Universidade de Brasília - UnB
Instituto de Ciências Biológicas - IB
Programa de Pós-Graduação em Biologia Molecular

Resolução de novas interações proteína-proteína a partir de paisagens de sequências artificiais

João Antonio Alves Nunes

Orientador: Prof. Dr. Werner Treptow

Brasília/DF

2022

João Antonio Alves Nunes

Resolução de novas interações proteína-proteína a partir de paisagens de sequências artificiais

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Biologia Molecular, da Universidade de Brasília para a obtenção do título de Mestre em Ciências Biológicas (Biologia Molecular).

Orientador: Prof. Dr. Werner Treptow

Brasília/DF

2022

Agradecimentos

Agradeço a minha mãe, Leocades, pelo ensino, educação, dedicação, pelos valores, por ser um exemplo de pessoa batalhadora e por sempre me incentivar a dar o melhor de mim em qualquer coisa na vida. Ao meu padrastro, Walter, por todos os ensinamentos paternais e esforço empreendidos em prol da minha educação. Também ao meu cachorro, Adamastor, por ser um amigo e companheiro durante oito anos e por me ajudar a enfrentar o período de isolamento durante a pandemia.

Agradeço a minha companheira, Laís, pelo amor, companheirismo, paciência, cuidado, inspiração, e motivação para seguir em frente e correr atrás dos meus sonhos. Agradeço também por tudo que compartilhamos e estamos construindo.

Agradeço aos meus avós, Israel, Maria e Madalena, e a minha primeira professora e tia, Quézia, por todo cuidado, compreensão, ternura, ensinamentos e por me instigarem a curiosidade e vontade de aprender, características tão fundamentais para qualquer cientista.

Agradeço ao meu orientador, Werner, por todos os ensinamentos, conselhos, apoio e por me guiar nos caminhos científicos. Seu profissionalismo, dedicação e empolgação com a ciência me inspiram a buscar as respostas para os problemas mais desafiadores.

Agradeço aos meus colegas de LBTC, Natália, Vinicíus e Letícia, pelo acolhimento e todo apoio no laboratório. Em especial, agradeço ao Leonardo por me introduzir nos caminhos da programação e pela paciência e disponibilidade com minhas dúvidas e ao Fiorote que contribuiu diretamente com este trabalho.

Agradeço a Universidade de Brasília por ter me proporcionado a experiência da graduação e pós-graduação, pelos auxílios financeiros, por disponibilizar a infra-estrutura e por participar da minha formação como ser humano.

Por fim, agradeço as agências de fomento, CNPq e FAPDF, pelo auxílio financeiro.

Um dia me disseram
Que as nuvens não eram de algodão
Um dia me disseram
Que os ventos às vezes erram a direção [...]

(Humberto Gessinger/Eng. do Hawaii)

Resumo

Teorias coevolutivas descrevem a distribuição de probabilidade de proteínas que interagem em termos de um modelo estatístico de Boltzmann. Como resultado de pressões seletivas, espera-se que essa distribuição se desvie acentuadamente da uniformidade apresentando um número relativamente pequeno de sequências muito prováveis em todo o espaço de sequências. Enquanto essa afirmação deva ser verdadeira para sistemas interólogos em geral, suas distribuições de sequência podem não ter sido totalmente moldadas por pressões seletivas, abrindo a possibilidade de que novas sequências interólogas possam ser selecionadas a partir de distribuições de menor entropia geradas artificialmente. O objetivo desse trabalho foi investigar o significado físico de novas sequências selecionadas a partir de paisagens de *fitness* artificiais. Para isso, exploramos um Algoritmo Genético, que resolve a distribuição maximizando os acoplamentos estatísticos, começando de um alinhamento múltiplo de sequências nativo e explorando o espaço de alinhamentos múltiplo de sequências embaralhados. Também resolvemos uma distribuição minimizando os acoplamentos estatísticos através do embaralhamento ao acaso do alinhamento múltiplo de sequências. Uma vez que as sequências artificiais foram selecionadas a partir das distribuições maximizadas e minimizadas, suas energias livre de ligação em uma pose de interação nativa fixa foram avaliadas de acordo com os cálculos de energia livre baseados no método MM/PBSA. Para avaliar o sentido físico das sequências nativas e artificiais calculamos a temperatura de seleção em relação a sequências aleatórias de mesma composição. Nossos resultados apontam que é possível selecionar novas sequências artificiais não-similares em temperaturas de seleção mais frias ou mais quentes que a temperatura de seleção nativa e, que as sequências artificiais selecionadas apresentam diferenças apenas quanto ao *design* de sequências, mas não em relação à afinidade de ligação. É possível concluir que a evolução molecular da interação de dímeros não-obrigatórias pode ser restrita somente pelo *design*, já que a afinidade deve ser apenas consequência da composição de aminoácidos determinada pelas restrições de enovelamento. Além disso, constata-se a possibilidade de encontrar novas interações proteína-proteína com características iguais ou melhores que as interações existentes na natureza.

Palavras-chave: interação proteína-proteína. coevolução molecular. paisagens de sequências.

Abstract

Coevolutionary theories describe the probability distribution of interacting proteins in terms of a Boltzmann statistical model. As a result of selective pressures, that distribution is expected to sharply deviate from uniformity by featuring a relatively small number of highly probable sequences across the entire sequence space. While that statement must be true for interolog systems in general, their sequence distributions may have not been fully shaped by selective pressures opening the possibility that novel interolog sequences could be selected from artificially generated lower entropy distributions. The goal of this work was to investigate the physical meaning of selected sequences from artificial fitness landscapes. For that, we explore a Genetic Algorithm, which solves the distributions by maximizing the statistical coupling, starting from the native multi-sequence alignments and exploring the space of scrambled multi-sequence alignments. We also solve a distribution by minimizing statistical couplings through random shuffling of multiple sequence alignment. Once likely artificial sequences are selected from maximized and minimized distributions, their binding free-energies at a fixed native bound state are evaluated according to free energy calculations based on the MM/PBSA method. To evaluate the physical meaning of native and artificial sequences, we calculated the selection temperature in relation to random sequences of the same composition. Our results indicate that it is possible to select new non-similar artificial sequences at colder or warmer selection temperatures than the native selection temperature, and that the selected artificial sequences show differences only in sequence design, but not in relation to binding affinity. It is possible to conclude that the molecular evolution of the interaction of non-obligate dimers can be restricted only by sequence design, since the affinity must only be a consequence of the amino acid composition determined by the folding restrictions. In addition, it is possible to find new protein-protein interactions with characteristics equal to or better than the interactions existing in nature.

Keywords: protein-protein interaction. molecular coevolution. sequence landscapes.

Lista de Figuras

Figura 1 – O espaço de sequências. As proteínas observadas na natureza são aquelas que são termodinamicamente e cineticamente enoveláveis, entretanto várias sequências que cumprem esses requisitos não foram exploradas no curso evolutivo. Adaptado de Onuchic, Luthey-Schulten e Wolynes (1997). 2
Figura 2 – A) Exemplo de mutação compensatória entre os aminoácidos amarelo e marrom das proteínas A e B, em dois organismos diferentes. B) As mutações compensatórias geram padrões, nas colunas do alinhamento múltiplo de sequências das proteínas A e B, que permite reconstituir a história coevolutiva dessas proteínas. 3
Figura 3 – Esquema do Modelo Aleatório de Energias no contexto de interação proteína-proteína. Esse modelo estabelece que a média da energia livre de ligação de sequências nativas em poses de interação não-nativas deve ser igual a energia livre de ligação de uma sequência aleatória, com a mesma composição de aminoácidos da sequência nativa, na pose de interação nativa. 6
Figura 4 – Representação das estruturas tridimensionais dos complexos considerados no estudo. A) 1BXR. B) 1EP3. C) 1ZUN. D) 3G5O. 14
Figura 5 – Esquema representando a otimização realizada pelo Algoritmo Genético. O processo de maximização dos acoplamentos estatísticos tem inicio na concatenação nativa do alinhamento múltiplo de sequências (rosa) e procede até convergir em uma concatenação maximizada (vermelho). . 17
Figura 6 – Distribuições de probabilidades conjuntas de Boltzmann para as sequências nativas dos quatro sistemas considerados no estudo. Aqui foi considerado apenas as top 10% sequências mais prováveis no espaço formado por todas as concatenações possíveis advindas do alinhamento múltiplo de sequências nativo. 20
Figura 7 – Relação entre os acoplamentos estatísticos, obtidos a partir da paisagem de *fitness* nativa, e a composição da interface para os quatro sistemas analisados. 21

Figura 8 – Cima: Representação das estruturas tridimensionais dos complexos nas poses de interação nativas (roxo e ciano) e algumas poses de interação não-nativas geradas por *docking* (vermelho e branco). Baixo: Distribuições das energias livre de ligação obtidas pelo MM/PBSA para as sequências aleatórias (vermelho claro), nativas em pose de interação nativa (verde) e nativa em poses de interação não-nativas (vermelho escuro). A) 1BXR. B) 1EP3. C) 1ZUN. D) 3G5O. 22
Figura 9 – O ajuste linear entre as medianas das sequências aleatórias (vermelho) e nativas (verde) para os quatro sistemas analisados. Para melhor visualização, os pontos foram transladados de tal forma que a mediana das sequências aleatórias coincidisse com a coordenada (0,0). 23
Figura 10 – Análise de erro para as temperaturas de seleção nativas. Os valores de temperatura de seleção obtidos são comparados com a média da temperatura de seleção para enovelamento descrita na literatura. . . . 24
Figura 11 – Trajetórias do Algoritmo Genético para a otimização dos acoplamentos estatísticos. Para cada sistema foram realizadas três réplicas que convergiram segundo critério de parada $\dot{\Delta_m} \leq 0.005$. 25
Figura 12 – Distribuições de probabilidades conjuntas de Boltzmann para as sequências nativas, maximizadas e embaralhadas, dos quatro sistemas analisados. Aqui foi considerado apenas as top 10% sequências mais prováveis no espaço formado por todas as concatenações possíveis advindas do alinhamento múltiplo de sequências nativo. 26
Figura 13 – Relação entre os acoplamentos estatísticos, obtidos a partir das paisagens de *fitness* nativa, embaralhada e maximizada, e a composição da interface para o sistema 1BXR. 27
Figura 14 – Relação entre os acoplamentos estatísticos, obtidos a partir das paisagens de *fitness* nativa, embaralhada e maximizada, e a composição da interface para o sistema 1EP3. 27
Figura 15 – Relação entre os acoplamentos estatísticos, obtidos a partir das paisagens de *fitness* nativa, embaralhada e maximizada, e a composição da interface para o sistema 1ZUN. 28
Figura 16 – Relação entre os acoplamentos estatísticos, obtidos a partir das paisagens de *fitness* nativa, embaralhada e maximizada, e a composição da interface para o sistema 3G5O. 28
Figura 17 – Distribuições das energias livre de ligação obtidas pelo MM/PBSA para as sequências aleatórias (vermelho claro), maximizadas (azul), embaralhadas (verde escuro), nativas em pose de interação nativa (verde claro) e nativa em poses de interação não-nativas (vermelho escuro), para todos os sistemas analisados. 29

Figura 18 – Análise das sequências similares (rosa) e não-similares (verde) que compõe as distribuições de energias livre de ligação das sequências embaralhadas (verde escuro) e maximizadas (azul). 30
Figura 19 – O ajuste linear entre as medianas das sequências aleatórias (vermelho), nativas (verde claro), embaralhadas (verde escuro) e maximizadas (azul), para os quatro sistemas analisados. Para melhor visualização, os pontos foram transladados de tal forma que a mediana das sequências aleatórias coincidisse com a coordenada (0,0). 31

Lista de Tabelas

Tabela 1 – Complexos proteicos considerados no estudo. L é o número de sequências no MSA, M+N é o tamanho da sequência e $|\Theta|$ é o número de contatos na interface, seguindo a definição de *cutoff* de 8 Å. 14
Tabela 2 – Temperaturas de seleção (K) obtidas para as sequências embaralhadas, nativas e maximizadas 32

Lista de Abreviações e Siglas

APBS	do inglês, *Adaptative Poisson-Boltzmann Solver*
mfDCA	Análise de Acoplamento Direto por campo médio, do inglês, *mean-field Direct Coupling Analysis*
GA	Algoritmo Genético, do inglês *Genetic Algorithm*
MM/PBSA	do inglês, *Molecular Mechanics/Poisson-Boltzmann Surface Area*
MSA	Alinhamento múltiplo de sequências, do inglês, *Multiple Sequence Alignment*
PDB	Banco de Dados de Proteínas, do inglês, *Protein Data Bank*
SASA	Área de superfície acessível à solvente, do inglês, *Solvent-accessible surface area*
REM	Modelo de Energias Aleatórias, do inglês, *Random Energy Model*

Sumário

1 INTRODUÇÃO 1
1.1 Interação proteína-proteína 1
1.2 Coevolução 2
1.2.1 Coevolução Molecular 3
1.2.2 Paisagens de *fitness* coevolutivas 4
1.2.3 Explorando o espaço de sequências com composição fixa 4
1.3 Modelo de Energias Aleatórias 5
1.4 Justificativa 6

2 OBJETIVOS 8
2.1 Objetivo geral 8
2.2 Objetivos específicos 8

3 TEORIA E MÉTODOS 9
3.1 Distribuição de sequências menos restrita 9
3.2 Temperatura de seleção 10
3.3 Em busca de distribuições de baixa entropia 11
3.4 Paisagem de energia das sequências artificiais 12
3.5 Métodos computacionais 13
3.5.1 Sistemas sob investigação 13
3.5.2 Composição de sequências e sequências aleatórias 13
3.5.3 Acoplamentos Estatísticos 15
3.5.4 Algoritmo Genético 15
3.5.5 Cálculo de Energias Livres de Ligação 16

4 RESULTADOS E DISCUSSÃO 19
4.1 Paisagem de *fitness* coevolutiva nativa 19
4.2 Paisagens de *fitness* coevolutivas artificiais 24

5 CONCLUSÃO E PERSPECTIVAS 33

REFERÊNCIAS BIBLIOGRÁFICAS 34

1 Introdução

1.1 Interação proteína-proteína

As interações proteína-proteína são fundamentais para diversos processos celulares, como a replicação, transcrição, tradução. A formação de complexos proteicos ocorre por meio do estabelecimento de interações não-covalentes entre aminoácidos na interface das moléculas parceiras, esse processo espontâneo permite a montagem de estruturas complexas que carregam novas funções biológicas (CHOTHIA; JANIN, 1975; JANIN, 1995). Duas características são essenciais para compreender e descrever as interações proteína-proteína, são elas: a afinidade e a especificidade. A afinidade pode ser definida como a diferença de energia livre entre o complexo proteico imerso em solvente e as proteínas livres e solvatadas pelo solvente (JANIN, 1995). Por outro lado, a especificidade é definida como a capacidade de uma proteína em discriminar seu parceiro molecular e sua pose de ligação nativa, evitando a interação com quaisquer outros parceiros e em poses de ligação não-nativas (JANIN, 1996).

Outro importante fator para a compreensão das interações proteína-proteína é sua classificação em diferentes tipos. Nessa ocasião, Nooren e Thornton (2003) classificaram os complexos proteicos quanto a sua composição, obrigatoriedade e duração da interação. No que diz respeito à composição, os complexos proteicos podem ser formados pela ligação entre proteínas idênticas ou proteínas diferentes, onde são denominados homo-oligômeros e hétero-oligômeros, respectivamente. Com relação à obrigatoriedade de formação dos complexos, esses podem ser classificados em obrigatórios, quando são formados por proteínas que não apresentam estabilidade por si próprias ou não-obrigatórios, onde as proteínas apresentam estabilidade para existir independentemente. Por fim, quanto à duração da interação, uma interação permanente possui alta afinidade e apenas existe em sua forma complexada, por outro lado, uma interação transiente se associa e desassocia continuamente, e são normalmente relacionadas a processos celulares regulatórios. Apesar dessas classificações, as interações proteína-proteína existem na forma de um contínuo e suas características dependem das condições fisiológicas e do ambiente (NOOREN; THORNTON, 2003).

Enquanto são selecionadas para serem termodinamicamente estáveis e cineticamente acessíveis em um enovalemento particular (ONUCHIC; LUTHEY-SCHULTEN; WOLYNES, 1997; GARCIA; TREPTOW; ARAÚJO, 2001; TREPTOW et al., 2002) (Fig. 1), duas proteínas interólogas não-obrigatórias A e B ainda devem garantir sua estabilidade de energia livre de ligação contra um vasto repertório de poses de interações não-nativas e parceiros não específicos. Assim, as interações proteína-proteína se apresentam como uma

importante restrição na evolução molecular de proteínas.

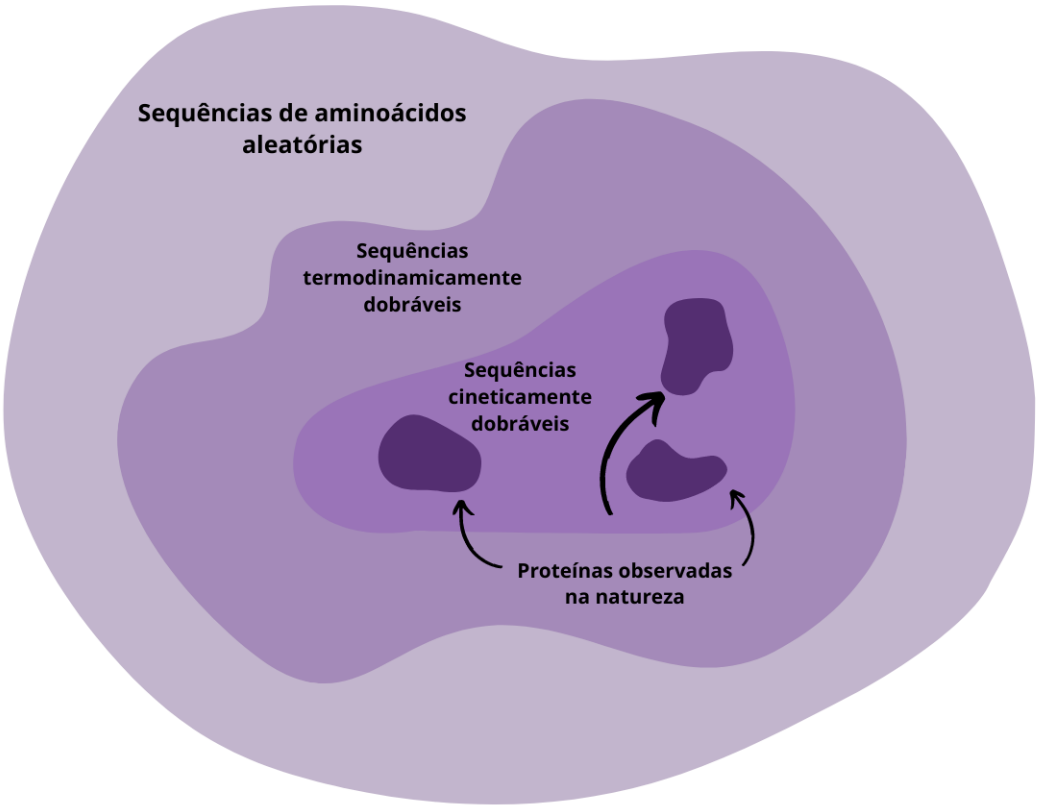

Figura 1 – O espaço de sequências. As proteínas observadas na natureza são aquelas que são termodinamicamente e cineticamente enoveláveis, entretanto várias sequências que cumprem esses requisitos não foram exploradas no curso evolutivo. Adaptado de Onuchic, Luthey-Schulten e Wolynes (1997).

