Universidade de Brasília
Instituto de Biologia
Laboratório de Biologia Teórica e Computacional

Investigação Estatística de Pareamentos de Proteínas em Abordagens Coevolutivas

José Antonio Fiorote

Orientador: Werner Treptow

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

Brasília - DF, Fevereiro de 2025

”A maior riqueza do homem é a sua incompletude.
Nesse ponto sou abastado.”

Manoel de Barros

Agradecimentos

Sinto-me enormemente agradecido a todos que contribuíram para a realização deste trabalho, seja com palavras de incentivo ou de forma mais direta. Em especial, gostaria de agradecer:

À Zélia, minha mãe, por sua ternura e apoio incondicional.

À Flávia, minha namorada, cujo carinho e paciência foram essenciais durante os últimos três anos.

Aos amigos que passaram pelo LBTC, cujos passos foram um guia para mim, em particular João Nunes, que desenvolveu parte deste trabalho comigo, e Caio Souza, que, mesmo em terras lusitanas, reservou um momento para as caferências.

Ao meu orientador, Werner Treptow, por sua paciência e cuidado, sempre presente e disposto a esclarecer as minhas dúvidas.

Aos membros da banca, que dispuseram de seu tempo para avaliar este trabalho.

Enfim, à Universidade de Brasília, ao Programa de Pós Graduação em Biologia Molecular e à Capes, pelo fomento e estrutura indispensáveis para a realização do trabalho.

Resumo

As interações físicas em proteínas são mantidas ao longo da evolução por meio de mutações compensatórias. Conforme extensivamente investigado nos últimos anos, esse sinal coevolutivo é de grande relevância para a resolução *ab initio* de parceiros proteicos específicos com base em alinhamentos múltiplos de sequências (MSAs). Neste trabalho, examinamos as condições estatísticas dos sinais de coevolução que permitem previsões algorítmicas de parceiros proteicos com base em sequências de aminoácidos. Apresentamos aqui um modelo estocástico do algoritmo genético que prevê o número de parceiros proteicos corretos com base em informações de coevolução. O modelo define as probabilidades de estado usando uma mistura de distribuições normais e de Poisson, com parâmetros de entrada que incluem o número total de sequências proteicas do sistema (M), a diferença de informação coevolutiva (α) e a variância da informação coevolutiva em sistemas com parceiros completamente embaralhados (σ_0^2). A análise do modelo aponta que estratégias algorítmicas baseadas na maximização da informação coevolutiva não são eficientes para encontrar os parceiros nativos em sistemas de proteínas com muitas sequências ($M \geq 100$), mas as taxas de verdadeiros positivos (TPs) podem ser consideravelmente maiores ao desconsiderar erros cometidos entre sequências semelhantes. Essa abordagem nos permite realizar uma classificação prévia de famílias de proteínas em que os parceiros podem ser previstos de forma confiável ao ignorar erros triviais de similaridade entre sequências.

Abstract

Physical interactions in proteins are maintained throughout evolution through compensatory mutations. As extensively investigated in recent years, this coevolutionary signal is highly relevant for the *ab initio* resolution of specific protein partners based on multiple sequence alignments (MSAs). In this work, we examine the statistical conditions of coevolutionary signals that enable algorithmic predictions of protein partners based on amino acid sequences. Here, we present a stochastic model of the genetic algorithm that predicts the number of correct protein partners based on coevolutionary information. The model defines state probabilities using a mixture of normal and Poisson distributions, with input parameters including the total number of protein sequences in the system (M), the difference in coevolutionary information (α), and the variance of coevolutionary information in systems with completely shuffled partners (σ_0^2). Model analysis indicates that algorithmic strategies based on maximizing coevolutionary information are not efficient for finding native partners in protein systems with many sequences ($M \geq 100$), but true positive rates (TPs) can be considerably higher when disregarding errors made among similar sequences. This approach allows us to perform a preliminary classification of protein families in which partners can be reliably predicted by ignoring trivial similarity-based errors between sequences.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

- **DNA** - *Deoxyribonucleic Acid*;
- **MSA** - *Multiple Sequence Alignment*;
- **PDF** - *Probability Density Function*;
- **pH** - Potencial Hidrogeniônico;
- **TP** - *True Pairs rate.*

Lista de Variáveis

- I - Informação mútua;
- M - Número de sequências de um sistema de proteínas;
- n - Número de pareamentos nativos formados em um sistema de proteínas;
- I_n - Informação mútua média para um dado número de pareamentos nativos;
- I' - Informação mútua de um sistema em que todos os pareamentos são nativos;
- σ_n^2 - Variância para um dado número de pareamentos nativos;
- α - Dado por $(I - I_0)$;
- c - Estado do modelo definido por (n, I);
- c_k - Estado absorvente do modelo que determina o fim da trajetória;
- TP - Dado por n/M;
- m - Número de pareamentos com parceiros similares;
- n' - Dado por $(n + m)$;
- I^* - I alcançado em uma simulação de otimização de I;
- α^* - Dado por $(I^* - I_0)$.

Glossário

- **Alelo** - (do grego *allelon* = de um a outro) Gene alternativo presente na cópia do cromossomo e que codifica a mesma característica;
- **Algoritmo** - Sequência de regras e instruções lógicas para realizar uma tarefa específica ou resolver um dado problema;
- **Clado** - Grupo de organismos (ou qualquer organização biológica) que compartilha um ancestral comum;
- ***Crossover*** - No contexto do algoritmo genético, é a troca de posições do genoma de um indivíduo. Neste trabalho, o indivíduo é representado por um sistema, o genoma pela rede r e uma posição do genoma é o pareamento entre duas sequências;
- **Filogenia** - (do grego *phylon* = tribo, raça + *geneia* = origem) Refere-se à história evolutiva de uma espécie, muitas vezes apresentada na forma de dendograma;
- ***Fitness*** - No contexto da teoria evolutiva, é a capacidade de sobrevivência e reprodução de uma organização biológica em seu meio;
- ***Gap*** - No contexto do alinhamento de proteínas, é a lacuna inserida nas sequências com o intuito de maximizar a semelhança entre as posições de aminoácidos;
- **Gene** - (do grego *genos* = descendência) Trecho da sequência do DNA que que codifica uma sequência polipeptídica;

- **Heurística** - Entende-se como a técnica que busca resolver problemas complexos com soluções aproximadas, geralmente utilizada quando não há métodos exatos viáveis;

- **Pares Nativos** - No escopo deste estudo, refere-se ao conjunto de proteínas ou sequências que interagem em condições celulares naturais;

- **Proteínas Homólogas** - Proteínas que evoluíram a partir de um ancestral comum e mantêm similaridades de estrutura e função. As proteínas homólogas pordem ser do tipo **ortólogas**, caso sua divergência tenha origem em um episódio de especiação, ou do tipo **parálogas**, caso tenham se diversificado a partir de um evento de duplicação gênica dentro da mesma espécie;

- ***Rencontres*** - Na matemática, ente conceito está relacionado ao problema de permutações com coincidências. No contexto deste trabalho, os *rencontres* são as ocorrências de pareamentos corretos entre pares de proteínas;

- **Sequência Primária** - Sequência de aminoácidos de uma proteína;

- **Variável Estocástica** - Quantidade de valor determinado por processos aleatórios e que varia de acordo com a distribuição de probabilidade a ela associada.

CAPÍTULO 1

Introdução

A evolução elucida a biologia[1]. A imensa variedade da vida só pode ser explicada por um processo de descendência com modificação, em que novas espécies emergem de formas ancestrais através de uma série de adaptações cumulativas, regidas por seleção natural. Essa ideia, que teve Charles Darwin como principal expoente[2], revolucionou as ciências naturais e deu uma dimensão nova à interação entre os organismos[3].

Quase um século após a publicação das teorias de Darwin, o advento da síntese evolutiva moderna[4] e da biologia molecular desvendou os mecanismos adaptativos da evolução. A molécula de DNA[5,6] é a herança transmitida entre as gerações, e as características dos indivíduos resultam da tradução das instruções genéticas do DNA em proteínas. A diversificação dos genes se dá por meio de mutações espontâneas e recombinações, enquanto pressões ambientais selecionam os genes que conferem as características mais vantajosas para a sobrevivência e reprodução dos indivíduos. Com o passar do tempo, a repetição desse processo causa uma mudança na composição alélica da população, reflexo do curso contínuo de adequação das espécies ao seu meio[7–9].

Um aspecto fundamental da teoria evolutiva é a coevolução. Conceituada por Ehrlich e Raven em 1964 como "mudanças evolutivas recíprocas entre espécies que interagem"[10], a coevolução estabelece um equilíbrio dinâmico nas relações entre espécies[11], onde a evolução de uma das espécies seleciona e provoca a evolução de outra espécie, que, por sua vez,

exerce uma pressão evolutiva sobre a primeira, também promovendo a sua evolução. O fenômeno coevolutivo, entretanto, permeia relações nas mais diversas esferas da natureza, de modo que sua definição pode ser expandida para compreender a evolução adaptativa em relações de qualquer classe de organização biológica, seja nas interações entre membros da mesma espécie, entre células ou até mesmo em níveis moleculares[12,13].

1.1 A coevolução entre proteínas

Proteínas raramente atuam sozinhas. Elas interagem fisicamente entre si, por meio do contato entre seus aminoácidos, para mediar os processos celulares[14,15]. A manutenção desses contatos ao longo da evolução é dada por mutações compensatórias entre essas proteínas. Caso duas proteínas interajam em um organismo, e uma delas sofra uma mudança em um dos seus aminoácidos, de forma que essa mudança diminua o valor adaptativo (ou *fitness*) do organismo em relação ao seu meio, é necessário que sua parceira passe por uma mudança adaptativa correspondente para que essa interação seja mantida e, consequentemente, possa ser observada[16–18].

O retrato da evolução de uma proteína pode ser visto em um alinhamento múltiplo de sequências (*Multiple Sequence Alignment*, MSA). A partir da sequência primária de uma proteína referência, realiza-se a busca de sequências de suas proteínas homólogas, dispondo-as de forma que seus aminoácidos estejam alinhados verticalmente. Dessa maneira, podemos observar padrões de mudança dos aminoácidos das proteínas no decorrer dos processos de especiação (Figura 1.1). Ademais, quando comparamos conjuntamente dois MSAs de proteínas que interagem, podemos vislumbrar o processo de coevolução que ocorreu entre elas (Figura 1.2)[19–21].

A dinâmica evolutiva das proteínas deixa um sinal que encontra aplicações em diversas questões dentro da biologia[22]. Dentre elas, destaca-se o desenvolvimento de modelos para predição de estruturas proteicas, que alcançaram grande notoriedade com o AlphaFold[23]. A aplicabilidade desse sinal, entretanto, se estende a outras áreas de investigação relevantes, como a predição de mutações patogênicas[24], descoberta de novos alvos para fármacos[25], desenvolvimento de novas proteínas no campo da biologia sintética[26], além de seu uso no mapeamento de interações em diversos grupos taxonômicos[27–29]

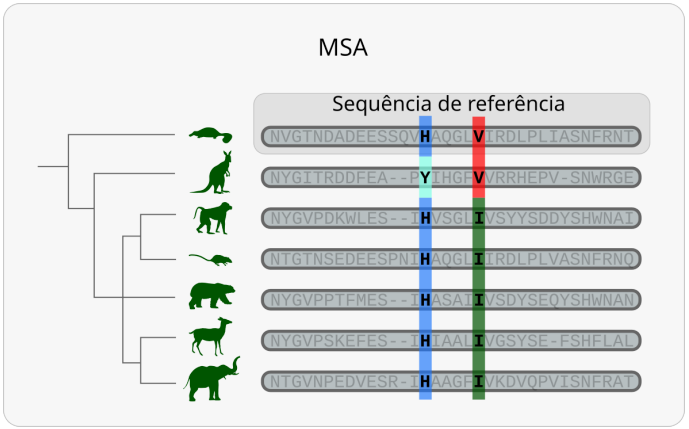

Figura 1.1: **Representação esquemática de um MSA.** Através do MSA, é possível visualizar a evolução de proteínas pela ótica das mudanças de aminoácidos. As colunas coloridas destacam o padrão correlato da mudança de aminoácidos em espécies de mamíferos.

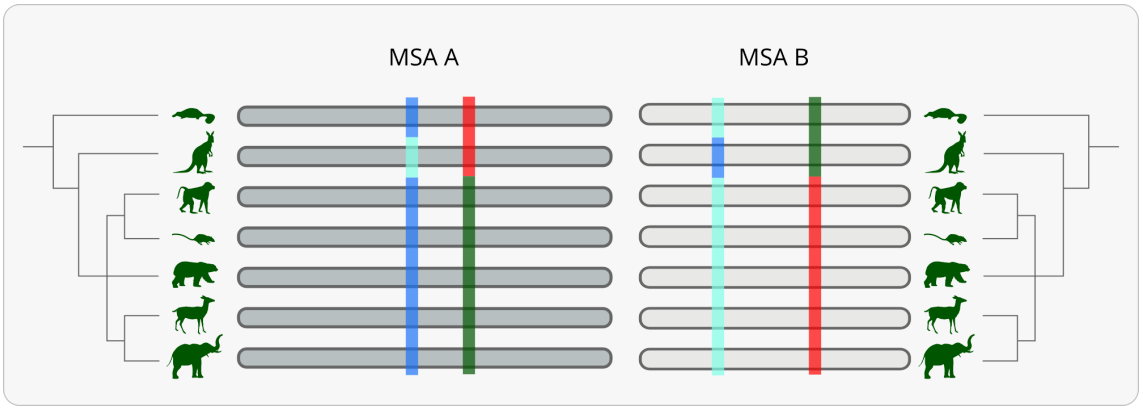

Figura 1.2: **Representação esquemática do sistema formado por duas famílias de proteínas** A **e** B**.** As colunas coloridas ilustram mudanças de aminoácidos relacionadas entre proteínas que interagem.

1.2 Quantificando a coevolução

Grande parte dos métodos utilizados para medir o sinal evolutivo é baseado em informação mútua (I)[30,31]. Essa medida tem como base a teoria da informação de Shannon e quantifica a dependência entre duas variáveis estocásticas. O diagrama de Venn na Figura 1.3 ilustra esse conceito: I representa a intersecção entre a informação (H) de duas variáveis que são, neste caso, colunas dos MSAs de duas proteínas, representadas por i_1 e i_2. Assim, quanto mais correlata é a variação de aminoácidos nas colunas do MSA, maior é o valor de I.

É preciso considerar, no entanto, que a covariância entre aminoácidos não surge apenas a partir da coevolução, mas também de outros elementos intrínsecos ao evento evolutivo. De forma que podemos decompor o valor de I nos seguintes fatores:

$$I = I_{coevolucao} + I_{filogenia} + I_{estocastico} \tag{1.1}$$

onde $I_{coevolucao}$ provém do resultado de pressões seletivas que agem sobre os dois sítios de aminoácidos e favorecem a manutenção da estrutura e função proteica, $I_{filogenia}$ representa a covariação devido à dependência histórica entre as espécies cujas proteínas compõem o sistema e $I_{estocastico}$ refere-se às mudanças acopladas pela simples aleatoriedade de eventos no decurso evolutivo[20,32,33].

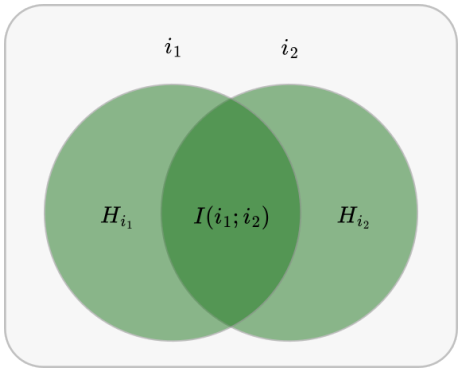

Figura 1.3: **Relação entre informação mútua** (I) **e entropia** (H). Considerando contexto deste trabalho, o diagrama de Venn ilustra a informação das colunas de aminoácidos (H) e a informação mútua (I) entre elas. Adaptado de Cover e Thomas, 2003[34].

Para distinguir a informação filogenética da coevolutiva, devemos considerar os pares formados entre todos os aminoácidos de duas proteínas. A parcela mais expressiva da informação de coevolução é dada pelo acoplamento entre aminoácidos que estão na interface entre as duas proteínas, cujos centros geométricos estão localizados dentro do limite de 8Å de distância. O acoplamento entre os demais aminoácidos, que se encontram a uma distância superior a 8Å, é responsável pela maior parcela da informação filogenética. A informação estocástica é inferida a partir da média de I para aminoácidos em MSAs com sequências completamente embaralhadas, quantificando o valor de I basal do sistema. Andrade *et al.*[33] mostraram que os aminoácidos da interface contêm maior valor de I médio por contato (Figura 1.4). À vista disso, para extrair o sinal coevolutivo mais significativo, tudo o que precisamos fazer é considerar os pares de contato da interface entre as proteínas.[33]

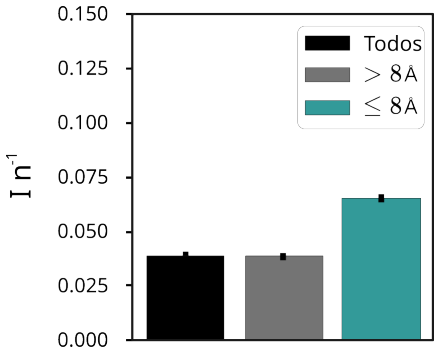

Figura 1.4: **Valor de I para pares de aminoácidos em uma proteína.** As barras indicam a média de I para pares de aminoácidos do sistema TusBCD, cadeias B e C, considerando três casos: os pares formados entre todos os aminoácidos (preto), entre aqueles que não participam do contato interproteico ($> 8\mathring{A}$) (cinza) e aminoácidos da interface ($\leq 8\mathring{A}$) (verde). Adaptado de Andrade *et al.*, 2019[33].

1.3 O Problema dos pares de proteínas

Se o sinal coevolutivo mais significativo é extraído do pareamento nativo de sequências em famílias de proteínas que interagem, seria possível recuperar os pares nativos de duas famílias de proteínas a partir de alinhamentos embaralhados aleatoriamente, por meio da reconstrução do sinal coevolutivo? Esse problema tem sido amplamente investigado nos últimos anos[35–37] e uma das abordagens para resolvê-lo consiste na otimização do sinal de coevolução. Por meio dessa metodologia, podemos encontrar os pareamentos corretos de proteínas identificando aqueles em que o valor da informação mútua é maximizado.

Uma das alternativas mais fáceis e intuitivas para navegar no imenso espaço de probabilidades de pares é o uso de algoritmos de otimização de I, como o algoritmo genético. O algoritmo genético simula o processo de seleção natural através das gerações e possui as seguintes etapas: seleção, morte e reprodução. Conforme mostrado na Figura 1.5, o algoritmo tem início com uma população inicial de arranjos r, que são gerados de forma aleatória e representam os indivíduos. A etapa de seleção tem por objetivo separar os indivíduos com o maior valor de I (maior *fitness*) e que integrarão o grupo da elite. Por sua vez, a etapa de morte tem por função descartar os indivíduos que não foram selecionados para a elite. A reposição desses indivíduos será feita na etapa de reprodução, onde são geradas as cópias dos indivíduos da elite com pequenas mudanças induzidas mediante a um *crossover* de posições sobre o genoma, dando origem a uma nova população. Esse processo é repetido por um determinado número de gerações até que haja a convergência do valor de *fitness*.

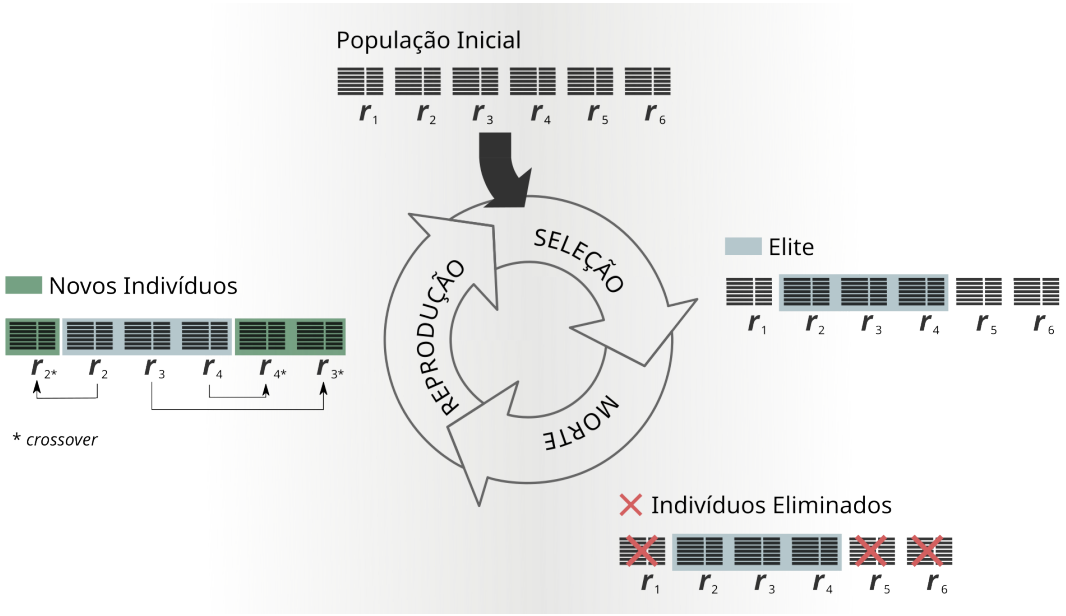

Figura 1.5: **Etapas de um algoritmo genético.** A figura ilustra a delineação do processo de um algoritmo genético com suas etapas de seleção, morte e reprodução.

Em um trabalho recente[38], conduzimos um estudo de otimização em diversas famílias de proteínas utilizando o algoritmo genético. Após a passagem de 50 mil gerações, o algoritmo falhou em parear corretamente os pares de proteínas em MSAs com muitas sequências, mesmo ao atingir valores de I similares ao nativo. As soluções degeneradas decorreram de duas fontes de erro: erros causados por sequências semelhantes, cuja distância de Hamming é de até 20%, e (ii) erros entre sequências não semelhantes.

Proteínas reconhecem suas parceiras em um ambiente denso, povoado por uma miríade de outras moléculas. A afinidade entre elas está condicionada à concentração das proteínas participantes, além das condições de pH e temperatura do compartimento onde se dá a interação. Apesar da força da especificidade da ligação nativa, proteínas podem estabelecer outras ligações, onde elas encontram complementaridade estereoquímica em outras proteínas que sejam semelhantes às suas parceiras. Isso posto, quando desconsideramos os erros triviais do tipo (i) devido à possibilidade da realização de interações promíscuas[39,40], a taxa de pares verdadeiros aumenta substancialmente, atingindo até 50%.

A distinção de soluções otimizadas com erros triviais de outras soluções degeneradas permite a classificação prévia de famílias proteicas onde a previsão precisa dos parceiros proteicos é plausível no nível de clados coevolutivos. Sua aplicação encontra-se em diversos propósitos biotecnológicos destinados à identificação de pares cognatos em eventos de cospeciação ou, ainda, entre genomas independentes. Alguns dos exemplos incluem

proteínas de fagos e receptores bacterianos[41], proteínas de patógenos e células hospedeiras[42], e também neurotoxinas e canais iônicos[43], o que justifica sua relevância.

Apesar dos avanços recentes nesse campo[15,44], o problema preditivo permanece sem solução para conjuntos de sequências em geral, especialmente porque não há uma heurística capaz de encontrar o conjunto correto de proteínas parceiras em MSAs com muitas sequências devido ao espaço astronômico de possibilidades de pareamento. Com o intuito de lançar uma luz sobre esse problema, apresentamos aqui um arcabouço estatístico para descrever a distribuição da informação mútua em sistemas de proteínas e explorar os fatores que possibilitam a conexão de parceiros nativos através do uso do algoritmo genético.

1.4 Objetivos

1.4.1 Objetivo geral

Estabelecer uma estrutura estatística para examinar o pareamento nativo de sequências de proteínas em simulações de otimização mediante o uso do algoritmo genético.

1.4.2 Objetivos específicos

1. Descrever a distribuição de probabilidades em modelos de interação entre famílias de proteínas;

2. Obter uma descrição quantitativa de uma simulação de otimização de informação mútua, em especial do processo embaralhado→nativo;

3. Discutir os parâmetros de um sistema de proteínas que favorecem a resolução dos problemas de pares por meio do algoritmo genético;

4. Analisar pareamentos de proteínas parceiras com a incorporação da reavaliação de erros triviais e seu impacto na precisão preditiva;

5. Comparar o modelo teórico com simulações de otimização da informação mútua realizadas em sistemas reais de proteínas.

CAPÍTULO 2

Teoria e Métodos

2.1 Modelo para um sistema de proteínas

Consideremos que um sistema é composto por dois conjuntos de proteínas de tamanho M, representados pelos MSAs A e B (Figura 2.1(a)). A interação entre eles se dá por meio de $M!$ arranjos distintos r, os quais são descritos por uma variável estocástica R com função massa de probabilidade $p(R)$. Para cada arranjo r, as proteínas pertencentes a A e B interagem por meio de $i = 1, ..., N$ contatos moleculares (Figura 2.1(b)), e suas sequências de aminoácidos são respectivamente descritas por um bloco N de variáveis estocásticas discretas $X \equiv (X_1, ..., X_n)$ e $Y \equiv (Y_1, ..., Y_n)$, com funções massa de probabilidade $p(x^N), p(y^N), p(x^N, y^N|r)$, onde

$$\begin{cases} p(x^N) = \sum\limits_{y^N} p(x^N, y^N|r) \\ p(y^N) = \sum\limits_{x^N} p(x^N, y^N|r) \end{cases} \tag{2.1}$$

e

$$\sum_{x^N, y^N} p(x^N, y^N|r) = 1 \tag{2.2}$$

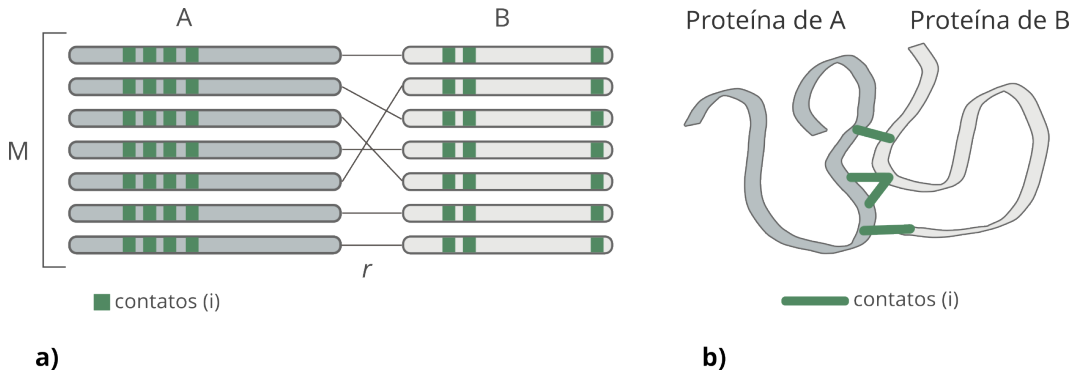

Figura 2.1: **Modelo do sistema de proteínas - a)**: Ilustração de um sistema formado por duas famílias de proteínas A e B; **b)**: Representa a interação entre as duas proteínas de referência do sistema por meio de i contatos.

em cada sequência conjunta $\{x^N, y^N\}_{\chi^{2N}}$ definida no alfabeto χ de tamanho $|\chi|$. No modelo atual, $|\chi| = 21$, referente aos 20 aminoácidos mais o símbolo que denota o *gap*.

2.2 Informação mútua entre famílias de proteínas

Como entendida na teoria da informação de Shannon[34], a informação mútua (I) entre A e B representa a redução da incerteza de A pelo conhecimento de B. Em outras palavras, ela é a quantidade de informação $I(X^N; Y^N|r)$ que uma família de proteínas guarda acerca da outra em cada arranjo r. Dado que as distribuições marginais dos blocos de variáveis de tamanho N $\{p(x^N), p(y^N)\}$ são independentes de r, pois a composição das proteínas é fixa em cada MSA, somente as distribuições conjuntas dependem do arranjo. Assim, para N contatos independentes, a informação mútua é definida extensivamente em termos de contribuições individuais:

$$I(X^N; Y^N|r) = \sum_{i=1}^{N} I(X_i; Y_i|r)\,, \tag{2.3}$$

e seu cálculo é feito da seguinte forma:

$$I(X^N; Y^N|r) = \sum_{x_i, y_i} p(x_i, y_i|r) \log\left(\frac{p(x_i, y_i|r)}{p(x_i)p(y_i)}\right)\,. \tag{2.4}$$

Assim, o valor de I atingirá o seu ponto máximo nos casos em que há um perfeito acoplamento entre A e B, onde $p(x^N, y^N|r) = p(x^N) = p(y^N)$. Por outro lado, seu valor será zero quando os conjuntos de proteínas estiverem totalmente desacoplados, com $p(x^N, y^N|r) = p(x^N)p(y^N)$.

