



**UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA**

DIEGO CAVALCANTE CIRNE

UM ESTUDO TEÓRICO SOBRE A PREVISIBILIDADE EM EVOLUÇÃO

Recife
2020

DIEGO CAVALCANTE CIRNE

UM ESTUDO TEÓRICO SOBRE A PREVISIBILIDADE EM EVOLUÇÃO

Dissertação apresentada ao Programa de Pós-Graduação em Física da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Física.

Área de Concentração: Física Teórica e Computacional

Orientador: Prof. Paulo Roberto de Araújo Campos

Recife
2020

Catálogo na fonte
Bibliotecária Mariana de Souza Alves CRB4-2105

C578e Cirne, Diego Cavalcante
Um estudo teórico sobre a previsibilidade em evolução / Diego Cavalcante Cirne.
– 2020.
88f.: il., fig.

Orientador: Paulo Roberto de Araújo Campos.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Física,
Recife, 2020.
Inclui referências.

1. Física Teórica e Computacional. 2. Previsibilidade. 3. Espaço de sequências.
4. Relevo de adaptação. I. Campos, Paulo Roberto de Araújo. (orientador) II. Título.

530.1

CDD (22. ed.)

UFPE-CCEN 2020-196

DIEGO CAVALCANTE CIRNE

UM ESTUDO TEÓRICO SOBRE A PREVISIBILIDADE EM EVOLUÇÃO

Dissertação apresentada ao Programa de Pós-Graduação em Física da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Física.

Aprovada em: 14/10/2020.

BANCA EXAMINADORA

Prof. Paulo Roberto de Araújo Campos
Orientador
Universidade Federal de Pernambuco

Prof. Ernesto Carneiro Pessoa Raposo
Examinador Interno
Universidade Federal de Pernambuco

Prof. José Fernando Fontanari
Examinador Externo
Universidade de São Paulo

A meus pais.

AGRADECIMENTOS

A minha família, pelo suporte e amor incondicional. A Paulo Campos, pelo acolhimento, orientação e paciência. A Fernando Parísio, Leonardo Cabral e Paulo Campos, pelos exemplos de pesquisador e docente nas disciplinas ministradas. A Jorge Gabriel, pela paixão contagiante pela ciência. Ao Departamento de Física da UFPB, pela minha formação básica em física. Ao Departamento de Física da UFPE, pelo ambiente acadêmico de excelência. Aos amigos, pela oportunidade de compartilhar o prazer e fascínio pela vida.

RESUMO

Neste estudo abordamos a importante questão da previsibilidade do processo evolutivo no contexto do espaço de sequências e relevo de adaptação, introduzidos originalmente por Sewall Wright. A topologia do espaço de sequências é tal que dois genótipos que diferem por uma única mutação são vizinhos. O relevo de adaptação surge através da atribuição do fitness a cada sequência, uma medida ou representante da aptidão do indivíduo. Nesse relevo multidimensional caminhadas adaptativas seguem um algoritmo de busca local por meio de reprodução, mutação e seleção em direção aos máximos locais, genótipos mais aptos que todos os seus vizinhos. Simulando diferentes dinâmicas populacionais num espaço de sequências binárias rastreamos a cada geração o indivíduo mais apto presente e ao fim tem-se armazenada uma caminhada adaptativa. O processo é claramente sujeito a fatores estocásticos e sua repetição sistemática produz um ensemble de trajetórias evolutivas, a informação fundamental a ser explorada. A partir destes ensembles sobre diferentes parâmetros realizamos uma série de medidas majoritariamente associadas à acessibilidade e reprodutibilidade de caminhadas evolutivas. Duas medidas são de especial interesse: *predictability* e divergência média de caminhos. Enquanto a primeira avalia a distribuição de probabilidade das caminhadas, representando um passo além em relação a medidas simplórias como o número de caminhos distintos, a segunda caracteriza tanto a distribuição quanto uma noção topológica de similaridade entre caminhos, baseada na distância de Hamming entre as sequências de cada trajeto. Entre os resultados relevantes destacamos como o fenômeno epistático de interação gênica está associado a relevos acidentados, onde o número de trajetórias é bastante reduzido ao mesmo tempo em que diminui sua similaridade, dois efeitos essencialmente opostos em termos de previsibilidade. No regime em que a evolução não pode ser simplificada como caminhadas adaptativas, devendo-se levar em consideração a demografia, observamos que tanto a previsibilidade quanto a divergência média de trajetórias apresentam uma dependência monotônica com o tamanho de população N , porém não monotônica com a taxa de mutação μ .