1.2 Coevolução

A evolução recíproca de características entre duas espécies é definida como coevolução (JANZEN, 1980; THOMPSON, 1994). Esse termo foi cunhado por Mode (1958) e popularizado por Ehrlich e Raven (1964) em um estudo sobre a interação entre borboletas e plantas. A coevolução é observada macroscopicamente em grupos de organismos que estabelecem algum tipo de interação ecológica como, por exemplo: presas-predadores, parasitas-hospedeiros, espécies mutualísticas e competição interespecífica. Nesse âmbito, essas interações podem ser perpetuadas ao longo dos anos através de um equilíbrio dinâmico (VALEN, 1973).

1.2.1 Coevolução Molecular

Em sistemas microscópicos, a coevolução pode ocorrer em várias biomoléculas, porém são as proteínas as mais estudadas (JUAN; PAZOS; VALENCIA, 2013) e o foco desse trabalho. Nesse cenário, a coevolução molecular decorre mediante a manutenção da interação entre resíduos de aminoácidos, principalmente na forma de mutações compensatórias (GÖBEL et al., 1994; PAZOS et al., 1997; LOVELL; ROBERTSON, 2010). Por exemplo, se um resíduo de aminoácido sofre mutação com perda de valor adaptativo, então seu resíduo de aminoácido contactante, seja ele intra ou inter-proteína, sofrerá pressão seletiva para reestabelecer o valor adaptativo da interação, por meio de outra mutação (Fig. 2A). Ao longo do tempo, esse processo é responsável por gerar um conjunto de sequências primárias variantes, que quando organizadas em um alinhamento múltiplo de sequências (MSA) permite reconstituir, com análises estatísticas, a história coevolutiva dessas proteínas (Fig. 2B).

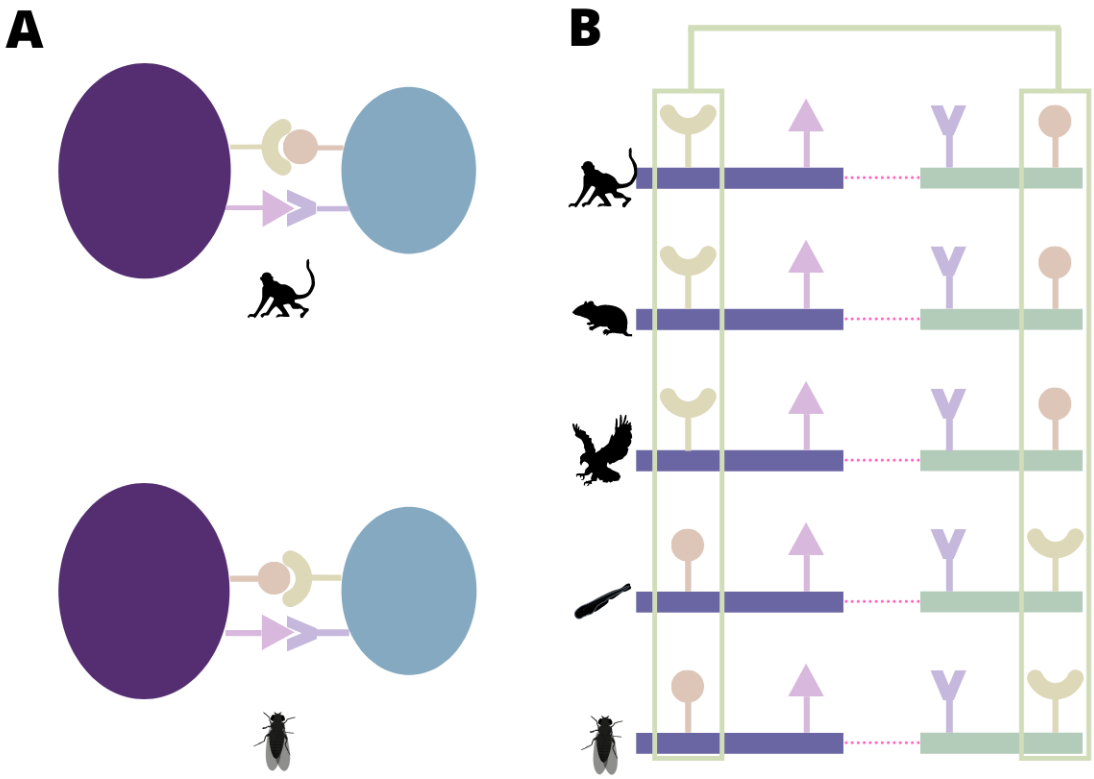

Figura 2 – A) Exemplo de mutação compensatória entre os aminoácidos amarelo e marrom das proteínas A e B, em dois organismos diferentes. B) As mutações compensatórias geram padrões, nas colunas do alinhamento múltiplo de sequências das proteínas A e B, que permite reconstituir a história coevolutiva dessas proteínas.

Teorias coevolutivas recentes descrevem a distribuição de probabilidade de sequên-

cias proteicas em termos de um modelo estatístico de Boltzmann; com o Hamiltoniano dado por um conjunto de campos locais e acoplamentos estatísticos (WEIGT et al., 2009; BURGER; NIMWEGEN, 2010; MORCOS et al., 2011; JONES et al., 2012; OVCHINNIKOV; KAMISETTY; BAKER, 2014). Desse modelo é extraída a informação coevolutiva, que é amplamente utilizada para: inferência de interações proteicas para um conjunto de parálogos (BITBOL et al., 2016; GUEUDRÉ et al., 2016; BITBOL, 2018), predição de interfaces proteína-proteína (SCHUG et al., 2009; OVCHINNIKOV; KAMISETTY; BAKER, 2014; HOPF et al., 2014; SANTOS et al., 2015) e predição de estruturas tridimensionais (SUŁKOWSKA et al., 2012; OVCHINNIKOV et al., 2017), além de ser um dos componentes do AlphaFold (JUMPER et al., 2021), a Inteligência Artificial que vem revolucionando o campo da predição de enovelamentos.

1.2.2 Paisagens de *fitness* coevolutivas

Ademais, outro problema do campo da Coevolução Molecular é explorar paisagens de sequências completas (MORCOS; ONUCHIC, 2019). Os avanços recentes realizados nessa área exploram o espaço de sequências variando a composição de aminoácidos em relação as sequências naturais. Nesse cenário, Figliuzzi et al. (2016) aplicaram um método coevolutivo para inferir o efeito de mutações na proteína TEM-1, envolvida com a resistência aos antibióticos em *Escherichia coli*. Cheng et al. (2016) observaram que uma paisagem coevolutiva era capaz de identificar os mutantes funcionais dentre o espaço de variantes das proteínas PhoQ/PhoP de *E. coli*. O mesmo grupo também usou a informação coevolutiva para realizar o design de interações entre as proteínas EnvZ de *E. coli* e Spo0F de *Bacillus subtilis*, que são parceiras não-cognatas (CHENG et al., 2018). Tian et al. (2018) utilizaram uma abordagem coevolutiva para realizar o design de novas sequências dos domínios GA e GB da proteína G estreptocócica e do domínio SH3 e constataram que essas se enovelaram de forma estável. Russ et al. (2020) aplicaram um modelo estatístico baseado em coevolução para explorar o espaço de sequências funcionais e projetar sequências artificiais, das enzimas da família de corismato mutases. Xie, Asadi e Warshel (2022) demonstraram uma correlação entre a atividade catalítica enzimática e a Hamiltoniana das sequências, sugerindo que paisagens coevolutivas poderiam ser utilizadas para guiar o design de enzimas mais eficientes. Chi et al. (2022) conectaram uma paisagem coevolutiva com ensaios experimentais de evolução direta para projetar uma Rodopsina fluorescente e sensível a cloreto.

1.2.3 Explorando o espaço de sequências com composição fixa

O presente trabalho fundamenta-se na seguinte afirmativa: pressões seletivas são responsáveis por desviar a distribuição de probabilidades de sequências da uniformidade, apresentando um número pequeno de sequências com alta probabilidade dentro de todo o

espaço de sequências, isto é equivalente a dizer que a seleção natural reduz a entropia da distribuição de sequências em relação ao máximo alcançável no caso uniforme. Embora essa afirmação deva ser verdadeira para sistemas interólogos em geral, suas distribuições de probabilidades de sequências podem não ser completamente moldadas por pressões seletivas, abrindo a possibilidade para que novas sequências interólogas possam ser selecionadas a partir de distribuições de probabilidade com menor entropia, geradas artificialmente. Intrínseco a essa suposição está o fato de que a minimização da entropia deve ser uma condição necessária, mas não suficiente para esse fim, já que as distribuições probabilísticas de sequências apresentam grande degeneração, que pode resultar na seleção de sequências prováveis sem significado físico. Portanto, estamos particularmente interessados em explorar o espaço de sequências mantendo a composição de aminoácidos das sequências fixa. A seleção artificial permitiria, em princípio, projetar um novo subconjunto de sequências mais prováveis que poderiam impactar na especificidade do estado nativo sobre os modos de ligação não-nativos e parceiros não-específicos.

1.3 Modelo de Energias Aleatórias

O Modelo de Energias Aleatórias (REM) foi desenvolvido para descrever sistemas *spin glass* (DERRIDA, 1980; DERRIDA, 1981), e é bem estabelecido no campo de Enovelamento de Proteínas, onde retrata estados desenovelados como heteropolímeros aleatórios (BRYNGELSON; WOLYNES, 1987; ONUCHIC; LUTHEY-SCHULTEN; WOLYNES, 1997). Isso deve-se ao fato de que estruturas desenoveladas são repletas de interações não-nativas entre resíduos de aminoácidos, que contribuem aleatoriamente - estabilizando ou desestabilizando - com a energia daquela conformação, e podem ser modeladas como as contribuições aleatórias das interações que ocorrem em qualquer conformação de um heteropolímero aleatório (ONUCHIC; LUTHEY-SCHULTEN; WOLYNES, 1997). Recentemente, Morcos et al. (2014) utilizaram o REM para conectar o funil de enovelamento proteico com a paisagem coevolutiva de sequências, em oito famílias proteicas, e assim mensurar, por meio da temperatura de seleção, a força das restrições de enovelamento durante a evolução proteica. Analogamente ao contexto de enovelamento, Janin (1996) extrapolou o REM para descrever a energética de modos de interações não-nativos no contexto de interações proteína-proteína (Fig. 3). De fato, outros autores demonstraram mais tarde que as interações proteína-proteína seguem o principio da frustração mínima e devem ser guiadas por uma paisagem de energia afunilada, assim como o enovelamento proteico (WANG; VERKHIVKER, 2003; LEVY; WOLYNES; ONUCHIC, 2004).

De acordo com o Modelo de Energias Aleatórias, no contexto de interações proteína-proteína, o processo artificial aplicado no presente trabalho deve envolver a realização de novas sequências, oriundas de distribuições de sequências de menor entropia, que:

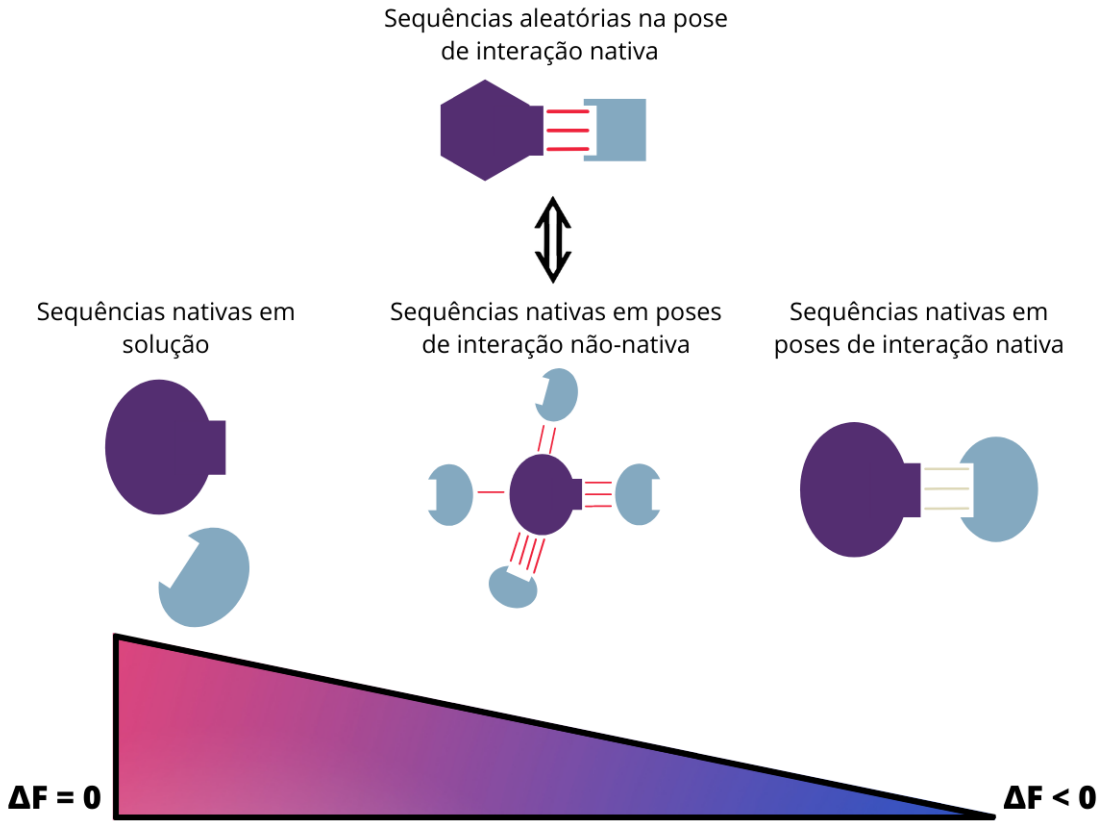

Figura 3 – Esquema do Modelo Aleatório de Energias no contexto de interação proteína-proteína. Esse modelo estabelece que a média da energia livre de ligação de sequências nativas em poses de interação não-nativas deve ser igual a energia livre de ligação de uma sequência aleatória, com a mesma composição de aminoácidos da sequência nativa, na pose de interação nativa.

melhoram a estabilidade de energia livre no estado de ligação nativo, em uma temperatura de seleção fixa; ou igualmente estáveis em temperaturas mais fria. Da mesma forma, esse processo artificial em sentido oposto deve envolver a realização de novas sequências, oriundas de distribuições de sequências de maior entropia, que: pioram a estabilidade de energia livre no estado de ligação nativo, em temperatura fixa; ou igualmente estáveis em temperaturas mais quentes.

1.4 Justificativa

Um tema recorrente nas investigações biológicas é compreender a relação entre sequência, estrutura e função de proteínas e como esses três elementos definem o espaço onde ocorre a evolução molecular dessas macromoléculas (LEVY, 2010). As teorias de coevolução molecular permitem resolver o padrão de aminoácidos em sequências proteicas selecionadas pela natureza, mas além disso, abre a possibilidade de explorar espaços

de sequências que não foram explorados pela seleção natural ou que ainda não foram resolvidos através de sequenciamento (MORCOS; ONUCHIC, 2019). O presente trabalho explorou espaços de sequências com composição fixa e ajuda a definir o sentido físico de novas sequências interólogas resolvidas a partir de cenários evolutivos distintos. A compreensão e resolução de novas interações entre proteínas possui importantes implicações na Biologia de Sistemas (CUSICK et al., 2005), no desenho racional de fármacos (WELLS; MCCLENDON, 2007) e na Biologia Sintética, principalmente no design de proteínas (KORTEMME; BAKER, 2004). Ademais, nossas investigações trazem novas perspectivas acerca do impacto das interações proteína-proteína na evolução de sistemas proteicos e esperamos que esse trabalho possa contribuir com o avanço da área em direção a uma teoria estatística quantitativa sobre evolução molecular.

2 Objetivos

2.1 Objetivo geral

Investigar o sentido físico de novas sequências provenientes de paisagens de *fitness* artificiais.

2.2 Objetivos específicos

1. Selecionar distribuições de probabilidades de sequências com menor e maior entropia, através de um Algoritmo Genético;
2. Avaliar a energia livre de ligação das sequências artificiais, sequências aleatórias e nativas;
3. Corroborar o Modelo de Energias Aleatórias no contexto de interação proteína-proteína;
4. Mensurar quão forte é a restrição da interação proteína-proteína ao longo da evolução proteica e comparar com as restrições de enovelamento;

3 Teoria e Métodos

Considere os interólogos A e B. Suas sequências de aminoácidos são respectivamente descritas por dois blocos de variáveis estocásticas discretas $X^M \equiv (X_1, ..., X_M)$ e $Y^N \equiv (Y_1, ..., Y_N)$ e função massa de probabilidade conjunta

$$\rho(x^M, y^N|z) \tag{3.1}$$

que satisfaz a condição de normalização

$$\sum_{x^M, y^N} \rho(x^M, y^N|z) = 1 \tag{3.2}$$

em cada sequência conjunta A e B definida no alfabeto χ de tamanho $|\chi|$. A probabilidade conjunta na eq. [3.1] é assumida como condicional a um dado processo evolutivo z descrito pela variável estocástica Z com função massa de probabilidade $\rho(z)$. É esperado que cada realização de z forme uma distribuição probabilística única de tamanho $|\chi|^{M+N}$

$$\boldsymbol{\rho} \stackrel{\text{def}}{=\!=} \{\rho(x^M, y^N|z)\}_{|\chi|^{M+N}} \tag{3.3}$$

com entropia

$$S(\boldsymbol{\rho}) = -\sum_{x^M, y^N} \rho(x^M, y^N|z) \ln \rho(x^M, y^N|z) \tag{3.4}$$

definida de acordo com a teoria da informação (SHANNON, 1948).