2.3 Distribuição da informação mútua

A distribuição estatística de valores de informação mútua (I) em um sistema de proteínas pode ser descrita de acordo com os pares nativos formados entre as proteínas desse sistema (n), em termos de um modelo de mistura de funções massa de probabilidade

$$f\left(I \mid w_1, \ldots, w_M, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\right) = \sum_{n=0}^{M} w_n f_n\left(I \mid \boldsymbol{\theta}_n\right) , \tag{2.5}$$

em que f_n é a função de distribuição com parâmetros $\boldsymbol{\theta}_n$ e $w_n > 0$ é o peso do enésimo componente da partição, tal que $\sum_{n=0}^{M} w_n = 1$.

Os valores de I resultam da soma de um grande número de distribuições independentes por contato, o que nos permite utilizar o teorema do limite central para inferir que esses valores são normalmente distribuídos

$$f_n\left(I \mid \boldsymbol{\theta}_n\right) = \frac{1}{\sigma_n\sqrt{2\pi}} e^{-\frac{(I-I_n)^2}{2\sigma_n^2}} , \tag{2.6}$$

que é definida pela média e variância para uma dada quantidade de pareamentos corretos $\boldsymbol{\theta}_n = \{I_n, \sigma_n^2\}$. Onde

$$\int_{0<I\leq I'} f\left(I \mid n, \boldsymbol{\theta}_n\right) dI = 1. \tag{2.7}$$

Como o número total de *rencontres* é bem definido quando consideramos todas as permutações possíveis e igualmente prováveis a partir do arranjo nativo

$$D_{M,n} = \frac{M!}{n!} \sum_{q=0}^{M-n} \frac{(-1)^q}{q!} , \tag{2.8}$$

o peso de cada distribuição normal w_n no modelo de mistura converge para uma função de massa de probabilidade da distribuição Poisson com valor esperado $\lambda = 1$

$$w_n = \lim_{M\to\infty} \frac{D_{M,n}}{M!} = \lambda^n \frac{e^{-\lambda}}{n!}\,, \tag{2.9}$$

desde que M seja suficientemente grande e $M! = \sum_{n=0}^{M} D_{M,n}$. Assim, podemos definir a função de distribuição de probabilidade como o enésimo componente da mistura de distribuições normais

$$w_n f\left(I \mid \boldsymbol{\theta}_n\right) = \frac{e^{-1}}{n!} \frac{1}{\sigma_n\sqrt{2\pi}} e^{-\frac{(I-I_n)^2}{2\sigma_n^2}}\,. \tag{2.10}$$

2.4 Equações paramétricas

Para um dado arranjo r com um número arbitrário de posições fixas $0 \leq n \leq M$, definimos $\boldsymbol{\theta}_n = \{I_n, \sigma_n^2\}$ a partir de parâmetros do arranjo completamente embaralhado $\boldsymbol{\theta}_0 = \{0 \leq I_0 < I', \sigma_0^2 > 0\}$ como funções polinomiais

$$\begin{cases} I_n = I_0 + \alpha\left(\frac{n}{M}\right)^2 \\ \sigma_n^2 = \gamma_n\sigma_0^2 + \beta_n \end{cases}\,, \tag{2.11}$$

com

$$\begin{cases} \alpha = I' - I_0 \\ \beta_n = \left(\frac{n}{M}\right)^a \left(1 - \frac{n}{M}\right)^b \\ \gamma_n = 1 - \left(\frac{n}{M}\right) \end{cases}\,, \tag{2.12}$$

se satisfeita a condição $\{I_M = I', \sigma_M^2 = 0\}$.

2.5 Modelo estocástico de Markov para o algoritmo genético

Uma trajetória de otimização no algoritmo genético pode ser encarada como uma sequência de variáveis estocásticas $\{C_t, t = 1, 2, 3, ...\}$ em um espaço de estados S, definidos pelos seus valores de n e I. Posto que a probabilidade de um estado futuro nessa trajetória depende somente do estado presente, mas não de qualquer estado passado, todo esse processo pode ser modelado por meio de uma cadeia de Markov:

$$P\Big(C_{t+1} = (n_{t+1}, I_{t+1})|C_1 = (n_1, I_1), ..., C_t = (n_t, I_t)\Big) = P\Big(C_{t+1} = (n_{t+1}, I_{t+1})|C_t = (n_t, I_t)\Big), \tag{2.13}$$

onde o índice t representa o tempo discretizado na forma das gerações.

2.6 Discretização do valor I e a probabilidade de estado

A fim de determinar as probabilidades de estado do sistema e utilizar o método de cadeia de Markov, calculamos a equação 2.10 dentro de um intervalo de valores I, que recebeu a nomenclatura de ***bin***, de forma que seus valores referem-se à probabilidade de se encontrar algum dos valores de I contidos nesse *bin* para um determinado n. Essa estratégia nos permite obter a probabilidade de cada estado no sistema utilizando apenas três parâmetros: M, I_0 e σ_0^2, que representam, respectivamente, o número de sequências do sistema de proteínas, e o valor médio de I e sua variância em um arranjo totalmente embaralhado.

O conceito de *bin* está associado ao tempo decorrido entre diferentes valores de I. Para que o algoritmo possa resolver os pares proteicos, o número de *bins* no modelo deve corresponder ao número de transições em n necessárias para atingir o arranjo nativo r'. Portanto, em um sistema com 10 sequências, por exemplo, o intervalo entre I_0 e I' é dividido em 11 *bins*, o que resulta em uma matriz quadrada de estados (Figura 2.2 (a)).

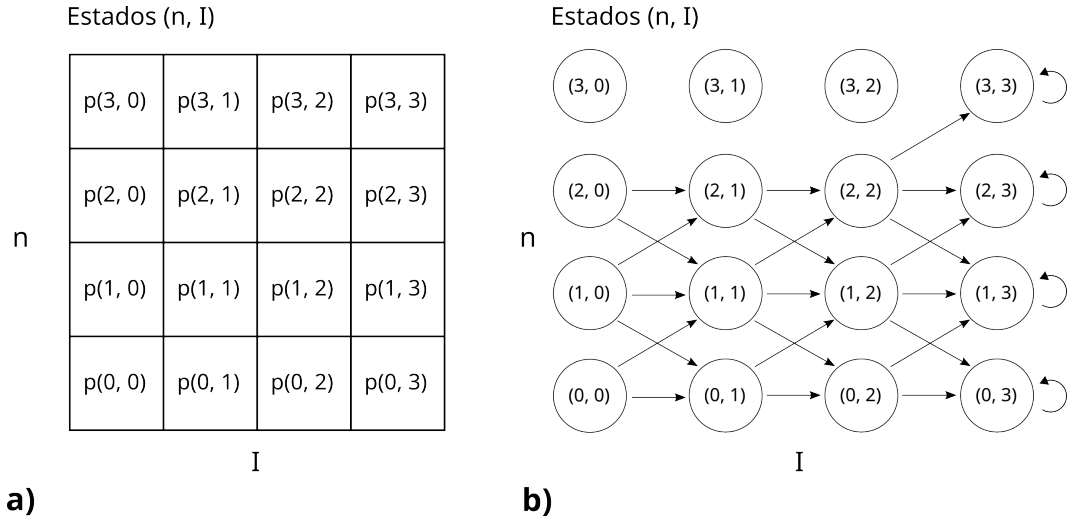

Figura 2.2: **Estados $c = (n, I)$ em um sistema de 3 sequências. a)**: mostra os estados $c = (n, I)$ no espaço do algoritmo genético; **b)**: Representa um diagrama de acessibilidade entre estados de acordo com as restrições do algoritmo genético.

2.7 A matriz de transições

A probabilidade de uma transição entre estados (p) será dada por:

$$p_{(n_t,I_t),(n_{t+1},I_{t+1})} = P\Big(C_{t+1} = (n_{t+1}, I_{t+1}) | C_t = (n_t, I_t)\Big), \quad n_t, n_{t+1}, I_t, I_{t+1} \in S\,. \quad (2.14)$$

Entretanto, é preciso observar que essa não é uma matriz irredutível, ou seja, alguns estados só são acessíveis a partir de estados específicos, como ilustrado na Figura 2.2 **b)**. Essa condição decorre das restrições impostas pelo processo do algoritmo genético. Primeiramente, lembremos que os valores de I são otimizados a cada geração t, então $I_t < I_{t+1}$. Além disso, o *crossover* realizado em cada etapa de reprodução é feito através da uma troca de duas sequências, por conseguinte consideremos que somente estados em que $n_{t+1} \in \{n_t - 1, n_t, n_t + 1\}$ são permitidos (vide seção 3.5). Por fim, estados onde o valor de I é máximo representam sempre o fim de uma trajetória e são considerados estados absorventes (c_k). Convém lembrar que, apesar da Figura 2.2 **b)** apresentar os estados simbólicos com $n = M$ e $I_n < I_M$, esses estados são impossíveis em termos práticos, pois somente o estado nativo atende a condição de $n = M$. Assim, a partir das condições discutidas, a Figura 2.3 mostra a matriz de transições construída para um

Estados (n, I)	(0, 0)	(1, 0)	(2, 0)	(3, 0)	(0, 1)	(1, 1)	(2, 1)	(3, 1)	(0, 2)	(1, 2)	(2, 2)	(3, 2)	(0, 3)	(1, 3)	(2, 3)	(3, 3)
(0, 0)	0	0	0	0	$\frac{p(1,0)}{\sum p}$	$\frac{p(1,1)}{\sum p}$	0	0	0	0	0	0	0	0	0	0
(1, 0)	0	0	0	0	$\frac{p(1,0)}{\sum p}$	$\frac{p(1,1)}{\sum p}$	$\frac{p(2,1)}{\sum p}$	0	0	0	0	0	0	0	0	0
(2, 0)	0	0	0	0	0	$\frac{p(1,1)}{\sum p}$	$\frac{p(2,1)}{\sum p}$	0	0	0	0	0	0	0	0	0
(3, 0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(0, 1)	0	0	0	0	0	0	0	0	$\frac{p(0,2)}{\sum p}$	$\frac{p(1,2)}{\sum p}$	0	0	0	0	0	0
(1, 1)	0	0	0	0	0	0	0	0	$\frac{p(0,2)}{\sum p}$	$\frac{p(1,2)}{\sum p}$	$\frac{p(2,2)}{\sum p}$	0	0	0	0	0
(2, 1)	0	0	0	0	0	0	0	0	0	$\frac{p(1,2)}{\sum p}$	$\frac{p(2,2)}{\sum p}$	0	0	0	0	0
(3, 1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(0, 2)	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{p(0,3)}{\sum p}$	$\frac{p(1,3)}{\sum p}$	0	0
(1, 2)	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{p(0,3)}{\sum p}$	$\frac{p(1,3)}{\sum p}$	$\frac{p(2,3)}{\sum p}$	0
(2, 2)	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{p(1,3)}{\sum p}$	$\frac{p(2,3)}{\sum p}$	$\frac{p(3,3)}{\sum p}$
(3, 2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(0, 3)	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
(1, 3)	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
(2, 3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
(3, 3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figura 2.3: **Matriz de transições para um sistema de 3 sequências**. $\sum p$ representa a soma das probabilidades de transição para estados acessíveis a partir daquele estado.

sistema composto por 3 sequências.

2.8 Como encontrar a trajetória mais provável?

Com as probabilidades de estado e a matriz de transições definidas, é possível calcular a probabilidade de uma trajetória específica no algoritmo genético por meio do produto da probabilidade do estado inicial e das probabilidades de transição entre os estados acessados na trajetória:

$$p_{\{C_t,t=1,2,3,...\}} = p_{(n_0,I_0)} \times p_{(n_0,I_0),(n_t,I_t)} \times ... \times p_{(n_t,I_t),(n_k,I_k)} \,. \tag{2.15}$$

Assim, para encontrar a probabilidade total da trajetória mais provável nesse espaço, é preciso partir do estado de maior probabilidade e acessar, a cada passo, o estado de maior probabilidade de transição. O último estado $c = (n, I)$ acessado na trajetória define a taxa de pares nativos (*True Pairs rate*, TP) dado como $TP = n/M$.

2.9 Reavaliação das sequências

Quando levamos em consideração um nível de promiscuidade nas interações entre as proteínas do modelo, a equação 2.10 pode ser reformulada para estabelecer n', onde uma sequência pode ser pareada com a sua parceira nativa (n) ou com outra sequência que seja semelhante a sua parceira (m), com base em uma distância predefinida. Consideremos a transformação invariante da distribuição de probabilidades de I:

$$\sum_{n=0}^{M} w_n f_n \left(I \mid \boldsymbol{\theta}_n\right) \Leftrightarrow \sum_{n'=0}^{M} w_{n'} f_{n'} \left(I \mid \boldsymbol{\theta}_{n'}\right), \; n' = n + m \,, \tag{2.16}$$

em que a densidade de probabilidade ponderada na equação 2.5 é expandida

$$w_n f_n \left(I \mid \boldsymbol{\theta}_n\right) \simeq \sum_{m=0}^{M-n} w_{nm} f_{nm} \left(I \mid \boldsymbol{\theta}_{nm}\right) \tag{2.17}$$

em termos das normais $f_{nm}(.)$ auxiliares de I através de um conjunto finito de processos de Bernoulli, onde arranjos de tamanho M e n pares nativos contêm $0 \leq m \leq M-n$ sequências similares com uma probabilidade p

$$\begin{cases} w_{nm} &= \begin{pmatrix} M-n \\ m \end{pmatrix} p^m (1-p)^{M-n-m} \frac{e^{-1}}{n!} \\ \boldsymbol{\theta}_n &= \sum_{m=0}^{M-n} w_{nm} \boldsymbol{\theta}_{nm} \end{cases} . \tag{2.18}$$

Assim, a distribuição de I pode ser reformulada

$$f\left(I \mid n', \boldsymbol{\theta}_{n'}\right) = \sum_{n'=0}^{M} w_{n'} f_{n'}\left(I \mid \boldsymbol{\theta}_{n'}\right) = \sum_{n'=0}^{M} \sum_{n=0}^{M} \sum_{m=0}^{M-n} \delta_{(n+m)n'} w_{nm} f_{nm}\left(I \mid \boldsymbol{\theta}_{nm}\right), \tag{2.19}$$

como uma combinação das contribuições individuais de $w_{nm} f_{nm}(.)$, no qual o numero de sequências similares satisfazem a condição do delta de Kronecker $n + m = n'$. Então, a equação se torna uma mistura reponderada de distribuições normais:

$$f\left(I \mid n', \boldsymbol{\theta}_n\right) = \sum_{n'=0}^{M} \sum_{n=0}^{M} \sum_{m=0}^{M-n} \delta_{(n+m)n'} \begin{pmatrix} M-n \\ m \end{pmatrix} p^m q^{M-n-m} \frac{e^{-1}}{n!} \frac{1}{\sigma_{nm}\sqrt{2\pi}} e^{-\frac{[I-I_{nm}]^2}{2\sigma_{nm}^2}} . \tag{2.20}$$

Como $I_n \approx I_{nm}$ e $\sigma_n^2 \approx \sigma_{nm}^2$ (vide seção 3.3), podemos realizar essas aproximações na equação 2.20:

$$f\left(I \mid n', \boldsymbol{\theta}_n\right) = \sum_{n'=0}^{M} \sum_{n=0}^{M} \sum_{m=0}^{M-n} \delta_{(n+m)n'} \begin{pmatrix} M-n \\ m \end{pmatrix} p^m q^{M-n-m} \frac{e^{-1}}{n!} \frac{1}{\sigma_n\sqrt{2\pi}} e^{-\frac{[I-I_n]^2}{2\sigma_n^2}} . \tag{2.21}$$

Essa substituição possibilita o uso do arcabouço estatístico dos estados $c = (n, I)$ para

investigar a redistribuição das probabilidades em estados $c = (n, I)$ reavaliados.

2.10 Sistemas reais de proteínas

Os sistemas reais de proteínas empregados neste estudo incluem famílias de proteínas formadas por sequências ortólogas e uma com sequências parálogas. Os sistemas ortólogos são descritos na Tabela 2.1 e no decorrer texto serão referenciados pelo padrão do PDB Id da proteína de referência mais as cadeias utilizadas para formação dos MSAs. Os MSAs dessas famílias de proteínas foram obtidos do trabalho de Ovchinnikov *et al.*[45], e as estruturas tridimensionais das proteínas de referência, utilizadas para a seleção das colunas do MSA referentes às posições dos contatos de interface (i), foram obtidas no Protein Data Bank[46] (PDB). Adicionalmente, o sistema parálogo HK-RR foi construído e validado por Bitbol *et al.*[15] e é formado por histidina quinases (HKs) e seus respectivos reguladores de resposta (RRs), que fazem parte de sistemas de sinalização bacterianos de dois componentes. A seleção dos contatos de interface desse sistema é feita com as proteínas *Thermotoga maritima* HK853 - RR468 (5UHT), cadeias A, B.

2.10.1 Simulações de otimização de I em sistemas reais

Obtivemos parte dos dados das simulações de otimização de I em sistemas ortólogos de outro trabalho desenvolvido anteriormente no Laboratório de Biologia Teórica e Computacional da Universidade de Brasília[38]. Especificamente, 6 das 12 réplicas totais para cada sistema. As demais réplicas das simulações foram feitas utilizando as mesmas condições encontradas no trabalho anterior, no qual a população foi composta por 8 indivíduos, que iniciaram os algoritmos com as sequências aleatoriamente embaralhadas. A elite foi de 50% dos indivíduos, com o *crossover* das proles da elite realizado sobre 2 sequências. O algoritmo foi finalizado após 50 mil gerações.

Em relação ao sistema parálogo HK-RR, o algoritmo foi ligeiramente diferente. Para este caso, o embaralhamento das seqûencias da população inicial foi feito somente dentro das sequências de uma mesma espécie. A cada geração, uma das espécies foi escolhida de forma randômica e o *crossover* de 2 sequências aconteceu dentro da espécie selecionada. Ao fim da simulação, o resultado da otimização foi avaliado para cada espécie de forma separada.

Tabela 2.1: Sistemas de proteínas ortólogas avaliados neste trabalho.

PDB Id	**Nome do Sistema**	**Cadeias**	**M**
1B70	Phenylalanyl-tRNA synthetase	A, B	1108
1BXR	Carbamoyl phosphate synthetase	A, B	1004
1EFP	Electron transfer flavoprotein	A, B	1347
1EP3	Dihydroorotate dehydrogenase B	A, B	552
1I1Q	Anthranilate synthase	A, B	1204
1QOP	Tryptophan synthase	A, B	1155
1RM6	4-hydroxybenzoyl-CoA reductase from *Thauera aromatica*	A, B	1604
1RM6	4-hydroxybenzoyl-CoA reductase from *Thauera aromatica*	A, C	1534
1RM6	4-hydroxybenzoyl-CoA reductase from *Thauera aromatica*	B, C	1481
1TYG	Thiazole synthase/ThiS complex	A, B	746
1W85	Pyruvate dehydrogenase	A, B	1537
1ZUN	GTP-Regulated ATP sulfurylase	A, B	649
2D1P	TusBCD	B, C	216
2NU9	Succinyl-CoA synthetase	A, B	798
2VPZ	Polysulfide reductase	A, B	676
2WDQ	Succinate:quinone oxidoreductase	C, D	221
2Y69	Bovine cytochrome C oxidase	A, B	1484
2Y69	Bovine cytochrome C oxidase	A, C	863
3G5O	Toxin-antitoxin complex RelBE2	A, B	904
3IP4	GatCAB	A, B	782
3IP4	GatCAB	A, C	879
3IP4	GatCAB	B, C	689
3MML	Hydrolase complex from *Mycobacterium smegmatis*	A, B	1067
3OAA	F1-ATP synthase	H, G	886
3PNL	Dha kinase DhaK-DhaL complex	A, B	902
3RRL	3-oxoadipate coA-transferase	A, B	1330

2.11 Recursos computacionais e acesso aos repositórios

2.11.1 Análise de dados e representação dos resultados

A extração de informações das estruturas tridimensionais das proteínas, manipulação dos MSAs, cálculos de I e análise dos dados gerados foram feitas utilizando a linguagem Python v3.8 e v3.11. Desenvolvemos a biblioteca Coevtools para maior reprodutibilidade dos resultados, que está disponível para *download* no seguinte repositório:

`https://github.com/jafiorote/coevtools`.

As simulações de algoritmo genético realizadas neste trabalho e o cálculo de I para arranjos aleatórios foram feitos por meio da biblioteca MIGA, que pode ser encontrada em:

`https://github.com/caioss/miga`.

O cálculo de I realizado pelo MIGA leva em conta os pares formados entre todos os aminoácidos da interface; o realizado neste trabalho, somente os pares formados entre aminoácidos da interface que estão na distância de até 8Å. Fizemos, então, uma pequena modificação no trecho do código relativo ao cálculo. Essa versão está disponível em:

`https://github.com/jafiorote/miga`.

Todos os gráficos deste trabalho foram gerados com o uso da biblioteca Matplolib v3.6.2. Em algumas ocasiões, representamos em um *heatmap* probabilidades que divergiam em muitas ordens de magnitude. Para isso, dividimos os valores a serem representados em percentis com a utilização do método *percentile* da biblioteca Numpy v1.21.

2.11.2 Implementação do Modelo

A implementação do modelo foi realizada com o uso da linguagem Python v3.8 e gerou biblioteca GA_error_sources, que contém um *notebook* com o fluxo de exploração de toda a estrutura estatística. Ela pode ser encontrada em:

`https://github.com/jafiorote/ga_error_sources`.

Por fim, todo o fluxo de trabalho foi realizado no Laboratório de Biologia Teórica e Computacional, no Instituto de Biologia - Universidade de Brasília.

CAPÍTULO 3

Resultados e Discussão

Apresentamos nesta seção uma investigação dos componentes do modelo descrito na metodologia em uma gama de parâmetros, incluindo os encontrados nos sistemas reais mostrados na Tabela 2.1. Os resultados aqui expostos nos auxiliam a compreender as ideias apreendidas por meio do modelo teórico, além de fundamentar as decisões metodológicas tomadas durante o processo de desenvolvimento. Com o intuito de facilitar o fluxo de leitura, alguns dos resultados suplementares serão exibidos no Capítulo 5.

3.1 Curvas teóricas das equações paramétricas

As curvas paramétricas foram embasadas no comportamento do valor médio de I em relação a n em sistemas reais. Dentre os sistemas ortólogos apresentados na Tabela 2.1, foi selecionada uma amostra de sistemas com valores de M que abrangiam todo o intervalo de M presente nas famílias de proteínas. Em cada um desses sistemas, foram definidos 50 pontos de n uniformemente distribuídos entre 0 e M, e calculamos o valor médio de I de 50 mil arranjos aleatórios em cada ponto de n. Os resultados de cada sistema são exibidos na Figura 3.1 **a)**.

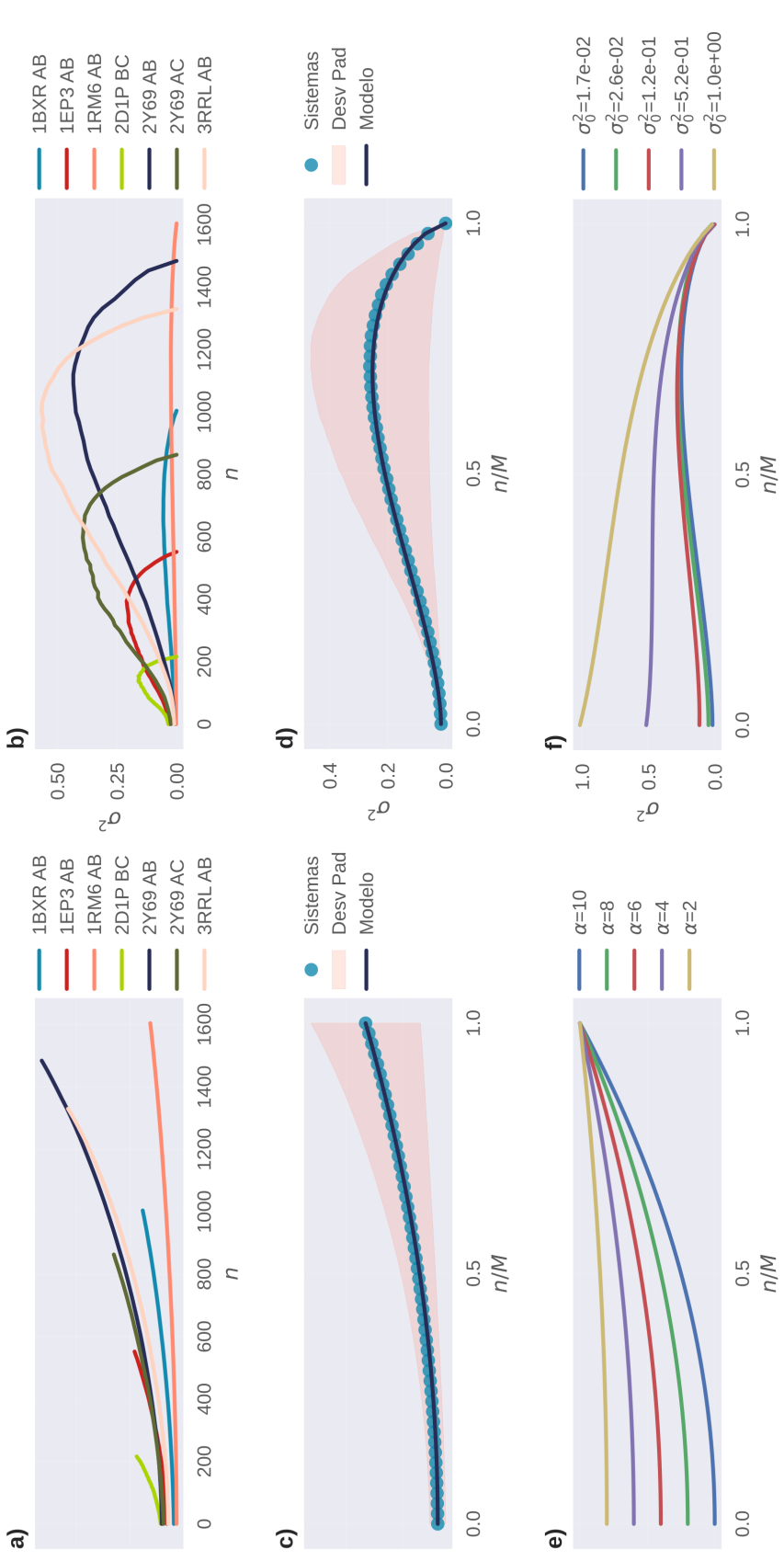

Figura 3.1: **Comparação das curvas paramétricas com curvas obtidas de sistemas reais.** Em **a)** e **b)** são mostrados os valores médios de I e σ^2 em função de n para um conjunto de sistemas de proteínas. **c)** e **d)** apresentam as curvas médias de I e σ^2 dos sistemas, seu desvio padrão, bem como as curvas teóricas da equação 2.11 calculada com valores de I_0 e σ_0^2 obtidos das curvas médias. **e)** e **f)** mostram a variação das curvas paramétricas em decorrência da mudança dos valores de entrada I_0 ($\alpha = I' - I_0$) e σ_0^2.

Foi realizada uma normalização das curvas da média de I em função de n para cada sistema e com isso obtivemos uma única curva média representativa de todos os sistemas (Figura 3.1 **c)**). Além dessa curva média, foram incluídos o desvio padrão dos valores médios de I e a curva teórica derivada da equação paramétrica de I_n (2.11), desenvolvida para descrever o comportamento da curva média. A curva paramétrica mostrou um ótimo ajuste quando calculada utilizando os valores médios de I_0 e I' dos sistemas. Na Figura 3.1 **e)**, é ilustrado o comportamento das curvas teóricas de I_n em resposta à variação do parâmetro α, definido como $I' - I_0$ (equação 2.12).

