_x e^{-\lambda_1 f_1(x) - \ldots - \lambda_m f_m(x)} \tag{2.14}$$

Modern approaches for contact prediction, such as Direct Coupling Analysis (DCA) [25], rely on Shannon Entropy Maximization. This, however, might not be the only framework possible. In the following, a generalization of Shannon statistics is presented, which might constitute a possible alternative to be used for sequence-based structural predictions in the future. As such, Tsallis statistics will be presented and regarded in comparison to

Shannon statistics throughout this work.

2.2.3 Tsallis statistics and q-exponential distributions

It is known that the statistical properties of the steady state of some complex systems are well described by q-exponential distributions [62]. These distributions are anomalous distributions from the point of view of conventional Statistical Mechanics, characterized by the Boltzmann exponential factor. The explicit form of a q-exponential distribution is as follows:

$$\rho(x) = \frac{1}{Z_q(\lambda)} e_q(-\lambda x), (x \in \Omega) \tag{2.15}$$

$$Z_q(\lambda) = \sum_x e_q(-\lambda x) \tag{2.16}$$

where Ω is the set of states accessible for a given system, $x \in \Omega$ is realization of a variable X, λ is a factor related to the Lagrange multiplier, and $e_q(t)$ is the q-exponential function

$$e_q(t) = \begin{cases} [1 + (1-q)t]^{1/(1-q)}, [1 + (1-q)t] > 0 \\ 0, \qquad\qquad\qquad\quad [1 + (1-q)t] \leq 0 \end{cases} \tag{2.17}$$

where q is a positive real number called entropic index. At the limit $q \to 1$, the q-exponential function converges to the ordinary exponential function and the q-exponential distribution for the Boltzmann distribution.

An upper limit for the Tsallis entropy [63] can be obtained as follows

$$H_q^{(T)}(\rho(X)) \leq \frac{1}{1-q} \sum_x [(\rho(x))^q - 1] \tag{2.18}$$

where X is a random variable, and q is entropic index ($q > 1$). The reason why this way of calculating the Tsallis entropy provides an upper limit instead of the exact value is due to its subadditive property when $q > 1$. In fact, this property might pose a challenge to working within the Tsallis statistics framework.

Here, Tsallis statistics will be used to calculate mutual information

$$I_q^{(T)}(\rho(X);\rho(Y)) \leq \frac{1}{1-q}\sum_{(x,y)}\rho(x,y)^q\left[\left(\frac{\rho(x,y)}{\rho(x)\rho(y)}\right)^{(1-q)}-1\right] \tag{2.19}$$

where X and Y are random variables, and q is entropic index ($q > 1$). To facilitate calculations, the particularity of subadditivity will be ignored in this work, and the exact value of $I_q^{(T)}$ will be considered approximately equal to its upper limit for $q > 1$.

The work of Conte et al. [35] shows that protein interfaces contain in average around 10 hydrogen bridges between residues. This means that, when considering the full range of possible amino acid site pairs possible for a given protein complex, it would be expected that only a very small number of these pairs present strong correlation, while a much bigger number of sites would present no correlation. Therefore, it can be expected that the pairwise correlation in the space of possible site pairs will decrease respecting either an exponential distribution or a power-law distribution. While Shannon statistics model the former case, Tsallis statistics models the latter (when $q > 1$).

Furthermore, in a letter written in 2008 [64], Bercher shows that the Tsallis distribution can be derived from a Maximum Entropy Shannon distribution, incorporating a constraint in the divergence between the distribution in question and another distribution that can be imagined as it's tail. This tail is identified as a power-law distribution. Therefore, we see that an important feature of the Maximum Entropy Tsallis distribution is its ability to model the tail of distributions with long tails. The natural selection process, in turn, has been modelled as a power-law in works on the evolution of microbial populations [65]. For this reason, Tsallis statistics might be more accurate than Shannon statistics for modelling the natural selection process.

2.2.4 Mutual information decomposition

It is known that mutual information (MI) in protein complexes can be derived from different sources, the main ones being coevolution, evolution and stochastic [54]. Here, it is considered that the average MI per contact contained in the interface of a complex,

$I_{d_c \leq 8}/N_{d_c \leq 8}$, can be decomposed as follows:

$$I_{d_c \leq 8}/N_{d_c \leq 8} = I_{coev} + I_{evol} + I_{rand} \tag{2.20}$$

where I_{coev} is the portion derived from coevolution, I_{evol} is the portion derived from evolution and I_{rand} is the portion derived from stochasticity.

The average stochastic MI per contact can be easily estimated from the average of the MI for different stochastic coevolution models, z_{rand}

$$I_{rand} = \frac{I_{d_c > 8}(X; Y \mid z_{rand})}{N_{d_c > 8}} \tag{2.21}$$

where the so called stochastic coevolution models z_{rand} are scrambled concatenations between the MSAs of the two interacting protein families (Figure 2.3), while the native coevolution model z^* is the correct concatenation between the MSAs.

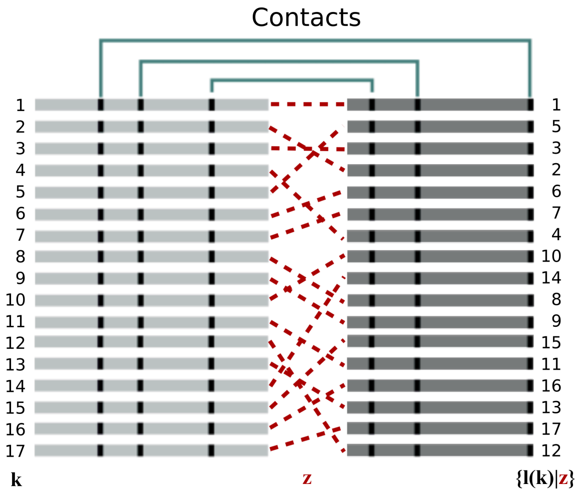


Figure 2.3: Stochastic coevolution model. Each sequence l in MSA B is randomly concatenated to sequence k in MSA A in a unique arrangement $\{l(k)|z\}$.

Now, taking as a premise the fact that coevolutive information is mostly stored in the complex interface at physically-coupled sites, *i.e.*, sites with less than 8Å between carbon betas (Figure 2.4), we can estimate I_{coev} average per contact as follows:

$$I_{coev} = \frac{I_{d_c \leq 8}(X; Y \mid z^*)}{N_{d_c \leq 8}} - \frac{I_{d_c > 8}(X; Y \mid z^*)}{N_{d_c > 8}} \tag{2.22}$$

Finally, to estimate I_{evol} average per contact, we do:

$$I_{evol} = \frac{I_{d_c>8}(X;Y \mid z^*)}{N_{d_c>8}} - \frac{I_{d_c>8}(X;Y \mid z_{rand})}{N_{d_c>8}} \qquad (2.23)$$

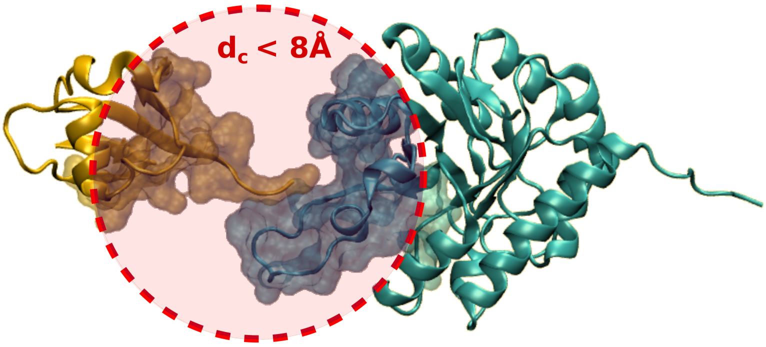


Figure 2.4: Thiazole synthase complex (in turquoise) / thiS (in yellow) (PDBID: 1TYG) with emphasis on the interaction interface. The distance cutoff for defining contacts between sites is shown. For example, given the cutoff $d_c < 8$Å, a set of contacts is obtained that contains all pairs of sites (i, j), with i in protein A and j in protein B, whose physical distance, $d(i, j)$, is less than 8Å (red circle).

As a comparison, in an analogous manner, the decomposition of the average Tsallis MI per contact was performed considering a value of $q = 1.75$. A lower value of $q = 1.50$ and higher value of $q = 2.00$ were also tested to evaluate the influence of this parameter in the results.

2.2.5 Systems under investigation

Protein complexes under investigation are shown in Table 2.1. Paired MSAs for all protein families were obtained from Ovchinnikov *et al.* [51]. The systems considered here were selected from within [51]'s dataset, using as a criteria a total number of possible site pairs (i, j), N, smaller than 100,000 (see Table 2.1). Amino acid contacts defining the discrete stochastic variables X^N and Y^N were identified from the crystal structure of the bound state of a representative complex of proteins A and B using a typical contact definition considering maximum separation distance of 8Å between amino acids carbon beta (carbon alpha for Glycine).

Table 2.1: Protein complexes considered in the study. M is the number of sequences in the multi-sequence alignment, $N_{d_c \leq 8}$ is the number of contacts in the interface, following an 8Å cutoff definition, and N is the total number of contacts.

#	Description	PDB ID	Chains	M	$N_{d_c \leq 8}$	N
1	Dihydroorotate dehydrogenase B (4-mer)	1EP3	A, B	552	91	80,910
2	Thiazole synthase/ThiS (8-mer)	1TYG	A, B	746	80	15,730
3	3-oxoadipate coA-transferase (4-mer)	3RRL	A, B	1,330	161	47,610
4	Toxin-antitoxin complex RelBE2 (4-mer)	3G5O	A, B	904	92	7,452
5	Electron transfer flavoprotein (2-mer)	1EFP	A, B	1,347	229	75,522
6	Anthranilate synthase (4-mer)	1I1Q	A, B	1,204	91	94,535
7	4-hydroxybenzoyl-CoA reductase (6-mer)	1RM6	B, C	1,481	93	49,742
8	GTP-Regulated ATP Sulfurylase (2-mer)	1ZUN	A, B	649	140	77,618
9	TusBCD proteins (6-mer)	2D1P	B, C	216	40	11,305
10	Succinate:quinone oxidoreductase (4-mer)	2WDQ	C, D	221	43	12,705
11	GatCAB (3-mer)	3IP4	A, C	879	146	44,620
12	Allophanate Hydrolase (4-mer)	3MML	A, B	1,067	116	60,401
13	F1-ATP synthase (8-mer)	3OAA	H, G	886	179	39,192
14	DhaK-DhaL (4-mer)	3PNL	A, B	902	113	74,195

2.3 Results and Discussion

In this section, results regarding the decomposition of mutual information (MI) in protein-protein interactions will be presented and discussed. The main results will be shown in detail for the 14 protein complexes present in Table 2.1.

2.3.1 Mutual information across structural distances

Figure 2.5 shows average MI per distance bin regarding both the native coevolution model, $I(X_i; Y_j \mid z^*)$ in turquoise, and a stochastic coevolution model, $I(X_i; Y_j \mid z_{rand})$ in gray. All possible pairs of sites (i, j) were considered for each different protein complex. Distance bins, $d(i, j)$ in Ångström, are shown in the x-axis. The Pearson's correlation between the average MI per bin and the mean value of each distance bin is shown on the top of each plot for both native (r_{nat}) and random (r_{rnd}) models. The MI was calculated using both Shannon and Tsallis statistics.

It is easy to notice that the MI is always greater for the native model, $\sum_{i,j} I(X_i; Y_j \mid z^*) > \sum_{i,j} I(X_i; Y_j \mid z_{rand})$. In addition, pairs with shorter physical distances tend to keep more information on average (higher negative values of r_{nat}). Error bars tend to be very big, since most of the MI values are close to zero (in all distance bins). In smaller distance bins, which contain some high values of MI, the distribution of values is then widely spread. In addition, it is possible to notice that in Tsallis statistics the stochastic information (in gray) is close to zero and better values of r_{nat} are obtained in 11 out of 14 cases.

A curious detail observed is that the stochastic information (in gray) tend to correlate positively with distance $d(i, j)$ in many cases (see values of r_{rnd}). Given this fact, a similar plot was reproduced in Figure 2.6 showing the subtraction of native MI and scrambled (random) MI, $I(X_i; Y_j \mid z^*) - I(X_i; Y_j \mid z_{rand})$. This subtraction improves the value r_{nat} for many systems, specially when applying Shannon statistics. Even considering this subtraction, however, there are still a coupled of systems for which the MI is not inversely correlated with distance, namely systems 1I1Q_AB, 2WDQ_CD and 1ZUN_AB.

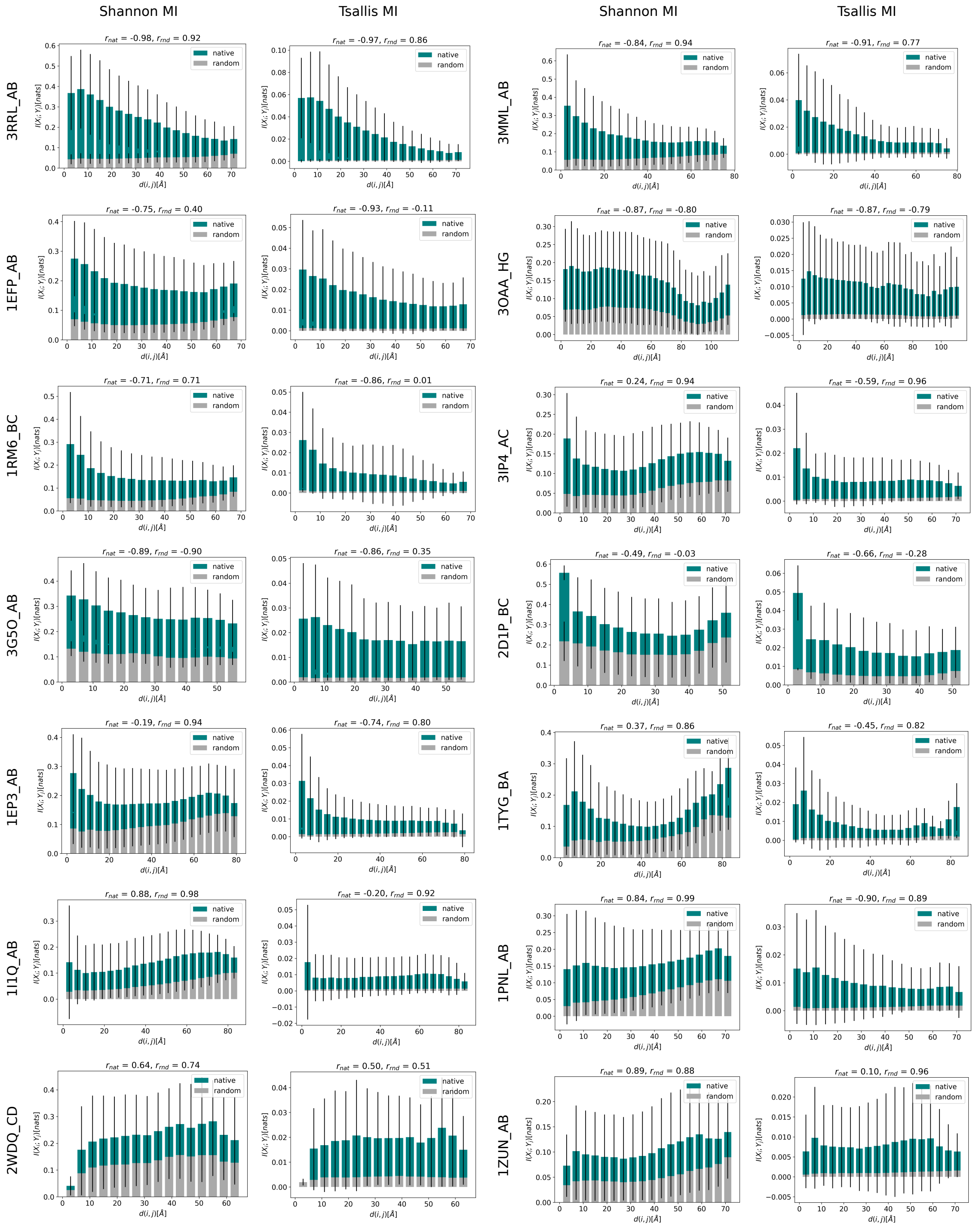


Figure 2.5: Mutual information for the native coevolution model, $I(X_i;Y_j \mid z^*)$ (in turquoise), and for a scrambled (random) coevolution model, $I(X_i;Y_j \mid z_{rand})$ (in gray), for all possible pairs (i,j) of sites in different protein complexes. On the x-axis are the distance bins in Ångström, $d(i,j)$. Results obtained using both Shannon and Tsallis ($q = 1.75$) statistics are shown.

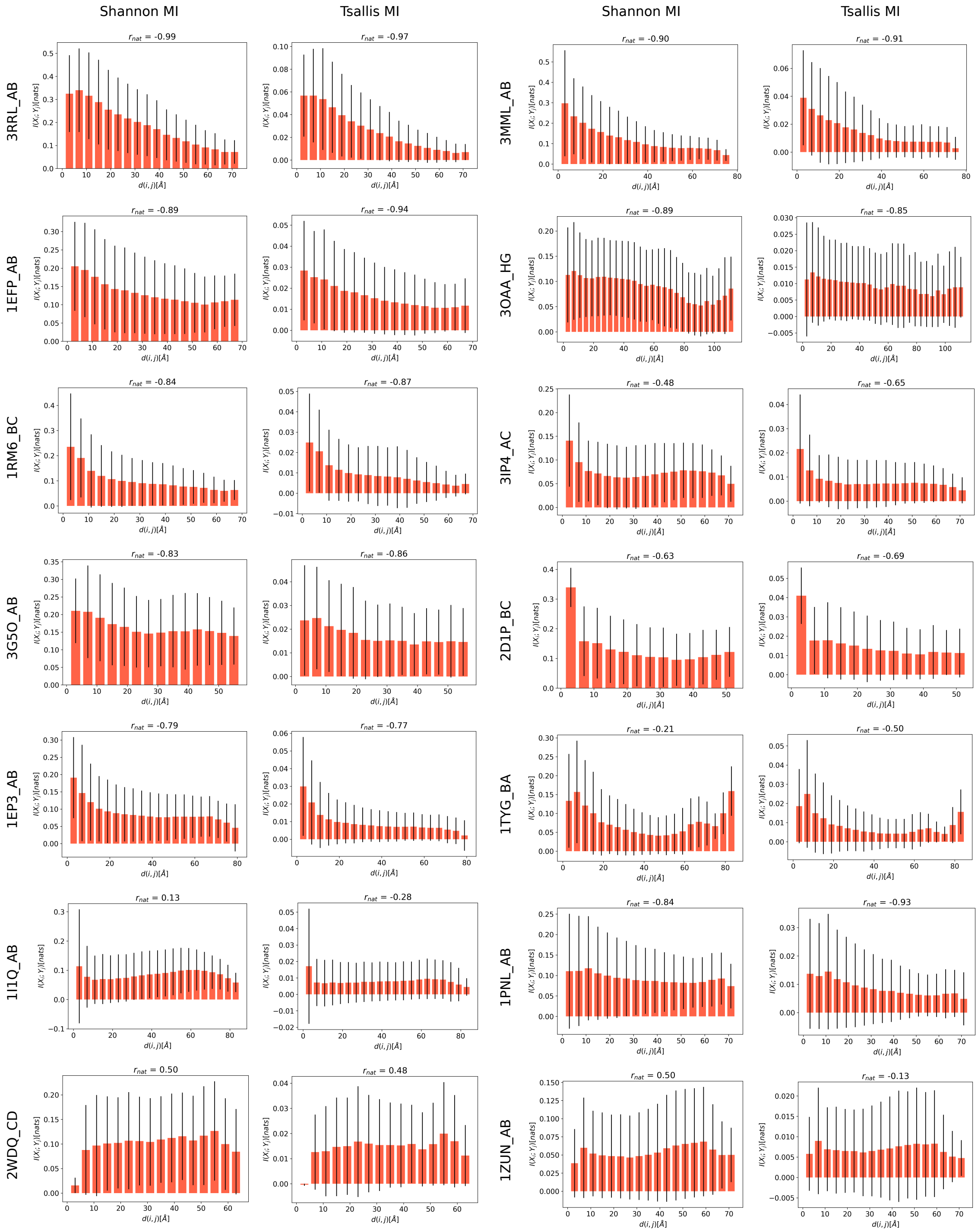


Figure 2.6: Subtraction of native mutual information and scrambled (random) mutual information, $I(X_i; Y_j \mid z^*) - I(X_i; Y_j \mid z_{rand})$, for all possible pairs (i, j) of sites in different protein complexes. On the x-axis are the distance bins in Ångström, $d(i, j)$. Results obtained using both Shannon and Tsallis ($q = 1.75$) statistics are shown.

To discard any influence caused by the Tsallis statistics parameter q in the results, different values q were tested (Figure 2.7). It is possible to observe that variations in the value of parameter q have a slight influence in the values of r_{nat} in some cases, but do not change the results qualitatively.

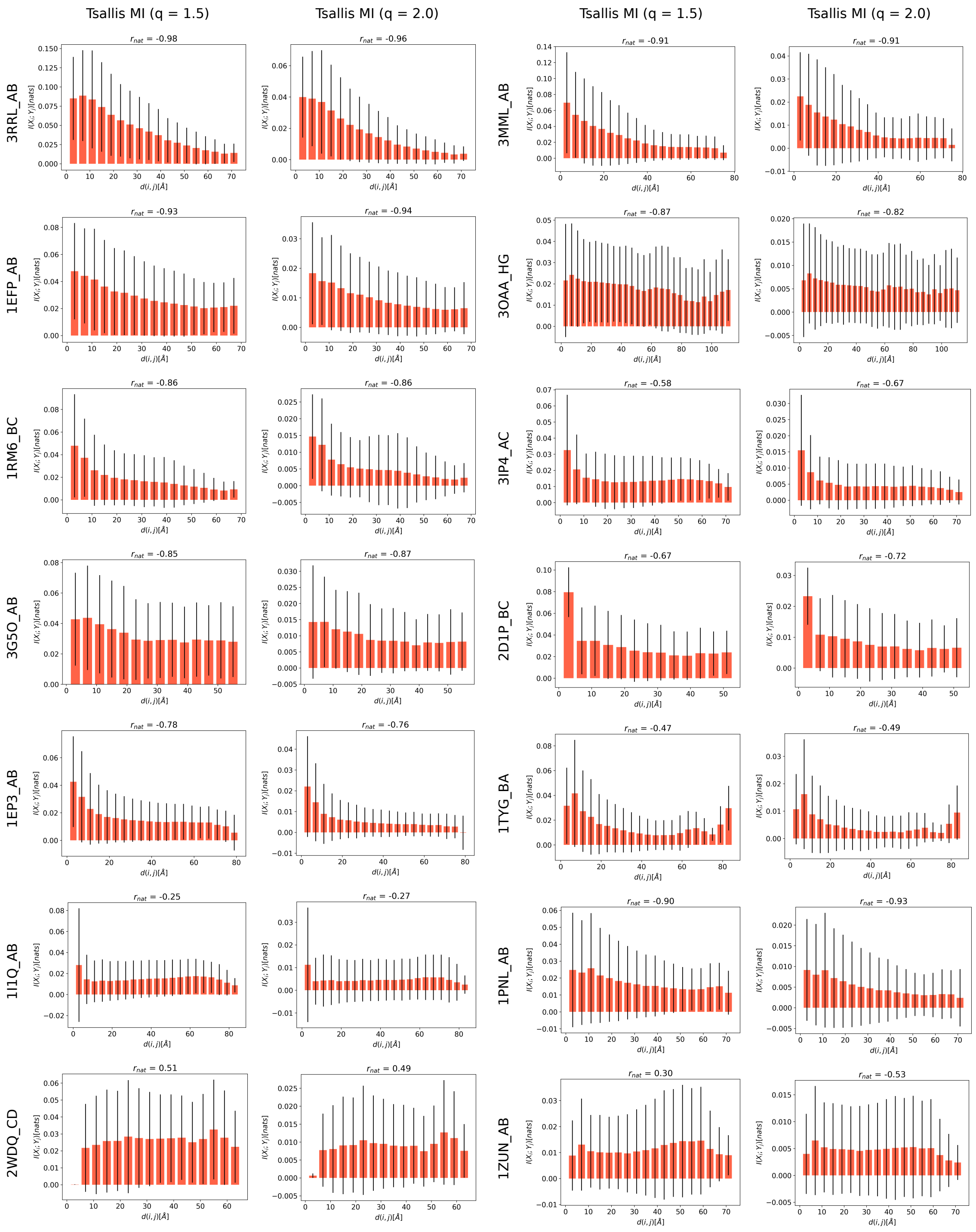


Figure 2.7: Subtraction of native mutual information and scrambled (random) mutual information, $I(X_i; Y_j \mid z^*) - I(X_i; Y_j \mid z_{rand})$, for all possible pairs (i, j) of sites in different protein complexes. On the x-axis are the distance bins in Ångström, $d(i, j)$. Results obtained using two alternative parameters ($q = 1.5$ and $q = 2.0$) for Tsallis' statistics are shown.

2.3.2 Coevolutive, evolutive and stochastic information

The values of I_{coev}, I_{evol} and I_{rand} were calculated as described in section 2.2.4 for all systems, using both Shannon and Tsallis statistics (Figure 2.8). For most of the systems studied, the average information content per contact in the interface, $N^{-1}_{d_c \leq 8} I_{d_c \leq 8}(X;Y \mid z^*)$ is greater than this same estimate for the rest of the protein, $N^{-1}_{d_c > 8} I_{d_c > 8}(X;Y \mid z^*)$. Since the raw information values obtained vary widely between systems, results are shown in Figure 2.9 in terms of fractions of the total MI. It is interesting to observe that there seems to be an inverse correlation between the coevolutive and evolutive information, which becomes especially clear when applying Tsallis statictics framework. In both cases, systems 2WDQ_CD, 3PNL_AB and 1EFP_AB figure among the ones with higher evolutive information, while systems 1RM6_BC, 1TYG_BA and 3MML_AB are among the ones with higher coevolutive information.

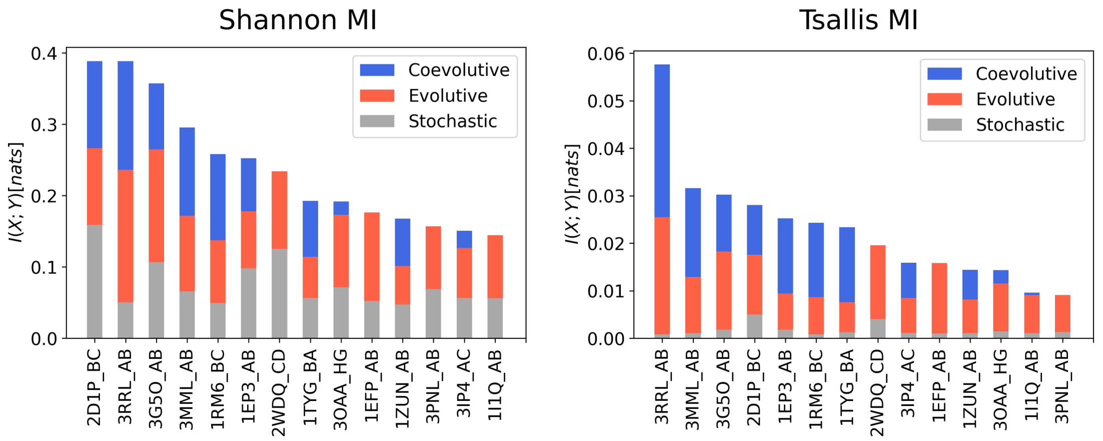


Figure 2.8: Raw values of coevolutive (in blue), evolutive (in orange) and stochastic (in gray) mutual information (MI), $I(X;Y)$, per system applying both Shannon and Tsallis ($q = 1.75$) statistics.

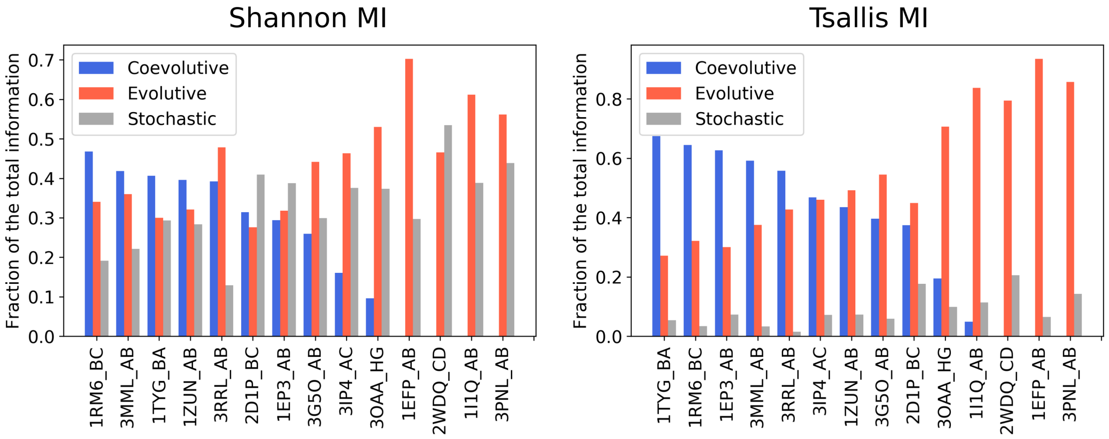


Figure 2.9: Relative values of coevolutive (in blue), evolutive (in orange) and stochastic (in gray) mutual information (MI), $I(X;Y)$, per system applying both Shannon and Tsallis ($q = 1.75$) statistics.

In systems 2D1P_BC, 1EP3_AB and 2WDQ_CD, the stochastic information is the one occurring in highest proportion when applying Shannon statistics. Curiously, these are the systems with the smallest number of sequences in the alignments. To test whether the small number of sequences in the alignment is related to high values of stochastic information, the Pearson correlation between these two values was calculated (Figure 2.10A). High values of correlation, $r = -0.77$ and $r = -0.81$ for Shannon and Tsallis statistics, respectivelly, were obtained, showing that the existence of stochastic information is probably related to effects of limited sample size (small number of sequences in the MSAs).