Estamos interessados em resolver um processo evolutivo artificial z para o qual a entropia da distribuição das sequências derivadas

$$S(\boldsymbol{\rho}) - S(\boldsymbol{\rho_0}) \leq 0 \tag{3.5}$$

é menor que o processo evolutivo de referência z_0 associado com a distribuição das sequências interólogas A e B na natureza $\boldsymbol{\rho_0}$.

3.1 Distribuição de sequências menos restrita

Em sua forma estatística menos restrita $\boldsymbol{\rho}^*$, a distribuição de probabilidade conjunta condicional é Boltzmann

$$\rho^*(x^M, y^N|z) \propto e^{-H(x^M, y^N|z)} \tag{3.6}$$

com Hamiltoniana

$$H(x^M, y^N|z) = -[\sum_{i=1}^{M} \lambda^*_{x_i|z} + \sum_{i=1}^{M-1} \sum_{j=i+1} \lambda^*_{x_i x_j|z} + \sum_{i=1}^{N} \lambda^*_{y_i|z} + \sum_{i=1}^{N-1} \sum_{j=i+1} \lambda^*_{y_i y_j|z} + \sum_{i=1}^{M} \sum_{j=1}^{N} \lambda^*_{x_i y_j|z}] \tag{3.7}$$

definida em termos de um conjunto de campos locais $\{\lambda^*_{x_i|z}, \lambda^*_{y_i|z}\}_{(M+N)|\chi|}$ e acoplamentos estatísticos $\{\lambda^*_{x_ix_j|z}, \lambda^*_{y_iy_j|z}, \lambda^*_{x_iy_j|z}\}_{\binom{M+N}{2}|\chi|^2}$.

A derivação do modelo probabilístico na eq. [3.6] ocorre por meio do principio de máxima entropia (JAYNES, 1957) e segue da solução do ponto crítico $(\boldsymbol{\rho}^*, \boldsymbol{\lambda}^*)$ da função *Lagrangiana*

$$\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\lambda}) = S(\boldsymbol{\rho}) - \boldsymbol{\lambda}[\boldsymbol{\nu}(\boldsymbol{\rho}) - \boldsymbol{f}] \tag{3.8}$$

de tal modo que,

$$(\frac{\partial \mathcal{L}}{\partial \rho})_{\rho^*} = 0 \quad e \quad (\frac{\partial \mathcal{L}}{\partial \lambda})_{\lambda^*} = 0 \tag{3.9}$$

para um conjunto de multiplicadores de *Lagrange*

$$\boldsymbol{\lambda}^* = \{\lambda^*_{x_i|z}, \lambda^*_{y_i|z}, \lambda^*_{x_ix_j|z}, \lambda^*_{y_iy_j|z}, \lambda^*_{x_iy_j|z}\}_{(M+N)|\chi|+\binom{M+N}{2}|\chi|^2} \tag{3.10}$$

que restringem as distribuições de probabilidades marginais de um e dois sítios da distribuição de probabilidade conjunta condicional

$$\boldsymbol{\nu}(\boldsymbol{\rho}^*) = \{\rho^*(x_i|z), \rho^*(y_i|z), \rho^*(x_i, x_j|z), \rho^*(y_i, y_j|z), \rho^*(x_i, y_j|z)\}_{(M+N)|\chi|+\binom{M+N}{2}|\chi|^2} \tag{3.11}$$

em um conjunto de frequências simples e conjuntas

$$\boldsymbol{f} = \{f_{x_i|z}, f_{y_i|z}, f_{x_ix_j|z}, f_{y_iy_j|z}, f_{x_iy_j|z}\}_{(M+N)|\chi|+\binom{M+N}{2}|\chi|^2} \tag{3.12}$$

a serem determinadas empiricamente através de um alinhamento múltiplo de sequências

$$\boldsymbol{m} \stackrel{\text{def}}{=\!=} \{x^M, y^N|z\}_L \tag{3.13}$$

de L sequências conjuntas A e B combinando com um processo evolutivo z.

Dado o posto da matriz *Jacobiana* $D\boldsymbol{\nu}(\boldsymbol{\rho})$, o número de restrições qualificadas não degeneradas que garantem uma solução única $(\boldsymbol{\rho}^*, \boldsymbol{\lambda}^*)$ é menor que o número total de parâmetros em $\boldsymbol{f}$. Logo a solução $(\boldsymbol{\rho}^*, \boldsymbol{\lambda}^*)$ envolve um certo número de graus de liberdade fixos em $\boldsymbol{\lambda}^*$.

Na presente formulação, a condicionalidade do processo evolutivo z do ponto crítico $(\boldsymbol{\rho}^*, \boldsymbol{\lambda}^*)$ da eq. [3.8] deriva essencialmente do conjunto de parâmetros externos $\boldsymbol{f}$ determinados a partir do alinhamento múltiplo de sequências $\boldsymbol{m}$.

3.2 Temperatura de seleção

De acordo com a Teoria de Paisagens baseada no Modelo de Energias Aleatórias (ONUCHIC; LUTHEY-SCHULTEN; WOLYNES, 1997), o significado físico das sequências nativas pode ser avaliado notando que sua distribuição corresponde a um conjunto canônico

$$\rho^*(x^M, y^N|z_0) \propto e^{-(k_BT_0)^{-1}F_N(x^M, y^N|z_0) - F_{nN}(x^M, y^N|z_0)} \tag{3.14}$$

caracterizado por uma temperatura de seleção efetiva T^* que dimensiona a energia livre de ligação $F_N(x^M, y^N|z_0)$ de cada realização de sequência na estrutura de ligação nativa das proteínas A e B em relação a média de energias livre de ligação das mesmas sequências em estruturas de ligação não-nativas $F_{nN}(x^M, y^N|z_0)$.

Das equações [3.6] e [3.14], nota-se a equivalência do peso de Boltzmann em ambas as descrições estatística e termodinâmica,

$$e^{H(x^M,y^N|z_0)-H_{rnd}(x^M,y^N|z_0)} \equiv e^{-(k_B T_0^*)^{-1} F_N(x^M,y^N|z_0)-F_{nN}(x^M,y^N|z_0)} \tag{3.15}$$

com a devida adição do termo $H_{rnd}(x^M, y^N|z_0)$, que corresponde ao valor da Hamiltoniana para uma sequência aleatória. O REM estabelece que a média das energias livre de ligação da sequência nativa em poses de interação não-nativas é equivalente a energia livre de ligação, na pose de interação nativa, de uma sequência aleatória de mesma composição,

$$F_{nN}(x^M, y^N|z_0) \equiv F_{rnd}(x^M, y^N|z_0). \tag{3.16}$$

Portanto, da equação [3.15] é possível obter a temperatura de seleção efetiva T_0^*, descrita por Morcos et al. (2014) para as sequências nativas em relação a sequências aleatórias, como uma razão entre as diferenças de energias livre de ligação e de Hamiltonianas,

$$T_0^* = \frac{-F_N(x^M, y^N|z_0) + F_{rnd}(x^M, y^N|z_0)}{k_B(H(x^M, y^N|z_0) - H_{rnd}(x^M, y^N|z_0))}. \tag{3.17}$$

3.3 Em busca de distribuições de baixa entropia

O cálculo exato da entropia $S(\boldsymbol{\rho}^*)$ na eq. [3.4] percorre por um grande espaço de sequências prováveis tornando impraticável sua minimização direta por qualquer método computacional. Uma alternativa é explorar a dependência implícita da entropia $S(\boldsymbol{\rho}^*(\boldsymbol{f}))$ com o conjunto de frequências simples e conjuntas de aminoácidos $\boldsymbol{f}$ para avaliar a eq. [3.5] em termos da maximização do seu gradiente. Dessa forma, assumindo que $\boldsymbol{\rho}^*(\boldsymbol{f})$ e $\boldsymbol{\lambda}^*(\boldsymbol{f})$ são funções diferenciáveis de $\boldsymbol{f}$, o Teorema do Envelope estabelece

$$\nabla_f S(\boldsymbol{\rho}^*(\boldsymbol{f})) = \nabla_f \mathcal{L}(\boldsymbol{\rho}^*(\boldsymbol{f}), \boldsymbol{\lambda}^*(\boldsymbol{f}), \boldsymbol{f}) = \boldsymbol{\lambda}^*(\boldsymbol{f}) \tag{3.18}$$

uma conexão direta entre o gradiente da entropia $\nabla_f S(\boldsymbol{\rho}^*(\boldsymbol{f}))$ e o multiplicador de *Lagrange* $\boldsymbol{\lambda}^*(\boldsymbol{f})$. Nesse caso, a eq. [3.18] apresenta os multiplicadores de *Lagrange* como mudanças marginais da entropia com frequências de aminoácidos e a condição

$$|\boldsymbol{\lambda}^*(\boldsymbol{f})| - |\boldsymbol{\lambda}^*(\boldsymbol{f_0})| \geq 0 \tag{3.19}$$

é equivalente a minimização da eq. [3.5].

Uma vez que as distribuições de baixa entropia podem ser altamente degeneradas, retratando sequências prováveis desprovidas de realidade física, queremos resolver a eq. [3.19] para o caso em que as distribuições marginais das variáveis X^M e Y^N são incondicionais

$$\begin{cases} \rho^*(x^M) = \sum\limits_{y^N} \rho^*(x^M, y^N|z) \\ \rho^*(y^N) = \sum\limits_{x^M} \rho^*(x^M, y^N|z) \end{cases} \tag{3.20}$$

significando que, para uma composição fixa de sequência das proteínas A e B, somente suas probabilidades conjuntas dependem do processo evolutivo z. Assim, o critério da eq. [3.19] se reduz a

$$[\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{x,y\in\chi} |\lambda^*_{x_i y_j|z}|^2]^{\frac{1}{2}} - [\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{x,y\in\chi} |\lambda^*_{x_i y_j|z_0}|^2]^{\frac{1}{2}} \;\geq\; 0 \tag{3.21}$$

e a minimização de entropia ocorre no sentido de uma distribuição de probabilidades $\boldsymbol{\rho}^*$ restringida por um conjunto de frequências de aminoácidos $\boldsymbol{f}$ associadas a sequências conjuntas de parceiros não-nativos das proteínas A e B, ou seja, derivadas de uma concatenação embaralhada dos blocos de aminoácidos M e N no alinhamento múltiplo de sequências $\boldsymbol{m}$.

3.4 Paisagem de energia das sequências artificiais

Dada a condição da eq. [3.20], somente sequências artificiais com a mesma composição de aminoácidos das sequências naturais serão resolvidas, com probabilidade mensurável através da eq. [3.21], implicando que todas as outras sequências são desprezíveis. Seguindo a mesma lógica da eq. [3.15], é possível escrever a identidade adimensional entre sequências nativas e artificiais como,

$$H(x^M, y^N|z) - H(x^M, y^N|z_0) = (k_B T)^{-1} F_N(x^M, y^N|z) - (k_B T_0)^{-1} F_N(x^M, y^N|z_0). \tag{3.22}$$

A referência da energia livre de ligação e da Hamiltoniana das sequências aleatórias devem ser constantes e análogas em ambas as realizações de sequências nativas e artificiais, portanto elas se cancelam e não aparecem na eq. [3.22].

Esperamos que as sequências prováveis, resolvidas a partir de uma distribuição de menor entropia, maximizem seus acoplamentos estatísticos de acordo com a eq. [3.21], logo a condição

$$H(x^M, y^N|z) - H(x^M, y^N|z_0) \leq 0 \tag{3.23}$$

impõe dois possíveis cenários para a realidade física das sequências artificiais. O cenário (i) envolve a realização de novas sequências que melhoram a estabilidade de energia livre

do estado de ligação nativo em uma temperatura fixa de seleção.

$$\begin{cases} T^* = T_0^* \\ F_N(x^M, y^N|z) < F_N(x^M, y^N|z_0) \end{cases} \tag{3.24}$$

Por outro lado, no cenário (ii) temos a seleção de sequências igualmente estáveis em temperaturas mais frias.

$$\begin{cases} T^* < T_0^* \\ F_N(x^M, y^N|z) = F_N(x^M, y^N|z_0) \end{cases} \tag{3.25}$$

A seguir exploramos as equações [3.21] e [3.22] para investigar sequências artificiais de menor entropia das proteínas A e B à luz dos cenários (i) e (ii). Iremos também utilizar as mesmas equações, com as devidas modificações dos sinais de desigualdade, para avaliar o sentido físico de sequências artificiais provenientes de uma paisagem de sequências de maior entropia.

3.5 Métodos computacionais

3.5.1 Sistemas sob investigação

O conjunto de proteínas A e B, que sabidamente interagem, considerado no estudo é apresentado na tabela [1]. A escolha desses sistemas visou abranger diferentes números de sequências e tamanhos de interfaces. Todos esses sistemas são hetero-dímeros não-obrigatórios. Seus alinhamentos múltiplos de sequências referência (nativo)

$$\boldsymbol{m_0} = \{x^M, y^N|z_0\}_L \tag{3.26}$$

foram reconstruídos a partir de alinhamentos de sequências primárias publicados por Ovchinnikov, Kamisetty e Baker (2014), cada qual contendo L sequências pareadas com interações proteína-proteína conhecidas e definidas no alfabeto de 20 aminoácidos mais o símbolo de *gap* ($|\chi| = 21$). Os contatos na interface foram identificados através da estrutura cristalográfica dos complexos proteicos e definidos como $\Theta = \{i, j|\ |r_i - r_j| \leq 8\text{Å}\}$, onde r_i e r_j são respectivamente os pontos de referência dos resíduos de aminoácidos nas posições i e j, com distância interatômica inferior ao *cutoff* de 8 Å (Fig. 4). No presente trabalho, o ponto de referência adotado para todos os aminoácidos foi o carbono beta ($C\beta$), com exceção da Glicina, onde a referência adotada foi o carbono alfa ($C\alpha$).

3.5.2 Composição de sequências e sequências aleatórias

Visto que a composição de aminoácidos das sequências é uma variável importante no presente trabalho, definimos que, para uma sequência l, sua composição em relação as

Tabela 1 – Complexos proteicos considerados no estudo. L é o número de sequências no MSA, M+N é o tamanho da sequência e $|\Theta|$ é o número de contatos na interface, seguindo a definição de *cutoff* de 8 Å.

Nome	**PDB ID**	**Cadeias**	**L**	**M + N**	$\|\Theta\|$
Carbamoil Fosfato Sintetase	1BXR	A, B	1004	1452	154
Diidroorotato Desidrogenase B	1EP3	A, B	552	572	91
ATP-Sulfurilase regulada por GTP	1ZUN	A, B	649	630	140
Complexo Toxina-Antitoxina RelBE2	3G5O	A, B	904	180	92

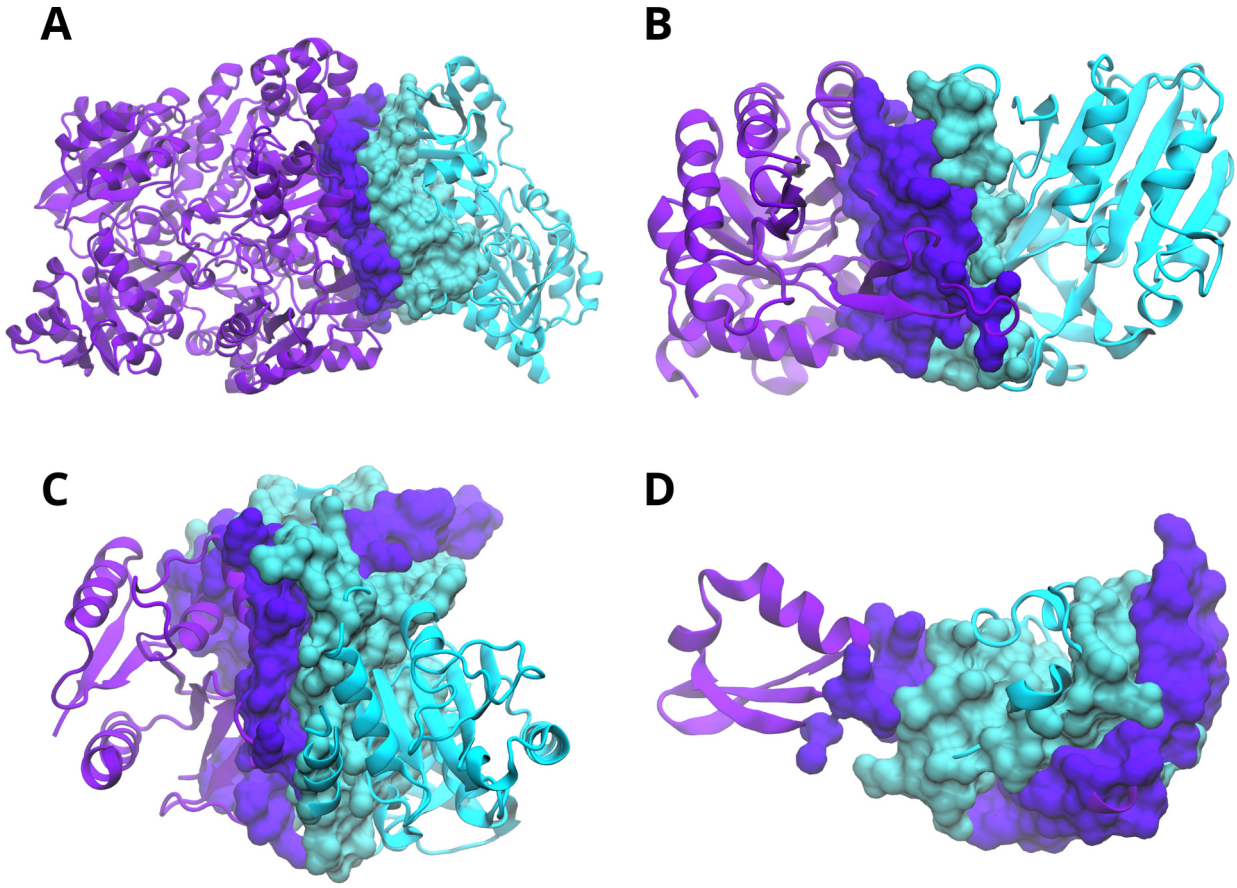

Figura 4 – Representação das estruturas tridimensionais dos complexos considerados no estudo. A) 1BXR. B) 1EP3. C) 1ZUN. D) 3G5O.

sequências nativas, é dada por

$$C = \frac{< \boldsymbol{f_l} \ , \ \frac{1}{L} \sum_{l=1}^{L} \boldsymbol{f_{l_0}} >}{|\boldsymbol{f_l}| \ |\frac{1}{L} \sum_{l=1}^{L} \boldsymbol{f_{l_0}}|} \tag{3.27}$$

onde $\boldsymbol{f_l} = \{f_{x_l}\}_{|\chi|}$ é a frequência de aminoácidos da sequência l e $\boldsymbol{f_{l_0}}$ é a frequência de aminoácidos de uma sequência nativa l_0.