Para o desenvolvimento da função de σ_n^2 da equação 2.11, utilizamos os valores obtidos na elaboração da curva de I_n. Calculamos a variância de I nas 50 mil redes aleatórias geradas em cada ponto n em cada um dos sistemas. Essas curvas de σ^2 em função de n são apresentadas na Figura 3.1, **b)**. Na Figura 3.1, **d)** são mostradas a curva média das variâncias, o desvio padrão dessas médias e a curva teórica dada pela equação de σ_n^2 com valores de σ_0^2 e σ_M^2 iguais aos encontrados nas médias dos sistemas. O ajuste da curva teórica sobre a curva média foi realizado com os valores de $a = 1{,}63$ e $b = 0{,}68$. Ao final, a família de curvas de σ_n^2 com diferentes valores de σ_0^2 podem ser vistas na Figura 3.1, **f)**.

3.2 Função de densidade de probabilidade

A partir dos valores de I e σ^2 fornecidos pelas equações paramétricas, a equação 2.6 foi utilizada para determinar a função de densidade de probabilidade (*Probability Density Function*, PDF) de I para um dado n. Na Figura 3.2 são mostradas as distribuições dos valores de I para 8000 redes aleatórias com valores de n fixos do sistema 1BXR-AB, além das curvas da equação 2.6, calculadas com parâmetros idênticos aos encontrados no sistema.

Em um sistema com duas famílias de M proteínas, n pode assumir valores de 0 a M, totalizando $M + 1$ n-normais. O conjunto dessas curvas nos permite vislumbrar o comportamento do sistema completo, em que os picos das normais evidenciam os pontos da equação paramétrica 2.11. Na Figura 3.3 apresentamos o comportamento das curvas normais sob diferentes condições de I_0 e σ_0^2, com os valores de $I' = 50$ e $M = 100$ fixos. As Figuras com os resultados para $M = 10$ (5.1), $M = 20$ (5.2) e $M = 50$ (5.3) podem ser encontradas na seção de resultados suplementares 5.1.

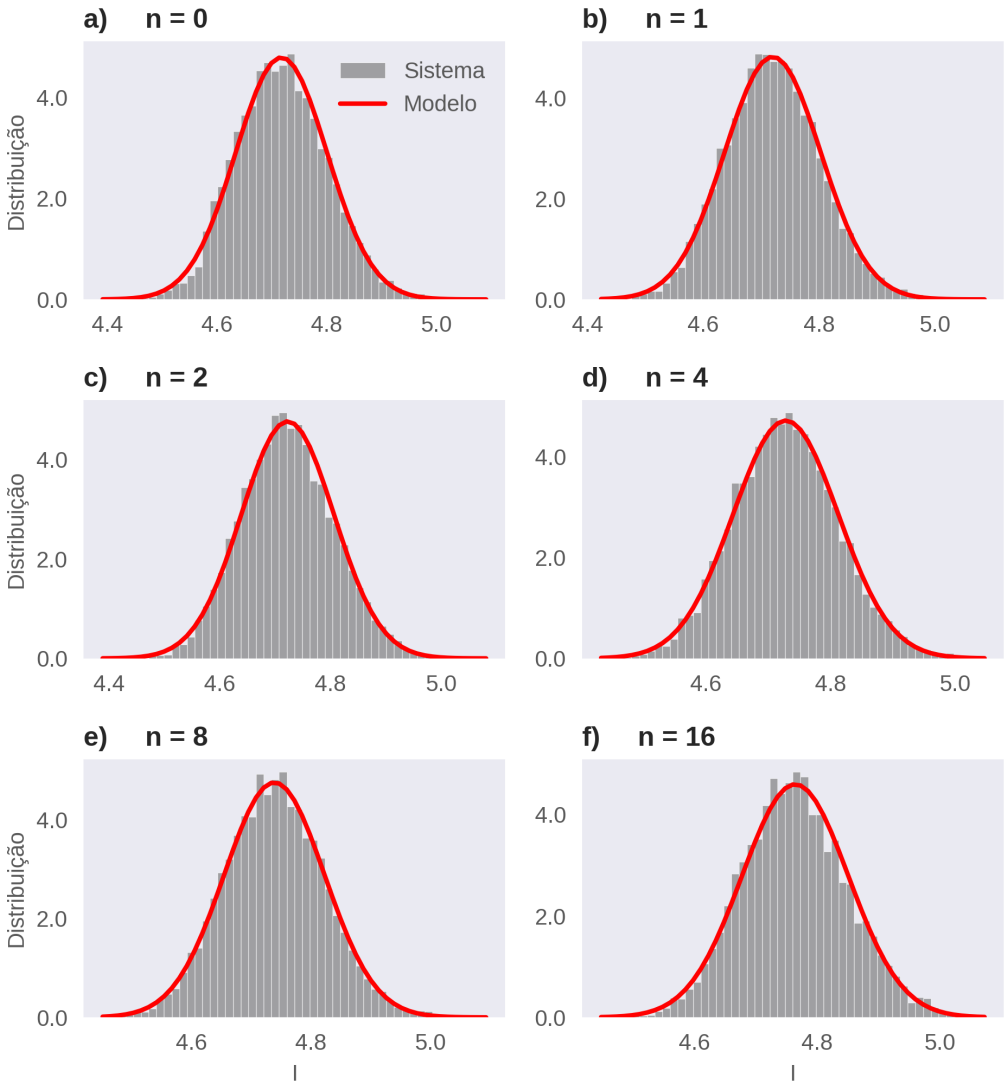

Figura 3.2: **Distribuição dos valores de I no sistema 1BXR-AB.** A figura exibe os histogramas normalizados com a distribuição dos valores de I em arranjos aleatórios com valores de n = **a)**:0, **b)**:1, **c)**:2, **d)**:4, **e)**:8 e **f)**:16. Junto a cada histograma está a curva normalizada do modelo.

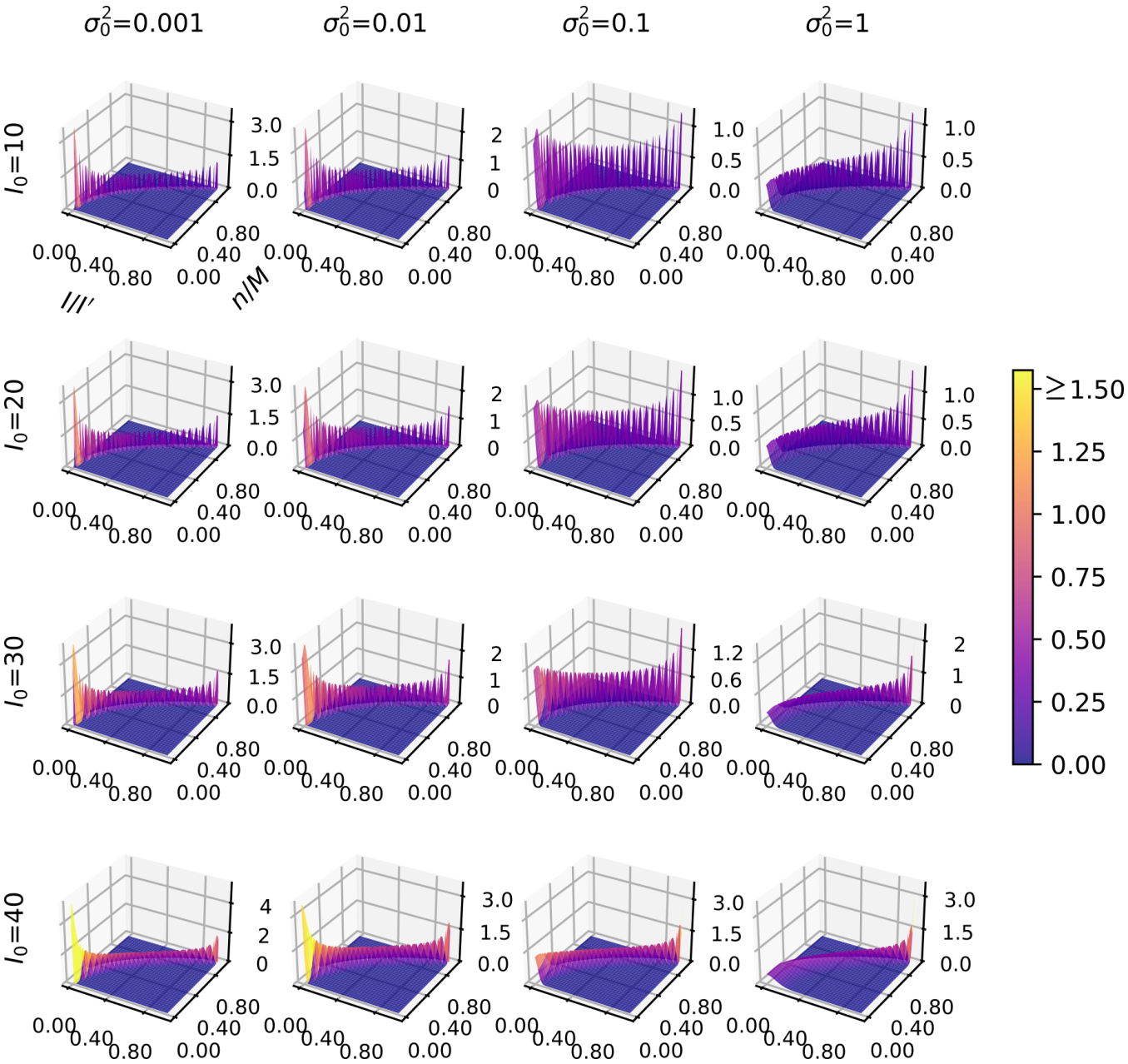

Figura 3.3: **Densidade de probabilidade;** $M = 100$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$. Nas linhas, os valores de I_0 aumentam de cima para baixo, enquanto, nas colunas, os valores de σ_0^2 aumentam da esquerda para a direita.

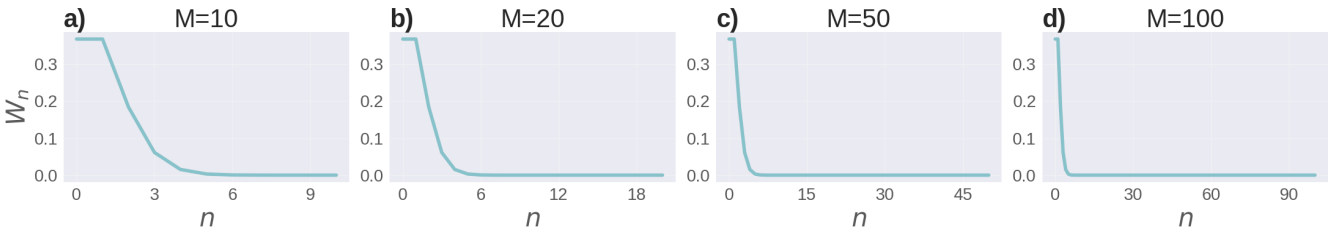

Figura 3.4: **Dependencia do peso de Poisson W_n com o número de pareamentos nativos n.** Os curvas consideram os valores de M = **a)**: 10, **b)**:20, **c)**:50 e **d)**:100.

No modelo de mistura de funções, as n-normais são ponderadas por uma distribuição Poisson. A Figura 3.4 exibe o comportamento da Poisson conforme o número de sequências cresce. Assim, ao adicionar o segundo componente da mistura, as PDFs exibidas na Figura 3.3 adquirem uma nova feição com a densidade de probabilidade concentrada nos menores valores de n e I, onde encontram-se a maioridade das redes possíveis, mas cujos pareamentos possuem pouca coerência coevolutiva (Figura 3.5). Como é intuitivamente esperado, os resultados mostram como é altamente improvável encontrar arranjos aleatórios com altos valores de I e n. Nos resultados suplementares também são mostradas as Figuras com os resultados para $M = 10$ (5.4), $M = 20$ (5.5) e $M = 50$ (5.6).

3.3 Aproximações para a reavaliação de sequências

Para realizar a reavaliação das n-normais da Figura 3.5, foi necessário investigar como se comportariam as variáveis de I_{nm} e σ^2_{nm} presentes na equação 2.21. Verificamos, então, a formação de pareamentos aleatório no sistema 1BRX-AB (Tabela 2.1). Foram gerados 8 mil arranjos aleatórios para arranjos com $n = 0$, 1, 2, 4, 6, 8 e 16, e averiguamos os pareamentos realizados entre uma sequência e outra sequência que seja próxima a sua parceira nativa (m), utilizando um ponto de corte correspondente ao 20° percentil da distância de Hamming, beaseado na distribuição das distâncias de Hamming das sequências do MSA B. Na Figura 3.6 os resultados são exibidos na forma de histograma, juntamente à curva de probabilidade da predição teórica, dada pela equação 2.18.

A partir do resultado anterior, calculamos a média e a variância de I em cada arranjo m do acumulado dos histogramas apresentados na Figura 3.6. Em seus respectivos escopos de n, esses valores foram comparados e correlacionados a I_n e σ^2_n. Os resultados mostram

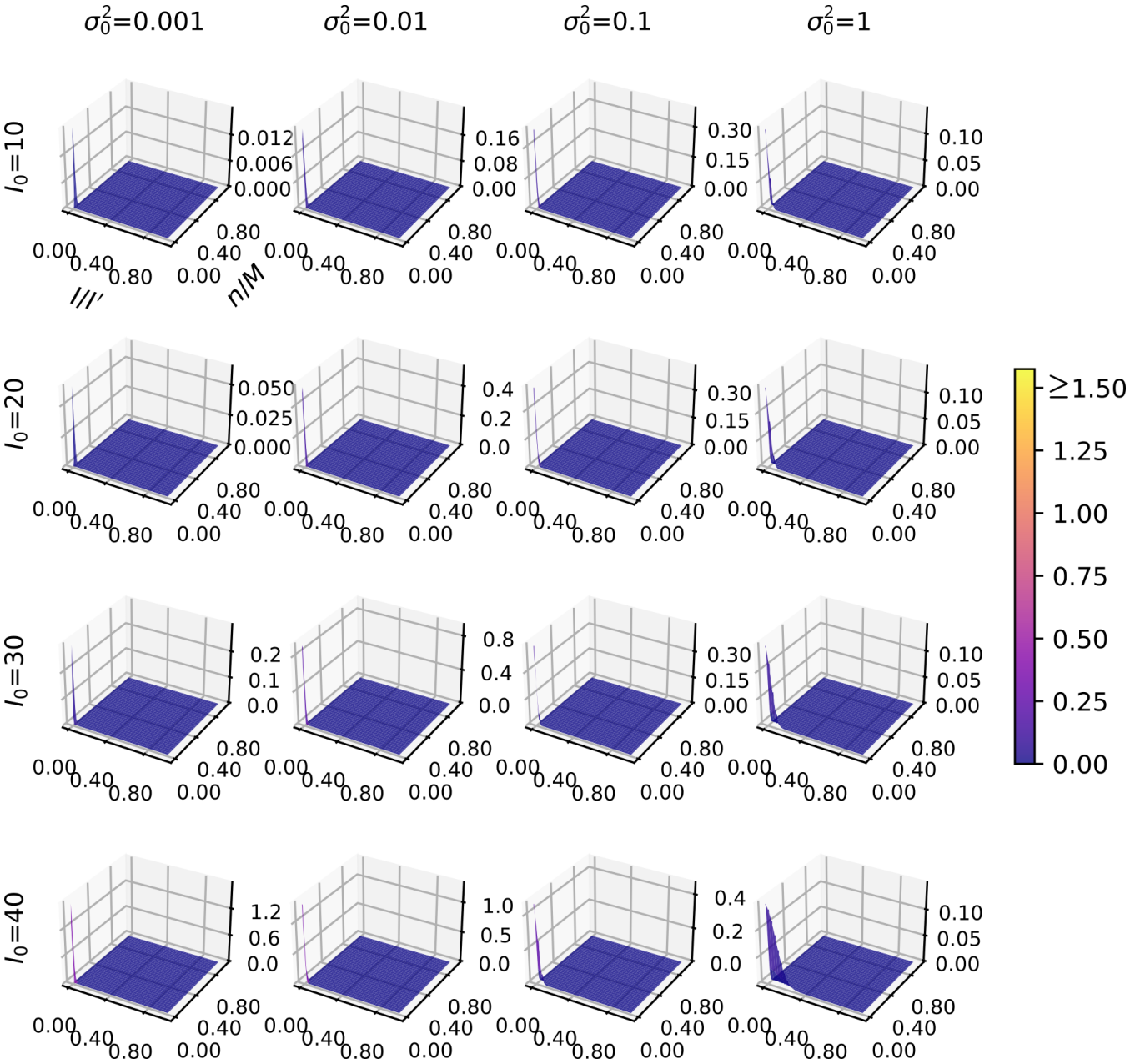

Figura 3.5: **Densidade de probabilidade ponderada por Poisson;** $M = 100$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$. Nas linhas, os valores de I_0 aumentam de cima para baixo, enquanto, nas colunas, os valores de σ_0^2 aumentam da esquerda para a direita.

na Figura 3.7 que $I_n \approx I_{nm}$ e $\sigma_n^2 \approx \sigma_{nm}^2$, o que nos permite realizar essas aproximações na equação 2.20.

Com o uso da equação 2.21, pudemos reavaliar as n-normais ponderadas por Poisson da Figura 3.5. Como ilustra a Figura 3.8, o pico de densidade continua localizado próximo a I_0, mas é deslocado para valores maiores n. Esse deslocamento da densidade depende do valor da probabilidade de se encontrar sequências similares em um pareamento, dado por p. Neste exemplo, assim como nos encontrados na seção 5.2 para $M = 10$ (Figura 5.7), $M = 20$ (Figura 5.8) e $M = 50$ (Figura 5.9), as PDFs foram reavaliadas com $p = 0.25$.

3.4 Probabilidades dos estados $c = (n, I)$

As probabilidades para os estados $c = (n, I)$ foram calculadas com a soma da função de densidade de probabilidade (equação 2.10) em um intervalo de I (bin) para um dado n. Como mencionado na seção 2.6, essa estratégia nos permite discretizar o valor de I para aplicação do modelo de Markov. Os valores em cada bin foram divididos pela probabilidade calculada no intervalo (I_0, I'), de modo a truncar a função no espaço de I do sistema. Apresentamos na Figura 3.9 a distribuição de probabilidades dos estados para sistemas com $10, 20, 50$ e 100 sequências, e valores fixos de $I_0 = 10$ e $I' = 50$.

3.5 Transições em um algoritmo genético

Com o propósito de entender a dinâmica de transições na simulação de otimização e determinar as restrições do modelo, fizemos uma simulação de otimização de I para os sistemas 1BXR-AB, 1ZUN-AB e 2D1P-BC. A escolha desses sistema se deve por conta do contraste na produção de erros triviais entre eles[38]. A cada geração do algoritmo, computou-se quantos pares corretos foram formados com a progressão de I. A Figura 3.10 mostra os resultados para os 3 sistemas, onde nenhum dos casos apresentou transições significativas para $n \pm 2$. Assim, restringimos a transição em n entre os estados do modelo como $n_{t+1} \in \{n_t - 1, n_t, n_t + 1\}$ (seção 2.7).

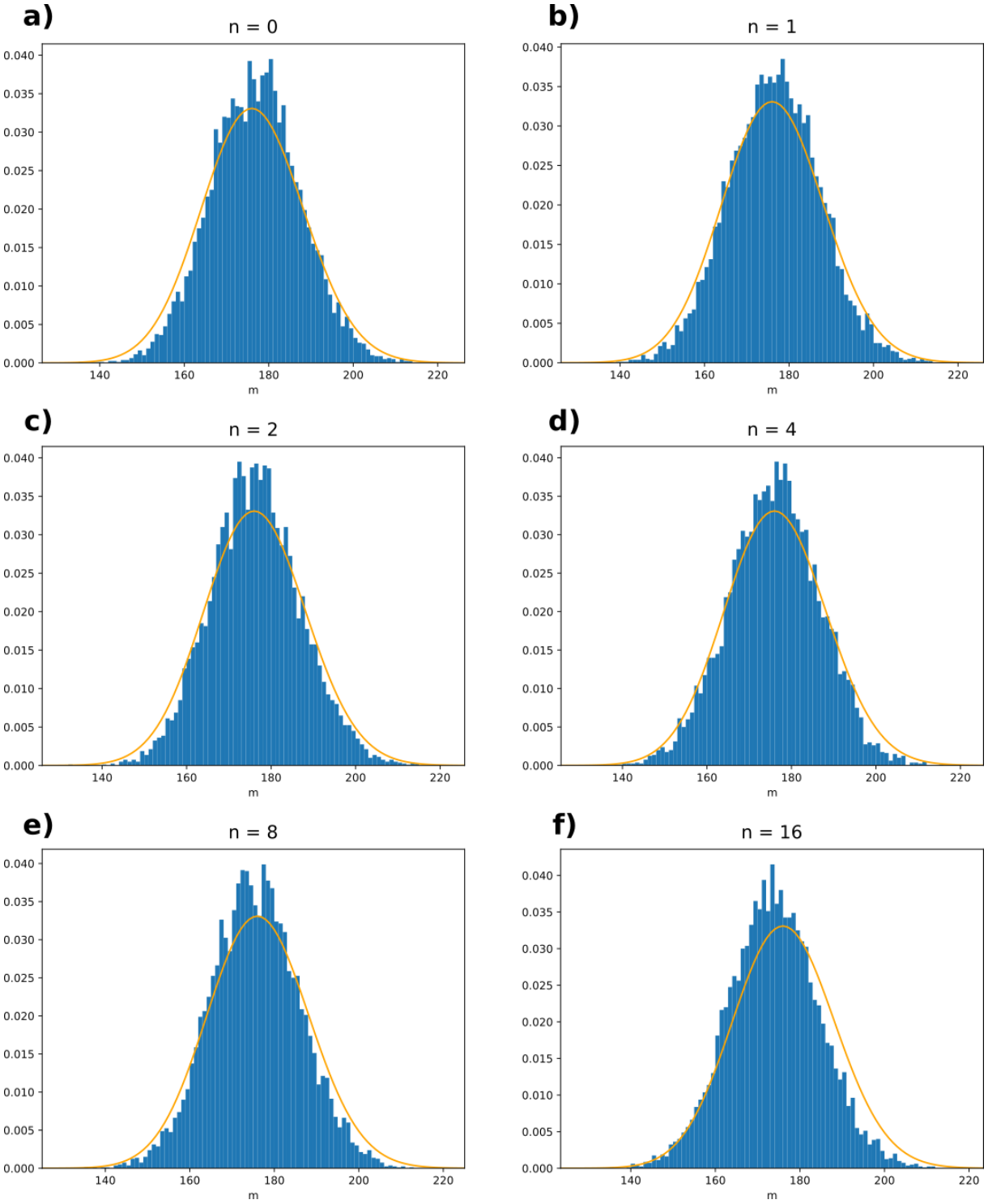

Figura 3.6: **Densidade de pareamentos similares** (m) **para o sistema 1BXR-AB.** A figura exibe os histogramas normalizados com a distribuição de m em arranjos aleatórios com valores de n = **a)**:0, **b)**:1, **c)**:2, **d)**:4, **e)**:8 e **f)**:16. Junto a cada histograma está a curva de probabilidade binomial dada pela equação 2.18.

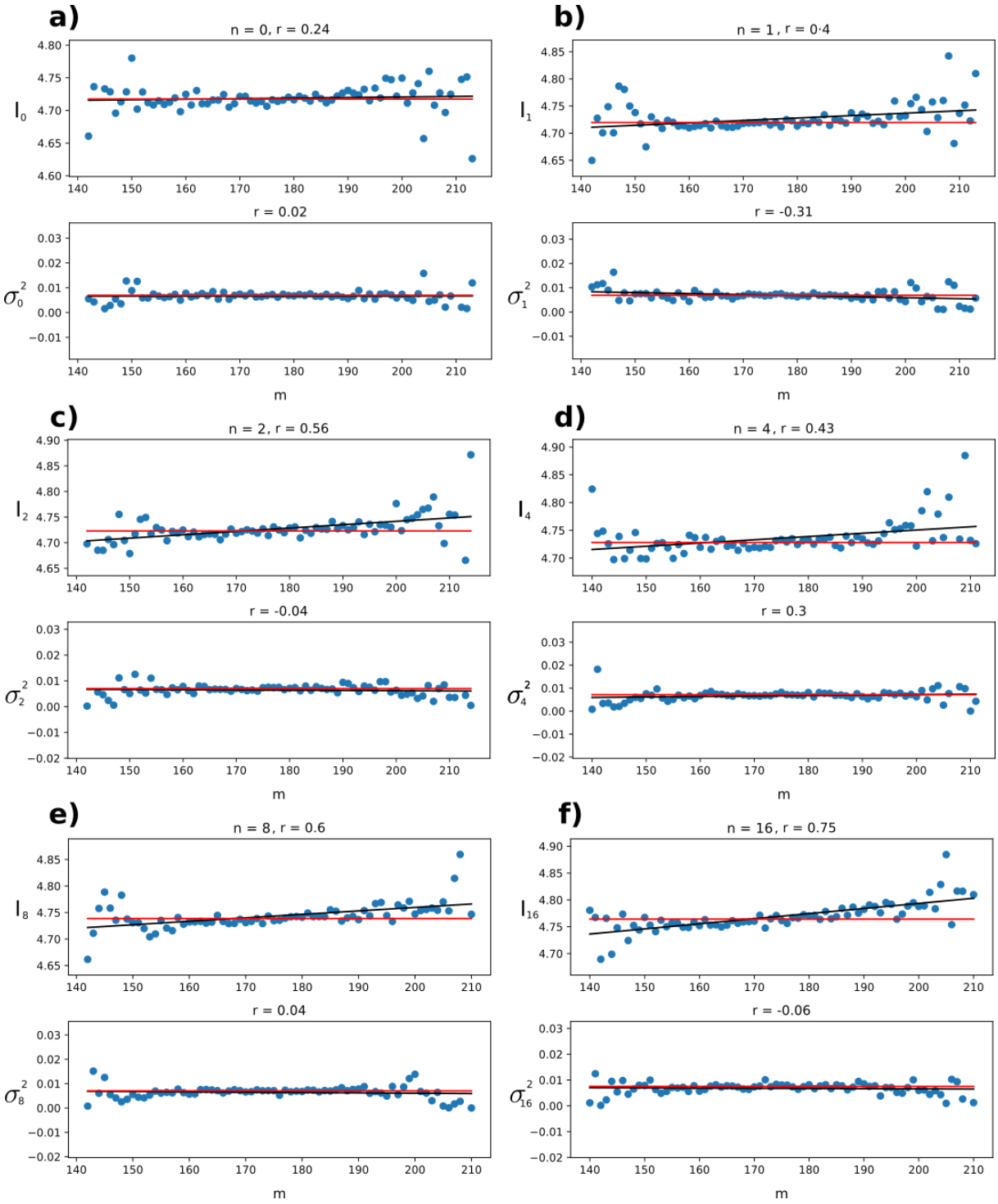

Figura 3.7: **Relação dos valores de I e σ^2 entre n e m.** Valores médios de $I - n$ (gráfico superior) e $\sigma^2 - n$ (gráfico inferior) para m a partir de diferentes ns no sistema 1BXR-AB, com n igual a **a)**: 0; **b)**: 1; **c)**: 2; **d)**: 4; **f)**: 8; **g)**: 16. Os pontos representam os valores de m. A reta vermelha marca o valor calculado para o dado n; a preta, o *fit* de valores de m. A correlação (r) entre as retas é apresentado acima de cada gráfico.

