Universidade de Brasília – UnB

Instituto de Ciências Biológicas – IB

Programa de Pós Graduação em Biologia Molecular

Teoria da Informação Aplicada

ao estudo de interações proteína-proteína

Miguel de Souza Andrade

Orientador: Prof. Dr. Werner Treptow

Brasília, 2020

Universidade de Brasília – UnB

Instituto de Ciências Biológicas – IB

Programa de Pós Graduação em Biologia Molecular

Teoria da Informação Aplicada ao estudo de interações proteína-proteína

Miguel de Souza Andrade

Orientador: Prof. Dr. Werner Treptow

Tese apresentada ao Programa de Pós-Graduação em Ciências Biológicas – Biologia Molecular, do Departamento de Biologia Celular, do Instituto de Ciências Biológicas da Universidade de Brasília como parte dos requisitos para obtenção do título de Doutor em Biologia Molecular.

MIGUEL DE SOUZA ANDRADE

Teoria da Informação Aplicada ao estudo de interações proteína-proteína

Tese apresentada ao Programa de Pós-Graduação em Ciências Biológicas – Biologia Molecular, do Departamento de Biologia Celular, do Instituto de Ciências Biológicas da Universidade de Brasília como parte dos requisitos para obtenção do título de Doutor em Biologia Molecular.

Banca Examinadora:

_____

Prof. Dr. Werner Treptow (Orientador) (CEL – UnB)

_____

Prof. Dr. Antônio Francisco Pereira de Araujo (CEL – UnB)

_____

Prof. Dr. Fernando Lucas de Melo (CEL – UnB)

_____

Prof. Dr. Leandro Martínez (UNICAMP)

“Nothing in biology makes sense except in the light of evolution”.

Theodosius Dobzhansky

“Ainda que eu andasse pelo vale da sombra da morte, não temeria mal algum, porque tu estás comigo; a tua vara e o teu cajado me consolam”.

Salmos 23:4

À minha filha, Clara

À minha esposa, Mara

Aos meus pais, João e Erenilsa

Às minhas irmãs, Eliene, Elaine e Camila

Agradecimentos

Agradeço primeiramente a Deus, por me permitir contemplar e estudar a sua obra.

Agradeço aos meus pais, Erenilsa e João, por me ensinarem o que é certo e errado. Por me educarem. Por me colocarem na escola. Por me cobrar ser correto. Por me cobrar ser estudioso. Por me incentivarem a ser estudioso. Por serem os melhores pais do Mundo. Eu sou o que sou devido a vocês e serei eternamente grato!

Agradeço às minhas irmãs Eliene, Elaine e Camila, pelas risadas e companheirismo. Vocês são as melhores irmãs que alguém pode ter!

Eu tive a sorte de fazer bons amigos ao longo da minha vida. Agradeço a todos os meus amigos e em especial aos amigos da biologia Rafael Correa e Raquel, por me aturarem desde o primeiro semestre e por estarem sempre ao meu lado. Agradeço em especial também aos meus amigos de escola (e da vida) Rafael Linhares e Wanderson Gonçalves pela amizade e pelo “passe de batalha”. Agradeço também aos amigos de mestrado Daniel e Roberto.

Agradeço ao meu orientador, Werner Treptow pela orientação, inspiração, por me apresentar problemas desafiadores, pela liberdade, pela confiança e por ter me introduzido no mercado de ações.

Agradeço aos colegas de laboratório, Caio, Camila, Leonardo, Natália, José, Letícia, Alessandra, Mônica, Vinícios, Pedro e Diego pelo melhor ambiente de trabalho possível! Caio e Camila tiveram uma contribuição importante para esse trabalho, um obrigado especial a vocês.

Agradeço a todos os professores que contribuíram para a minha caminhada até aqui. Em especial ao Edson, Heitor, Euterlúcia, Bergmann e Fernando.

Agradeço aos membros da banca, Leandro Martínez, Fernando Lucas, Antônio Francisco, por aceitarem ler e avaliar o meu trabalho.

Agradeço à Universidade de Brasília, por me abrigar nesses 10 anos, por ser tão diversa, por ser inspiradora, por ter uma biblioteca fantástica. Obrigado! Agradeço também ao Restaurante Universitário, sem ele eu não estaria aqui.

Agradeço à agência de fomento CAPES, que pagou minha bolsa e à FAPDF por eventuais apoios em congressos.

Agradeço ao Brasil, meu amado país. Agradeço à educação pública, gratuita e de qualidade que tive o privilégio de ter acesso.

E por último, mas não menos importante, agradeço à minha família, minha esposa Mara e minha filha Clara, que me dão força para continuar, que são minha inspiração, que fazem os meus dias mais felizes! As palavras não são suficientes e como dizia um poeta, descrever a felicidade é diminuí-la. Eu amo vocês!

Prefácio

Esta é a minha Tese de Doutorado apresentada ao Programa de Pós Graduação em Biologia Molecular como parte dos requisitos para obtenção do título de doutor. Optei por dividi-la em três Capítulos: *(1)* Coevolução e interação proteína-proteína, uma revisão; *(2)* Informação coevolutiva, filogenética e estocástica em interações proteína-proteína; *(3)* Perspectivas no estudo de interações proteína-proteína. No primeiro capítulo apresento uma revisão bibliográfica, cuja versão em inglês será publicada como um artigo de revisão. No segundo apresento os principais resultados obtidos durante o doutorado e que compõe um artigo já publicado (ver Anexo I). No terceiro mostro alguns dos resultados mais recentes e promissores, que compõe um outro artigo em fase de preparação do manuscrito. Optei por escrever um documento compacto e de maneira simples, para que o leitor compreenda sem dificuldade parte do que foi realizado durante o meu doutorado. Uma parte do trabalho desenvolvido durante o meu doutorado (como centenas de linhas de código em *python*, avanços teóricos relacionados ao assunto, diversas colaborações com colegas de laboratório) não estão presentes neste documento mas serão publicadas em breve na forma de artigos científicos e códigos abertos no *git hub*.

Resumo

Interações proteína-proteína (PPI) são processos-chave para a manutenção da vida. As proteínas sofrem diversas pressões seletivas para serem funcionais o que inclui pressões para a manutenção das interações. Nesse contexto, observamos a coevolução de pares de proteínas, que é quando a evolução de uma delas afeta a evolução da outra e vice versa. A Coevolução pode ser quantificada por medidas de correlação entre as substituições de aminoácidos, como por exemplo a *transinformação* (*I*). Essas medidas podem ser utilizadas para resolver diferentes problemas biológicos no contexto de PPI, a saber: *(i)* predição de contatos tridimensionais, *(ii)* predição de interfaces de interação, *(iii)* predição de redes de interação e *(iv)* proposição de pareamento de sequências homólogas de proteínas que interagem. É sabido que diversas fontes contribuem para os valores de *I*, entretanto entender como cada uma dessas fontes contribui para a resolução desses problemas permanece um desafio. Nesse sentido, esta Tese de Doutorado tem como objetivo entender como a evolução, a coevolução e a estocasticidade influenciam nos valores *I* e mais do que isso, compreender qual a usabilidade dessas diferentes fontes na resolução do problema *(iv)*. Utilizando alinhamentos de sequências e estruturas disponíveis em bancos de dados, eu mostro aqui que a quantidade de *I* resultado da coevolução pode ser a mais interessante para a resolução do problema *(iv)*. Isso porque a essa quantidade é a que possui menor degeneração, isto é, nos permite distinguir melhor entre arranjos de pareamento que fazem sentido biológico de arranjos aleatórios e deve ser a única fonte em sistemas que coevoluem em organismos diferentes. Além disso essa fonte de informação coevolutiva pode ser captada apenas em interfaces de interação corretas e de curta distância. Por fim, apresento perspectivas em otimizações para resolução de problemas do tipo (*IV)*.

Abstract

Protein-protein interactions (PPI) are a key process for maintenance of life. Proteins are under diverse selective pressures to maintain their function and interactions. I this context, we observe the coevolution of protein pairs, which is when the evolution of one affect the evolution of other and vice-versa. The coevolution can be quantified by correlation metrics between the amino acids substitution, like mutual information (*I*). These metrics can be used to solve different biological problems like: *(i)* structural contacts prediction, *(ii)* interaction interface prediction, *(iii)* network interaction prediction and *(iv)* interactinng proteins homologous pairing. It has been known that different sources contributes to *I* values, however, to understand how each one of these sources contributes for these problems remains a challenge. In this way, this PhD thesis has as goal understand how the evolution, the coevolution and the stochastic influences *I* values and more than that, understand the usability of this different sources to solve problem *(iv)*. Using multiple sequence alignments and structures available in data bases, I show here that the *I* resulted of coevolution can be the best choice to solve the problem *(iv)*. Because, these quantity has less degeneration than others, that is, *I* resulted from coevolution allow us to distinguish better biological pairing from random pairing and must be the unique source in systems who evolved in different organisms. Furthermore, this source of coevolutive information could be found just in correct interfaces defined by short distances. finaly, I show perspectives in maximization to solve the problem *(iv)*.

Índice

Lista de Figuras

.

Lista de Tabelas

Lista de Anexos

Lista de Abreviações

AA	Amino ácido
DCA	*Direct Coupling Analisys*
DI	*Direct Information*
GREMLIN	*Generative REgularized ModeLs of proteINs*
HIV	Vírus da Imunodeficiência Humana
MI	*Mutual Information*
MSA	Alinhamento Múltiplo de Sequências
PDB	*Protein Data Bank*
PPI	Interação proteína-proteína
PSICOV	*Protein Sparse Inverse COVariance*
TMV	*Tobaco Mosaic Virus*

Capítulo 1. Coevolução e interação proteína-proteína, uma revisão

O estudo de biologia molecular vem avançando a cada ano e junto a descobertas importantes está o desenvolvimento e aprimoramento de técnicas de sequenciamento de biomoléculas. Nos últimos anos o número de sequências disponíveis nos bancos de dados cresceu (e continua crescendo) exponencialmente. Há quem diga, inclusive, que já estamos vivendo a “era pós genômica”.

Agora, um desafio muito maior surge: o que fazer com esse enorme número de sequências biológicas? De maneira bem consolida, essa abundante quantidade de informação pode ser usada para reconstrução da história filogenética dos mais diversos organismos. Outra forma de utilizar esses dados vem se desenvolvendo nos últimos anos e lançam mão de técnicas advindas da Teoria da Informação para estudar interações entre proteínas. Utilizando apenas alinhamentos múltiplos de sequências (MSA, sigla em inglês de *multiple sequence alignments*) é possível, por exemplo, predizer com uma certa acurácia, quais aminoácidos estão em contato tridimensional na estrutura de uma proteína. O presente capítulo trata-se de uma revisão sobre o uso da Teoria da Informação (e algumas derivações), para o estudo de interações entre biomoléculas.

1. Evolução de proteínas

Em 1859, Charles Darwin propôs a teoria da evolução das espécies mesmo sem conhecer a base genética que permitia que a evolução acontecesse. Hoje, sabemos que a evolução biológica é resultado da mudança na frequência alélica (William Klug, 2010). Mutações e outros mecanismos evolutivos geraram toda a variabilidade de espécies e sequências que observamos hoje.

Embora as mutações possam ocorrer de forma aleatória ao longo de todo o genoma, nós não observamos na natureza todas as mutações que acontecem, isso porque parte delas causa substituição do aminoácido codificado, podendo, em muitos casos, ser

deletéria, o que impede que o organismo sobreviva. Desse modo, as substituições aminoacidicas que observamos na natureza são resultados de diferentes pressões seletivas que variam dependendo da localização no genoma ou região codificadora (Netto and Menck, 2012). Além disso, algumas proteínas estão mais suscetíveis a substituições (mutações que causam troca de aminoácido) do que outras e, ainda, dentro de uma mesma proteína existem regiões que são mais variáveis e outras mais conservadas (Echave et al., 2016; Kimura, 1968; Kimura and Ohta, 1973).

A explicação por trás da variação nas taxas de substituição está, como dito acima, nas diferentes pressões evolutivas. Por exemplo, suponha que o aminoácido da posição 27 de uma proteína participa de uma interação com o aminoácido da posição 56 da mesma proteína e essa interação é fundamental para a manutenção da estrutura dessa proteína. Muito provavelmente, mutações que alterem qualquer um desses aminoácidos causará um dobramento errado dessa proteína causando perda de função, dessa maneira essas mutações serão deletérias e não serão observadas na natureza. Suponha um aminoácido localizado em um sítio catalítico de uma enzima, mutações nesse aminoácido podem afetar a capacidade catalítica dessa proteína causando um mal funcionamento, temos aqui outro exemplo de um sítio que tem alta restrição para mutações.

A **Figura 1.1** exemplifica este aspecto da evolução de proteínas. De modo geral, sítios localizados no interior de proteínas (importantes para a manutenção da estrutura) e sítios catalíticos tendem a ter uma menor taxa evolutiva, isto é, mudam com pouca frequência. Sítios com alta variação costumam estar localizados na superfície proteica. As diferenças nas taxas evolutivas podem ser observadas nos alinhamentos de sequências homólogas da proteína em questão.

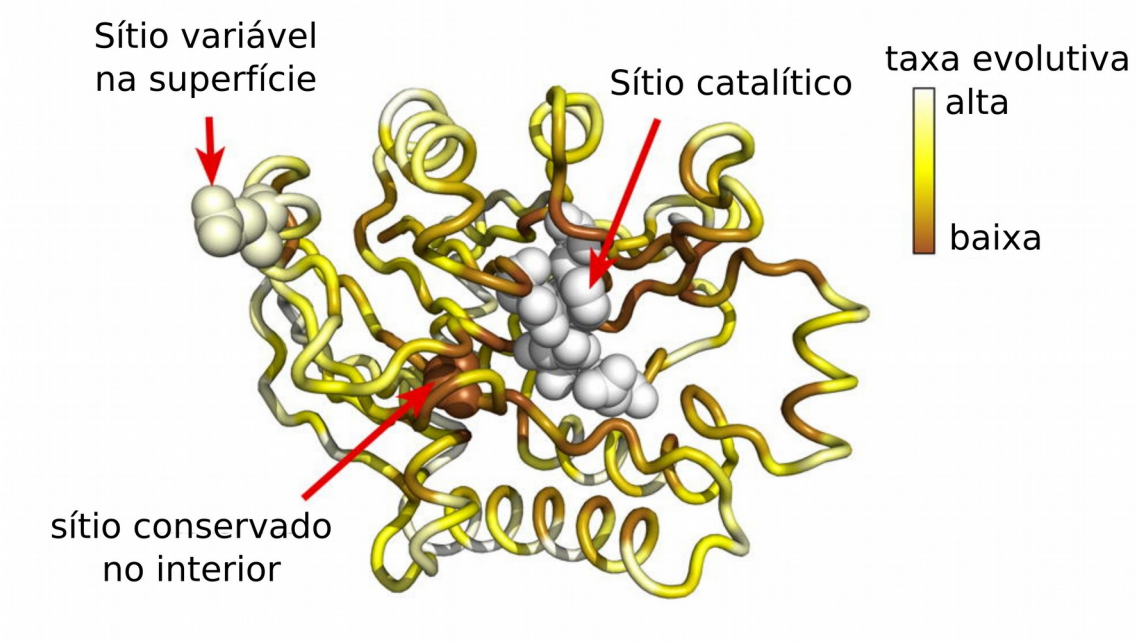


Figura 1.1. Diferentes Taxas evolutivas em uma proteína globular. *Representação esquemática da estrutura de Exonuclease III de Escherichia coli (código do PDB: 1AKO) colorida por taxa de evolução. As setas vermelhas destacam as diferentes restrições estruturais e funcionais sobre a taxa evolutiva. Figura traduzida de (Echave et al., 2016).*

Um fato interessante é que há vezes em que diferentes sítios (ou posições em um alinhamento) apresentam mutações correlacionadas. Essas mutações são chave para o entendimento de como proteínas se dobram e interagem, conforme veremos no tópico a seguir.

2. Da Coevolução entre espécies para a Coevolução Molecular

Darwin estudou a evolução do tamanho de corolas de flores de orquídeas e o tamanho de probóscides de polinizadores. Ele observou que mariposas que se alimentavam em orquídeas com corolas longas tinham probócides longas e mariposas que se alimentavam de orquídeas pequenas possuiam probócides pequenas. Tal fenômeno é um exemplo de dois caracteres biológicos, em organismos diferentes, que evoluem juntos, ou, coevoluem. Esse exemplo descrito por Darwin é tão importante que ilustra diversas capas de livros didáticos de evolução utilizados nos cursos de Biologia no mundo.

Diversos autores estudaram evolução conjunta de dois ou mais organismos (Dobzhansky, 1950; Wallace, 1953) mas foram Ehrlich e Raven que usaram o termo **coevolução** pela primeira vez (Ehrlich and Raven, 1964). A motivação apresentada no início do artigo de 1964 é simples e muito lógica: as espécies na natureza não estão sozinhas e interagem com diversas outras, de modo que a evolução de uma deve influenciar na evolução da outra. Os autores, então, descreveram diversos aspéctos da coevolução entre borboletas e plantas.

A coevolução é tão importante que pode moldar a macroevolução, por meio de um tipo de coevolução denominada rainha vermelha, em uma referência ao livro de Lewis Carroll, "Alice através do espelho", onde o autor escreve: "*Now here, you see, it takes all the running you can do, to keep in the same place*" (em tradução livre para o português: "Aqui, como você vê, é preciso correr o máximo possível, para permanecer no mesmo lugar"). Essa é uma analogia ao fato de que muitas espécies estão sempre mudando mas as interações permanecem (CARROLL, 1960). Por exemplo um hospedeiro muda para escapar de um parasita e o parasita muda para superar a resistência recém adquirida pelo hospedeiro. É Como uma corrida armamentística.

Todos esses aspectos de coevolução entre espécies foram, mais tarde, naturalmente agregados ao estudo de interações entre biomoléculas. Em outras palavras, proteínas também coevoluem e podemos, agora, utilizar um conceito mais moderno de coevolução, que pode ser explicada como *mudanças coordenadas que ocorrem em organismos ou* ***biomoléculas****, geralmente para manter ou refinar interações funcionais entre os pares* (revisto por de Juan et al., 2013).

3. Coevolução entre aminoácidos intra-proteína, uma perspectiva histórica

Neste tópico apresentaremos abordagens do que chamaremos de **problema 1**: **predizer quais aminoácidos estão próximos tridimensionalmente em uma proteína.**

Em 1987, Altschuh e colaboradores observaram que algumas posições do alinhamento de 7 sequências homólogas da proteína da capa proteica do *Tobacco Mosaic Virus* (TMV) apresentavam um padrão de substituição similar (Altschuh et al.,

1987). Eles observaram também, que essas posições correlacionadas estavam fisicamente próximas na estrutura da proteína. Essa foi a primeira identificação de posições que coevoluiam utilizando alinhamento de sequências. No ano seguinte, 1988, o mesmo grupo publicou um novo artigo expandindo as análises para outras famílias de proteínas (Altschuh et al., 1988).

Em 1993, (Korber et al., 1993) usou sequências de aminoácidos do *loop* V3 do vírus da imunodeficiência humana tipo 1 (HIV-1) e uma medida chamada *Mutual information* ou transinformação (*I*), para quantificar a dependência entre mutações nessa região da proteína. Este é o primeiro trabalho que utiliza essa medida derivada da teoria da informação (Cover and Thomas, 2006; Shannon, 1948).

A transinformação é uma medida simples de dependência entre duas variáveis estocásticas muito utilizada em diversas áreas, principalmente na comunicação. A *I* de duas variáveis estocásticas quaisquer (*X* e *Y*) é dada pela equação $I(X,Y)=\sum_{x,y}\rho(x,y)\log\frac{\rho(x,y)}{\rho(x)\rho(y)}$ onde $\rho(x,y)$ é a probabilidade conjunta e $\rho(x)$ e $\rho(y)$ são as probabilidades individuais. O valor de *I* é pequeno para variáveis independentes (por exemplo o resultado de cara ou coroa de duas moedas comuns, uma moeda não influencia a outra) e grande quando as variáveis não são independentes (por exemplo o resultado de cara ou coroa de duas moedas mágicas em que sempre que a primeira dá coroa a segunda dá cara).