The relationship of both evolutive and coevolutive information with the mirror-tree correlation [2] between the two interacting families was also investigated (Figure 2.10B-C). The mirror-tree correlation, as a measure that captures the similarity between the phylogenetic trees of two MSAs, was calculates as follows:

$$r = \frac{\sum_{i=1}^{m}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{m}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{m}(B_i - \bar{B})^2}} \tag{2.24}$$

where $m = (M^2 - M)/2$, *i.e.*, the number of sequences in the triangular matrix of all-vs-all sequences, A_i are the elements of the distance matrix of MSA A, B_i is the corresponding value for MSA B, and $\bar{A}$ and $\bar{B}$ are the means of A_i and B_i, respectively. Interestingly, the evolutive information presented a positive correlation with values of mirror-tree correlation, while the coevolutive information did not. This observation corroborates the hypothesis that evolutive and coevolutive information are, in fact, distinct and non-clearly related quantities.

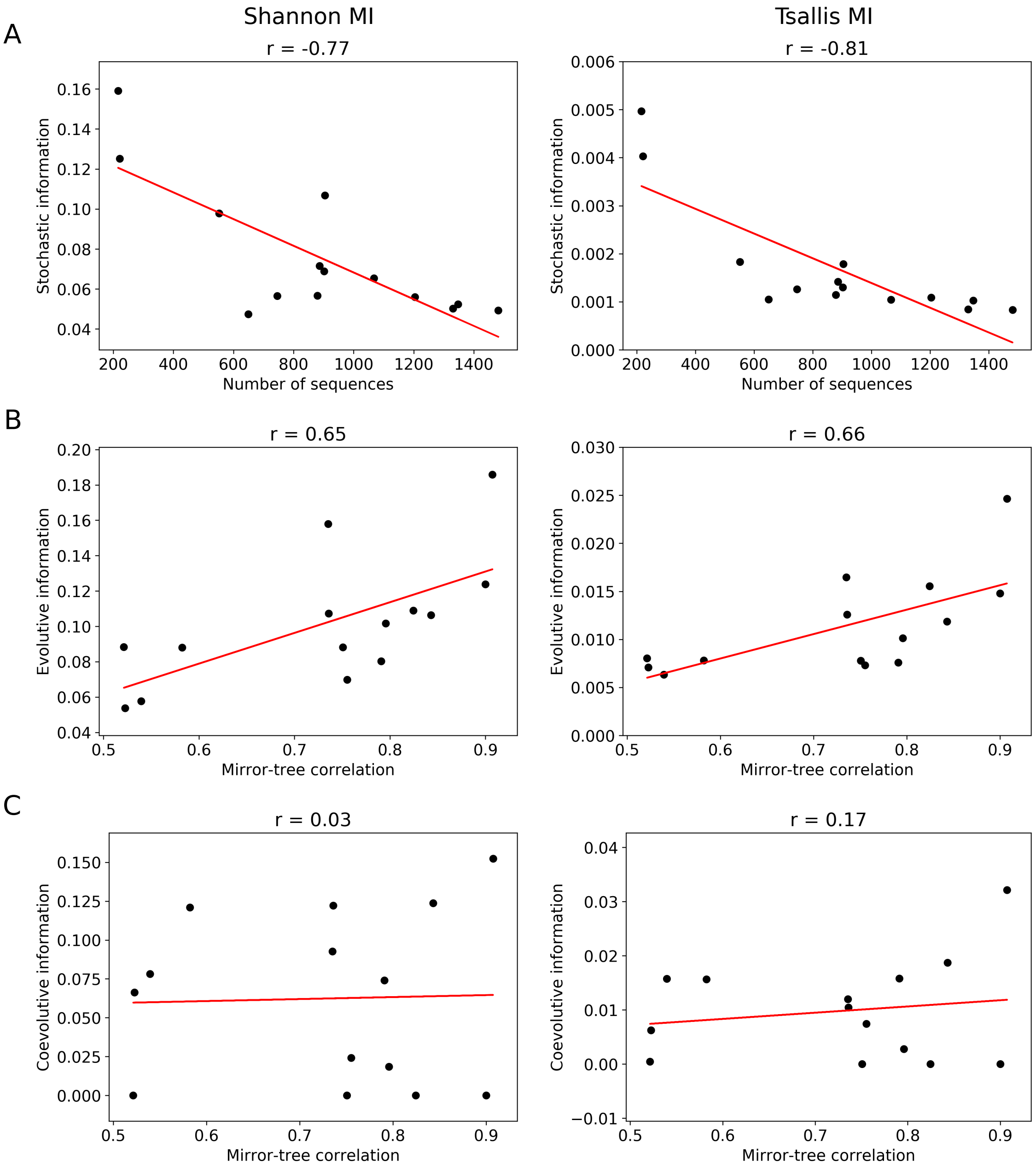


Figure 2.10: Correlation between the values of stochastic information and the number of sequences in the paired MSAs (A). Correlation between the evolutive (B) and coevolutive (C) information and the mirror-tree correlation. A total of 14 protein complexes were considered (n = 14). The values of mutual information (MI) were obtained using both Shannon and Tsallis statistics.

Finally, related to the results presented in this chapter, there is a work published in the *Computational and Structural Biotechnology Journal* [1], in which the contributions of coevolutive, evolutive and stochastic information in determining specific protein partners are investigated. We show that physically-coupled amino acid sites at short range distances store the largest per-contact MI content, with a significant fraction of that content resulting from coevolutive sources alone. The information stored in coupled amino acid sites is shown further to discriminate MSAs with the largest expectation fraction of protein partners matches (Figure 2.11) – a conclusion that holds against various definitions of intermolecular contacts and binding modes.

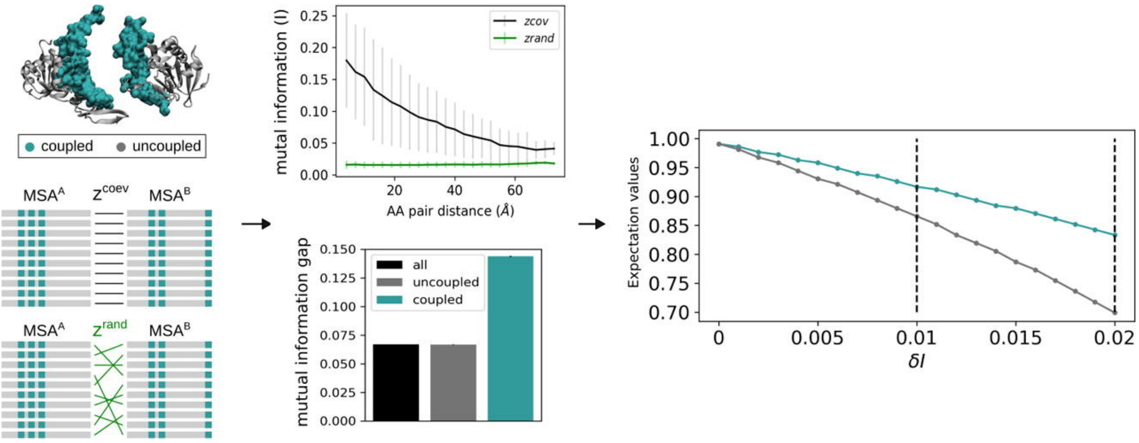


Figure 2.11: Graphic abstract of [1] showing that the mutual information stored in physically coupled amino acid sites contains the strongest signal to infer specific protein partners (highest expectation fraction of true positives).

2.4 Conclusions

In this chapter, the mutual information (MI) stored in protein complexes was decomposed according to three different possible sources: coevolution, evolution and stochasticity. First, it was shown that, as expected, the MI content of individual amino acid site pairs (i, j) tends to be anti-correlated with the structural distance $d(i, j)$. This leads to the observation that there is a surplus of MI stored in physically-coupled amino acids, which was hypothesized to come from coevolutive sources. It was then observed that different protein systems present a variable content of evolutive and coevolutive information. In addition, it was shown that evolutive information is correlated with the similarity between the interacting protein families phylogenetic trees (quantified by the mirror-tree method [2]), while the coevolutive information is not. This observation is evidence of the distinct character of these two quantities.

Moreover, two different theoretical frameworks were compared in all analyses performed: Shannon and Tsallis statistics. Overall, Tsallis statistics yielded more robust results, in the sense that, for example, MI values correlated better with structural distances. Also, within Tsallis statistics framework, the MI fraction resulting from stochastic sources was much smaller than the one observed within Shannon statistics framework. These results indicate that, despite issues raised by the subadditivity property of Tsallis statistics, it might be a suitable framework to be applied in sequence-based structural inferences.

Finally, it was shown that the information stored in coupled amino acid sites discriminates MSAs with the largest expectation fraction of protein partners matches. This means that, when compared to the informational content resulting from evolution at long-range interactions, the MI stored in physically-coupled amino acid sites is the strongest signal to distinguish partners in protein families with a shared evolutionary history. It is also likely the unique signal in case of molecular coevolution in independent genomes, as the evolutive information is expected to vanish for protein families that did not share an evolutionary history. In the next chapter, the utility of the interface MI for the inference of specific protein partners will be thoroughly investigated.

Chapter 3

Protein Partners Inference using Mutual Information-Based Approaches

3.1 Introduction

The coevolution process of two protein families translates itself into an *ensemble* of primary-sequence variants encoding coordinated compensatory mutations and therefore, a specific set of protein partners. There are, among others, two classic types of theoretical problems related to the inference of structural information based on sequences of interacting protein families: I - prediction of structural contacts from a paired MSA of proteins with know binding partners in the same genome; II - prediction of a set of preferential affinities from a pair of MSAs of proteins with unknown specific partners (Figure 3.1). The first type of problem has already been widely explored by several research groups, which have been able to accurately infer inter-protein structural contacts from paired MSAs using sophisticated theoretical approaches, thereby showing that there is a strong correlation between statistical coupling and structural coupling [25, 50, 51].

The problem of inferring specific protein partners based on MSAs, in turn, has received less attention over the years, with most of the works focusing on paralog matching. Ingenious approaches based on the correlation of phylogenetic trees [2, 29, 66] and profiles [67], gene colocalization [68] and fusions [69], maximum coevolutionary interde-

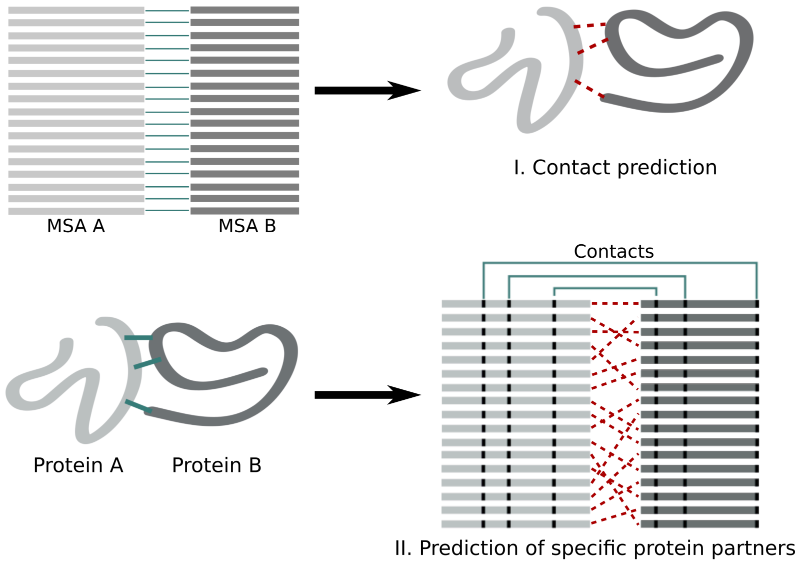


Figure 3.1: Two types of classical theoretical problems related to protein coevolution. I - Prediction of structural contacts from a paired MSA of proteins with know binding partners in the same genome; II - Prediction of a set of preferential affinities from a pair of MSAs of proteins with unknown specific partners and a set of structural contacts.

pendencies [70] and correlated mutations [43, 71], maximization of interfamily coevolutionary signal [32], iterative paralog matching based on sequence energies [3] and expectation–maximization [72] have been developed and applied to infer interaction partners in protein families with single or multiple (paralogous) gene copies in the same genome. Despite these advances, this problem remains unsolved for large sequence *ensembles* in general, especially for the case of protein coevolution across independent genomes - examples are phage proteins and bacterial receptors, pathogen and host-cell proteins, neurotoxins and ion channels, to mention a few. The problem lacks any suitable solution specially because an effective heuristic to search for specific partners across the space of potential matches is still missing for case of large protein families (Figure 3.2).

In Chapter 2 and in the work entitled "Coevolutive, Evolutive and Stochastic Information in Protein-Protein Interactions" (Appendix 1), it is shown that coevolutive information encoded on the interacting amino acids of two proteins can be more useful to discriminate the correct set of protein partners than other evolutive and stochastic sources spread over their sequences [1]. When compared to other sources, the coevolutive information is the strongest signal to distinguish protein partners derived from evolution within the same genome and, likely, the unique indication available in the case of protein interac-

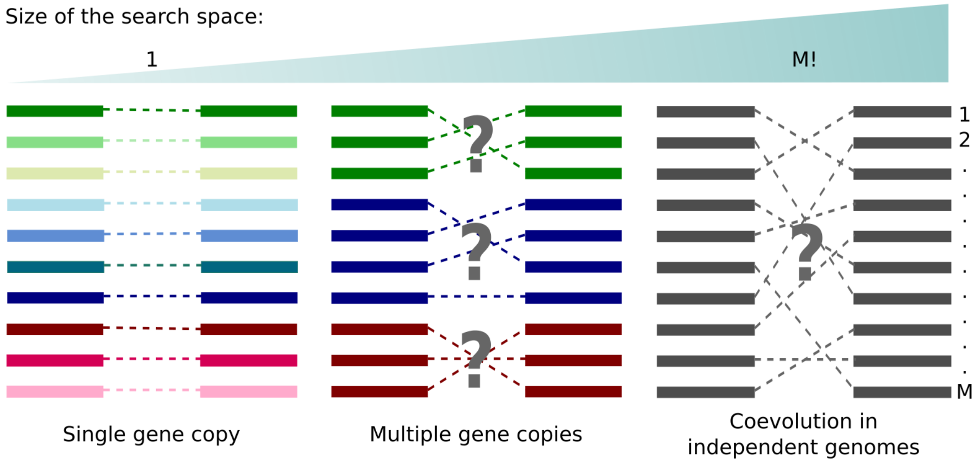


Figure 3.2: Different scenarios for protein partners inference based on multi-sequence alignments. The correct set of partners is known for systems with a single gene copy per genome and unknown for systems involving multiple (paralogous) sequences within the same genome or multiple sequences across independent genomes.

tions in independent genomes. We showed that physically-coupled amino acid sites at the molecular interface store the largest per-contact mutual information ($\hat{I}_{AB}$) to discriminate scrambled MSA concatenations with the largest expectation fraction of correct interaction partners - a result that was found to hold for various definitions of intermolecular contacts and binding modes. Although that information content might be of practical relevance in the search of an effective heuristic to resolve specific protein partners, the degeneracy, *i.e.*, the number of MSA concatenations with a similar value of $\hat{I}_{AB}$ to the native concatenation is expected to be very large, imposing severe limitations to that purpose.

In this chapter, that hypothesis will be investigated for a variety of protein families, including obligate and non-obligate heteromers. It is worth emphasizing that the aim of this work is not to provide a method for prediction of protein-protein interactions nor protein-protein interfaces, hence it differs from the studies in which sequence covariance is used to predict three-dimensional contacts or to infer specific interactions for a set of paralogs. Instead, the objective is to to qualitatively explore the MI degeneracy in the space of possible protein partners associations in two protein families. To approach that, we analyze a set of MSA concatenation solutions obtained by a genetic algorithm (GA) that maximizes $\hat{I}_{AB}$ starting from scrambled MSA concatenations of protein families with known partners in the same genome. Results obtained using two different theoretical frameworks (Shannon and Tsallis statistics) are compared, in a similar manner to what was done in the previous chapter. In the following, the general and specific objectives of this chapter are presented.

3.1.1 Objectives

General objective

The general objective of this chapter is to assess the utility of mutual information (MI) stored in protein-protein interfaces to infer specific protein partners.

Specific objectives

The specific objectives of this chapter are the following:

1. evaluate MSA concatenation solutions obtained by a genetic algorithm that maximizes interface MI starting from a population of scrambled concatenations;
2. investigate which intrinsic properties explain the patterns observed in the sets of solutions obtained for each protein system;
3. understand if the conclusions obtained also hold for the well-studied paralog matching problem;
4. compare the results obtained using Shannon and Tsallis statistics.

3.2 Theory and Methods

In this section, details about the theoretical framework and methods used in this chapter are presented. First, the calculation of the interface mutual information $\hat{I}_{AB}$ is introduced, followed by a description of the protein systems investigated. Then, the genetic algorithm used to search for MSA concatenation solutions with near-native values of $\hat{I}_{AB}$ is described. Finally, details are given about how the accuracy of MSA concatenation solutions is assessed.

3.2.1 Interface mutual information

Consider two M-length MSAs containing sequences from interacting protein families A and B, respectively. A specific coevolution process z associates each sequence l in MSA B to a sequence k in MSA A in a unique arrangement of size M (see Figure 3.3). Given that members of A and B interact via formation of N independent amino-acid contacts at molecular level, it is possible to extract from these MSAs only the columns corresponding to sites that are in contact, belonging to the complex interface. In this context, the interacting amino-acids of families A and B are described by two N-length blocks of discrete stochastic variables, $X^N = (X_1, ..., X_N)$ and $Y^N = (Y_!, ..., Y_N)$, with associated probability mass functions (PMFs) $\{\rho(x_1...x_N), \rho(y_1...y_N), \rho(x_1...x_N, y_1...y_N|z)|x_i, y_i \in \Omega, \forall i \in \{1, ..., N\}\}$. Here, the alphabet Ω has size 21 and contains all 20 amino acids and the gap symbol '-'. Note that only the joint PMF will depend on process z.

Here, we approximate each site-specific PMF $\{\rho(x_i), \rho(y_i), \rho(x_i, y_i|z)|i \in \{1, ..., N\}\}$ by the empirical amino acid frequencies $\{f(x_i), f(y_i), f(x_i, y_i|z)|i \in \{1, ..., N\}\}$ obtained from the concatenated MSAs. Note that each coevolution process z determines a specific concatenation, as illustrated in Figure 3.3. It means that, essentially, the search will be guided by the amount of information X^N stored about Y^N conditional to different coevolution processes z.

The Shannon mutual information contained on the interface of interacting proteins

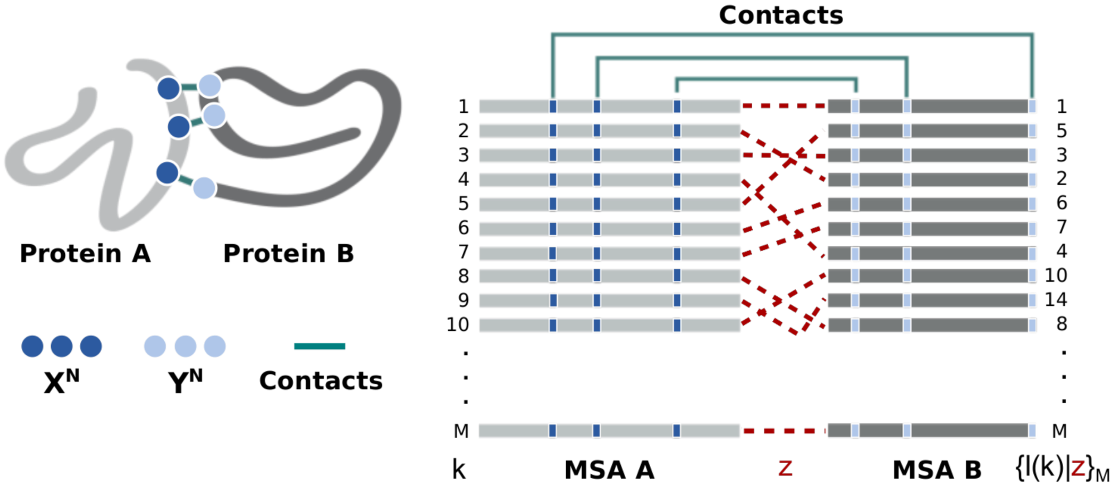


Figure 3.3: Structural contacts mapped into M-long multi-sequence alignment of protein interologs A and B. A set of pairwise protein-protein interactions is defined by associating each sequence l in MSA B to a sequence k in MSA A in one unique arrangement, $l(k)|z$, determined by the coevolution process z to which these protein families were subjected.

A and B conditional to a given coevolution process z is calculated as follows:

$$\hat{I}_{AB} = \frac{1}{N} \sum_{\Omega \times \Omega} f(x_i, y_i|z) \ln \left(\frac{f(x_i, y_i|z)}{f(x_i) f(y_i)} \right) \tag{3.1}$$

where N is the number of contacts at the complex AB interface, $f(x_i)$ is the empirical frequency of x_i as a realization of X_i, $f(y_i)$ is the empirical frequency of y_i as a realization of Y_i, and $f(x_i, y_i|z)$ is the empirical frequency of pair (x_i, y_i) as a realization for the i-th contact given a specific coevolution process z.

Similarly, Tsallis mutual information contained on the interface of interacting proteins A and B conditional to a given coevolution process z is calculated as follows

$$\hat{I}^{(T)}_{AB} = \frac{1}{N} \left(\frac{1}{1-q} \right) \sum_{\Omega \times \Omega} f(x_i, y_i|z)^q \left[\left(\frac{f(x_i, y_i|z)}{f(x_i) f(y_i)} \right)^{(1-q)} - 1 \right] \tag{3.2}$$

where N is the number of contacts at the complex AB interface, $f(x_i)$ is the empirical frequency of x_i as a realization of X_i, $f(y_i)$ is the empirical frequency of y_i as a realization of Y_i, $f(x_i, y_i|z)$ is the empirical frequency of pair (x_i, y_i) as a realization for the i-th contact given a specific coevolution process z, and q is the entropic index. In line with the previous chapter, a value of $q = 1.75$ was considered in all analyses performed in the present chapter.

The empirical values of single and joint frequencies were corrected considering a pseudocount, as follows

$$f(x_i) \leftarrow (1-\lambda)f(x_i) + \frac{\lambda}{Q} \tag{3.3}$$

$$f(x_i, y_i|z) \leftarrow (1-\lambda)f(x_i, y_i|z) + \frac{\lambda}{Q^2} \tag{3.4}$$

where Q is the size of alphabet Ω and λ is the pseudocount parameter. Here, we adopt a small pseudocount of $\lambda = 0.001$.

3.2.2 Systems under investigation

Protein complexes under investigation are shown in Table 3.1. MSAs and for all protein families were obtained from Ovchinnikov and coworkers [51]. Amino-acid contacts defining the discrete stochastic variables were identified from the x-ray crystal structure of the bound state of a representative protein pair from families using a typical contact definition considering maximum separation distance of 8Å between amino acids carbon beta. The full dataset of protein systems validated in [51] was considered here, with the exception of systems 2Y69_BC, 2ONK_AB, 3A0R_AB, 3RPF_AB and 4HR7_AB, which were considered outliers in terms of M/N value (Table 3.1), with M/N values of 469.3, 87.7, 192.3, 150.6 and 45.3, respectively.

Additionally, the HK-RR standard dataset containing around 5,000 sequences, coming from around 450 bacterial genomes from the P2CS database [73–75] was included. This paired MSA was produced and validated by Bitbol and coworkers [3, 33] in paralog matching experiments. The PDB complex 5UHT (chains A and B) was selected as a representative for this system. The reason for including this system containing paralogous proteins is to have a baseline for comparison with previous related studies.

Table 3.1: Protein complexes considered in the study. M is the number of sequences in the multi-sequence alignment and N is the number of contacts in the interface, following an 8Å cutoff definition.

#	Description	PDB ID	Chains	M	$N_{d_c \leq 8}$	M/N
1	Dihydroorotate dehydrogenase B (4-mer)	1EP3	A, B	552	91	6.1
2	Thiazole synthase/ThiS (8-mer)	1TYG	A, B	746	80	9.3
3	Carbamoyl phosphate synthetase (8-mer)	1BXR	A, B	1,004	154	6.5
4	3-oxoadipate coA-transferase (4-mer)	3RRL	A, B	1,330	161	8.3
5	Toxin-antitoxin complex RelBE2 (4-mer)	3G5O	A, B	904	92	9.8
6	Bovine cytochrome C oxidase (13-mer)	2Y69	A, B	1,484	246	6.0
7	Bovine cytochrome C oxidase (13-mer)	2Y69	A, C	863	210	4.1
8	Phenylalanyl-tRNA synthetase (4-mer)	1B70	A, B	1,108	255	4.3
9	Electron transfer flavoprotein (2-mer)	1EFP	A, B	1,347	229	5.9
10	Anthranilate synthase (4-mer)	1I1Q	A, B	1,204	91	13.2
11	Tryptophan synthase (4-mer)	1QOP	A, B	1,155	102	11.3
12	4-hydroxybenzoyl-CoA reductase (6-mer)	1RM6	A, B	1,604	71	22.6
13	4-hydroxybenzoyl-CoA reductase (6-mer)	1RM6	A, C	1,534	154	10.0
14	4-hydroxybenzoyl-CoA reductase (6-mer)	1RM6	B, C	1,481	93	15.9
15	Pyruvate dehydrogenase E1 (5-mer)	1W85	A, B	1,537	121	12.7
16	GTP-Regulated ATP Sulfurylase (2-mer)	1ZUN	A, B	649	140	4.6
17	TusBCD proteins (6-mer)	2D1P	B, C	216	40	5.4
18	Succinyl-CoA Synthetase (4-mer)	2NU9	A, B	798	144	5.5
19	Polysulfide reductase (6-mer)	2VPZ	A, B	676	119	5.7
20	Succinate:quinone oxidoreductase (4-mer)	2WDQ	C, D	221	43	5.1
21	GatCAB (3-mer)	3IP4	A, B	782	94	8.3
22	GatCAB (3-mer)	3IP4	A, C	879	146	6.0
23	GatCAB (3-mer)	3IP4	B, C	689	122	5.6
24	Allophanate Hydrolase (4-mer)	3MML	A, B	1,067	116	9.2
25	F1-ATP synthase (8-mer)	3OAA	H, G	886	179	4.9
26	DhaK-DhaL (4-mer)	3PNL	A, B	902	113	8.0
27	*Thermotoga maritima* HK853-RR468 (4-mer)	5UHT	A, B	5,110	33	

3.2.3 Genetic algorithm

Both Shannon and Tsallis mutual information (MI) contained on the interface of protein complexes (Table 3.1), calculated as described in eq. 3.1 and in eq. 3.2, respectively, were maximized using a genetic algorithm (GA, Figure 3.4, Algorithm 3.1). For each of the protein complexes considered, six independent optimization trajectories were obtained, starting from different randomly generated populations. Each optimization was performed with a population of eight individuals with unique genomes encoding a specific concatenation of MSAs A and B. In each generation, the elite (top-50$\%$ individuals with best fitness) reproduces and replaces the remaining 50$\%$ individuals with lower fitness by new individuals with genomes that are mutated copies of the elite. A mutation in the genome of an individual consists of swapping positions of two sequences on MSA B, and thereby slightly changing the concatenation z. The fitness of the individuals is calculated in each generation and corresponds to the total interface MI, $\hat{I}_{AB}$, obtained considering an individual unique genome, *i.e.*, a specific concatenation of MSAs A and B. The optimization is stopped after a predefined number of 50,000 generations is reached.

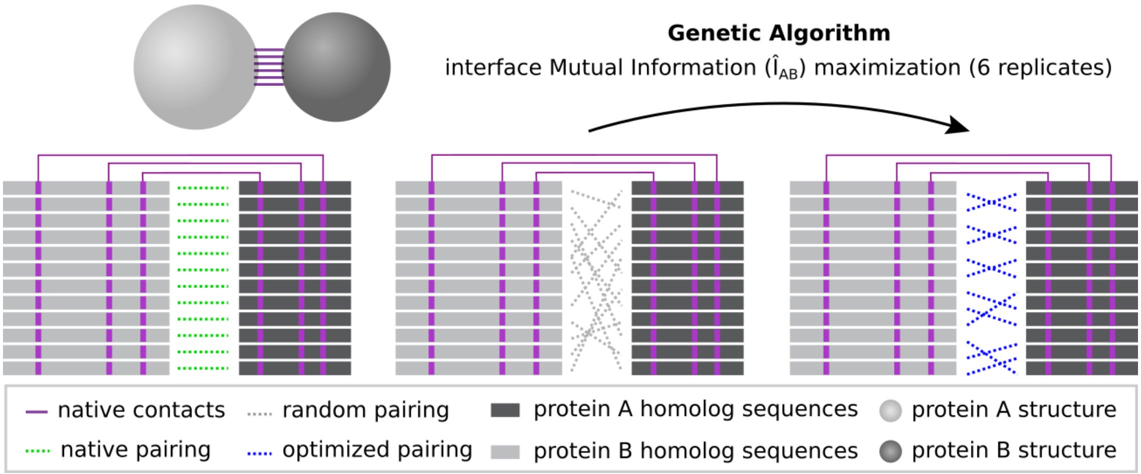

Genetic Algorithm
interface Mutual Information ($\hat{I}_{AB}$) maximization (6 replicates)
native contacts
native pairing
random pairing
optimized pairing
protein A homolog sequences
protein B homolog sequences
protein A structure
protein B structure

Figure 3.4: Scheme showing interface mutual information ($\hat{I}_{AB}$) optimization process starting from a scrambled MSA concatenation (in gray) and reaching an optimized concatenation (in blue). Only physically coupled MSA position pairs (shown in purple) are taken into account.

A slightly different optimization procedure was implemented for the special case of the HK-RR standard dataset (Figure 3.5). In this case, the initial population is composed by within-species scrambled solutions and, in each generation, only within-species changes are allowed. More specifically, each time a new mutated individual is generated, one of the species that composes the MSA is randomly selected, and a change in the concatenation within this species is performed. The optimization is stopped after a predefined number of 100,000 generations is reached.