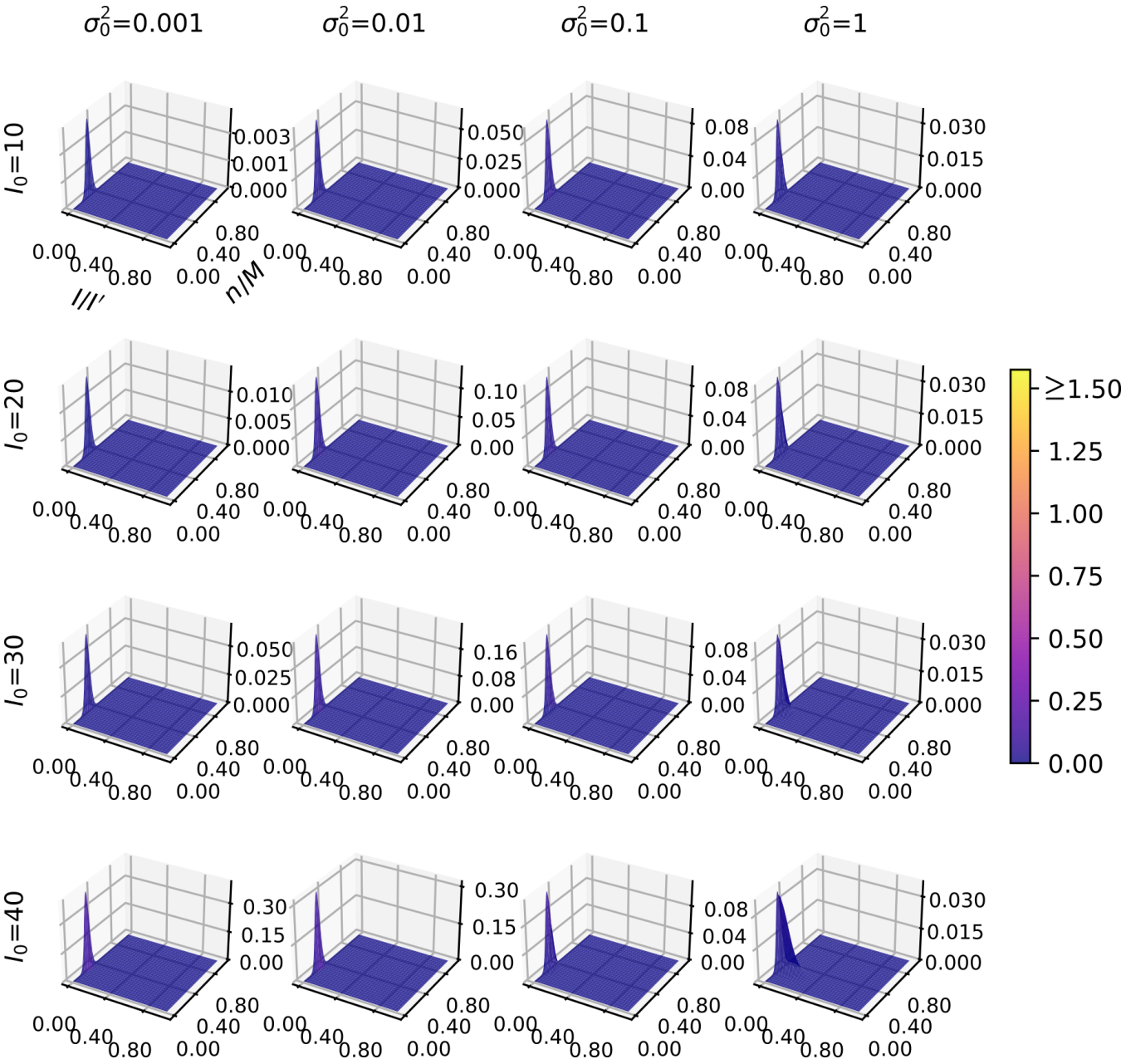

Figura 3.8: **Reavaliação da densidade de probabilidade ponderada por Poisson;** $M = 100$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$. Nas linhas, os valores de I_0 aumentam de cima para baixo, enquanto, nas colunas, os valores de σ_0^2 aumentam da esquerda para a direita.

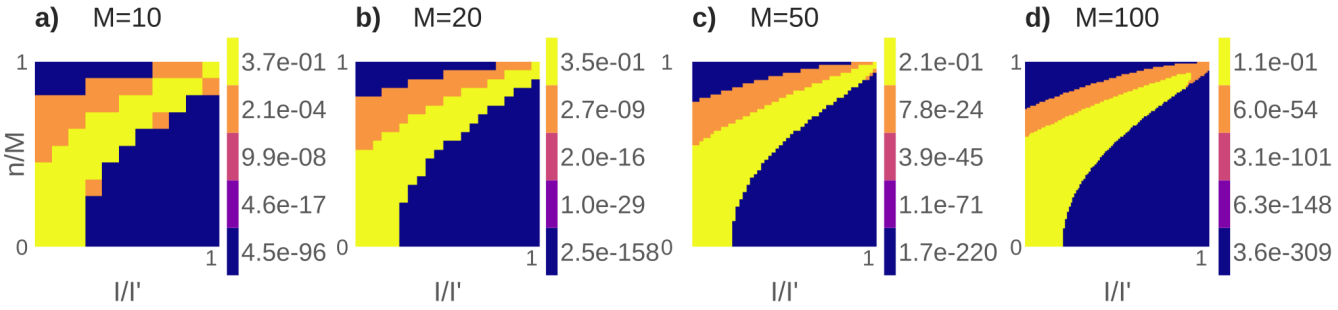

Figura 3.9: ***Heatmaps* de probabilidade de estados $c = (n, I)$.** A figura apresenta os percentis de probabilidades de estado para $M =$ **a):10**, **b)**:20, **c)**:50 e **d)**:100, com $\alpha = 40$ e $\sigma_2^0 = 0.1$.

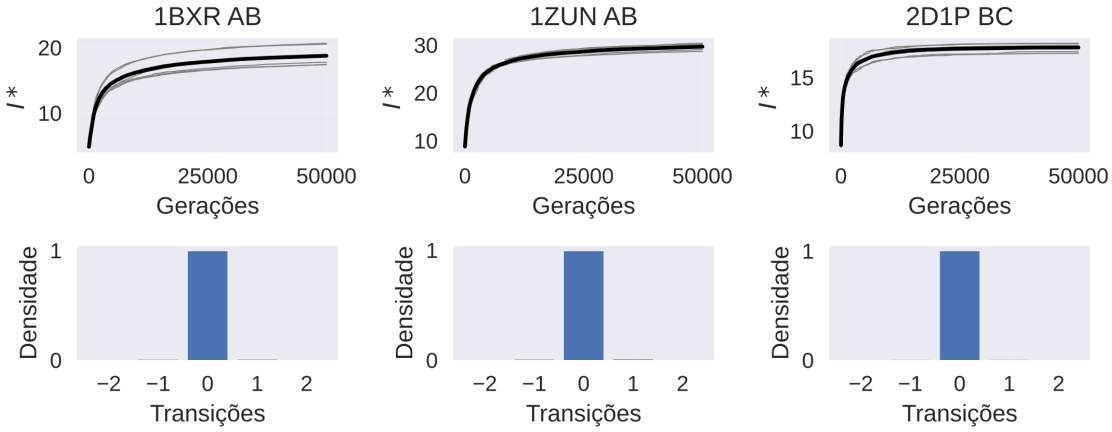

Figura 3.10: **Densidade de transições em n realizadas em simulações de algoritmo genético.** Durante as simulações de otimização realizadas sobre os sistemas 1BXR-AB, 1ZUN-AB e 2D1P-BC não foram computadas transições significativas em $n \pm 2$. As curvas em cinza representam as replicatas da simulação; a curva preta, a média das replicatas. O valor de I^* é o I mais alto a cada geração da simulação.

3.6 Probabilidade de um caminho dentro do algoritmo genético

Com a probabilidade dos estados $c = (n, I)$ e as regras de transição entre esses estados, definimos a matriz de transições do modelo estocástico de Markov e determinamos o caminho mais provável em um algoritmo genético. No exemplo exposto nas Figura 3.11, o caminho parte do estado $c = (n, I)$ em $t = 0$, que é o mais provável do sistema e possui valores baixos tanto de n, quanto de I, e realiza as transições de maior probabilidade a cada passo na otimização. Essa trajetória vai até o estado absorvente c_k, que é o estado

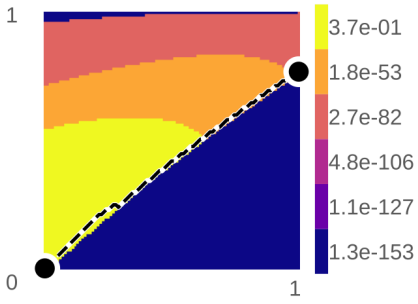

Figura 3.11: **Caminho mais provável entre os estados $c = (n, I)$ em um sistema de 100 sequências.** O *heatmap* retrata as probabilidades de estado para os parâmetros $I_0 = 40$, $I' = 50$ e $\alpha_0^2 = 0.001$. O caminho mais provável é dado pela linha pontilhada, que parte do estado mais provável $c = (n, I)$ e vai até o estado absorvente c_k

final e indica o TP alcançado pelo valores de n/M.

Ao calcular as probabilidades de transição em estados com probabilidades muito baixas, encontramos algumas limitações na resolução numérica. Assim, alguns dos valores representados no percentis mais baixos da Figuras 3.9 e 3.11 podem estar subestimados. Apesar de tais limitações, é notável a dependência das taxas de TP encontradas no modelo com os parâmetros de entrada M, α e σ_0^2, como mostraremos a seguir.

3.7 A influência de M, σ_0^2 e α no TP *rate*

Atentemo-nos, primeiramente, às entradas do modelo σ_0^2 e α, que são dados pela variância de I quando temos $n = 0$ e pela diferença entre I' e I_0, respectivamente. Na Figura 3.12 é mostrada a probabilidade dos estados I_n para sistemas com diferentes combinações de I_0 e σ_0^2, juntamente ao caminho mais provável em sua otimização via algoritmo genético. À medida que o valor de σ_0^2 aumenta, o peso de Poisson impulsiona a probabilidade de I em ns mais baixos, o que faz com que o TP diminua consideravelmente e a trajetória percorrida por entre os estados $c = (n, I)$ fique à deriva depois de perder o rastro dos picos das n-normais. Nota-se também que parece haver um razão inversa entre o aumento do TP *rate* e o valor de α. Nesse caso, os menores valores de α distribuem melhor a densidade de I nas n-normais (equação 2.6). A diluição das probabilidades oferece um tipo de guia à trajetória, que impede seu distanciamento dos estados que localizam-se nos auges das PDFs. Mais exemplos com $M = 10$ (Figura 5.10), $M = 20$ (Figura 5.11) e $M = 50$ (Figura 5.12) são exibidos no seção 5.3.

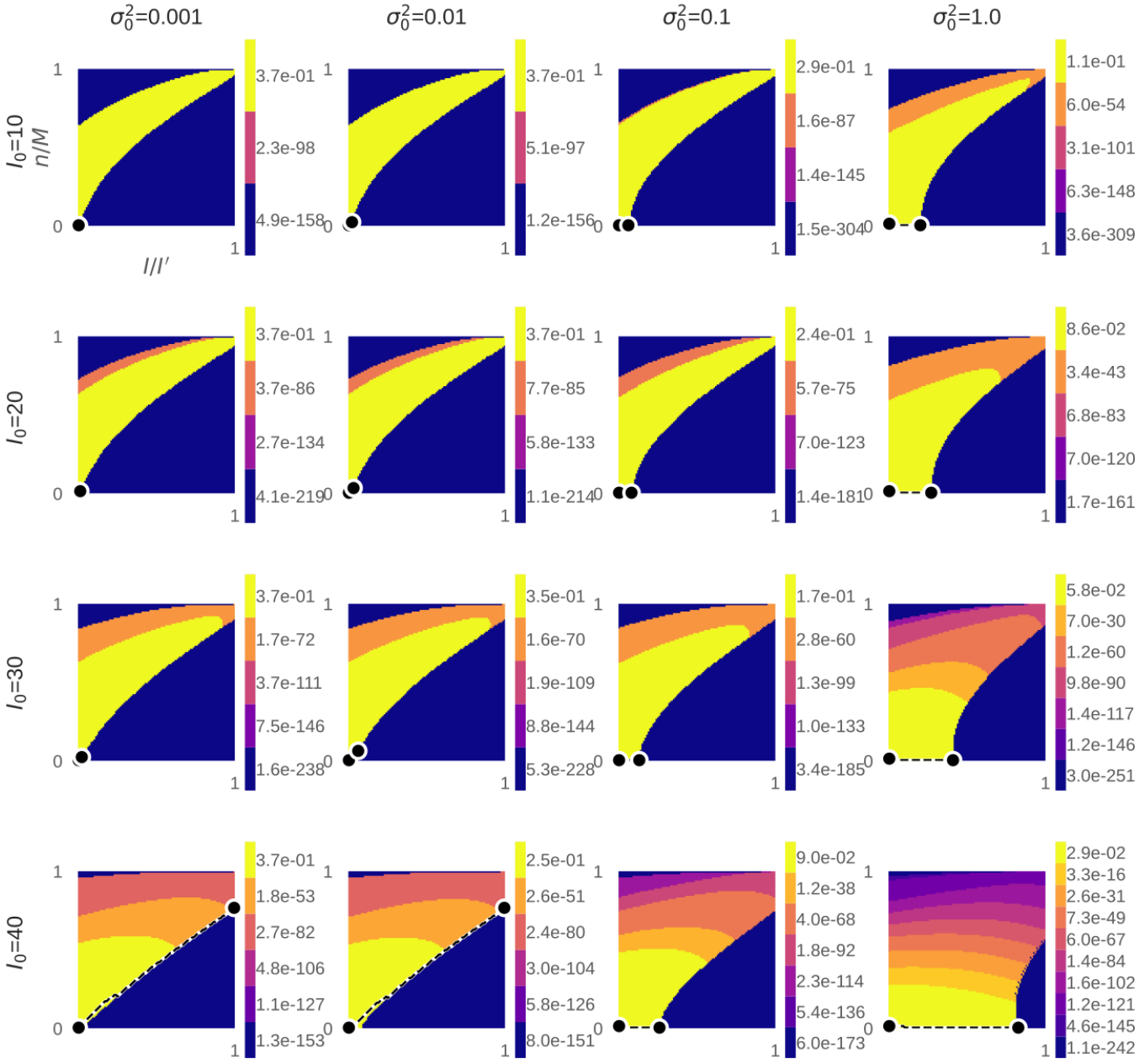

Figura 3.12: **Caminho mais provável entre os estados $c = (n, I)$ para diversas condições de α_0^2 e I_0, com $M = 100$.** Os *heatmaps* mostram as probabilidades de estados em um sistema com o valor de $I' = 50$. Nas linhas, os valores de I_0 aumentam de cima para baixo, enquanto, nas colunas, os valores de σ_0^2 aumentam da esquerda para a direita. O caminho mais provável é dado pela linha pontilhada, que parte do estado mais provável $c = (n, I)$ e vai até o estado absorvente c_k.

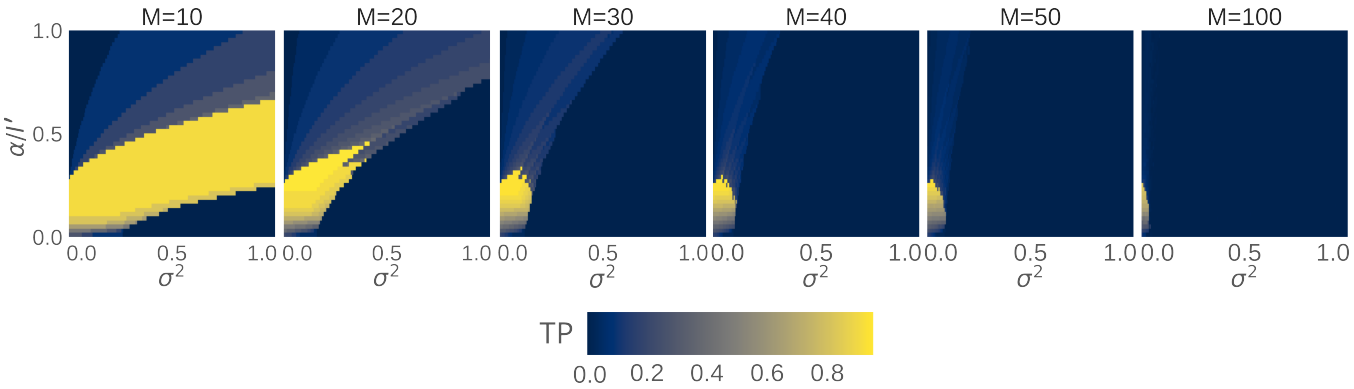

Figura 3.13: **Taxas de TP em função de σ_0^2 e α.** Os valores de TP foram calculados com as condições de $I' = 50$ e $M = 10$, 20, 30, 40, 50 e 100. Os pontos dos *heatmaps* têm passo de 0.005 no eixo σ_0^2 e de 1 no eixo α.

Outra variável a considerar é o número de sequências do sistema (M). Para avaliar sua influência, verificamos o comportamento dos valores de TP no domínio de α e σ_0^2, com diferentes valores de M. A Figura 3.13 mostra o resultado para M = 10, 20, 30, 40, 50 e 100. Embora a região com altos valores de TP para sistemas com 10 sequências seja generosa, é notável o quanto a faixa ótima dentro do domínio $\{\sigma_0^2, \alpha\}$ deteriora-se rapidamente conforme M aumenta. Isso se deve, principalmente, por conta do peso de Poisson (equação 2.10). O fatorial de M esmaga a probabilidade em ns mais altos e extingue as chances de evitar a deriva da trajetória em meio aos estados $c = (n, I)$. Assim, a partir de $M = 100$, o modelo mostra uma convergência para um cenário que virtualmente descarta a ocorrência de TPs significativos em todo o espaço de parâmetros.

Juntamente aos TPs simulados, são representados nas Figura 3.14 os TPs das simulações de otimização de informação coevolutiva realizadas com o sistema parálogo HK-RR – M = 10, 20, e 30 – e também os resultados das simulações com ortólogos – M = 100. Diferentemente dos resultados dos sistemas ortólogos (Tabela 3.2), que apresentam uma simples média sobre as 12 replicatas de otimização, os resultados do HK-RR (Tabela 3.1) referem-se às médias calculadas em 3 grupos: o primeiro, com 16 espécies e $M = 10$; o segundo, com 4 espécies e $M = 20$; e o terceiro, com 6 espécies e $M = 30$.

Quando comparamos os dados de predições do modelo com o dados das Tabelas 3.1 e 3.2, é possível afirmar que os resultados apoiam as previsões obtidas pelo modelo. Com exceção das famílias HK-RR com 10 ou menos cópias de proteínas por genoma, que alcançam taxas de TP significativas após a otimização I ($\geq 0,5$), todos os sistemas simulados com parâmetros fora da faixa ótima falham consistentemente em resolver

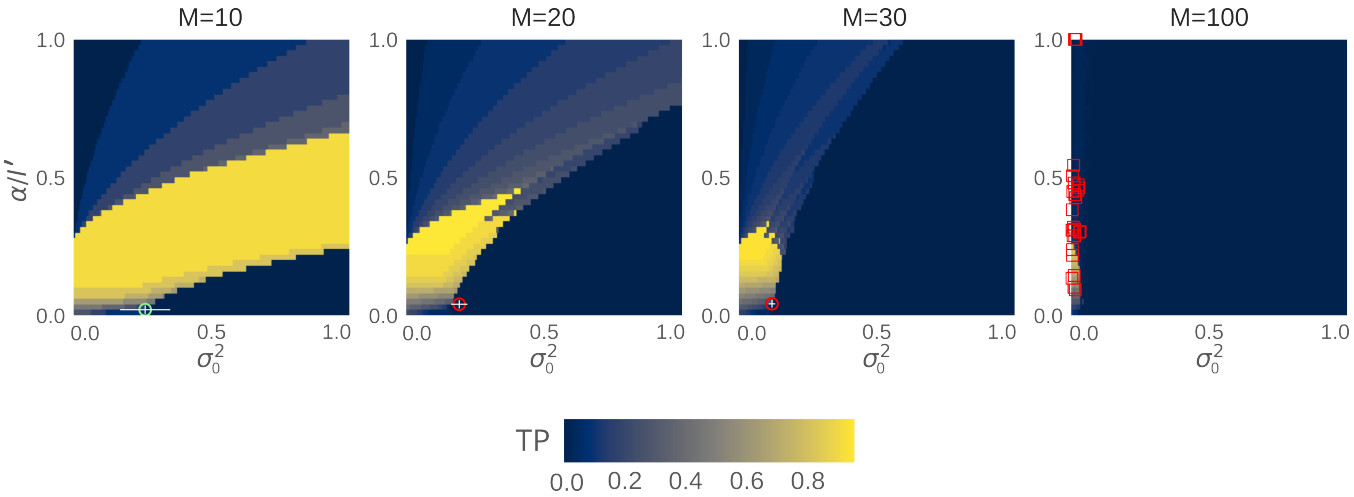

Figura 3.14: **Comparação entre TPs* dos sistemas e TPs do modelo no espaço de σ_0^2 e α.** A figura exibe a projeção dos pontos de sistemas reais sobre os *heatmaps* da Figura 3.13. Em $M = 10$, 20 e 30, os pontos são referentes à família HK-RR (Tabela 3.1) e sua representação é feita por meio de um círculo, com seus erros de σ_0^2 e α. Em $M = 100$, os pontos descrevem os resultados dos sistemas ortólogos (Tabela 3.2), com seus pontos representados por um quadrado. A cor dos pontos indica o TP* da simulação: os vermelhos indicam $TP < 0.5$; os verdes, $TP \geq 0.5$.

Tabela 3.1: Resultados da simulação de otimização do sistema HK-RR.

Sistema (Réplicas)	M	I'	I_0	σ_0^2	α/I'	I*(nat)	TP* (TP)	TP*$_{reav}$ (TP_{reav})
HK-RR(16)	10	15.913±9.960	14.861 ±11.110	0.263±0.010	0.021	15.790±10.050	0.500 (0.100±0.075)	0.687 (0.350±0.106)
HK-RR(4)	20	16.592±9.660	14.537 ±6.902	0.195±0.001	0.041	16.026±5.362	0.3125 (0.012±0.021)	0.500 (0.337±0.064)
HK-RR(6)	30	14.319±8.255	12.200 ±5.965	0.122±0.000	0.042	13.791±6.070	0.205 (0.016±0.016)	0.583 (0.333±0.083)

corretamente os parceiros proteicos. Convém observar que, embora o exemplo com 100 sequências na Figura 3.13 mostre uma estreita faixa com TPs expressivos para $M = 100$, esta região ruma ao desaparecimento à medida que M cresce até alcançar o número de sequências das famílias de ortólogos.

3.8 A reavaliação de sequências no domínio de M, σ_0^2 e α

Ao desconsiderar os erros de correspondência entre sequências semelhantes, foi possível atingir maiores TPs em condições de α e σ_0^2 que pareciam desfavoráveis. A Figura 3.15 mostra os resultado da reavaliação dos estados considerando $p = 0.25$. Quando deslocamos o estado mais provável para ns mais altos, permitimos que a trajetória siga o caminho

Tabela 3.2: Resultados das simulações de otimização realizadas sobre os sistemas ortólogos descritos na Tabela 2.1.

Sistema (Réplicas)	M	I'	I_0	σ_0^2	α/I'	I^*(nat)	TP^* (TP)	TP^*_{reav} (TP_{reav})
1B70-AB(12)	1108	35.671	11.761	0.011	0.480	25.989	0.004 (0.04)	0.295 (0.980)
1BXR-AB(12)	1004	19.179	4.717	0.007	0.289	15.965	0.006 (0.13)	0.312 (0.900)
1EFP-AB(12)	1347	65.050	14.305	0.018	1.015	46.174	0.006 (0.02)	0.362 (1.000)
1EP3-AB(12)	552	22.945	7.762	0.033	0.304	22.916	0.009 (0.03)	0.373 (0.900)
1I1Q-AB(12)	1204	9.752	3.014	0.003	0.135	12.449	0.001 (0.63)	0.200 (0.720)
1QOP-AB(12)	1155	14.741	2.773	0.005	0.239	13.380	0.002 (0.82)	0.215 (0.860)
1RM6-AB(12)	1604	19.673	4.212	0.003	0.309	14.228	0.001 (0.07)	0.158 (0.900)
1RM6-AC(12)	1534	27.484	5.020	0.006	0.449	20.867	0.003 (0.04	0.267 (0.970)
1RM6-BC(12)	1481	24.015	4.863	0.004	0.383	17.185	0.003 (0.04	0.316 (0.940)
1TYG-AB(12)	746	15.397	4.515	0.005	0.218	13.074	0.003 (0.79	0.272 (0.840)
1W85-AB(12)	1537	31.162	5.879	0.005	0.506	20.469	0.002 (0.03	0.315 (0.980)
1ZUN-AB(12)	649	23.727	8.612	0.030	0.302	26.153	0.011 (0.04	0.390 (0.900)
2D1P-BC(12)	216	15.541	8.268	0.011	0.145	15.948	0.010 (0.66	0.203 (0.740)
2NU9-AB(12)	798	29.521	5.879	0.026	0.473	27.721	0.011 (0.04	0.315 (0.980)
2VPZ-AB(12)	676	36.305	13.581	0.025	0.454	31.400	0.003 (0.04	0.166 (0.970)
2WDQ-CD(12)	221	7.864	3.192	0.015	0.093	9.145	0.018 (0.36	0.194 (0.590)
2Y69-AB(12)	1484	65.845	10.219	0.012	1.113	48.620	0.006 (0.01	0.500 (1.000)
2Y69-AC(12)	863	32.533	9.405	0.029	0.463	30.417	0.006 (0.03	0.388 (0.980)
3G5O-AB(12)	904	32.880	10.978	0.013	0.438	28.374	0.004 (0.04	0.313 (0.970)
3IP4-AB(12)	782	7.168	2.081	0.010	0.102	8.578	0.005 (0.42	0.160 (0.620)
3IP4-AC(12)	879	22.003	6.375	0.013	0.313	17.204	0.009 (0.09	0.336 (0.910)
3IP4-BC(12)	689	21.928	7.552	0.013	0.288	18.430	0.004 (0.08	0.248 (0.900)
3MML-AB(12)	1067	34.283	7.071	0.009	0.544	24.969	0.004 (0.03	0.390 (0.990)
3OAA-HG(12)	886	34.312	12.954	0.017	0.427	28.435	0.002 (0.05	0.206 (0.970)
3PNL-AB(12)	902	20.740	4.756	0.008	0.320	17.766	0.004 (0.09	0.303 (0.910)
3RRL-AB(12)	1330	62.525	7.554	0.015	1.099	47.312	0.006 (0.01	0.340 (1.000)

pelos picos das n-normais mais prontamente. Outros exemplos para $M = 10$ (Figura 5.13), $M = 20$ (Figura 5.14) e $M = 50$ (Figura 5.15) são mostrados nos resultados suplementares, seção 5.4.

Quando examinamos as taxas de TP no domínio de α e σ_0^2 para sequências reavaliadas, o modelo encontra taxas de TPs que são sistematicamente maiores do que as encontradas anteriomente. A Figura 3.16 apresenta a reavaliação das probabilidade para 10, 20, 30, 40, 50 e 100 sequências. Em comparação com a Figura 3.13, a faixa ótima de TP é consistentemente mais ampla para todos os valores de M. Como foi discutido anteriormente, a reavaliação das sequências causa uma mudança drástica na evolução temporal da variável estocástica devido à reponderação das probabilidades.

Muito embora a análise da reavaliação das sequências não pareça apresentar uma convergência completa no exemplo com $M = 100$ da Figura 3.16, a extrapolação dos dados para $M \geq 100$ sugere que soluções otimizadas com erros triviais devem estar confinadas, aproximadamente, na região do espaço de parâmetros ($\sigma_0^2 \geq 15, \alpha \leq 0.5$)(Figura 3.16; região demarcada pela linha tracejada). Essa conclusão é particularmente útil, pois nos fornece a localização potencial das soluções otimizadas com erros triviais no espaço de parâmetros, o que torna mais clara sua distinção estatística de outras soluções degeneradas. Como σ_0^2 e α podem ser razoavelmente conhecidos a partir de simulações de algoritmos genéticos iniciadas a partir de um conjunto de arranjos embaralhados, nosso achado quantitativo permite a classificação a *priori* de famílias de proteínas que podem ter os parceiros proteicos efetivamente resolvidos ao desconsiderar o erro entre sequências similares produzidos pela otimização do sinal coevolutivo (Figura 3.17).

3.9 Considerações Finais

A identificação de parceiros proteicos continua a ser um problema desafiador, que requer, idealmente, uma avaliação da energia livre de ligação pela utilização de *docking*[47] ou estudos atomísticos avançados[48]. Entretanto, quando tratamos de famílias com um grande número de proteínas, essa abordagem se torna inviável em níveis práticos. Esse cenário faz com que sinal deixado pela coevolução se torne uma alternativa interessante para ajudar inferência dos parceiros específicos.