Para avaliar a influência da composição nos acoplamentos estatísticos, para cada sistema considerado no estudo foram geradas três populações de sequências aleatórias, por meio de mutações ao acaso. As duas primeiras tiveram suas composições, para toda a sequência, restritas a $C_{M+N} \leq 0.3$ e $C_{M+N} \geq 0.8$ e a última teve a composição, da interface de interação, restrita a $C_{\Theta} \geq 0.8$.

3.5.3 Acoplamentos Estatísticos

Os acoplamentos estatísticos foram calculados de acordo com a metodologia de Análise de Acoplamento Direto por campo médio (mfDCA), descrita por Morcos et al. (2011). Como explicado por Morcos et al. (2014), a DCA pode ser ruidosa por efeito de uma amostragem finita, assim, o cálculo dos acoplamentos estatísticos foi restrito ao conjunto Θ de contatos na interface de interação proteica. Além disso, é nessa interface que está armazenada a informação coevolutiva (ANDRADE; PONTES; TREPTOW, 2019).

O cálculo das frequências de aminoácidos simples e conjuntas, necessárias para o cômputo dos acoplamentos estatísticos, também foi baseado na metodologia proposta por Morcos et al. (2011), de tal forma que

$$\begin{cases} f_A = (M_{eff} + \kappa)^{-1}[\kappa|\chi|^{-1} + \sum_{l=1}^{L} \delta_{A,A(l)}(n_l)^{-1}] \ \ \forall \ \ A = x_i, y_i \\ f_{AB} = (M_{eff} + \kappa)^{-1}[\kappa|\chi|^{-2} + \sum_{l=1}^{L} \delta_{AB,A(l)B(L)}(n_l)^{-1}] \ \ \forall \ \ AB = x_i x_j, y_i y_j, x_i y_j | z \end{cases} \tag{3.28}$$

com,

$$n_l = |\{k \mid 1 \leq k \leq L, \ d(k,l) \geq \theta\}| \tag{3.29}$$

denotando o número de sequências similares k dentro de um certo *cutoff* de distância de Hamming θ da sequência l e

$$M_{eff} = \sum_{l=1}^{L} (n_l)^{-1} \tag{3.30}$$

o número efetivo de sequências distinguíveis no *cutoff* de distância θ. O delta de Kronecker δ assume valor um em caso de índices iguais, e zero no caso contrário. O pseudocontador κ restringe a ocorrência de frequências de aminoácidos iguais a zero, e portanto corrige possíveis vieses de amostragem.

No presente trabalho, todos os cálculos de frequências consideraram $\theta = 1.0$ e $\kappa = 0.5$.

3.5.4 Algoritmo Genético

Começando de um alinhamento múltiplo de sequências nativos $\boldsymbol{m_0}$, um Algoritmo Genético (GA) foi aplicado para resolver a distribuição de sequências de menor entropia (Fig.

5). O algoritmo utilizado foi adaptado do código disponibilizado por Pontes et al. (2021) e para cada complexo proteico considerado nesse estudo, três trajetórias independentes de GA foram realizadas. Cada trajetória foi otimizada com uma população de oito indivíduos e em cada geração t, o indivíduo com melhor *fitness* foi replicado na geração seguinte, e os sete indivíduos com piores fitness foram substituídos por cópias mutadas do melhor individuo. A mutação de um indivíduo consiste na troca de posição de duas sequências da proteína B, que resulta em uma nova concatenação de sequências, ou seja, um novo processo evolutivo z. Dessa forma, a simulação do GA explorou o espaço de alinhamentos múltiplos de sequências embaralhadas,

$$\boldsymbol{m_m} = \{x^M, y^N | z_t\}_L \tag{3.31}$$

com um conjunto de parâmetros $\boldsymbol{f_m} = \{f_x, f_y, f_{xy}\}$, onde somente as frequências de aminoácidos conjuntas entre as sequências A e B dependem de sua concatenação.

O fitness Δ_m calculado em cada geração corresponde aos acoplamentos estatístico, restrito aos contatos da interface Θ, e seguem a condição da eq. [3.21]

$$\Delta_m = [\sum_{i,j\in\Theta}\sum_{x,y\in\Omega} |\lambda^*_{x_i y_j|z}|^2]^{\frac{1}{2}} - [\sum_{i,j\in\Theta}\sum_{x,y\in\Omega} |\lambda^*_{x_i y_j|z_0}|^2]^{\frac{1}{2}} \tag{3.32}$$

onde Ω é o subconjunto de χ, que corresponde apenas aos resíduos de aminoácidos observados nas posições i e j. A otimização foi encerrada quando cada trajetória atingiu 50.000 gerações ou até a convergência ser obtida, seguindo um critério de $\dot{\Delta_m} \leq 0.005$.

Para a resolver a distribuição de sequências de maior entropia, o alinhamento múltiplo de sequências foi embaralhado ao acaso, com a restrição de que a concatenação final não apresentasse nenhuma sequência em posição nativa. Esse procedimento deve ser suficiente para cumprir a condição da equação [3.21], com a devida modificação do sinal de desigualdade, com menor custo computacional.

3.5.5 Cálculo de Energias Livres de Ligação

O sentido físico das sequências artificiais maximizadas foram avaliadas por meio do cálculo de suas energias livres de ligação, no estado nativo de interação $F_N(x^M, y^N|z)$. A energia livre de ligação no estado nativo de ligação também foi avaliada para: as sequências artificiais provenientes do MSA embaralhado ao acaso, as sequências nativas e as sequências aleatórias com composição de aminoácidos da interface similar a nativa. Por fim, para verificar a validade do Modelo de Energias Aleatórias no contexto de interação proteína-proteína, também calculamos a energia livre de ligação das sequências nativas em estados de ligação não-nativos $F_{nN}(x^M, y^N|z)$.

Para cada sequência considerada acima, foi utilizado o Modeller 9.25 (ŠALI; BLUNDELL, 1993), para modelar a interface de ligação nativa baseada na estrutura cristalográfica

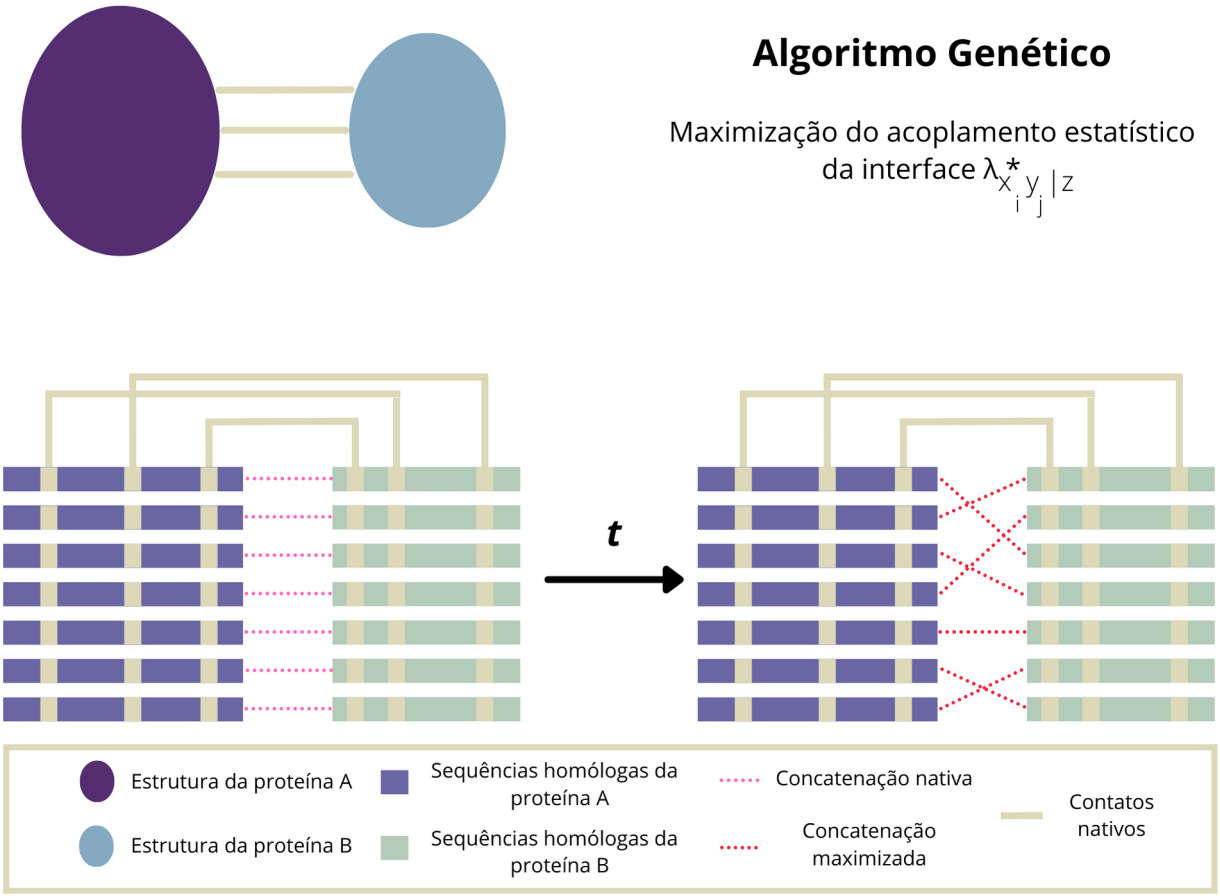

Figura 5 – Esquema representando a otimização realizada pelo Algoritmo Genético. O processo de maximização dos acoplamentos estatísticos tem inicio na concatenação nativa do alinhamento múltiplo de sequências (rosa) e procede até convergir em uma concatenação maximizada (vermelho).

do modelo homólogo apresentado na Tab. [1]. No caso das sequências que apresentavam o símbolo de *gap* (-) em algum sítio da interface, esse foi substituído pelo aminoácido Alanina (A), afim de não reduzir o tamanho da interface de ligação. Já os modelos tridimensionais das sequências nativas em estados de ligação não-nativos foram gerados por cálculos de *Docking*, no servidor online HDOCK (YAN et al., 2017). Os arquivos de coordenadas de cada cadeia apresentada na Tab. [1] foram submetidos separadamente e nenhum outro parâmetro ou restrição foi indicado. No total, foram selecionados J poses de interações não nativas, de tal forma que essas correspondessem a duas vezes o número de sequências $J = 2L$.

As estruturas geradas por meio de modelagem por homologia e *Docking* foram submetidas a uma etapa de minimização e simulação por Dinâmica Molecular. Ambas foram realizadas por meio do NAMD (PHILLIPS et al., 2005) atribuindo o campo de força CHARMM36 (HUANG; JR, 2013) como parâmetro para proteínas. A simulação dos complexos foi realizada em vácuo, com temperatura constante de 300 K, à pressão

constante de 1 atm e com constante dielétrica do meio de 80 $F\ m^{-1}$. A minimização durou 2 ps e a Dinâmica Molecular durou 20 ps. Ao final da trajetória, o último *frame* foi salvo e considerado para os cálculos posteriores.

As energias livres de ligação $F(x^M, y^N|z)$ foram avaliadas por meio da metodologia de MM/PBSA (KOLLMAN et al., 2000; LUO; SHARP, 2002), de acordo com

$$F(x^M, y^N|z) = F_{AB}(x^M, y^N|z) - F_A(x^M, y^N|z) - F_B(x^M, y^N|z) \tag{3.33}$$

onde a energia livre de ligação total $F(x^M, y^N|z)$ é a diferença entre a energia livre de ligação do complexo proteico $F_{AB}(x^M, y^N|z)$ e as energias livres de ligação das proteínas A $F_A(x^M, y^N|z)$ e B $F_B(x^M, y^N|z)$ separadamente. Essa energia livre de ligação total pode ser decomposta

$$F(x^M, y^N|z) = U - TS \approx U_{MM} + U_{PBSA} - TS \tag{3.34}$$

em um termo proveniente da energia de Mecânica Molecular (U_{MM}), uma contribuição da solvatação polar e não polar (U_{PBSA}) e a entropia conformacional (TS), que pode ser negligenciada tendo em vista que irá ser cancelada na subtração da equação 3.33.

A componente da Mecânica Molecular é definida

$$U_{MM} = U_{lig} + U_{\hat{a}ngulo} + U_{diedro} + U_{elec} + U_{vdW} \tag{3.35}$$

em termos de potenciais ligados: U_{lig} - potencial harmônico relacionado ao comprimento da ligação, $U_{\hat{a}ngulo}$ - potencial relacionado ao ângulo entre três átomos ligados, U_{diedro} - potencial relacionado à torção entre quatro átomos ligados, e potenciais não ligados: U_{elec} - potencial eletrostático relacionado com a Lei de Coulomb e U_{vdW} - interações de van der Waals relacionadas ao potencial de Lennard-Jones 12-6. O cálculo dessa componente ocorreu por meio do *plugin namdenergy* vinculado ao VMD (HUMPHREY; DALKE; SCHULTEN, 1996) e com parâmetros para proteína definido pelo campo de força CHARMM36 (HUANG; JR, 2013).

A componente de solvatação polar e não polar é definida como

$$U_{PBSA} = U^p_{PB} + U^{np}_{SA} \tag{3.36}$$

no qual a energia de solvatação polar U^p_{PB} foi calculada pelo *Adaptative Poisson-Boltzmann Solver* (APBS) por meio de um esquema de diferenças finitas (JURRUS et al., 2018), considerando uma caixa cúbica de 300Å e um *grid* de 1.0 x 1.0 x 1.0 Å^3. Seguindo a definição de superfície molecular, a constante dielétrica interna de todos os complexos foi considerada 15 $F\ m^{-1}$. A solução eletrolítica do meio foi representada por uma constante dielétrica externa de 80 $F\ m^{-1}$ e solução salina de 100 mM. Por sua vez, a energia de solvatação não polar $U^{np}_{SA} = \gamma(SASA)$ foi avaliada por meio de um modelo de Área de Superfície Acessível à Solvente (SASA) com raio de sondagem do solvente de 1.4 Å, onde a tensão de superfície foi $\gamma = 0.006\ kcal\ mol^{-1}$ Å^{-2}.

4 Resultados e Discussão

Os resultados e discussões apresentadas nessa seção são dividas em duas partes. Primeiramente, são descritas e analisadas as paisagens de *fitness* nativa, dos quatro sistemas considerados no estudo. A seguir, os resultados referentes as paisagens de *fitness* otimizada e minimizada são apresentados e comparados com a paisagem nativa.

4.1 Paisagem de *fitness* coevolutiva nativa

Como mencionado anteriormente, o presente trabalho fundamenta-se em uma afirmação: pressões seletivas são responsáveis por desviar a distribuição de probabilidades de sequências da uniformidade. Com a finalidade de verificar essa afirmação, analisamos a distribuição de probabilidade de Boltzmann nativa dos quatro sistemas considerados no estudo em relação a uma distribuição uniforme onde todas as sequências seriam equiprováveis. A figura 6 demonstra que a afirmação é verdadeira, uma vez que todas as distribuições de probabilidades nativas apresentaram menor entropia do que o caso uniforme, onde a entropia é a máxima alcançável, ou seja, as pressões seletivas atuantes sobre esses sistemas são responsáveis por moldar essas distribuições de tal forma que as sequências apresentam probabilidades diferenciais dentro do espaço de sequências composto pelas concatenações possíveis do MSA.

Visto que a composição das sequências é uma importante premissa do REM e é uma restrição imposta para a seleção das sequências artificiais, avaliamos a influência da composição da interface no cálculo dos acoplamentos estatísticos. Não encontramos qualquer relação entre a composição e o acoplamentos estatísticos, já que as três populações de sequências aleatórias, geradas com diferentes restrições de composição em relação a média das sequências nativas, apresentaram acoplamentos estatísticos similares entre si e inferiores as sequências nativas (Fig. 7). Tendo em vista que, para alguns sistemas, restringir a composição de toda a sequência não garante sequências aleatórias com composição da interface igual a das nativas, os cálculos posteriores consideraram as sequências aleatórias que foram restritas apenas quanto a composição da interface. Ainda, vale ressaltar na figura 7 que dentro do espaço de sequências do MSA, que é representado por todas as possíveis concatenações de sequências A e B do MSA, há um grande variação de valores de acoplamentos estatísticos, ainda que a maioria possua valores próximos aos do MSA nativo, além disso, para os sistemas 1EP3 e 1ZUN, podemos observar que muitas sequências desse espaço perdem composição em relação a composição média do MSA nativo.