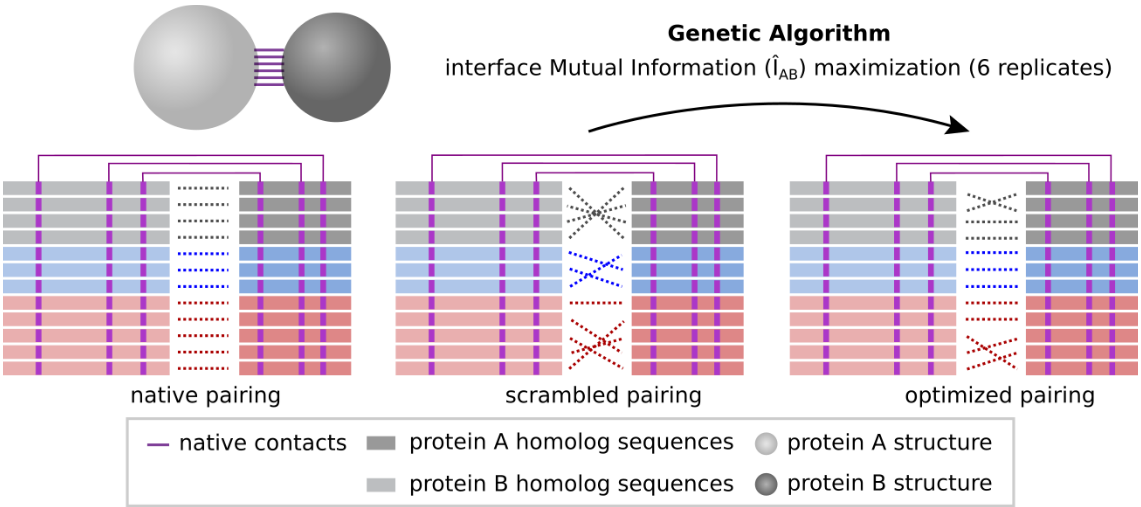


Figure 3.5: Scheme showing interface mutual information ($\hat{I}_{AB}$) optimization process for the HK-RR standard dataset. It starts from a within-species scrambled MSA concatenation and reaches an optimized concatenation. Different species are shown in different colors. Only physically coupled MSA position pairs (shown in purple) are taken into account and only within-species changes are made in each generation.

The optimal set of parameters for the GA were derived from a series of tests performed on six representative systems. In each test, one of these parameters varied, assuming a range of values while all other parameters remained fix (Table 3.2). All tests were performed with a predefined seed for the random number generator, which means that the starting point and the sequence of mutations performed are constant for all trajectories of the same system. This was done to ensure that any effects observed in the final results were due solely to variations in the GA parameters.

Figure 3.6 shows how parameter values correlated with relative $\hat{I}_{AB}$ at the end of test trajectories. Given that both the number of individuals and the elite proportion correlated positively with relative $\hat{I}_{AB}$ (Figure 3.6A-B), the values selected for these parameters were the maximum tested, *i.e.*, 8 and 0.5, respectively. The number of mutations, on the other hand, correlated negatively with relative $\hat{I}_{AB}$ (Figure 3.6C), thus the value selected for this parameter was 1. Results for parameter λ were not so conclusive (Figure 3.6D)

Table 3.2: Genetic algorithm parameters values tested on representative systems 1BXR_AB, 3MML_AB, 2NU9_AB, 1RM6_AB, 3IP4_AB and 3G5O_AB. In each test, one of these parameters varied to assume all its possible values while all other parameters remained fix in the value marked in bold.

Population size	**Elite (% of population)**	**Number of mutations**	**Pseudocount parameter λ**
2	12.5%	**1**	**0.001**
4	25%	2	0.01
6	37.5%	3	0.1
8	**50%**	4	0.5

and, since this parameter was set to 0.001 in previous work [1], its value was maintained the same. However, in future work, it might be interesting to test higher values of λ.

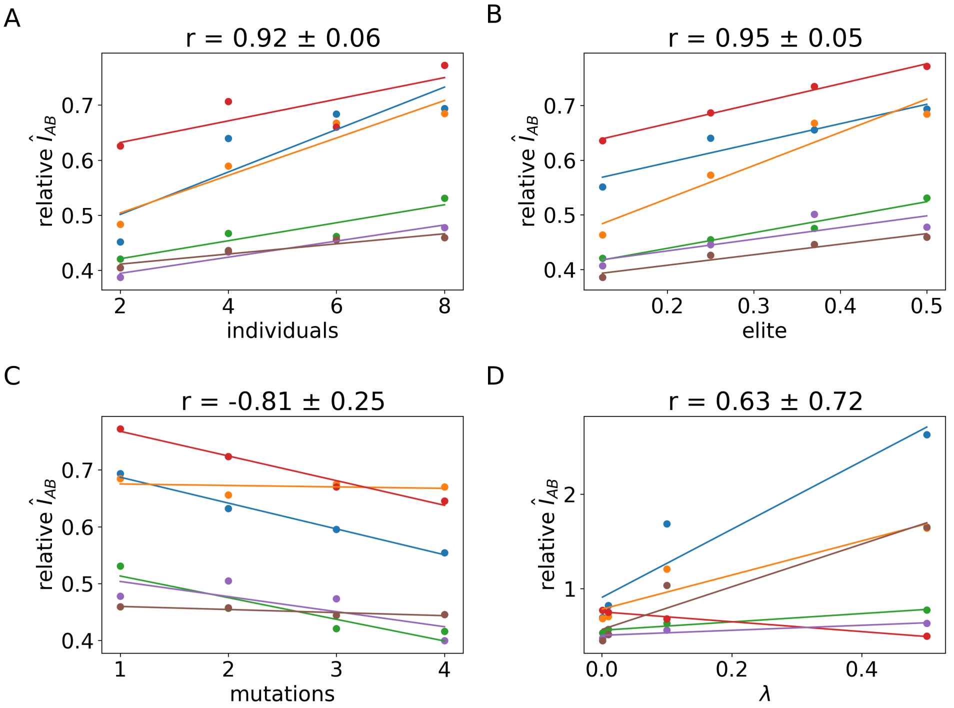


Figure 3.6: Analysis of relative $\hat{I}_{AB}$ values reached at the end of test trajectories considering six representative systems: 1BXR_AB (blue), 3MML_AB (green), 2NU9_AB (orange), 1RM6_AB (purple), 3IP4_AB (brown) and 3G5O_AB (red). The parameters tested were: population size (A), elite (B), number of mutations (C), and pseudocount parameter λ (D). While one parameter was tested, the others were fixed in the following default values: 8, 0.5, 1 and 0.001, respectively. All trajectories ended after 5,000 generations. The average Pearson correlation is shown on top of each plot considering all systems (n = 6).

Algorithm 3.1: Simplified Python implementation of the genetic algorithm used for mutual information optimization.

```python
# User defined parameters
GENERATIONS = 50000
MUTATIONS = 1
INDIVIDUALS = 8
ELITE = 4
LAMBDA = 0.001

# Read and encode MSA
msa_a = readMSA("msa_a.fasta")
msa_b = readMSA("msa_b.fasta")

# Read and map contacts to MSA
contacts = readContacts("contacts_8A.txt")
col_pairs = mapContactsToMSAs(contacts, msa_a, msa_b)

# Fitness function
def fitness(genome):
  # Concatenates MSA A and B and extract relevant column pairs
  coev_model = getCoevolutionModel(genome, msa_a, msa_b, col_pairs)
  # Calculate single-site amino acid frequencies
  site_freqs = calculateSiteFreqs(coev_model, LAMBDA)
  # Calculate double-site amino acid frequencies
  pair_freqs = calculatePairFreqs(coev_model, LAMBDA)
  # Calculate Shannon’s mutual information matrix
  mi = calculateMI(site_freqs, pair_freqs)

  return sum(mi)

# Prepare initial population
population = []
for n in range(INDIVIDUALS):
  # Generate random concatenation for MSA A and B
  new_genome = generateRandomGenome(len(msa_a))
  population.append([new_genome, fitness(new_genome)])
population.sort(key = lambda x : x[1], reverse = True)
```

```python

# Run optimization
for g in range(GENERATIONS):
  for i in range(ELITE):
    genome = population[i][0]
    for j in range(MUTATIONS):
      # Swap two indexes in genome
      mutate(genome)
    # Replace low fitness individuals
    population[INDIVIDUALS - ELITE + i] = [genome, fitness(genome)]
  population.sort(key = lambda x : x[1], reverse = True)
```

3.2.4 Assessment of optimized solutions accuracy

The true positive (TP) rates of optimized concatenations obtained at the end of the genetic algorithm (GA) $\hat{I}_{AB}$ maximization trajectories was calculated in two different manners: with and without mismatch discounting. TP rate assessment without mismatch discounting consists simply in counting how many sequence partners were correctly paired in the target solution and divide by the total number of sequences (Figure 3.7A). TP rate assessment with mismatch discounting, on the other hand, consists in counting how many sequences were paired either with their correct partner or with a partner that is close enough to the correct one in terms of Hamming distance (Figure 3.7B). Hence, mismatch discounting depends on a predefined Hamming distance cutoff, below which sequences are considered similar enough for the mistakes to be forgiven. In this chapter, we consider the 20th percentile of a given protein family B distance distribution as the predefined cutoff for mismatch discounting. However, different cutoffs are considered at a certain point to show that this parameter does not affect qualitatively the results.

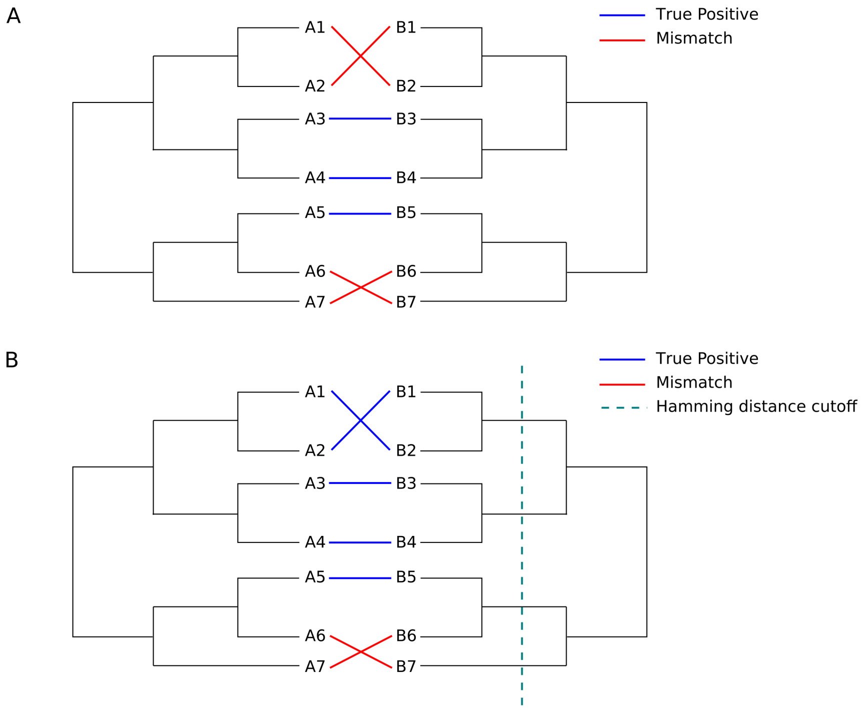


Figure 3.7: Mismatch discounting based on a Hamming distance cutoff. Scheme showing how the accuracy of the same MSA concatenation would be assessed with (B) and without (A) mismatch discounting.

3.3 Results and Discussion

In this section, results concerning protein partners inference based on interface mutual information $\hat{I}_{AB}$ are presented and discussed. First, MSA concatenation solutions obtained by a genetic algorithm (GA) maximizing $\hat{I}_{AB}$ based on Shannon statistics are presented and analyzed. Then, similar results obtained for the HK-RR standard dataset of paralogous sequences are compared with previous results in literature. Finally, MSA concatenation solutions found by a GA maximizing $\hat{I}_{AB}$ based on Tsallis statistics are analyzed and compared to the results obtained using Shannon statistics.

3.3.1 Protein partners inference using Shannon statistics

In search for an effective heuristic to resolve specific protein partners based on MSAs with large numbers of sequences, degeneracy of the per-contact mutual information $\hat{I}_{AB}$ was investigated across 26 independent protein families with known interaction partners in the same genome (Theory and Methods - Table 3.1). To approach that, optimization trajectories were produced by a genetic algorithm (GA) that starts from a random concatenation of MSA A and MSA B, and maximizes $\hat{I}_{AB}$ by performing small changes in the MSA concatenation iteratively (Theory and Methods - Figure 3.4). Accordingly, Figure 3.8A shows 156 optimization trajectories with convergence obtained after 50,000 generations as indicated by their average time derivative $\delta\hat{I}_{AB} \leq 0.0001$ in Figure 3.8B. The average trajectory converges at around 100% of the reference $\hat{I}_{AB}$ value calculated for the native concatenation z^{*}.

Despite presenting near-native values of $\hat{I}_{AB}$, optimized solutions fail at pairing sequences correctly in consequence of the degeneracy of the space of possible concatenations constrained by the mutual information maximization criteria. As made clear in Figure 3.9A, optimized solutions appear separated from random solutions in terms of relative $\hat{I}_{AB}$, but not in terms of true positive (TP) rate. Careful inspection of the data reveals that the presence of similar sequences in MSA B contribute to that high error rate by generating solutions with indistinguishable values of $\hat{I}_{AB}$. Indeed, reassessment of TP rates by disregarding mistakes made among sequences below the 20th percentile of the

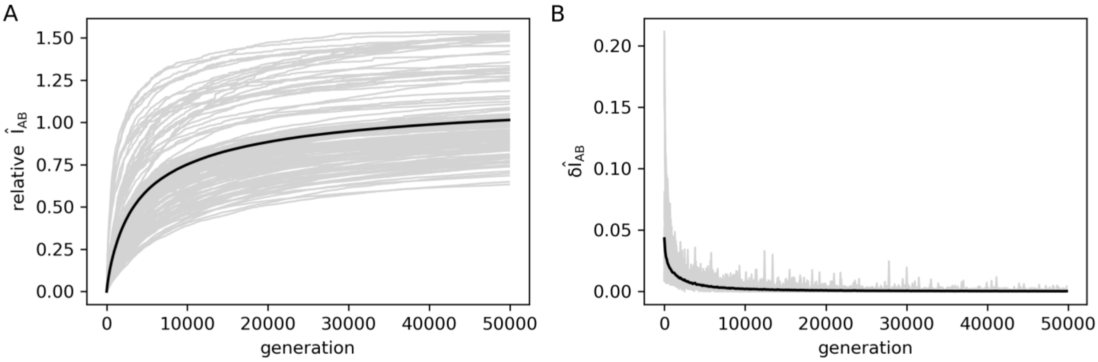


Figure 3.8: $\hat{I}_{AB}$ optimization trajectories. For each of the 26 systems, there are six trajectories with different starting points (n = 156). A, The value of $\hat{I}_{AB}$ normalized by the native $\hat{I}_{AB}$ (relative $\hat{I}_{AB}$) is plotted against the number of generations of the genetic algorithm (gray lines). The average trajectory is shown in black. B, First-order derivative of the optimization trajectories shown in A. The derivatives of individual trajectories are shown in gray, while the average derivative over all trajectories is shown in black.

Hamming distances distribution allows the classification of solutions into type-(i) with significant TP rates over 30% (p-value = 0.0005), and type-(ii) with TP rates below the significance value of 30% (Figure 3.9B).

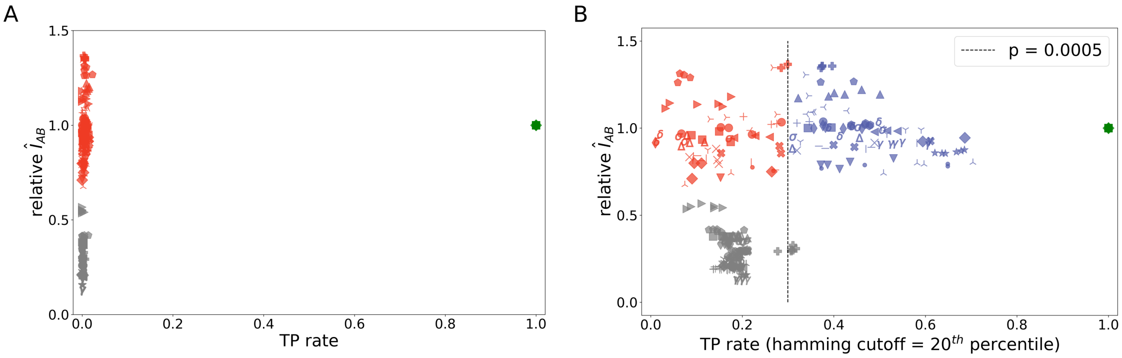


Figure 3.9: Evaluation of the accuracy of MSA concatenation solutions. A, The relative interface mutual information $\hat{I}_{AB}$ is plotted against the true positive (TP) rate of random (gray), optimized (red) and native (green) MSA concatenations. B, Reassessed values of TP rate of random, optimized and native MSA concatenations discounting wrong pairings among related sequences, with Hamming distance within the 20th percentile of the distance distribution. Optimized solutions with significant TP rate over 30% (p-value = 0.0005) are shown in blue, while optimized solutions with non-significant TP rate below 30% are shown in red. Random solutions are shown in gray. Each symbol represents a different protein system (n = 26).

As a measure of correlation, it is not surprising that $\hat{I}_{AB}$ is degenerate given this trivial error source arising from mismatches among similar sequences. Unexpected however is the fact that the degeneracy does also generate this other subspace of type-(ii) optimized solutions, which contain many non-trivial mismatches among sequences at larger Hamming distances. As shown in Figure 3.10, conclusions about subspaces (i) and (ii) hold for mismatch discounting using other Hamming distance cutoffs.

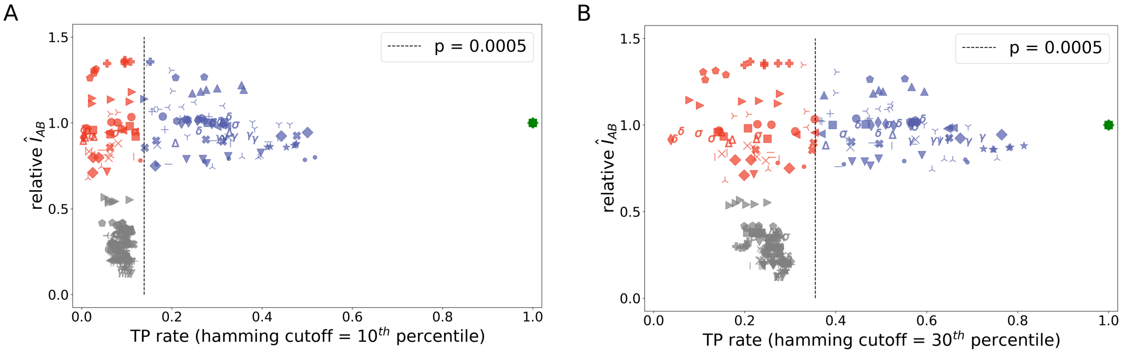


Figure 3.10: Alternative Hamming distance cutoffs. The relative interface mutual information $\hat{I}_{AB}$ is plotted against the true positive (TP) rate of random, optimized and native MSA concatenations. Wrong pairings among related sequences, with Hamming distance within the 10th (A) and 30th (B) percentiles of the distance distribution were disregarded. Optimized solutions with significant TP rates over 14% (A) and 36% (B) (p-value = 0.0005) are shown in blue, while the remaining solutions are shown in red. Random solutions are shown in gray. Each symbol represents a different protein system (n = 26).

To further investigate these results, the 26 protein systems were classified in five groups, according to the results obtained at the end of the $\hat{I}_{AB}$ maximization trajectories. Group 1 is composed by systems with only type-(i) solutions (Figure 3.11). A total of four systems fall in this group, namely 3RRL_AB, 2Y69_AB, 2Y69_AC and 1ZUN_AB. 2Y69_AB achieved the best average TP rate (around 60%), while 1ZUN_AB achieve the worst (around 40%).

Group 2 is composed by systems with a majority of type-(i) solutions (Figure 3.12). A total of seven systems belong to this group, namely 1BXR_AB, 1B70_AB, 1EFP_AB, 1EP3_AB, 1W85_AB, 3PNL_AB and 3G5O_AB. While some systems like 1EFP_AB achieved presented big differences in TP rates of type-(i) and type-(ii) solutions, other systems, like 1EP3_AB presented a smaller gap between TP rates of the two kinds of solutions.

Group 3 is composed by systems with the same proportions of type-(i) and type-

(ii) solutions (Figure 3.13). A total of four systems are found in this group, namely 3IP4_AC, 3IP4_BC, 1RM6_BC and 2NU9_AB. This group apparently does no present any distinguishable features, since all systems present variable TP rates for both type-(i) and type-(ii) solutions.

Group 4 is composed by systems with a majority of type-(ii) solutions (Figure 3.14). A total of six systems belong to this group, namely 1RM6_AB, 1RM6_AC, 2VPZ_AB, 2WDQ_CD, 3MML_AB and 3OAA_HG. It is possible to notice all systems present somewhat prominent differences in the TP rates of type-(i) and type-(ii) solutions, with most type-(ii) solutions presenting lower TP rates than random solutions.

Finally, group 5 is composed by systems in which optimized concatenations did not differentiate from the scrambled ones (Figure 3.15). A total of five systems are found in this group, namely 1I1Q_AB, 1QOP_AB, 1TYG_BA, 2D1P_BC and 3IP4_AB. Systems 1I1Q_AB and 3IP4_AB present Hamming distance distributions shifted to the left, which might be one of the features influencing the observed TP rates. System 2D1P_BC presents especially low TP rates.

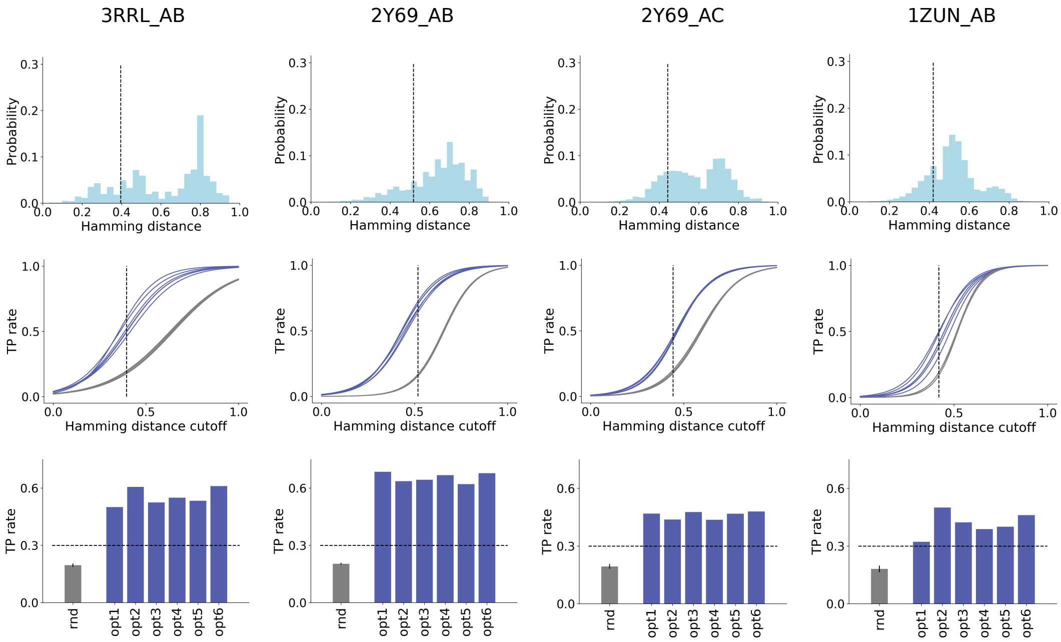


Figure 3.11: Group 1 - all solutions are type-(i), blue. (1st) Hamming distance distribution of MSA B. (2nd) True positive (TP) rate for different Hamming distance discounts. The 20th percentile is shown with a dashed line, random solutions in gray, optimized solution in blue. (3rd) TP rates of random (rnd) and optimized (opt1-6) solutions at 20th percentile Hamming distance cutoff. The significance value is shown with a dashed line (p=0.0005).

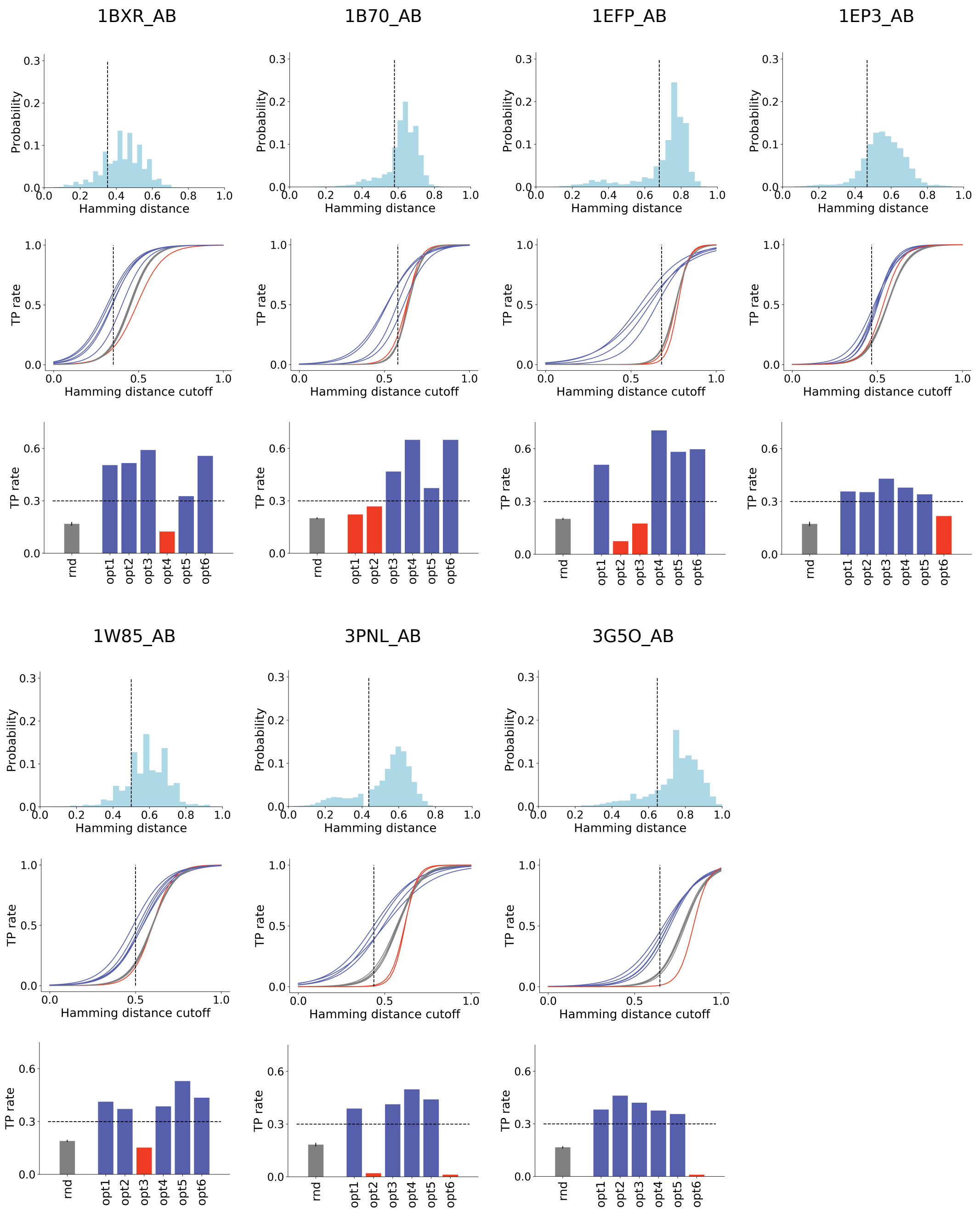


Figure 3.12: Group 2 - majority of solutions type-(i), blue. (1st) Hamming distance distribution of MSA B. (2nd) True positive (TP) rate for different Hamming distance discounts. The 20th percentile is shown with a dashed line, random solutions in gray, optimized solution in blue. (3rd) TP rates of random (rnd) and optimized (opt1-6) solutions at 20th percentile Hamming distance cutoff. The significance value is shown with a dashed line (p=0.0005).

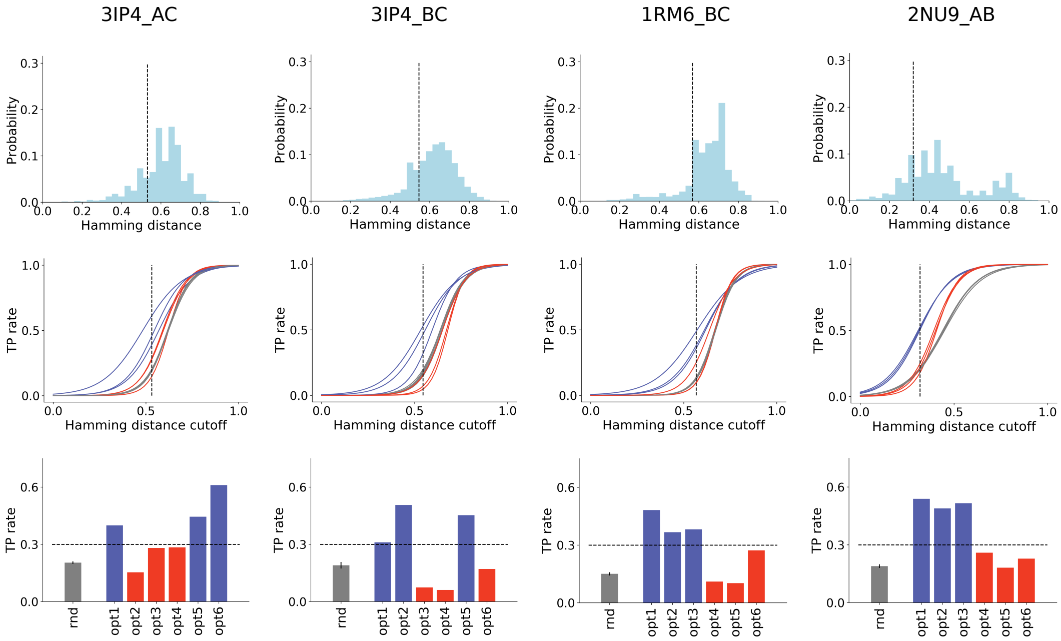


Figure 3.13: Group 3 - mixed type-(i) blue and type-(ii) red solutions. (1st) Hamming distance distribution of MSA B. (2nd) True positive (TP) rate for different Hamming distance discounts. The 20th percentile is shown with a dashed line, random solutions in gray, optimized solution in blue. (3rd) TP rates of random (rnd) and optimized (opt1-6) solutions at 20th percentile Hamming distance cutoff. The significance value is shown with a dashed line (p=0.0005).