A investigação das condições estatísticas que permitem previsões mais acertadas dos

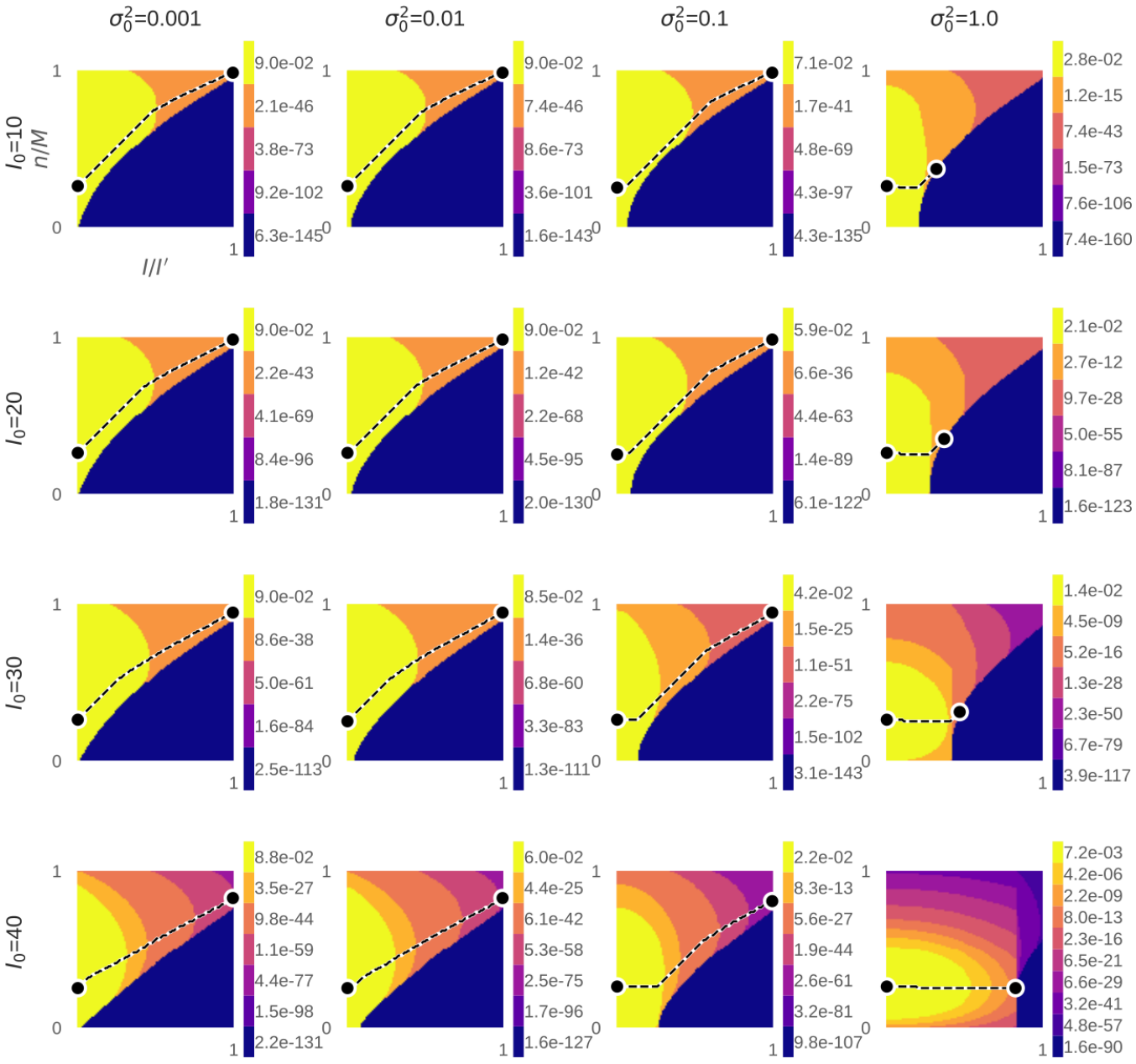

Figura 3.15: **Caminho mais provável entre os estados $c = (n, I)$ reavaliados para diversas condições de α_0^2 e I_0, com $M = 100$.** Os *heatmaps* mostram as probabilidades de estados em sistemas com o valor de $I' = 50$. Nas linhas, os valores de I_0 aumentam de cima para baixo, enquanto, nas colunas, os valores de σ_0^2 aumentam da esquerda para a direita. O caminho mais provável é dado pela linha pontilhada, que parte do estado mais provável $c = (n, I)$ e vai até o estado absorvente c_k.

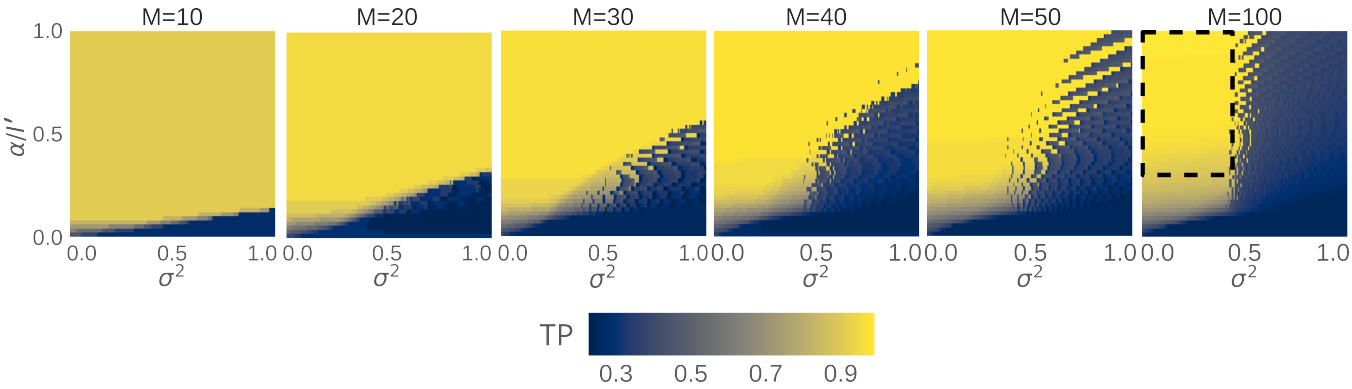

Figura 3.16: **Taxas de TP reavaliadas em função de σ_0^2 e α.** Os valores de TP foram calculados com as condições de $I' = 50$ e $M = 10, 20, 30, 40, 50$ e 100. Os pontos dos *heatmaps* têm passo de 0.005 no eixo σ_0^2 e de 1 no eixo α. A região prevista pelo modelo no espaço de parâmetros, que provavelmente contém soluções otimizadas com erros triviais, é indicada pela região tracejada.

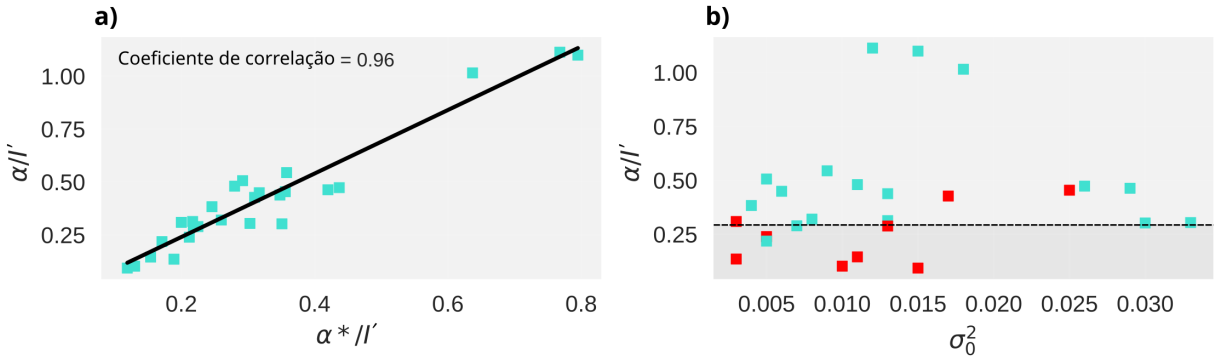

Figura 3.17: **Distinção estatística de soluções otimizadas com erros triviais. a)** Relação entre α e α^*, que foi calculado com base a informação coevolutiva otimizada I^* (Tabela 3.2). **b)** Localização das famílias de proteínas no espaço de parâmetros σ_0^2 e I_0. A maioria dos sistemas ortólogos, em que as taxas simuladas melhoraram em mais de 25% após a reavaliação de sequências similares (azul), encontra-se dentro da região prevista pelo modelo no espaço de parâmetros, provavelmente contendo soluções otimizadas com erros mínimos (linha tracejada). Para cada família de proteínas, σ_0^2 e I_0 foram estimados a partir de aproximadamente 8.000 arranjos embaralhados gerados aleatoriamente, com o número fixo de posições $n = 0$.

parceiros proteicos a partir da informação mútua é pertinente e nos ajuda a entender melhor o escopo de aplicabilidade desse método. É notável que modelo estocástico do algoritmo genético reteve características essenciais das simulações de otimização realizadas sobre sistemas reais de proteínas, sugerindo que taxas significativas de TP só são alcançadas para pequenos valores de M em domínios de parâmetros σ_0^2 e α específicos. Esse domínio é alargado quando levamos em consideração interações promíscuas entre proteínas por meio da reavaliação de sequências, em que desconsideramos erros que são cometidos entre sequências similares, dentro de um limite pré estabelecido.

Este estudo apresenta um avanço significativo na compreensão das condições estatísticas que permitem a predição de parceiros proteicos por meio da informação mútua e, até onde sabemos, é o primeiro estudo que investiga a produção estatística de pareamentos a partir de uma perspectiva de modelagem. Embora nosso modelo se baseie na maximização da informação coevolutiva via algoritmo genético, vale ressaltar que sua estrutura pode ser generalizada para a explorar outras heurísticas, como o algoritmo de Metropolis. Dessa forma, acreditamos que os resultados são de amplo interesse, pois os parâmetros σ_0^2 e α parecem ser críticos para abordagens coevolutivas em geral.

CAPÍTULO 4

Referências Bibliográficas

[1] Dobzhansky. "Nothing in Biology Makes Sense except in the Light of Evolution". Em: *The American Biology Teacher* 35.3 (1 de mar. de 1973), pp. 125–129. ISSN: 0002-7685. DOI: 10.2307/4444260. URL: https://doi.org/10.2307/4444260 (Acessado em 10/11/2021).

[2] Charles Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. 1st. London: John Murray., 1859.

[3] H. Allen Orr. "Darwin and Darwinism: The (Alleged) Social Implications of The Origin of Species". Em: *Genetics* 183.3 (nov. de 2009), pp. 767–772. ISSN: 0016-6731. DOI: 10.1534/genetics.109.110445. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2778974/ (Acessado em 09/03/2024).

[4] Julian Huxley. *Evolution: the modern synthesis*. 1942.

[5] Oswald T. Avery, Colin M. MacLeod e Maclyn McCarty. "STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES". Em: *The Journal of Experimental Medicine* 79.2 (1 de fev. de 1944), pp. 137–158. ISSN: 0022-1007. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2135445/ (Acessado em 20/01/2025).

[6] J. D. Watson e F. H. C. Crick. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". Em: *Nature* 171.4356 (abr. de 1953), pp. 737–738. ISSN: 1476-4687. DOI: 10.1038/171737a0. URL: https://www.nature.com/articles/171737a0 (Acessado em 20/01/2025).

[7] Theodosius Dobzhansky. "Genetics and the origin of species". Em: *RU Authors* (1 de jan. de 1951). URL: https://digitalcommons.rockefeller.edu/ru-authors/37.

[8] Sewall Wright. "EVOLUTION IN MENDELIAN POPULATIONS". Em: *Genetics* 16.2 (1 de mar. de 1931), pp. 97–159. ISSN: 1943-2631. DOI: 10.1093/genetics/16.2.97. URL: https://doi.org/10.1093/genetics/16.2.97 (Acessado em 20/01/2025).

[9] John Burdon Haldane. *The Causes of Evolution*. Princeton University Press, 10 de out. de 1990. 251 pp. ISBN: 9780691024424.

[10] Paul R. Ehrlich e Peter H. Raven. "Butterflies and Plants: A Study in Coevolution". Em: *Evolution* 18.4 (1964), pp. 586–608. ISSN: 0014-3820. DOI: 10.2307/2406212. URL: https://www.jstor.org/stable/2406212 (Acessado em 09/02/2024).

[11] L. Van Valen. "Molecular evolution as predicted by natural selection". Em: *Journal of Molecular Evolution* 3.2 (1974), pp. 89–101. ISSN: 0022-2844.

[12] Tanmay Dixit. "A synthesis of coevolution across levels of biological organization". Em: *Evolution* 78.2 (1 de fev. de 2024), pp. 211–220. ISSN: 0014-3820. DOI: 10.1093/evolut/qpad082. URL: https://doi.org/10.1093/evolut/qpad082 (Acessado em 22/02/2024).

[13] Simon C. Lovell e David L. Robertson. "An Integrated View of Molecular Coevolution in Protein–Protein Interactions". Em: *Molecular Biology and Evolution* 27.11 (1 de nov. de 2010), pp. 2567–2575. ISSN: 0737-4038. DOI: 10.1093/molbev/msq144. URL: https://doi.org/10.1093/molbev/msq144 (Acessado em 22/02/2024).

[14] S Jones e J M Thornton. "Principles of protein-protein interactions." Em: *Proceedings of the National Academy of Sciences of the United States of America* 93.1 (9 de jan. de 1996), pp. 13–20. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC40170/ (Acessado em 13/04/2024).

[15] Anne-Florence Bitbol et al. "Inferring interaction partners from protein sequences". Em: *Proceedings of the National Academy of Sciences* 113.43 (25 de out. de 2016), pp. 12180–12185. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1606762113. URL: https://www.pnas.org/content/113/43/12180 (Acessado em 24/05/2020).

[16] Michael P. H. Stumpf et al. "Evolution at the system level: the natural history of protein interaction networks". Em: *Trends in Ecology & Evolution* 22.7 (1 de jul. de 2007), pp. 366–373. ISSN: 0169-5347. DOI: 10.1016/j.tree.2007.04.004. URL: http://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(07)00131-0 (Acessado em 17/10/2017).

[17] Raphaël Champeimont et al. "Coevolution analysis of *Hepatitis C* virus genome to identify the structural and functional dependency network of viral proteins". Em: *Scientific Reports* 6 (20 de mai. de 2016), p. 26401. ISSN: 2045-2322. DOI: 10.1038/srep26401. URL: https://www.nature.com/articles/srep26401 (Acessado em 18/07/2018).

[18] Julian Echave, Stephanie J. Spielman e Claus O. Wilke. "Causes of evolutionary rate variation among protein sites". Em: *Nature Reviews Genetics* 17.2 (fev. de 2016), pp. 109–121. ISSN: 1471-0064. DOI: 10.1038/nrg.2015.18. URL: https://www.nature.com/articles/nrg.2015.18 (Acessado em 18/07/2018).

[19] William R. Taylor e Kerr Hatrick. "Compensating changes in protein multiple sequence alignments". Em: *Protein Engineering, Design and Selection* 7.3 (1 de mar. de 1994), pp. 341–348. ISSN: 1741-0126. DOI: 10.1093/protein/7.3.341. URL: https://academic.oup.com/peds/article/7/3/341/1469718 (Acessado em 18/07/2018).

[20] Gregory B. Gloor et al. "Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions". Em: *Biochemistry* 44.19 (1 de mai. de 2005), pp. 7156–7165. ISSN: 0006-2960. DOI: 10.1021/bi050293e. URL: https://doi.org/10.1021/bi050293e (Acessado em 16/11/2021).

[21] Duccio Malinverni e Alessandro Barducci. "Coevolutionary Analysis of Protein Sequences for Molecular Modeling". Em: *Biomolecular Simulations: Methods and Protocols*. Ed. por Massimiliano Bonomi e Carlo Camilloni. Methods in Molecular Biology. New York, NY: Springer, 2019, pp. 379–397. ISBN: 9781493996087. DOI: 10.1007/978-1-4939-9608-7_16. URL: https://doi.org/10.1007/978-1-4939-9608-7_16 (Acessado em 09/04/2020).

[22] David Ochoa e Florencio Pazos. "Practical aspects of protein co-evolution". Em: *Frontiers in Cell and Developmental Biology* 2 (22 de abr. de 2014). ISSN: 2296-634X. DOI: 10.3389/fcell.2014.00014. URL: https://www.frontiersin.org/journals/cell-and-developmental-biology/articles/10.3389/fcell.2014.00014/full (Acessado em 27/10/2024).

[23] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". Em: *Nature* 596.7873 (15 de jul. de 2021), p. 583. DOI: 10.1038/s41586-021-03819-2. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8371605/ (Acessado em 27/10/2024).

[24] Panagiotis Katsonis e Olivier Lichtarge. "A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness". Em: *Genome Research* 24.12 (dez. de 2014), p. 2050. DOI: 10.1101/gr.176214.114. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4248321/ (Acessado em 27/10/2024).

[25] Juan Xie et al. "Coevolution-based prediction of key allosteric residues for protein function regulation". Em: *eLife* 12 (17 de fev. de 2023). Ed. por Shozeb Haider, Volker Dötsch e Sarath Dantu, e81850. ISSN: 2050-084X. DOI: 10.7554/eLife.81850. URL: https://doi.org/10.7554/eLife.81850 (Acessado em 27/10/2024).

[26] Mohammed AlQuraishi. "End-to-End Differentiable Learning of Protein Structure". Em: *Cell Systems* 8.4 (24 de abr. de 2019), 292–301.e3. ISSN: 2405-4712. DOI: 10.1016/j.cels.2019.03.006. URL: https://www.sciencedirect.com/science/article/pii/S2405471219300766 (Acessado em 27/10/2024).

[27] Pierre C. Havugimana et al. "A census of human soluble protein complexes". Em: *Cell* 150.5 (31 de ago. de 2012), pp. 1068–1081. ISSN: 1097-4172. DOI: 10.1016/j.cell.2012.08.011.

[28] Gregory W. Clark et al. "Using coevolution to predict protein-protein interactions". Em: *Methods in Molecular Biology (Clifton, N.J.)* 781 (2011), pp. 237–256. ISSN: 1940-6029. DOI: 10.1007/978-1-61779-276-2_11.

[29] David Juan, Florencio Pazos e Alfonso Valencia. "High-confidence prediction of global interactomes based on genome-wide coevolutionary networks". Em: *Proceedings of the National Academy of Sciences* 105.3 (22 de jan. de 2008), pp. 934–939. DOI: 10.1073/pnas.0709671105. URL: https://www.pnas.org/doi/10.1073/pnas.0709671105 (Acessado em 27/10/2024).

[30] David de Juan, Florencio Pazos e Alfonso Valencia. "Emerging methods in protein co-evolution". Em: *Nature Reviews Genetics* 14.4 (abr. de 2013), pp. 249–261. ISSN: 1471-0064. DOI: 10.1038/nrg3414. URL: https://www.nature.com/articles/nrg3414 (Acessado em 22/02/2024).

[31] Faruck Morcos et al. "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". Em: *Proceedings of the National Academy of Sciences* 108.49 (6 de dez. de 2011), E1293–E1301. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1111471108. URL: https://www.pnas.org/content/108/49/E1293 (Acessado em 09/04/2020).

[32] Francisco M. Codoñer e Mario A. Fares. "Why Should We Care about Molecular Coevolution?" Em: *Evolutionary Bioinformatics* 4 (1 de jan. de 2008), p. 117693430800400003. ISSN: 1176-9343. DOI: 10.1177/117693430800400003. URL: https://doi.org/10.1177/117693430800400003 (Acessado em 17/03/2024).

[33] Miguel Andrade, Camila Pontes e Werner Treptow. "Coevolutive, evolutive and stochastic information in protein-protein interactions". Em: *Computational and Structural Biotechnology Journal* 17 (20 de nov. de 2019), pp. 1429–1435. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2019.10.005. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6906720/ (Acessado em 16/11/2021).

[34] Thomas M. Cover e Joy A. Thomas. *Elements of Information Theory, 2nd Edition*. Wiley, 2006. ISBN: 978-0-471-24195-9. (Acessado em 24/01/2024).

[35] Miguel Correa Marrero et al. "Improved inference of intermolecular contacts through protein–protein interaction prediction using coevolutionary analysis". Em: *Bioinformatics* 35.12 (15 de jun. de 2019), pp. 2036–2042. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty924. URL: https://doi.org/10.1093/bioinformatics/bty924 (Acessado em 13/04/2024).

[36] Thomas Gueudré et al. "Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis". Em: *Proceedings of the National Academy of Sciences* 113.43 (25 de out. de 2016), pp. 12186–12191. DOI: 10.1073/pnas.1607570113. URL: https://www.pnas.org/doi/full/10.1073/pnas.1607570113 (Acessado em 13/04/2024).

[37] Alfonso Valencia e Florencio Pazos. "Computational methods for the prediction of protein interactions". Em: *Current Opinion in Structural Biology* 12.3 (1 de jun. de 2002), pp. 368–373. ISSN: 0959-440X. DOI: 10.1016/S0959-440X(02)00333-0. URL: http://www.sciencedirect.com/science/article/pii/S0959440X02003330 (Acessado em 13/10/2017).

[38] Camila Pontes et al. "Trivial and nontrivial error sources account for misidentification of protein partners in mutual information approaches". Em: *Scientific Reports* 11.1 (25 de mar. de 2021), p. 6902. ISSN: 2045-2322. DOI: 10.1038/s41598-021-86455-0. URL: https://www.nature.com/articles/s41598-021-86455-0 (Acessado em 26/01/2024).

[39] Emmanuel D. Levy, Subhajyoti De e Sarah A. Teichmann. "Cellular crowding imposes global constraints on the chemistry and evolution of proteomes". Em: *Proceedings of the National Academy of Sciences of the United States of America* 109.50 (11 de dez. de 2012), pp. 20461–20466. ISSN: 1091-6490. DOI: 10.1073/pnas.1209312109.

[40] Muyoung Heo, Sergei Maslov e Eugene Shakhnovich. "Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions". Em: *Proceedings of the National Academy of Sciences of the United States of America* 108.10 (8 de mar. de 2011), pp. 4258–4263. ISSN: 1091-6490. DOI: 10.1073/pnas.1009392108.

[41] Zheng Zhang et al. "Phage protein receptors have multiple interaction partners and high expressions". Em: *Bioinformatics* 36.10 (15 de mai. de 2020), pp. 2975–2979. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa123. URL: https://doi.org/10.1093/bioinformatics/btaa123 (Acessado em 03/02/2025).

[42] Charlotte Nicod, Amir Banaei-Esfahani e Ben C Collins. "Elucidation of host–pathogen protein–protein interactions to uncover mechanisms of host cell rewiring". Em: *Current Opinion in Microbiology*. Antimicrobials * Bacterial Systems Biology 39 (1 de out. de 2017), pp. 7–15. ISSN: 1369-5274. DOI: 10.1016/j.mib.2017.07.005. URL: https://www.sciencedirect.com/science/article/pii/S1369527417301029 (Acessado em 03/02/2025).

[43] Qiwen Liao et al. "Cnidarian peptide neurotoxins: a new source of various ion channel modulators or blockers against central nervous systems disease". Em: *Drug Discovery Today* 24.1 (1 de jan. de 2019), pp. 189–197. ISSN: 1359-6446. DOI: 10.1016/j.drudis.2018.08.011. URL: https://www.sciencedirect.com/science/article/pii/S1359644618301673 (Acessado em 03/02/2025).

[44] Anne-Florence Bitbol. "Inferring interaction partners from protein sequences using mutual information". Em: *PLOS Computational Biology* 14.11 (13 de nov. de 2018), e1006401. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006401. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006401 (Acessado em 24/05/2020).

[45] Sergey Ovchinnikov, Hetunandan Kamisetty e David Baker. "Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information". Em: *eLife* 3 (1 de mai. de 2014), e02030. ISSN: 2050-084X. DOI: 10.7554/eLife.02030. URL: https://elifesciences.org/articles/02030 (Acessado em 04/12/2017).

[46] RCSB Protein Data Bank. *RCSB PDB: Homepage*. URL: https://www.rcsb.org/ (Acessado em 07/11/2024).

[47] Prithviraj Nandigrami et al. "Computational Assessment of Protein–Protein Binding Specificity within a Family of Synaptic Surface Receptors". Em: *The Journal of Physical Chemistry B* 126.39 (6 de out. de 2022), pp. 7510–7527. ISSN: 1520-6106. DOI: 10.1021/acs.jpcb.2c02173. URL: https://doi.org/10.1021/acs.jpcb.2c02173 (Acessado em 03/02/2025).

[48] Siddharth Bhadra-Lobo, Georgy Derevyanko e Guillaume Lamoureux. "Dock2D: Synthetic Data for the Molecular Recognition Problem". Em: *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 21.6 (30 de mai. de 2024), pp. 2580–2586. ISSN: 1545-5963. DOI: 10.1109/TCBB.2024.3407477. URL: https://dl.acm.org/doi/10.1109/TCBB.2024.3407477 (Acessado em 03/02/2025).

CAPÍTULO 5

Resultados Suplementares

5.1 Função de densidade de probabilidade

Figuras 5.1, 5.2, 5.3.

Figuras 5.4, 5.5 e 5.6.

5.2 Reavaliação de sequências

Figuras 5.7, 5.8 e 5.9

5.3 Probabilidade de um caminho dentro do algoritmo genético

. Figuras 5.10, 5.11 e 5.12

5.4 A reavaliação de sequências no domínio de M, σ_0^2 e α

Figuras 5.13, 5.14 e 5.15.

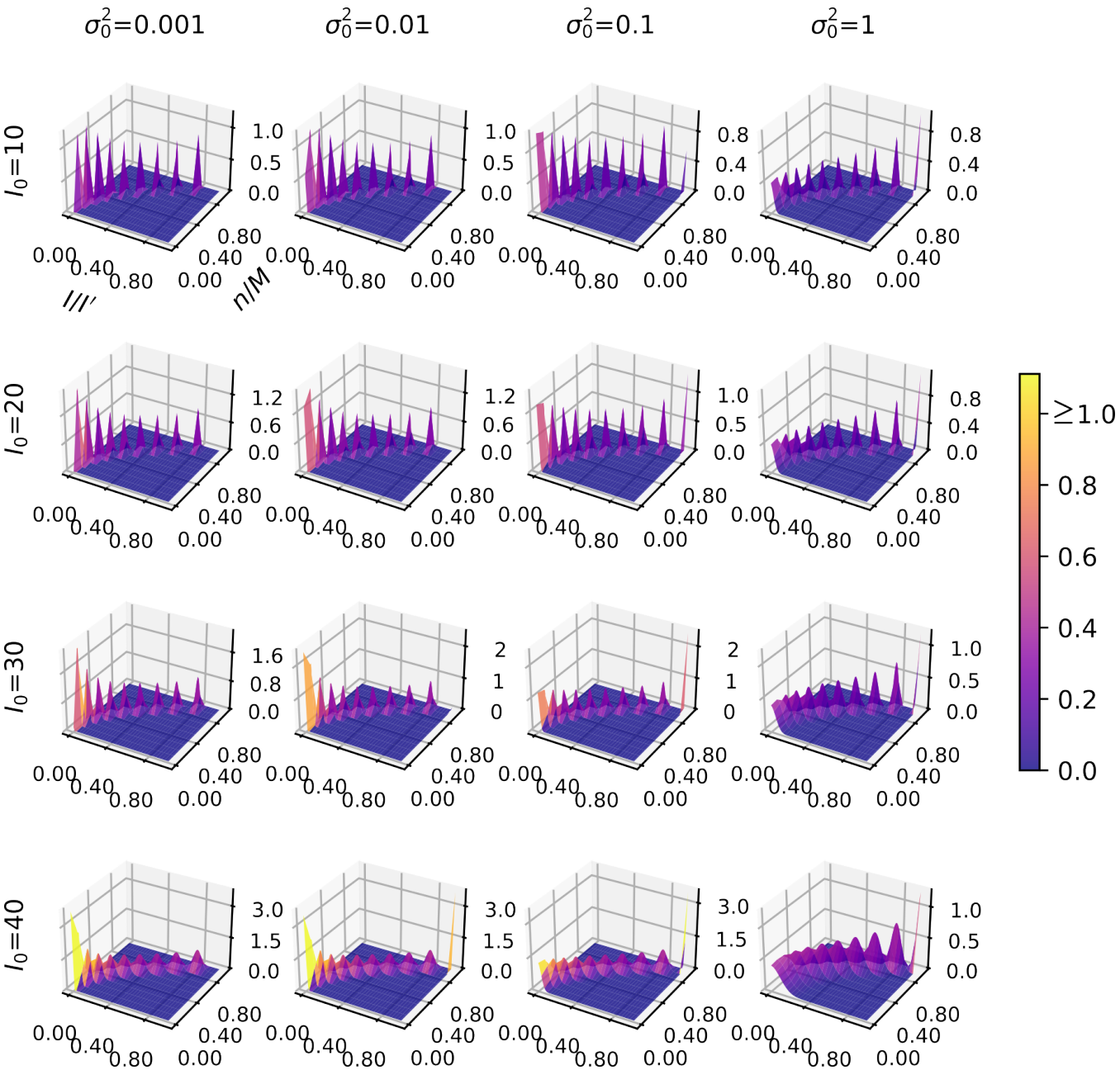

Figura 5.1: **Densidade de probabilidade;** $M = 10$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$. Nas linhas, os valores de I_0 aumentam de cima para baixo, enquanto, nas colunas, os valores de σ_0^2 aumentam da esquerda para a direita.