Em 1994, foram publicados três artigos importantes no estudo de covariação de aminoácidos em proteínas. Os dois primeiros foram publicados em um mesmo volume da revista *protein engineering design & selection* em março 1994 (Shindyalov et al., 1994; Taylor and Hatrick, 1994) e o outro foi publicado em abril daquele ano na revista *proteins* (Göbel et al., 1994). Nota-se que em todos esses artigos o termo utilizado é “mutações correlacionadas” e a palavra “coevolução”, que já era aplicada para os estudos de organismos, ainda não aparece utilizada para biomoléculas.

Também em 1994, Neher apresentou uma estatística capaz de medir correlações em sequências de famílias de proteínas alinhadas atribuindo uma métrica

escalar (como carga ou volume da cadeia lateral) para cada aminoácido e calculando coeficientes de correlação dessas métricas para diferentes posições. Uma abordagem nova, quando comparada à que vinha sendo utilizada, que levava em conta apenas o aminoácido em si.

Shindyalov et al., 1994 mostraram por meio de um método baseado em análises estatísticas da distribuição de mutações nos ramos de árvores filogenéticas construídas com base no MSA, que existia uma tendência (significante, mas fraca, segundo os próprios autores) de pares de resíduos correlacionados estarem próximos tridimensionalmente. Os autores utilizaram um banco de MSAs criado por (Sander and Schneider, 1991).

Taylor e colaboradores citaram o trabalho de Altschuh et al., 1987 e argumentaram que o método proposto por eles poderia selecionar erroneamente pares de posições simplesmente por serem conservadas, eles então propuseram um método que, além de evitar esse problema, adicionou uma medida relacionada a características físico-químicas dos aminoácidos e, considerou muito mais proteínas com o objetivo de ganhar significância estatística. Eles mostram que o sinal de covariação, quando presente, é difícil de ser discriminado da clusterização resultado da conservação, principalmente em resíduos do núcleo e do sítio ativo da proteína.

A ideia apresentada em todos esses trabalhos listados acima é que se dois aminoácidos covariam, provavelmente esses aminoácidos estão próximos na estrutura tridimensional e o que Altschuh et al., 1987 e Shindyalov et al., 1994 propunham era a possibilidade de utilizar, no futuro, esse tipo de informação para melhorar a predição *ab initio*[1] de estruturas de proteínas. Göbel et al. (1994) consegue prever, com acurácia de 37 a 68%, o mapa de contatos de 11 famílias de proteínas, utilizando um método baseado em mutações correlacionadas em alinhamentos de sequências, de modo que esse é um dos primeiros artigos a propor um método automático para extrair os padrões de correlações de mutações de aminoácidos com o objetivo de predizer resíduos próximos estruturalmente. Vale ressaltar que essa acurácia ainda não era suficiente para predizer uma estrutura *de novo*, embora fosse melhor do que fazer isso randomicamente.

1 Predição de estrutura do zero, sem o uso de estruturas de sequências homólogas.

Cabe dizer também, que o número de sequências disponíveis ainda não era nada perto do que temos hoje. A **Figura 1.2** ilustra um mapa de contatos.

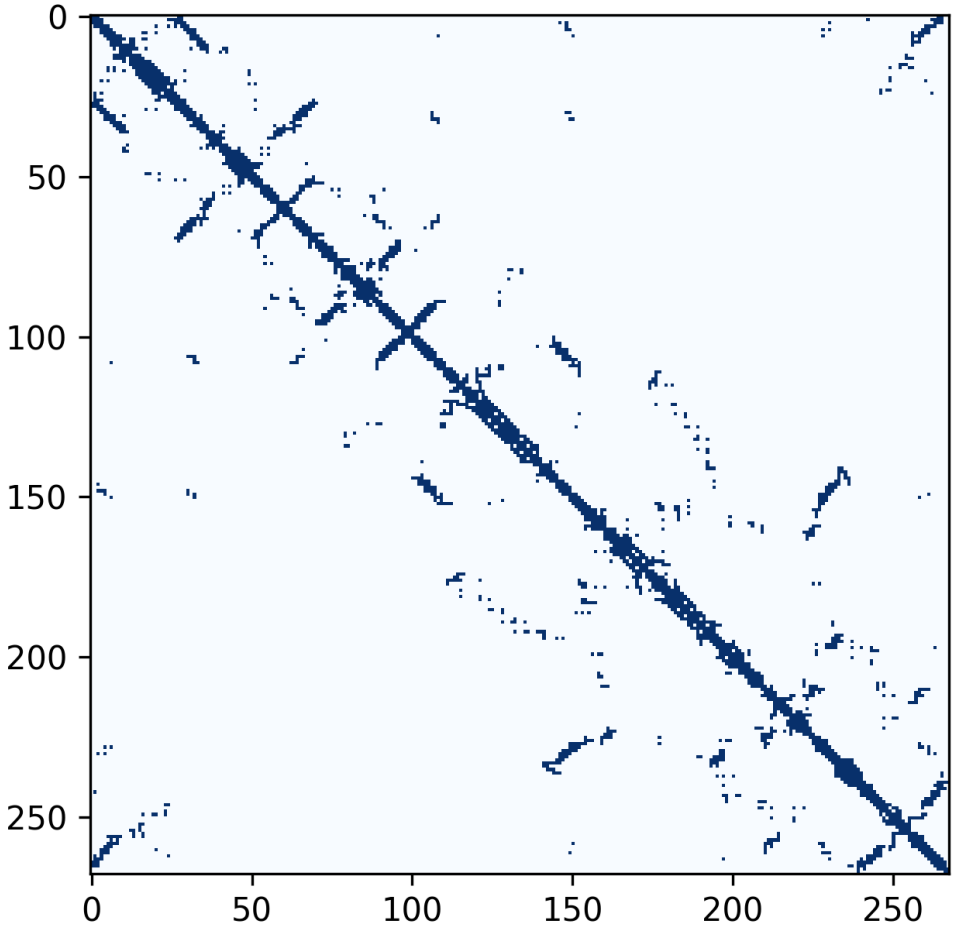


Figura 1.2 Mapa de contatos para a proteína Exonuclease III de Escherichia coli (1AKO). *Esse mapa corresponde a uma matriz de todos os aminoácidos contra todos, os pares que possuem distância menor ou igual a 8Å estão pintados de azul.*

Olmea and Valencia (1997) aperfeiçoaram o método proposto por Göbel et al. (1994). No Artigo de 1997 os autores propuseram a primeira abordagem sistemática que combina mutações correlacionadas com outras fontes de informação das sequências. Eles mostraram que os seguintes procedimentos aumentavam a acurácia das predições: *(i)* excluir os pares de resíduos que aparecem em <80% dos experimentos de *bootsrapping*; *(ii)* usar uma combinação linear dos valores de correlação com os valores de conservação (aminoácidos completamente conservados possuem correlação zero por

definição); *(iii)* filtrar a lista de contatos preditos com o critério de ocupancia (medida de quantos contatos um aminoácido pode fazer). Apesar das melhoras dos resultados ainda havia muito trabalho a ser feito e, como os próprios autores sugeriram, a combinação de diferentes métricas pode ser o caminho.

Depois de muitos anos com avanços pontuais no uso de teoria da informação para a predição de interações aminoácido-aminoácido (Atchley et al., 2000; Martin et al., 2005), incluindo aplicações para identificação de sítios importantes para modulação alostética (Süel et al., 2003) mas foi em 2009 que Weigt *et al.* propuseram a *DI (Direct Information,* ou informação direta, em português) motivados pelo fato da *I* não ser capaz de discriminar correlações diretas de indiretas, isto é, se um aminoácido *i* interage com *j* e com *k* e *k* e *j* não interagem (ver **Figura 1.3**), as mutações de K e j podem estar indiretamente correlacionadas, gerando um alto valor de MI indireto. Nesse exemplo, as correlações i-j e i-k são diretas, enquanto a correlação k-j é indireta. Embora essa métrica tenha nascido em um estudo de interação inter-proteína (ver próximo tópico) ela foi utilizada também para estudo de interações entre aminoácidos intra-proteína (Morcos et al., 2011). Nesse trabalho, os autores implementaram uma ferramenta computacional chamada de *Direct-Coupling Analysis* (DCA), que utiliza a DI para inferir os mapas de contato de aminoácidos de uma dada proteína. A DI é capaz de predizer contatos tridimencionais com melhor acurácia do que a MI.

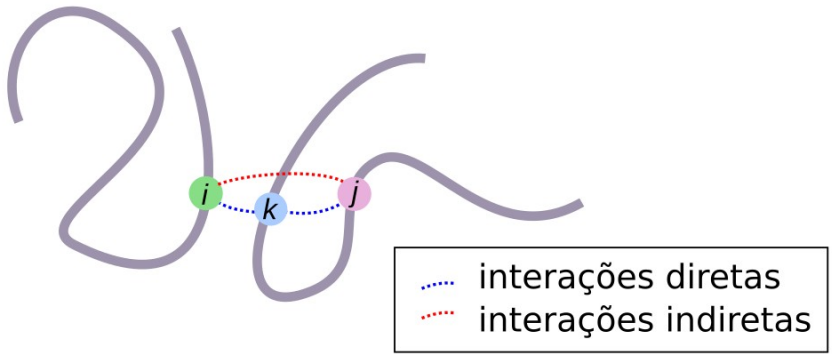


Figura 1.3 Esquema representativo de mutações indiretas. O aminoácido *k* interage com *j* e com *i*, entretanto *i* e *j* não interagem. Pode haver uma correlação indireta entre *i* e *j*.

O algorítimo proposto por Weigt et al. (2009) para calcular DI, chamado de *message-passing algorithm* ou mpDCA demandava muito poder computacional e era

lento, o que tornava difícil a sua aplicação para um número grande de famílias ou domínios. Morcos et al. (2011) usa outro método para o cálculo da DI, chamado *mean-field aproximation of DCA*, ou mfDCA, que é de 10^3 a 10^4 vezes mais rápido que a mpDCA. Mais tarde, em 2013 uma terceira abordagem de cálculo da DI foi proposta, usando agora *pseudolikelihoods* para calcular a DI e é chamada de PLMDCA (Ekeberg et al., 2013).

Antes da PLMDCA, Balakrishnan et al. (2011) propôs o método chamado GREMLIN (*Generative Regularized Models of proteINs*), que consiste em aprender um modelo probabilístico gráfico (um grafo onde os nós correspondem às colunas do alinhamento e as arestas especificam as independências condicionais entre as colunas) da composição de aminoácidos dentro do MSA. Esse método, que também utiliza *pseudolikelihoods*, é uma forma de mapear a informação a respeito do contexto estrutural de proteínas, que permite, por exemplo, identificar funções de uma dada sequência. Esse método foi utilizado mais tarde para predição de contatos estruturais (Kamisetty et al., 2013). Nesse artigo de 2013, os autores mostraram que o GREMLIN e a PLMDCA possuem acurácias similares, entretanto o GREMLIN é de 5 a 20 vezes mais rápido.

Diversos métodos recentes vem sendo desenvolvidos para a predição de mapas de contatos e possível utilização desses mapas para predição de estruturas 3D, como por exemplo as abordagens que utilizam *machine learning* e *neural networks (Shackelford and Karplus, 2007; Xie et al., 2020)*. Entretanto, essas abordagens não fazem parte do escopo desta tese.

4. Coevolução entre aminoácidos inter-proteína

Enquanto o problema apresentado no tópico 2 consistia em identificar aminoácidos próximos tridimensionalmente na estrutura de uma proteína, o problema apresentado neste tópico se preocupa em identificar aminoácidos próximos tridimensionalmente na interação *entre* proteínas. Dessa maneira, o **problema 2 é identificar os aminoácidos da interface de interação sem a necessidade de cocristalização.**

Em 1997 Pazos et al., 1997 estenderam o pensamento de mutações compensatórias intra-proteínas para o estudo de interações inter-proteína (e inter-domínios). O princípio aqui é similar ao das interações intra-proteína, considerando um par de proteínas que interagem, as mudanças acumuladas durante a evolução em uma devem ser compensadas por mudanças na outra. Esse trabalho foi pioneiro na utilização de mutações compensatórias para predição de interfaces de interação. Os autores mostraram que, em média, os pares de resíduos detectados como "correlacionados" estão mais próximos tridimensionalmente do que a média total dos pares de resíduos. Nesse mesmo trabalho, utilizando os valores de correlação entre os pares de aminoácidos das interfaces propostas, soluções corretas de *docking* foram discriminadas de soluções erradas. Para experimentos de *docking* proteína-proteína é necessário que se conheçam as estruturas tridimensionais das duas proteínas em questão. Seria possível identificar os aminoácidos participantes de uma interação proteína-proteina sem o uso de estruturas? Nesse mesmo trabalho, os autores identificaram os aminoácidos participantes da interação entre dois domínios da proteína HSC70 (*heat-shock protein*) utilizando somente o alinhamento de sequências, mostrando que era sim possível.

Como apresentado no tópico anterior, Weigt et al. (2009) propuseram a DI, que limpava o sinal indireto que aparecia na MI. Utilizando como modelo de estudo a interação de um par de proteínas envolvidas na transdução de sinal em bactérias [*sensor kinase* (HK) e *response regulator* (RR)]. Eles mostraram que pares de aminoácidos com baixo valor de MI também possuem baixo valor de DI, entretanto, aminoácidos com alto valor de MI não necessariamente possuem alto valor de DI (esses últimos, fruto de transinformação indireta). Além disso, os pares com alto valor de DI estão majoritariamente próximos na estrutura tridimensional do complexo cocristalizado.

Outros dois métodos foram propostos para diferenciar mutações correlacionadas diretas de indiretas. O primeiro dentro da abordagem *bayesian network method* (Burger and van Nimwegen, 2010) e o segundo chamado PSICOV, no qual os contatos diretos são preditos por meio da inversão da matriz de covariância (processo que limpa os contatos indiretos - Jones et al., 2012). Esse último os autores mostraram

aumento de acurácia em relação aos métodos anteriores, mas não compararam os seus resultados com a DCA.

O GREMLIN, apresentado no tópico acima, também foi utilizado para a predição de interações resíduo-resíduo em interfaces de proteínas (Ovchinnikov et al., 2014). Nesse trabalho, o GREMLIN foi utilizado para predizer aminoácidos participantes da ligação das proteínas que formam o complexo da subunidade ribossomal 50S bacteriano. A maioria dos aminoácidos preditos estavam a uma distância inferior a 8Å e todos a uma distância inferior a 12Å.

Como visto acima o uso de correlação entre mutações pode ser utilizado para ranqueamento de soluções de *docking*. Diversos trabalhos recentes propuseram a combinação dessas métricas para melhorar os recorrentes problemas de falso-positivos de docking proteína-proteína (Othersen et al., 2011).

5. Predição de redes de interação proteína-proteína

Considerando que duas proteínas que interagem em um mesmo organismo evoluem sobre condições parecidas de modo a manter a ligação fisico-química dessa interação, é razoável pensar que essas proteínas devam ter filogenias parecidas, isto é, a árvore filogenética calculada com base no alinhamento de homólogos da proteína A deve ser parecida com a calculada com o alinhamento de homólogos da B, foi o que Goh et al. (2000) mostraram para as filogenias dos dois domínios da *phosphoglicerate kinase (PKG)*. Os autores mostraram que o coeficiente de correlação linear (R) entre as matrizes de distâncias dos dois alinhamentos era alto (0,79). Os autores sugeriram, então, que observar a correlação filogenética de duas proteínas poderia ser útil para encontrar parceiros candidatos para “ligantes órfãos” e “receptores órfãos” como no caso do sistema *chemokine-receptor*. Esse é um problema parecido com o que será abordado nessa tese e mais detalhes serão dados adiante.

Mais tarde, em 2001, Pazos and Valencia selecionaram um grupo de pares de proteínas que sabidamente interagem, outro grupo de pares que não interagem e calcularam o coeficiente de correlação linear entre alinhamentos. Eles observaram que o

R tendia a ser alto para os pares que interagem e baixo para os pares que não interagem e propuseram utilizar o R para predição de redes de interação de proteínas dada uma coleção de alinhamentos. A ideia consiste em calcular o coeficiente de correlação entre os alinhamentos de todos os pares possíveis e aqueles pares que possuem alto valor de R são pares que potencialmente interagem. Os autores propõe com base em correlações entre domínios em uma mesma proteína, que obrigatoriamente interagem, um valor de corte de 0,8, isto é, pares de proteínas com R superior a 0,8 provavelmente interagem.

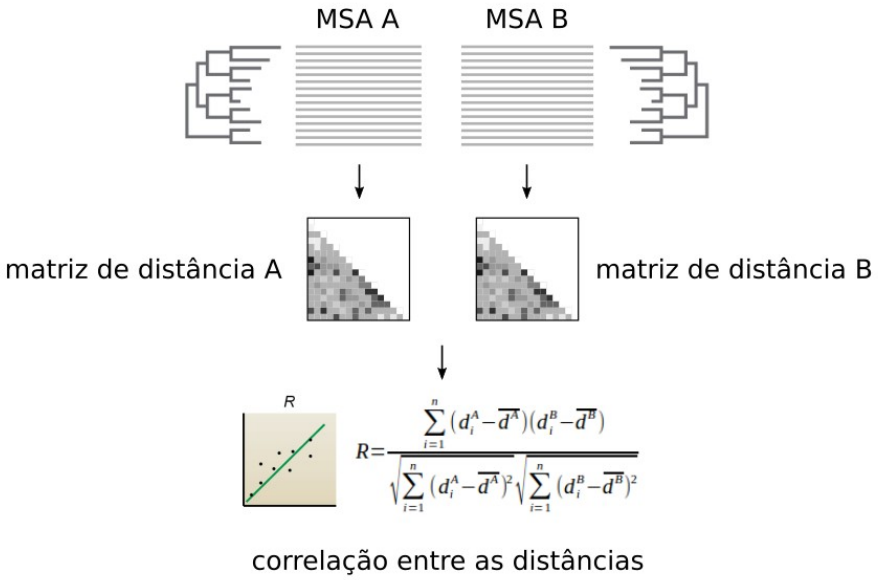


Figura 1.4. Correlação filogenética entre proteínas que interagem. *A figura acima ilustra a metodologia aplicada para calcular a correlação entre as matrizes de distâncias vindas dos alinhamentos dos homólogos da proteína A e B, onde d_i é a i-ésima entrada da matriz de distância (A ou B).*

No ano seguinte, os dois mesmos autores publicaram um artigo entitulado "*In Silico Two-Hybrid System for the Selection of Physically Interacting Protein Pairs*" (Pazos and Valencia, 2002), fazendo analogia à técnica de predição de interações entre proteínas *Two-Hybrid System*. Nesse artigo eles propuseram um possível interactoma completo de *Escherichia coli*, isto é, toda a rede de interações entre todas as proteínas do organismo.

Toda essa abordagem apresentada neste tópico consiste em uma forma diferente de pensar a coevolução, olhando agora para correlações filogenéticas e não

para correlações entre colunas específicas de aminoácidos, como na MI e na DI. Além disso, essa abordagem abriu o que chamaremos de **problema 3: Dados dois alinhamentos de proteínas homólogas A e B, é possível predizer se essas proteínas interagem?** Esse problema pode ainda ser estendido para um conjunto de proteínas de um organismo, no qual essa pergunta é feita para todos os pares possíveis e, então, é possível predizer uma rede de interações para quele conjunto de proteínas, como fez Pazos e Valencia (2002).