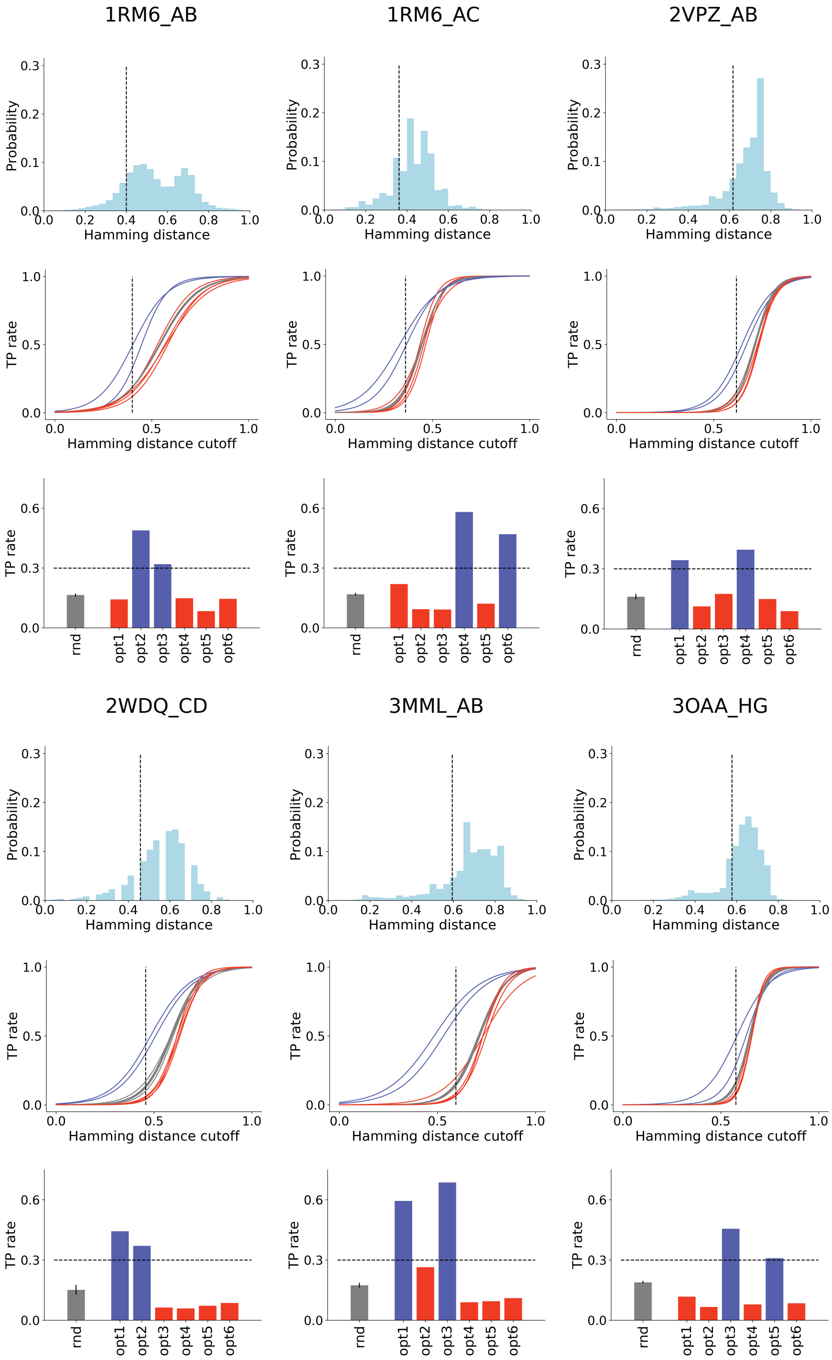


Figure 3.14: Group 4 - majority of solutions type-(ii), red. (1st) Hamming distance distribution of MSA B. (2nd) True positive (TP) rate for different Hamming distance discounts. The 20th percentile is shown with a dashed line, random solutions in gray, optimized solution in blue. (3rd) TP rates of random (rnd) and optimized (opt1-6) solutions at 20th percentile Hamming distance cutoff. The significance value is shown with a dashed line (p=0.0005).

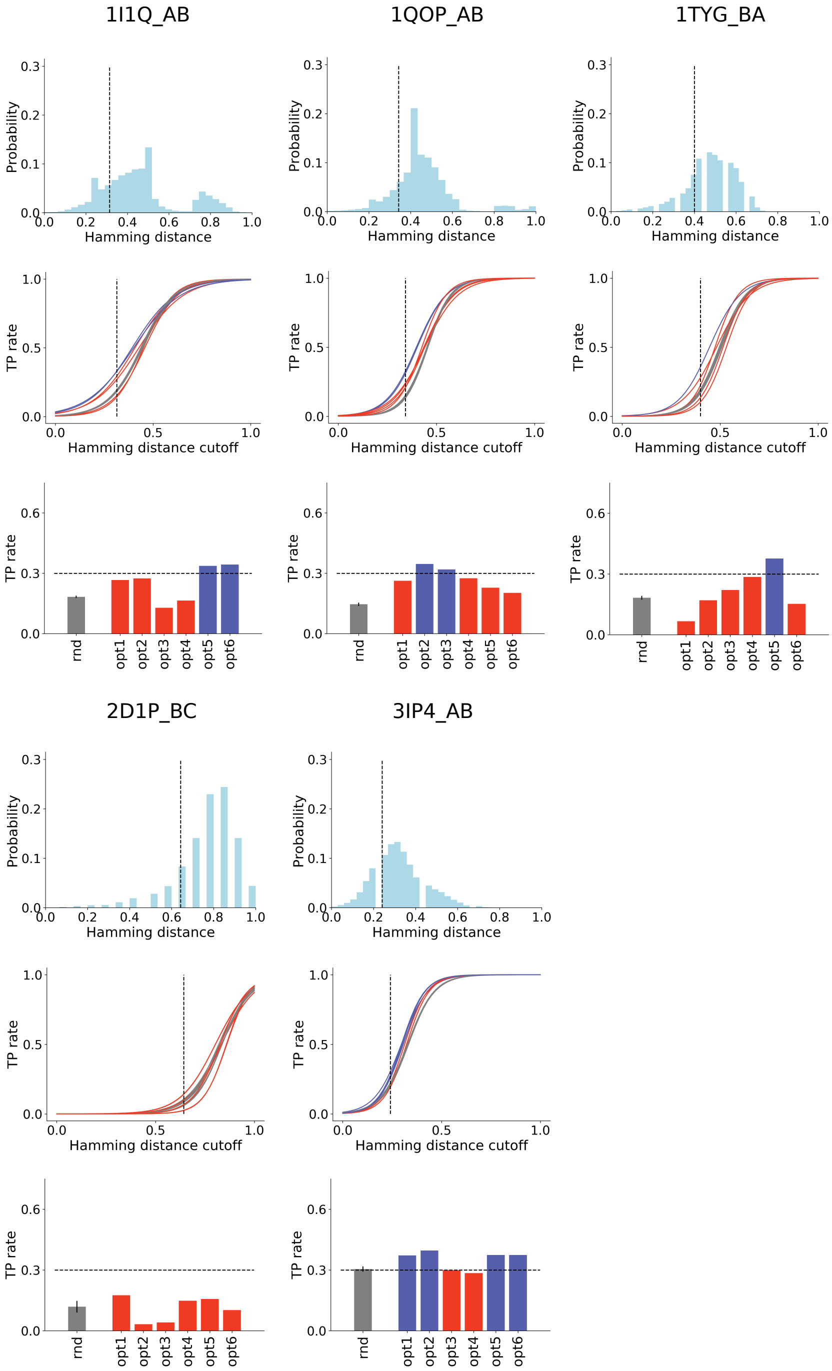


Figure 3.15: Group 5 - optimized solutions non-distinguishable from random. (1st) Hamming distance distribution of MSA B. (2nd) True positive (TP) rate for different Hamming distance discounts. The 20th percentile is shown with a dashed line, random solutions in gray, optimized solution in blue. (3rd) TP rates of random (rnd) and optimized (opt1-6) solutions at 20th percentile Hamming distance cutoff. The significance value is shown with a dashed line (p=0.0005).

At this point, it is important to try to identify which intrinsic system properties drive optimization results towards subspace (i), *i.e.* towards solutions with higher TP rates, or subspace (ii), *i.e.* towards solutions with lower TP rates. Different properties were investigated, namely MSA depth (number of sequences), interface size (number of contacts), native $\hat{I}_{AB}$ value, and Mirror-Tree correlation [2] between MSA A and MSA B, trying to elucidate the observed heterogeneous behavior of the different systems upon optimization (Figure 3.16).

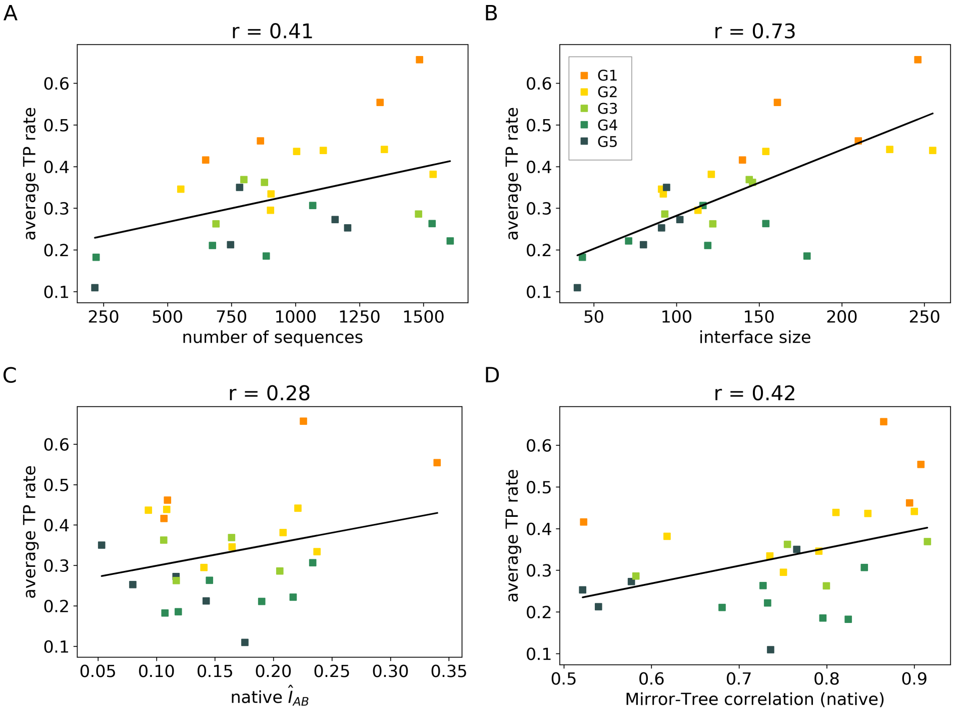


Figure 3.16: Pearson correlation between the average true positive (TP) rate of the optimized solutions (n = 6 for each system) and the number of sequences in the alignment (A), number of contacts on the interface (interface size) (B), mutual information per contact on the interface $\hat{I}_{AB}$ of the native solution (C), and Mirror-Tree correlation [2] of the native solution (D). Systems are colored by groups G1-5.

It can be seen that both the number of sequences in the MSA and the native Mirror-Tree correlation are weakly correlated to the average TP rates of the studied systems (Figure 3.16A,D), while the interface size (number of contacts) is strongly correlated to the average TP rate (Figure 3.16B). In addition to that, it is possible to observe a pattern in the distributions of groups G1-5, with G1 and G2 concentrated on the upper right quadrant and G3-5 in the lower left quadrant. This result indicates that having a larger block of interface contacts might help maintaining the stability of the optimizations, avoiding great deviations from the expected average value, and making the trajectories grow in the right direction (towards the native solution).

3.3.2 A particular case study: HK-RR paralogs

So far, results were obtained for a set of protein families involving unique sequence pairs per genome that may not have evolved under strong selective pressures towards specificity. To better understand any implicit dependence of the results with that experimental condition, error sources (i) and (ii) were then further investigated in the context of the bacterial two-component system HK-RR featuring highly specific protein-protein interactions across multiple protein copies per genome. More specifically, histidine kinase (HK) and their respective response regulator (RR) are paralogous gene families, each consisting of multiple sequences sharing significant homology at the primary and tertiary levels. Despite that signature, HK-RR pairs are highly specific within the same genome in consequence of evolutive pressures avoiding crosstalk between independent two-component pathways [76] - as shown by Rowland and Deeds, evolution of new HK-RR pairs follows rapid sequence divergence immediately after duplication events [77].

Accordingly, Figure 3.17 presents a series of optimizations performed on a HK-RR MSA containing around 5,000 sequences, coming from 450 bacterial genomes from the P2CS database [73–75]. Optimizations were performed with 6 replicates each, starting from a paired alignment with a randomized pairing within each species. All species were optimized together, which means that each optimization step benefited from the cumulative changes that happened in previous steps (see Figure 3.5). As shown in Figure 3.17A, optimization to near-native values of $\hat{I}_{AB}$ is attained after 100,000 generations, with $\delta\hat{I}_{AB} \leq 0.001$.

When analysing the true positive (TP) rates for species with different numbers of paralogs, optimized MSA solutions presented an improvement over the initial concatenations (Figure 3.17B). In this case, TP rates are not null because the degeneracy of $M \leq 32$ paired sequences of paralogs is expected to be significantly smaller than that of $M > 200$ paired sequences in Figures 3.11-3.15. Is is interesting to notice that TP rates obtained

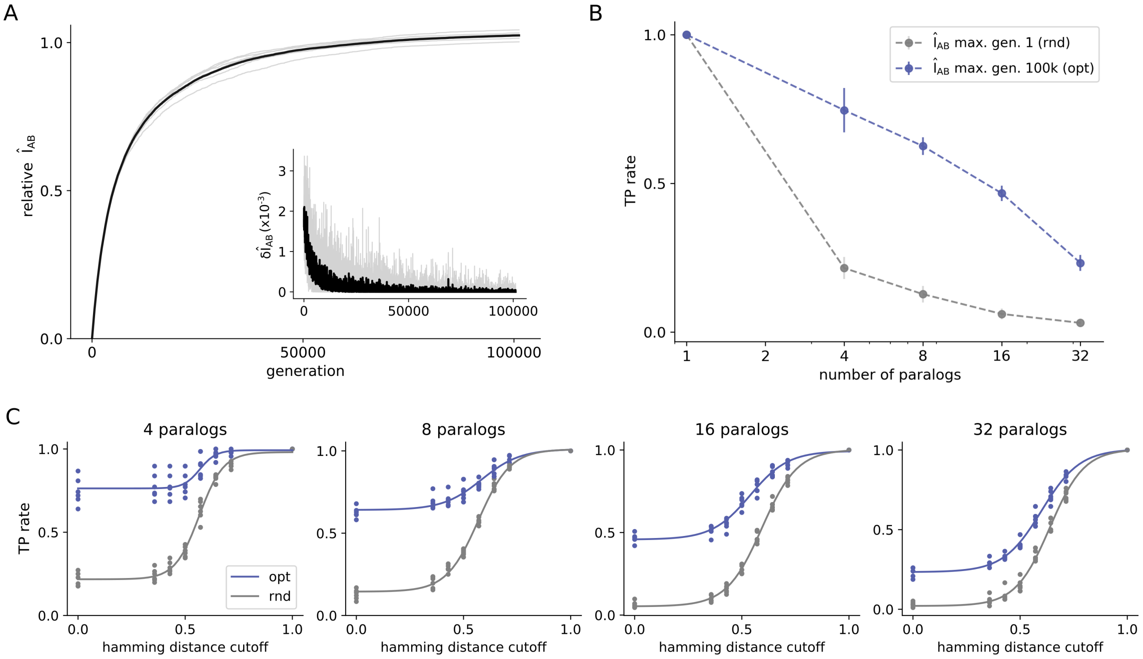


Figure 3.17: Evaluation of optimized MSA concatenations of HK-RR paralogs dataset. A, Optimization trajectories for the HK-RR standard dataset [3]. The interface mutual information normalized by the native interface mutual information (relative $\hat{I}_{AB}$) is plotted against the number of generations for optimizations (n = 6) starting from a solution with a scrambled concatenation within each species. The first derivative of the trajectory is shown in the smaller plot. B, True positive (TP) rate of start (in gray) and final (in blue) solutions after 100,000 generations. The TP rate is shown in average for bacterial species containing different numbers of paralogs. C, TP rates obtained for mismatch discounts at different Hamming distance cutoffs both random (rnd) and optimized (opt) MSA concatenations.

here by optimizing only the interface MI are only slightly inferior to the same estimates obtained considering full protein MI found in literature [33], especially for genomes with a higher number of paralogs.

Figure 3.17C shows the TP rate of optimized and random MSA concatenations considering mismatch discounts at different Hamming distance cutoffs for bacterial genomes with different numbers of paralogs. It is possible to observe that random and optimized curves approximate with increasing number of paralogs. Extrapolating for cases with more than 32 paralogs, the two curves will probably overlap, a behavior similar to what is observed for group 5 systems (Figures 3.15). This supports the conclusion that type (i) errors do not contribute to $\hat{I}_{AB}$ degeneracy in HK-RR system. It is unclear, however, if this lack of type-(i) error (originated from mismatches among similar sequences) is due to the high specificity of this system or due to the small interface size (33 contacts).

In previous work, Bitbol *et al.* developed an iterative pairing algorithm (IPA) capable of inferring protein partners using either direct coupling analysis (DCA-IPA) [3], mutual information (MI-IPA) [33] or phylogeny (Mirrortree-IPA) [78]. When benchmarked for paralog matching on the standard HK-RR dataset, DCA-IPA was as accurate as MI-IPA, and Mirrortree-IPA was even more accurate. The performance of these algorithms, however, drops considerably for species with more than 32 paralogs. The tendency is that the TP rate also drops to zero in a hypothetical genome with hundreds of paralogs [78]. This is the same situation observed here for systems in groups 1-5. In conclusion, results presented in Figure 3.17 suggest that paralog matching is a problem that is only solvable due to the small number of sequences involved and, when extended to genomes with more sequences, does probably present only type-(ii) solutions, leaving virtually no room for improvement of TP rates.

3.3.3 Protein partners inference using Tsallis statistics

In the previous chapter, it was shown that Tsallis statistics seem to capture better structural properties of protein systems. Thus, this alternative theoretical framework was also investigated regarding its utility to infer specific protein partners. Figure 3.18 shows $\hat{I}_{AB}$ maximization trajectories obtained using Tsallis statistics. It is possible to see a few trajectories decoupling from the others and reaching relative values of relative $\hat{I}_{AB}$ up to 2.5 times the native $\hat{I}_{AB}$.

Similarly to what happens in Shannon statistics, optimized solutions obtained after 50,000 generations present negligible true positive (TP) rates (Figure 3.19A). However, after reassessing the accuracy of solutions with mismatch discount, it is possible to observe that most of the optimized solutions present TP rates above the significance threshold on around 30%. Overall, up to this point, this looks like a better result than the one obtained using Shannon MI.

To better evaluate these results, the 26 protein systems were classified in three groups based on the pattern of optimized solutions presented by them. Group 1, composed by

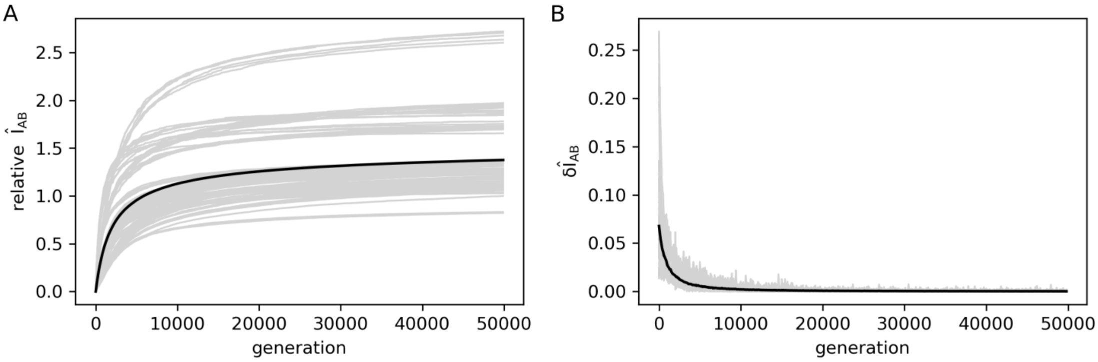


Figure 3.18: Tsallis $\hat{I}_{AB}$ optimization trajectories. For each of the 26 systems, there are six trajectories with different starting points (n = 156). A, The value of $\hat{I}_{AB}$ normalized by the native $\hat{I}_{AB}$ (relative $\hat{I}_{AB}$) is plotted against the number of generations of the genetic algorithm (gray lines). The average trajectory is shown in black. B, First-order derivative of the optimization trajectories shown in A. The derivatives of individual trajectories are shown in gray, while the average derivative over all trajectories is shown in black.

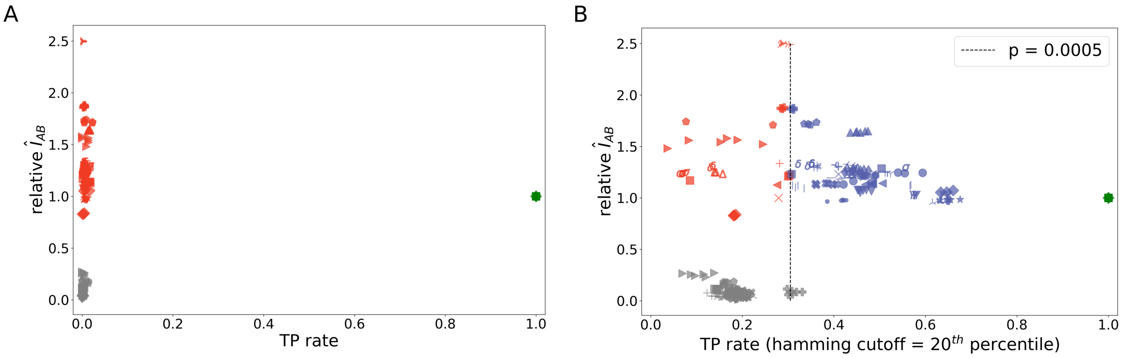


Figure 3.19: Evaluation of the accuracy of MSA concatenation solutions obtained using Tsallis statistics. A, The relative interface mutual information $\hat{I}_{AB}$ is plotted against the true positive (TP) rate of random (gray), optimized (red) and native (green) MSA concatenations. B, Reassessed values of TP rate of random, optimized and native MSA concatenations discounting wrong pairings among related sequences, with Hamming distance within the 20th percentile of the distance distribution. Optimized solutions with significant TP rate over 30% (p-value = 0.0005) are shown in blue, while optimized solutions with non-significant TP rate below 30% are shown in red. Random solutions are shown in gray. Each symbol represents a different protein system (n = 26).

14 systems, contains systems which only produced type-(i) solutions upon optimization (Figure 3.20). It is interesting to notice that Tsallis group 1 contains more than three times more systems than Shannon group 1. Group 2 contains systems which produced both type-(i) and type-(ii) solutions upon optimization, namely 1QOP_AB, 1RM6_AB, 2NU9_AB, 3PNL_AB, 3IP4_BC, 3MML_AB and 3OAA_HG (Figure 3.21). Finally, group 3 i composed by systems which failed to produce solutions that clearly differenti-

ated from random solutions, namely 1I1Q_AB, 2D1P_AB, 2VPZ_AB, 2WDQ_AB and 3IP4_AB (Figure 3.22). Three systems are in the intersection of Shannon group 5 and Tsallis group 3, namely 1I1Q_AB, 2D1P_BC and 3IP4_AB, being the most difficult systems for no clear reason.

As a comparison, the same correlations shown in Figure 3.16 were reassessed within the framework of Tsallis statistics to make sense of the solutions obtained by maximizing Tsallis $\hat{I}_{AB}$ (Figure 3.23). The correlation between TP rate and number of sequences in the MSA went from 0.41 to 0.56, a moderate increase (Figure 3.23A). Meanwhile, the correlation between TP rate and interface size went from 0.73 to 0.61, a moderate decrease (Figure 3.23B). This means that, within the Tsallis statistics framework, TP rate seems to be less determined by the interface size, even though the correlation is still strong. The correlation between TP rate and native $\hat{I}_{AB}$ and went from 0.28 to 0.53, an increase of almost 100% (Figure 3.23C). This is noteworthy and desirable, since the optimization is guided by the informational content on the interface. The fact that this correlation is not so strong within the Shannon statistics framework might be an indication that the informational signal is not being captured well. Lastly, the correlation between TP rate and Mirror-Tree correlation went from 0.42 to 0.26, a significant decrease (Figure 3.23D). This indicates that optimization results became less dependent on the similarity of phylogenetic trees of the interacting protein families. This is a highly desirable feature for cases of host-pathogen protein interactions, in which no similarity of phylogenetic trees is expected.

Finally, building a connection with Chapter 2, the Pearson correlation between TP rate and the different kinds of information stored in protein systems interface was computed (Table 3.3). It is possible to see that within both statistical frameworks, negative correlation between stochastic information and TP rate is strong ($r < -0.5$). This makes sense, since the presence of stochastic information is expected to be a confounding factor in the process of searching for near-native solutions. Meanwhile, the correlation between TP rate and evolutive information is mild in both cases, presenting no differences when considering Shannon or Tsallis statistics. Notably, however, the correlation between TP rate and coevolutive information is significantly stronger within Tsallis statistics framework, reaching up to 0.52 when considering maximum TP rate, compared to -0.05 for Shannon. This last finding indicates that the fundamental difference between Tsallis and Shannon statistics lies in the latter differential capability of capturing signal from coevolutive information stored in protein systems interfaces.

Table 3.3: Pearson correlation between true positive rate (TPR), both average (n = 6 trajectories) and maximum, of MSA concatenation solutions obtained after Shannon or Tsallis $\hat{I}_{AB}$ maximization and different kinds of mutual information (stochastic, evolutive, coevolutive) stored on the interface. Correlation values were obtained for the set of 14 interacting protein systems considered in Chapter 2.

	Max. TPR Shannon	Avg. TPR Shannon	Max. TPR Tsallis	Avg. TPR Tsallis
Stochastic info.	-0.62	-0.59	-0.58	-0.64
Evolutive info.	0.24	0.38	0.36	0.35
Coevolutive info.	-0.05	0.19	0.52	0.43

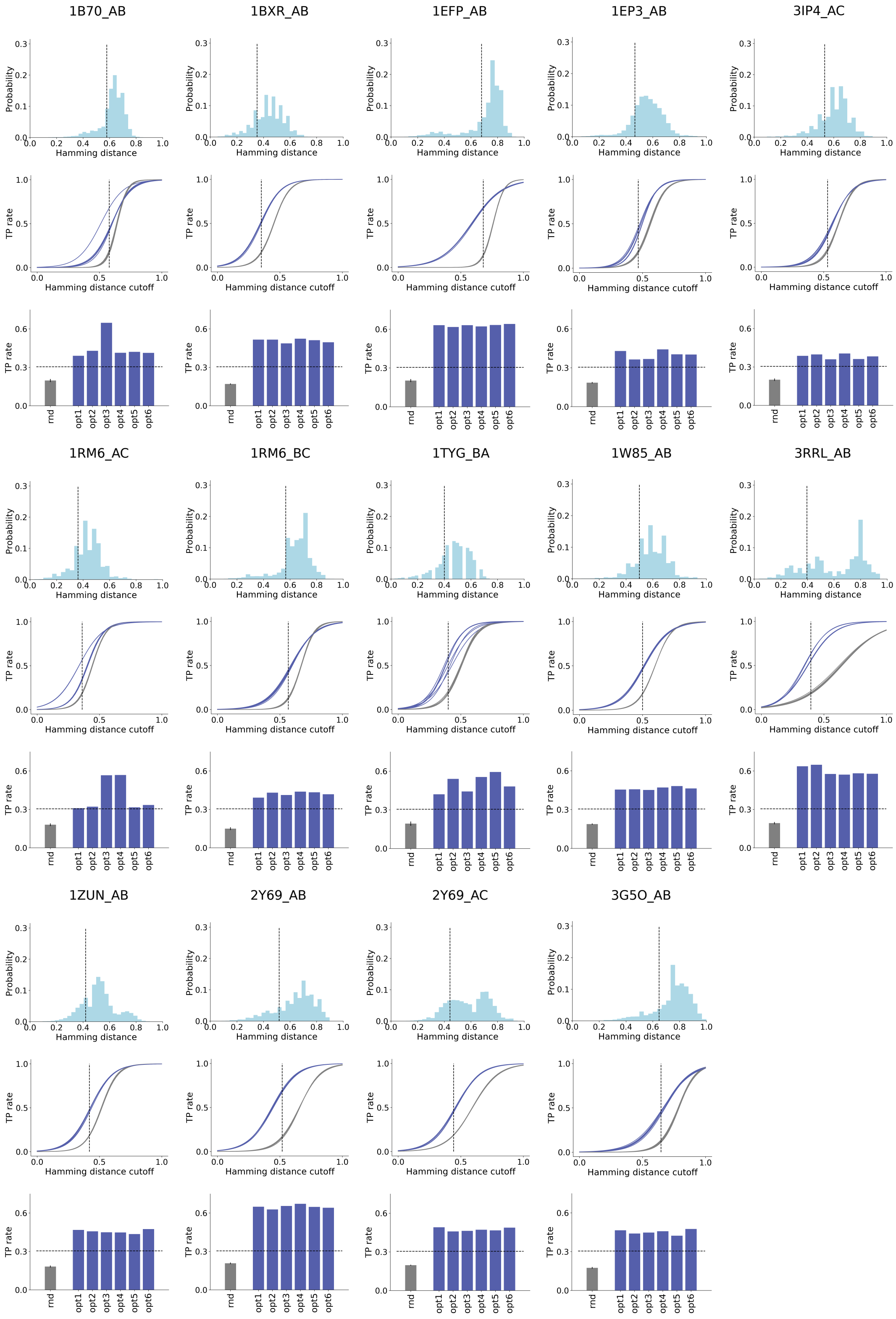


Figure 3.20: Tsallis group 1 - all solutions are type-(i), blue. (1st) Hamming distance distribution of MSA B. (2nd) True positive (TP) rate for different Hamming distance discounts. The 20th percentile is shown with a dashed line, random solutions in gray, optimized solution in blue. (3rd) TP rates of random (rnd) and optimized (opt1-6) solutions at 20th percentile Hamming distance cutoff. The significance value is shown with a dashed line (p=0.0005).