Para cada sistema, as sequências aleatórias escolhidas a partir do conjunto de

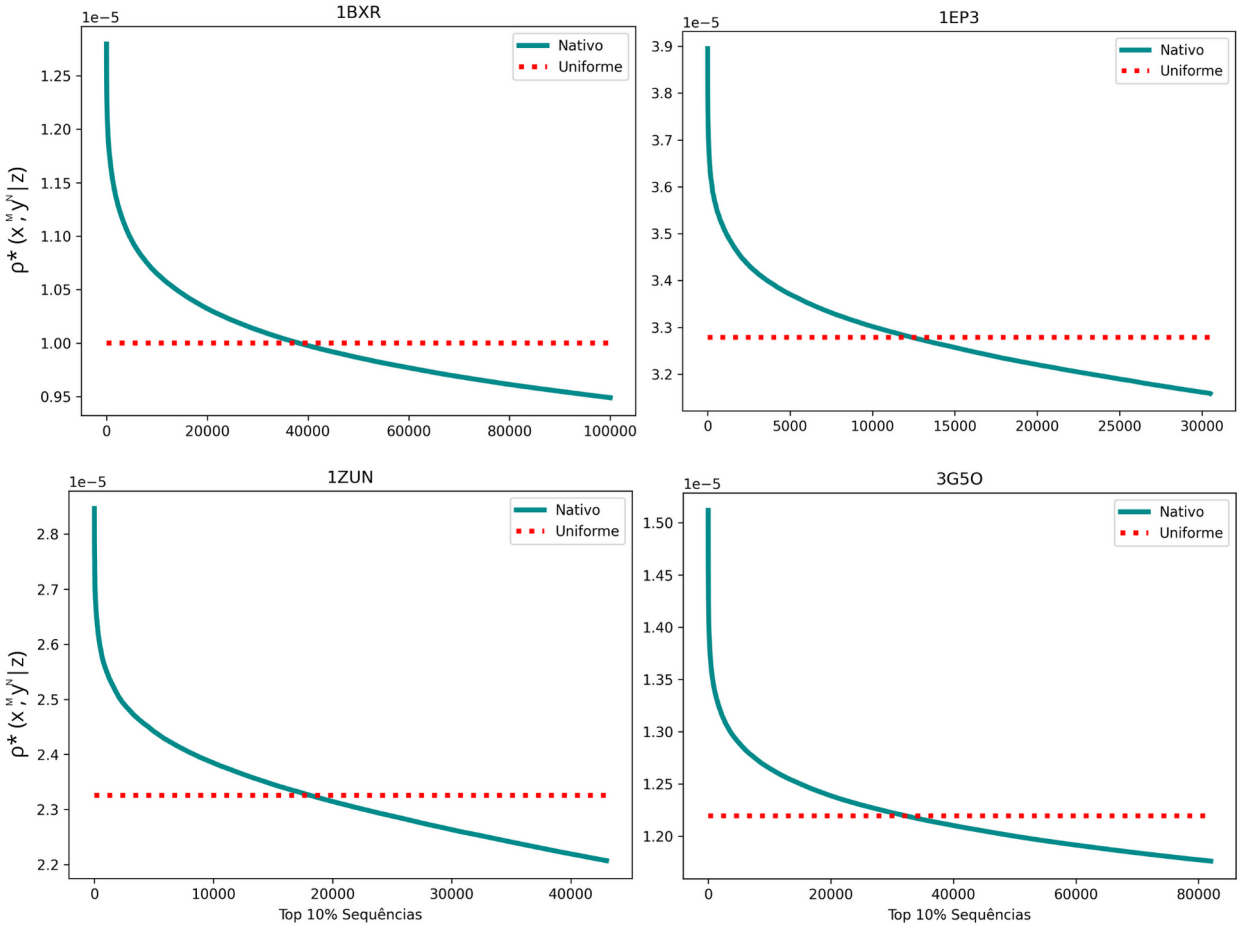

Figura 6 – Distribuições de probabilidades conjuntas de Boltzmann para as sequências nativas dos quatro sistemas considerados no estudo. Aqui foi considerado apenas as top 10% sequências mais prováveis no espaço formado por todas as concatenações possíveis advindas do alinhamento múltiplo de sequências nativo.

sequências com composição da interface igual foram modeladas na estrutura de ligação nativa do raio-x e tiveram suas energias livre de ligação avaliadas conforme os cálculos de MM/PBSA. Para corroborar o Modelo de Energias Aleatórias, geramos através de *docking* um conjunto de poses de interação não-nativas para a sequência nativa do raio-x (Fig. 8). Essas poses de interação não-nativas também tiveram suas energias de ligação livre avaliadas por meio do MM/PBSA, tal como as sequências nativas do MSA. Os resultados estão sumarizados na figura 8, onde é possível observar que para todos os sistemas, a média das energias das poses de interação não-nativa não se diferenciam da média das energias das sequências aleatórias, o que condiz com a premissa do REM. No mais, é importante notar que apenas o sistema 1BXR apresentou diferença entre a média das energias das sequências nativas em relação ao grupo das sequências aleatórias e poses não-nativas.

Sabendo que o REM aplica-se aos sistemas em questão, podemos utilizar a equação [3.17] para calcular a temperatura efetiva de seleção como o ajuste linear entre as medianas

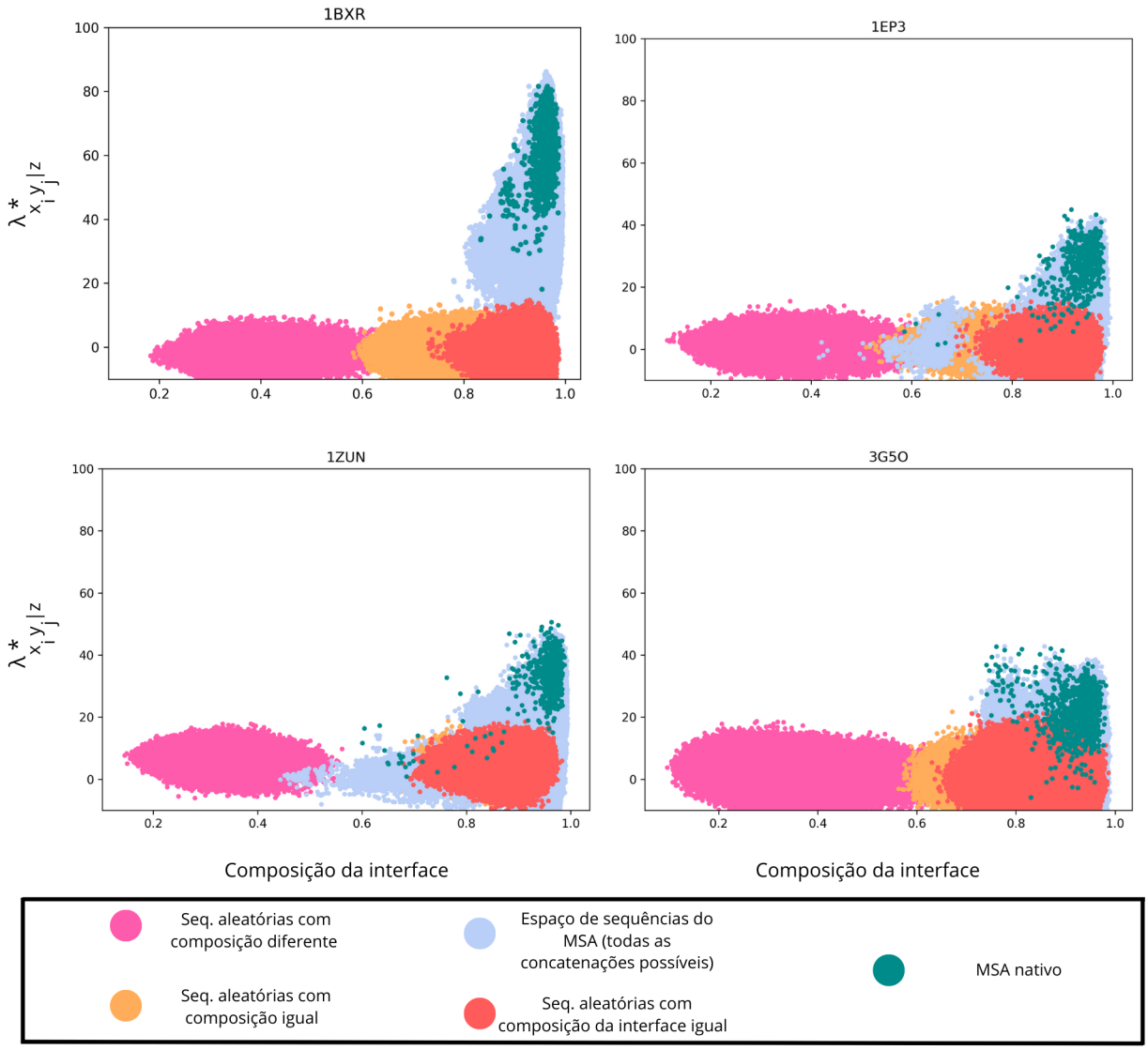

Figura 7 – Relação entre os acoplamentos estatísticos, obtidos a partir da paisagem de *fitness* nativa, e a composição da interface para os quatro sistemas analisados.

de energia livre e acoplamento estatístico, das sequências nativas em relação as sequências aleatórias (Figura 9). As temperaturas de seleção encontradas para a 1BXR, 1EP3, 1ZUN e 3G5O foram de $3740K$, $2541K$, $4957K$, $3073K$, respectivamente. Entretanto, é sabido que o método de MM/PBSA para avaliação da energia livre de ligação pode errar os valores de energia preditos em algum grau (CHEN et al., 2016), e dessa forma os valores para temperatura de seleção seriam afetados. Levando isso em consideração, realizamos uma análise para avaliar de que forma possíveis erros no cálculo de energia livre de ligação impactaria os valores da temperatura de seleção. A figura 10 mostra que ainda que a energia livre de ligação seja superestimada em 1000%, a temperatura de seleção dos quatro sistemas ainda seria maior que a média da temperatura de seleção do enovelamento de

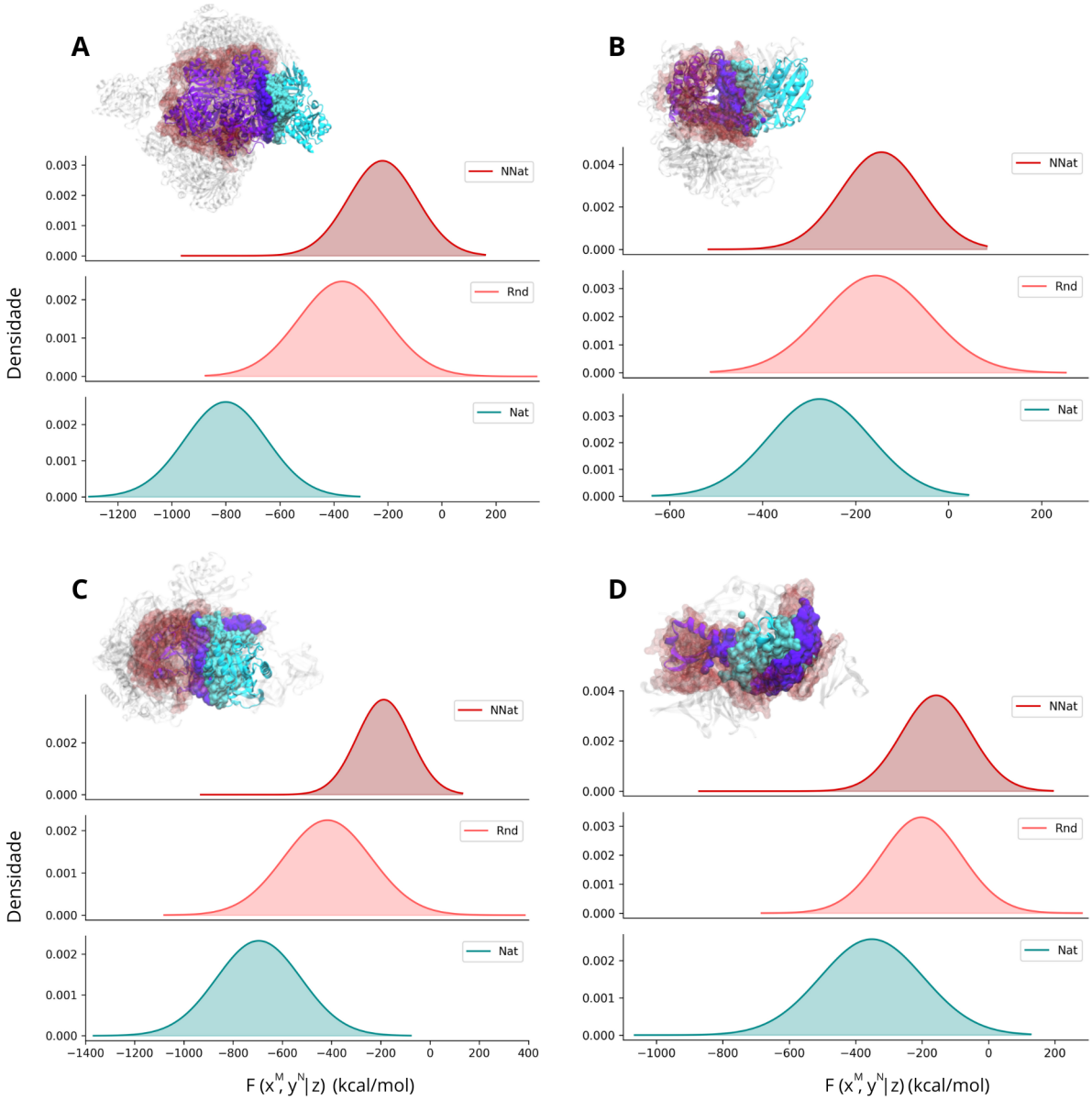

Figura 8 – Cima: Representação das estruturas tridimensionais dos complexos nas poses de interação nativas (roxo e ciano) e algumas poses de interação não-nativas geradas por *docking* (vermelho e branco). Baixo: Distribuições das energias livre de ligação obtidas pelo MM/PBSA para as sequências aleatórias (vermelho claro), nativas em pose de interação nativa (verde) e nativa em poses de interação não-nativas (vermelho escuro). A) 1BXR. B) 1EP3. C) 1ZUN. D) 3G5O.

proteínas, para os oito sistemas analisados por Morcos et al. (2014). Isso indica que a restrição imposta pelas interações de proteína-proteína na evolução proteica é menor que a restrição imposta pelo enovelamento, o que condiz com a classificação dos quatro sistemas em dímeros não-obrigatórios.

Sabendo que as restrições estruturais são um dos principais fatores que determinam a taxa evolutiva das sequências proteicas (ECHAVE; SPIELMAN; WILKE, 2016). É

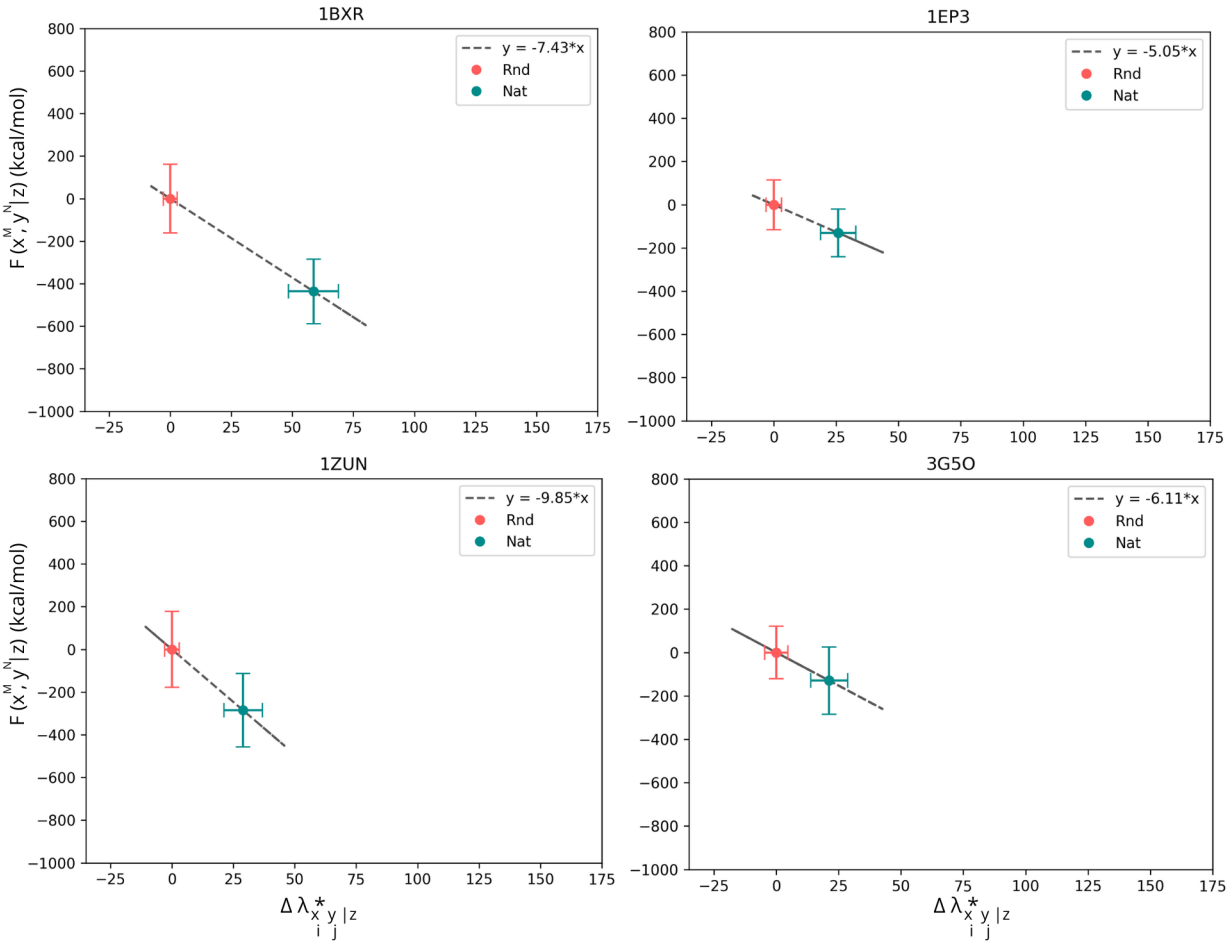

Figura 9 – O ajuste linear entre as medianas das sequências aleatórias (vermelho) e nativas (verde) para os quatro sistemas analisados. Para melhor visualização, os pontos foram transladados de tal forma que a mediana das sequências aleatórias coincidisse com a coordenada (0,0).

plausível imaginar que os monômeros A e B, de cada sistema, foram fortemente selecionados em composições de aminoácidos específicas que garantem seu enovelamento. A alteração dessa composição pode levar essas sequências a serem incapazes de se enovelar. Portanto, partindo do principio de que a evolução já selecionou a melhor composição de aminoácidos, podemos nos perguntar: "Qual é o sentido físico de sequências interólogas, com a composição de aminoácidas fixa, selecionadas a partir de paisagens artificiais?".