Ainda nessa linha de raciocínio,o artigo de Ovchinnikov et al. (2014) levanta a seguinte questão: para um complexo grande formado por interações proteína-proteína, a soma da força dos *couplings* entre os pares de proteínas do complexo pode ser utilizada para distinguir diretamente pares de proteínas que interagem de pares que não interagem? Segundo os dados apresentados pelo próprio artigo, todos os pares cuja soma dos *couplings* dá um valor maior do que 1,5 são considerados pares que de fato interagem. Há, entretanto, vários pares de proteínas que estão em contato físico que não apresentam covariação e o valor dessa soma não é maior do que 1,5 (sendo inclusive, zero, em muitos casos). Dessa forma é possível que a soma dos *couplings* maior do que 1,5 seja um forte indicativo de interação, entretanto uma soma inferior a esse valor não é suficiente para descartar a interação. Essa observação é extremamente importante. Nem todos os aminoácidos em contato covariaram e muitas vezes os métodos não são capazes de identificar a coevolução entre os pares.

6. Predição de pares de interação

Todas as abordagens apresentadas até aqui, quando utilizam alinhamentos de duas famílias de proteínas, necessitam que estes alinhamentos estejam corretamente pareados (ou concatenados), uma vez que sem isso não seria possível calcular a frequência dupla de aminoácidos. Há entretanto, uma classe de problemas na natureza em que é sabido que duas famílias de proteínas interagem, mas, o pareamento entre os homólogos das duas famílias nem sempre é conhecido. Por exemplo interações entre proteínas virais e receptores celulares ou entre neurotoxinas e canais iônicos. Esses são exemplos de interações entre proteínas de organismos diferentes (que pdoem ser

patógeno e hospedeiro) e existem algumas abordagens experimentais para predição dos pares de interação, como *Two hybrid system* e coimunoprecipitação (Bonetta, 2010). Entretanto esses métodos são caros e laboriosos, tornando a sua aplicação impraticável para estudos em larga escala. Existem algumas abordagens *in silico* para predição de pares de interação em um contexto mais simples, no qual são consideradas apenas uma espécie de patógeno e uma espécie de hospedeiro e busca-se quais proteínas interagem (Davis et al., 2007; Krishnadev and Srinivasan, 2011; Lewis et al., 2010). Essas abordagens, entretanto se enquadram no problema 3 apresentado anteriormente.

Dessa maneira, faltam ferramentas que permitam identificar o pareamento correto de duas famílias de proteínas que interagem e que estão em tipos de organismos diferentes, em outras palavras, encontrar a concatenação correta entre dois alinhamentos. Aparece aqui, então, o quarto e último problema abordado nesta tese, problema esse que é o principal foco das investigações apresentadas nos capítulos subsequentes. O **problema 4** é: **Dados um par de proteínas que sabidamente interagem, é possível predizer o pareamento correto entre os alinhamentos A e B?**

Esse problema é extremamente difícil, dado que o número de soluções possíveis é *M!*, onde *M* é o número de sequências em cada alinhamento (e isso assumindo que cada sequência *A* irá interagir apenas com uma sequência de *B*). Uma visão mais simples desse problema foi tratada algumas vezes na literatura: o problema dos **parálogos**[2]. Considerando duas proteínas que interagem e considerando que os genes codificadores para essas proteínas aparecem duplicados nos diversos genomas que possuem esse gene, não é possível saber qual parálogo interage com qual. Perceba que este é um subproblema do **problema 4**, porque aqui embora eu saiba que as proteínas parálogas de *A* da espécie 1 interagem apenas com os parálogas de *B* da espécie 1, eu não sei como essas interações estão distribuídas. A **Figura 1.5** ilustra o problema dos parálogos e o problema 4.

2 Genes em um mesmo genoma que sofreram duplicação, sendo portanto um subtipo de gene homólogo.

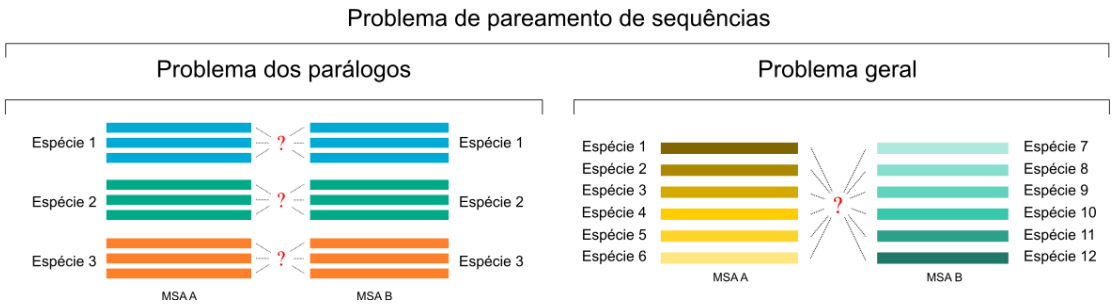


Figura 1.5. Diferentes abordagens do problema 4. *O problema dos parálogos consiste em identificar qual dos parálogos interage com qual, é um problema mais simples do que o problema geral, uma vez que as proteínas de uma mesma espécie são codificadas pelo mesmo genoma e no problema geral as proteínas estão em organismos diferentes.*

Embora alguns poucos trabalhos tenham abordado o problema 4 de maneira ampla (Pazos and Valencia, 2002; Tillier et al., 2006), ele permanece como um desafio a ser resolvido e métodos capazes de predizer a concatenação correta de alinhamentos de pares de proteínas que interagem se fazem necessários. Nesse contexto, o trabalho apresentado no próximo capítulo tem como objetivo caracterizar diferentes fontes de informação (coevolutiva, filogenética e estocástica) e identificar o papel de cada uma delas em diferenciar arranjos nativos de pareamento de arranjos sem significado biológico, o que poderá ajudar os pesquisadores a encontrar uma forma de resolver o problema 4.

Capítulo 2. Informação coevolutiva, filogenética e estocástica em interações proteína-proteína

1. Introdução

Como visto, interações proteína-proteína são processos-chave para a existência e manutenção da vida. As proteínas sofrem pressões seletivas para a manutenção de sua estrutura e para a manutenção das interações. As pressões seletivas para manutenção da interação entre duas proteínas A e B podem ocasionar covariação de aminoácidos na interface de interação destas proteínas.

Otimizações de transinformação (*I*) e correlação filogenética (*R*) já foram utilizadas no passado para a predição de pares de interação proteína-proteína(Pazos and Valencia, 2002; Tillier et al., 2006). Entretanto, tais otimizações foram realizadas com um número baixo de sequências (em geral artificiais). Solucionar o problema 4 permanece como um desafio importante para o estudo de interações proteína-proteína bem como para a biologia de sistemas. É sabido que os valores de *I* entre pares de aminoácidos não são exclusivamente resultados da coevolução. Embora no passado muito esforço tenha sido dedicado em eliminar a influência filogenética dos valores de transinformação (Dunn et al., 2008) a compreensão da contribuição das fontes (coevolutivas, filogenéticas e estocásticas) de transinformação ainda não está elucidada e o trabalho apresentado neste capítulo se insere justamente nessa abordagem.

Dessa forma, em um cenário atual, onde temos um grande número de sequências homólogas[3] para diversas famílias disponíveis em bancos de dados, estamos revisitando o campo, quantificando quanto que uma proteína *A* guarda de informação sobre uma proteína *B*, o quanto que é decorrente de cada uma dessas fontes e, mais importante, as suas contribuições para determinar o pareamento correto de PPI baseado em um MSA. Este trabalho se enquadra no **problema 4** apresentado no **Capítulo 1**, que

3 Sequências que possuem uma origem evolutiva comum.

se preocupa com a predição dos pares de interação proteína-proteína dado dois alinhamentos de homólogos de duas famílias A e B.

2. Objetivos

2.1. Objetivo geral

O **objetivo geral** deste trabalho é caracterizar o papel de diferentes fontes de transinformação na capacidade de predição de arranjos de interação proteína-proteína.

2.2. Objetivos específicos

1. Medir os valores de transinformação para diferentes valores de R_c;
2. Isolar as contribuições das diferentes fontes de transinformação;
3. Avaliar a degeneração das correlações;
4. Avaliar a dependência dos valores de *I* para diferentes modelos de *docking.*

3. Teoria, materiais e métodos

3.1. Transinformação

Considere duas proteínas *A* e *B* que interagem formando $i = 1, \ldots, N$ contatos tridimensionais de aminoácidos. Homólogos das proteínas A e B podem ser alinhados formando o alinhamento A (MSA^A) e o alinhamento B (MSA^B). No nosso modelo, cada sequência do MSA^A pode ser conectada com uma sequência do MSA^B. O arranjo dessas conexões é chamado de *z* e é descrito como uma variável estocástica *Z* com uma função de massa de probabilidade $\rho(z)$, $\forall\, z \in \{1, 2, 3, \ldots, M\}$, onde *M* é o número de sequências em cada alinhamento. Em termos práticos, *z* é um vetor que mapeia as sequências B nas sequências A. Os contatos tridimensionais são extraídos do arquivo de estrutura do complexo formado pela ligação das duas proteínas (.pdb) e são definidos por uma distância de proximidade entre os centros geométricos, isto significa que um par de

aminoácidos (um da proteína *A* e outro da proteína *B*) estão em "contato" se o centro geométrico desses aminoácidos estiverem a uma distância menor ou igual a R_c^*. Usaremos inicialmente a distância $R_c^* = 8Å$, . A **Figura 2.1** ilustra a identificação dos contatos na estrutura e no alinhamento de um complexo hipotético.

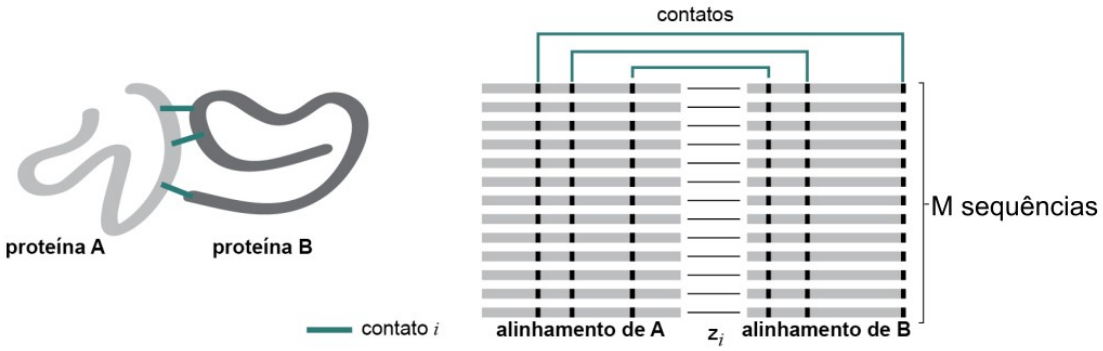


Figura 2.1. Representação esquemática do modelo. *À esquerda temos duas estruturas de proteínas que interagem, A e B, com os contatos de interação inter-proteína marcados em verde. À direita temos os alinhamentos (cinza) de M homólogos de A e de B, com os contatos destacados em verde e pintados em preto na sequência. Conectando os dois alinhamentos temos o arranjo* z_i.

Dado um arranjo específico *z*, as posições (colunas do alinhamento) de aminoácidos que interagem podem ser extraídas gerando um sub-alinhamento. Neste alinhamento, as sequências podem ser descritas como variáveis estocásticas $X^N \equiv (X_1, \ldots, X_N)$ que descreve as sequências de A e $Y^N \equiv (Y_1, \ldots, Y_N)$ que descreve as sequências de B. Essas variáveis estocásticas possuem funções de massa de probabilidade $\{\rho(X^N), \rho(Y^N), \rho(X^N, Y^N|z)\}$, tais que

$$\begin{cases} \rho(X^N) = \sum_{Y^N} \rho(X^N, Y^N|z) \\ \rho(Y^N) = \sum_{X^N} \rho(X^N, Y^N|z) \end{cases} \quad (1)$$

e

$$\sum_{X^N, X^N} \rho(X^N, Y^N|z) = 1 \quad (2)$$

para cada sequência conjunta $\{x_1, \ldots, x_N, y_1, \ldots, y_N\}_{|\chi|^{2N}}$ definida no alfabeto χ de tamanho $|\chi|$. Sob essas considerações, a quantidade de informação que a proteína *A*

guarda sobre a proteína *B* é dada pela informação mútua (ou transinformação) $I(X^N;Y^N|z)$ entre X^N e Y^N condicionadas ao arranjo *z*. Como explicitado na equação (1), nós estamos particularmente interessados em quantificar $I(X^N;Y^N|z)$ para a situação em que o bloco de variáveis $\{\rho(X^N),\rho(Y^N)\}$ são assumidas como independentes do arranjo *z*, isto é, para uma composição fixa de sequências *A* e *B* somente a distribuição conjunta depende do arranjo *z*. Além disso, assumindo *N*-contatos independentes, nós queremos que a informação seja quantificada para o modelo com vínculos mínimos $\{\rho^*(X^N,Y^N|z)\}$ que maximiza a entropia condicional conjunta entre *A* e *B* – essa condição garante que a transinformação seja escrita exatamente em termos das contribuições individuais dos contatos *i*.

Para a distribuição crítica $\{\rho^*(X^N,Y^N|z)\}$ a informação mútua

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

pode ser escrita em termos da entropia de Shannon

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

associada com a distribuição conjuntas $\{\rho^*(X^N,Y^N|z)\}$ e as marginais derivadas $\{\rho^*(X^N),\rho^*(Y^N)\}$. A partir da propriedade de maximização da entropia, a distribuição crítica $\{\rho^*(X^N,Y^N|z)\}$ fatoriza na marginal condicional de cada contato *i*

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

permitindo assim que a equação [4] possa ser escrita extensivamente, em termos das contribuições entrópicas individuais

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

de maneira que

$$I(X^N; Y^N|z) = \sum_i^N I(X_i, Y_i|z) \quad (7)$$

Na equação (7) a transinformação alcança limite inferior de zero se X^N e Y^N são condicionalmente independentes dado *z*, por exemplo, $\rho^*(x^N, y^N|z) = \rho^*(X^N) \times \rho^*(Y^N)$. Para o caso de variáveis perfeitamente correlacionadas $\rho^*(x^N, y^N|z) = \rho^*(X^N) = \rho^*(Y^N)$, a transinformação está ligada ao máximo e não pode exceder a entropia de cada um dos blocos $H(X^N)$ e $H(Y^N)$.

O desenvolvimento acima, nos permite dizer que a transinformação (I) entre o bloco X^N e o bloco Y^N é dada pela soma das entropias das variáveis X_i e Y_i, que são as colunas extraídas do alinhamento. A $I(X_i, Y_i|z)$ é calculada na forma

$$I(X_i, Y_i|z) = \sum_{X_i Y_j} \rho(x_i, y_i|z) \ln \frac{\rho(x_i, y_i|z)}{\rho(x_i)\rho(y_i)} \quad (8)$$

onde o somatório percorre todos os pares de aminoácidos possíveis (21 x 21), note que as probabilidades simples $\rho(x_i)$ e $\rho(y_i)$ não dependem do arranjo *z*, já a probabilidade conjunta $\rho(x_i, y_i|z)$ depende. O alfabeto de valores que as variáveis X_i e Y_i podem assumir tem tamanho 21, os 20 aminoácidos mais o *gap* (-).

Dado um conjunto de contatos de aminoácidos conhecidos de um par de proteínas e a sua distribuição de sequências primárias subjacentes definindo as variáveis estocásticas X^N e Y^N, a Eq. [7] estabelece a dependência formal da transinformação com qualquer processo *z* dado.

Em termos práticos, a transinformação, Eq. [7], foi calculada normalizada pela entropia $H(X_i, Y_i|z)^{-1} I(X_i; Y_i|z)$ de modo a superar as contribuições superestimadas de sítios com alta variabilidade (Dunn et al., 2008).

3.2. Probabilidades conjuntas

Para um dado MSA, as probabilidades conjuntas $\rho^*(x_i, y_i|z) \equiv f_{x_i, y_i|z}$ são definidas a partir das frequências observadas $f_{x_i, y_i|z}$ regularizadas por uma fração efetiva de pseudocontagem λ^* no caso de amostragem insuficiente, como proposto por Morcos et al. (2011). Mais especificamente, as frequencias duplas são calculadas de acordo com

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

Onde $n_z^m = |\{m' | 1 \leqslant m' \leqslant M, \textit{distância de hamming}(m, m') \geqslant \delta h\}|$ é o número de sequências similares *m'* com uma certa distância de hamming δh da sequência m e

$M_z^{eff} = \sum_{m=1}^{M} (n_z^m)^{-1}$ é o número efetivo de sequências a um limite de distância.

3.3. Estruturas e alinhamentos

Os alinhamentos utilizados neste estudo foram retirados do trabalho de Ovchinnikov et al. (2014) e as estruturas foram baixadas do PDB. A **Tabela 2.1** apresenta todos os pares de interação a serem estudados.

Tabela 2.1. Relação das estruturas utilizadas. *O sistema 1BXR foi destacado em alguns momentos no texto. M corresponde ao número de sequências no alinhamento e o Tamanho apresentado é o número de posições no alinhamento concatenado.*

Descrição do complexo	PDB ID	Proteína A	Proteína B	M	Tamanho
Carbamoyl Phosphate Synthetase	1BXR	**Chain A:** Carbamoyl-Phosphate Synthetase large subunit	**Chain B:** Carbamoyl-Phosphate Synthetase small subunit	1004	1452
Lactococcus Lactis Dihydroorotate	1EP3	**Chain A:** Dihydroorotate Dehydrogenase B (PYRD	**Chain B:** Dihydroorotate Dehydrogenase B (pyrk	552	572

Dehydrogenase B.		Subunit)	Subunit)		
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
Toxin-antitoxin complex RelBE2 from Mycobacterium tuberculosis	3G5O	ChainA: Protein Rv2865	ChainB: Protein Rv2866	904	173

A interação entre as subunidades A e B do sistema 1BXR (*Carbamoyl Phosphate Synthetase*, uma enzima hétero dimérica composta de duas subunidades (*small* e *large*) que se organizam formando a estrutura representada na **Figura 2.2** será utilizada como modelo principal de estudo deste trabalho.

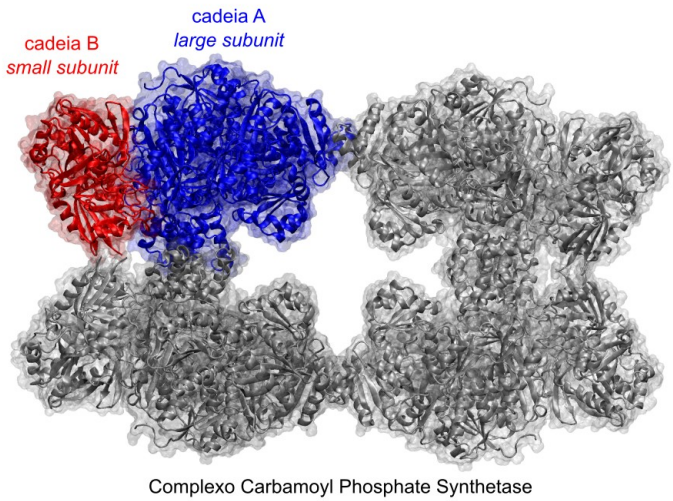


Figura 2.2. Representação da estrutura tridimensional do sistema modelo 1BXR. *Em destaque a cadeia A (vermelho) e a cadeia B (azul). A interação entre estas duas cadeias será o modelo de estudo deste trabalho.*

4. Resultados e discussão

Para o estudo aqui apresentado, selecionei pares de proteínas que sabidamente interagem e cuja estrutura tridimensional está resolvida por co-cristalização e disponível no *Protein Data Bank* (PDB), que serão chamados aqui de "sistema". Os sistemas analisados são: 1BXR-AB, 1EP3-AB, 2VPZ-AB, 2D1P-BC, 3RRL-AB, 2Y69-AB e 3G5O-AB. Embora os resultados sejam robustos e consistentes para todos os sistemas irei eventualmente enfatizar o 1BXR-AB, apenas para facilitar/simplificar a visualização dos dados.