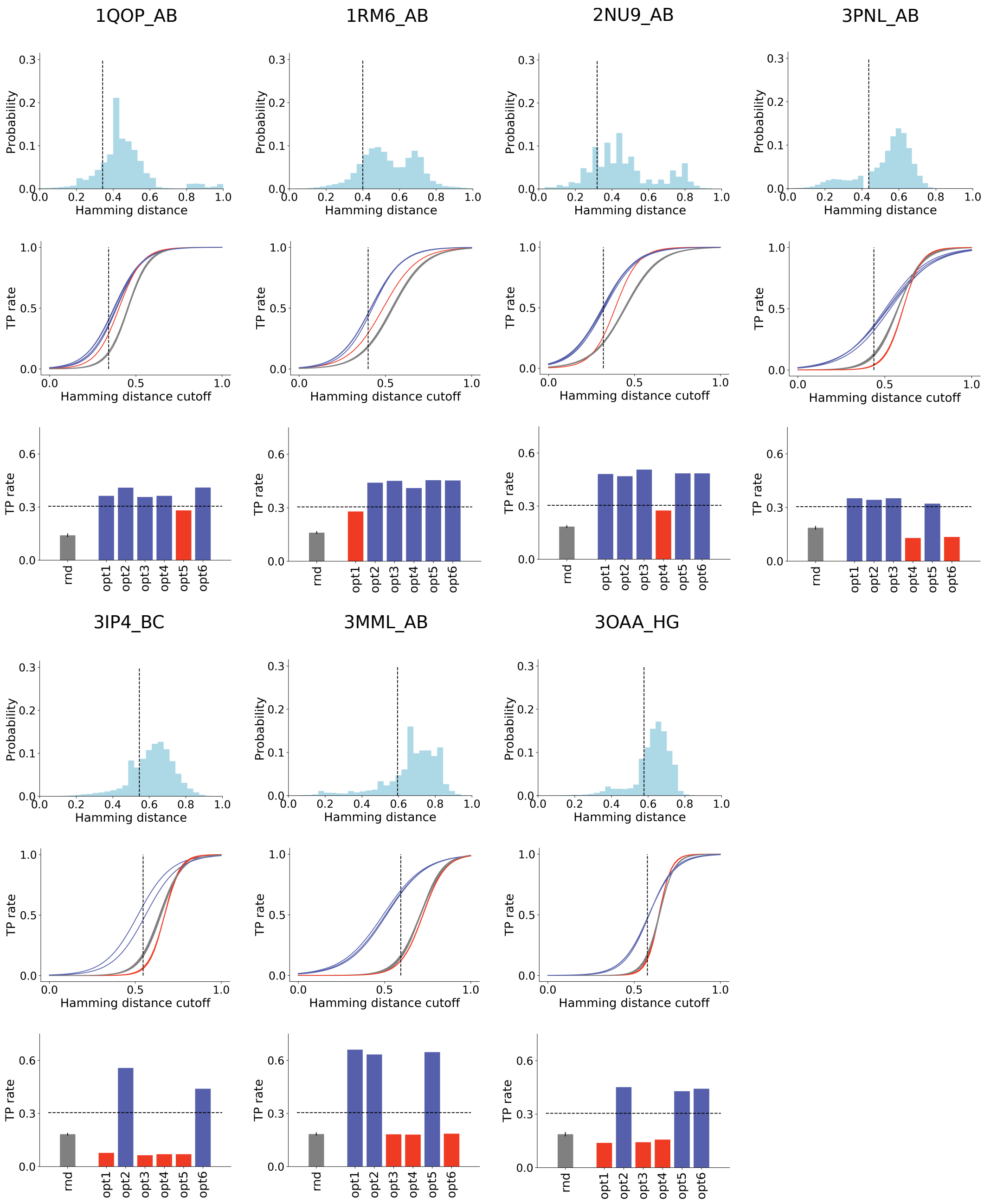


Figure 3.21: Tsallis group 2 - mixed type-(i) blue and type-(ii) red solutions. (1st) Hamming distance distribution of MSA B. (2nd) True positive (TP) rate for different Hamming distance discounts. The 20th percentile is shown with a dashed line, random solutions in gray, optimized solution in blue. (3rd) TP rates of random (rnd) and optimized (opt1-6) solutions at 20th percentile Hamming distance cutoff. The significance value is shown with a dashed line (p=0.0005).

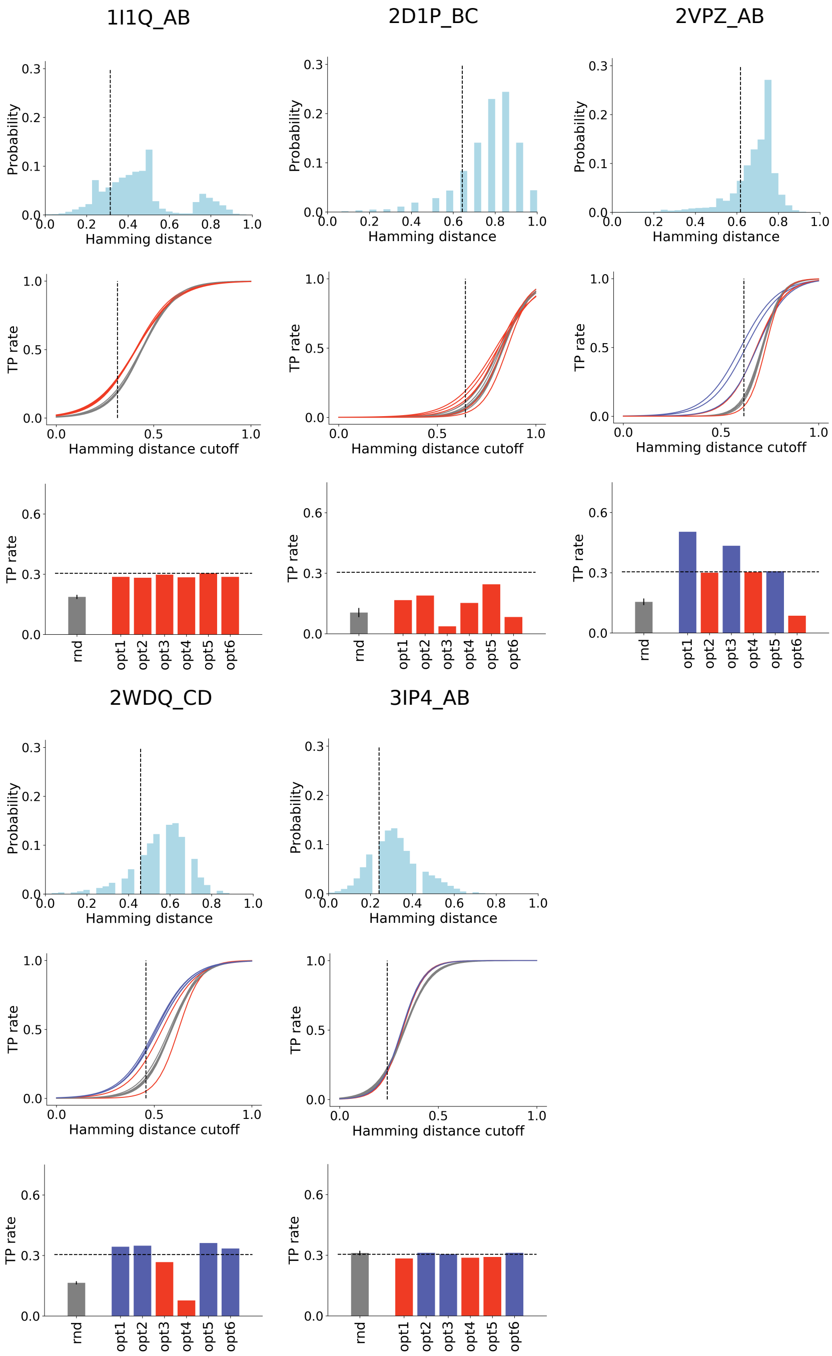


Figure 3.22: Tsallis group 3 - most solutions indistinguishable from random. (1st) Hamming distance distribution of MSA B. (2nd) True positive (TP) rate for different Hamming distance discounts. The 20th percentile is shown with a dashed line, random solutions in gray, optimized solution in blue. (3rd) TP rates of random (rnd) and optimized (opt1-6) solutions at 20th percentile Hamming distance cutoff. The significance value is shown with a dashed line (p=0.0005).

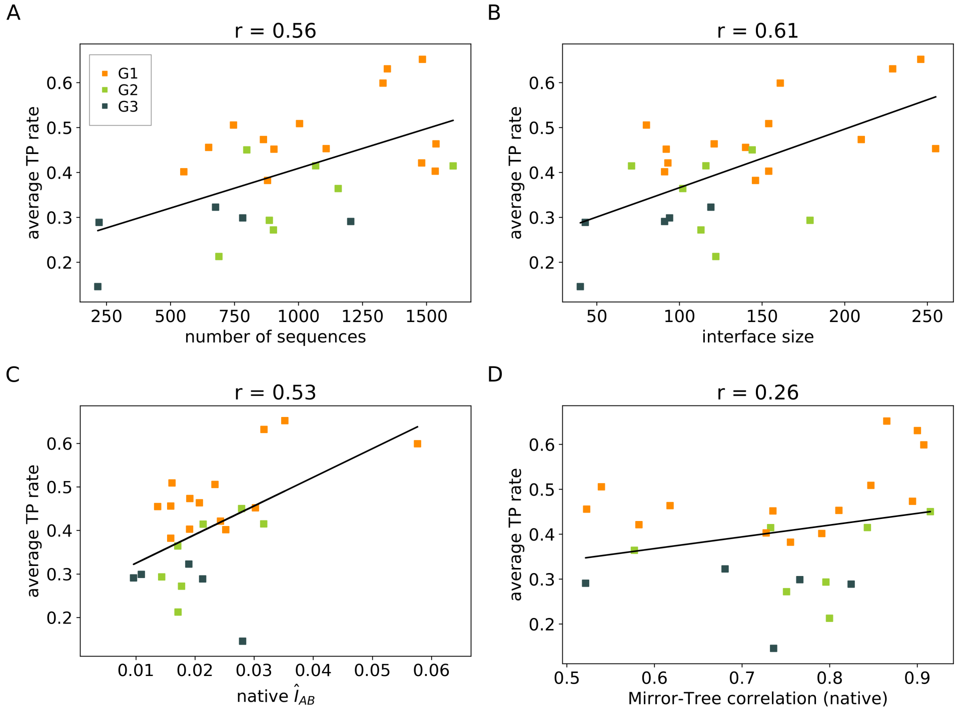


Figure 3.23: Pearson correlation between the average true positive (TP) rate of the optimized solutions (n = 6 for each system) and the number of sequences in the alignment (A), number of contacts on the interface (interface size) (B), mutual information per contact on the interface $\hat{I}_{AB}$ of the native solution (C), and Mirror-Tree correlation [2] of the native solution (D). Systems are colored by groups G1-3.

3.4 Conclusions

In this chapter, the hypothesis that mutual information (MI) encoded on the interacting amino acids of two proteins can be used to correctly discriminate protein partners based on long MSAs was investigated. It was previously found that the interface MI ($\hat{I}_{AB}$) has the strongest signal to distinguish protein partners and is likely the unique signal in a case of proteins in independent genomes [15], *e.g.* host-pathogen interactions. Inferring the correct set of specific protein partners, however, becomes increasingly complicated for longer MSAs, as the degeneracy of $\hat{I}_{AB}$ is expected to be large and may impose severe limitations to practical applications.

Indeed, $\hat{I}_{AB}$ maximization starting from scrambled MSA concatenations is shown here to resolve partners at very low true positive (TP) rates in consequence of two different error sources, called type-(i) and type-(ii). It is not surprising that $\hat{I}_{AB}$ is degenerated due to the existence of very similar sequences in the MSAs (type-(i) error). Unexpected, however, is the fact that degeneracy may also arise due to type-(ii) errors, which arise from mismatches among non-similar sequences. If type-(i) error sources are disregarded, further analysis indicates that the correct MSA pairing can be resolved at best TP rates of 70%. This shows that maximization of $\hat{I}_{AB}$ may be of some utility to obtain solutions that at least approximate the native one in terms of phylogenetic accuracy.

In a further step, the influence of different system properties on the results was analyzed, and it was observed that interface size is the factor most strongly correlated with TP rate (r = 0.73). This raises the question of whether considering only interface sites to calculate MI is really the best strategy. Even though the average per-contact coevolutive signal is stronger at the interface, when considering the information as a whole, it might be worth including more (weaker) signal sources. There might, however, be a trade-off involved in this decision, which will be investigated in future work.

The HK-RR system was investigated as a special case of a highly specific system of interacting paralogs. TP rates recovered through $\hat{I}_{AB}$ maximization were similar to TP rates reported in literature [78], which were reached with other more complex optimization algorithms, such as DCA-IPA [3]. Additionally, it was observed that the HK-RR system does not contain type-(i) errors, either due to its small interface or because of the high specificity of the protein partners. This suggests another layer of complexity that sequence diversity and specificity may add to the problem. In fact, type-(i) errors might only arise in systems with a single pair of interacting proteins per genome, since in this cases there will be no selective pressure to avoid cross-binding homologs occurring in other species, assuming that the interacting proteins have never been and will never be in contact.

Finally, upon comparison with similar results obtained using Tsallis statistics, it was observed that this novel proposed framework provides significantly better optimized solutions. In fact, native $\hat{I}_{AB}$ is more correlated with TP rates within Tsallis statistics (increase from 0.28 to 0.53), while the interface size is less correlated (decrease from 0.73 to 0.61). Also, TP rates are less correlated with Mirror-Tree correlation (decrease from 0.42 to 0.26) within Tsallis statistics, what indicates that the optimization is probably not being guided that shared evolutionary signal. Indeed, looking at the correlation of TP rates with the coevolutive component of the interface MI, it is observed that r changes from -0.05 in Shannon to 0.52 in Tsallis framework. This means that Tsallis MI is probably able to capture better and be guided by coevolutive signal, independent of shared evolutionary history of the interacting protein families.

Overall, the investigations performed in this chapter provide some clarifications into the general problem of protein coevolution from the perspective of sequence diversity. It is difficult to say to which point homologous sequences were selected to selectively bind to its native partners, since it appears to exist a huge degeneracy in the space of possible sets of partners. Despite the intrinsic complexity of the problem of specific partners prediction for large sequence *ensembles*, the novel theoretical insights presented in the present work provide relevant information for future studies and should contribute to advancing our knowledge in the field.

Chapter 4

Conclusions

In the present work, protein-protein interactions were investigated from a theoretical point of view regarding evolutionary and coevolutionary aspects using solely the information stored in primary sequences. In Chapter 2, a decomposition of the interface mutual information $\hat{I}_{AB}$ of different interacting proteins was carried out considering the hypothesis that $\hat{I}_{AB}$ is originated from a combination of coevolutive, evolutive and stochastic sources. This decomposition revealed that different protein systems present a variable content of evolutive and coevolutive information. Afterwards, it was shown that evolutive information is correlated with the similarity between the interacting protein families phylogenetic trees, while the coevolutive information is not. This observation is evidence of the distinct character of these two quantities, showing that they are not merely theoretical constructs. In the following, Shannon and Tsallis statistical frameworks were compared in their abilities to capture informational signal from MSAs, and Tsallis statistics yielded better results. Finally, this investigation originated (in its early stages) a co-first author publication (Appendix 1), in which we show that the interface of protein complexes stores the largest per-contact information to discriminated the correct set of protein partners for two interacting families.

Building on this last conclusion, in Chapter 3 a $\hat{I}_{AB}$-maximizing genetic algorithm (GA) was implemented in an attempt to find the correct set of specific protein partners for a set of interacting protein families, starting from a randomized solution. Optimized solutions, however, yielded negligible true positive (TP) values due to two different error sources, type-(i) and type-(ii). Solutions containing type-(i) error attained TP rates of up to 70% when mismatches among similar sequences were disregarded. This means that this type of optimized solution somehow contains a coarse-grained phylogenetic structure. Error type-(ii), in turn, happened due to mistakes made among non-similar sequences and could not be disregarded. Upon further analysis, it was observed that TP rates of optimized solutions strongly correlate with the complex interface size, raising doubts about whether considering only the signal extracted from the interface (instead of the whole protein) is really the better way to reach more accurate solutions. Also in this chapter, Shannon and Tsallis statistics were compared, with the latter yielding much better results than the first. In fact, Tsallis statistics seems to capture better the coevolutive signal, and not rely so much on the shared phylogenetic history of the two interacting protein families. This feature might be important to when solving the problem of finding specific protein partners among interacting proteins in independent genomes, like in host-pathogen interactions.

Finally, the work presented in Chapter 3 has also originated a paper, which was published in *Scientific Reports*. In conclusion, the present work provides some clarifications into the general problem of protein coevolution, both by characterizing different kinds of mutual information stored on the interface of protein complexes, and by making some sense of the complexity involved in finding the correct set of protein partners given a pair of interacting families. Notably, a different statistical framework tested here (Tsallis statistics) seems to be better than Shannon statistics at capturing important information from primary sequences, posing the question of whether the currently used standard statistical models could not be replaced by more accurate models. Despite the intrinsic complexity of the problems investigated here, the novel theoretical insights presented provide relevant information for future studies and should contribute to advancing our knowledge in the field.

Bibliography

[1] Miguel Andrade, Camila Pontes, and Werner Treptow. Coevolutive, evolutive and stochastic information in protein-protein interactions. *Computational and structural biotechnology journal*, 17:1429–1435, 2019.

[2] Florencio Pazos and Alfonso Valencia. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein engineering*, 14(9):609–614, 2001.

[3] Anne-Florence Bitbol, Robert S Dwyer, Lucy J Colwell, and Ned S Wingreen. Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences*, 113(43):12180–12185, 2016.

[4] L Van Valen. A new evolutionary law. 1973.

[5] Paul R Ehrlich and Peter H Raven. Butterflies and plants: a study in coevolution. *Evolution*, pages 586–608, 1964.

[6] Arthur M Lesk and Cyrus Chothia. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of molecular biology*, 136(3):225–270, 1980.

[7] Cyrus Chothia and Arthur M Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823–826, 1986.

[8] Cyrus Chothia and Joël Janin. Principles of protein–protein recognition. *Nature*, 256(5520):705–708, 1975.

[9] Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

[10] Aviva Presser, Michael B Elowitz, Manolis Kellis, and Roy Kishony. The evolutionary dynamics of the saccharomyces cerevisiae protein interaction network after duplication. *Proceedings of the National Academy of Sciences*, 105(3):950–954, 2008.

[11] Kirill Evlampiev and Hervé Isambert. Conservation and topology of protein interaction networks under duplication-divergence evolution. *Proceedings of the National Academy of Sciences*, 105(29):9863–9868, 2008.

[12] Michael PH Stumpf, William P Kelly, Thomas Thorne, and Carsten Wiuf. Evolution at the system level: the natural history of protein interaction networks. *Trends in Ecology & Evolution*, 22(7):366–373, 2007.

[13] John W Pinney, Grigoris D Amoutzias, Magnus Rattray, and David L Robertson. Reconstruction of ancestral protein interaction networks for the bzip transcription factors. *Proceedings of the National Academy of Sciences*, 104(51):20449–20453, 2007.

[14] Luke Hakes, Simon C Lovell, Stephen G Oliver, and David L Robertson. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences*, 104(19):7999–8004, 2007.

[15] Johannes Berg, Michael Lässig, and Andreas Wagner. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC evolutionary biology*, 4(1):51, 2004.

[16] DANIÈLE Altschuh, AM Lesk, AC Bloomer, and A Klug. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of molecular biology*, 193(4):693–707, 1987.

[17] Neil D Clarke. Covariation of residues in the homeodomain sequence family. *Protein Science*, 4(11):2269–2278, 1995.

[18] Gregory B Gloor, Louise C Martin, Lindi M Wahl, and Stanley D Dunn. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44(19):7156–7165, 2005.

[19] Chen-Hsiang Yeang and David Haussler. Detecting coevolution in and among protein domains. *PLoS Comput Biol*, 3(11):e211, 2007.

[20] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[21] Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317, 1994.

[22] Florencio Pazos, Manuela Helmer-Citterich, Gabriele Ausiello, and Alfonso Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of molecular biology*, 271(4):511–523, 1997.

[23] John Moult, Jan T Pedersen, Richard Judson, and Krzysztof Fidelis. A large-scale experiment to assess protein structure prediction methods, 1995.

[24] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[25] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[26] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[27] Simon C Lovell and David L Robertson. An integrated view of molecular coevolution in protein–protein interactions. *Molecular biology and evolution*, 27(11):2567–2575, 2010.

[28] David De Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.

[29] Chern-Sing Goh, Andrew A Bogan, Marcin Joachimiak, Dirk Walther, and Fred E Cohen. Co-evolution of proteins with their interaction partners. *Journal of molecular biology*, 299(2):283–293, 2000.

[30] Georg Casari, Chris Sander, and Alfonso Valencia. A method to predict functional residues in proteins. *Nature structural biology*, 2(2):171–178, 1995.

[31] Olivier Lichtarge, Henry R Bourne, and Fred E Cohen. An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, 257(2):342–358, 1996.

[32] Thomas Gueudré, Carlo Baldassi, Marco Zamparo, Martin Weigt, and Andrea Pagnani. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proceedings of the National Academy of Sciences*, 113(43):12186–12191, 2016.

[33] Anne-Florence Bitbol. Inferring interaction partners from protein sequences using mutual information. *PLoS computational biology*, 14(11):e1006401, 2018.

[34] Emmanuel D Levy. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of molecular biology*, 403(4):660–670, 2010.

[35] Loredana Lo Conte, Cyrus Chothia, and Joël Janin. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285(5):2177–2198, 1999.

[36] Susan Miller, Joel Janin, Arthur M Lesk, and Cyrus Chothia. Interior and surface of monomeric proteins. *Journal of molecular biology*, 196(3):641–656, 1987.

[37] Cyrus Chothia. The nature of the accessible and buried surfaces in proteins. *Journal of molecular biology*, 105(1):1–12, 1976.

[38] Erwin Neher. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences*, 91(1):98–102, 1994.

[39] William R Taylor and Kerr Hatrick. Compensating changes in protein multiple sequence alignments. *Protein Engineering, Design and Selection*, 7(3):341–348, 1994.

[40] Anthony A Fodor and Richard W Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2):211–221, 2004.

[41] Julian Mintseris and Zhiping Weng. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences*, 102(31):10930–10935, 2005.

[42] Hocine Madaoui and Raphaël Guerois. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proceedings of the National Academy of Sciences*, 105(22):7708–7713, 2008.

[43] Lukas Burger and Erik Van Nimwegen. Accurate prediction of protein–protein interactions from sequence alignments using a bayesian method. *Molecular systems biology*, 4(1):165, 2008.

[44] Lukas Burger and Erik Van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, 6(1):e1000633, 2010.

[45] Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, 2011.

[46] David T Jones, Daniel WA Buchan, Domenico Cozzetto, and Massimiliano Pontil. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, 2012.

[47] Chan-Seok Jeong and Dongsup Kim. Reliable and robust detection of coevolving protein residues. *Protein Engineering, Design & Selection*, 25(11):705–713, 2012.

[48] Dennis R Livesay, Kyle E Kreth, and Anthony A Fodor. A critical evaluation of correlated mutation algorithms and coevolution within allosteric mechanisms. In *Allostery*, pages 385–398. Springer, 2012.

[49] Thomas A Hopf, Lucy J Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, and Debora S Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–1621, 2012.

[50] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, 2013.

[51] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife*, 3:e02030, 2014.

[52] Christoph Feinauer, Hendrik Szurmant, Martin Weigt, and Andrea Pagnani. Inter-protein sequence co-evolution predicts known physical interactions in bacterial ribosomes and the trp operon. *PloS one*, 11(2):e0149166, 2016.

[53] Raphaël Champeimont, Elodie Laine, Shuang-Wei Hu, Francois Penin, and Alessandra Carbone. Coevolution analysis of hepatitis c virus genome to identify the structural and functional dependency network of viral proteins. *Scientific reports*, 6(1):1–20, 2016.

[54] Francisco M Codoñer and Mario A Fares. Why should we care about molecular coevolution? *Evolutionary Bioinformatics*, 4:117693430800400003, 2008.

[55] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[56] Nikolay V Dokholyan, Leonid A Mirny, and Eugene I Shakhnovich. Understanding conserved amino acids in proteins. *Physica A: Statistical Mechanics and its Applications*, 314(1-4):600–606, 2002.

[57] William SJ Valdar. Scoring residue conservation. *Proteins: structure, function, and bioinformatics*, 48(2):227–241, 2002.

[58] John A Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.

[59] Fredrik Johansson and Hiroyuki Toh. A comparative study of conservation and variation scores. *BMC bioinformatics*, 11(1):388, 2010.

[60] Julian Echave, Stephanie J Spielman, and Claus O Wilke. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17(2):109, 2016.

[61] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[62] Sumiyoshi Abe. Stability of tsallis entropy and instabilities of rényi and normalized tsallis entropies: A basis for q-exponential distributions. *Physical Review E*, 66(4):046134, 2002.

[63] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.

[64] J-F Bercher. Tsallis distribution as a standard maximum entropy solution with 'tail'constraint. *Physics Letters A*, 372(35):5657–5659, 2008.

[65] Richard E Lenski. What is adaptation by natural selection? perspectives of an experimental microbiologist. *PLoS genetics*, 13(4):e1006668, 2017.

[66] Jason Gertz, Georgiy Elfond, Anna Shustrova, Matt Weisinger, Matteo Pellegrini, Shawn Cokus, and Bruce Rothschild. Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, 19(16):2039–2045, 2003.

[67] Matteo Pellegrini, Edward M Marcotte, Michael J Thompson, David Eisenberg, and Todd O Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[68] Thomas Dandekar, Berend Snel, Martijn Huynen, and Peer Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences*, 23(9):324–328, 1998.

[69] Cynthia J Verjovsky Marcotte and Edward M Marcotte. Predicting functional linkages from gene fusions with confidence. *Applied bioinformatics*, 1(2):93–100, 2002.

[70] Elisabeth RM Tillier, Laurence Biro, Ginny Li, and Desiree Tillo. Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. *Proteins: Structure, Function, and Bioinformatics*, 63(4):822–831, 2006.

[71] Florencio Pazos and Alfonso Valencia. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins: Structure, Function, and Bioinformatics*, 47(2):219–227, 2002.

[72] Miguel Correa Marrero, Richard GH Immink, Dick de Ridder, and Aalt DJ van Dijk. Improved inference of intermolecular contacts through protein–protein interaction prediction using coevolutionary analysis. *Bioinformatics*, 35(12):2036–2042, 2019.

[73] Mohamed Barakat, Philippe Ortet, Cécile Jourlin-Castelli, Mireille Ansaldi, Vincent Méjean, and David E Whitworth. P2cs: a two-component system resource for prokaryotic signal transduction research. *BMC genomics*, 10(1):1–10, 2009.

[74] Mohamed Barakat, Philippe Ortet, and David E Whitworth. P2cs: a database of prokaryotic two-component systems. *Nucleic acids research*, 39(suppl_1):D771–D776, 2011.

[75] Philippe Ortet, David E Whitworth, Catherine Santaella, Wafa Achouak, and Mohamed Barakat. P2cs: updates of the prokaryotic two-component systems database. *Nucleic acids research*, 43(D1):D536–D541, 2015.

[76] Michael T Laub and Mark Goulian. Specificity in two-component signal transduction pathways. *Annu. Rev. Genet.*, 41:121–145, 2007.

[77] Michael A Rowland and Eric J Deeds. Crosstalk and the evolution of specificity in two-component signaling. *Proceedings of the National Academy of Sciences*, 111(15):5550–5555, 2014.

[78] Guillaume Marmier, Martin Weigt, and Anne-Florence Bitbol. Phylogenetic correlations can suffice to infer protein partners from sequences. *PLoS computational biology*, 15(10):e1007179, 2019.

Appendix 1: Paper published at the Computational and Structural Biotechnology Journal

Computational and Structural Biotechnology Journal 17 (2019) 1429–1435

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Coevolutive, evolutive and stochastic information in protein-protein interactions

Miguel Andrade, Camila Pontes, Werner Treptow *

Laboratório de Biologia Teórica e Computacional (LBTC), Universidade de Brasília DF, Brazil

ARTICLE INFO

Article history:
Received 1 August 2019
Received in revised form 19 October 2019
Accepted 22 October 2019
Available online 20 November 2019

Keywords:
Coevolution
Mutual information
Protein-protein interaction
Protein network
Evolution

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.