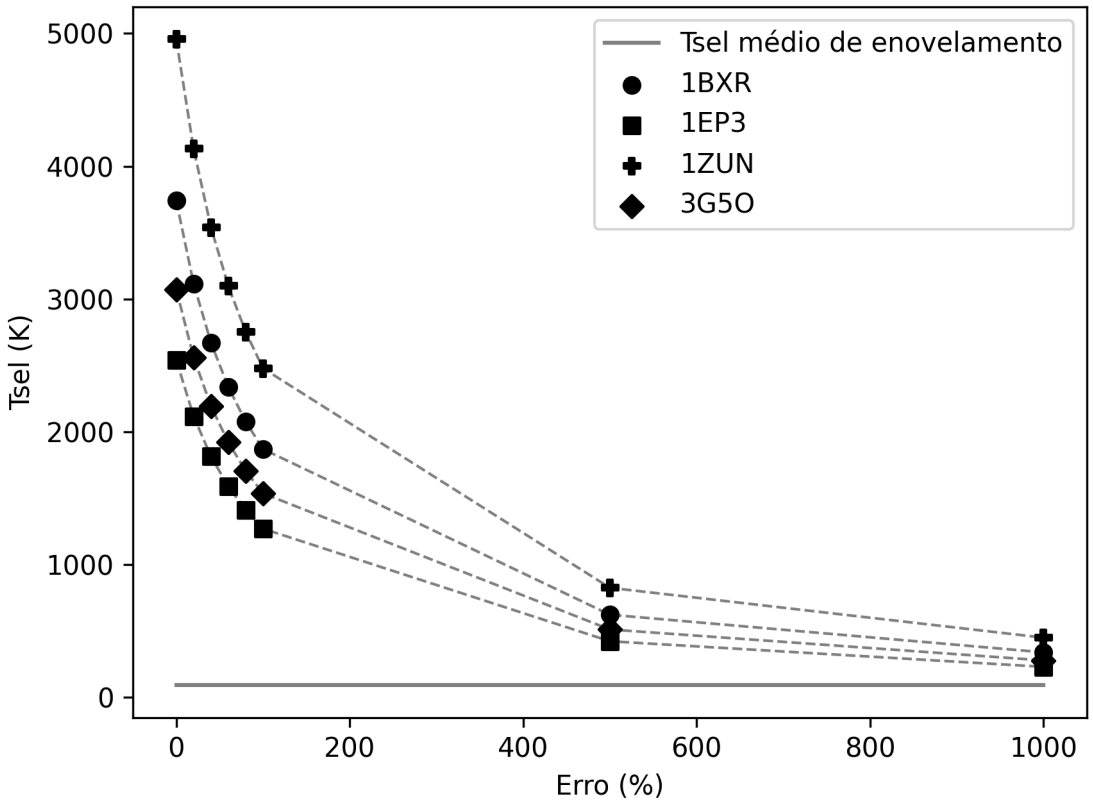

Figura 10 – Análise de erro para as temperaturas de seleção nativas. Os valores de temperatura de seleção obtidos são comparados com a média da temperatura de seleção para enovelamento descrita na literatura.

4.2 Paisagens de *fitness* coevolutivas artificiais

As paisagens artificiais consideradas nesse tópico correspondem as distribuições de probabilidade de sequências de menor e maior entropia. A distribuição menos entrópica foi gerada através de um GA que maximizou a diferença de acoplamentos estatísticos da interface através de pequenas mudanças na concatenação do MSA, iterativamente. As três trajetórias geradas pela GA para cada sistema são apresentadas na figura 11. Em todos os casos, a convergência foi alcançada seguindo o critério de parada definido como $\dot{\Delta_m} \leq 0.005$.

Por outro lado, para cada sistema, a distribuição mais entrópica foi gerada através do embaralhamento ao acaso do MSA nativo, com a restrição de que nenhuma sequência permanecesse na concatenação nativa. A figura 12 mostra que para todos os sistemas, a distribuição de probabilidade proveniente da GA é de fato menos entrópica que a nativa enquanto que a distribuição de probabilidades proveniente do embaralhamento é mais entrópica que a nativa. Logo, podemos concluir que as distribuições nativas, de fato, não foram completamente moldadas pela seleção natural, e a pressão artificial imposta pelo

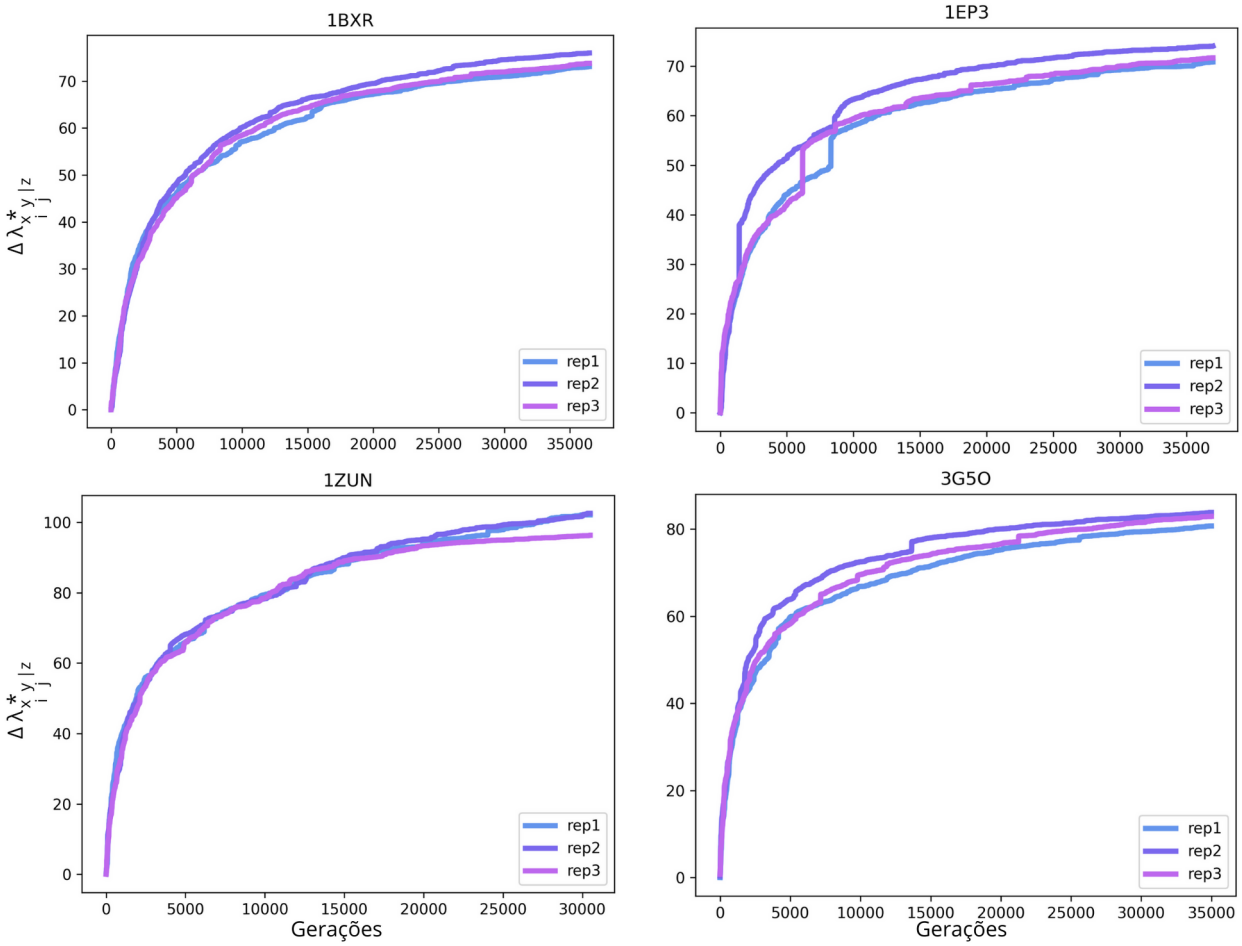

Figura 11 – Trajetórias do Algoritmo Genético para a otimização dos acoplamentos estatísticos. Para cada sistema foram realizadas três réplicas que convergiram segundo critério de parada $\dot{\Delta}_m \leq 0.005$.

GA é responsável por desviar esse processo ainda mais da uniformidade. No caso contrário, a minimização dos acoplamentos estatísticos cria um cenário evolutivo distinto e mais próximo ao caso uniforme.

A minimização e maximização dos acoplamentos estatísticos em relação aos acoplamentos nativos pode ser visualizada nas figuras 13, 14, 15 e 16, respectivamente para os sistemas 1BXR, 1EP3, 1ZUN e 3G5O. Nota-se da figura que, para todos os sistemas, a minimização dos acoplamentos estatísticos no MSA embaralhado também reduz os acoplamentos das sequências presentes no espaço de todas as concatenações possíveis e não produz qualquer efeito nas sequências aleatórias, que mantém os mesmos valores de acoplamento estatístico do espaço nativo. Já para os alinhamentos maximizados, várias sequências do espaço de concatenações possíveis melhoram seus acoplamentos e as sequências aleatórias também apresentam ligeiro ganho de acoplamento estatístico, com destaque para o sistema 3G5O, onde esse acréscimo por parte das sequências aleatórias é notável.

As sequências do MSA embaralhado, bem como as sequências não redundantes

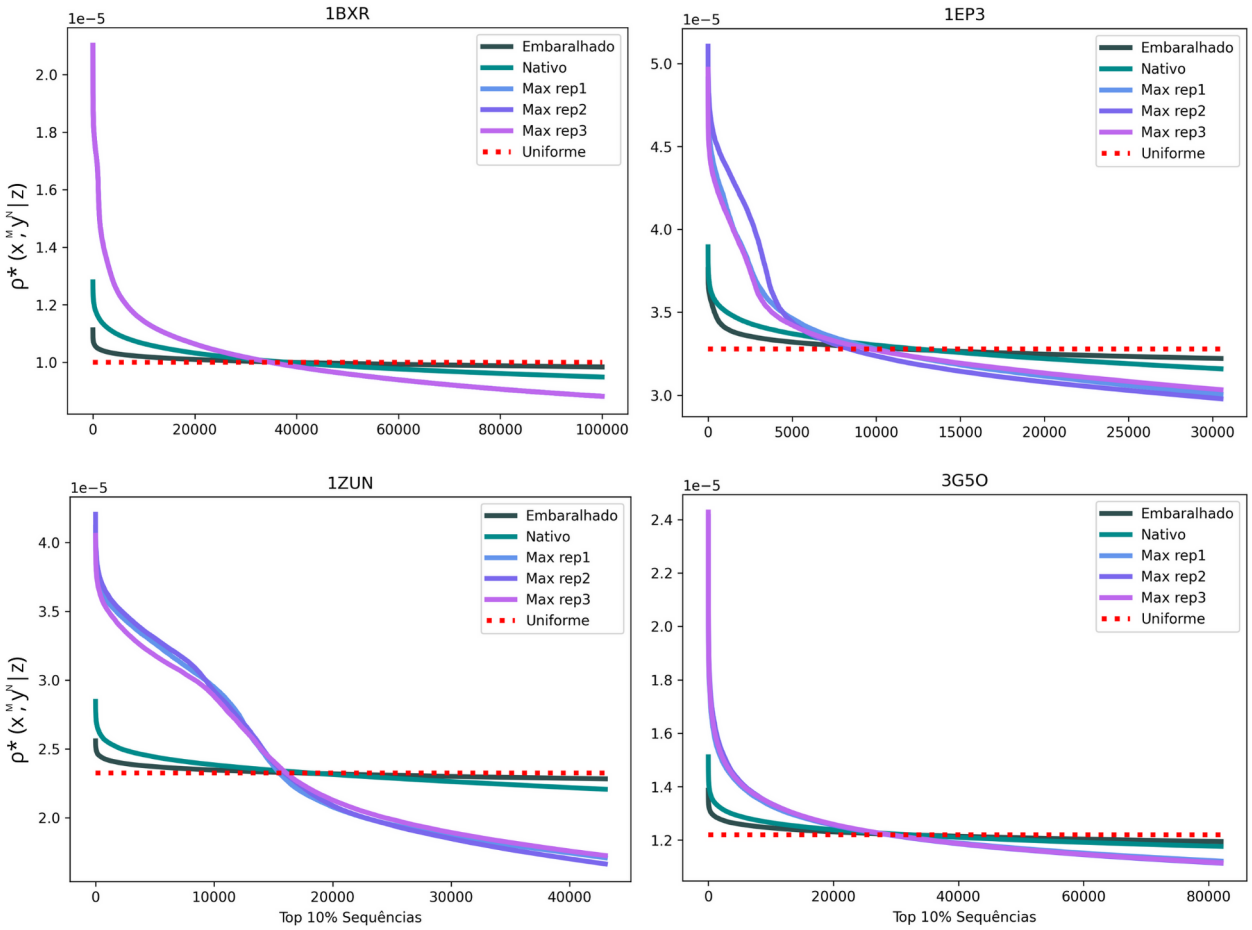

Figura 12 – Distribuições de probabilidades conjuntas de Boltzmann para as sequências nativas, maximizadas e embaralhadas, dos quatro sistemas analisados. Aqui foi considerado apenas as top 10% sequências mais prováveis no espaço formado por todas as concatenações possíveis advindas do alinhamento múltiplo de sequências nativo.

dos três MSAs maximizados tiveram suas energias livre de ligação na pose de interação nativa avaliadas por meio do MM/PBSA. A figura 17 apresenta, para todos os sistemas, os valores encontrados em comparação com as energias já descritos no tópico anterior para as sequências nativas e aleatórias, além das poses de interação não-nativas. De forma geral, não observamos diferença nas médias de energia livre de ligação das sequências maximizadas ou embaralhadas em relação a nativa. Dado o interesse em selecionar novas sequências, calculamos a similaridade das sequências artificiais dos MSAs maximizados e embaralhado, em relação as sequências do MSA nativo, por meio da distância de Hamming, seguindo o mesmo *cutoff* de distância estabelecido para esses sistemas por Pontes et al. (2021). Os resultados indicam que a maioria dessas sequências são não similares quando comparadas com as sequências do MSA nativo (Figura 18). Portanto, podemos concluir que a seleção artificial aplicada no presente trabalho, bem como o processo de embaralhamento ao acaso é suficiente para produzir novas sequências interólogas artificiais. Além do mais, vale ressaltar que não há diferença de energia livre de ligação entre as sequências similares

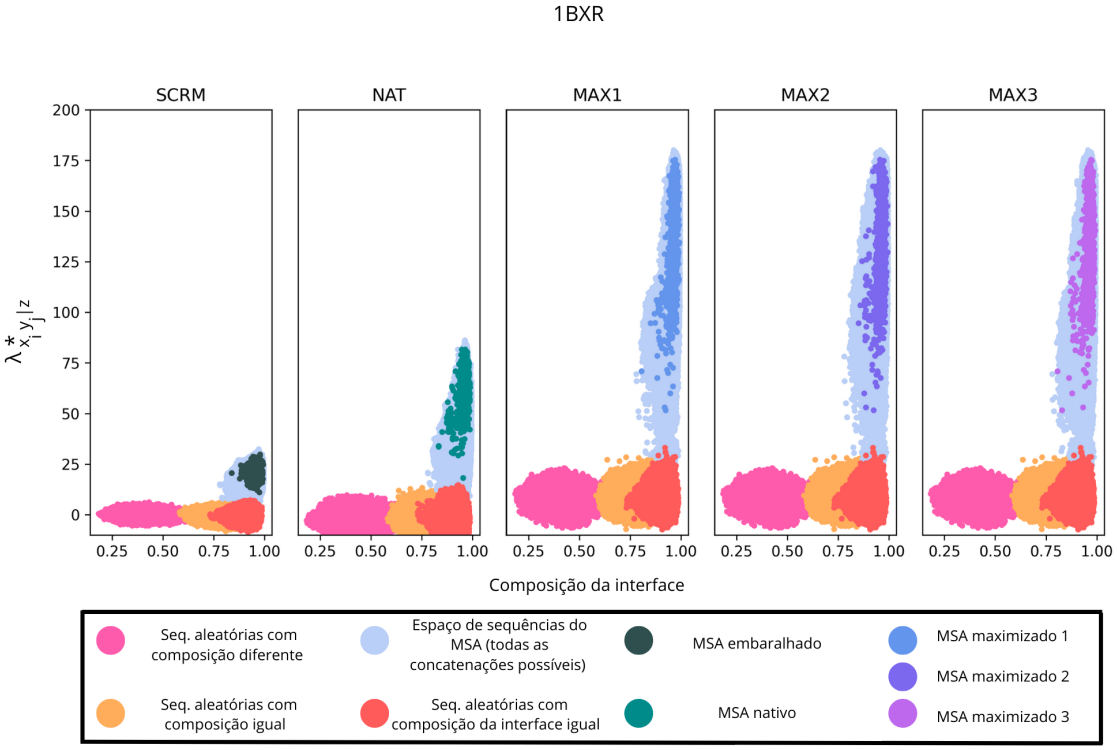

Figura 13 – Relação entre os acoplamentos estatísticos, obtidos a partir das paisagens de *fitness* nativa, embaralhada e maximizada, e a composição da interface para o sistema 1BXR.

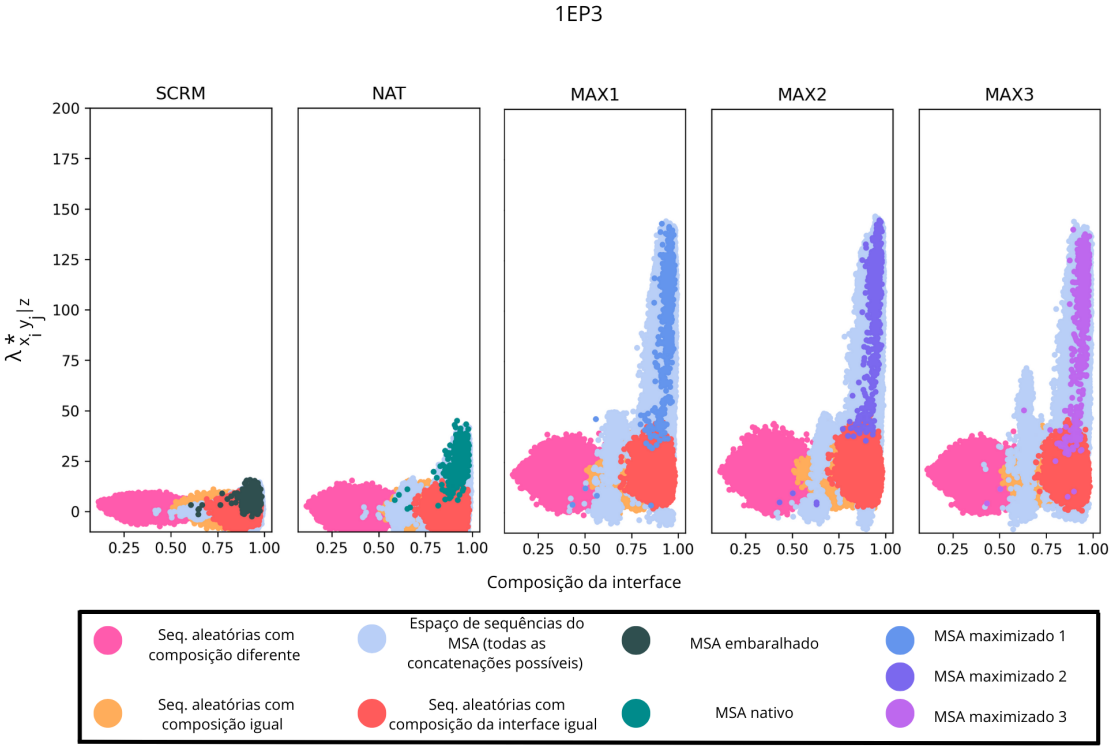

Figura 14 – Relação entre os acoplamentos estatísticos, obtidos a partir das paisagens de *fitness* nativa, embaralhada e maximizada, e a composição da interface para o sistema 1EP3.