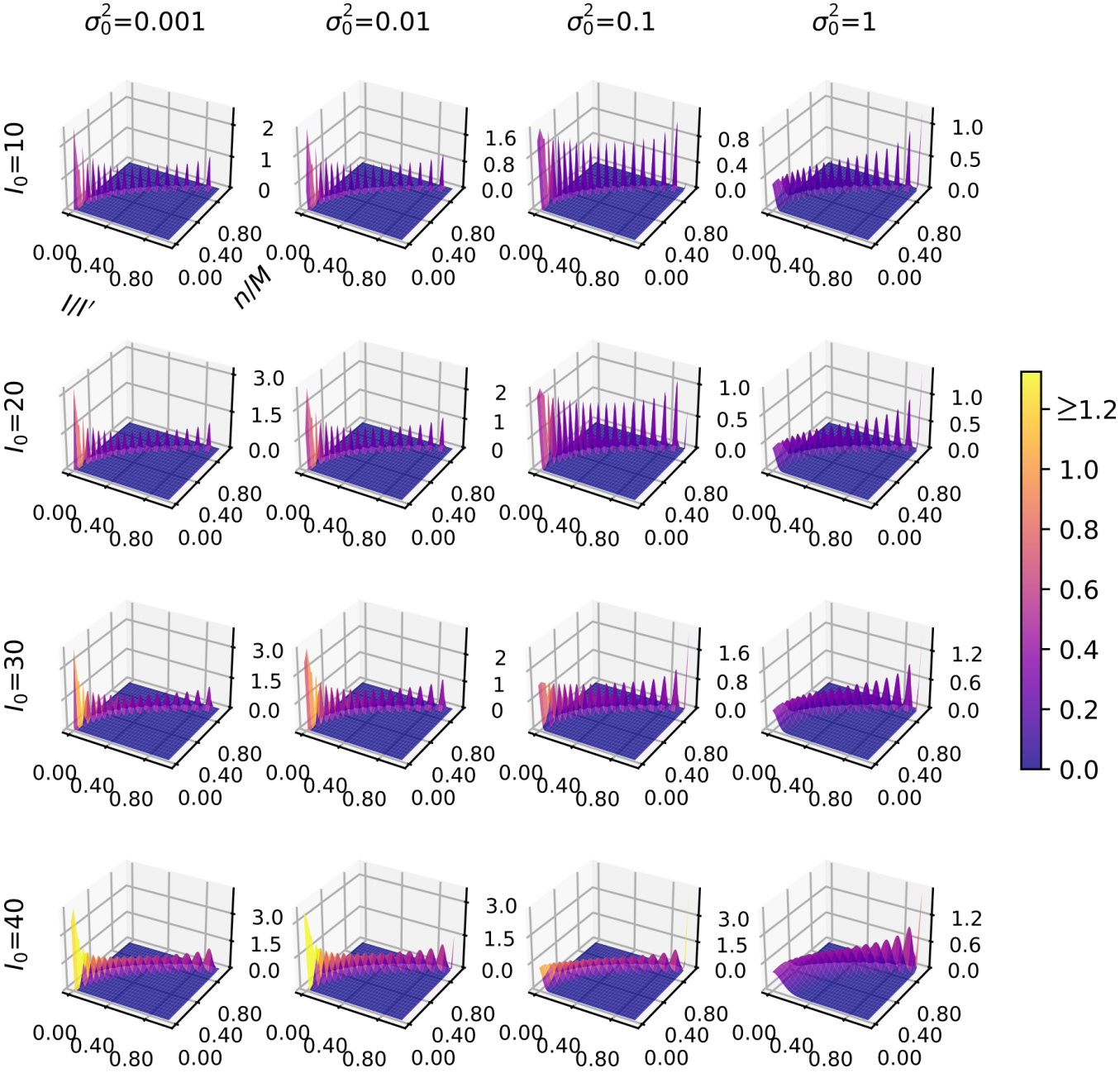

Figura 5.2: **Densidade de probabilidade;** $M = 20$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$.

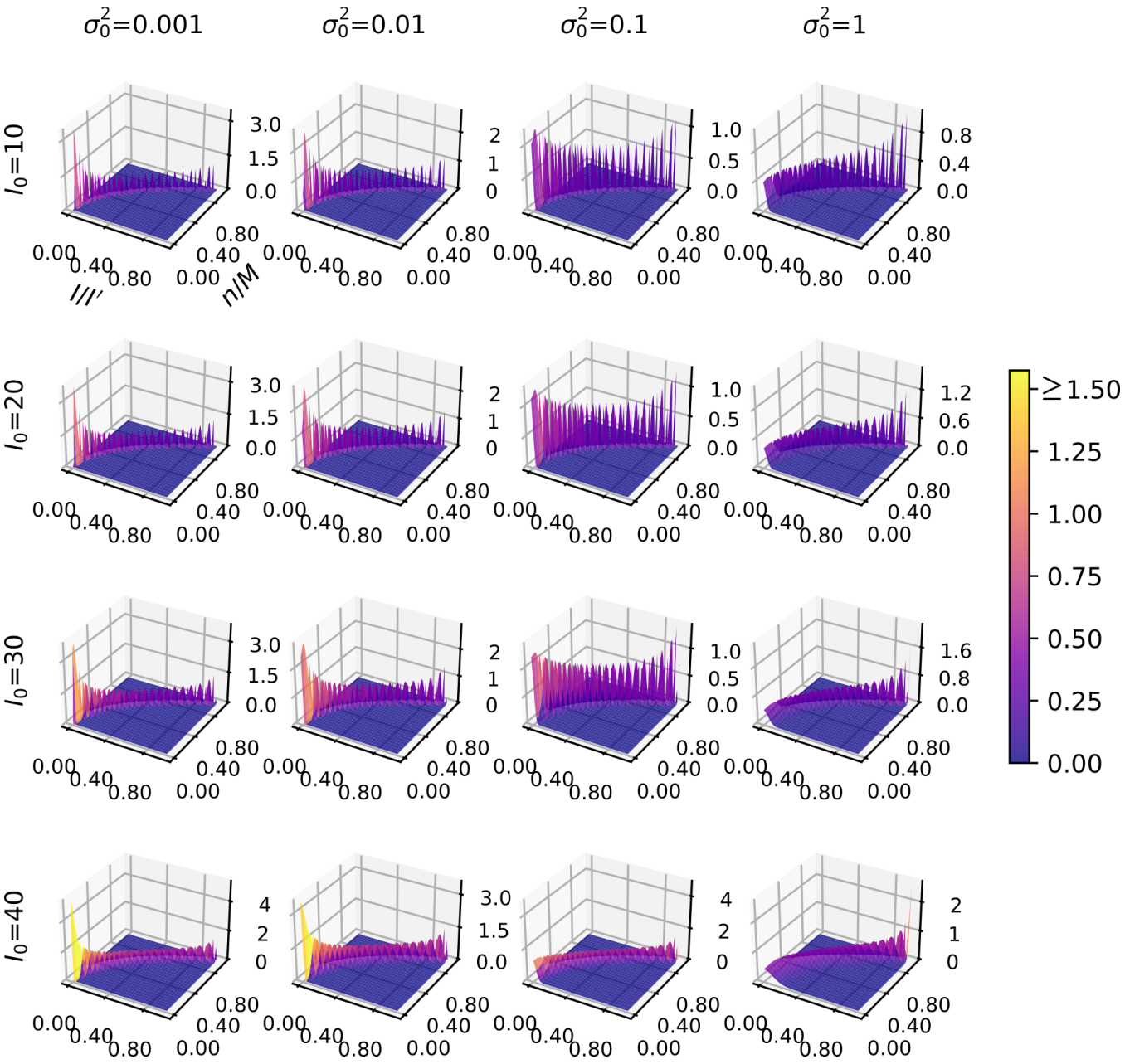

Figura 5.3: **Densidade de probabilidade;** $M = 50$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$.

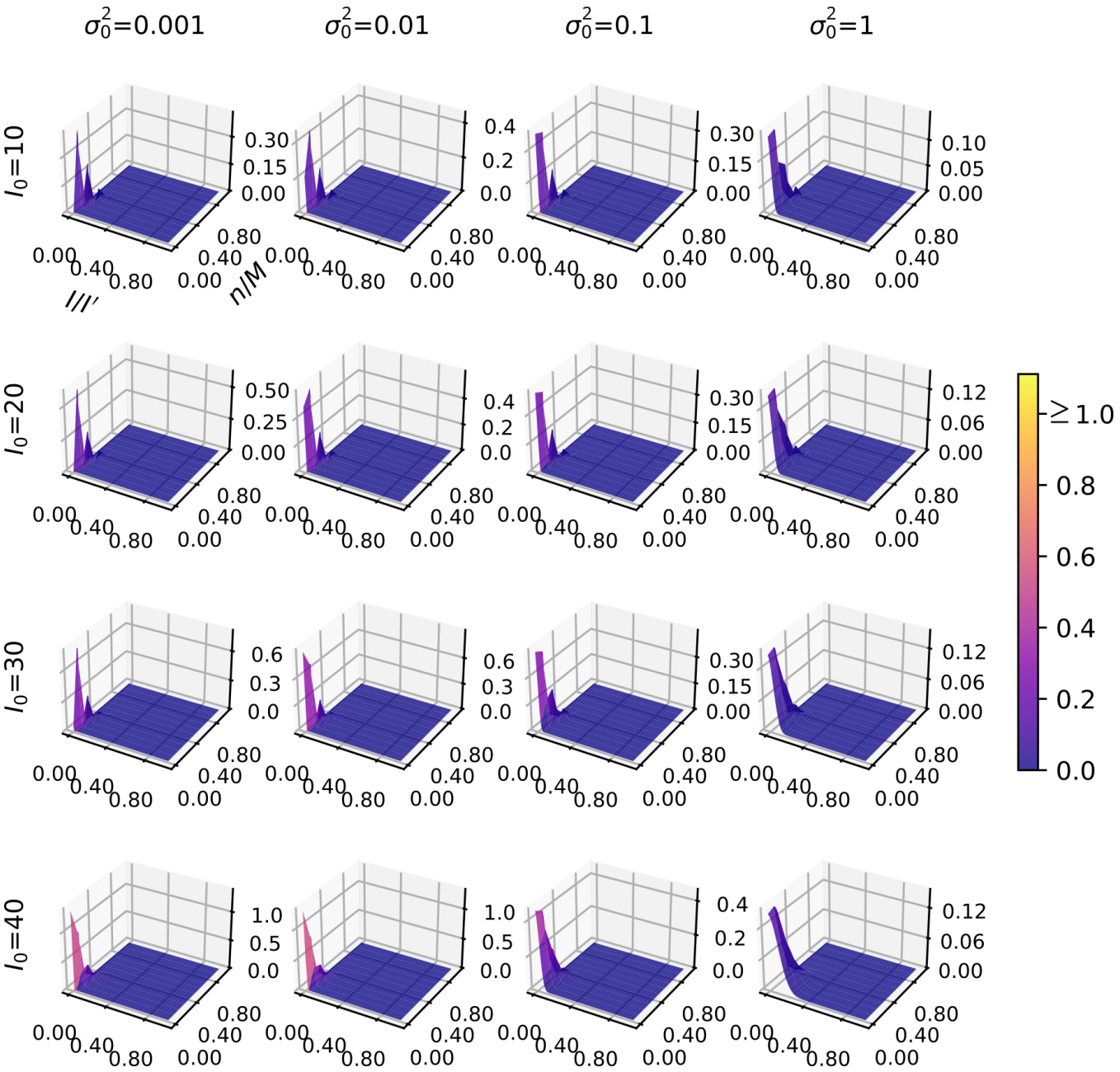

Figura 5.4: **Densidade de probabilidade ponderada por Poisson;** $M = 10$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$. Nas linhas, os valores de I_0 aumentam de cima para baixo, enquanto, nas colunas, os valores de σ_0^2 aumentam da esquerda para a direita.

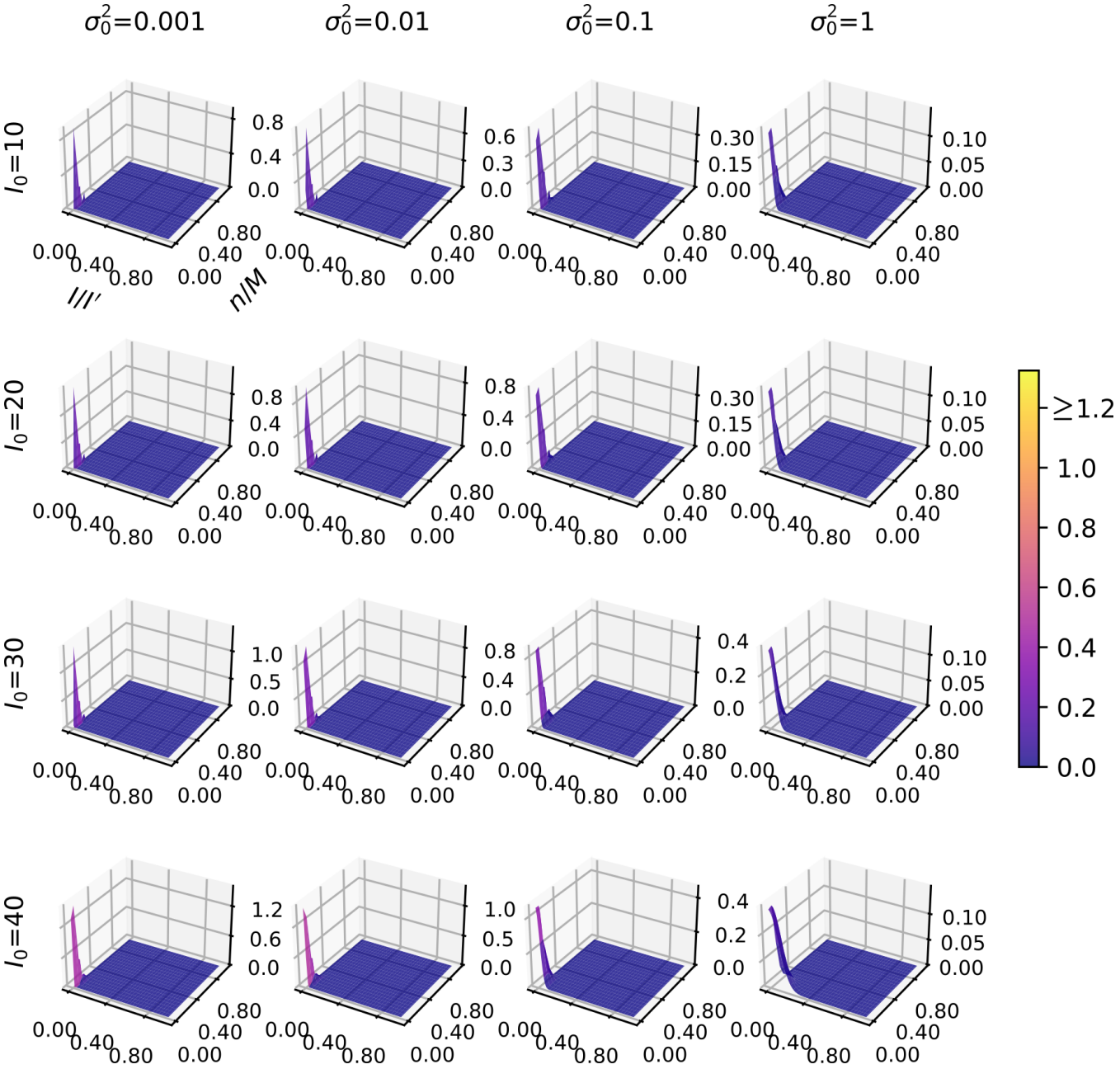

Figura 5.5: **Densidade de probabilidade ponderada por Poisson;** $M = 20$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$.

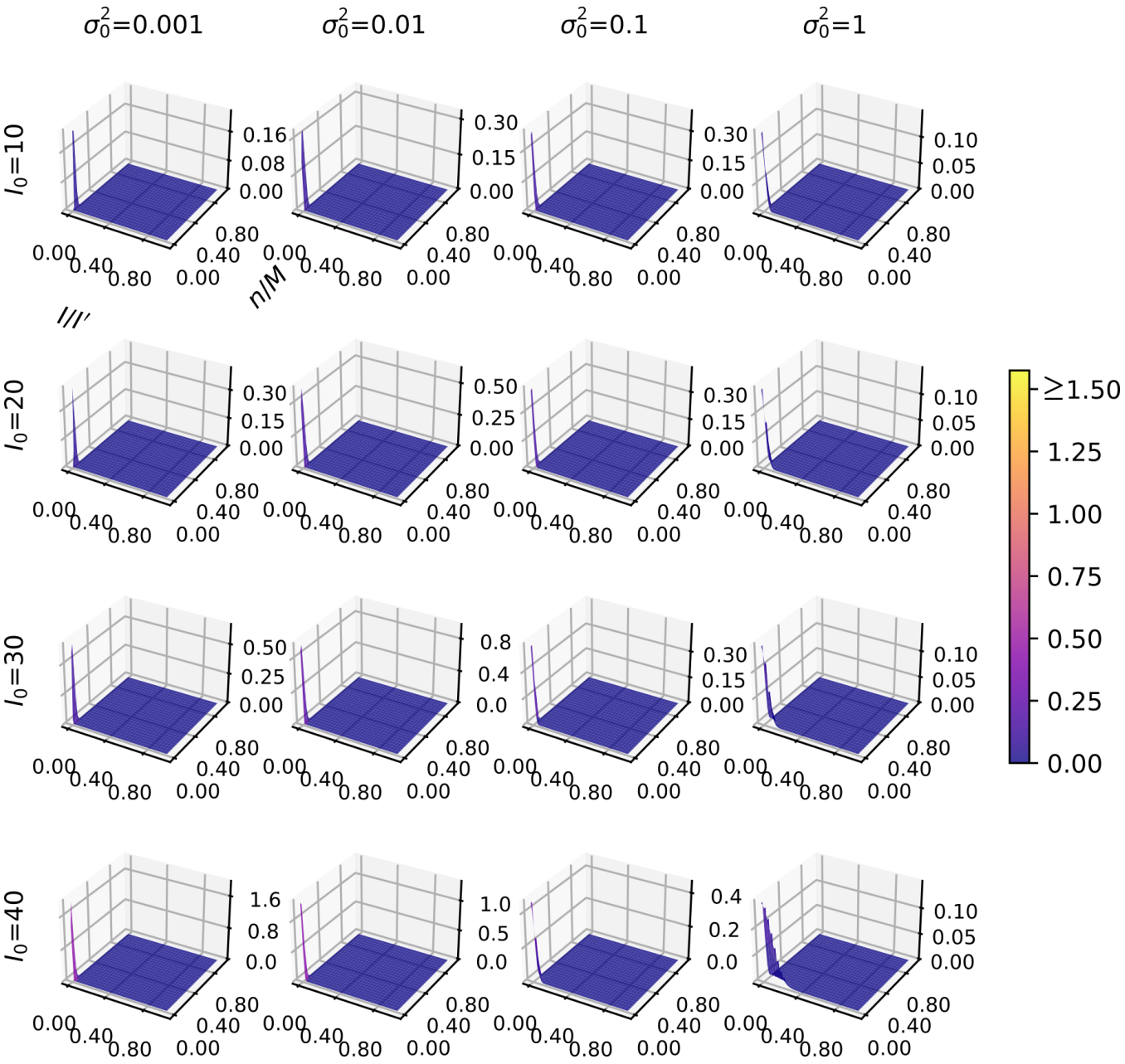

Figura 5.6: **Densidade de probabilidade ponderada por Poisson;** $M = 50$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$.

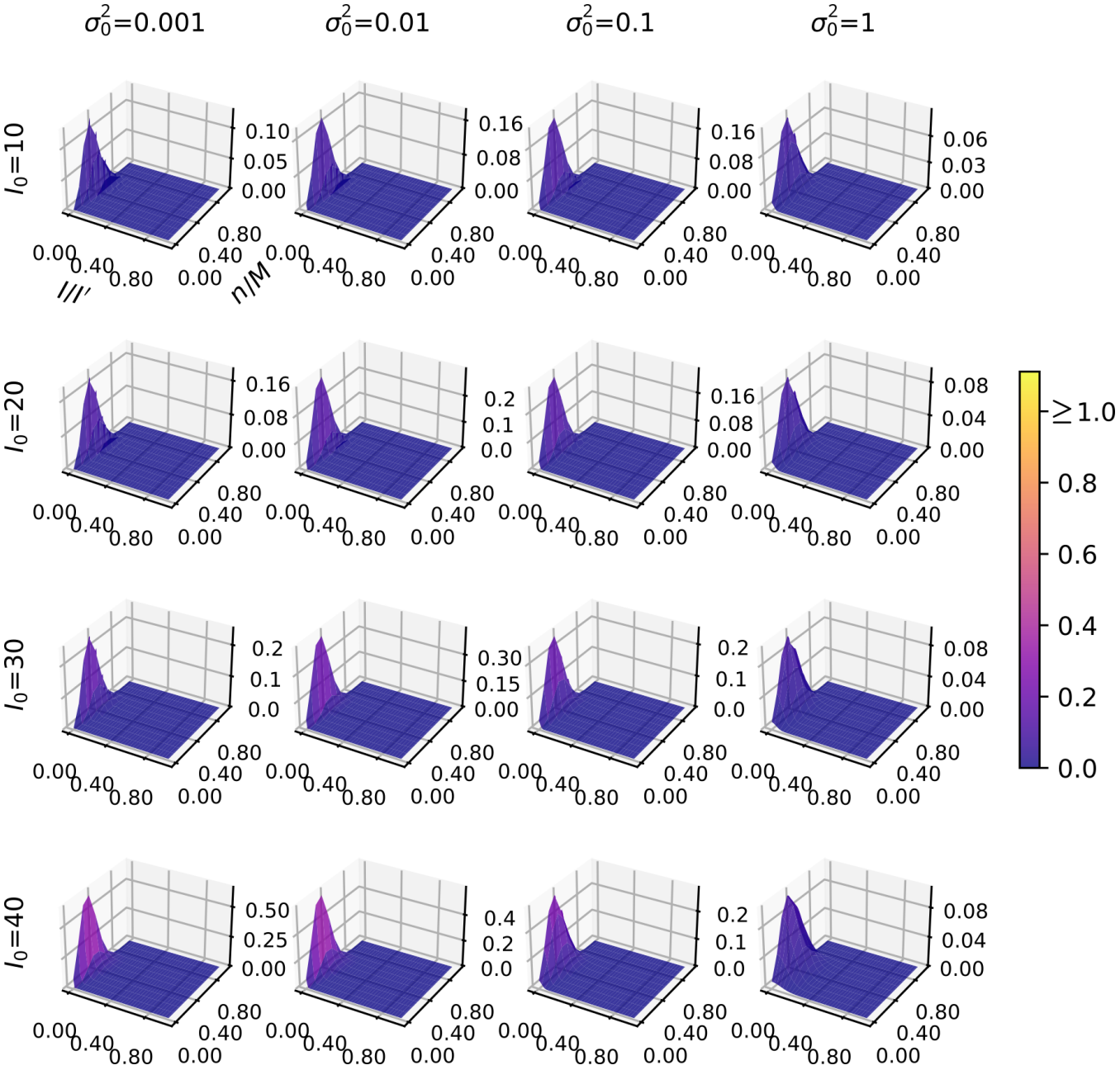

Figura 5.7: **Reavaliação da densidade de probabilidade ponderada por Poisson;** $M = 10$ **e** $p = 0.25$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$.

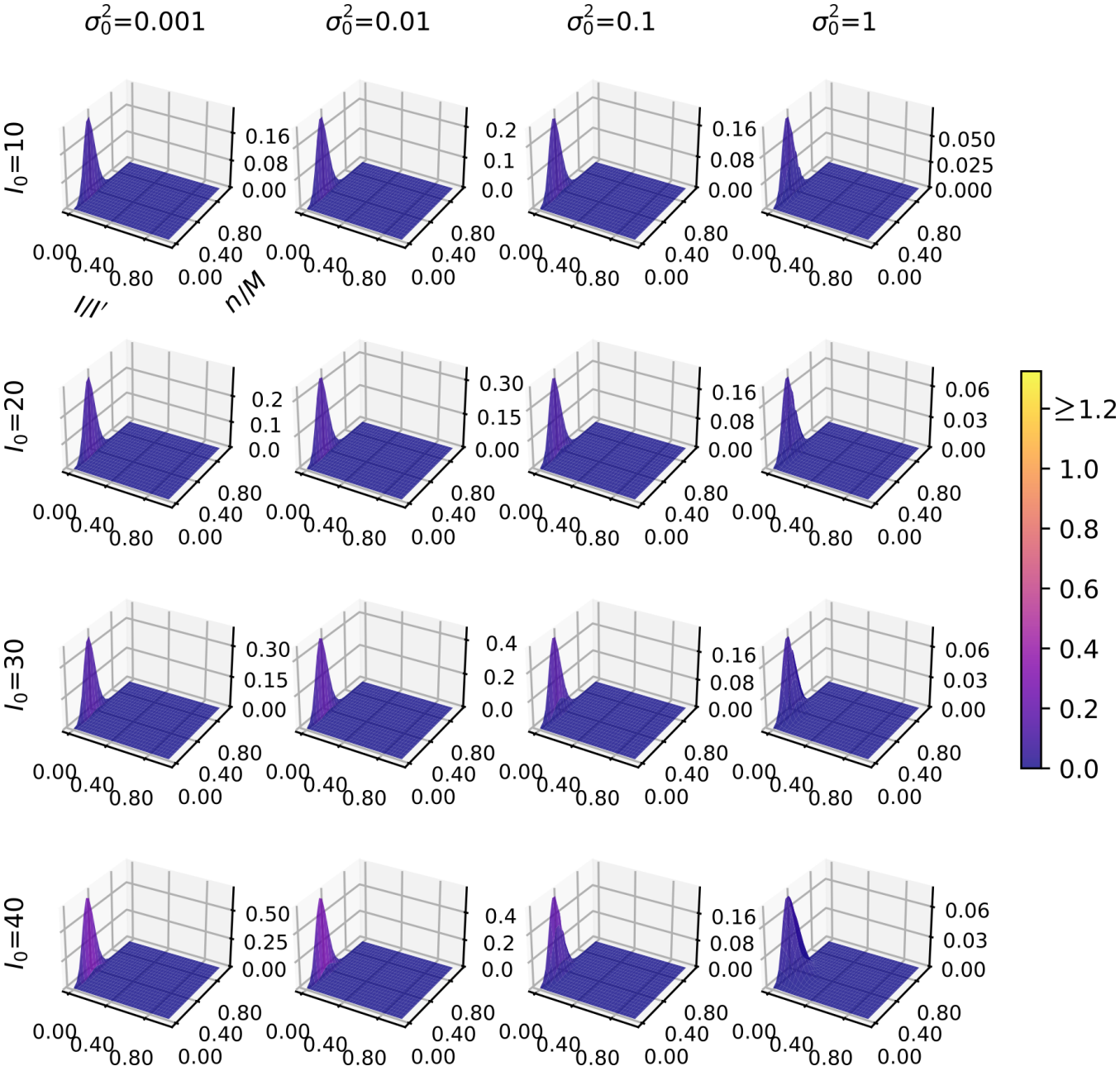

Figura 5.8: **Reavaliação da densidade de probabilidade ponderada por Poisson;** $M = 20$ **e** $p = 0.25$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$.