Palavras-chaves: Previsibilidade. Espaço de sequências. Relevo de adaptação. Caminhadas adaptativas. Epistasia.

ABSTRACT

In this study, we approach the important question of the predictability in the evolutionary process in the context of sequence space and fitness landscape, originally introduced by Sewall Wright. The topology of the sequence space is such that two genotypes that differ by a single mutation are linked. The fitness landscape comes through the attribution of fitness values to each sequence, a measure, or a proxy of the individual's adaptation. In this multidimensional landscape, the evolutionary dynamics follow a local search algorithm through reproduction, mutation, and selection towards local maximums, genotypes that are fitter than all their neighbors. Simulating different versions and limits of population dynamics in this binary space of sequences, we keep track of the fittest individual present at each generation, and in the end, the evolutionary pathway is stored. The process is clearly subject to stochastic factors and its systematic repetition produces an ensemble of evolutionary trajectories, the fundamental information to be explored. From these ensembles for different sets of parameter values, we perform a series of measures mostly associated with the accessibility and reproducibility of evolutionary walks. Two measures are of special interest: predictability and mean path divergence. While the first assesses the probability distribution of the walks, representing a step further with respect to simplistic measures such as the number of distinct paths, the second characterizes both the distribution and a topological notion of path similarity based on the Hamming distance between the sequences of each path. Among our findings, we highlight how the epistatic phenomenon of gene interaction is associated with rugged landscapes, where the number of accessible trajectories is greatly reduced while reducing their similarity, two essentially opposite effects in terms of predictability. In the regime in which the evolution can not be simplified as adaptive walks, but instead one has to take into account the demography, we observe that both the predictability and mean path divergence display a monotonic dependence on the population size N , but non-monotonic on the mutation rate μ .

Keywords: Predictability. Sequence space. Adaptive landscape. Adaptive walks. Epistasis.

LISTA DE FIGURAS

- Figura 1 – Espaço de sequências binárias de comprimento $L = 3$. Para um alfabeto binário e comprimento arbitrário, o espaço de sequências constitui um hiper-cubo. 20
- Figura 2 – Representação bidimensional de relevo de adaptação fictício. Sinais de + e - representam máximos e mínimos de fitness, respectivamente. 22
- Figura 3 – Dois relevos experimentais são representados por sequências binárias onde 0 e 1 representam, respectivamente, ausência e ocorrência de uma determinada mutação. Setas coloridas apontam para o vizinho de maior fitness e um conjunto de setas da mesma cor indica uma bacia de atração para um máximo local. À esquerda, o relevo de mutações benéficas para a bactéria *Methylobacterium extorquens* e, à direita, o relevo de mutações associadas à resistência a medicamentos contra malária. 23
- Figura 4 – Populações sofrem diferenciação a partir de um ancestral desconhecido sujeitas a determinadas condições ambientais (linhas). C e D apresentam repetibilidade, embora com padrões diferentes de evolução, paralela e convergente, respectivamente. A e B mostram divergência, onde diferenças iniciais são amplificadas por condições ambientais diferentes em A, e por condições semelhantes em B. Em C tais diferenças são mantidas ao longo do tempo. 28
- Figura 5 – Relevo adaptativo de dois genes com possibilidade de apenas uma única mutação em cada. O tamanho das circunferências representa o fitness do genótipo e mudanças adaptativas favoráveis são indicadas pelas setas. À esquerda, um relevo suave, onde qualquer genótipo inicial inevitavelmente leva ao único máximo “AB”. À direita, um relevo rugoso, com dois máximos “AB” e “ab”. Populações centradas em “Ab” ou “aB” podem atingir ambos os picos, dependendo da ordem em que sofrem as mutações benéficas. Uma população iniciada em “ab” não é capaz de alcançar o máximo global. . . . 30