1. Introduction

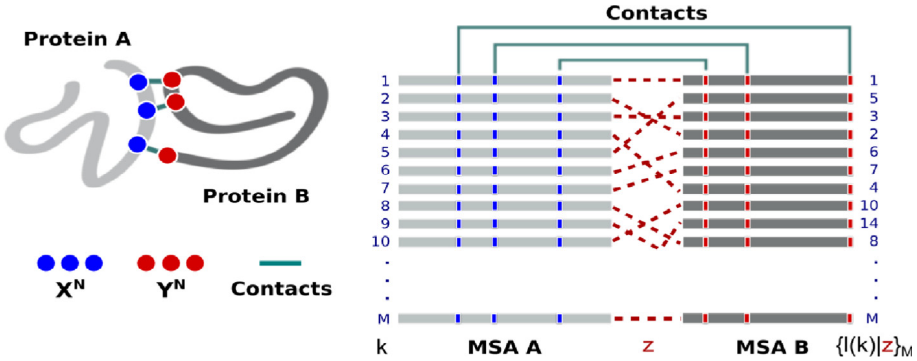
While being selected to be thermodynamically stable and kinetically accessible in a particular fold [1,2], interacting proteins *A* and *B* coevolve to maintain their bound free-energy stability against a vast repertoire of non-specific partners and interaction modes. Protein coevolution, in the form 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 coevolutive, evolutive and stochastic information in resolving the problem. Differentiating functional coevolution from stochastic and phylogenetic sources remains looked for in the research field and may help introducing models capable of accurately detecting protein-protein interactions and interfaces, especially when the number of sequences or the amount of biological information are limited [25].

Here, by benefiting from much larger data sets 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 discriminating PPIs based on MSAs (Scheme 1). Specifically, under the assumption that the coevolutive 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 the information to be dissected in terms of compensatory mutations at

* Corresponding author.
E-mail address: treptow@unb.br (W. Treptow).

https://doi.org/10.1016/j.csbj.2019.10.005
2001-0370/© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.



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 **l** in MSA B to a sequence **k** in MSA *A* in one unique arrangement, $\{l(k)|z\}_M$, determined by the coevolution process *z* to which these protein families were subjected. Shown is a "scrambled" concatenated MSA of A and B associated to a given process *z* (red dashes).

physically-coupled and uncoupled amino acids of A and B. Given a known set of protein three-dimensional amino-acid contacts and their underlying primary sequences we seek therefore differentiating functional coevolution from stochastic and phylogenetic signals for subsequent evaluation of their contributions in PPI recognition from primary sequences. It is worth emphasizing our study is not aimed at providing a method for prediction of protein-protein interactions nor protein-protein interfaces, hence it differs from previous studies in which sequence covariance is used to predict three-dimensional amino-acid contacts across interfaces and assemble models of protein complexes [26] or protein docking [27]. 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, including native and non-native decoy structures. A significant fraction of that information results from coevolutive sources alone. Although, our analysis involved protein interologs under cospeciation that is, proteins evolving in the same genome, the derived conclusions are likely general to cases of non-cospeciating interologs given that the underlying thermodynamical principles must be the same for all cases.

2. Theory and methods

2.1. Decomposition of mutual information

In detail, consider two proteins *A* and *B* that interact via formation of *i = 1,...,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 $X^N \equiv (X_1,\ldots,X_N)$ and $Y^N \equiv (Y_1,\ldots,Y_N)$ with probability mass functions $\{\rho(x^N),\rho(y^N),\rho(x^N,y^N|z)\}$ such that,

$$\begin{cases} \rho\left(x^N\right) = \sum_{y^N} \rho\left(x^N, y^N|z\right) \\ \rho\left(y^N\right) = \sum_{x^N} \rho\left(x^N, y^N|z\right) \end{cases} \tag{1}$$

and

$$\sum_{x^N,y^N} \rho\left(x^N, y^N|z\right) = 1 \tag{2}$$

for every joint sequence $\{x^N, y^N\}_{|\chi|^{2N}}$ defined in the alphabet χ 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* [28]. 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

$$I\left(X^N; Y^N|z\right) = H\left(X^N\right) + H\left(Y^N\right) - H\left(X^N, Y^N|z\right) \tag{3}$$

writes in terms of the Shannon's information entropies

$$\begin{cases} H\left(X^N\right) = \sum_{x^N} \rho^*\left(x^N\right) 1n\rho^*\left(x^N\right) \\ H\left(Y^N\right) = -\sum_{y^N} \rho^*\left(y^N\right) 1n\rho^*\left(y^N\right) \\ H\left(X^N, Y^N|z\right) = -\sum_{x^N,y^N} \rho^*\left(x^N, y^N|z\right) 1n\rho^*\left(x^N, y^N|z\right) \end{cases} \tag{4}$$

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*

$$\rho^*\left(x^N, y^N|z\right) = \prod_{i=1}^{N} \rho^*(x_i, y_i|z) \tag{5}$$

then allowing Eq. (4) to be written extensively, in terms of the individual entropic contributions

$$\begin{cases} H\left(X^N\right) = \sum_i H(X_i|z) \\ H\left(Y^N\right) = \sum_i H(Y_i|z) \\ H\left(X^N, Y^N|z\right) = \sum_i H(X_i, Y_i|z) \end{cases} \tag{6}$$

such that,

$$I\left(X^N;Y^N|z\right)=\sum_{i=1}^{N}I(X_i;Y_i|z) \tag{7}$$

(*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 perfectly correlated variables $\rho^*(x^N, y^N|z) = \rho^*(x^N) = \rho^*(y^N)$, the conditional mutual information is bound to a maximum which cannot exceed the entropy of either block variables $H(X^N)$ and $H(Y^N)$.

Given a known set of protein amino-acid contacts and their underlying primary sequence distributions defining the stochastic variables X^N and Y^N, 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)

$$\rho^*(x_i,y_i|z)=\sum_{x'_1,\ldots,x'_N,y'_1,\ldots,y'_N}\delta_{x'_iy'_ix_iy_i}\rho^*\left(x'_1,\ldots,x'_N,y'_1,\ldots,y'_N|z\right)\equiv f_{x_iy_i|z} \tag{8}$$

from the observed frequencies $\mathbf{f}=\{\mathrm{f}_{x_i,y_i|z}\}$ in the multiple-sequence alignment

$\{x_k^N, y_l^N|z\}_M$

where the N-length amino-acid block l of protein B is joint to block k of protein A in one unique arrangement $\{l(k)|z\}_M$ for $1 \le k \le M$ (*cf.* Scheme 1 and Computational Methods).

2.2. Computational methods

Table 1 details the interacting protein systems considered in the study. For each system under investigation, amino-acid contacts defining the discrete stochastic variables X^N and Y^N including physically coupled amino acids at short-range cut-off distances ($r_c \le 8.0$ Å) and physically uncoupled amino-acids at long-range cut-off distances ($r_c > 8.0$ Å) were identified from the x-ray crystal structure of the bound state of proteins A and B. The reference (native) multi-sequence alignment $\{x_k^N, y_l^N|z^*\}_M$ of the joint amino-acid blocks associated to X^N and Y^N was reconstructed from annotated primary-sequence alignments published by Baker and coworkers [22], containing M paired sequences with known protein-protein interactions and defined in the alphabet of 20 amino acids plus the gap symbol ($|\chi|$=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 $\rho^*(x_i,y_i|z)\equiv f_{x_iy_i|z}$ were defined from the observable frequencies $f_{x_iy_i|z}$ regularized by a pseudocount effective fraction λ^* in case of insufficient data availability as devised by Morcos and coauthors [19]. More specifically, two-site frequencies were calculated according to

$$f_{x_iy_i|z}=\frac{\lambda^*}{|\chi|^2}+(1-\lambda^*)\frac{1}{M_z^{eff}}\sum_{m=1}^{M}\frac{1}{n_z^m}\delta_{x_i^my_i^m|z,x_iy_i|z} \tag{9}$$

where, $n_z^m = |\{m'|1 \le m' \le M, \mathit{Hamming\ Disatnce}(m,m') \ge \delta h\}|$ is the number of similar sequences m' within a certain Hamming distance δh of sequence m and $M_z^{eff}=\sum_{m=1}^{M}\left(n_z^m\right)^{-1}$ is the effective number of distinguishable primary sequences at that distance threshold – the Kronecker delta $\delta_{x_i^my_i^m|z,x_iy_i|z}$ 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 $\frac{1}{|x|^2}$ for $\lambda^* = 1$; Eq. (9) is identical to the equation devised by Morcos and coauthors [19] by rewriting $\lambda^* = \lambda/(\lambda + M_z^{eff})$. Here, two-site probabilities $\rho^*(x_i,y_i|z)\equiv f_{x_i,y_i|z}$ 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 $M_z^{eff} = M$ (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 $\{\rho(x^N), \rho(y^N)\}$ were derived from $\rho^*(x_i, y_i|z)$ 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 $I(X_i; Y_i|z)$ is bound to the conditional joint entropy, Eq. (7) was computed in practice as a per-contact entropy-weighted conditional mutual information [29], $H(X_i; Y_i|z)^{-1}\, I(X_i; Y_i|z)$, to avoid that contributions of $H(X_i, Y_i|z)$ contacts between highly variable sites are overestimated. Because $H(X_i, Y_i \mid z)$ and $I(X_i, Y_i|z)$ have units of *nats*, Eq. (7) is dimensionless in the present form.

3. Results and discussion

Details of all protein systems under investigation are presented in Table 1. Each system involves two families of protein interologs A and B with known PPIs derived from cospeciation in the same genome [26]. We denote by $\{x_k^N, y_l^N|z^*\}_M$ 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*

Table 1
Protein system A and B considered in the study.

	Complex description	PDB ID	Protein A	Protein B	M	MSA length
Obligate Dimers	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
Non-Obligate Dimer	Toxin-antitoxin complex RelBE2 from Mycobacterium tuberculosis	3G5O	**ChainA:** Protein Rv2865	**ChainB**: Protein Rv2866	904	173

coli. Similar results and conclusions hold for all other systems in Table 1 as presented in supplementary Figs. S1 through S4 (cf. SI).

3.1. Decomposition of mutual information

Fig. 1A shows the three-dimensional representation of stochastic variables 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 ($r_c \leq 8.0$ Å) and physically uncoupled amino-acids at long-range cutoff distances ($r_c > 8.0$ Å). In Fig. 1B, the total mutual information (coupled + uncoupled) across every possible amino-acid pairs of *A* and *B* amounts to 987.88 in the reference (native) MSA. As estimated from a generated ensemble of "scrambled" MSA models, expectation values for the mutual information $<I(X^N;\ Y^N|z)>_{M-n}$ decreases significantly as decorrelation or the number of mismatched proteins in the reference MSA increases. The result also holds at the level of individual protein contacts *i* as the mutual information $I(X_i;\ Y_i|z^*)$ for the reference alignment is systematically larger than the mutual information expectation value for "scrambled" MSA models full of sequence mismatches that is, with a total number *M* of mismatched sequences (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, the mutual information of fully "scrambled" models is found here to be very similar to the same estimate from randomized sequence alignments featuring aleatory swapping of lines within columns. Subtraction of that stochastic source from the native mutual information, as computed in the form of an information gap

$$\Delta I_{M-n} \equiv \left| I\left(X^N;\ Y^N|z^*\right) - \left\langle I\left(X^N;\ Y^N|z\right)\right\rangle_{M-n} \right| \tag{10}$$

between the reference MSA and "scrambled" models full of sequence mismatches, then reveals the isolated nonstochastic contributions to the total correlation between proteins *A* and *B*. Here, the information gap amounts to ~440 for every possible amino-acid pairs of *A* and *B*.

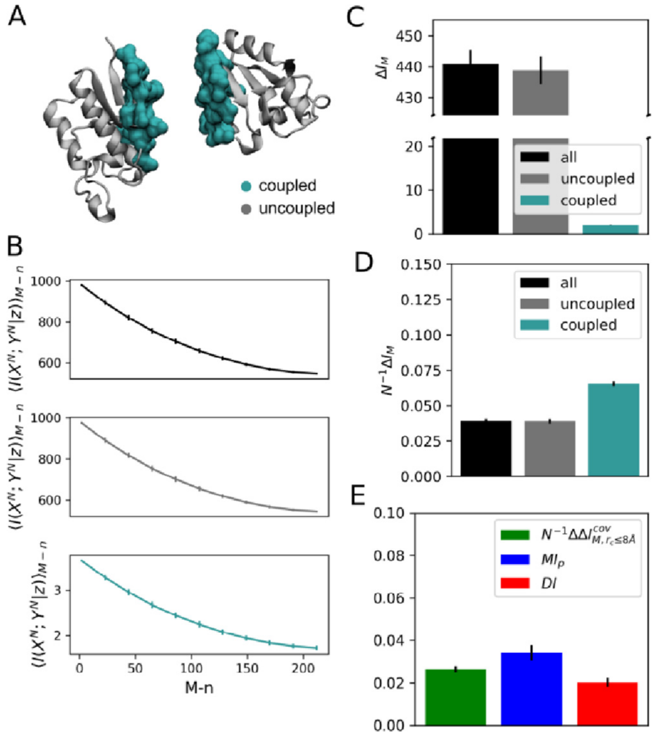


Fig. 1. Informational analysis of protein complex TusBCD, chains *B* and *C*. (A) Three-dimensional representation of stochastic variables X^N and Y^N as defined from physically coupled amino acids at short-range cutoff distances $r_c \leq 8.0\ \mathring{A}$ (turquoise) and physically uncoupled amino-acids at long-range cutoff distances $r_c > 8.0\ \mathring{A}$ (gray). Calculation of r_c involved C^β-C^β atomic separation distances. (B) Conditional mutual information $<I(X^N;\ Y^N|z)>_{M-n}$ as a function of the number $M - n$ of randomly paired proteins in the reference (native) MSA, for $0 \leq n \leq M$. $<\mathrm{I}(X^N;\ Y^N|z)>_{M-n}$ are expectation values estimated from a generated ensemble of 500 MSA models. Mutual information of fully "scrambled" models featuring *M* unpaired sequences is similar to that calculated from randomized sequence alignments generated by aleatory swapping of lines within columns. (C) Mutual information gap $\Delta \mathrm{I}_\mathrm{M}$ between reference and 100 fully "scrambled" models featuring M unpaired sequences. (D) Per-contact mutual information gap $\mathrm{N}^{-1}\Delta I_{M,rc}$. (E) Mutual information decomposition $\left(N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\text{Å}}\right)$ according to Eq. (11) and comparison with functional mutual information ($MI_{\mathrm{p},\mathrm{rc}\leq 8\mathring{A}}$) and direct information ($DI_{rc\leq 8\mathring{A}}$). In B, C, D and E error bars correspond to standard deviations.

Fig. 1C shows the individual contributions of physically coupled and uncoupled amino acids to the total mutual information gap, $\Delta I_M = \Delta I_{M,rc\leq 8.0\text{Å}} + \Delta I_{M,rc>8.0\text{Å}}$. As a direct consequence of the extensive property of Eq. (7), individual contributions to the total mutual information gap ($\Delta I_{M,r_c}$) increase with cutoff distances defining amino-acid contacts (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 ± 4.5159). When normalized by the block length or the number of amino-acid contacts (Fig. 1D), the mutual-information contribution N^{-1}-$\Delta I_{M,rc}$ reveals a distinct dependence being larger for physically coupled amino acids than uncoupled ones (0.0653 ± 0.0015 versus 0. 039 ± 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 information-gap values at short 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,

$$N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\text{Å}} \overset{def}{\equiv} N^{-1}\Delta I_{M,r_c\leq 8\text{Å}} - N^{-1}\Delta I_{M,r_c\leq 8\text{Å}} \tag{11}$$

where, $N^{-1}\Delta_{IM,r_c>8\text{Å}}$ represents the per-contact mutual information resulting from evolution. As shown in Fig. 1E, the information content resulting from coevolution alone amounts to 0.0264 ± 0.0014 which compares well to independent measures of coevolutionary information i.e., functional mutual information ($MI_{p,r_c\leq 8\text{Å}}$) [29] and direct information ($DI_{r_c\leq 8\text{Å}}$) [19], 0.0340 ± 0.0037 and 0.0202 ± 0.0019. More specifically, MI_p is a metric formulated by Dunn and coworkers [29] in which mutual information is subtracted from structural or functional relationships 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 ~40% of the information content stored in physically coupled amino acids of the protein complex TusBCD results from coevolutive sources alone.

3.2. 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 the total mutual-information component $N^{-1}\Delta I_{M,r_c}$ provides us with an unbiased (intensive) estimate for proper comparison of the information content between coupled and uncoupled amino acids, in the following, we focus our attention on $N^{-1}\Delta I_{M,r_c}$ to dissect their effective contributions in determining PPIs based on sequence alignments. Accordingly, let us define the total number ω_S of native-like MSA models generated by scrambling of $M-n$ sequence pairs in the reference alignment

$$\omega_{S(r_c)} \equiv \sum_{n\in S(r_c)} \omega_{M,n} \tag{12}$$

in terms of *rencontres* numbers $\omega_{M,n}$

$$\omega_{M,n} = \frac{M!}{n!}\sum_{q=0}^{M-n}\frac{(-1)^q}{q!} \tag{13}$$

or permutations of the reference sequence set $\{l(k)|z^*\}_M$ with n fixed positions satisfying $\sum_{n=0}^{M} \omega_{M,n} = M!$ (in combinatorics language). Here, $S(r_c)$ denotes the set of fixed positions n

$$S(r_c) \equiv \left\{ n | 0 \le n \le M,\ N^{-1}\Delta I_{M-n,r_c} \le \delta I \right\} \tag{14}$$

for which the mutual information gap $N^{-1}\Delta I_{M-n,r_c}$ is smaller than a certain resolution δI independently from the corresponding block length N or the number of amino-acid contacts. In simple terms, ω_S in Eq. (12) informs us on the degeneracy or the number of "scrambled" MSA models with a similar amount of mutual information of that in the reference (native) alignment.

As shown in Table S1, rencontres numbers $\omega_{M,n}$ is an astronomically increasing function of $M-n$, identical for any definition of the stochastic variables 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, the total number ω_S of native-like MSA models depends on the stochastic variables at various resolutions δI (Fig. 2A). That number 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 unpaired sequences required to perturb $N^{-1}\Delta I_{M-n,r_c}$ of a fixed change δI such that ω_S accumulates less over MSA models satisfying the condition $N^{-1}\Delta I_{M-n,r_c} \le \delta I$ in Eq. (14) (Fig. 2B).

The degeneracy of *native-like* MSA models at a given resolution depends on the cutoff distance defining stochastic variables (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

$$\langle \varepsilon \rangle_S = \sum_{n\in S}\left(M \sum_{n\in S} \omega_{M,n} \right)^{-1} n\,\omega_{M,n} \tag{15}$$

for the fraction $M^{-1}n$ of primary sequence matches among native-like MSA models decreases substantially with the degeneracy of such models meaning that $<\varepsilon>_S$ is systematically larger for physically-coupled amino-acids at various mutual-information resolutions δI (Fig. 2C). For instance, the fraction of matches at $\delta I = 0.02$ is ~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 mutual-information resolutions suggests even larger differences in the expectation fraction of PPI matches between short and long-range correlations of *A* and *B*.

3.3. Dependence with contact definition and docking decoys

So far, "contact" is actually any given pair of residues "i" in protein *A* and "j" in protein *B* within a given distance r_c^* which can be redefined for a variety of cutoff distances. Specifically, our results were determined by defining physically coupled amino acids at short-range cutoff distances ($r_c \le r_c^*$) and physically uncoupled amino-acids at long-range cutoff distances ($r_c > r_c^*$) for a typical "contact" geometrical definition involving C^β-C^β atomic separation distances of 8.0 Å (that is, $r_c^* \overset{def}{\equiv} 8.0$ Å). In the following, amino-acid "contacts" are loosely redefined for a variety of cutoff distances to study the dependence of the information encoded in short and long-range protein interactions with r_c^*. Further analysis shows a clear dependence of the per-contact mutual information gap (N^{-1}-$\Delta I_{M,rc}$) 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 $r_c^* \approx 8.0$ 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 subtracted from structural-functional relationships (MI_P) as a function of r_c^*.

Still, the information encoded in short and long-range amino-acid interactions was analyzed across the native binding interface

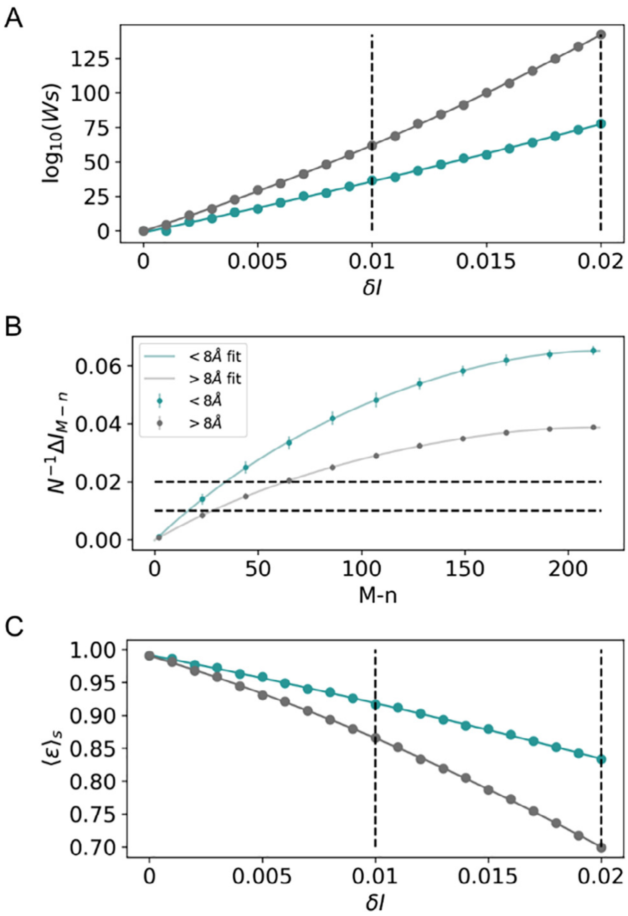


Fig. 2. Degeneracy and error analysis for stochastic variables X^N and Y^N involving interacting amino acids at short-range distances $r_c \le 8.0$ Å (turquoise) and long-range distances $r_c > 8.0$ Å (gray). (A) Total number ω_S of native-like MSA models at various mutual-information resolutions δI. (B) Per-contact gaps of mutual information $N^{-1}\Delta I_{M-n,rc}$ as a function of the number $M-n$ of "scrambled" sequence pairs in the reference native alignment. (C) Expectation values $<\varepsilon>_S$ (Eq. (15)) for the fraction of sequence matches across native-like MSA models at various mutual-information resolutions δI. Dashed lines highlight differences at δI values of 0.01 and 0.02.

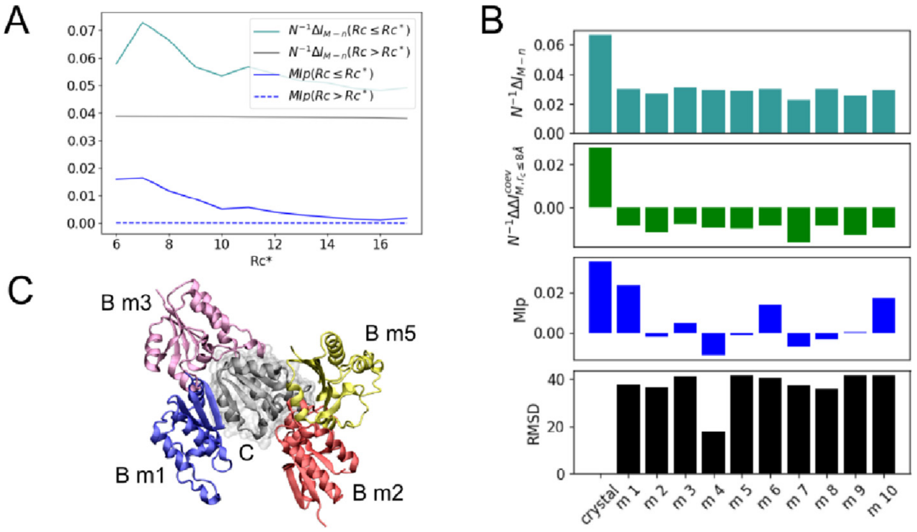


Fig. 3. Dependence with contact definition r_c* and docking decoys. (A) Per-contact mutual information gap $N^{-1}\Delta I_{M,r_c}$ and mutual information subtracted from structural-functional relationships $MI_{p,rc}$ at various r_c*. (B) Per-contact mutual information gap $N^{-1}\Delta I_{M,r_c}$ (turquoise), information content resulting from coevolution alone $N^{-1}\Delta\Delta I^{Cov}_{M,r_c}$ (green) and mutual information subtracted from structural or functional relationships MI_{p,r_c} (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 in Å units) between the native bound structure and docking decoys as generated by GRAMM-X [30]. Docking solutions were selected following a stability binding-energy criterium according to the scoring function of GRAMM – all docking decoys considered in the study are low-energy configurations despite large RMSD values relative to the native structure. (C) Illustration of four docking decoys of chain B in the protein complex TusBCD (chain C is shown in gray).

between proteins as revealed by x-ray crystallography experiments. The dependence of the per-contact mutual information gap with non-native binding modes or docking decoys of proteins *A* and *B* was then analyzed further, at the typical definition of amino-acid contacts ($r_c^* \overset{def}{\equiv} 8.0$ Å). Shown in Fig. 3B, there is a clear dependence of the information gap with binding modes – the per-contact mutual information gap reaches a maximum at the experimentally-determined native 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 in Fig. 3B by noticing that the isolated coevolutive content for the bound configuration of *A* and *B* or the associated mutual information subtracted from structural-functional relationships are larger than the very same estimates for any docking decoys.

4. 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 at $r_c^* \approx 8.0$ 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 of native-like MSA models and, therefore, to the largest expectation fraction of PPI matches across such models. These findings hold against any other definition of protein contacts, either across a variety of limitrophe distances discriminating coupled and uncoupled amino acids or alternative binding interfaces in docking decoys. Although presented for the protein complex TusBCD, results and discussion also extent to other protein systems, including obligate and non-obligate dimers, as shown in supplementary Figs. 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 protein, neurotoxins and ion channels, to mention a few. Accordingly, to add efforts in the field, we have addressed the following questions in our study: knowing three-dimensional amino-acid contacts from x-ray crystal structures, what would be the information encoded by them in terms of stochastic, evolutive and coevolutive sources, and what would be the utility of such pieces of information in resolving PPIs from "scrambled" multi-sequence alignments. Since the *Direct Information* derived from modern coevolutionary approaches [19,22] already filters out most of the information sources, the decomposition as proposed here does only make sense by considering the Mutual Information embodying unfiltered information. In this regard, it is worth emphasizing that our goals are neither the resolution of pair of residues highly-correlated via direct physical coupling [19,22] nor to provide with a method for prediction of protein-protein interactions and interfaces [26,27].

Although our study is not aimed at providing an approach for PPI prediction, the largest amount of non-stochastic information available in primary sequences helpful to differentiate MSA models with the largest expectation fraction of sequence matches as found here, might be of practical relevance in search of more effective heuristics to resolve protein-protein interactions from "scrambled" 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 the non-stochastic information of uncoupled amino acids must vanish in independent proteins – indeed, low information between amino acid positions of multiple sequence alignments is typically indicative of independently evolved proteins. Developments of more effective heuristics based on that signal would be applied for resolution of the more general problem of PPIs under coevolution in independent genomes, providing us with a highly welcome advance in the field.

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.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Comments of Antônio Francisco P. de Araújo. Fernando Melo, Georgios Pappas 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.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2019.10.005.

References

[1] Garcia LG, Treptow WL, Pereira de Araújo AF. Folding simulations of a three-dimensional protein model with a nonspecific hydrophobic energy function. Phys Rev E 2001;64:011912.

[2] Treptow WL, Barbosa MAA, Garcia LG, de Araújo AFP. Non-native interactions, effective contact order, and protein folding: A mutational investigation with the energetically frustrated hydrophobic model. Proteins Struct Funct Bioinforma 2002;49:167–80.

[3] Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–17.

[4] de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet 2013;14:249–61.

[5] Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Co-evolution of proteins with their interaction partners. J Mol Biol 2000;299:283–93.

[6] Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng 2001;14:609–14.

[7] Gertz J et al. Inferring protein interactions from phylogenetic distance matrices. Bioinforma Oxf Engl 2003;19:2039–45.

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

[9] Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 1998;23:324–8.

[10] Marcotte CJV, Marcotte EM. Predicting functional linkages from gene fusions with confidence. Appl Bioinform 2002;1:93–100.

[11] Tillier ERM, Biro L, Li G, Tillo D. Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. Proteins 2006;63:822–31.

[12] Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 2002;47:219–27.

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

[14] Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. Proc Natl Acad Sci 2016;113:12186–91.

[15] Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS. Inferring interaction partners from protein sequences. Proc Natl Acad Sci 2016;113:12180–5.

[16] Várnai C, Burkoff NS, Wild DL. Improving protein-protein interaction prediction using evolutionary information from low-quality MSAs. PLoS ONE 2017;12:e0169356.

[17] Bitbol A-F. Inferring interaction partners from protein sequences using mutual information. PLOS Comput Biol 2018;14:e1006401.

[18] Correa Marrero M, ImminkRGH, de Ridder D, van Dijk ADJ. Improved inference of intermolecular contacts through protein–protein interaction prediction using coevolutionary analysis. Bioinformatics https:/doi.org/10.1093/bioinformatics/bty924 (May 15, 2019).

[19] Morcos F et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci 2011;108:E1293–301.

[20] Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28:184–90.

[21] Jeong C-S, Kim D. Reliable and robust detection of coevolving protein residues. Protein Eng Des Sel 2012;25:705–13.

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

[23] Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLOS Comput Biol 2010;6:e1000633.

[24] Juan D, Pazos F, Valencia A. Co-evolution and co-adaptation in protein networks. FEBS Lett 2008;582:1225–30.

[25] Codoñer FM, Fares MA. Why should we care about molecular coevolution? Evol Bioinform 4, 117693430800400000 (2008).

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

[27] Nadaradjane AA, Guerois R, Andreani J. Protein-protein docking using evolutionary information. Methods Mol Biol Clifton NJ 2018;1764:429–47.

[28] MacKay DJC, Information theory, inference and learning algorithms, 1st ed., Cambridge University Press; 2003.