Vale ressaltar também que, como esse é um estudo de caracterização e decomposição das fontes de transinformação e não temos (ainda) o objetivo de resolver o **problema 4** apresentado no Capítulo 1, esses sistemas possuem o arranjo de pareamento de sequências (*z*) conhecido (dado que as proteínas que interagem são codificadas por genes no mesmo genoma).

4.1. A transinformação média em função da distância do par de aminoácidos

Para avaliar como os valores de transinformação entre os pares de aminoácidos (AA) se comportam em função da distância entre eles calculei para cada par a transinformação do par dado o arranjo correto $I(X_i, Y_i|z^*)$ e a transinformação do par dado um arranjo aleatório qualquer $I(X_i, Y_i|z^{\text{rand}})$ e plotei o valor médio de transinformação para cada intervalo de distância em Å (**Figura 2.3**).

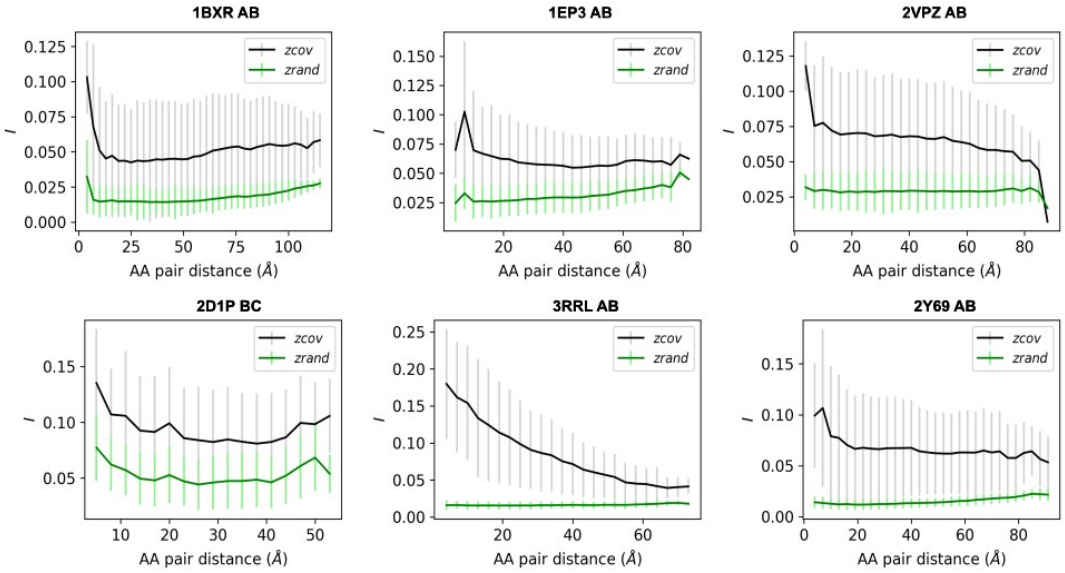


Figura 2.3. Transinformação vs distância do par de aminoácidos. *Transinformação média* $\langle I(X^N; Y^N|z)\rangle$ *calculada para diferentes janelas de distância entre pares de aminoácidos. As barras indicam os desvios.*

A transinformação média tende a ser maior em distâncias menores, resultado similar às descrições da literatura (Ovchinnikov et al., 2014). Tendo como base esses resultados e a literatura, é possível definir que aminoácidos em contato são aqueles que estão a uma distância menor ou igual a 8Å. Os pares de aminoácidos com $R_c \leq 8Å$ serão chamados de "fisicamente acoplados" e os demais de "fisicamente não acoplados". A **Figura 2.4** mostra os aminoácidos fisicamente acoplados destacados em verde e os não acoplados em cinza na estrutura tridimensional de cada complexo.

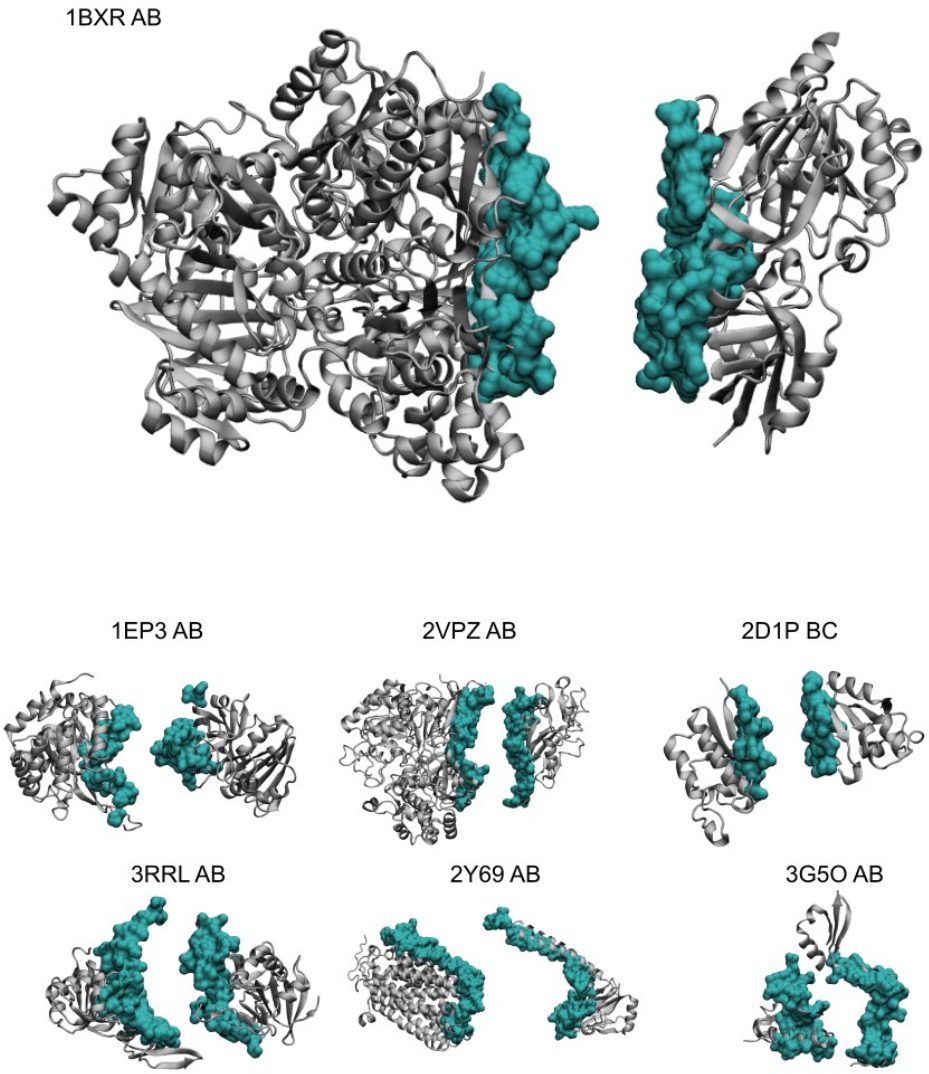


Figura 2.4. Estrutura tridimensional dos sistemas. *Em verde são os aminoácidos fisicamente acoplados (Rc ≤ 8Å) e em cinza os fisicamente não acoplados.*

4.2. A transinformação total cai conforme *z* é embaralhado

Com o objetivo de entender como é o comportamento da transinformação total entre as duas proteínas calculei a transinformação dos blocos $I\left(X^N;Y^N|z\right)$ para diferentes níveis de embaralhamento de *z*. Os valores foram plotados em função de $M-n$, onde *M* é o número de pares de sequências no alinhamento e n é o número de sequências pareadas corretamente ($M-n$ grande significa que muitas sequências estão

trocadas e o nível de embaralhamento é maior). Os valores médios $\langle I(X^N;Y^N|z)\rangle$ para diferentes $M-n$ do sistema 1BXR-AB estão apresentados na **Figura 2.5.**

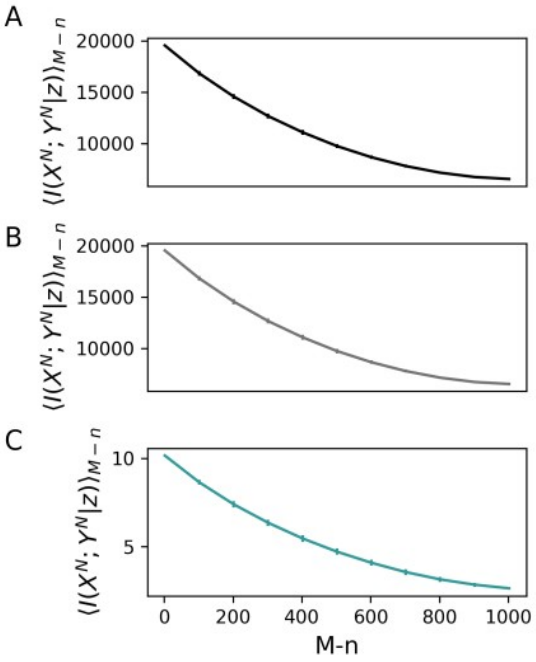


Figura 2.5. Transinformação do bloco dado o nível de embaralhamento. *Valores de* $\langle I(X^N;Y^N|z)\rangle$ *vs* $M-n$ *para o bloco contendo todos os pares (A), apenas os pares fisicamente não acoplados (B) e apenas os pares fisicamente acoplados (C).*

De maneira esperada, os valores de $\langle I(X^N;Y^N|z)\rangle$ caem conforme o nível de embaralhamento aumenta, uma vez que as mutações correlacionadas vão perdendo sentido conforme as frequências duplas são alteradas. Curiosamente $\langle I(X^N;Y^N|z)\rangle$ não vai para zero nos arranjos *z* completamente embaralhados ($M-n=1004$ para o sistema 1BXR-AB). Valores similares de transinformação são encontrados quando os MSAs são completamente embaralhados, isto é, trocas completamente aleatórias entre as linhas e as colunas. Dessa maneira, podemos sugerir que esses valores de $\langle I(X^N;Y^N|z)\rangle$ são gerados ao acaso e depende da distribuição total de aminoácidos no alinhamento.

A subtração dessa fonte estocástica da transinformação nativa, computada na forma de um *gap* informacional($\Delta I_{M-n}\equiv\left|I(X^N;Y^N|z^*)-\langle I(X^N;Y^N|z)\rangle_{M-n}\right|$)entre o MSA referência e o "completamente embaralhado", revela as contribuições não

estocásticas isoladas da correlação total entre as proteínas *A* e *B*. A **Figura 2.6** mostra os valores de ΔI_{M-n} para os diferentes sistemas.

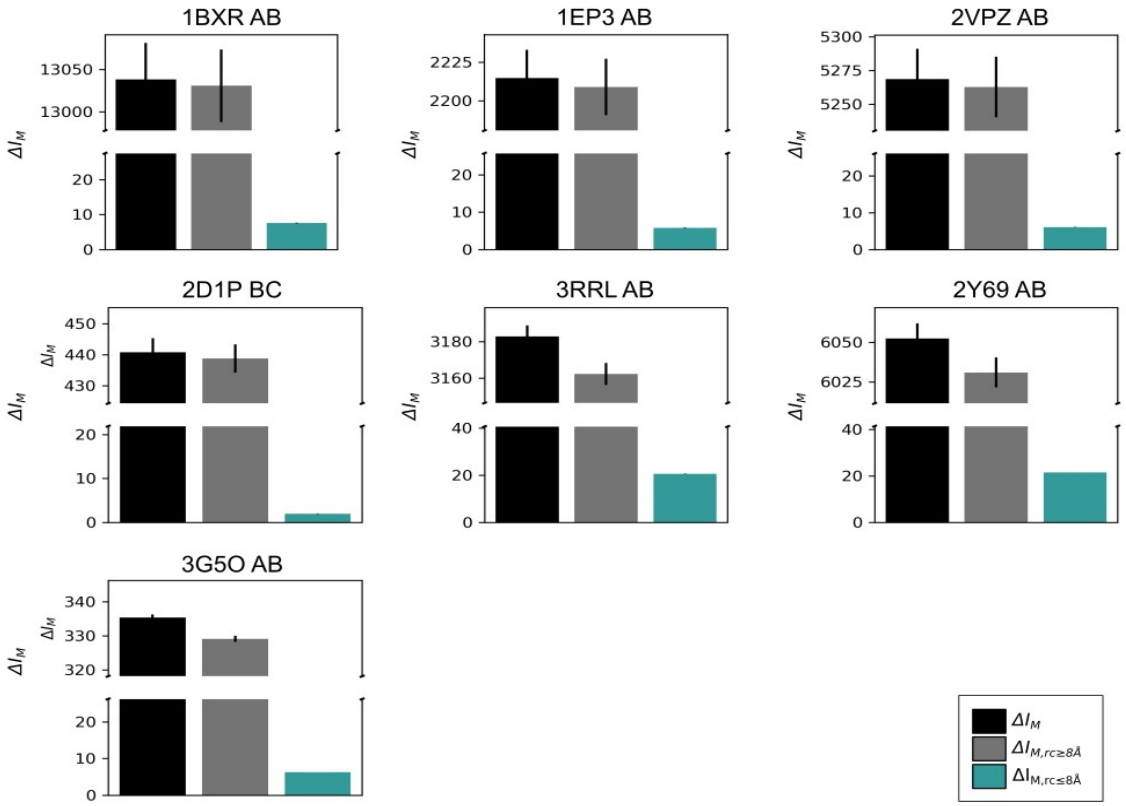


Figura 2.6. Gap de transinformação entre o arranjo natural e o aleatório. *Valores de* ΔI_{M-n} *para os diferentes sistemas. As barras de erro indicam os desvios, uma vez que* $\langle I(X^N;Y^N|z)\rangle_{M-n}$ *é uma média de 5 redes aleatórias. As cores indicam os diferentes blocos conforme a legen*da.

4.3. A interface guarda mais transinformação média por contato

Os resultados apresentados na **Figura 2.6** apresentam a propriedade extensiva da Eq. (7), de modo que $\Delta I_M = \Delta I_{M,rc>8Å} + \Delta I_{M,rc\leq 8Å}$. Devido a essa propriedade, as contribuições individuais crescem com o tamanho do bloco (*N*). Quando normalizamos os valores pelo tamanho do bloco, a contribuição $N^{-1}\Delta I_{M,rc}$ fica diferente e passa a ser maior para o bloco dos aminoácidos fisicamente acoplados, conforme pode ser visto na **Figura 2.7**. Dessa forma, concluímos que a interface é a região da proteína que guarda maior transinformação média por contato.

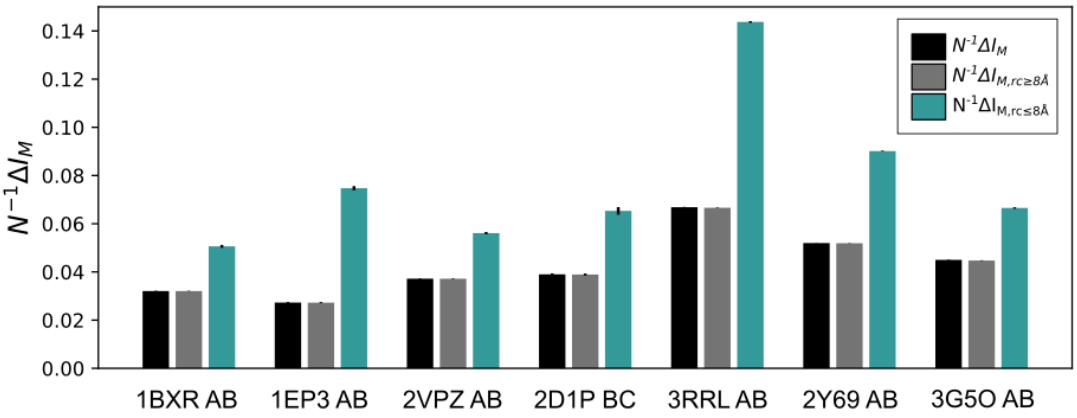


Figura 2.7. Gap de transinformação normalizado entre o arranjo natural e o aleatório. *Valores de* $N^{-1}\Delta I_{M-n}$ *para os diferentes sistemas. As barras de erro indicam os desvios. As cores indicam os diferentes blocos conforme a legenda.*

4.4. Decomposição da transinformação.

Codoñer e Fares (2008) em uma revisão apresentaram que o *coupling* entre um aminoácido *i* e *j* pode ser $C_{ij}=C_{filogenia}+C_{estrutura}+C_{função}+C_{interações}+C_{estocástico}$. Utilizando tal trabalho como inspiração, propusemos aqui a decomposição de $I\left(X^N;Y^N|z\right)$, ou apenas I, como $I=I_{filogenia}+I_{estrutura}+I_{função}+I_{interações}+I_{estocástico}$. Considerando que a interação em questão é entre duas proteínas, podemos assumir que $I_{coevolução}=I_{estrutura}+I_{função}+I_{interações}$, uma vez que as pressões seletivas que fazem com que os aminoácidos covariem para a manutenção da interação está relacionada com a manutenção da estrutura (para ligar a estrutura precisa ser compatível) e da função (para funcionar a estrutura precisa estar compatível e a ligação deve acontecer). Podemos decompor I, então, como $I=I_{filogenia}+I_{coevolução}+I_{estocástico}$. Os resultados apresentados até então nos permite conhecer o valor de $I_{estocástico}$ e $I_{filogenia}+I_{coevolução}$, que é o $N^{-1}\Delta I_{M-n}$. Assumindo que os aminoácidos fisicamente não acoplados não possuem covariações fruto de coevolução, apenas fruto da filogenia (ou evolução), podemos decompor $N^{-1}\Delta I_{M-n}$ da seguinte forma:

$$N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\,\mathring{A}} \overset{\text{def}}{=} N^{-1}\Delta I_{M,r_c\leq 8\,\mathring{A}} - N^{-1}\Delta I_{M,r_c>8\,\mathring{A}} \quad (10)$$

Os valores $N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\,\mathring{A}}$ para os diferentes sistemas analisados estão plotados na **Figura 2.8** em verde. De acordo com a definição da Equação (11), podemos concluir que aproximadamente 40% do conteúdo de informação guardado nos aminoácidos fisicamente acoplados do sistema 2D1P, por exemplo, resulta somente da coevolução. Outros trabalhos anteriores eliminaram o efeito filogenético da transinformação (Dunn et al., 2008) e o efeito das correlações indiretas (Morcos et al., 2011). com o objetivo de comparar $N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\,\mathring{A}}$ com essas duas abordagens desenvolvidas no passado e que tem efeito de, teoricamente, capturar apenas o que é fruto de covariação por contato direto, ou seja, coevolução, calculei a *MIp* para todos os sistemas e a DI apenas para o sistema 2D1P (em razão de limitações computacionais) e apresento esses resultados também na **Figura 2.6**. Podemos notar que os valores de $N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\,\mathring{A}}$ são compatíveis com os valores de *MIp* e *DI*.