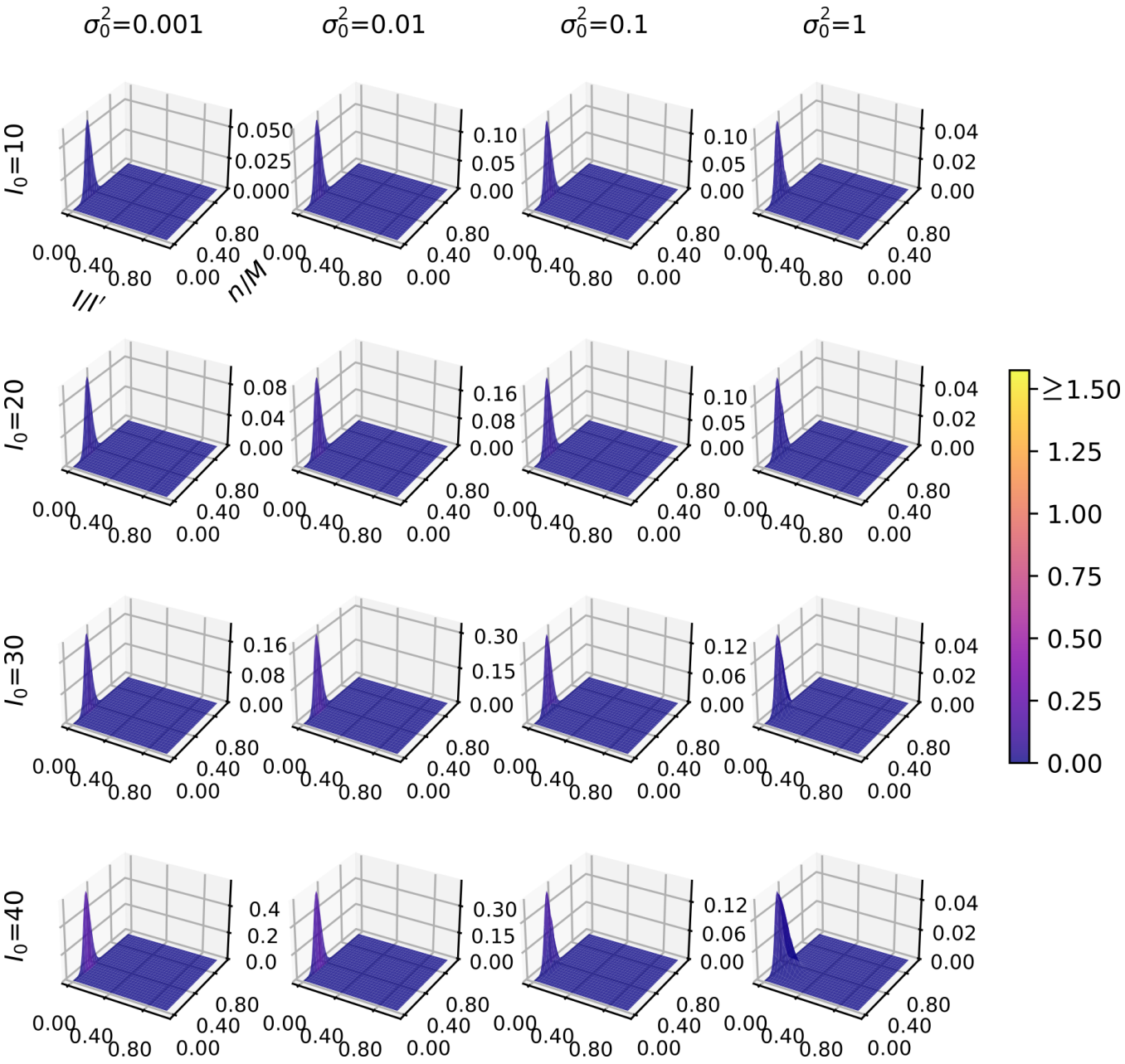

Figura 5.9: **Reavaliação da densidade de probabilidade ponderada por Poisson;** $M = 50$ **e** $p = 0.25$**.** Os gráficos consideram diversas condições de I_0 e σ_0^2 ($f_n(I|\theta_n)$), com $I' = 50$.

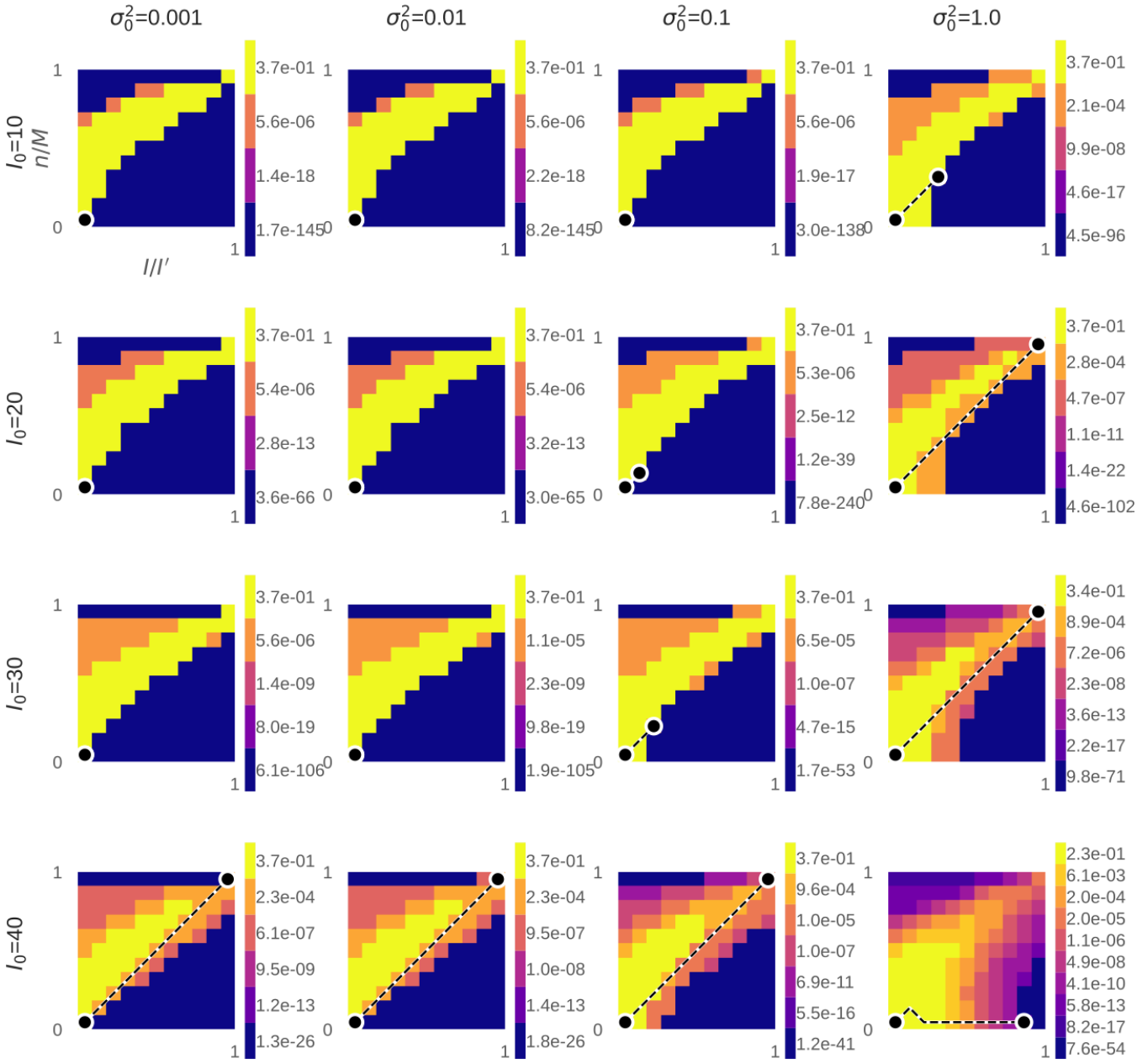

Figura 5.10: **Caminho mais provável entre os estados $c = (n, I)$ reavaliados para diversas condições de α_0^2 e I_0, com $M = 10$.** Os *heatmaps* mostram as probabilidades de estados em sistemas com o valor de $I' = 50$. O caminho mais provável é dado pela linha pontilhada, que parte do estado mais provável $c = (n, I)$ e vai até o estado absorvente c_k.

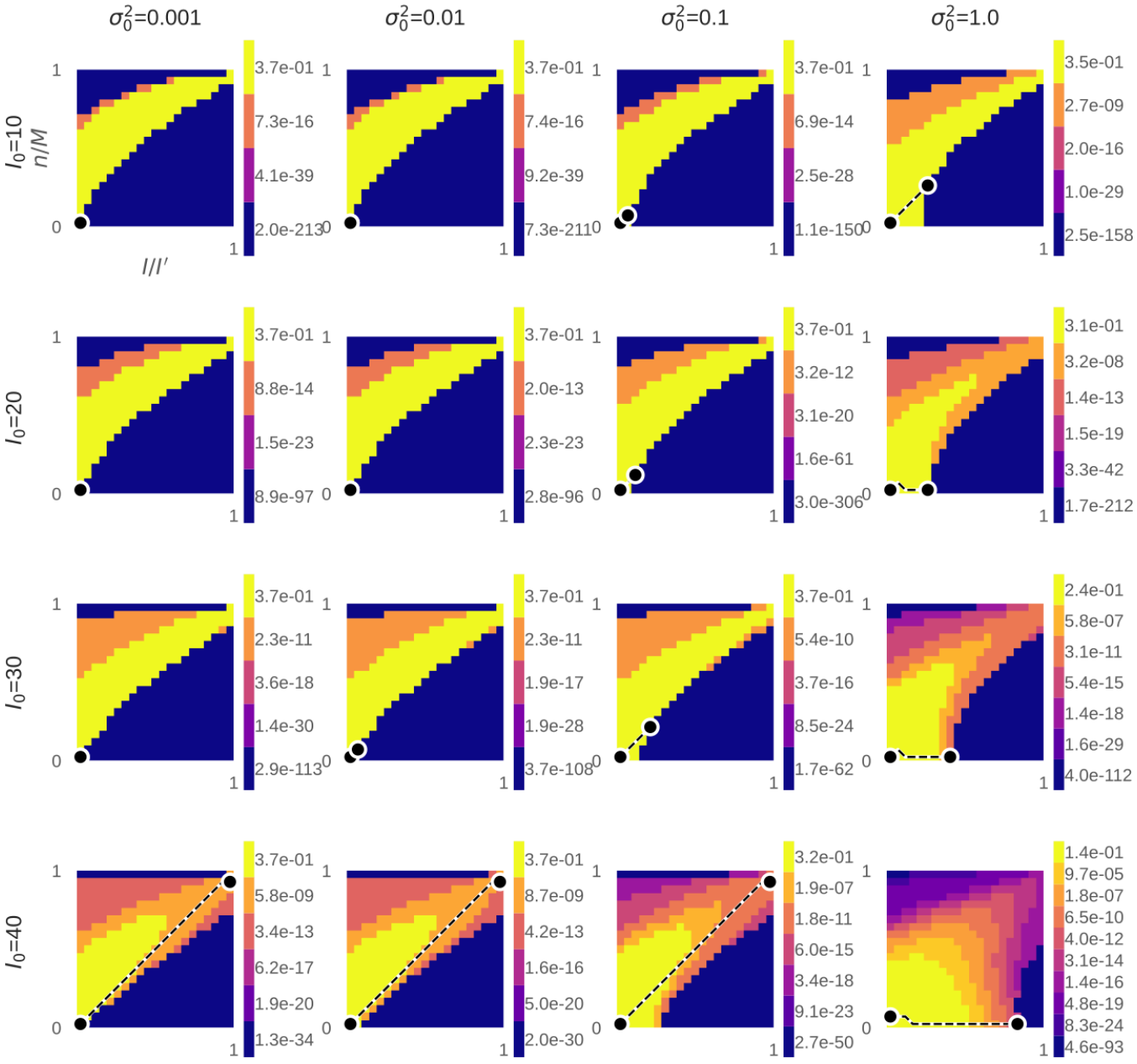

Figura 5.11: **Caminho mais provável entre os estados $c = (n, I)$ para diversas condições de α_0^2 e I_0, com $M = 20$.** Os *heatmaps* mostram as probabilidades de estados em sistemas com o valor de $I' = 50$. O caminho mais provável é dado pela linha pontilhada, que parte do estado mais provável $c = (n, I)$ e vai até o estado absorvente c_k.

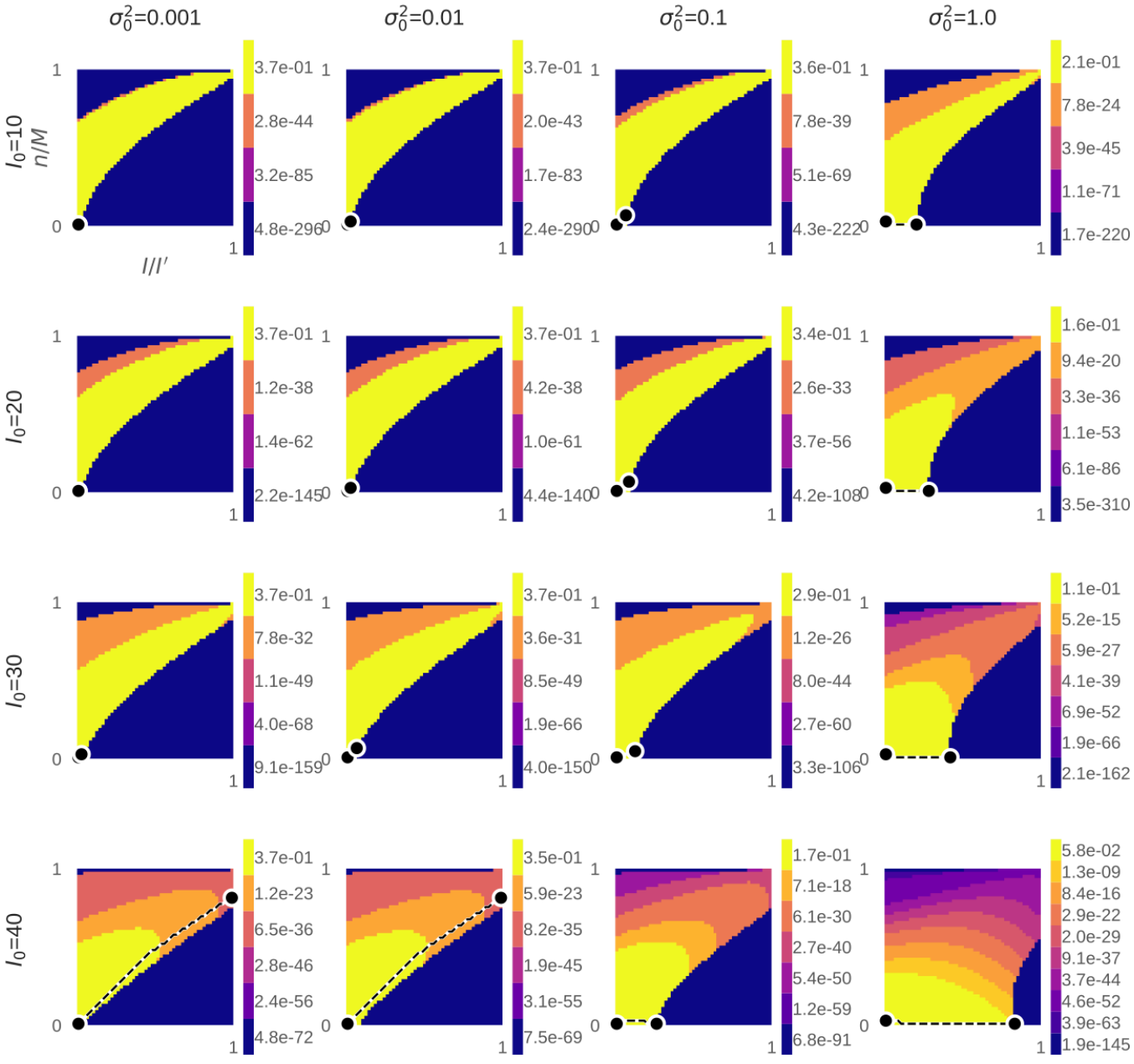

Figura 5.12: **Caminho mais provável entre os estados $c = (n, I)$ reavaliados para diversas condições de α_0^2 e I_0, com $M = 50$.** Os *heatmaps* mostram as probabilidades de estados em sistemas com o valor de $I' = 50$. O caminho mais provável é dado pela linha pontilhada, que parte do estado mais provável $c = (n, I)$ e vai até o estado absorvente c_k.

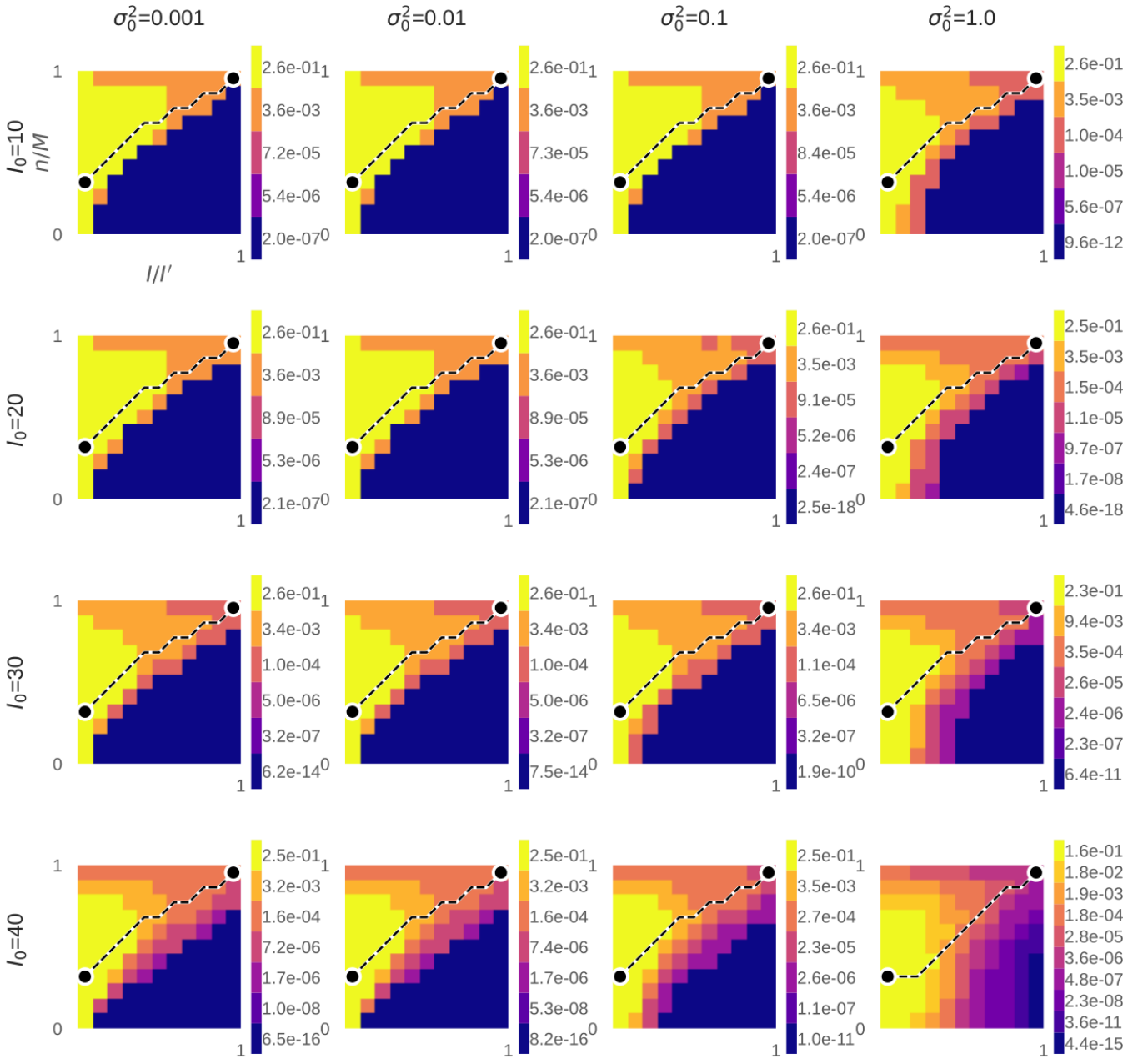

Figura 5.13: **Caminho mais provável entre os estados $c = (n, I)$ reavaliados para diversas condições de α_0^2 e I_0, com $M = 10$ e $p = 0.25$.** Os *heatmaps* mostram as probabilidades de estados em sistemas com o valor de $I' = 50$. O caminho mais provável é dado pela linha pontilhada, que parte do estado mais provável $c = (n, I)$ e vai até o estado absorvente c_k.

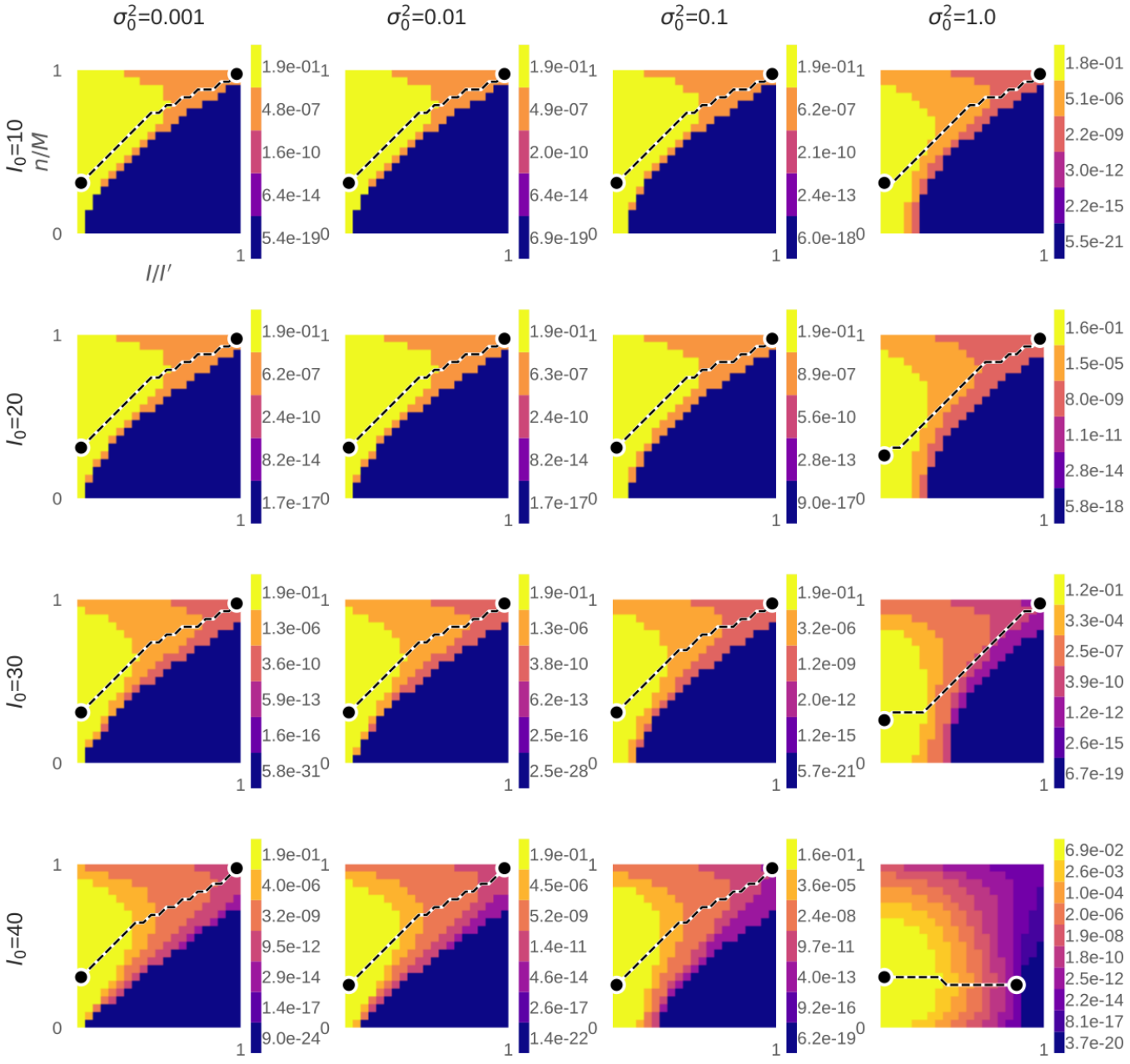

Figura 5.14: **Caminho mais provável entre os estados $c = (n, I)$ reavaliados para diversas condições de α_0^2 e I_0, com $M = 20$ e $p = 0.25$.** Os *heatmaps* mostram as probabilidades de estados em sistemas com o valor de $I' = 50$. O caminho mais provável é dado pela linha pontilhada, que parte do estado mais provável $c = (n, I)$ e vai até o estado absorvente c_k.

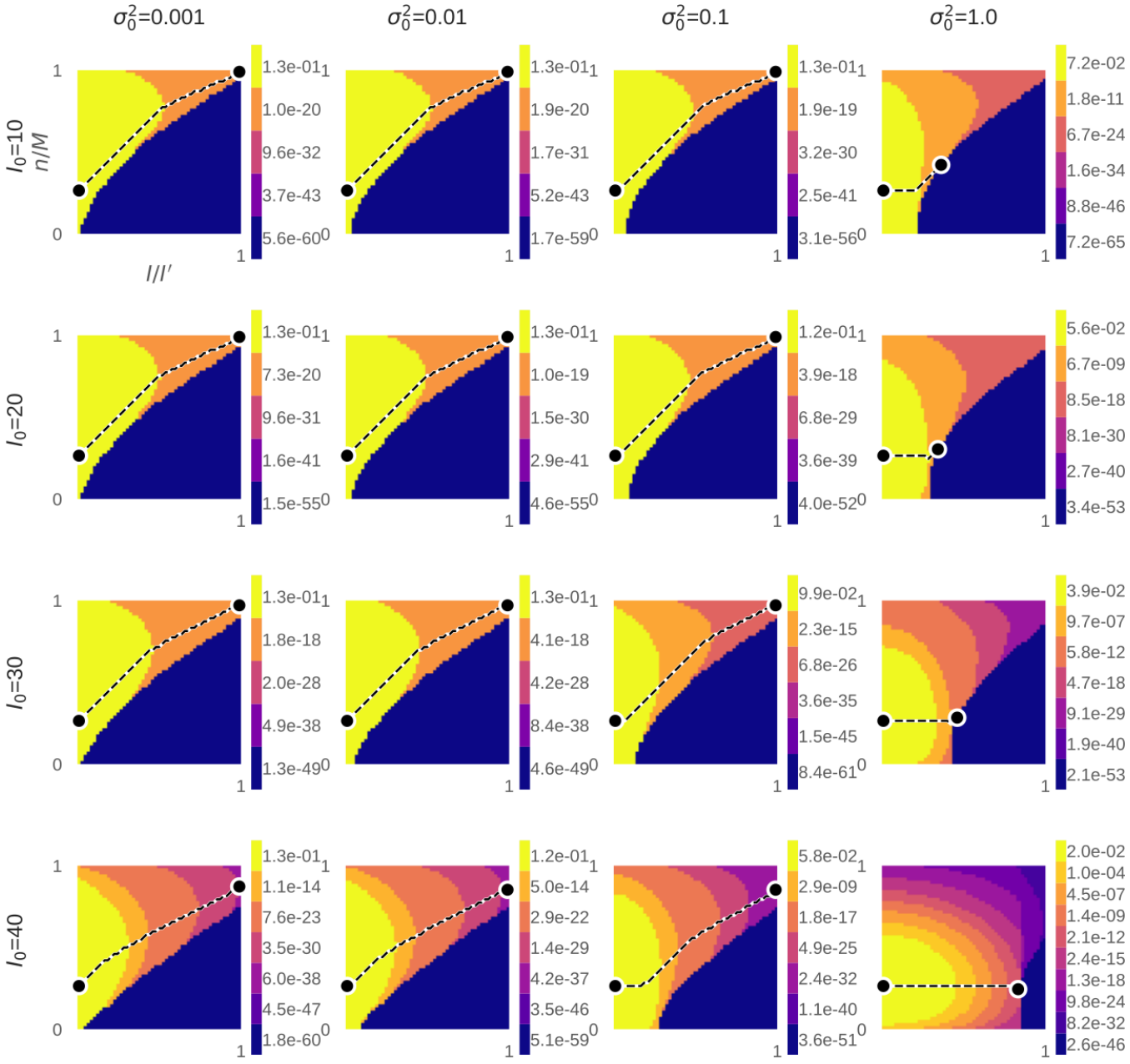

Figura 5.15: **Caminho mais provável entre os estados $c = (n, I)$ reavaliados para diversas condições de α_0^2 e I_0, com $M = 50$ e $p = 0.25$.** Os *heatmaps* mostram as probabilidades de estados em sistemas com o valor de $I' = 50$. O caminho mais provável é dado pela linha pontilhada, que parte do estado mais provável $c = (n, I)$ e vai até o estado absorvente c_k.