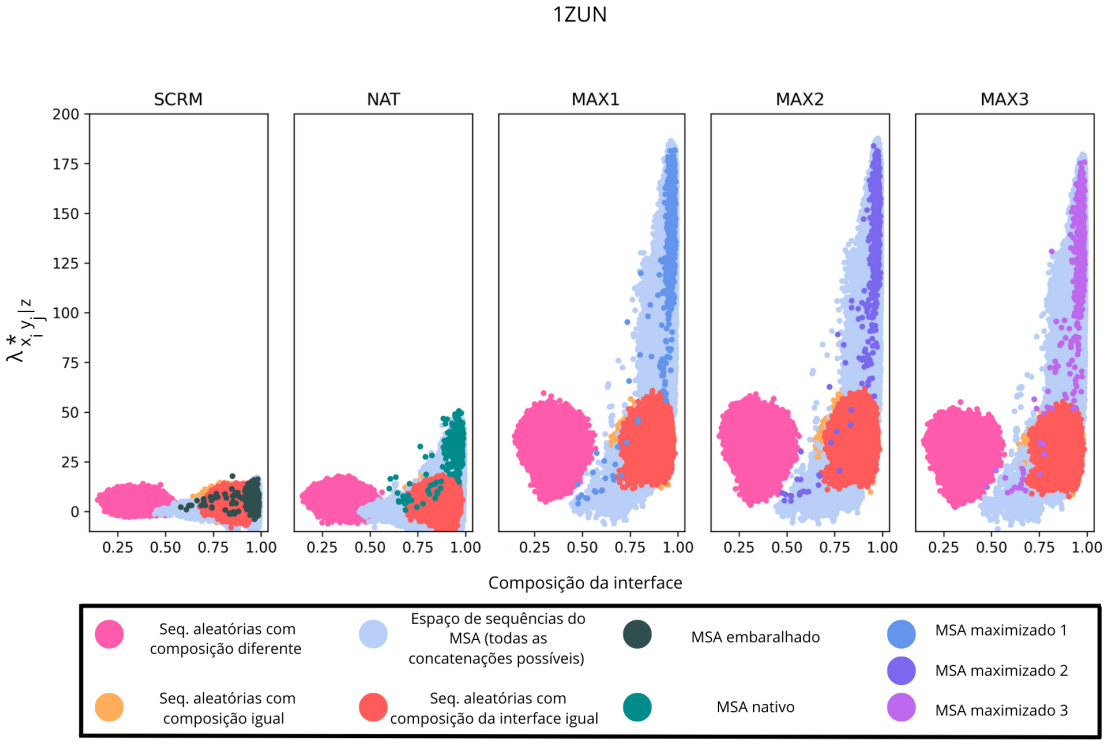

Figura 15 – Relação entre os acoplamentos estatísticos, obtidos a partir das paisagens de *fitness* nativa, embaralhada e maximizada, e a composição da interface para o sistema 1ZUN.

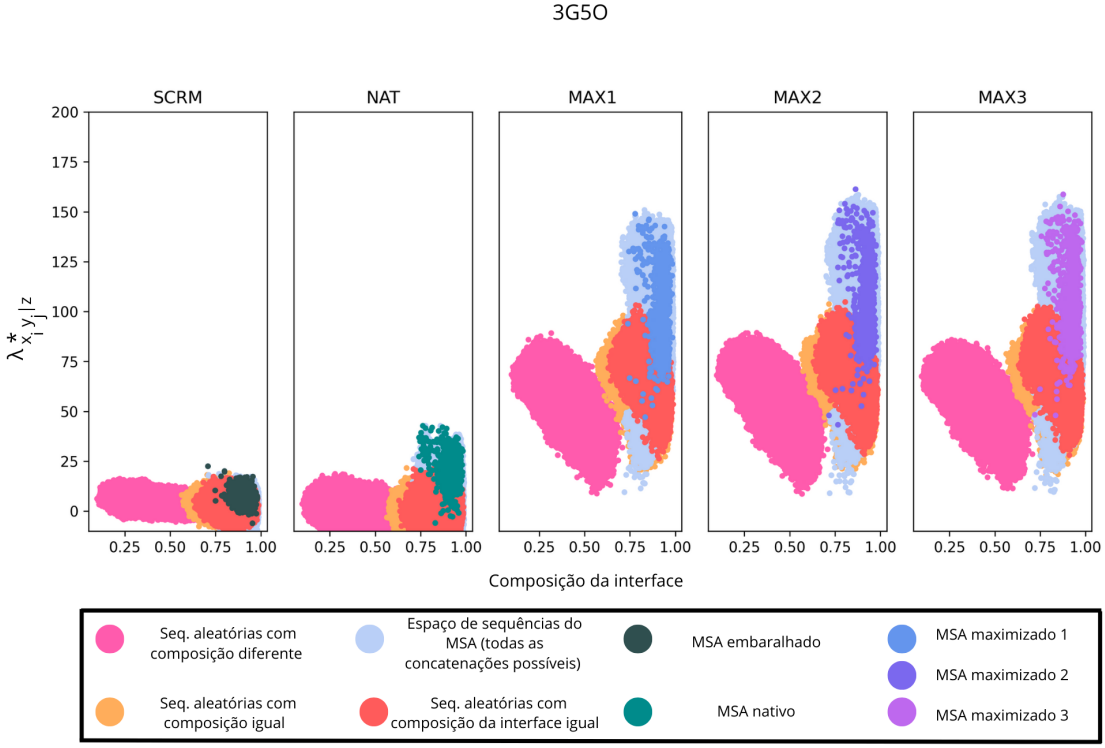

Figura 16 – Relação entre os acoplamentos estatísticos, obtidos a partir das paisagens de *fitness* nativa, embaralhada e maximizada, e a composição da interface para o sistema 3G5O.

e não-similares que compõe as distribuições de energia apresentadas na figura 18.

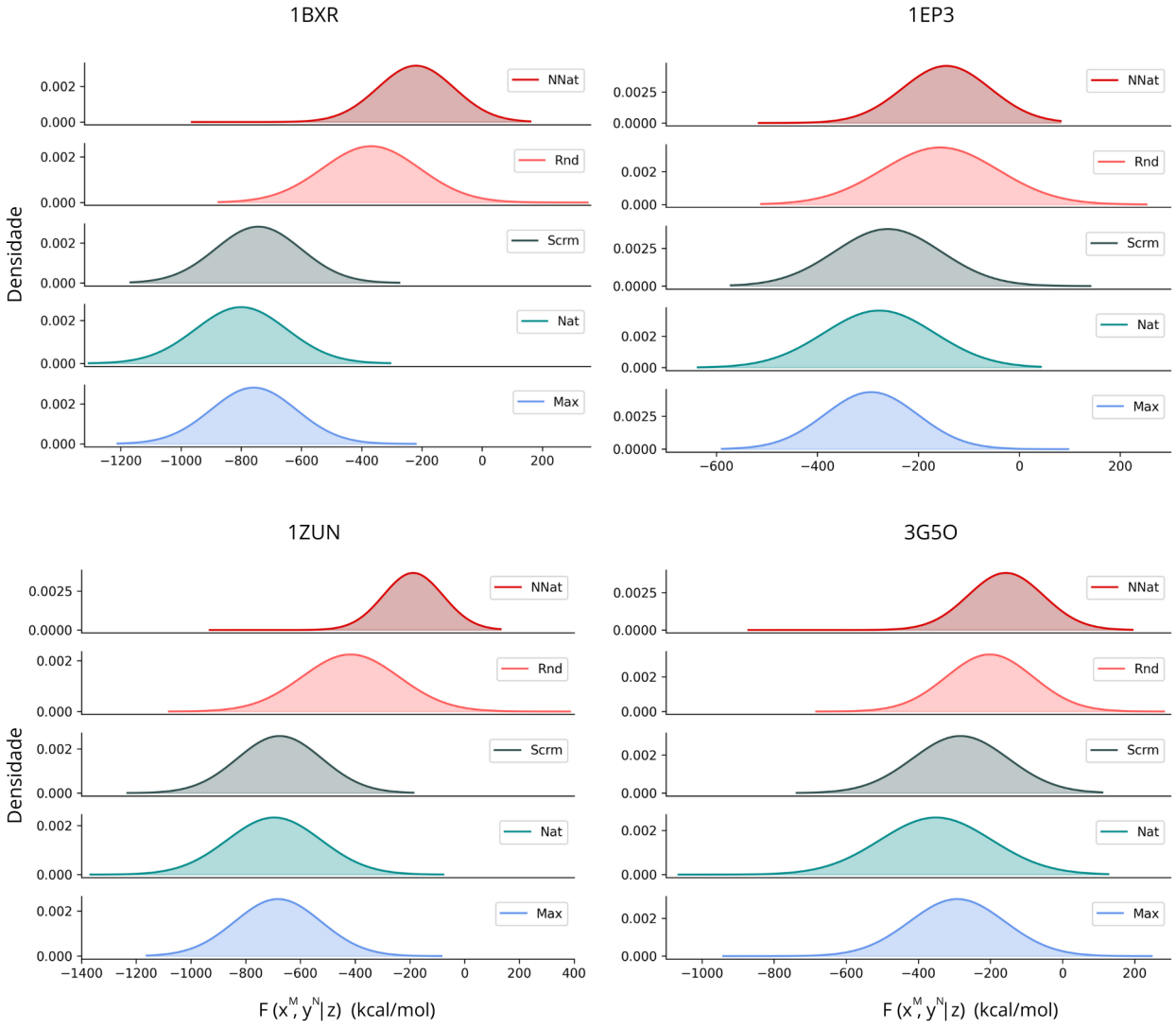

Figura 17 – Distribuições das energias livre de ligação obtidas pelo MM/PBSA para as sequências aleatórias (vermelho claro), maximizadas (azul), embaralhadas (verde escuro), nativas em pose de interação nativa (verde claro) e nativa em poses de interação não-nativas (vermelho escuro), para todos os sistemas analisados.

O ajuste linear entre as medianas de energia livre e acoplamento estatístico, das sequências maximizadas e embaralhadas em relação as sequências aleatórias é representado na figura 19, onde é comparada com o mesmo ajuste linear realizado para a mediana das sequências nativas. Para os quatro sistemas, as temperaturas de seleção encontradas para os três conjuntos de sequências considerados são apresentados na Tabela 2. Com base nesses resultados, observamos que o Algoritmo Genético aplicado selecionou novas sequências igualmente estáveis em temperatura mais fria que as sequências nativas. Já,

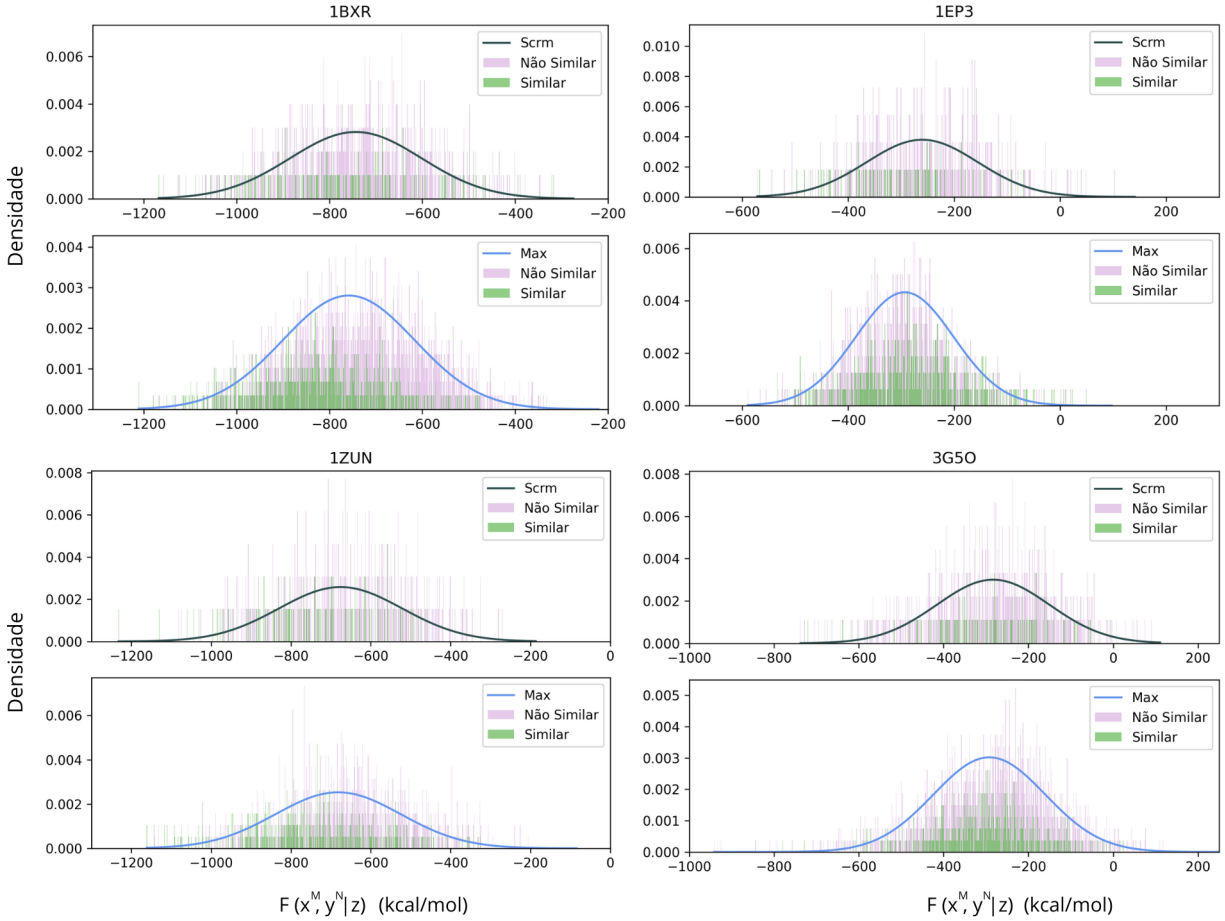

Figura 18 – Análise das sequências similares (rosa) e não-similares (verde) que compõe as distribuições de energias livre de ligação das sequências embaralhadas (verde escuro) e maximizadas (azul).

o processo de embaralhamento também envolveu a realização de sequências igualmente estáveis, mas em temperaturas mais quentes que a nativa.

Podemos concluir então que os quatro sistemas analisados possuem a capacidade de resolver a mesma afinidade da interação de proteína-proteína em diferentes temperaturas de seleção. Uma vez que esses sistemas já devem ter sofrido forte pressão seletiva em suas composições de aminoácidos para garantirem um enovelamento estável, a plasticidade evolutiva observada para a afinidade deve ser consequência da restrição imposta sobre a composição de aminoácidos fixa realizada no presente estudo. Logo, a afinidade não sofreria pressão seletiva por si só e seria apenas uma propriedade emergente desses sistemas evoluindo sobre restrições de enovelamento.

Por outro lado, é esperado que a *design* da sequência seja maior nas sequências provenientes da paisagem artificial de menor entropia, já que essas possuem a menor temperatura de seleção. Entretanto, o ganho de *design* da sequência não é similar em todos os sistemas, já que visivelmente ele é menor para a 3G5O (Fig. 19) e, isso está

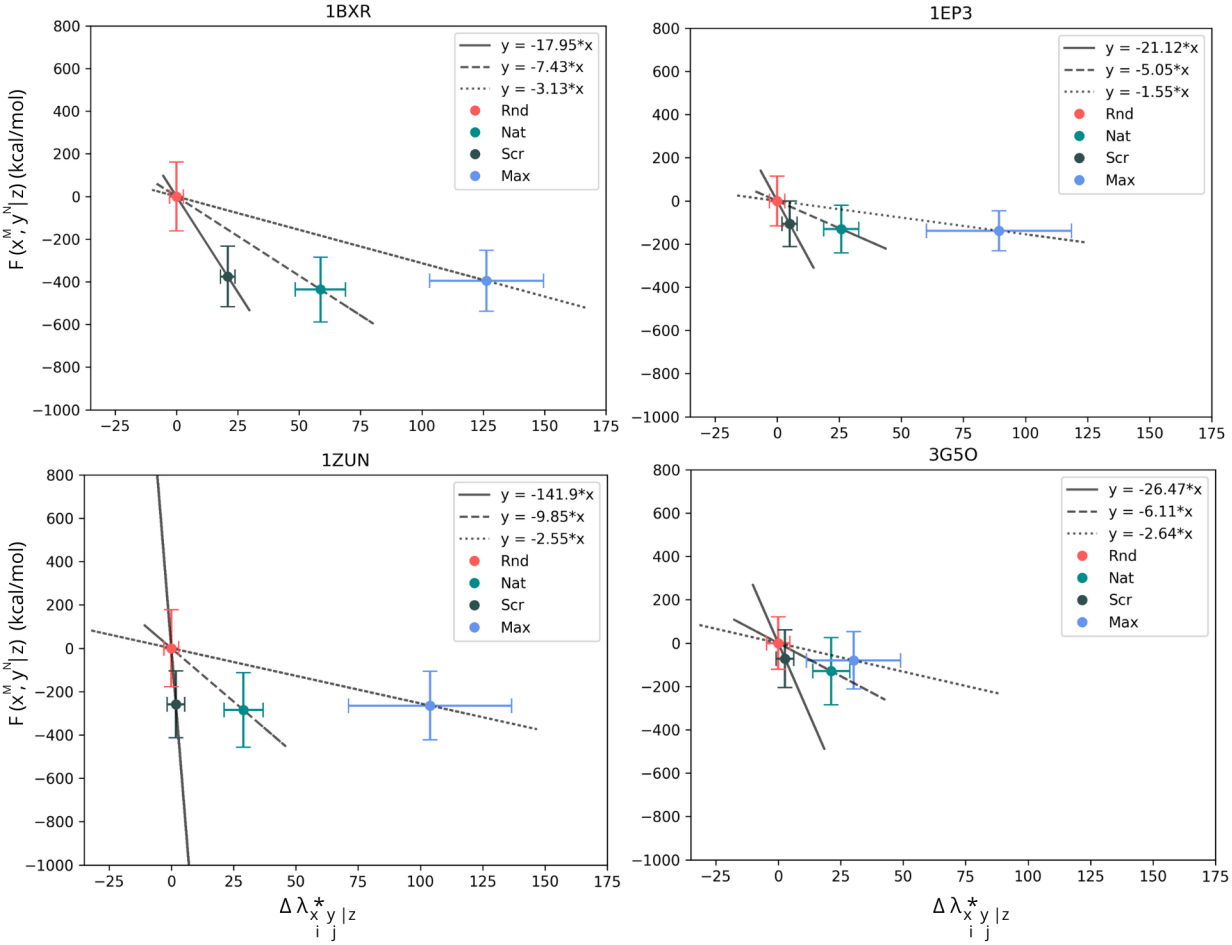

Figura 19 – O ajuste linear entre as medianas das sequências aleatórias (vermelho), nativas (verde claro), embaralhadas (verde escuro) e maximizadas (azul), para os quatro sistemas analisados. Para melhor visualização, os pontos foram transladados de tal forma que a mediana das sequências aleatórias coincidisse com a coordenada (0,0).

relacionado ao ganho notável de acoplamento estatístico por parte das sequências aleatórias na paisagem maximizada (Fig. 16). Esse achado indica que possivelmente a natureza já resolveu esse complexo toxina-antitoxina de forma bastante otimizada, e portanto os esforços de maximização aplicados não resulta em grande aumento de *design* da sequência. Nesse cenário, Aakre et al. (2015) e Lite et al. (2020) apresentam conclusões similares sobre a alta otimização em termos de especificidade já existente em outros complexos de toxina-antitoxina bacterianos.