- Figura 6 – Diagrama adaptativo das 10 trajetórias mais prováveis entre o genótipo original TEM^{wt} e o mutante resistente TEM^* . Circunferências representam o genótipo, cuja identidade é descrita por uma sequência de símbolos + e – que correspondem respectivamente à presença ou ausência das mutações g4205a, A42G, E104K, M182T e G238S, nesta ordem. Números representam a resistência ao antibiótico cefotaxima em mg/ml. Setas representam as mutações nelas rotuladas. A probabilidade de ocorrência destas mutações é exibida em escala logarítmica tanto em cor quanto largura da seguinte forma: verde/largo, 0,316 a 1,0; azul/médio, 0,1 a 0,316; roxo/estrito, 0,0316 a 0,1; vermelho/muito estrito, menor que 0,0316. Onde há setas duplas, contínuas e segmentadas representam as probabilidades sob os modelos de probabilidade de fixação correlacionada e descorrelacionada com o fitness, respectivamente. 34
- Figura 7 – *Predictability* em função da probabilidade de mutações neutras relativamente à probabilidade de mutações benéficas. Linhas preta e cinza são obtidas excluindo e admitindo a possibilidade de mutações não unitárias, respectivamente. Quando as mutações neutras são muito menos prováveis que mutações benéficas (1%), $\mathcal{P}_2$ sofre pouca alteração (5%). Quando as probabilidades são aproximadamente iguais, no entanto, $\mathcal{P}_2$ diminui significativamente, em torno de quatro vezes. 35
- Figura 8 – Ilustração do cálculo da divergência $d(\phi_1, \phi_2)$ entre um par de caminhos ϕ_1 e ϕ_2 . Para cada ponto $\sigma_1 \in \phi_1$ procura-se o ponto mais próximo pertencente a ϕ_2 e determina-se a distância $h(\sigma_1, \phi_2)$. O mesmo é feito sobre o caminho ϕ_2 e a divergência $d(\phi_1, \phi_2)$ é calculada como a média de $h(\sigma_1, \phi_2)$ e $h(\sigma_2, \phi_1)$ conjuntamente. 36
- Figura 9 – À esquerda, sob uma baixa taxa de suprimento de mutantes por geração ($N\mu \ll 1$) a população se comporta de maneira homogênea em praticamente todos os instantes, e mutações são fixadas sequencialmente como passos. À direita, com o aumento do número de mutações por geração diversos genótipos passam a coexistir, cada um seguindo sua própria trajetória, de forma que a evolução da população é melhor representada por um conjunto de caminhos que se espalha a partir do genótipo inicial e volta a convergir no máximo. A cada geração a estrutura da população pode ser descrita por uma distribuição de probabilidade no espaço de sequências. . . 38

Figura 10 – No mundo natural toda a informação fenotípica é armazenada no DNA, escrita em sequências de um alfabeto de quatro bases genéticas $\{A, C, T, G\}$. Em experimentos microbiais, no entanto, é comum o uso de uma base binária $\{0, 1\}$ na análise do processo de adaptação, dado que na maioria das vezes a unidade de interesse são mutações específicas em cada gene, de um alelo para outro, situação melhor representada pela ausência (0) ou presença (1) da mutação. No alfabeto natural os genótipos tem três vizinhos para cada locus, isto é, acessíveis por uma única mutação (destacadas em vermelho), enquanto sob o alfabeto binário só há um.	40
Figura 11 – Para sequências binárias o espaço de sequências de dimensão L é construído através da duplicação do espaço de dimensão $L - 1$, como ilustrado da esquerda para direita em ordem crescente de dimensionalidade.	41
Figura 12 – Função de auto-correlação no Modelo NK para $L = 32$. A função $\rho(h)$ no gráfico difere sutilmente da Equação 3.10 e é dada na forma $(L - K)!(L - h)!/[L!(L - K - h)]$, onde aparentemente o autor da publicação original desconsidera que uma mutação no próprio gene i também altera o valor de f_i . O comportamento de ambas, no entanto, é basicamente o mesmo, onde valores baixos de K correspondem a decaimentos lentos de $\rho(h)$, enquanto valores altos apresentam decaimentos acentuados, culminando no caso extremo $K = L - 1$, onde $\rho(h)$ vai a zero já para $h = 1$ (não ilustrado).	48
Figura 13 – Quantidade de máximos locais em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.	51
Figura 14 – Acessibilidade do máximo global em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.	52
Figura 15 – Fração de máximos locais acessíveis em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.	53
Figura 16 – Quantidade de caminhos monotonicamente crescentes em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda. . . .	54
Figura 17 – <i>Predictability</i> de pontos finais em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.	56
Figura 18 – <i>Predictability</i> de caminhos em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.	57