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

scientific reports

Check for updates

OPEN

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

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

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

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

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

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

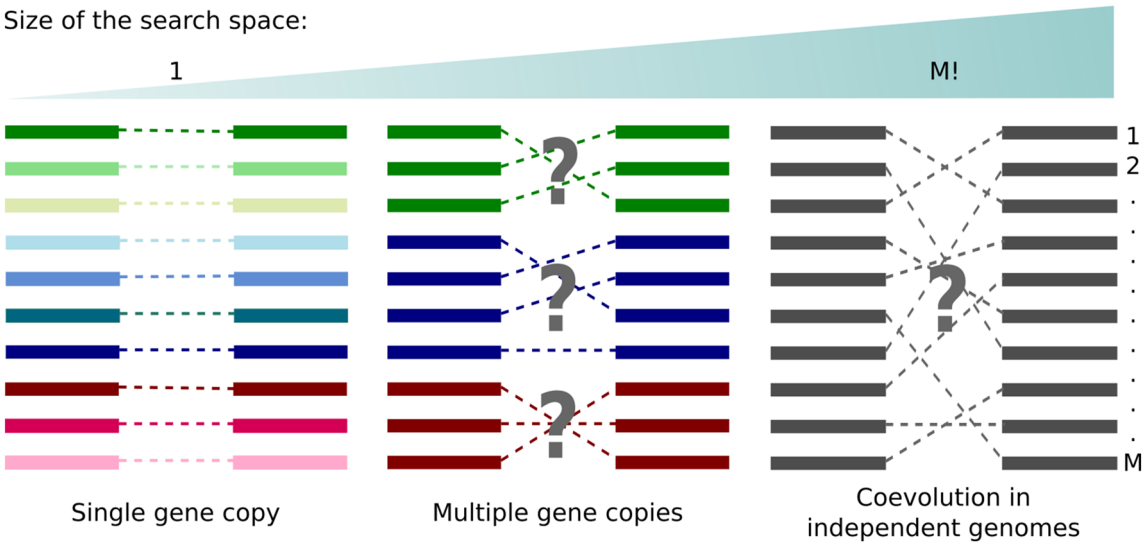

Figure 1. Different scenarios for protein partners determination from multi-sequence alignments. The correct set of partners is known for systems with a single gene copy per genome and unknown for systems involving multiple (paralogous) sequences within the same genome or multiple sequences across independent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

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

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

Results and discussion

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

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

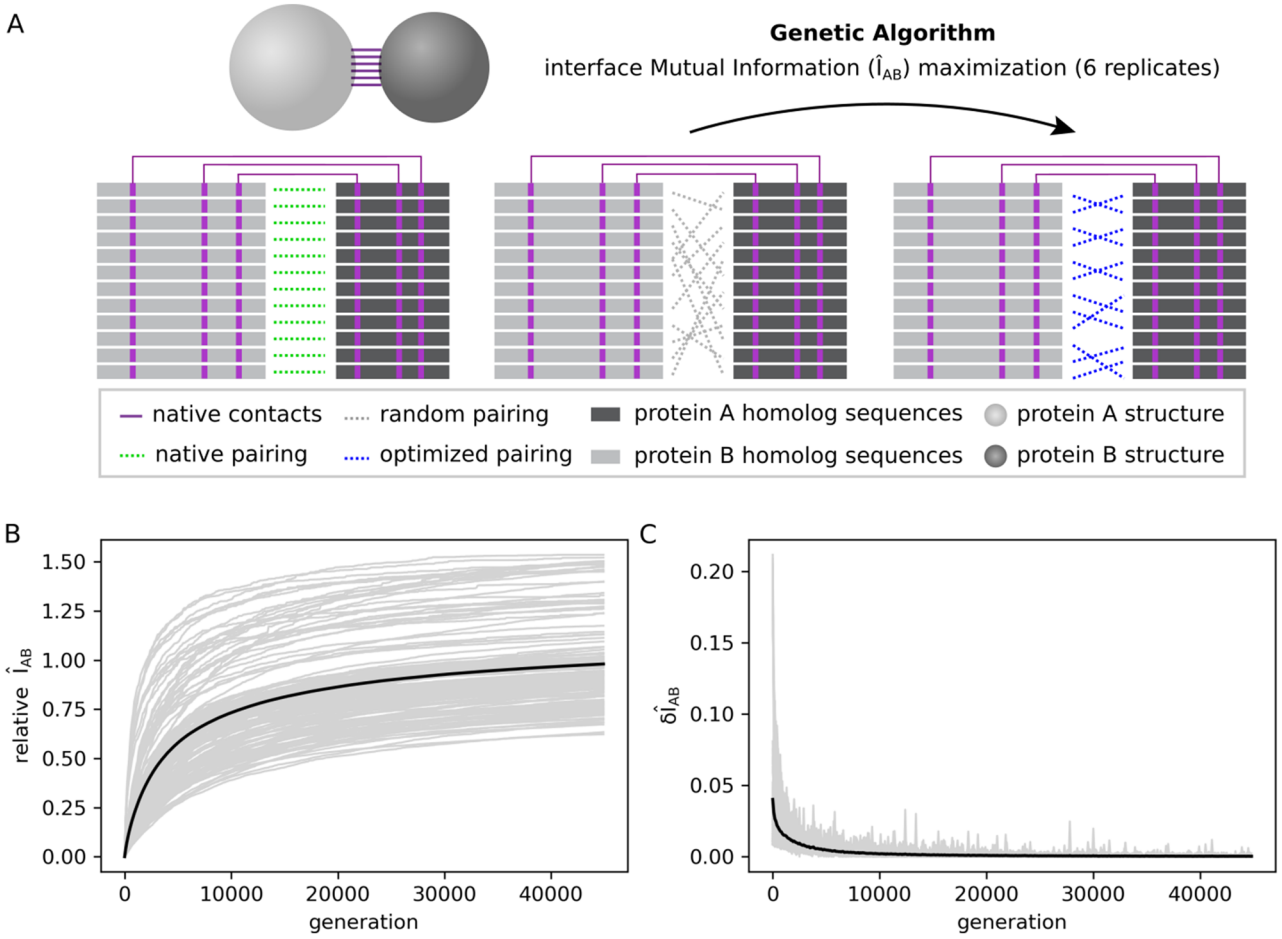

Figure 2. Interface mutual information ($\hat{I}_{AB}$) optimization trajectories. (**A**) Scheme showing $\hat{I}_{AB}$ optimization process starting from a scrambled multi-sequence alignment (MSA) concatenation (in gray) and reaching an optimized concatenation (in blue). Only physically coupled MSA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

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

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

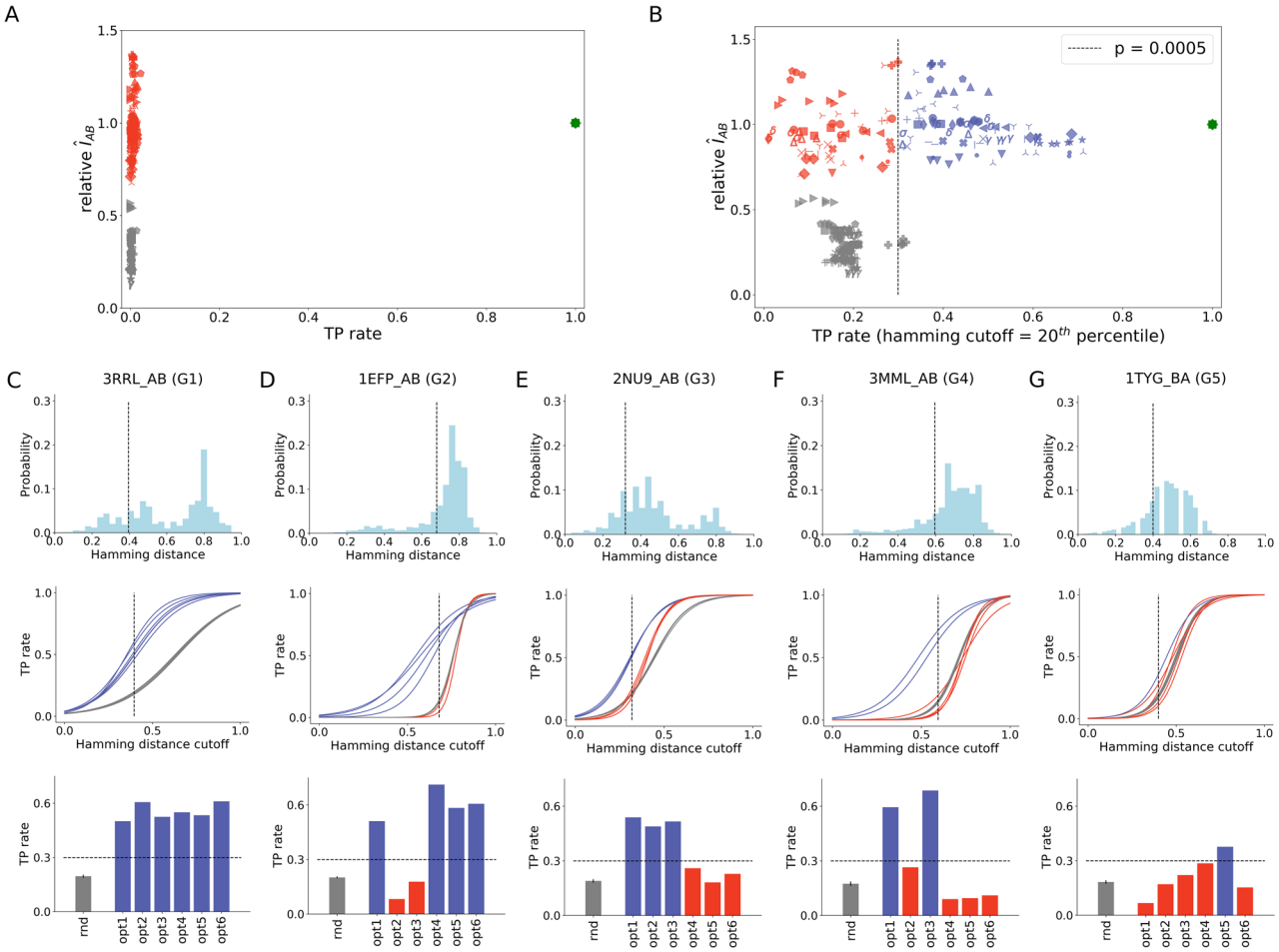

Figure 3. Evaluation of optimized MSA concatenations. (**A**) True positive (TP) rate of random, optimized and native MSA concatenations. (**B**) Reassessed TP rate of random, optimized and native MSA concatenations by discounting wrong pairings among sequences with Hamming distance within the 20th percentile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

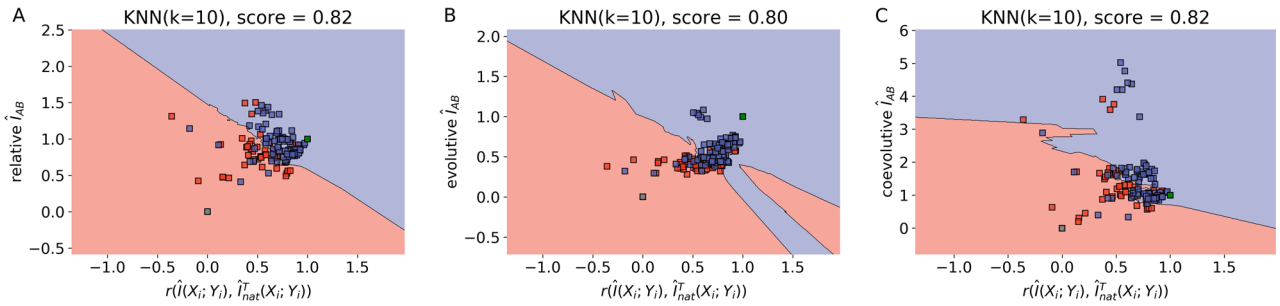

Figure 4. (**A**) Optimized concatenation solutions scattered across the space of relative interface mutual information (MI), $\hat{I}_{AB}$, against Pearson correlation between optimized and native MI vectors, $r(\hat{I}(X_i;\ Y_i),\ \hat{I}^T_{nat}(X_i;\ Y_i))$. Type-(i) solutions are shown in red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

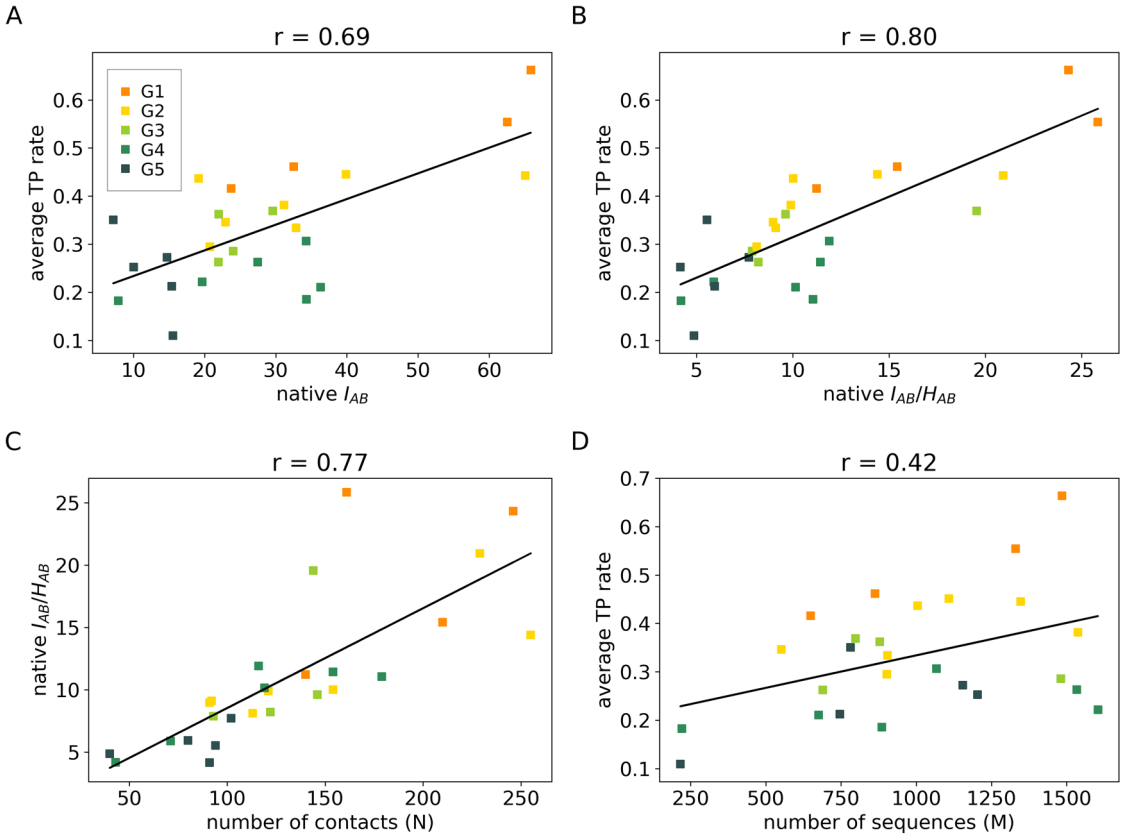

Figure 5. (**A**) Correlation between the true positive (TP) rate of optimized solutions and mutual information (MI) on the interface I_{AB}. (**B**) Correlation between TP rate of optimized solutions and I_{AB} regularized by the joint entropy on the interface, I_{AB}/H_{AB}. (**C**) Correlation between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

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

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

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

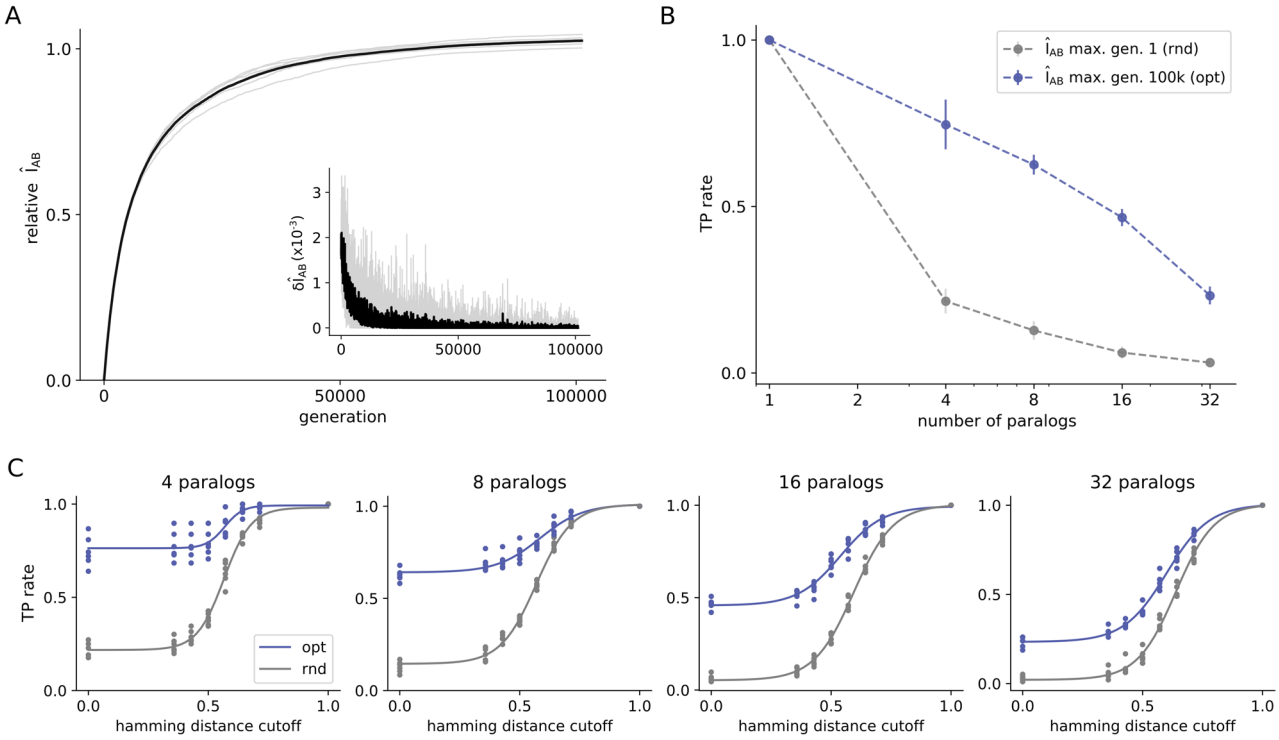

Figure 6. Evaluation of optimized MSA concatenations of the HK-RR paralogs dataset. (**A**) Optimization trajectories for the HK-RR standard dataset. The interface mutual information normalized by the native interface mutual information (relative $\hat{I}_{AB}$) is plotted against the number of generations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

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

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

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

Results in Fig. 6 appear to rationalize the sharp deterioration of TP rates with the number of sequences in recent investigations of paralogous systems[12–14,18,19], by hypothesizing it is due to the lack of type-(i) mismatches and the great degeneracy involved. In previous works, Bitbol and coworkers developed an iterative pairing

algorithm (IPA) capable of inferring protein partners using either direct coupling analysis (DCA-IPA)[13], mutual information (MI-IPA)[18], or phylogeny (Mirrortree-IPA)[19]. When benchmarked for paralog matching on the standard HK-RR dataset, DCA-IPA was as accurate as MI-IPA, and Mirrortree-IPA was even more accurate. The performance of these algorithms, however, drops considerably for species with more than 32 paralogs. The tendency is that the TP rate also drops to zero in a hypothetical genome with hundreds of paralogs[19], a situation analogous to the results in Fig. 6. In conclusion, results presented in Fig. 6 suggest that paralog matching is only possible because there is usually a small number of paralogous sequences per genome. When extended to genomes with more paralogs, this problem tends to present only type-(ii) solutions, leaving virtually no room for improvement of TP rates.

Conclusions and future work

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

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

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

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

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

Methods

Consider two interacting protein families, A and B. It is possible to construct two MSAs, MSA A and MSA B, containing M sequences from families A and B, respectively. A specific coevolution process $z \in \{1, \ldots, M!\}$ associates each sequence l in MSA B to a sequence k in MSA A in a unique arrangement of size M (see Fig. 7). Given that members of A and B interact via formation of N independent amino acid contacts at molecular level, it is possible to extract from these MSAs only the columns corresponding to sites that are in contact, belonging to the complex interface. In this context, the interacting amino acids of families A and B are described by two N -length blocks of discrete stochastic variables, $X^N = (X_1, \ldots, X_N)$ and $Y^N = (Y_1, \ldots, Y_N)$, with associated probability mass functions (PMFs) $\{\rho(x_1 \ldots x_N),\ \rho(y_1 \ldots y_N),\ \rho(x_1 \ldots x_N,\ y_1 \ldots y_N|z)|x_i,\ y_i \in \Omega,\ \ \forall i \in \{1, \ldots, N\}\}$. Here, the alphabet Ω has size 21 and contains all 20 amino acids and the gap symbol '–'. Note that only the joint PMF will depend on process z.

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

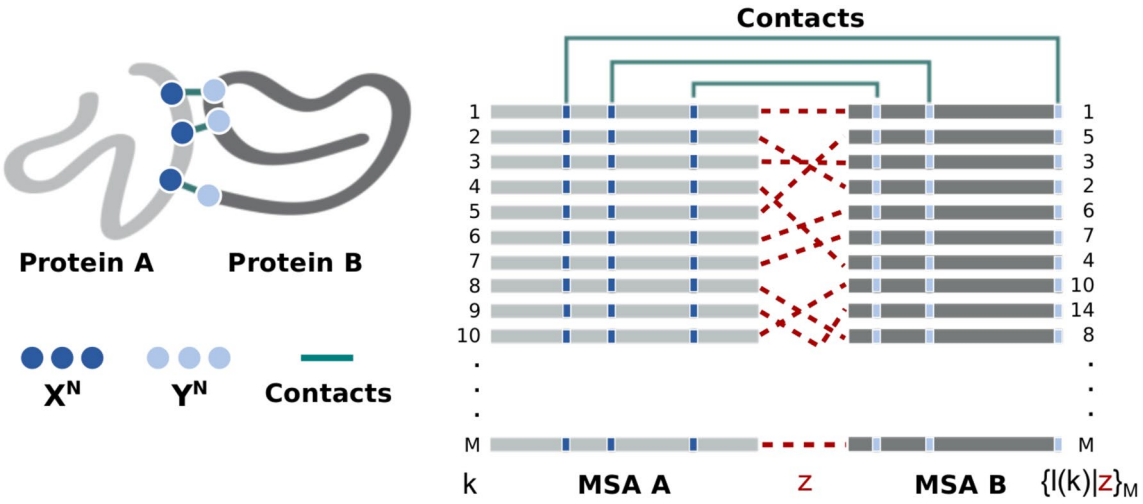

Figure 7. Structural contacts mapped into M-long multi-sequence alignment of protein interologs A and B. A set of pairwise protein–protein interactions is defined by associating each sequence l in MSA B to a sequence k in MSA A in one unique arrangement, $\{l(k)|z\}$, determined by the coevolution process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

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

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

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

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

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

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

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

The joint entropy of the interface was calculated for individual contacts

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

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

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

where N is the number of contacts.

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

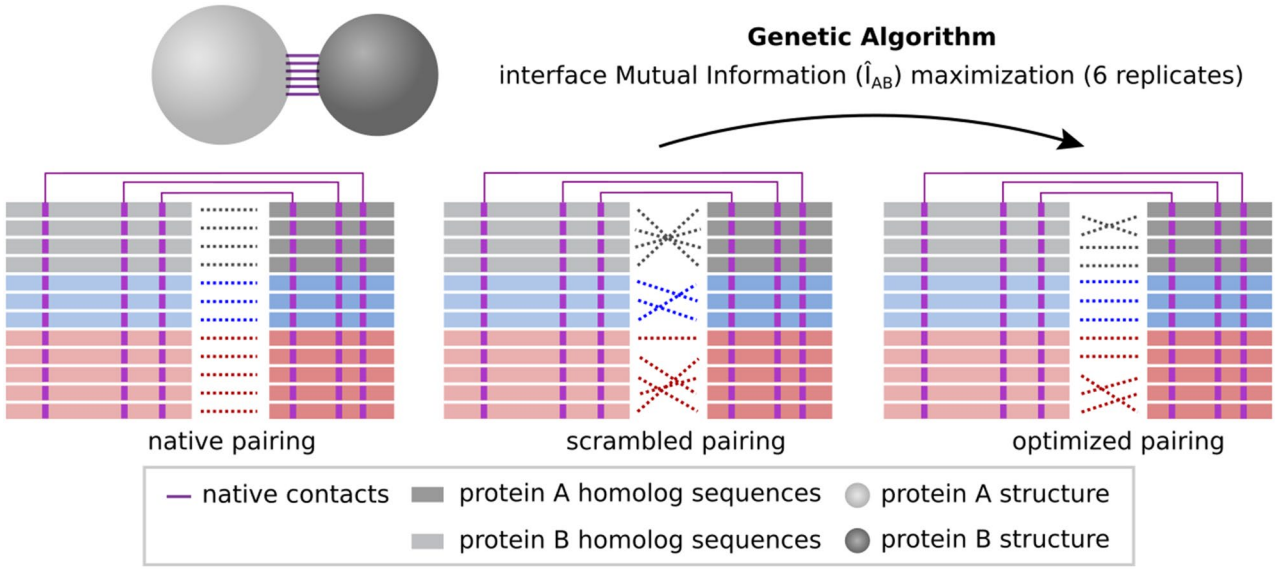

Figure 8. Scheme showing interface mutual information ($\hat{I}_{AB}$) optimization process for the HK-RR standard dataset. It starts from a within-species scrambled MSA concatenation and reaches an optimized concatenation. Different species are shown in different colors. Only physically coupled MSA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

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

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

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

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

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

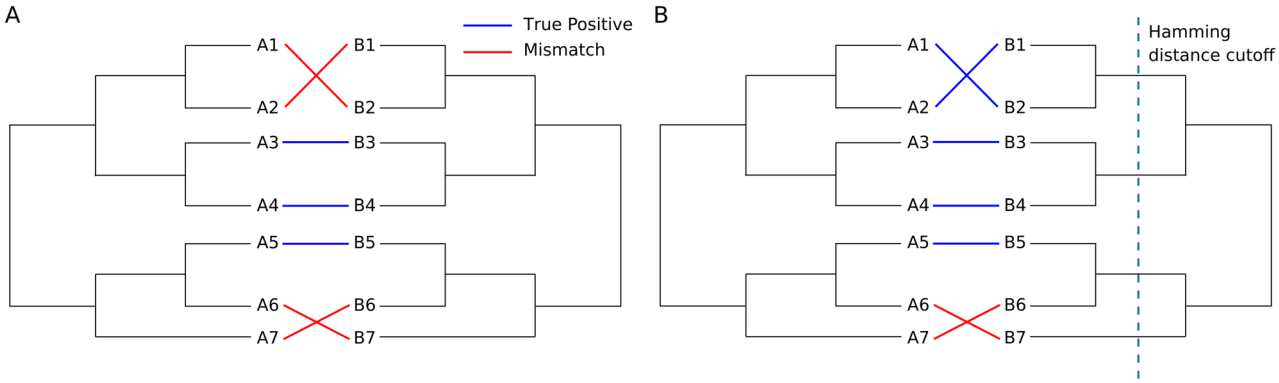

Figure 9. Mismatch discounting based on a Hamming distance cutoff. Scheme showing how the accuracy of the same MSA concatenation would be assessed with (**B**) and without (**A**) mismatch discounting. This figure was created with Inkscape (https://inkscape.org/).

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

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

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

References

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

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

Acknowledgements

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

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

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

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

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

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