De forma geral, os resultados aqui apresentados estão de acordo com os achados de Mintseris e Weng (2005), que explicam que a natureza dos dímeros não-obrigatórios exige uma rápida adaptação as possíveis mutações que ocorrem na interface da proteína parceira e, portanto há a necessidade de uma plasticidade de interfaces transientes. Para mais, os mesmos autores demonstraram que os dímeros não-obrigatórios sofrem menos restrição evolutiva de interações proteína-proteína, do que os dímeros obrigatórios. Considerando

Tabela 2 – Temperaturas de seleção (K) obtidas para as sequências embaralhadas, nativas e maximizadas

PDB ID	T_{sel} **scr**	T_{sel} **nat**	T_{sel} **max**
1BXR	$9034K$	$3740K$	$1573K$
1EP3	$10626K$	$2541K$	$779K$
1ZUN	$71404K$	$4957K$	$1281K$
3G5O	$13389K$	$3073K$	$1326K$

que nossos resultados evidenciam que sistemas interólogos não-obrigatórios sofrem seleção somente quanto ao *design* da sequência, mas não quanto a afinidade, então fica claro a explicação para a menor restrição evolutiva de interação quando comparado com os dímeros obrigatórios. Além disso, tendo em mente nosso resultado de que os dímeros não-obrigatórios sofrem restrição de enovelamento mais forte que de interação e que a afinidade de ligação deve ser consequência disso, é plausível imaginar que os sistemas interólogos obrigatórios representem o caso contrário, onde a restrição da interação proteína-proteína é o fator determinante para a composição de aminoácidos, e portanto esses devem ser selecionados em ambas as características de afinidade e *design* da sequência.

5 Conclusão e Perspectivas

A exploração do espaço de sequências permite não só revelar novas interações proteína-proteína que ainda não foram descobertas, ou que não foram amostradas pela seleção natural, como também compreender como as pressões e restrições seletivas determinam a evolução molecular de proteínas. Em nossa pesquisa, constatamos que é possível selecionar novas interações proteína-proteína em temperaturas de seleção mais frias ou mais quentes que a temperatura de seleção das sequências nativas e, que as sequências artificiais selecionadas apresentam diferenças apenas quanto ao *design* da sequência, mas não em relação à afinidade de ligação. Esses resultados demonstram que a evolução molecular da interação dos quatro dímeros não-obrigatórios analisados foi restrita somente pelo *design* da sequência, já que a afinidade deve ser apenas consequência da composição de aminoácidos determinada pelas restrições de enovelamento. Além disso, os resultados evidenciam a possibilidade de encontrar novas interações proteína-proteína com características iguais ou melhores que as interações existentes na natureza, o que pode representar um avanço na área do design de proteínas. Apesar dos avanços apresentados, investigações futuras são necessárias para fundamentar: as restrição de enovelamento atuante nos monômeros, os critérios utilizados para a seleção artificial das sequências, as funções de energia física utilizada e a razão para os altos valores de temperaturas de seleção. Também, para compreender como as restrições de enovelamento e interação proteína-proteína atuam em dímeros obrigatórios e, para analisar se as sequências artificiais desses sistemas apresentam diferentes comportamentos com relação as temperaturas de seleção..

Referências Bibliográficas

AAKRE, C. D.; HERROU, J.; PHUNG, T. N.; PERCHUK, B. S.; CROSSON, S.; LAUB, M. T. Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell*, Elsevier, v. 163, n. 3, p. 594–606, 2015.

ANDRADE, M.; PONTES, C.; TREPTOW, W. Coevolutive, evolutive and stochastic information in protein-protein interactions. *Computational and structural biotechnology journal*, Elsevier, v. 17, p. 1429–1435, 2019.

BITBOL, A.-F. Inferring interaction partners from protein sequences using mutual information. *PLoS computational biology*, Public Library of Science San Francisco, CA USA, v. 14, n. 11, p. e1006401, 2018.

BITBOL, A.-F.; DWYER, R. S.; COLWELL, L. J.; WINGREEN, N. S. Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 113, n. 43, p. 12180–12185, 2016.

BRYNGELSON, J. D.; WOLYNES, P. G. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of sciences*, National Acad Sciences, v. 84, n. 21, p. 7524–7528, 1987.

BURGER, L.; NIMWEGEN, E. V. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, Public Library of Science San Francisco, USA, v. 6, n. 1, p. e1000633, 2010.

CHEN, F.; LIU, H.; SUN, H.; PAN, P.; LI, Y.; LI, D.; HOU, T. Assessing the performance of the mm/pbsa and mm/gbsa methods. 6. capability to predict protein–protein binding free energies and re-rank binding poses generated by protein–protein docking. *Physical Chemistry Chemical Physics*, Royal Society of Chemistry, v. 18, n. 32, p. 22129–22139, 2016.

CHENG, R. R.; HAGLUND, E.; TIEE, N. S.; MORCOS, F.; LEVINE, H.; ADAMS, J. A.; JENNINGS, P. A.; ONUCHIC, J. N. Designing bacterial signaling interactions with coevolutionary landscapes. *PloS one*, Public Library of Science San Francisco, CA USA, v. 13, n. 8, p. e0201734, 2018.

CHENG, R. R.; NORDESJÖ, O.; HAYES, R. L.; LEVINE, H.; FLORES, S. C.; ONUCHIC, J. N.; MORCOS, F. Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. *Molecular biology and evolution*, Oxford University Press, v. 33, n. 12, p. 3054–3064, 2016.

CHI, H.; ZHOU, Q.; TUTOL, J. N.; PHELPS, S. M.; LEE, J.; KAPADIA, P.; MORCOS, F.; DODANI, S. C. Coupling a live cell directed evolution assay with coevolutionary landscapes to engineer an improved fluorescent rhodopsin chloride sensor. *ACS Synthetic Biology*, v. 11, n. 4, p. 1627–1638, 2022. PMID: 35389621.

CHOTHIA, C.; JANIN, J. Principles of protein–protein recognition. *Nature*, Nature Publishing Group, v. 256, n. 5520, p. 705–708, 1975.

CUSICK, M. E.; KLITGORD, N.; VIDAL, M.; HILL, D. E. Interactome: gateway into systems biology. *Human molecular genetics*, Oxford University Press, v. 14, n. suppl_2, p. R171–R181, 2005.

DERRIDA, B. Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, APS, v. 45, n. 2, p. 79, 1980.

DERRIDA, B. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, APS, v. 24, n. 5, p. 2613, 1981.

ECHAVE, J.; SPIELMAN, S. J.; WILKE, C. O. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, Nature Publishing Group, v. 17, n. 2, p. 109–121, 2016.

EHRLICH, P. R.; RAVEN, P. H. Butterflies and plants: a study in coevolution. *Evolution*, JSTOR, v. 18, n. 4, p. 586–608, 1964.

FIGLIUZZI, M.; JACQUIER, H.; SCHUG, A.; TENAILLON, O.; WEIGT, M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Molecular biology and evolution*, Oxford University Press, v. 33, n. 1, p. 268–280, 2016.

GARCIA, L. G.; TREPTOW, W. L.; ARAÚJO, A. F. P. de. Folding simulations of a three-dimensional protein model with a nonspecific hydrophobic energy function. *Physical Review E*, APS, v. 64, n. 1, p. 011912, 2001.

GÖBEL, U.; SANDER, C.; SCHNEIDER, R.; VALENCIA, A. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 18, n. 4, p. 309–317, 1994.

GUEUDRÉ, T.; BALDASSI, C.; ZAMPARO, M.; WEIGT, M.; PAGNANI, A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 113, n. 43, p. 12186–12191, 2016.

HOPF, T. A.; SCHÄRFE, C. P.; RODRIGUES, J. P.; GREEN, A. G.; KOHLBACHER, O.; SANDER, C.; BONVIN, A. M.; MARKS, D. S. Sequence co-evolution gives 3d contacts and structures of protein complexes. *elife*, eLife Sciences Publications Limited, v. 3, p. e03430, 2014.

HUANG, J.; JR, A. D. M. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *Journal of computational chemistry*, Wiley Online Library, v. 34, n. 25, p. 2135–2145, 2013.

HUMPHREY, W.; DALKE, A.; SCHULTEN, K. Vmd: visual molecular dynamics. *Journal of molecular graphics*, Elsevier, v. 14, n. 1, p. 33–38, 1996.

JANIN, J. Elusive affinities. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 21, n. 1, p. 30–39, 1995.

JANIN, J. Quantifying biological specificity: the statistical mechanics of molecular recognition. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 25, n. 4, p. 438–445, 1996.

JANZEN, D. H. When is it coevolution?. *Evolution*, JSTOR, v. 34, n. 3, p. 611–612, 1980.

JAYNES, E. T. Information theory and statistical mechanics. *Physical review*, APS, v. 106, n. 4, p. 620, 1957.

JONES, D. T.; BUCHAN, D. W.; COZZETTO, D.; PONTIL, M. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, Oxford University Press, v. 28, n. 2, p. 184–190, 2012.

JUAN, D. D.; PAZOS, F.; VALENCIA, A. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, Nature Publishing Group, v. 14, n. 4, p. 249–261, 2013.

JUMPER, J.; EVANS, R.; PRITZEL, A.; GREEN, T.; FIGURNOV, M.; RONNEBERGER, O.; TUNYASUVUNAKOOL, K.; BATES, R.; ŽÍDEK, A.; POTAPENKO, A. et al. Highly accurate protein structure prediction with alphafold. *Nature*, Nature Publishing Group, v. 596, n. 7873, p. 583–589, 2021.

JURRUS, E.; ENGEL, D.; STAR, K.; MONSON, K.; BRANDI, J.; FELBERG, L. E.; BROOKES, D. H.; WILSON, L.; CHEN, J.; LILES, K. et al. Improvements to the apbs biomolecular solvation software suite. *Protein Science*, Wiley Online Library, v. 27, n. 1, p. 112–128, 2018.

KOLLMAN, P. A.; MASSOVA, I.; REYES, C.; KUHN, B.; HUO, S.; CHONG, L.; LEE, M.; LEE, T.; DUAN, Y.; WANG, W. et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research*, ACS Publications, v. 33, n. 12, p. 889–897, 2000.

KORTEMME, T.; BAKER, D. Computational design of protein–protein interactions. *Current opinion in chemical biology*, Elsevier, v. 8, n. 1, p. 91–97, 2004.

LEVY, E. D. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of molecular biology*, Elsevier, v. 403, n. 4, p. 660–670, 2010.

LEVY, Y.; WOLYNES, P. G.; ONUCHIC, J. N. Protein topology determines binding mechanism. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 101, n. 2, p. 511–516, 2004.

LITE, T.-L. V.; GRANT, R. A.; NOCEDAL, I.; LITTLEHALE, M. L.; GUO, M. S.; LAUB, M. T. Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *Elife*, eLife Sciences Publications, Ltd, v. 9, 2020.

LOVELL, S. C.; ROBERTSON, D. L. An integrated view of molecular coevolution in protein–protein interactions. *Molecular biology and evolution*, Oxford University Press, v. 27, n. 11, p. 2567–2575, 2010.

LUO, H.; SHARP, K. On the calculation of absolute macromolecular binding free energies. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 99, n. 16, p. 10399–10404, 2002.

MINTSERIS, J.; WENG, Z. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 102, n. 31, p. 10930–10935, 2005.

MODE, C. J. A mathematical model for the co-evolution of obligate parasites and their hosts. *Evolution*, JSTOR, p. 158–165, 1958.

MORCOS, F.; ONUCHIC, J. N. The role of coevolutionary signatures in protein interaction dynamics, complex inference, molecular recognition, and mutational landscapes. *Current opinion in structural biology*, Elsevier, v. 56, p. 179–186, 2019.

MORCOS, F.; PAGNANI, A.; LUNT, B.; BERTOLINO, A.; MARKS, D. S.; SANDER, C.; ZECCHINA, R.; ONUCHIC, J. N.; HWA, T.; WEIGT, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 108, n. 49, p. E1293–E1301, 2011.

MORCOS, F.; SCHAFER, N. P.; CHENG, R. R.; ONUCHIC, J. N.; WOLYNES, P. G. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 111, n. 34, p. 12408–12413, 2014.

NOOREN, I. M.; THORNTON, J. M. Diversity of protein–protein interactions. *The EMBO journal*, John Wiley & Sons, Ltd, v. 22, n. 14, p. 3486–3492, 2003.

ONUCHIC, J. N.; LUTHEY-SCHULTEN, Z.; WOLYNES, P. G. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 48, n. 1, p. 545–600, 1997.

OVCHINNIKOV, S.; KAMISETTY, H.; BAKER, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *elife*, eLife Sciences Publications Limited, v. 3, p. e02030, 2014.

OVCHINNIKOV, S.; PARK, H.; VARGHESE, N.; HUANG, P.-S.; PAVLOPOULOS, G. A.; KIM, D. E.; KAMISETTY, H.; KYRPIDES, N. C.; BAKER, D. Protein structure determination using metagenome sequence data. *Science*, American Association for the Advancement of Science, v. 355, n. 6322, p. 294–298, 2017.

PAZOS, F.; HELMER-CITTERICH, M.; AUSIELLO, G.; VALENCIA, A. Correlated mutations contain information about protein-protein interaction. *Journal of molecular biology*, Elsevier, v. 271, n. 4, p. 511–523, 1997.

PHILLIPS, J. C.; BRAUN, R.; WANG, W.; GUMBART, J.; TAJKHORSHID, E.; VILLA, E.; CHIPOT, C.; SKEEL, R. D.; KALE, L.; SCHULTEN, K. Scalable molecular dynamics with namd. *Journal of computational chemistry*, Wiley Online Library, v. 26, n. 16, p. 1781–1802, 2005.

PONTES, C.; ANDRADE, M.; FIOROTE, J.; TREPTOW, W. Trivial and nontrivial error sources account for misidentification of protein partners in mutual information approaches. *Scientific reports*, Nature Publishing Group, v. 11, n. 1, p. 1–11, 2021.

RUSS, W. P.; FIGLIUZZI, M.; STOCKER, C.; BARRAT-CHARLAIX, P.; SOCOLICH, M.; KAST, P.; HILVERT, D.; MONASSON, R.; COCCO, S.; WEIGT, M. et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, American Association for the Advancement of Science, v. 369, n. 6502, p. 440–445, 2020.

ŠALI, A.; BLUNDELL, T. L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, Elsevier, v. 234, n. 3, p. 779–815, 1993.

SANTOS, R. N. D.; MORCOS, F.; JANA, B.; ANDRICOPULO, A. D.; ONUCHIC, J. N. Dimeric interactions and complex formation using direct coevolutionary couplings. *Scientific reports*, Nature Publishing Group, v. 5, n. 1, p. 1–10, 2015.

SCHUG, A.; WEIGT, M.; ONUCHIC, J. N.; HWA, T.; SZURMANT, H. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 106, n. 52, p. 22124–22129, 2009.

SHANNON, C. E. A mathematical theory of communication. *The Bell system technical journal*, Nokia Bell Labs, v. 27, n. 3, p. 379–423, 1948.

SUŁKOWSKA, J. I.; MORCOS, F.; WEIGT, M.; HWA, T.; ONUCHIC, J. N. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 109, n. 26, p. 10340–10345, 2012.

THOMPSON, J. N. *The Coevolutionary Process*. [S.l.]: University of Chicago Press, 1994.

TIAN, P.; LOUIS, J. M.; BABER, J. L.; ANIANA, A.; BEST, R. B. Co-evolutionary fitness landscapes for sequence design. *Angewandte Chemie International Edition*, Wiley Online Library, v. 57, n. 20, p. 5674–5678, 2018.

TREPTOW, W. L.; BARBOSA, M. A. A.; GARCIA, L. G.; ARAUJO, A. F. Pereira de. Non-native interactions, effective contact order, and protein folding: a mutational investigation with the energetically frustrated hydrophobic model. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 49, n. 2, p. 167–180, 2002.

VALEN, L. V. A new evolutionary law. *Evol theory*, v. 1, p. 1–30, 1973.

WANG, J.; VERKHIVKER, G. M. Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. *Physical review letters*, APS, v. 90, n. 18, p. 188101, 2003.

WEIGT, M.; WHITE, R. A.; SZURMANT, H.; HOCH, J. A.; HWA, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 106, n. 1, p. 67–72, 2009.

WELLS, J. A.; MCCLENDON, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*, Nature Publishing Group, v. 450, n. 7172, p. 1001–1009, 2007.

XIE, W. J.; ASADI, M.; WARSHEL, A. Enhancing computational enzyme design by a maximum entropy strategy. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 119, n. 7, p. e2122355119, 2022.

YAN, Y.; ZHANG, D.; ZHOU, P.; LI, B.; HUANG, S.-Y. Hdock: a web server for protein–protein and protein–dna/rna docking based on a hybrid strategy. *Nucleic acids research*, Oxford University Press, v. 45, n. W1, p. W365–W373, 2017.