[29] Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008;24:333–40.

[30] Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein–protein docking. Nucleic Acids Res 2006;34:W310–4.

Appendix 2: Paper published at the Scientific Reports

scientific reports

OPEN

Trivial and nontrivial error sources account for misidentification of protein partners in mutual information approaches

Camila Pontes[1,2], Miguel Andrade[1,2], José Fiorote[1] & Werner Treptow[1]✉

The problem of finding the correct set of partners for a given pair of interacting protein families based on multi-sequence alignments (MSAs) has received great attention over the years. Recently, the native contacts of two interacting proteins were shown to store the strongest mutual information (MI) signal to discriminate MSA concatenations with the largest fraction of correct pairings. Although that signal might be of practical relevance in the search for an effective heuristic to solve the problem, the number of MSA concatenations with near-native MI is large, imposing severe limitations. Here, a Genetic Algorithm that explores possible MSA concatenations according to a MI maximization criteria is shown to find degenerate solutions with two error sources, arising from mismatches among (i) similar and (ii) non-similar sequences. If mistakes made among similar sequences are disregarded, type-(i) solutions are found to resolve correct pairings at best true positive (TP) rates of 70%—far above the very same estimates in type-(ii) solutions. A machine learning classification algorithm helps to show further that differences between optimized solutions based on TP rates are not artificial and may have biological meaning associated with the three-dimensional distribution of the MI signal. Type-(i) solutions may therefore correspond to reliable results for predictive purposes, found here to be more likely obtained via MI maximization across protein systems having a minimum critical number of amino acid contacts on their interaction surfaces (N > 200).

Coevolution of proteins A and B translates itself into a series of homologous primary-sequence variants encoding coordinated compensatory mutations and, therefore, a specific set of protein–protein interactions between members of family A and members of family B. The problem of resolving specific protein partners based on multi-sequence alignments (MSAs) has received great attention over the years[1,2]. Ingenious approaches based on the correlation of phylogenetic trees[3–5] and profiles[6], gene colocalization[7] and fusions[8], maximum coevolutionary interdependencies[9] and correlated mutations[10,11], maximization of the interfamily coevolutionary signal[12], iterative paralog matching based on sequence energies[13] and expectation–maximization[14] have been developed and applied to resolve interaction partners in single or multiple (paralogous) gene copies in the same genome. Despite these advances, the problem of protein partners prediction remains unsolved for large sequence ensembles in general, especially for the case of protein coevolution across independent genomes—examples are phage proteins and bacterial receptors, pathogen and host-cell proteins, neurotoxins and ion channels, to mention a few. The problem lacks any suitable solution especially because an effective heuristic to search for the correct set of protein partners across the space of M! potential matches still misses in case of large number of sequences M (Fig. 1).

In a previous investigation, we showed that the coevolutive information encoded on the interacting amino acids of proteins A and B can be useful to discriminate the correct set of protein partners based on MSAs, in contrast to other evolutive and stochastic sources spread over their sequences[15]. When compared to other sources, the coevolutive information is the strongest signal to distinguish protein partners derived from coevolution within the same genome and, likely, the unique indication available in the case of protein interactions in independent genomes. We showed that physically-coupled amino acids at the molecular interface of A and B store the largest per-contact mutual information ($\hat{I}_{AB}$) to discriminate MSA concatenations with the largest expectation fraction of correct interaction partners—a result that was found to hold for various definitions of intermolecular contacts and binding modes. Although that information content might be of practical relevance in the search of an effective heuristic to resolve specific protein partners, the degeneracy ω, i.e., the number of

[1]Laboratório de Biologia Teórica e Computacional (LBTC), Universidade de Brasília DF, Brasília, Brazil. [2]These authors contributed equally: Camila Pontes and Miguel Andrade. ✉email: treptow@unb.br

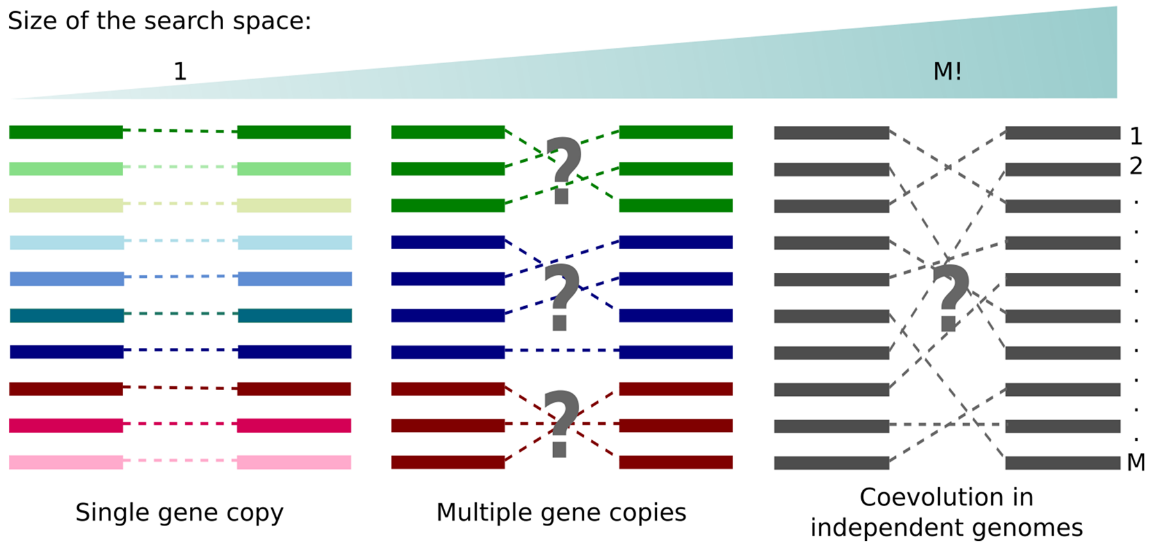


Figure 1. Different scenarios for protein partners determination from multi-sequence alignments. The correct set of partners is known for systems with a single gene copy per genome and unknown for systems involving multiple (paralogous) sequences within the same genome or multiple sequences across independent genomes. This figure was created with Inkscape (https://inkscape.org/).

MSA concatenations with a similar amount of $\widehat{I}_{AB}$ to the native concatenation is expected to be large ($\omega \gg M$), imposing severe limitations to that purpose.

Here, we investigate that hypothesis accordingly for a variety of protein families, including obligate and non-obligate complexes. It is worth emphasizing that the aim of this work is not to provide a method for the prediction of protein–protein interactions nor protein–protein interfaces, hence it differs from the studies in which sequence covariance is used to predict three-dimensional amino acid contacts or to infer specific interactions for a set of paralogs. Instead, we want to qualitatively explore the MI degeneracy in the space of possible protein partners associations between two interacting protein families. To approach that, we analyze a set of converged trajectories produced by a Genetic Algorithm (GA) that maximizes $\widehat{I}_{AB}$ starting from scrambled MSA concatenations of protein families with known partners in the same genome. Consistent with the expected degeneracy of $\widehat{I}_{AB}$, GA optimizations show two subspaces of MSA concatenation solutions: subspace (i), which consists of optimized solutions with a trivial error source arising from mismatches among similar sequences; and subspace (ii), which consists of optimized solutions with a non-trivial error source due to mismatches among non-similar sequences. By disregarding mistakes made among similar sequences, protein partners are resolved at best true-positive (TP) rates of ~ 70% in type-(i) optimizations – far above best TP rates in type-(ii). Type-(i) and -(ii) solutions are found to be functionally distinct from each other, with the former presenting a larger near-native content of mutual information correctly distributed among amino acid contacts. Particularly important, that finding supports the notion that differences between optimized solutions based on TP rates have a biological meaning associated with the amount of functional information and its spatial distribution. Type-(i) solutions may therefore correspond to reliable results for predictive purposes[1], more likely obtained via $\widehat{I}_{AB}$ maximization across protein systems found here to have a minimum critical number of amino acid contacts on their interaction surfaces (N > 200).

Results and discussion

In search of an effective heuristic to resolve specific protein partners based on MSAs with large numbers of sequences, the degeneracy of the per-contact mutual information $\hat{I}_{AB}$ was investigated here across 26 independent protein families with known interaction partners in the same genome (see "Methods" and Table S1). To approach that, we have performed optimization trajectories produced by a Genetic Algorithm (GA, see "Methods" and Algorithm S1) that starts from a random concatenation of MSA A and MSA B, and maximizes $\hat{I}_{AB}$ by performing small changes in the MSA concatenation iteratively (Fig. 2A). Accordingly, Fig. 2B shows 156 optimization trajectories with convergence obtained after 45,000 generations as indicated by their average time derivative $\delta\hat{I}_{AB} \leq 0.001$ in Fig. 2C. The average trajectory converges at ~ 98% of the $\hat{I}_{AB}$ reference value in the native concatenation z^{*}.

Despite presenting near-native values of $\hat{I}_{AB}$, optimized solutions fail at pairing sequences correctly in consequence of the degeneracy of the space of possible MSA models constrained by the $\hat{I}_{AB}$ maximization criteria. As made clear in Fig. 3A, there are three groups of solutions: one group of scrambled concatenations with 0% TP rate and low values of $\hat{I}_{AB}$ (in gray), one group of optimized concatenations with 0% TP rate and near-native $\hat{I}_{AB}$ (in red), and one group of native concatenations with 100% TP rate and native $\hat{I}_{AB}$ (in green). Careful inspection of the data reveals that the presence of similar sequences in MSA B contributes to that high error rate by yielding similar optimized values of $\hat{I}_{AB}$ when paired with a given sequence in MSA A. Indeed, reassessment of TP rates by disregarding mistakes made among sequences at the 20th percentile of Hamming distances distribution (see "Methods"—Fig. 9) allows regrouping of solutions into a subspace (i) with TP rates larger than 30% (Fig. 3B).

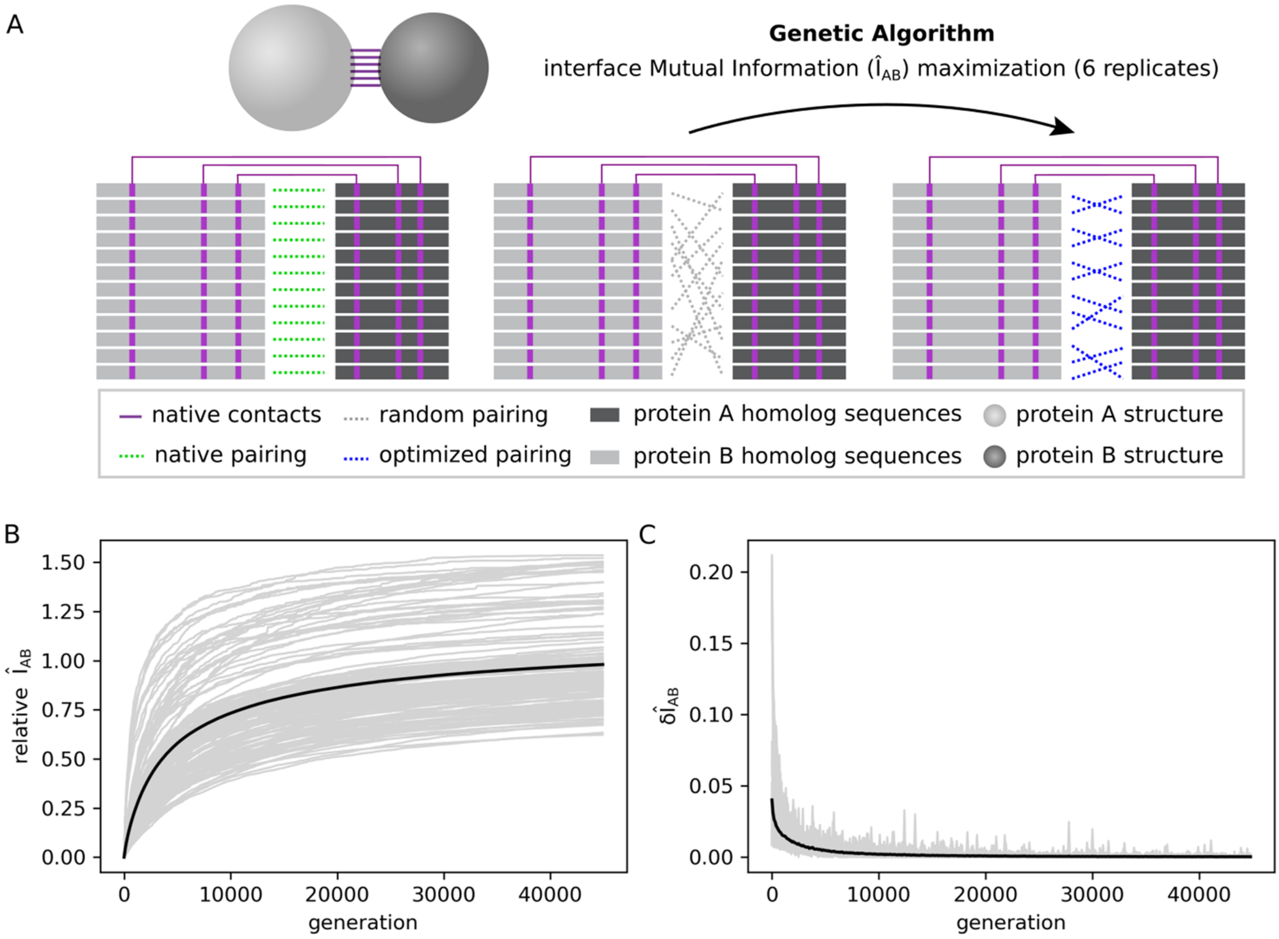


Figure 2. Interface mutual information ($\hat{I}_{AB}$) optimization trajectories. (**A**) Scheme showing $\hat{I}_{AB}$ optimization process starting from a scrambled multi-sequence alignment (MSA) concatenation (in gray) and reaching an optimized concatenation (in blue). Only physically coupled MSA position pairs (shown in purple) are taken into account. (**B**) Optimization trajectories for 26 protein systems. For each system, there are six trajectories with different starting points. The $\hat{I}_{AB}$ normalized by the native interface mutual information (relative $\hat{I}_{AB}$) is plotted against the number of generations of the genetic algorithm (gray lines). The average trajectory over all complexes is shown in black. (**C**) First-order derivative of the optimization trajectories shown in (**B**). The derivatives of individual trajectories are shown in gray, while the average derivative over all trajectories is shown in black. This figure was generated with Inkscape (https://inkscape.org/) and matplotlib v3.1.2 (https://matplotlib.org/).

As a measure of correlation, it is not surprising that mutual information is degenerate given that trivial source of error. Unexpected however is the fact that degeneracy may also involve another subspace of optimized solutions (ii) related to the non-trivial mismatch of sequences at larger Hamming distances. Supporting that notion, protein partners prediction at better TP rates (> 30%) demands a larger fraction of sequence mismatches (above the 20th percentile) to be discounted in optimized solutions (ii). As shown in Supporting Information, conclusions about subspaces (i) and (ii) hold for mismatches definitions using other Hamming distance cutoffs (Figure S1).

To get further insights on the mismatch problem reported in Fig. 3, the functional distinction of solutions type-(i) and (ii) was then analyzed according to the three-dimensional distribution of evolutive and coevolutive sources of the mutual information signal. Implicit in the analysis is the assumption that type-(i) solutions must necessarily have a near-native content of mutual information correctly distributed among amino acid contacts i.e., a near-native information content with a high correlation $r(\hat{I}(X_i;\ Y_i),\ \hat{I}^T_{nat}(X_i;\ Y_i))$ between the optimized solution vector $\hat{I}(X_i;\ Y_i)$ and its native conjugate $\hat{I}^T_{nat}(X_i;\ Y_i)$. Consistent with that assumption, Fig. 4 shows that the k-nearest neighbor (KNN) machine learning algorithm[16] discriminates type-(i) and -(ii) solutions with high accuracy ~ 82%, according to their nativelikeness across the space $\hat{I}_{AB} \times r$. A further decomposition analysis reveals the information recovered from type-(i) solutions has larger contents of the evolutive (phylogenetic) and coevolutive signals encoded on the native interacting amino acids of proteins A and B[15]—as also indicated by the high accuracy ~ 82% in which such solutions are effectively classified by the KNN algorithm applied on the correlation space redefined in terms of the specific signals. Here, what is meant by coevolutive signal, as explained in[15], is the surplus of MI stored in residue pairs at the interface (on average) when compared to the MI stored in residue pairs in general (on average), which is the evolutive, or phylogenetic, signal. For all cases, differentiation

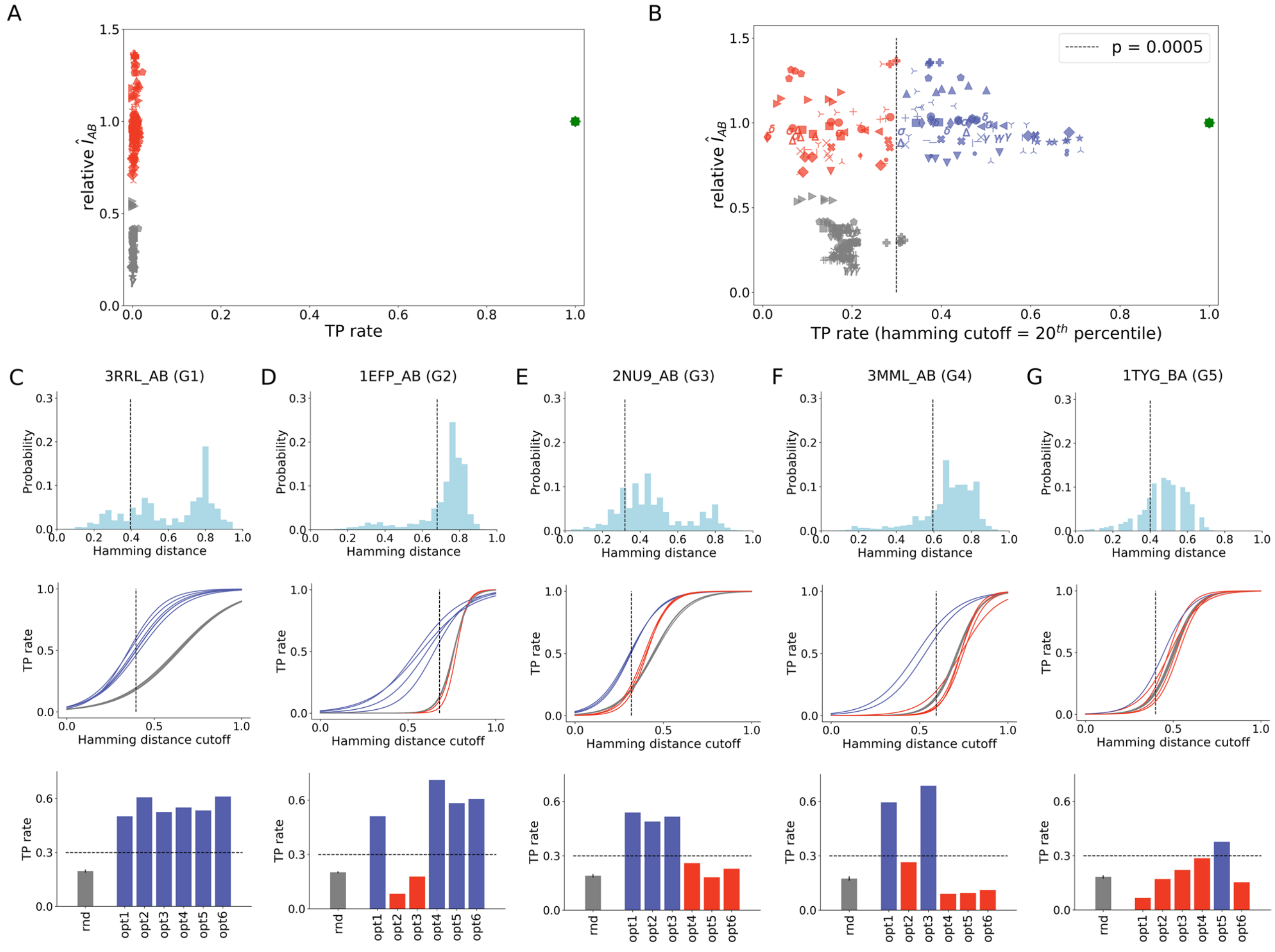


Figure 3. Evaluation of optimized MSA concatenations. (**A**) True positive (TP) rate of random, optimized and native MSA concatenations. (**B**) Reassessed TP rate of random, optimized and native MSA concatenations by discounting wrong pairings among sequences with Hamming distance within the 20th percentile of the distance distribution. Optimized solutions with TP rate greater than 30% (p = 0.0005) are shown in blue, while optimized solutions with TP rate lower than 30% are shown in red. Random solutions are shown in gray. (**C**–**G**) Hamming distance distribution of MSA B, TP rates versus Hamming distance discounts (the 20th percentile is shown with a dashed line), and TP rates of random (rnd) and optimized (opt1–6) solutions for the 20th percentile Hamming distance cutoff shown for representative systems: 3RRL_AB (**C**), 1EFP_AB (**D**), 2NU9_AB (**E**), 3MML_AB (**F**), and 1TYG_BA (**G**). This figure was generated using matplotlib v3.1.2 (https://matplotlib.org/).

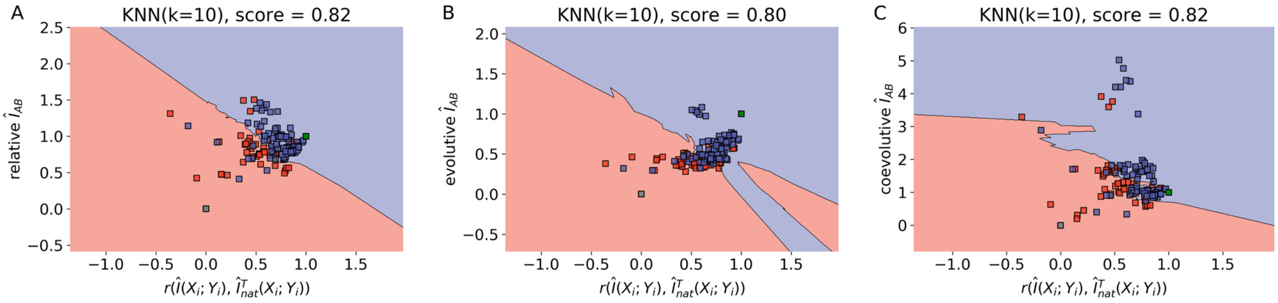


Figure 4. (**A**) Optimized concatenation solutions scattered across the space of relative interface mutual information (MI), $\hat{I}_{AB}$, against Pearson correlation between optimized and native MI vectors, $r(\hat{I}(X_i;\ Y_i),\ \hat{I}^T_{nat}(X_i;\ Y_i))$. Type-(i) solutions are shown in red and type-(ii) solutions are shown in blue. The bidimensional space was separated by a k-nearest neighbors (KNN) classification algorithm[16] (default Python 3 scikit-learn implementation, k = 10, for other k values see Figure S2). Native and scrambled concatenations were plotted afterwards in the same space and are shown in green and gray, respectively. Analogous plots were generated for the evolutive (**B**) and coevolutive (**C**) components of $\hat{I}_{AB}$. The decomposition was performed according to[15]. This figure was generated using sci-kit learn v0.22.2 (https://scikit-learn.org) and mlxtend v0.18.0 (http://rasbt.github.io/mlxtend/).

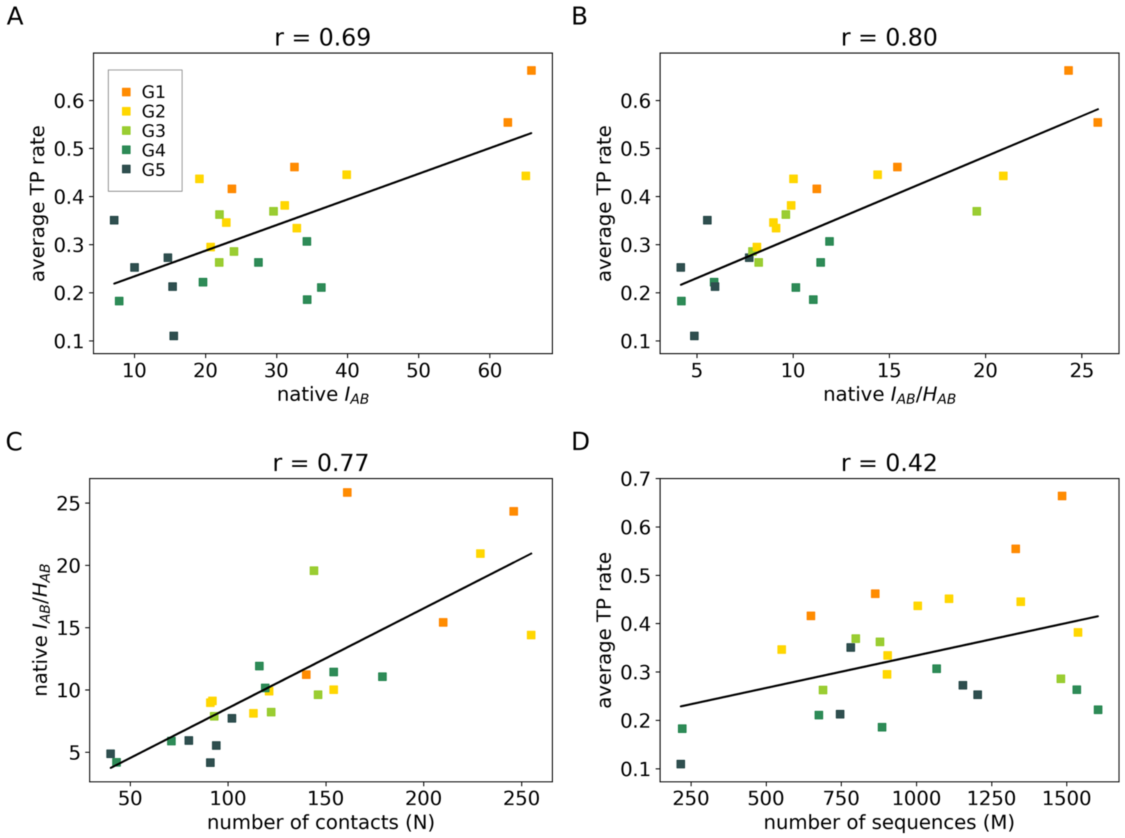


Figure 5. (**A**) Correlation between the true positive (TP) rate of optimized solutions and mutual information (MI) on the interface I_{AB}. (**B**) Correlation between TP rate of optimized solutions and I_{AB} regularized by the joint entropy on the interface, I_{AB}/H_{AB}. (**C**) Correlation between native I_{AB}/H_{AB} and the number of contacts on the interface (N). (**D**) Correlation between TP rate and number of sequences in the alignment (M). Values on the x-axis in A-B were calculated considering the native pairing. TP rates are shown as averages (n = 6) for each system. Systems were colored based on groups G1–5: group 1 is composed by systems with only type-(i) solutions (Fig. 3C and Fig. S3), group 2 by systems with a majority of type-(i) solutions (Fig. 3D and Fig. S4), group 3 by systems with the same proportions of type-(i) and type-(ii) solutions (Fig. 3E and Fig. S5), group 4 by systems with a majority of type-(ii) solutions (Fig. 3F and Fig. S6), and group 5 by systems in which optimized concatenations did not differentiate from the scrambled ones (Fig. 3G and Fig. S7). This figure was generated using matplotlib v3.1.2 (https://matplotlib.org/).

is far above the non-significant value of 50% thus supporting the conclusion that differences between optimized solutions based on TP rates may have a biological meaning associated with the amount of functional information recovered and its spatial distribution.

Given the importance that native-like solutions may have in predictive purposes, the propensity of protein systems to produce such optimized solutions was further analyzed according to the content of non-trivial errors. As shown in Fig. 5A,B, protein systems were found to cluster into five distinct groups with average TP rates that strongly correlate with the amount of mutual information at the interaction surface of proteins, with or without regularization by the local joint entropy H_{AB} (see "Methods"). According to that analysis, lower contents of mutual information appear to account for the higher propensity of the system in producing type-(ii) solutions. Because the mutual information content is proportional to the number of amino acid contacts at the protein surface, N (Fig. 5C), this result appears to be consistent with the statistical expectation that the distribution of MI values is broader over systems with fewer degrees of freedom (contacts). More importantly, it indicates N as an important parameter to discriminate suitable protein systems for which maximization of $\hat{I}_{AB}$ may likely produce near-native type-(i) solutions with biological meaning as reported in Fig. 4. The relevance of that parameter becomes clear by noting that the number of MSA sequences (M) does not explain well the content of non-trivial errors across protein clusters (Fig. 5D), despite the well-documented fact that M may significantly impact the accuracy of coevolutionary approaches[17]. The condition N > 200 thus emerges here as one plausible threshold criteria for the classification of protein systems that are suitable for maximization of $\hat{I}_{AB}$ and resolution of protein partners via type-(i) solutions.