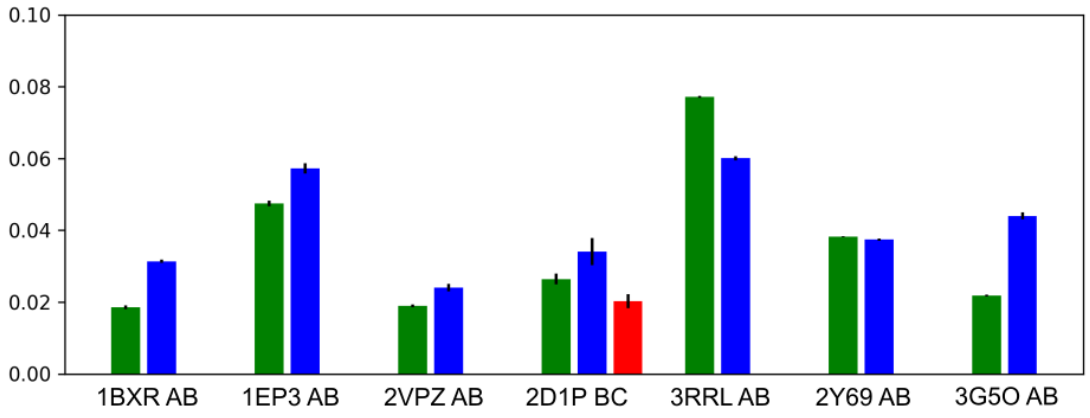


Figura 2.8. Decomposição da transinformação. *Valores de $N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\,\mathring{A}}$ plotados em verde (verde) comparados com valores de Mip (azul) e DI (vermelho, apenas para o sistema 2D1P devido a limitações computacionais).*

4.5. Análise de degeneração e erro

A presente análise revela as diferenças quantitativas entre as correlações provenientes dos aminoácidos fisicamente acoplados e dos fisicamente não acoplados. Considerando que o total de transinformação não dependente do acaso $N^{-1}\Delta I_{M,rc}$ é uma medida não enviesada (intensiva) do acoplamento, nós podemos comparar a contribuição efetiva de $N^{-1}\Delta I_{M,r_c\leq 8\mathring{A}}$ e $N^{-1}\Delta I_{M,r_c>8\mathring{A}}$ em discriminar arranjos de concatenação de MSAs, chamaremos a partir de agora cada arranjo de concatenação de modelo de MSA.

Dessa maneira, podemos definir o número total $\omega_{\boldsymbol{S}}$ de soluções *native-like* de modelos de MSA gerados pela randomização dos pares de sequência $M-n$ no alinhamento referência

$$\omega_{\boldsymbol{S}(r_c)} \equiv \sum_{n\in \boldsymbol{S}(r_c)} \omega_{M,n} \tag{11}$$

em termos de *rencontres numbers* temos:

$$\omega_{M,n} = \frac{M!}{n!}\sum_{q=0}^{M-n}\frac{(-1)^q}{q!} \tag{12}$$

ou permutações no pareamento refereência com n posições fixas satisfazendo $\sum_{n=0}^{M}\omega_{M,n} = M!$ (em linguagem combinatória). Aqui, $\boldsymbol{S}(r_c)$ denota o conjunto de posições fixas n

$$\boldsymbol{S}(r_c) \equiv \{n | 0\leq n\leq M, N^{-1}\Delta I_{M-n,r_c}\leq \delta I\} \tag{13}$$

Para o qual o gap de transinformação $N^{-1}\Delta I_{M,rc}$ é menor do que uma certa resolução δI independentemente do bloco de tamanho N correspondente ou do número de aminoácidos em contato. De maneira simples, $\omega_{\boldsymbol{S}}$ nos informa a degeneração do número

de modelos de MSA com uma quantidade de transinformação similar ao do MSA referência.

Tabela 2.2. Rencontres numbers para alguns dos sistemas analisados.

M-n	1BXR AB	1EP3 AB	2VPZ AB	2D1P BC	3RRL AB	2Y69 AB
0	1	1	1	1	1	1
1	0	0	0	0	0	0
2	503506	152076	228150	23220	883785	1100386
3	336342008	55761200	102515400	3312720	782444320	1087181368
4	378763143759	34439511150	77616972225	793810530	1168091564220	1811380056759
5	370346185008800	18453455396640	50999525216640	164548102752	1514469649396704	2621268206581024
…	…	…	…	…	…	…
40	1,963993579054E+119	4,115347994062E+108	1,788169031032E+112	1,8626849992297E+91	1,830259053207E+124	1,558622316946E+126

Como mostrado na **Tabela 2.2**, os valores de *rencontres numbers* $\omega_{M,n}$ crescem astronomicamente em função de $M-n$, de maneira idêntica independente da definição de contatos utilizada e do tamanho do bloco, uma vez que esses valores dependem apenas do *M* e *n*. Entretanto, o número de modelos de MSA *native like* ω_s depende da utilização de aminoácidos acoplados ou não, conforme pode ser observado na **Figura 2.9A**. Esse número é muito menor para as definições de X^N e Y^N envolvendo os aminoácidos fisicamente acoplados, isso implica que o número de trocas $M-n$ para perturbar o mesmo valor δI é maior (**Figura 2.9B**) para as definições de X^N e Y^N baseadas nos aminoácidos fisicamente não acoplados.

Podemos calcular o valor esperado

$$\langle \epsilon \rangle_S = \sum_{n \in S} \left(M \sum_{n \in S} \omega_{M,n} \right)^{-1} \omega_{M,n} \times n \qquad (14)$$

de conexões iguais às do modelo de MSA referência para diferentes valores de δI. Na **Figura 2.9C** podemos observar que $\langle \epsilon \rangle_S$ é maior para as definições levando em conta os aminoácidos fisicamente acoplados para diversos valores δI.

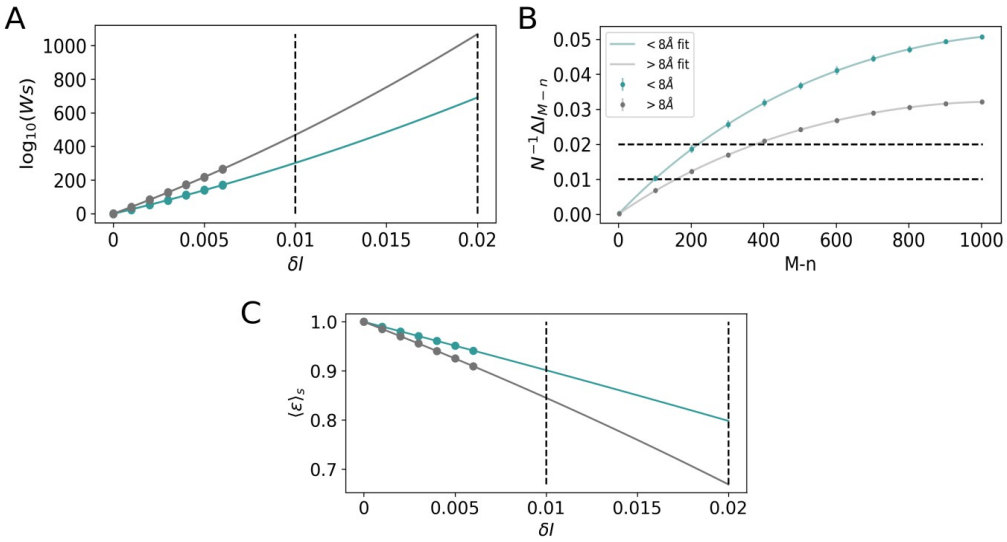


Figura 2.9. Análise de degeneração e erro das variáveis estocásticas X^N e Y^N envolvendo amonoácidos fisicamente acoplados (turquesa) e não acoplados (cinza). *(A) número total* ω_S *de modelos de MSA native-like em diferentes resoluções* δI *(B) gap de transinformação por contato* $N^{-1}\Delta I_{M,rc}$ *em função do número de sequencias embaralhadas M-n. (C) Valores esperados* $\langle\epsilon\rangle_S$ *para diferentes valores de* δI *.*

Esses resultados sugerem que as definições baseadas nos aminoácidos fisicamente acoplados($N^{-1}\Delta I_{M,r_c\leq 8\mathring{A}}$) são melhores do que as baseadas em aminoácidos fisicamente não acoplados($N^{-1}\Delta I_{M,r_c>8\mathring{A}}$) em discernir entre pareamentos *coevoluton-like* de pareamentos gerados ao acaso, uma vez que possuem menor degeneração e um valor esperado de conexões certas maior. Isso significa que na vizinhança de valores de $N^{-1}\Delta I_{M,r_c\leq 8\mathring{A}}$ existem menos redes possíveis do que na vizinhança de $N^{-1}\Delta I_{M,r_c>8\mathring{A}}$.

4.6. Os valores de transinformação dependem da definição de contatos e da interface escolhida.

Definimos até agora o valor de corte para considerarmos um par de aminoácidos como fisicamente acoplados um Rc* de 8Å, de modo que os aminoácidos de distâncias pequenas (chamados até agora de fisicamente acoplados) são aqueles cujas

distâncias são menores ou iguais a Rc* e os aminoácidos de distância grande (chamados até agora de fisicamente não acoplados) são aqueles cujas distâncias são maiores do que Rc*. Com o objetivo de avaliar a dependência dos *gaps* de transinformação com a definição de contatos, calculei $N^{-1}\Delta I_{M,r_c \le r_c^*}$ e $N^{-1}\Delta I_{M,r_c > r_c^*}$ para diferentes valores de R_c^* (**Figura 2.10**). O *gap* de transinformação $N^{-1}\Delta I_{M,r_c \le r_c^*}$ depende de R_c^* (com picos entre 6 e 8Å), entretanto $N^{-1}\Delta I_{M,r_c > r_c^*}$ não depende de R_c^*, o que faz muito sentido, uma vez que $N^{-1}\Delta I_{M,r_c > r_c^*}$ capta justamente o efeito da filogenia (ou da evolução) na transinformação. Enquanto $N^{-1}\Delta I_{M,r_c \le r_c^*}$ capta o efeito da coevolução mais os efeitos evolutivos.

A título de comparação, também calculei os valores de *Mip* utilizando dois grupos de pares de aminoácidos. Os valores de *MIp* seguiram a mesma lógica, entretanto com valores menores, uma vez que não há o efeito da evolução. Interessantemente, os valores de *MIp* calculados para os "contatos" de longa distância são sempre muito próximos de zero, conforme pode ser visto no gráfico tracejado azul na **Figura 2.10**.

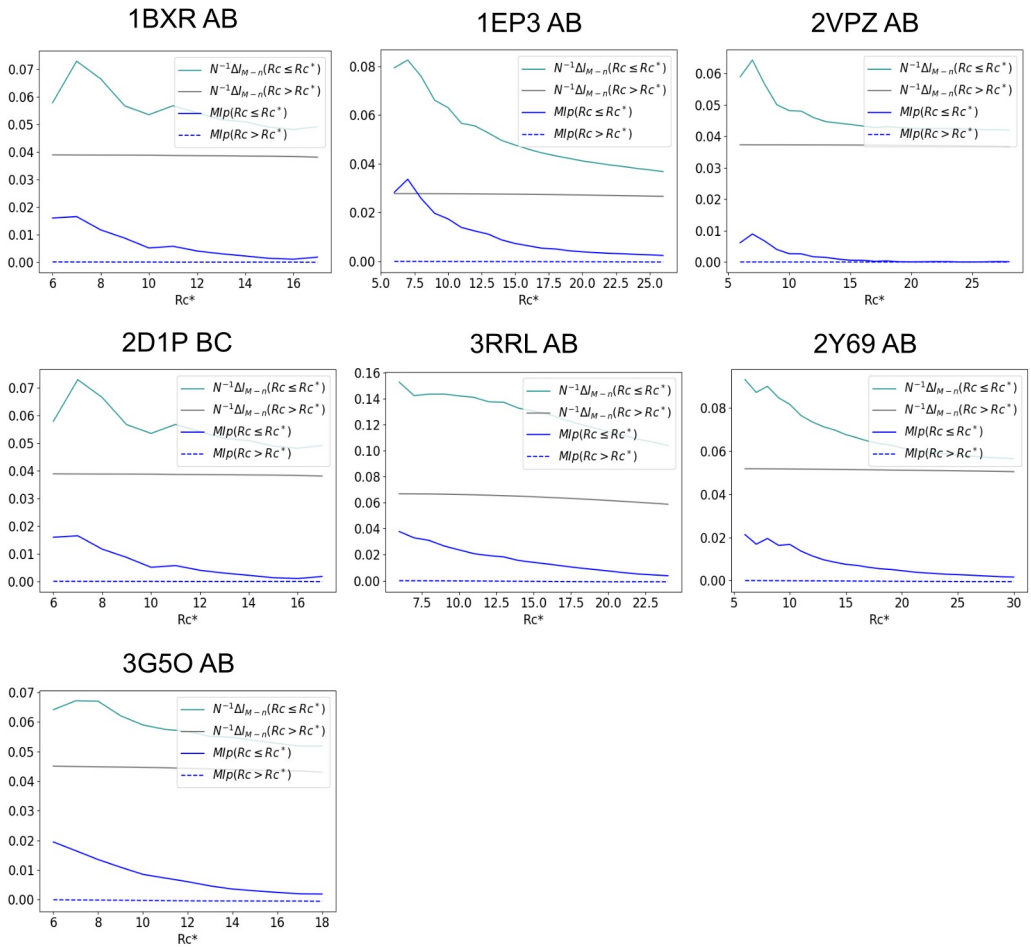


Figura 2.10. Transinformação em função de R_c^*. *$N^{-1}\Delta I_{M,r_c\le r_c^*}$ (turquesa), $N^{-1}\Delta I_{M,r_c>r_c^*}$ (cinza), $MIp_{r_c\le r_c^*}$ (azul) e $MIp_{r_c\ge r_c^*}$ (azul serrilhado).*

Esses resultados se mostraram robustos para diferentes sistemas. Agora, o que aconteceria se fizéssemos os mesmos cálculos para interfaces de interação alternativas (erradas). Para avaliar tal situação, selecionei 10 interfaces alternativas geradas por *docking* molecular (GRAMM-X, Tovchigrechko and Vakser, 2006) para todos os sistemas e calculei para cada interface alternativa o *gap* de transinformação $N^{-1}\Delta I_{M-n}$, a transinformação fruto da coevolução $N^{-1}\Delta\Delta I^{Cov}_{M,r_c\le 8\,\mathring{A}}$ e *MIp*. A **Figura 2.11** mostra os resultados para o sistema 1BXR-AB e a **Figura 2.12** para os demais sistemas. Os valores de $N^{-1}\Delta\Delta I^{Cov}_{M,r_c\le 8\,\mathring{A}}$ são altos apenas para interfaces com baixo são parecidas com a interface original, interfaces que possuem baixo RMSD quando comparadas ao cristal.

É importante notar que a transinformação fruto da coevolução $N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\,\mathring{A}}$ e *Mip* possuem valores mais próximos de zero do que o *gap* de transinformação $N^{-1}\Delta I_{M-n}$ para modelos ruins (alto RMSD), isso porque o *gap* contem as contribuições filogenéticas, que estão presentes em toda a proteína, enquanto a transinformação fruto da coevolução não possui tais contribuições. Esses resultados nos levam a propor que essas medidas podem ser úties em conjunto para o ranqueamento de soluções de Docking.

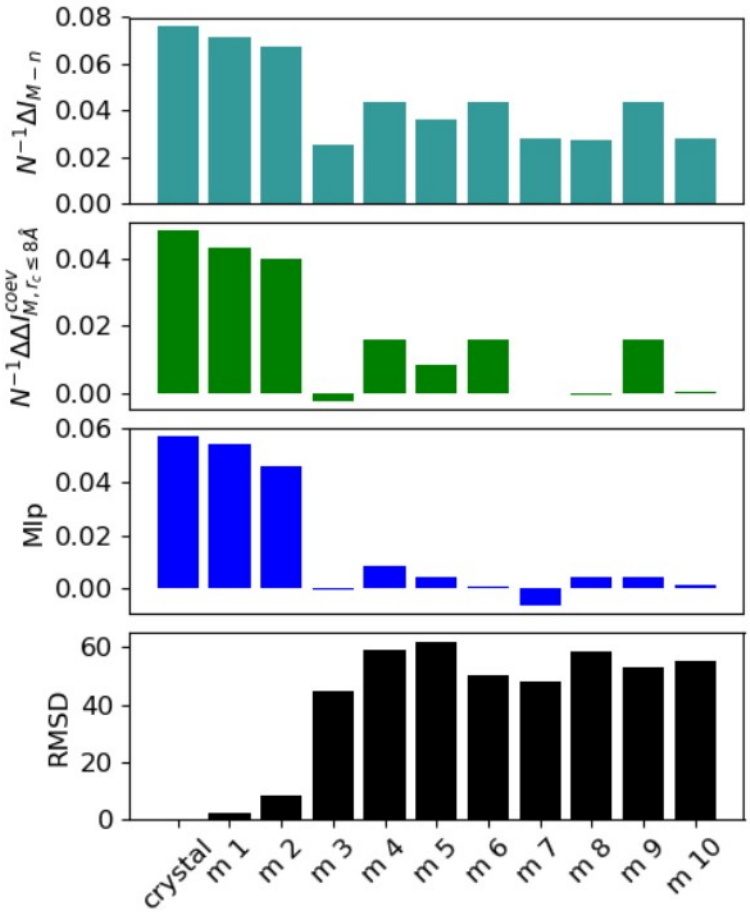


Figura 2.11. Transinformação em função de interfaces alternativas. *Gap de transinformação* $N^{-1}\Delta I_{M-n}$*, transinformação fruto da coevolução* $N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\,\mathring{A}}$*, Mip e RMSD em relação ao complexo para diferentes interfaces alternativas para o sistema 1BXR-AB.*

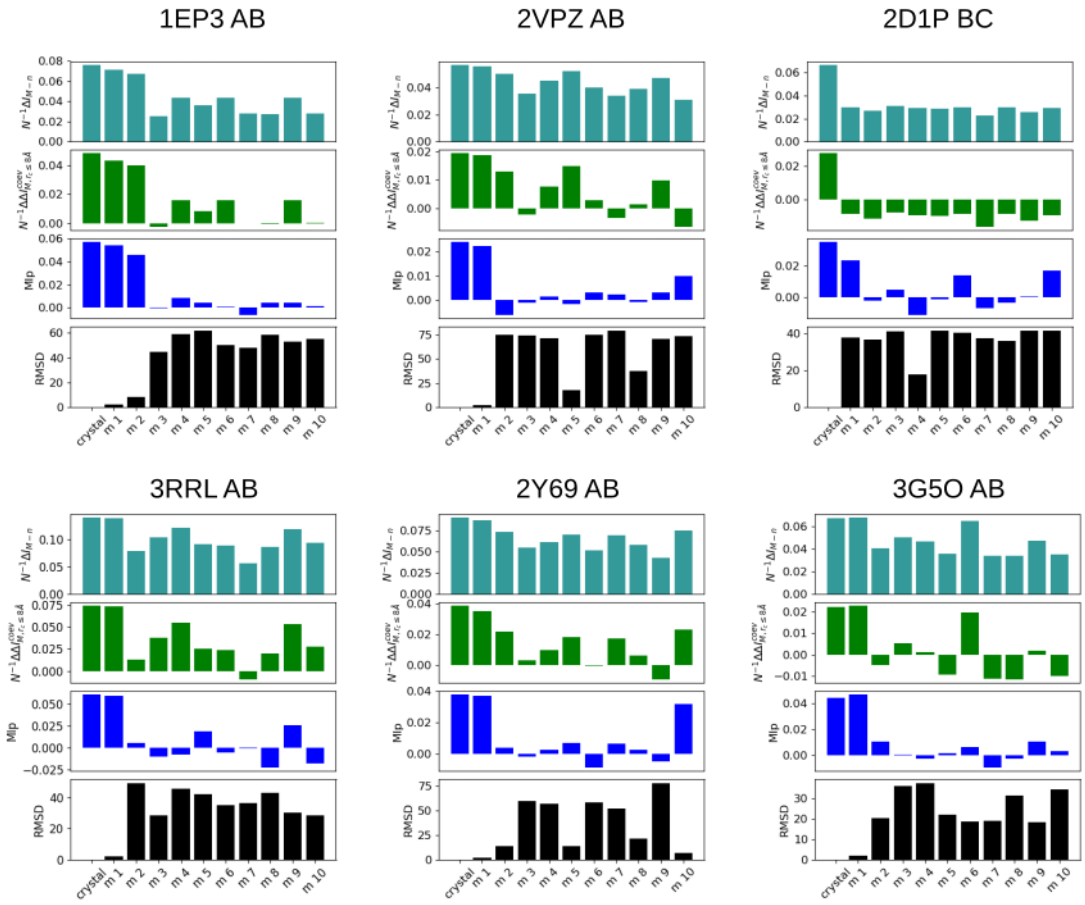


Figura 2.12. Transinformação em função de interfaces alternativas para os demais sistemas. *Gap de transinformação* $N^{-1}\Delta I_{M-n}$, *transinformação fruto da coevolução* $N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\text{Å}}$, *Mip e RMSD em relação ao complexo para diferentes interfaces alternativas para os sistemas 1EP3-AB, 2VPZ-AB, 2D1P-BC, 3RRL-AB, 2Y69-AB e 3G5O-AB.*

5. Conclusão

De modo geral, a coevolução molecular entre duas proteínas que interagem resulta em pares de aminoácidos (a uma curta distância menor ou igual a 8Å) com alto valor de transinformação em interfaces de interações. Esse valor de transinformação se deve, além da coevolução, a fontes estocásticas e filogenéticas. Entretanto, a transinformação entre os aminoácidos fisicamente acoplados é melhor em distinguir arranjos de pareamentos corretos de arranjos randômicos, possuindo o menor erro esperado e uma menor degeneração. Dessa maneira, otimizações que utilizem a

maximização da transinformação dos aminoácidos fisicamente acoplados devem ser a melhor opção para a resolução do problema 4, sendo uma melhor alternativa do que a maximização da transinformação total como poroposto por Tillier et al. (2006).