Figura 19 – Comprimento médio de caminhos em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.	58
Figura 20 – Acessibilidade por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.	59
Figura 21 – Quantidade de caminhos monotônicos por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.	61
Figura 22 – <i>Predictability</i> por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.	62
Figura 23 – Divergência média de caminhos por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.	63
Figura 24 – Comprimento médio de caminhadas por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.	64
Figura 25 – Quantidade de caminhos em função do tamanho N de população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	65
Figura 26 – <i>Predictability</i> em função do tamanho N de população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	66
Figura 27 – Divergência média de caminhos em função do tamanho N de população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	67
Figura 28 – Comprimento médio de caminhos em função do tamanho N da população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	68

Figura 29 – Quantidade de caminhos em função da taxa μ de mutação para diferentes valores de tamanho N de população e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	69
Figura 30 – <i>Predictability</i> em função da taxa μ de mutação para diferentes valores de tamanho N de população e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	70
Figura 31 – Divergência média de caminhos em função do tamanho N de população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	71
Figura 32 – <i>Predictability</i> em função do tamanho N de população para diferentes valores do número esperado $N\mu$ de mutantes e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	72
Figura 33 – Divergência média de caminhos em função do tamanho N de população para diferentes valores do número esperado $N\mu$ de mutantes e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	73
Figura 34 – <i>Predictability</i> em função da taxa μ de mutação para diferentes valores do número esperado $N\mu$ de mutantes e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	73
Figura 35 – Divergência média de caminhos em função da taxa μ de mutação para diferentes valores do número esperado $N\mu$ de mutantes e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$	74
Figura 36 – <i>Predictability</i> sob diferentes dinâmicas adaptativas. As duas primeiras barras à esquerda de cada gráfico mostram os valores referentes às caminhadas adaptativas aleatória e probabilística, nesta ordem. O restante das barras está associado às dinâmicas de população finita para diversos valores de N e μ indicados no eixo horizontal. Preto, azul e vermelho representam K igual a 1, 2 e 3, respectivamente. Cada barra representa uma média sobre 10 relevos de tamanho $L = 8$. Todas as dinâmicas partem do antípoda.	75

Figura 37 – Divergência média de caminhos sob diferentes dinâmicas adaptativas. As duas primeiras barras à esquerda de cada gráfico mostram os valores referentes às caminhadas adaptativas aleatória e probabilística, nesta ordem. O restante das barras está associado às dinâmicas de população finita para diversos valores de N e μ indicados no eixo horizontal. Preto, azul e vermelho representam K igual a 1, 2 e 3, respectivamente. Cada barra representa uma média sobre 10 relevos de tamanho $L = 8$. Todas as dinâmicas partem do antípoda.	76
Figura 38 – Acessibilidade por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 1.000.000 caminhadas no relevo experimental do fungo <i>Aspergillus niger</i>	77
Figura 39 – <i>Predictability</i> por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 1.000.000 caminhadas no relevo experimental do fungo <i>Aspergillus niger</i>	78
Figura 40 – Divergência média de caminhos por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 1.000.000 caminhadas no relevo experimental do fungo <i>Aspergillus niger</i>	79
Figura 41 – <i>Predictability</i> em função do tamanho de população N para diferentes valores da taxa de mutação μ . Foram simuladas 1.000 caminhadas partindo do antípoda no relevo experimental do fungo <i>Aspergillus niger</i>	80
Figura 42 – Divergência média de caminhos em função do tamanho de população N para diferentes valores da taxa de mutação μ . Foram simuladas 1.000 caminhadas partindo do antípoda no relevo experimental do fungo <i>Aspergillus niger</i>	81
Figura 43 – <i>Predictability</i> em função da taxa de mutação μ para diferentes valores de tamanho de população N . Foram simuladas 1.000 caminhadas partindo do antípoda no relevo experimental do fungo <i>Aspergillus niger</i>	82
Figura 44 – Divergência média de caminhos em função da taxa de mutação μ para diferentes valores de tamanho de população N . Foram simuladas 1.000 caminhadas partindo do antípoda no relevo experimental do fungo <i>Aspergillus niger</i>	83