So far, our results were obtained from a set of protein families involving unique sequence pairs per genome that may not have coevolved under strong selective pressures towards specificity. To better understand any implicit dependence of the results with that experimental condition, error sources (i) and (ii) were then further

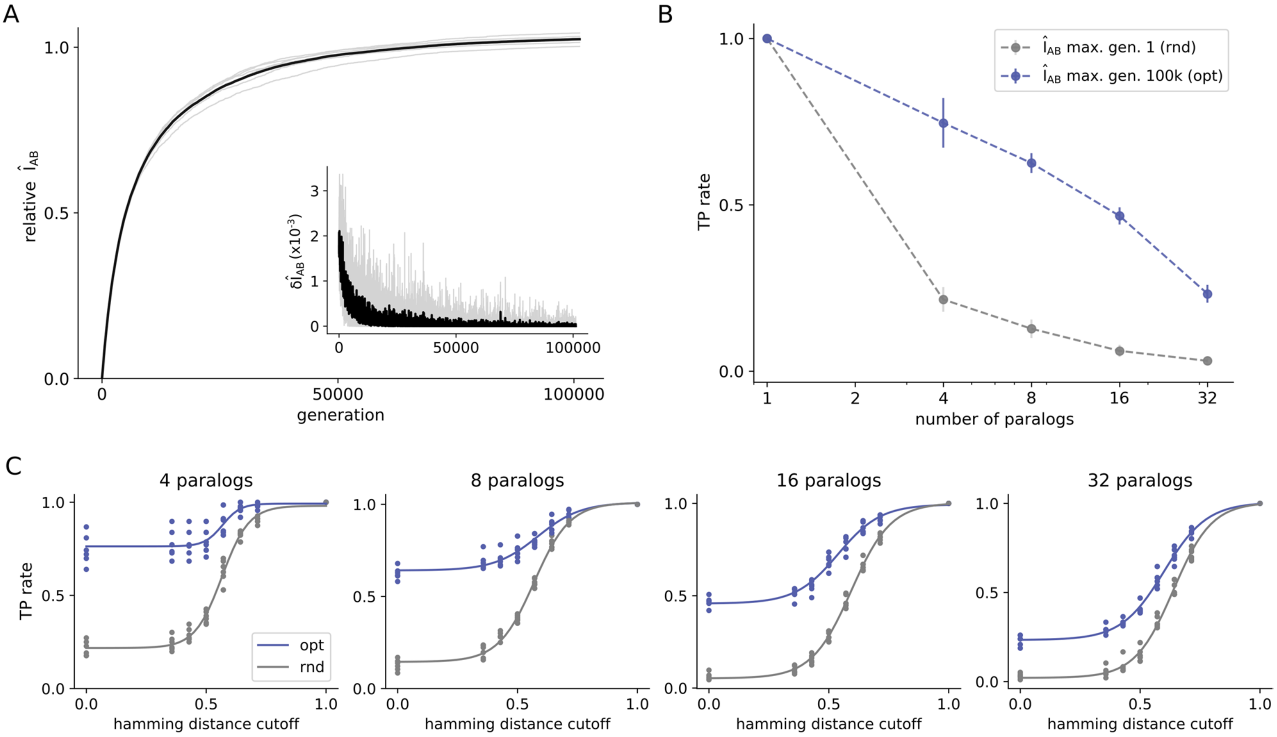


Figure 6. Evaluation of optimized MSA concatenations of the HK-RR paralogs dataset. (**A**) Optimization trajectories for the HK-RR standard dataset. The interface mutual information normalized by the native interface mutual information (relative $\hat{I}_{AB}$) is plotted against the number of generations for optimizations (with 6 replicates each) starting from a solution with a scrambled concatenation within each species. The first derivative of the trajectory is shown in the smaller plot. (**B**) True positive (TP) rate of start (in gray) and final (in blue) solutions after ~ 100,000 rounds of $\hat{I}_{AB}$ maximization. The TP rate is shown in average for bacterial species containing different numbers of paralogs. (**C**) TP rate after disregarding mismatches among sequences considering different Hamming distance cutoffs for bacterial genomes with different numbers of paralogs in the standard HK-RR dataset. The TP rate is shown for both random (rnd) and optimized (opt) MSA concatenations. This figure was generated using matplotlib v3.1.2 (https://matplotlib.org/).

investigated in the context of the bacterial two-component system HK-RR featuring highly specific protein–protein interactions across multiple protein copies per genome. More specifically, histidine kinase (HK) and their respective response regulator (RR) are paralogous gene families[13,18,19], each consisting of multiple sequences sharing significant homology at the primary and tertiary levels. Despite that signature, HK-RR pairs are highly specific within the same genome in consequence of evolutive pressures avoiding crosstalk between independent two-component pathways[20]—as shown by Rowland and Deeds, the evolution of new HK-RR pairs follows rapid sequence divergence immediately after duplication events[21].

Accordingly, Fig. 6 presents another series of $\hat{I}_{AB}$ optimizations performed on the HK-RR dataset containing around 5000 sequences, coming from ~ 450 bacterial genomes from the P2CS database[22–24]. Optimizations were performed with 6 replicates each, starting from a paired alignment with a randomized pairing within each species. All species were optimized together, which means that each optimization step benefits from the cumulative changes that happened in previous steps (see "Methods"—Fig. 8). As shown in Fig. 6A, optimization to near-native values of $\hat{I}_{AB}$ is attained after ~ 100,000 generations, with $\delta\hat{I}_{AB} < 0.001$.

When analyzing the TP rate for species with different numbers of paralogs, optimized MSA solutions present an improvement over the initial concatenations (Fig. 6B). In this case, TP rates are not null because the degeneracy of (M ≤ 32) paired sequences of paralogs is expected to be significantly smaller than that of (M > 200) paired sequences in Fig. 3. It is interesting to notice that TP rates obtained here by optimizing only the interface MI are only slightly inferior to the same estimates obtained considering full protein MI found in the literature[18], especially for genomes with a higher number of paralogs. Figure 6C shows further the TP rate of optimized and random MSA concatenations, considering a 20th percentile Hamming distance discount cutoff, for bacterial genomes with different numbers of paralogs. It is possible to observe that random and optimized curves approximate with increasing numbers of paralogs. Extrapolating for cases with more than 32 paralogs, the two curves tend to overlap similarly to what occurs in protein systems in which optimized concatenations did not differentiate from the scrambled ones (Fig. 3G and Fig. S7) and therefore, suggesting that type (i) errors do not contribute to $\hat{I}_{AB}$ degeneracy in HK-RR system. We hypothesize that the lack of type-(i) error originated from mismatches among similar sequences is due to the high specificity of this system.

Results in Fig. 6 appear to rationalize the sharp deterioration of TP rates with the number of sequences in recent investigations of paralogous systems[12–14,18,19], by hypothesizing it is due to the lack of type-(i) mismatches and the great degeneracy involved. In previous works, Bitbol and coworkers developed an iterative pairing algorithm (IPA) capable of inferring protein partners using either direct coupling analysis (DCA-IPA)[13], mutual information (MI-IPA)[18], or phylogeny (Mirrortree-IPA)[19]. When benchmarked for paralog matching on the standard HK-RR dataset, DCA-IPA was as accurate as MI-IPA, and Mirrortree-IPA was even more accurate. The performance of these algorithms, however, drops considerably for species with more than 32 paralogs. The tendency is that the TP rate also drops to zero in a hypothetical genome with hundreds of paralogs[19], a situation analogous to the results in Fig. 6. In conclusion, results presented in Fig. 6 suggest that paralog matching is only possible because there is usually a small number of paralogous sequences per genome. When extended to genomes with more paralogs, this problem tends to present only type-(ii) solutions, leaving virtually no room for improvement of TP rates.

Conclusions and future work

Here, we investigate the hypothesis that the coevolutive information encoded on the interacting amino acids of proteins A and B ($\hat{I}_{AB}$) can be useful to discriminate protein partners based on large multi-sequence alignments (MSAs). When compared to evolutive and stochastic sources, $\hat{I}_{AB}$ was previously found as the strongest signal to distinguish protein partners derived from coevolution within the same genome and likely the unique indication in the case of independent genomes[15]. In contrast to other coevolutionary signals that may also be considered in purpose[9,10,12–14], $\hat{I}_{AB}$ thus corresponds to a small and still important fraction of the total information available in protein sequences making it especially suitable for specific partners inference via fast algorithmic routines. Despite these aspects, the degeneracy of $\hat{I}_{AB}$ is expected to be large and may impose severe limitations to practical applications.

Indeed, $\hat{I}_{AB}$ optimization across the space of possible MSA concatenations is shown here to resolve specific protein partners at very low true positive (TP) rates in consequence of error sources (i) and (ii). As a measure of correlation, it is not surprising that $\hat{I}_{AB}$ is degenerate given trivial mismatches (i) among similar sequences. Unexpected however is the fact that degeneracy may also involve another subspace of optimized solutions (ii) with the non-trivial mismatch of sequences at larger Hamming distances. If trivial error sources are disregarded, further analysis indicates, however, that protein partners may be resolved in the context of type-(i) solutions at best TP rates of ~ 70%—far above the same estimates in type-(ii) solutions.

Type-(i) and -(ii) solutions are found to be functionally distinct from each other, with the former presenting a larger near-native content of mutual information correctly distributed among amino acid contacts. Particularly important, that finding supports the notion that their differentiation based on TP rates is not just a theoretical construct but instead has a biological meaning associated with how much functional information is recovered and how accurately distributed this information is. Type-(i) solutions may therefore correspond to reliable results for predictive purposes[1], more likely obtained via $\hat{I}_{AB}$ maximization across protein systems with a minimum critical number of amino acid contacts on their interaction surfaces (N > 200).

Finally, as a special case of a highly specific system of paralogs, HK-RR interactions are resolved here at very low TP rates following $\hat{I}_{AB}$ maximization, which is consistent with TP rates reported in the literature[19] employing other more complex optimization algorithms, such as DCA-IPA[13]. As shown in Fig. 6, the HK-RR system was found not to present type-(i) degeneracy and, as such, its TP rates sharply deteriorate with M≥32 sequences per genome and cannot be improved by any means. Exclusive existence of type-(ii) errors in the HK-RR system thus suggests another layer of complexity that sequence diversity and specificity may add to the problem. Investigation of these aspects as key determinants for error sources (i) and (ii) is therefore another important perspective of the presented work. In this direction, we speculate that HK-RR pairs within the same genome are highly specific and this is the reason why there is no type (i) error in this system. In contrast, systems with only one pair of interacting proteins per genome do not suffer selective pressure to avoid cross-binding homologs occurring in other species and, therefore, present both type (i) and type (ii) errors.

Overall, the investigations performed in this work provide some clarifications into the general problem of protein coevolution from the perspective of sequence diversity. It is difficult to say to which point homologous sequences were selected to selectively bind to their native partners since there is a huge degeneracy in the space of possible sets of partners. Despite the intrinsic complexity of the problem of specific protein partners prediction for large sequence ensembles, the novel theoretical insights presented here might provide relevant information for future studies and should contribute to advancing our knowledge in the field.

Methods

Consider two interacting protein families, A and B. It is possible to construct two MSAs, MSA A and MSA B, containing M sequences from families A and B, respectively. A specific coevolution process $z \in \{1, \ldots, M!\}$ associates each sequence l in MSA B to a sequence k in MSA A in a unique arrangement of size M (see Fig. 7). Given that members of A and B interact via formation of N independent amino acid contacts at molecular level, it is possible to extract from these MSAs only the columns corresponding to sites that are in contact, belonging to the complex interface. In this context, the interacting amino acids of families A and B are 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 associated probability mass functions (PMFs) $\{\rho(x_1 \ldots x_N),\ \rho(y_1 \ldots y_N),\ \rho(x_1 \ldots x_N,\ y_1 \ldots y_N|z)|x_i,\ y_i \in \Omega,\ \ \forall i \in \{1, \ldots, N\}\}$. Here, the alphabet Ω has size 21 and contains all 20 amino acids and the gap symbol '–'. Note that only the joint PMF will depend on process z.

Here, we approximate each site-specific PMF $\{\rho(x_i),\ \rho(y_i),\ \rho(x_i,\ y_i|z)|i \in \{1, \ldots, N\}\}$ by the empirical amino acid frequencies $\{f(x_i),\ f(y_i),\ f(x_i,\ y_i|z)|i \in \{1, \ldots, N\}\}$ obtained from the concatenated MSAs. Note that each coevolution process z determines a specific concatenation, as illustrated in Fig. 7. It means that, essentially, the search will be guided by the amount of information X^N stored about Y^N conditional to different coevolution processes z.

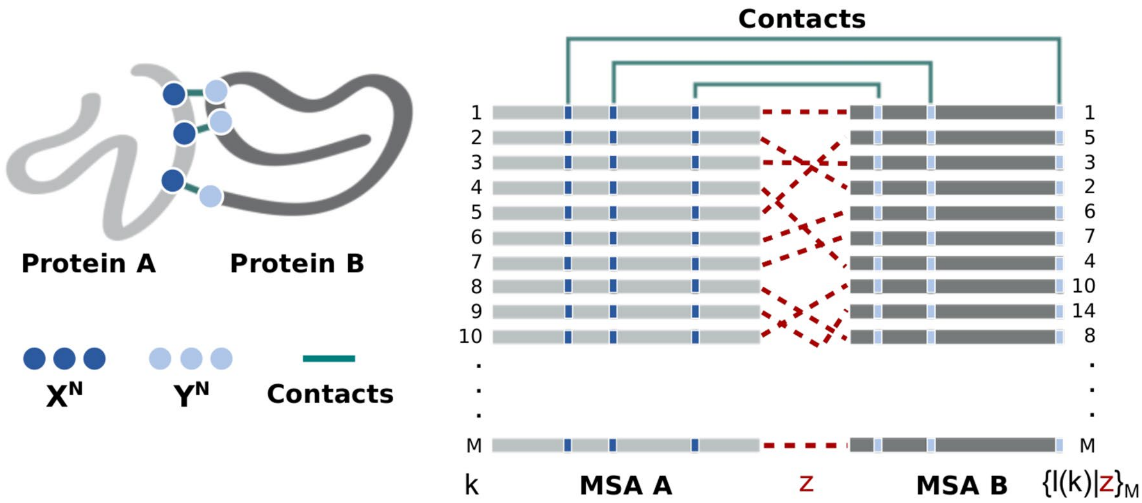


Figure 7. Structural contacts mapped into M-long multi-sequence alignment of protein interologs A and B. A set of pairwise protein–protein interactions is defined by associating each sequence l in MSA B to a sequence k in MSA A in one unique arrangement, $\{l(k)|z\}$, determined by the coevolution process z to which these protein families were subjected. This figure was created with Inkscape (https://inkscape.org/).

Shannon mutual information. The Shannon mutual information contained on the interface of interacting proteins A and B conditional to a given coevolution process z is calculated as follows

$$\begin{aligned}\hat{I}_{AB} &= \frac{1}{N} I(X^N;\ Y^N|z) = \frac{1}{N}\sum_{i=1}^{N} I(X_i;\ Y_i|z) \\ &= \frac{1}{N}\sum_{\Omega x \Omega} f(x_i,\ y_i|z)\ln\left(\frac{f(x_i,\ y_i|z)}{f(x_i)f(y_i)}\right), \quad x_i,\ y_i \in \Omega\end{aligned} \tag{1}$$

where N is the number of contacts at the AB complex interface, $f(x_i)$ is the empirical frequency of x_i as a realization of X_i, $f(y_i)$ is the empirical frequency of y_i as a realization of Y_i, and $f(x_i,\ y_i|z)$ is the empirical frequency of pair $(x_i,\ y_i)$ as a realization for the i-th contact given a specific coevolution process z.

The empirical values of single and joint frequencies were corrected considering a pseudocount, as follows

$$f_i(x_i) \leftarrow (1-\lambda)f_i(x_i) + \frac{\lambda}{Q}$$

$$f_{ij}(x_i,\ x_j|z) \leftarrow (1-\lambda)f_{ij}(x_i,\ x_j|z) + \frac{\lambda}{Q^2}$$

where, Q is the size of alphabet Ω and λ is the pseudocount parameter. In this work, we adopt a small pseudocount of $\lambda = 0.001$.

The joint entropy of the interface was calculated for individual contacts

$$H(X_i,\ Y_i|z) = f(x_i,\ y_i|z)\ln(f(x_i,\ y_i|z))$$

where $f(x_i,\ y_i|z)$ is the empirical frequency of pair $(x_i,\ y_i)$ as a realization for the i-th contact given a specific coevolution process z. Afterwards, the regularization I_{AB}/H_{AB} was obtained according to

$$I_{AB}/H_{AB} = \sum_{i=1}^{N} I(X_i;\ Y_i|z)/H(X_i,\ Y_i|z)$$

where N is the number of contacts.

Systems under investigation. Protein complexes under investigation are shown in Table S1. MSAs A and B for all protein families were obtained from Ovchinnikov and coworkers[25]. Amino acid contacts defining the discrete stochastic variables X^N and Y^N were identified from the x-ray crystal structure of the bound state of a representative protein pair from families A and B using a typical contact definition considering maximum separation distance of 8 Å between amino acids carbon beta. The full dataset of protein systems validated in[25] was considered here, except for systems 2Y69_BC, 2ONK_AC, 3A0R_AB, 3RPF_AC, and 4HR7_AB, which were considered outliers in terms of M/N values 469.3, 87.7, 192.3, 150.6, and 45.3 significantly larger than their typical estimates described in Table S1.

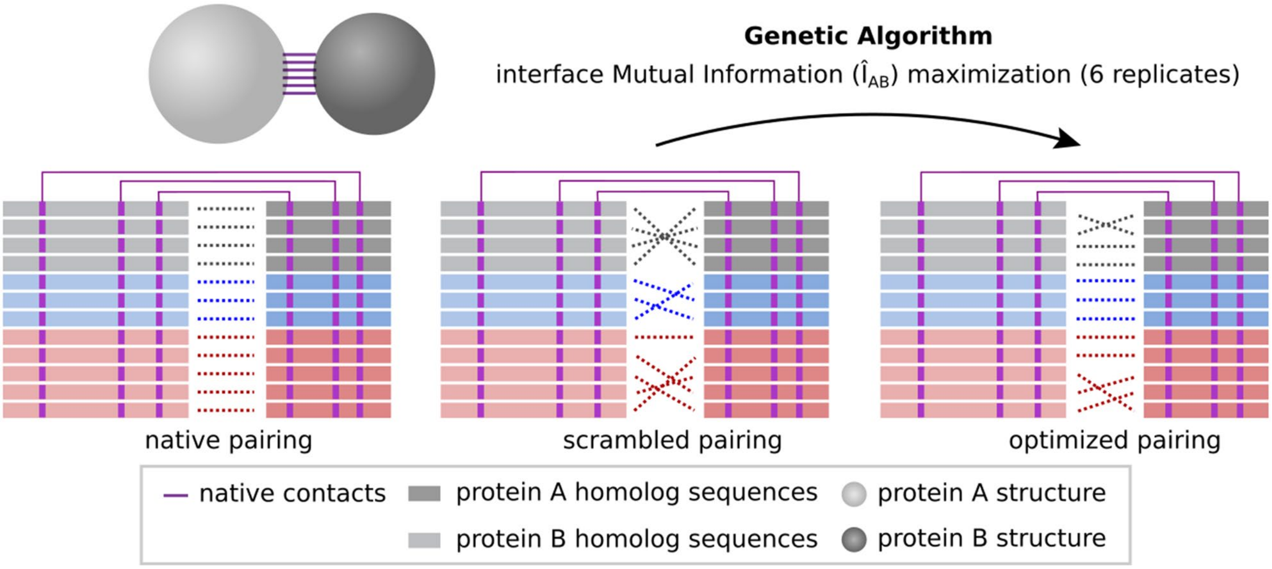

Genetic Algorithm
interface Mutual Information ($\hat{I}_{AB}$) maximization (6 replicates)
native pairing
scrambled pairing
optimized pairing
native contacts
protein A homolog sequences
protein B homolog sequences
protein A structure
protein B structure

Figure 8. Scheme showing interface mutual information ($\hat{I}_{AB}$) optimization process for the HK-RR standard dataset. It starts from a within-species scrambled MSA concatenation and reaches an optimized concatenation. Different species are shown in different colors. Only physically coupled MSA position pairs (shown in purple) are taken into account and only within-species changes are made in each generation. This figure was created with Inkscape (https://inkscape.org/).

Additionally, the HK-RR standard dataset containing around 5000 sequences, coming from around 450 bacterial genomes from the P2CS database[22–24] was included. This paired MSA was produced and validated by Bitbol and coworkers[13] in paralog matching experiments. The PDB complex 5UHT (chains A and B) was selected as a representative for this system. The reason for including this system containing paralogous proteins is to have a baseline for comparison with previous related studies.

Genetic algorithm. The mutual information contained on the interface of the protein complexes, calculated as described in Eq. (1), was maximized using a Genetic Algorithm (GA, Algorithm S1). For each of the protein complexes considered, six independent optimization trajectories were obtained, starting from different randomly generated populations. Each optimization was performed with a population of eight individuals with unique genomes encoding a specific concatenation z of MSAs A and B. In each generation, the elite (top-50% individuals with the best fitness) reproduces and replaces the remaining 50% individuals with lower fitness with new individuals with genomes that are mutated copies of the elite. A mutation in the genome of an individual consists of swapping positions of two sequences on MSA B, and thereby slightly changing the concatenation z. The fitness of the individuals is calculated in each generation and corresponds to the total interface mutual information obtained considering an individual unique genome, i.e., a specific concatenation of MSAs A and B. The optimization was stopped after a predefined number of 50,000 generations was reached.

A slightly different optimization procedure was implemented for the special case of the HK-RR standard dataset (Fig. 8). In this case, the initial population is composed of within-species scrambled solutions and, in each generation, only within-species changes are allowed. More specifically, each time a new mutated individual is generated, one of the species that compose the MSA is randomly selected, and a change in the concatenation within this species is performed. The optimization was stopped after a predefined number of 100,000 generations was reached.

The optimal set of parameters for the GA were derived from a series of tests performed on six representative systems. In each test, one of these parameters varied, assuming a range of values while all other parameters remained fixed (Table S2). All tests were performed with a predefined seed for the random number generator, which means that the starting point and the sequence of mutations performed are constant for all trajectories of the same system. This was done to ensure that any effects observed in the final results were due solely to variations in the GA parameters.

Figure S8 shows how parameter values correlated with relative $\hat{I}_{AB}$ at the end of test trajectories. Given that both the number of individuals and the elite proportion correlated positively with relative $\hat{I}_{AB}$ (Figure S8A,B), the values selected for these parameters were the maximum tested, i.e., 8 and 0.5, respectively. The number of mutations, on the other hand, correlated negatively with relative $\hat{I}_{AB}$ (Figure S8C), thus the value selected for this parameter was 1. Results for parameter λ were not so conclusive (Figure S8D) and, since this parameter was set to 0.001 in previous work[15], its value was maintained the same. As shown in Figure S9, GA parameters do not influence TP rates observed at the end of trajectories thus supporting that our conclusions are robust over GA parameters, with the possible exception of λ, which will be investigated in future work.

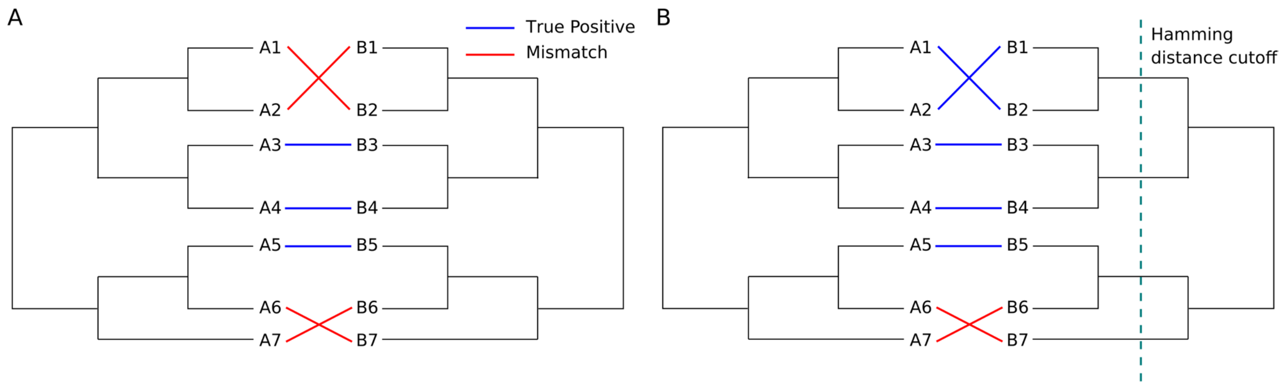


Figure 9. Mismatch discounting based on a Hamming distance cutoff. Scheme showing how the accuracy of the same MSA concatenation would be assessed with (**B**) and without (**A**) mismatch discounting. This figure was created with Inkscape (https://inkscape.org/).

Assessment of optimized solutions accuracy. The true positive (TP) rates of optimized concatenations obtained at the end of the genetic algorithm (GA) $\hat{I}_{AB}$ maximization trajectories were calculated in two different manners: with and without mismatch discounting. TP rate assessment without mismatch discounting consists simply of counting how many sequence partners were correctly paired in the target solution and divided by the total number of sequences (Fig. 9A). TP rate assessment with mismatch discounting, on the other hand, consists of counting how many sequences were paired either with their correct partner or with a partner that is close enough to the correct one in terms of Hamming distance (Fig. 9B). Hence, mismatch discounting depends on a predefined Hamming distance cutoff, below which sequences are considered similar enough for the mistakes to be forgiven. Here, we consider the 20th percentile of a given protein family B distance distribution as the predefined cutoff for mismatch discounting. Figure S1 shows that the relaxation of that parameter does not affect qualitatively the results.

A K-Nearest Neighbors (KNN) classifier was used to investigate if MSA pairing solutions with trivial and non-trivial error sources scattered differently in the space of relative $\hat{I}_{AB}$ against correlation of individual MI values with the native solution, $r(\hat{I}(X_i;\ \ Y_i),\ \hat{I}^T_{nat}(X_i;\ \ Y_i))$. All type-(i) and type-(ii) solutions obtained were used to train a KNN classifier with default scikit-learn (https://scikit-learn.org) parameters, except for the number of neighbors (K). Values of K were tested ranging from 2 to 20, but little variation in the accuracy score was observed, with scores ranging from 0.76 to 0.87. Therefore a value of K = 10 was chosen as a compromise between a possible overfit when considering too few neighbors and losing accuracy when considering too many neighbors (results for other values of K are shown in Figure S2). The accuracy score was calculated using the scikit-learn function *.score*() on the model inferred by the KNN classifier. This function indicates how well the model fits the provided data points, i.e., it calculates the accuracy on the training set.

Received: 14 October 2020; Accepted: 15 March 2021
Published online: 25 March 2021

References

1. Morcos, F. & Onuchic, J. N. The role of coevolutionary signatures in protein interaction dynamics, complex inference, molecular recognition, and mutational landscapes. *Curr. Opin. Struct. Biol.* **56**, 179–186 (2019).
2. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14**, 249–261 (2013).
3. Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293 (2000).
4. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng. Design Select.* **14**, 609–614. https://doi.org/10.1093/protein/14.9.609 (2001).
5. Gertz, J. *et al.* Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics* **19**, 2039–2045 (2003).
6. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4285–4288 (1999).
7. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
8. Marcotte, C. J. V. & Marcotte, E. M. Predicting functional linkages from gene fusions with confidence. *Appl. Bioinform.* **1**, 93–100 (2002).
9. Tillier, E. R. M., Biro, L., Li, G. & Tillo, D. Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. *Proteins* **63**, 822–831 (2006).
10. Pazos, F. & Valencia, A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins Struct. Funct. Genet.* **47**, 219–227. https://doi.org/10.1002/prot.10074 (2002).
11. Burger, L. & van Nimwegen, E. Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.* https://doi.org/10.1038/msb4100203 (2008).
12. Gueudré, T., Baldassi, C., Zamparo, M., Weigt, M. & Pagnani, A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12186–12191 (2016).
13. Bitbol, A.-F., Dwyer, R. S., Colwell, L. J. & Wingreen, N. S. Inferring interaction partners from protein sequences. *Proc. Natl. Acad. Sci.* https://doi.org/10.1101/050732 (2016).

14. Marrero, M. C., Immink, R. G. H., de Ridder, D. & van Dijk, A. D. J. Improved inference of intermolecular contacts through protein–protein interaction prediction using coevolutionary analysis. *Bioinformatics* **35**, 2036–2042. https://doi.org/10.1093/bioinformatics/bty924 (2019).
15. Andrade, M., Pontes, C. & Treptow, W. Coevolutive, evolutive and stochastic information in protein-protein interactions. *Comput. Struct. Biotechnol. J.* **17**, 1429–1435. https://doi.org/10.1016/j.csbj.2019.10.005 (2019).
16. Dasarathy BV. Nearest Neighbor (NN) Norms: Nn Pattern Classification Techniques (1991).
17. Mao, W., Kaya, C., Dutta, A., Horovitz, A. & Bahar, I. Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution. *Bioinformatics* **31**, 1929–1937 (2015).
18. Bitbol, A.-F. Inferring interaction partners from protein sequences using mutual information. *PLoS Comput. Biol.* **14**, e1006401 (2018).
19. Marmier, G., Weigt, M. & Bitbol, A.-F. Phylogenetic correlations can suffice to infer protein partners from sequences. *PLoS Comput. Biol.* **15**, e1007179 (2019).
20. Laub, M. T. & Goulian, M. Specificity in two-component signal transduction pathways. *Annu. Rev. Genet.* **41**, 121–145. https://doi.org/10.1146/annurev.genet.41.042007.170548 (2007).
21. Rowland, M. A. & Deeds, E. J. Crosstalk and the evolution of specificity in two-component signaling. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5550–5555 (2014).
22. Barakat, M. *et al.* P2CS: A two-component system resource for prokaryotic signal transduction research. *BMC Genomics* **10**, 315 (2009).
23. Barakat, M., Ortet, P. & Whitworth, D. E. P2CS: A database of prokaryotic two-component systems. *Nucleic Acids Res.* **39**, D771–D776 (2011).
24. Ortet, P., Whitworth, D. E., Santaella, C., Achouak, W. & Barakat, M. P2CS: Updates of the prokaryotic two-component systems database. *Nucleic Acids Res.* **43**, D536–D541 (2015).
25. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).

Acknowledgements

We would like to thank Caio Souza for his work in the early stages of this project and Antônio Francisco Pereira de Araújo for useful discussions. This work was supported by National Council for Scientific and Technological Development CNPq [Grant number 302089/2019-5 (WT)], Coordenação de Aperfeiçoamento de Pessoal de Nível Superior CAPES [Grant number 23038.010052/2013-95 (WT)], and Fundação de Apoio à Pesquisa do Distrito Federal FAPDF [Grant number 193.001.202/2016 (WT)].

Author contributions

C.P., M.A. and W.T. designed research; C.P., M.A. and J.F. performed research; C.P., M.A., J.F. and W.T. analyzed data; C.P. and W.T. wrote the original and the reviewed manuscript; C.P. and M.A. contributed equally to this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-86455-0.

Correspondence and requests for materials should be addressed to W.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.