Os sistemas apresentados neste trabalho, como são estudos de caso em que o pareamento correto deve ser conhecido para as devidas validações, possuem os genes para as proteínas A e B codificadas no mesmo genoma, ou seja, sofreram especiação sempre da mesma forma e tem alta taxa de correlação filogenética (dados não mostrados) e a fonte filogenética é muito presente. Cabe ressaltar aqui que os problemas reais, como interações proteínas virais e proteínas de hospedeiro ou ainda toxinas e canais iônicos, a fonte filogenética não deve existir, uma vez que esses organismos, em geral, não compartilham a história filogenética.

Embora mais de um par de proteínas tenham sido utilizados como estudo de caso, mais sistemas precisam ser analizados para entender melhor se de fato tais conclusões são universais e, se não são, em quais classes de proteínas elas são verdadeiras e quais não são. Para isso a elaboração de uma ferramenta de construção de alinhamentos para esse tipo de estudo se faz necessária.

A observação de que o *gap* de transinformação e a transinformação coevolutiva são altas majoritariamente apenas para soluções de docking com baixo RMSD nos permite propor combinações dessas, com outras métricas, para o reranqueamento de soluções de docking.

Embora o trabalho apresentado neste capítulo não proponha uma metodologia para a resolução do problema 4, o fato de a informação coevolutiva deve ter maior resolução na discriminação de PPI é uma importante contribuição para o campo. Além abordagens de otimização de transinformação e caracterização do espaço de PPI estão sendo feitas pelo grupo de trabalho e alguns *insights* serão apresentados no próximo capítulo.

Capítulo 3.Perspectivas no uso de transinformação para estudo de interações proteína-proteína

1. Introdução

Como vimos até aqui, a coevolução de duas proteínas pode resultar em um sinal coevolutivo que pode ser capturado pela transinformação. Outras métricas se desenvolveram ao longo dos anos e conferiram avanços consideráveis tanto em acurácia como em velocidade computação. Entretanto existem alguns desafios a serem superados no estudo de coevolução proteica.

Tillier et al. (2006) propôs o método chamado Codep para predição dos pares de interação de dois alinhamentos de sequências. Ela utilizou um algorítimo de maximização de transinformação que era capaz de resolver com uma boa acurácia problemas simples envolvendo sequências controladas e com acurácia moderada problemas em sistemas biológicos reais. Tal trabalho, entretanto, foi pouco citado e o método quase nunca efetivamente utilizado. Um dos problemas desse método é a limitação computacional, ele não suporta alinhamentos grandes (tanto em tamanho como em número de sequências).

O Pazos e Alfonso propuseram a utilização do Mirror Tree, método que utiliza correlações entre matrizes de distância, para o pareamento de dois alinhamentos (Pazos and Valencia, 2002). Esse método parece eficiente e resolve bem alguns casos e utiliza uma métrica computacionalmente barata. A correlação entre duas matrizes de distância provenientes de alinhamentos dos homólogos da proteína A e B é calculada por:

$$R = \frac{\sum_{i=1}^{n} (d_i^A - \overline{d^A})(d_i^B - \overline{d^B})}{\sqrt{\sum_{i=1}^{n} (d_i^A - \overline{d^A})^2} \sqrt{\sum_{i=1}^{n} (d_i^B - \overline{d^B})^2}}$$

onde d_i^A e d_i^B são a *i*-ésima entrada da matriz de distância A e B respectivamente e, $\overline{d^A}$ e $\overline{d^A}$ são os valores médios das respectivas matrizes.

Em três trabalhos recentes (Bitbol, 2018; Bitbol et al., 2016; Marmier et al., 2019), o grupo da pesquisadora Anne Bitbol avaliou a usabilidade da *I*, da *DI* e do *R*. para resolver um caso mais simples do problema 4, que consiste na determinação dos pares de interação entre parálogos. O caso dos parálogos é um problema mais simples do que o problema 4 de um modo geral (a **Figura 1.5** do Capítulo 1 ilustra a diferença entre esses problemas) as três métricas testadas apresentaram alta acurácia para a predição correta dos pares de parálogos que interagem.

Podemos perceber aqui, que apesar dos trabalhos citados, incluindo o estudo do caso dos parálogos, o problema 4 geral avançou muito pouco nos últimos anos. E é nesse contexto que revisito esse problema, agora em uma situação onde o número de sequências disponíveis nos bancos de dados é muito maior e nos possibilita tratar com maior estatística esse problema e em uma situação onde isolamos a contribuição de cada uma das fontes de transinformação (Andrade et al., 2019). O objetivo deste capítulo é apresentar os resultados preliminares do estudo da degeneração do espaço de transinformação de arranjos de pareamento de alinhamentos de proteínas que interagem. Embora preliminar, esse estudo revelou que a degeneração desse espaço é mais complexa do que parecia até o trabalho anterior. Antecipando os resultados, existe uma fonte de erro trivial relacionada à similaridade das sequências e essa é facilmente tratada perdoando conexões entre sequências filogeneticamente relacionadas. Entretanto, existe uma segunda fonte de erro, não trivial, que gera arranjos de pareamento com alto valor de transinformação aparentemente sem significado biológico, que são inclusive, piores do que os arranjos aleatórios. Esses resultados sugerem cautela no uso de transinformação para a resolução do problema 4.

2. Teoria, materiais e métodos

2.1. Transinformação

A transinformação entre dois blocos de sequências X^N e X^N foi definida conforme teoria apresentada no capítulo anterior:

$$I\left(X^N;Y^N|z\right)=\sum_i^N I\left(X_i,Y_i|z\right)$$

e foi calculada para os blocos correspondentes aos pares de aminoácidos fisicamente acoplados (cuja distância entre os centros geométricos dos aminoácidos é menor ou igual a 8 Å). Para mais detalhes do cálculo ver o capítulo anterior.

2.2. Algorítimo genético

O número de possibilidades de conexões entre as sequências de dois aminoácidos é *M!*. Isso torna impossível a enumeração completa de todos os arranjos possíveis. Dessa forma, com o objetivo de caracterizar melhor o espaço de transinformação propusemos um algorítimo genético que maximiza uma função objetiva, a transinformação ($I\left(X^N;Y^N|z\right)$) considerando os aminoácidos fisicamente acoplados (ver capítulo anterior para detalhes do cálculo). Neste estudo, O algorítimo genético foi feito baseado na Dissertação de Mestrado (Pontes, 2017). O algorítimo começa com uma população de indivíduos (arranjos *z*) gerados aleatoriamente e que a cada geração sofrem uma determinada fração de mutações (mudanças no arranjos *z*). A cada geração a função objetiva é calculada para os diferentes indivíduos da população e aqueles que possuem valores maiores são selecionados para o próximo evento de mutação. Esse processo segue recursivamente até que a simulação sature, isto é, os valores da função objetiva ficam estáveis ao longo de muitas gerações significando que um máximo foi encontrado.

2.3. Alinhamentos e estruturas

Os alinhamentos utilizados neste trabalho foram retirados de (Ovchinnikov et al., 2014). As estruturas utilizadas para definir os aminoácidos fisicamente acoplados foram retirados do *Protein Data* Bank (PDB). Os sistemas analisados estão apresentados na **Tabela 3.1**.

Tabela 3.1. Relação das estruturas utilizadas e alinhamentos utilizados. *O sistema 1BXR foi destacado em alguns momentos no texto. M corresponde ao número de sequências no alinhamento e o Tamanho apresentado é o número de posições no alinhamento concatenado.*

Descrição do complexo	PDB ID	Proteína A	Proteína B	M	Tamanho
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
Structure of the thiazole synthase/ThiS complex	1TYG	**Chain B:** yjbS	**Chain A:** Thiazole biosynthesis protein thiG	746	307
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

2.4. Matrizes de distância de hamming

As matrizes de distância de hamming são matrizes que levam em conta apenas a diferença entre os aminoácidos nos alinhamentos, por exemplo, suponhamos que um alinhamento tenha 100 posições, e uma sequência *i* possui 10 aminoácidos diferentes da sequência *j* e 90 posições iguais, isso significa que a distância entre essas duas sequências é 0,1. Essa é a forma mais simples de medir distância entre sequências e não leva em conta as diferentes probabilidades de substituição de aminoácidos. Dessa maneira, é possível calcular uma matriz de distâncias entre todas contra todas

sequências de um alinhamento, gerando uma matriz (*M* x *M*) onde *M* é o número de sequências do alinhamento.

2.5. Correlação entre matrizes de alinhamentos (*R*)

Com o objetivo de encontrar um segundo descritor que pudesse ajudar a caracterizar os arranjos de pareamento z, calculei a correlação entre matrizes de distância, para o pareamento de dois alinhamentos (Pazos and Valencia, 2002). A correlação entre duas matrizes de distância (de hamming, ver tópico anterior) provenientes de alinhamentos dos homólogos da proteína A e B é calculada por:

$$R = \frac{\sum_{i=1}^{n} (d_i^A - \overline{d^A})(d_i^B - \overline{d^B})}{\sqrt{\sum_{i=1}^{n} (d_i^A - \overline{d^A})^2} \sqrt{\sum_{i=1}^{n} (d_i^B - \overline{d^B})^2}}$$

onde d_i^A e d_i^B são a *i*-ésima entrada da matriz de distância A e B respectivamente e, $\overline{d^A}$ e $\overline{d^A}$ são os valores médios das respectivas matrizes.

3. Resultados e discussão

3.1. Os valores de transinformação ao longo das gerações

Aqui apresento os valores de transinformação ao longo das gerações do algorítimo genético para o sistema 1BXR até a geração 20 mil (esses dados ainda são preliminares e as simulações estão sendo estendidas até 100 mil gerações para verificação de convergência, ou seja, essas otimizações aina não convergiram mas revelam aspectos interessantes das otimizações). A **Figura 3.1** apresenta os resultados de três réplicas para o sistema 1BXR e revela um crescimento rápido no início que vai diminuindo no final. Essa diminuição na taxa de crescimento deve cair conforme o número de gerações aumenta e o valor de transinformação se aproxima do valor do arranjo *z* nativo.

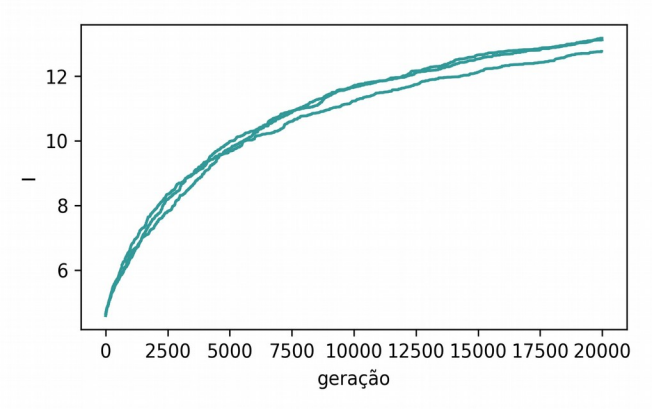


Figura 3.1. Transinformação em função do tempo (gerações). *As linhas turquesas indicam os valores para três réplicas de otimizações da transinformação para o sistema 1BXR.*

3.2. Tratando a fonte de erro trivial devido à similaridade das sequências

Como visto no capítulo anterior, existe uma degeneração nos valores de transinformação, isto é, mais de um arranjo pode gerar o memso valor de transinformação. A degeneração na vizinhança do arranjo correto (z^*) é menor para os aminoácidos fisicamente acoplados mas ainda é muito grande. Isso significa que uma otimização de transinformação deve chegar a valores muito próximos ao valor do arranjo z^* mesmo sem ter conexões par a par exatamente iguais a z^*, entretanto, isso não significa necessariamente que o pareamento não tenha significado biológico. Por exemplo, arranjos que conectam determinada sequência A a uma sequência B errada, porem muito parecida com a sequência B correta, tem alto valor de transinformação e fazem mais sentido biológico do que a conexão aleatória que liga uma determinada sequência A a uma sequência B distante da correta. É como se conectássemos sequências de um chimpanzé com as de humanos. Esses arranjos embora não sejam corretos quando analisados par a par, fazem mais sentido biológico do que os arranjos aleatórios.

Para tratar o senário descrito no parágrafo acima elaboramos uma medida de acurácia que “perdoa” as conexões próximas. Para tanto, analisamos o histograma da matriz de distâncias do alinhamento B (ver metodologia) e traçamos uma linha de corte no valor de distância no ponto em que a soma do número de entradas da matriz atinge valor igual a 10% do número de entradas da matriz, começando da menor distância.

Dessa maneira, nosso objetivo é perdoar qualquer pareamento de uma sequência A com uma sequência B errada, desde que essa sequência tenha uma distância menor do que esse valor de corte da sequência B correta. Em outras palavras, se a sequência B pareada erroneamente possuir distância menor ou igual a esse valor de corte esse pareamento é considerado certo, e se a distância for maior o pareamento é considerado errado.

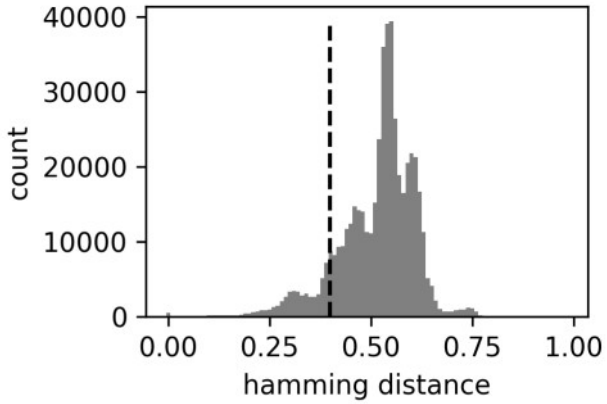


Figura 3.2. Histograma de distâncias de hamming do sistema 1BXR. *À esquerda da linha tracejada estão as menores distâncias (10% do total).*

3.3. O algorítimo genético melhora o pareamento do sistema 1BXR

A análise da acurácia das três réplicas do sistema 1BXR revelou que a otimização do algorítimo genético que maximiza a transinformação conforme apresentado na metodologia melhora a acurácia de pareamento. A **Figura 3.3** apresenta a acurácia relativa para o sistema 1BXR em função do valor de corte discutido no tópico anterior. É importante notar que quando o valor de corte é maior do que 10% muitas sequências são perdoadas e o valor da acurácia relativa sobe inclusive para arranjos aleatórios (cinco arranjos aleatórios são plotados na **Figura 3.3** em vermelho). Quando o valor de corte é mito menor, por exemplo zero, nada é perdoado e mesmo as otimizações do algorítimo genético possuem baixa acurácia. Podemos observar aqui a fonte de erro esperada apresentada no tópico anterior que é esperada devido à similaridade de algumas sequências.

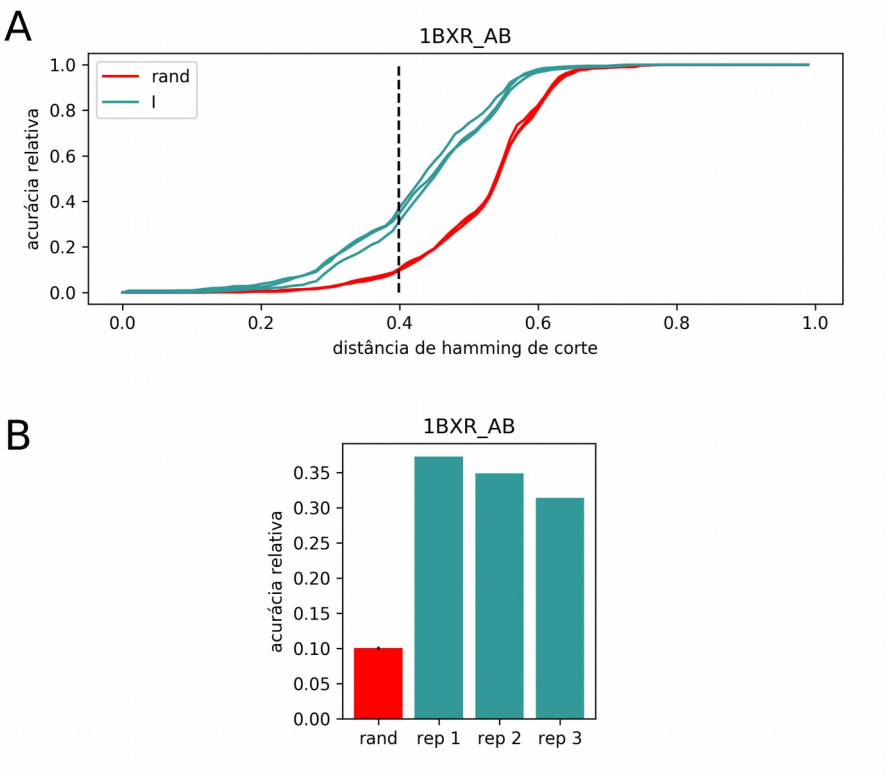


Figura 3.3. Acurácia em função do valor de distância de hamming de "perdão" para o sistema 1BXR-AB*. Em A a linha tracejada indica o valor de corte da acurácia em 10%. Em vermelho temos 5 arranjos randômicos, em turquesa, roxo e amarelo temos o arranjo da geração 20.000 do algorítimo genético que maximiza I, R do sublinhamento da interface e R do alinhamento completo, respectivamente. Em B, são plotados os valores de acurácia relativa ao corte da linha tracejada da figura A.*

3.4. O algorítimo genético revela uma região com erro não trivial para alguns sistemas

Podemos observar que a escolha de 10% como valor de corte faz muito sentido biológico e por isso foi o valor escolhido aqui para ser utilizado para a medida de acurácia, uma vez que trata a fonte de erro trivial resultado da similaridade entre algumas sequências. Os valores de acurácia relativa para cada uma das réplicas de cada um dos sistemas analisados estão apresentados na **Figura 3.4.**

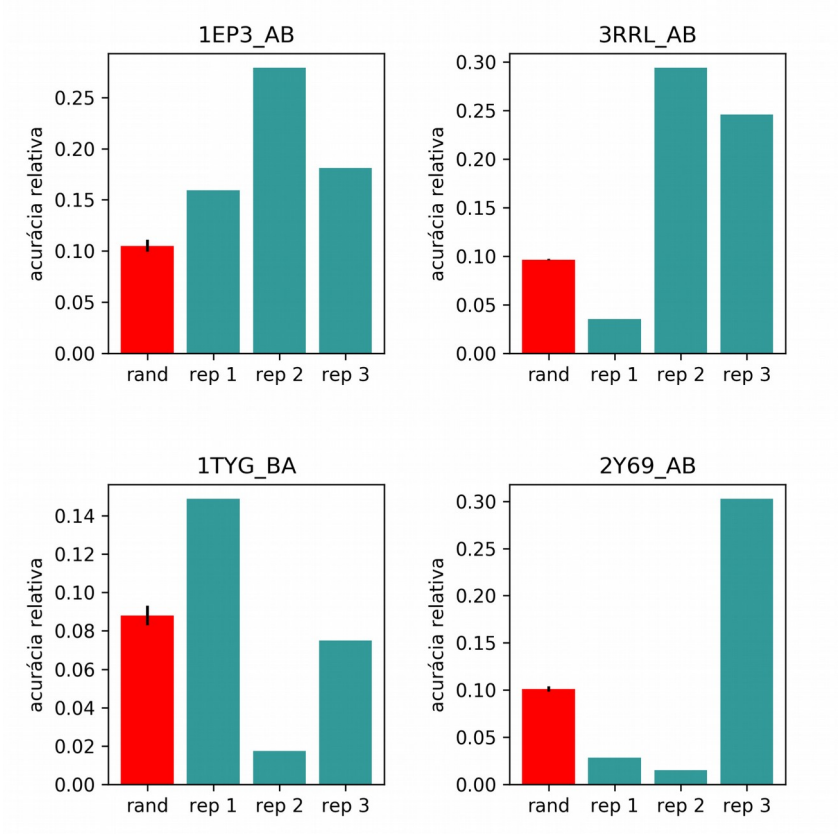


Figura 3.4. Acurácia em função do valor de distância de hamming de "perdão" para os sistemas 1EP3, 3RRL, 1TYG e 2Y69. *As barras vermelhas indicam a acurácia relativa média de 5 arranjos z randômicos e as barras turquesa indicam a acurácia de três réplicas da otimização de transinformação no algorítimo genético (20 mil gerações).*

A análise desses resultados revelou a existência de arranjos z que possuem alto valor de transiformação e baixo significado biológico (baixa acurácia), conforme pode ser observado para a réplica 1 do sistema 3RRL, para a réplica 2 do sistema 1TYG e para as réplicas 1 e 2 do sistema 2Y69 na **Figura 3.4**.