SUMÁRIO

1	INTRODUÇÃO	15
2	FUNDAMENTAÇÃO GERAL	18
2.1	EVOLUÇÃO	18
2.2	ESPAÇO DE SEQUÊNCIAS	19
2.3	RELEVO DE ADAPTAÇÃO	21
2.4	RELEVOS EXPERIMENTAIS	22
2.5	RELEVOS TEÓRICOS	24
2.6	QUASI-ESPÉCIE	25
2.7	PREVISIBILIDADE EVOLUTIVA	27
3	FUNDAMENTAÇÃO ESPECÍFICA	32
3.1	MEDIDAS DE PREVISIBILIDADE	32
3.2	DINÂMICA ADAPTATIVA	37
3.2.1	Caminhante adaptativo	41
3.2.2	População finita	43
3.3	MODELO NK	45
4	RESULTADOS	49
4.1	CAMINHADA ADAPTATIVA E RUGOSIDADE	49
4.1.1	Valores médios em função de L e K	49
4.1.2	Medidas por máximo local	58
4.2	DINÂMICA POPULACIONAL	64
4.2.1	Tamanho da população	65
4.2.2	Taxa de mutação	69
4.2.3	Fluxo de mutantes fixo	72
4.2.4	Caminhadas adaptativas VS dinâmicas populacionais	74
4.3	RELEVO EXPERIMENTAL: <i>ASPERGILLUS NIGER</i>	77
4.3.1	Caminhadas adaptativas	77
4.3.2	Dinâmica populacional	79
5	CONCLUSÕES	84
	REFERÊNCIAS	86

1 INTRODUÇÃO

A Teoria da Evolução e Seleção Natural é uma das grandes joias da ciência moderna. Na verdade, sua formulação e consolidação se confunde com a do próprio método científico. Não obstante, a evolução das espécies está longe de configurar uma teoria completa e fechada (MORRIS, 2009). Devido à grande gama de variáveis envolvidas, sua estrutura tem sido continuamente reformulada através da elucidação de diversos dos mecanismos inerentes.

Entre as questões mais importantes nesse contexto destacamos a da previsibilidade do processo evolutivo, considerando o conflito entre o fator determinístico, promovido pela seleção, e o fator estocástico, intrínseco da mutação (SZENDRO et al., 2013a). O evolucionista Stephen Jay Gould sintetizou o problema numa simples pergunta: se o filme da vida pudesse ser reproduzido, teríamos ao final seres e estruturas semelhantes aos atuais, ou sistemas completamente diferentes (GOULD, 1990)?

Se por um lado a ideia é de uma clareza conceitual enorme, do ponto de vista experimental é uma tarefa inviável. Por esse motivo o estudo em previsibilidade evolutiva foi por muito tempo fortemente limitado ao trabalho teórico (SZENDRO et al., 2013b). Entretanto, os avanços tecnológicos em genética do final do século XX e início do século XXI permitiram finalmente uma abordagem experimental do problema, pelo menos em escala microevolutiva, principalmente devido à capacidade de identificação rápida e precisa de genomas e da aplicação da mutagênese (VISSER; KRUG, 2014).

Paralelamente, o desenvolvimento teórico ocorreu por meio da abstração do conjunto de mutantes possíveis nos conceitos de espaço de sequências e relevo de adaptação, cujo arquétipo inicial foi proposto por Sewall Wright (WRIGHT, 1932), com contribuições fundamentais de Maynard Smith e Manfred Eigen posteriormente. O espaço de sequências introduz a ideia de vizinhança, associada a genótipos que diferem por uma única mutação (NOWAK, 2006). O relevo de adaptação, por sua vez, é a simples atribuição de um parâmetro de aptidão no espaço de sequências, conhecido como fitness (NOWAK, 1992). Conjuntamente, as ideias de vizinhança e aptidão formam o terreno base para o estudo de caminhadas adaptativas e trajetórias evolutivas.

Dentro desse cenário o presente trabalho aborda a questão de maneira simulacional. Seguindo a ideia de Gould (GOULD, 1990), simulamos repetidamente o processo evolutivo sobre relevos de adaptação, gerando ensembles de caminhadas independentes, fruto da replicação de populações sujeitas às mesmas condições iniciais e parâmetros. Utilizando-se de uma forma particular de identificação e armazenamento das caminhadas, os ensembles são sujeitos a medidas específicas de previsibilidade, sendo um dos nossos principais objetivos a caracterização da repetibilidade de trajetórias e de sua relação com as quantidades envolvidas, com ênfase para os papéis da epistasia (interação gênica), do tamanho da população, e da taxa de mutação.