Esses resultados acendem um alerta no uso da transinformação como função objetiva para algorítimos que busquem resolver o problema 4. Ou seja, é possível achar arranjos com alto valor de transinformação e nenhum significado biológico. Embora essa métrica tenha sido utilizada para a resolução do problema dos parálogos (Bitbol, 2018), um subproblema do problema 4, esses resultados revelaram que, até o momento, ela não é confiável para a resolução do problema mais geral.

4. Perspectivas

Conforme apresentado até então, a otimização da transinformação por si só não é capaz de garantir a resolução do problema de pareamento de sequências de homólogos que interagem (problema 4). Dessa forma testamos algumas outras métricas que poderiam ser utilizadas como um segundo parâmetro de validação e/ou otimização. Conforme apresentado na introdução deste capítulo (Pazos and Valencia, 2002) utilizaram um algorítimo de maximização de *R* com o objetivo de resolver o problema 4. A combinação de *R*, bem como a *Direct Information – DI* como funções auxiliares para a maximização da transinformação podem ser uma possível alternativa. Essas métricas estão sendo testadas e os resultados ainda não são conclusivos e por isso apresento a utilização dessas métricas em conjunto com a transinformação como perspectiva futura.

Vale ressaltar porém que, embora a combinação da transinformação com *R* possa ser uma boa escolha para os sistemas testes, ela pode não funcionar para problemas reais, nos quais os alinhamentos devem ter baixa correlação *R*, uma vez que as proteínas *A* e *B* estão em organismos diferentes e em geral não compartilham a história filogenética.

Um resultado muito interessante que obtive é que a DI por si só, utilizada como função objetiva do algorítimo genético, não é capaz de melhorar a acurácia dos pareamentos *z*, o que significa que embora ela seja uma boa métrica para o problema dos parálogos (Bitbol et al., 2016) ela parece não funcionar bem para o problema mais geral (em nenhum dos sistemas testados – dados não mostrados). Isso se deve ao fato de que no problema apresentado aqui os algorítimos partem de arranjos completamente aleatórios, o que faz com que o sinal de DI não tenha nenhum significado, isto é, o calculo por aproximação [*mean-field aproximation of DCA* - Weigt et al. (2009)] não faz nenhum sentido em redes aleatórias e a maximização da DI gera arranjos que são tão bons quanto arranjos randômicos. Estes resultados, embora preliminares são importantes e a conclusão desses experimentos também são perspectivas do trabalho aqui apresentado.

Referências

Altschuh, D., Lesk, A.M., Bloomer, A.C., and Klug, A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J. Mol. Biol. *193*, 693–707.

Altschuh, D., Vernet, T., Berti, P., Moras, D., and Nagai, K. (1988). Coordinated amino acid changes in homologous protein families. Protein Eng. Des. Sel. *2*, 193–199.

Andrade, M., Pontes, C., and Treptow, W. (2019). Coevolutive, evolutive and stochastic information in protein-protein interactions. Comput. Struct. Biotechnol. J. *17*, 1429–1435.

Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W., and Dress, A.W. (2000). Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis. Mol. Biol. Evol. *17*, 164–178.

Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.-I., and Langmead, C.J. (2011). Learning generative models for protein fold families. Proteins Struct. Funct. Bioinforma. *79*, 1061–1078.

Bitbol, A.-F. (2018). Inferring interaction partners from protein sequences using mutual information. PLoS Comput. Biol. *14*, e1006401.

Bitbol, A.-F., Dwyer, R.S., Colwell, L.J., and Wingreen, N.S. (2016). Inferring interaction partners from protein sequences. Proc. Natl. Acad. Sci. U. S. A. *113*, 12180–12185.

Bonetta, L. (2010). Interactome under construction. Nature *468*, 851–852.

Burger, L., and van Nimwegen, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput. Biol. *6*, e1000633.

CARROLL, L. (1960). The Annotated Alice: Alice’s Adventures in Wonderland and Through the Looking-Glass (New York).

Codoñer, F.M., and Fares, M.A. (2008). Why should we care about molecular coevolution? Evol. Bioinforma. Online *4*, 29–38.

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

Davis, F.P., Barkan, D.T., Eswar, N., McKerrow, J.H., and Sali, A. (2007). Host–pathogen protein interactions predicted by comparative modeling. Protein Sci. *16*, 2585–2596.

Dobzhansky, T. (1950). Genetics of Natural Populations. Xix. Origin of Heterosis through Natural Selection in Populations of Drosophila Pseudoobscura. Genetics *35*, 288–302.

Dunn, S.D., Wahl, L.M., and Gloor, G.B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics *24*, 333–340.

Echave, J., Spielman, S.J., and Wilke, C.O. (2016). Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. *17*, 109–121.

Ehrlich, P.R., and Raven, P.H. (1964). Butterflies and Plants: A Study in Coevolution. Evolution *18*, 586–608.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Phys. Rev. E *87*, 012707.

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

Goh, C.-S., Bogan, A.A., Joachimiak, M., Walther, D., and Cohen, F.E. (2000). Co-evolution of proteins with their interaction partners11Edited by B. Honig. J. Mol. Biol. *299*, 283–293.

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

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

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. U. S. A. *110*, 15674–15679.

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. Nature *217*, 624–626.

Kimura, M., and Ohta, T. (1973). Mutation and evolution at the molecular level. Genetics *73*, Suppl 73:19-35.

Korber, B.T., Farber, R.M., Wolpert, D.H., and Lapedes, A.S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc. Natl. Acad. Sci. *90*, 7176–7180.

Krishnadev, O., and Srinivasan, N. (2011). Prediction of protein–protein interactions between human host and a pathogen and its application to three pathogenic bacteria. Int. J. Biol. Macromol. *48*, 613–619.

Lewis, A.C.F., Saeed, R., and Deane, C.M. (2010). Predicting protein-protein interactions in the context of protein evolution. Mol. Biosyst. *6*, 55–64.

Marmier, G., Weigt, M., and Bitbol, A.-F. (2019). Phylogenetic correlations can suffice to infer protein partners from sequences. PLOS Comput. Biol. *15*, e1007179.

Martin, L.C., Gloor, G.B., Dunn, S.D., and Wahl, L.M. (2005). Using information theory to search for co-evolving residues in proteins. Bioinformatics *21*, 4116–4124.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. *108*, E1293–E1301.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci. *91*, 98–102.

Netto, L.E.S., and Menck, C.F.M. (2012). Capítulo 5. Estabilidade do material genético: mutagênese e reparo. In Biologia Molecular e Evolução, (Ribeirão Preto: Sociedade Brasileira de Genética), p.

Olmea, O., and Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold. Des. *2*, S25–S32.

Othersen, O.G., Stefani, A.G., Huber, J.B., and Sticht, H. (2011). Application of information theory to feature selection in protein docking. J. Mol. Model. *18*, 1285–1297.

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

Pazos, F., and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng. Des. Sel. *14*, 609–614.

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

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction 11Edited by A. R. Fersht. J. Mol. Biol. *271*, 511–523.

Pontes, C. (2017). Coevolução molecular em canais iônicos e neurotoxinas. Universidade de Brasília.

Sander, C., and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins Struct. Funct. Bioinforma. *9*, 56–68.

Shackelford, G., and Karplus, K. (2007). Contact prediction using mutual information and neural nets. Proteins *69 Suppl 8*, 159–164.

Shannon, C. (1948). A Mathematical Theory of Communication.

Shindyalov, I.N., Kolchanov, N.A., and Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng. Des. Sel. *7*, 349–358.

Süel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat. Struct. Biol. *10*, 59–69.

Taylor, W.R., and Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. Protein Eng. Des. Sel. *7*, 341–348.

Tillier, E.R.M., Biro, L., Li, G., and Tillo, D. (2006). Codep: Maximizing co-evolutionary interdependencies to discover interacting proteins. Proteins Struct. Funct. Bioinforma. *63*, 822–831.

Tovchigrechko, A., and Vakser, I.A. (2006). GRAMM-X public web server for protein–protein docking. Nucleic Acids Res. *34*, W310–W314.

Wallace, B. (1953). On Coadaptation in Drosophila. Am. Nat. *87*, 343–358.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. Proc. Natl. Acad. Sci. *106*, 67–72.

William Klug (2010). Capítulo 27. Genética de Populações. In Conceitos de Genética, (Porto Alegre: Artmed), pp. 710–736.

Xie, Z., Deng, X., and Shu, K. (2020). Prediction of Protein–Protein Interaction Sites Using Convolutional Neural Network and Improved Data Sets. Int. J. Mol. Sci. *21*, 467.

Anexo 1 – Artigo

Coevolutive, evolutive and stochastic information in protein-protein interactions

MiguelAndrade, CamilaPontes e WernerTreptow

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

COMPUTATIONAL AND STRUCTURAL 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

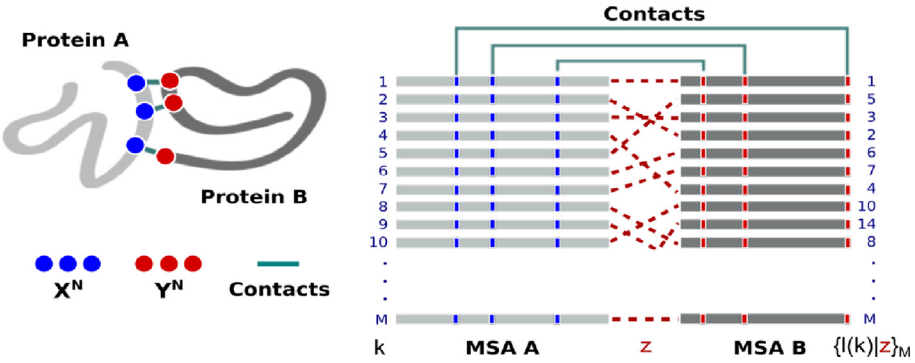
While being selected to be thermodynamically stable and kinetically accessible in a particular fold [1,2], interacting proteins *A* and *B* coevolve to maintain their bound free-energy stability against a vast repertoire of non-specific partners and interaction modes. Protein coevolution, in the form 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



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 $i = 1,\ldots,N$ amino-acid contacts at the molecular level. Proteins *A* and *B* are assumed to coevolve throughout *M!* distinct processes *z* described by the stochastic variable *Z* with an uniform probability mass function $\rho(z)$, $\forall z \in \{1,\ldots,M!\}$. Given any specific process *z*, their interacting amino-acid sequences are respectively described by two *N*-length blocks of discrete stochastic variables $X^N \equiv (X_1,\ldots,X_N)$ 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,

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

and

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

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

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

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*

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

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

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

such that,

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

(*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)

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

from the observed frequencies $\mathbf{f}=\{f_{x_i,y_i|z}\}$ 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 $\rho^*(x_i,y_i|z) \equiv f_{x_i y_i|z}$ were defined from the observable frequencies $f_{x_i y_i|z}$ 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

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

where, $n_z^m = |\{m'|1 \le m' \le M, \textit{Hamming Disatnce(m,m')} \ge \delta h\}|$ 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 $\delta_{x_i^m y_i^m|z,x_i y_i|z}$ 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 $\frac{1}{|\chi|^2}$ 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 $\langle I(X^N; Y^N|z)\rangle_{M-n}$ 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

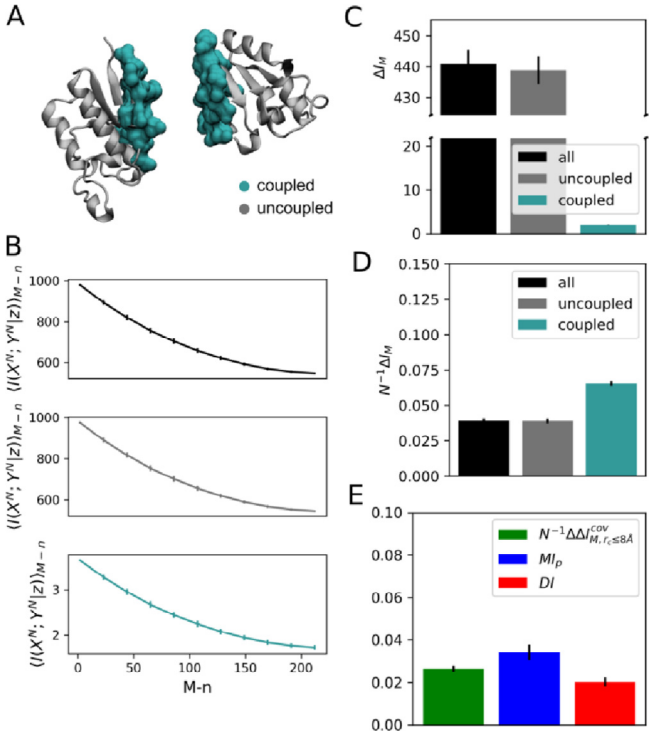


Fig. 1. Informational analysis of protein complex TusBCD, chains *B* and *C*. (A) Three-dimensional representation of stochastic variables X^N and Y^N as defined from physically coupled amino acids at short-range cutoff distances $r_c \leq 8.0$ Å (turquoise) and physically uncoupled amino-acids at long-range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

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. 1C shows the individual contributions of physically coupled and uncoupled amino acids to the total mutual information gap, $\Delta I_M = \Delta I_{M,rc\leq 8.0Å} + \Delta I_{M,rc>8.0Å}$. 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,

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

where, $N^{-1}\Delta_{IM,r_c>8Å}$ 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 ($MI_{p,r_c\leq 8Å}$) [29] and direct information ($DI_{r_c\leq 8Å}$) [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 $\{I(k)|z^*\}_M$ 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

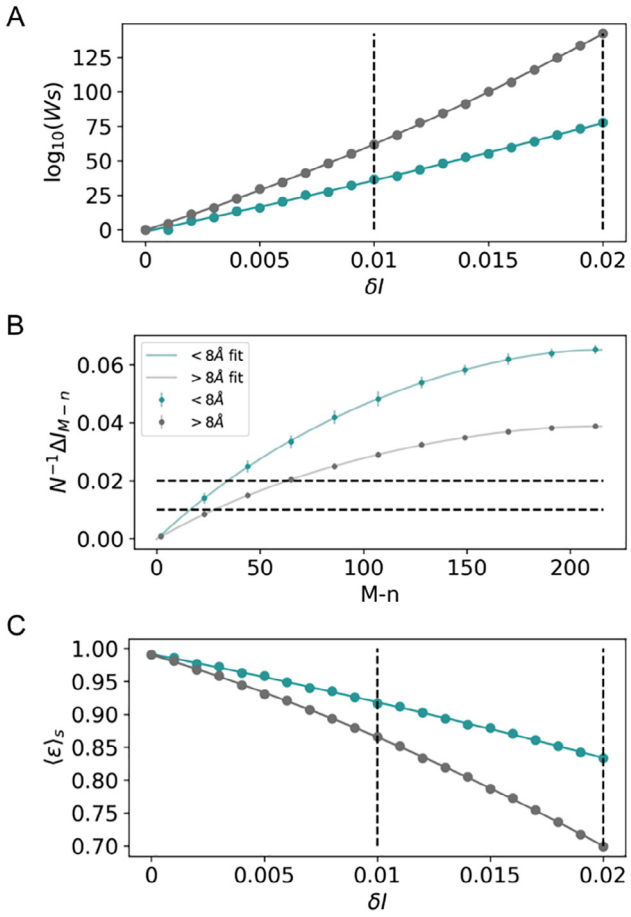


Fig. 2. Degeneracy and error analysis for stochastic variables X^N and Y^N involving interacting amino acids at short-range distances $r_c \le 8.0$ Å (turquoise) and long-range distances $r_c > 8.0$ Å (gray). (A) Total number ω_S of native-like MSA models at various mutual-information resolutions δI. (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.

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

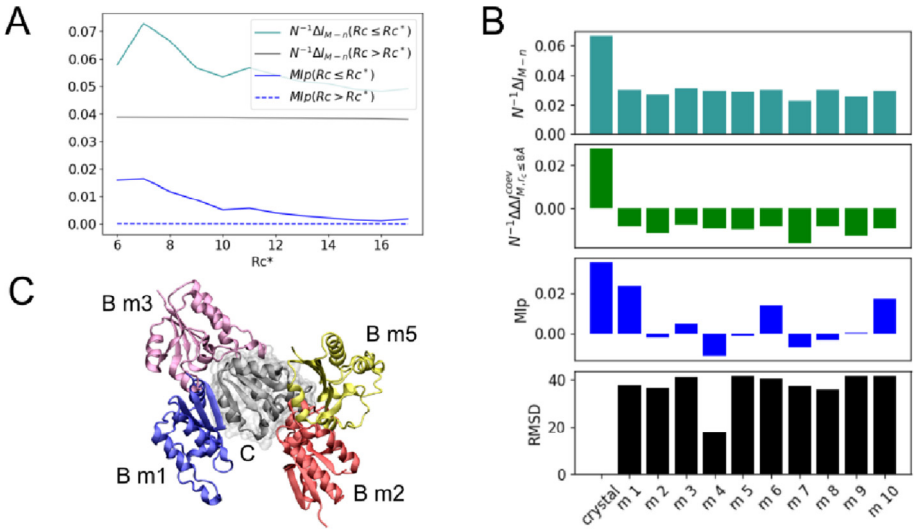


Fig. 3. Dependence with contact definition r_c^* and docking decoys. (A) Per-contact mutual information gap $N^{-1}\Delta I_{M,r_c}$ and mutual information subtracted from structural-functional relationships $MI_{p,rc}$ at various r_c^*. (B) Per-contact mutual information gap $N^{-1}\Delta I_{M,r_c}$ (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 ($r_c^* \stackrel{def}{\equiv} 8.0\ \AA$). 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.