Dois tipos de relevos são utilizados, um teórico e outro experimental. O primeiro é obtido do Modelo NK, um modelo de campo aleatório no qual é possível controlar o grau de epistasia e,

consequentemente, aspectos topográficos (KAUFFMAN; WEINBERGER, 1989). O segundo relevo é fruto de um estudo experimental sistemático sobre combinações de 8 mutações do fungo *Aspergillus niger* (VISSER; HOEKSTRA; ENDE, 1997). A relação da previsibilidade evolutiva com o grau de epistasia, assim como a comparação entre relevos experimental e teórico, configuram o restante dos objetivos primários deste trabalho.

O corpo do texto é dividido em três capítulos: Fundamentação Geral (Capítulo 2), Fundamentação Específica (Capítulo 3) e Resultados (Capítulo 4). Na Fundamentação Geral começamos abordando conceitos gerais de evolução (Seção 2.1) assim como o problema teórico que dá origem à formulação do espaço de sequências e do relevo de adaptação. O espaço de sequências é explorado na Seção 2.2, onde desenvolve-se o caráter topológico desse espaço através da ideia de vizinhança entre sequências, intimamente relacionada à noção de caminhada adaptativa. O relevo de adaptação é abordado logo a seguir (Seção 2.3), onde define-se o fitness como representante da aptidão de um indivíduo e evidenciamos a natureza topográfica fruto da atribuição desta quantidade ao espaço de sequências. Nas duas Seções seguintes (2.4 e 2.5) distinguimos entre os relevos experimentais e modelos teóricos de relevo. Sobre os relevos experimentais descrevemos o avanço histórico das técnicas e limitações, assim como as perspectivas experimentais atuais. Os relevos teóricos, por sua vez, também são abordados de uma maneira histórica, e alguns dos modelos mais conhecidos são apresentados. Na Seção 2.6 abordamos as quasi-espécies. Nela apresentamos as equações de quasi-espécie, muito embora não se apliquem a populações finitas, objeto de estudo deste trabalho. Temos bastante interesse, no entanto, em aspectos qualitativos da quasi-espécie compartilhados por populações finitas. Por último, apresentamos na Seção 2.7 o problema geral da previsibilidade em evolução.

A Fundamentação Específica (Capítulo 3) introduz informações particulares deste trabalho, isto é, medidas, dinâmicas e relevos utilizados. A Seção 3.1 traz as duas principais medidas de caracterização da reprodutibilidade de trajetórias, a *predictability* e a divergência média de caminhos, onde são apresentadas suas definições operacionais e contexto de aplicação. Na Seção 3.2 define-se precisamente os algoritmos de reprodução, mutação e seleção dos dois tipos de dinâmicas adaptativas utilizadas, assim como as motivações teóricas e experimentais. Esta seção é dividida entre Caminhante Adaptativo (Subseção 3.2.1) e População Finita (Subseção 3.2.2), os dois tipos de dinâmica aplicados aqui. A caminhada adaptativa está associada, na verdade, a um limite de baixas mutações da população finita e, na prática, se resume à caminhada de um único indivíduo, representante de toda a população. Encerrando o capítulo, a Seção 3.3 traz definição e propriedades do Modelo NK. Nele a intensidade das interações epistáticas são reguláveis e, indiretamente, também torna-se ajustável a rugosidade do relevo. A origem dessa relação entre epistasia e rugosidade é devidamente esclarecida e está diretamente associada ao fenômeno da frustração.

Nos Resultados (Capítulo 4) temos três seções. A primeira delas (Seção 4.1) trata diretamente da relação entre previsibilidade e aspectos topográficos, utilizando somente o Modelo

NK. No primeiro tópico nos dedicamos a explorar essa relação de maneira geral, através de médias sobre relevos, e em função dos parâmetros do modelo. Já no tópico seguinte 4.1.2 lidamos com relevos individuais, onde pôde-se analisar a relação da previsibilidade por máximo local, correlacionando-a com propriedades destes picos como distância ao ponto de partida ou fitness. Seguindo para a Seção 4.2 temos a análise sistemática da dependência entre medidas de previsibilidade e os parâmetros populacionais: tamanho da população, taxa de mutação e número esperado de mutantes. É feita ainda a comparação entre caminhantes adaptativos e população finita. A terceira e última Seção (4.3) traz a mesma abordagem dos resultados anteriores desta vez sob o relevo experimental do fungo *Aspergillus niger*.