Anexo 2 – Material suplementar do artigo

Coevolutive, evolutive and stochastic information in protein-protein interactions

MiguelAndrade, CamilaPontes e WernerTreptow

Supporting Information

Coevolutive, Evolutive and Stochastic Information in Protein-Protein Interactions

Miguel Andrade, Camila Pontes and Werner Treptow
Laboratório de Biologia Teórica e Computacional (LBTC), Universidade de Brasília DF, Brasil

Derivation of Main Text Equations

Consider two proteins A and B that interact via formation of $i=1,\ldots,N$ independent amino-acid contacts at the molecular level. Proteins A and B are assumed to evolve throughout M distinct coevolution processes z described by the stochastic variable Z with probability mass function $\rho(z)$, $\forall z\in\{1,\ldots,M\}$. Given any specific process z, their interacting amino-acid sequences are respectively described by two N-length blocks of discrete stochastic variables $(X_1,\ldots,X_N)$ and $(Y_1,\ldots,Y_N)$ with probability mass functions $\{\rho(x_1,\ldots,x_N),\rho(y_1,\ldots,y_N),\rho(x_1,\ldots,x_N,y_1,\ldots,y_N|z)\}$ such that

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

and

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

over every joint sequence $\{x_1,\ldots,x_N,y_1,\ldots,y_N\}_{|\chi|^{2N}}$ defined in the alphabet χ of size $|\chi|$.

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

Given its entropy-maximization property [1], $\rho^*(x_1,\ldots,x_N,y_1,\ldots,y_N|z)$ factorizes into the conditional joint distributions of individual contacts i

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

such that

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

are marginals for any specific N-block sequence of proteins A and B. Eq. [S3] ensures the conditional joint entropy to be written extensively in terms of entropic contributions of contact i

$$
\begin{aligned}
H(X_1,\ldots,X_N,Y_1,\ldots,Y_N|z) &= -\sum_{x_1',\ldots,x_N',y_1',\ldots,y_N'} \rho(x_1',\ldots,x_N',y_1',\ldots,y_N'|z)\ln\rho(x_1',\ldots,x_N',y_1',\ldots,y_N'|z) \\
&= -\sum_{x_1',y_1'} \rho^*(x_1',y_1'|z)\ln\rho^*(x_1',y_1'|z)\times[\overbrace{\sum_{x_2',\ldots,x_N',y_2',\ldots,y_N'} \rho^*(x_2',\ldots,x_N',y_2',\ldots,y_N'|z)}^{=1}]\ldots \\
&\quad -[\overbrace{\sum_{x_1',\ldots,x_{N-1}',y_1',\ldots,y_{N-1}'} \rho^*(x_1',\ldots,x_{N-1}',y_1',\ldots,y_{N-1}'|z)}^{=1}]\times\sum_{x_N',y_N'} \rho^*(x_N',y_N'|z)\ln\rho^*(x_N',y_N'|z) \\
&= \sum_i -\sum_{x_i',y_i'} \rho^*(x_i',y_i'|z)\ln\rho^*(x_i',y_i'|z) \\
&= \sum_i H(X_i,Y_i|z)
\end{aligned}
\tag{S5}
$$

given that

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

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

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

where

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

are normalized probabilities of $(N-1)$ -block sequences.

Throughout any specific coevolution process z , the amount of information that protein A stores about the interacting amino-acids of protein B is given by the conditional mutual information $I(X_1,\ldots,X_N;Y_1,\ldots,Y_N|z)$ between the stochastic variables $(X_1,\ldots,X_N)$ and $(Y_1,\ldots,Y_N)$.

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

$$
I(X_1,\ldots,X_N;Y_1,\ldots,Y_N|Z) = \sum_z \rho(z')I(X_1,\ldots,X_N;Y_1,\ldots,Y_N|z')
\tag{S9}
$$

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

$$
I(X_1,\ldots,X_N;Y_1,\ldots,Y_N|Z) = I(X_1,\ldots,X_N;Y_1,\ldots,Y_N) + I(X_1,\ldots,X_N,Y_1,\ldots,Y_N|Z) - I(X_1,\ldots,X_N|Z) - I(Y_1,\ldots,Y_N|Z)
\tag{S10}
$$

in terms of the information entropies

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

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

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

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

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

thus simplifying eq. [S10]

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

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

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

implying

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

as a direct consequence of eq. [S9].

REFERENCES

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

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

SUPPLEMENTARY FIGURES AND TABLES

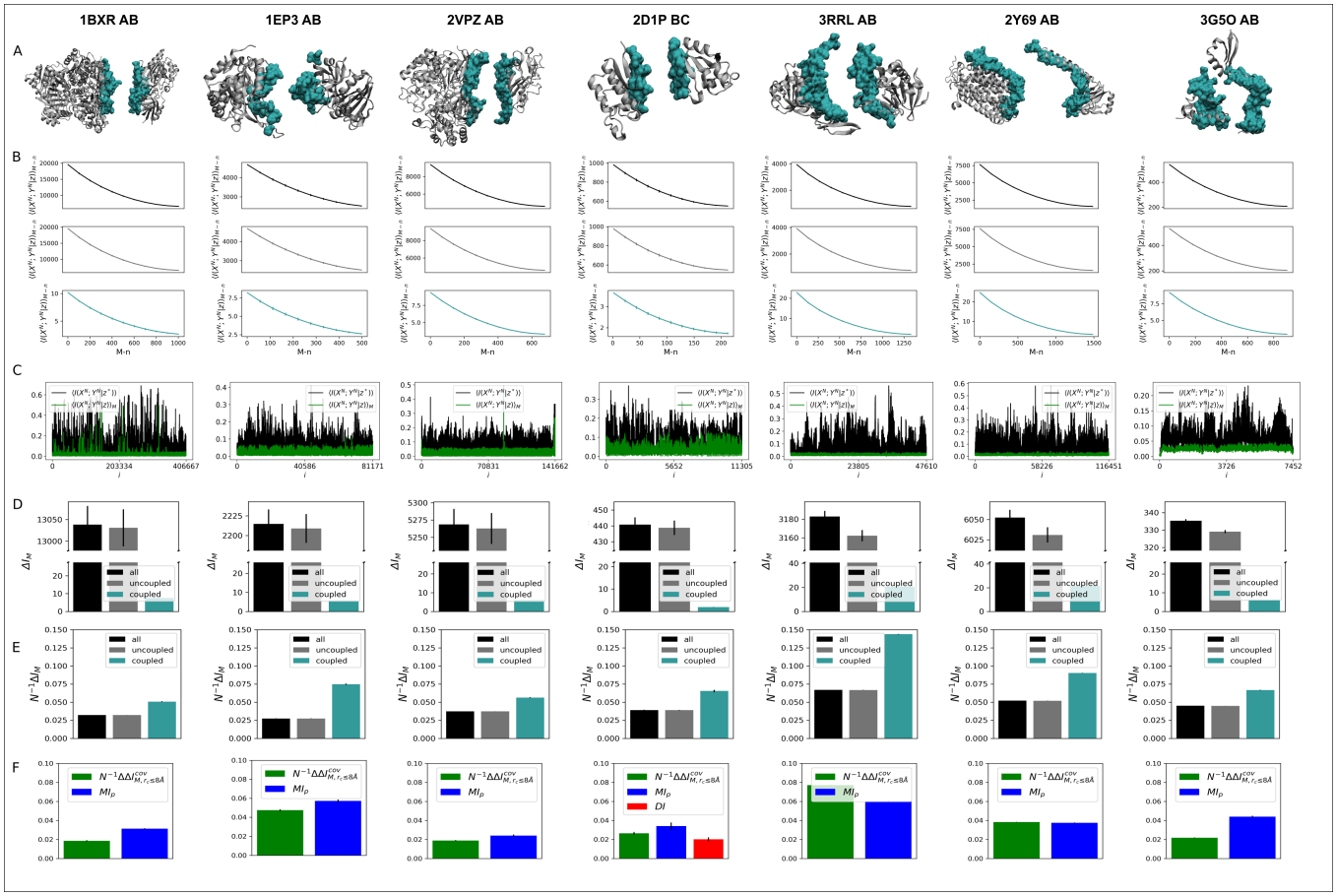


Fig. S1. Informational analysis of protein-protein complexes used in Baker and coworkers.[2] All protein complexes but 2G5O are obligate dimers. (A) Three-dimensional representation of stochastic variables X^N and Y^N as defined from physically coupled amino acids at short-range cutoff distances $r_c \leq 8.0\,\mathring{A}$ (turquoise) and physically uncoupled amino-acids at long-range cutoff distances $r_c > 8.0\,\mathring{A}$ (gray). (B) Conditional mutual information $\langle I(X^N;Y^N|z)\rangle_{M-n}$ as a function of the number $M-n$ of randomly paired proteins in the reference MSA. $\langle I(X^N;Y^N|z)\rangle_{M-n}$ are expectation values estimated from a generated ensemble of ~ 100 MSA models. (C) Conditional mutual information as a function of protein contact i. Mutual information $I(X_i;Y_i|z^*)$ for the reference alignment (black) is systematically larger than $\langle I(X_i;Y_i|z)\rangle_M$ for scrambled models (green) along every contact i. (D) Mutual information gap ΔI_M between reference and 100 random models featuring M randomly paired sequences. (E) Per-contact mutual information gap $N^{-1}\Delta I_{M,r_c}$. (F) Mutual information decomposition ($N^{-1}\Delta\Delta I^{Cov}_{M,r_c\leq 8\mathring{A}}$) and comparison with functional mutual information ($MI_{p,r_c\leq 8\mathring{A}}$) and direct information ($DI_{r_c\leq 8\mathring{A}}$). In C, D, E and F error bars correspond to standard deviations.

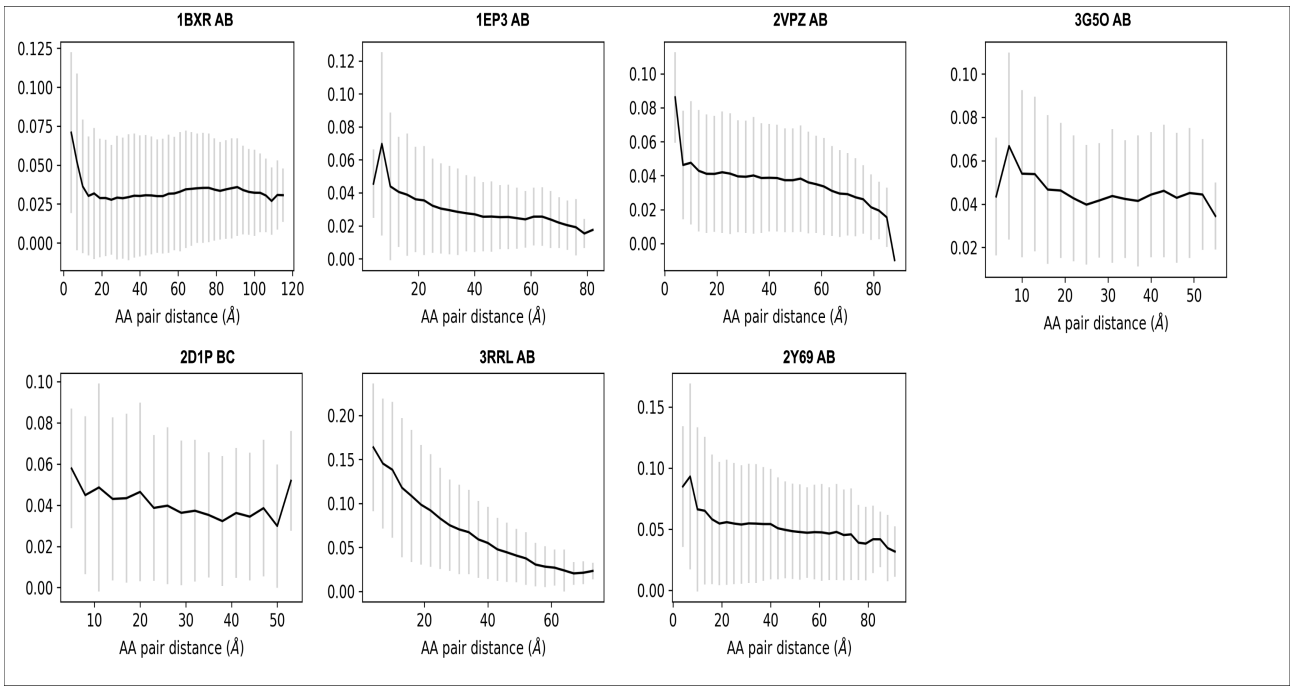


Fig. S2. Information gap ΔI_M profile as a function of amino-acid (AA) pair distances. Shown are average values and the associated standard deviations (error bars) of ΔI_M at various pair distances. The profile shows few larger values of ΔI_M at short distances in contrast to many smaller ones at long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

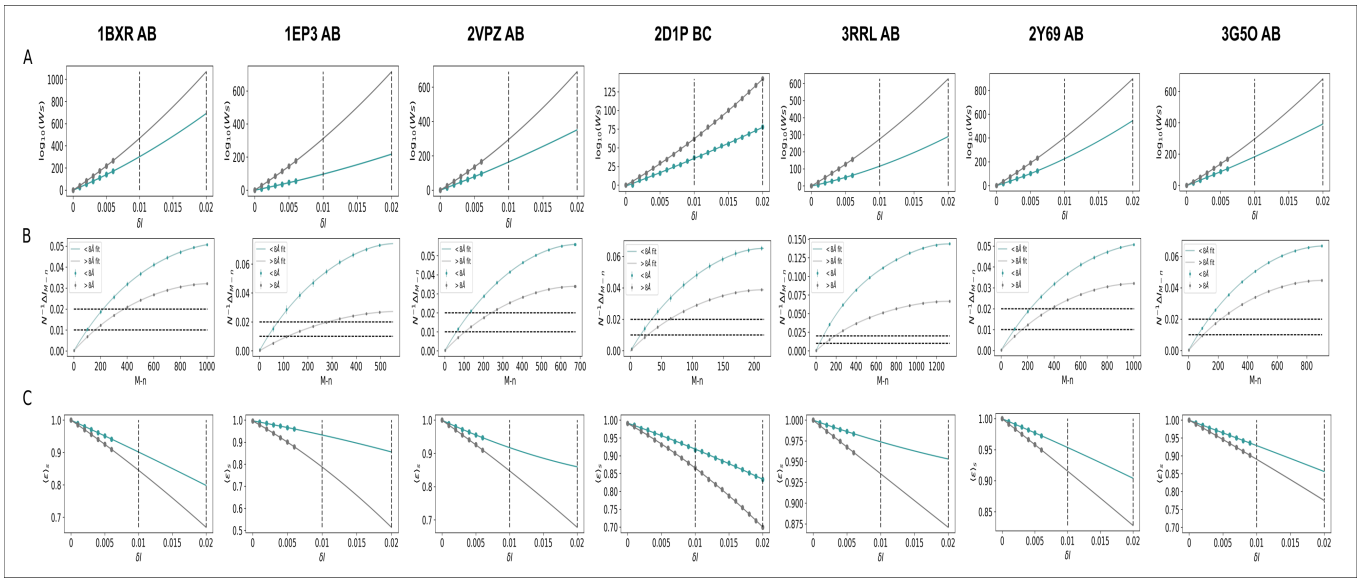


Fig. S3. Degeneracy and error analysis for X^N and Y^N involving interacting amino acids at short-range distances $r_c \leq 8.0\,\mathring{A}$ (blue), long-range distances $r_c > 8.0\,\mathring{A}$ (red), or both (green). (A) Total number ω_S of native-like models at various resolutions δI. (B) Per-contact gaps of mutual information $N^{-1}\Delta I_{M-n,r_c}$ as a function of the number $M-n$ of randomly paired sequences in the reference alignment. Error bars correspond to standard deviations. (C) Expectation values $\langle \epsilon \rangle_S$ for the fraction of sequence matches at various resolutions δI.

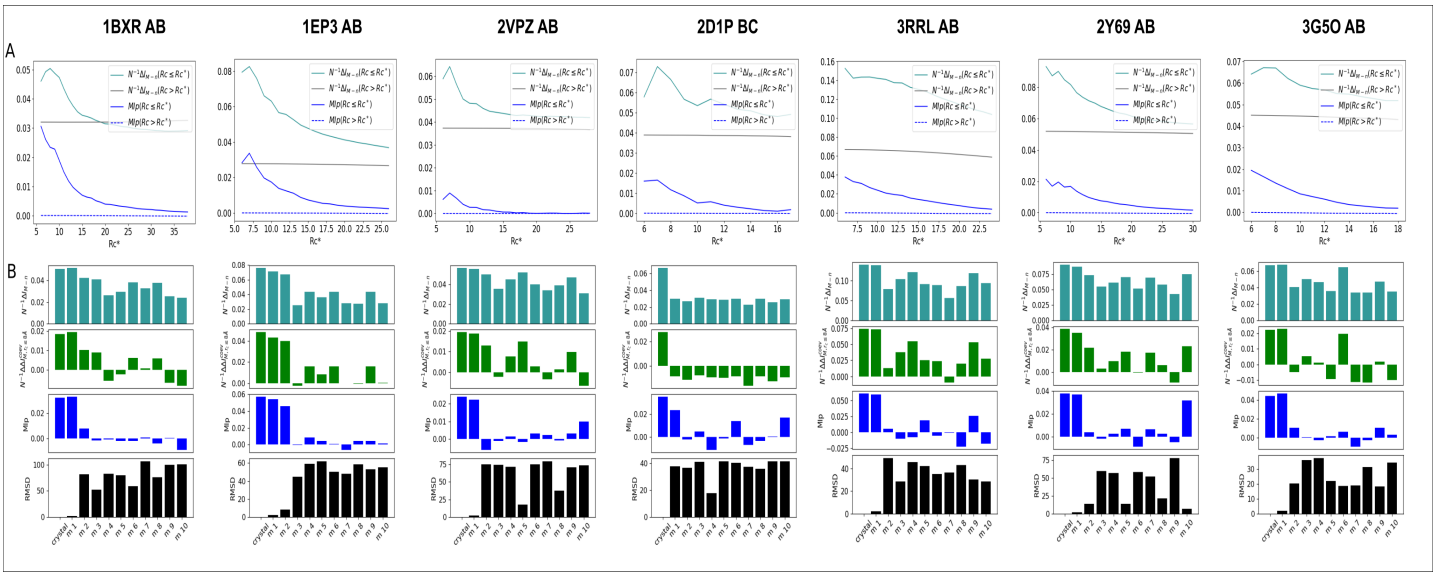


Fig. S4. Dependence with contact definition r_c^* and docking decoys. (A) $N^{-1}\Delta I_{M,r_c}$ and MI_{p,r_c} at various r_c^*. (B) $N^{-1}\Delta I_{M,r_c}$ (turquoise), $N^{-1}\Delta\Delta I_{M,r_c}^{Cov}$ (green), MI_{p,r_c} (blue) at alternative interfaces generated by docking – only physically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

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

M-n	1BXR AB	1EP3 AB	2VPZ AB	2D1P BC	3RRL AB	2Y69 AB
0	1	1	1	1	1	1
1	0	0	0	0	0	0
2	503506	152076	228150	23220	883785	1100386
3	336342008	55761200	102515400	3312720	782444320	1087181368
4	378763143759	34439511150	77616972225	793810530	1168091564220	1811380056759
5	370346185008800	18453455396640	50999525216640	164548102752	1514469649396704	2621268206581024
…	…	…	…	…	…	…
40	1,963993579054E+119	4,115347994062E+108	1,788169031032E+112	1,8626849992297E+91	1,830259053207E+124	1,558622316946E+126