2 FUNDAMENTAÇÃO GERAL

A teoria da evolução é, sem dúvidas, um divisor de águas na história das ciências. A hipótese da natureza estocástica do fenômeno foi fundamental na consolidação dos sistemas probabilísticos como uma disciplina essencial, numa época dominada pela corrente de pensamento determinista. O caráter dinâmico das espécies evidenciou, por sua vez, a relevância e generalidade dos fenômenos fora do equilíbrio termodinâmico. O aspecto coletivo abriu as portas para uma visão científica recém nascida, o emergentismo, ao mesmo tempo em que reforçou a importância da interdisciplinaridade na compreensão do mundo natural. A Teoria da Evolução e Seleção Natural é, portanto, um marco profundo no desenvolvimento do próprio método científico.

2.1 EVOLUÇÃO

O processo evolutivo é baseado em três pilares: reprodução, mutação e seleção. Todos os organismos vivos foram e estão sendo moldados continuamente por esses fatores, nas mais diversas escalas de organização, espaço e tempo. O paradigma básico é de que em um ambiente favorável, um ser vivo deve se reproduzir. Quando dois tipos de indivíduos competem por recursos e se reproduzem a taxas diferentes, ocorre seleção, que favorece aquele com a maior taxa (NOWAK, 2006). Ocorre que, a nível de DNA e RNA reprodução significa replicação, isto é, cópia de informação, um processo intrinsecamente propenso a erros (DOMINGO; SCHUSTER, 2015). Estes erros ocasionais na transmissão do material genético produzem novos tipos de indivíduos e, com eles, diversidade, que estará sujeita à seleção perpetuando o ciclo. É importante ressaltar que a capacidade de evoluir não é de genes, células ou organismos, mas de populações. Só populações conseguem manter e transmitir a informação gerada no processo (NOWAK, 2006).

Existem duas hipóteses básicas sustentando essa dinâmica. Primeiro, as mutações são aleatórias no que diz respeito à sua efetividade ou funcionalidade (SMITH, 1970). Deve ficar claro que nada se afirma, no entanto, sobre aleatoriedade na forma como são geradas, em termos de suas origens químicas e físicas. Nesse aspecto, na verdade, o processo pode ser muito tendencioso em geral, traduzindo-se por exemplo, na existência de certas regiões nas sequências genéticas fortemente acometidas por mutações (SMITH, 1970; EIGEN; MCCASKILL; SCHUSTER, 1989). A segunda hipótese assera a seleção natural de mutações favoráveis. No caso de mutações neutras, a fixação ou não ocorre ao acaso (SMITH, 1970).

A força das hipóteses citadas acima repousa principalmente sobre o fato de que nenhuma alternativa plausível foi jamais sugerida e nenhuma evidência para invalidá-las encontrada (SMITH, 1970). Entretanto, uma aparente incompatibilidade na estrutura desse sistema é fruto de bastante discussão e vale a pena ser mencionada.

Para introduzir o problema é necessário antes de tudo ter uma ideia das grandezas envolvidas no número de possibilidades de sequências genômicas e proteicas. Se tomarmos, por exemplo, um comprimento de 100 aminoácidos, que é bastante modesto para uma proteína, e considerando os 20 aminoácidos existentes, isso resulta em 20^{100} proteínas possíveis, cerca de 10^{130} . Para efeito de comparação estima-se que haja no Universo por volta de 10^{80} prótons (NOWAK, 2006). Em uma outra estimativa, dessa vez para uma sequência de DNA com 1000 loci, a chance de uma combinação aleatória de genes produzir um indivíduo tão apto quanto o encontrado na natureza seria de aproximadamente 10^{-100} (SMITH, 1970). A conclusão imediata é de que a evolução não foi nem será capaz de explorar mais que um subconjunto ínfimo de todas as possibilidades.

Consciente deste fato, SALISBURY enxergou uma inconsistência: os números astronômicos de possibilidades associados a genomas e proteínas parece tornar insustentável a adaptação via mutações aleatórias, isto é, mutações favoráveis seriam raras demais para serem encontradas ao acaso. Nesse cenário a seleção natural não teria sobre o que atuar.

Embora bastante apelativo, esse conflito é apenas aparente e fruto de uma maneira incorreta de interpretar as probabilidades em questão. De fato, a probabilidade de uma sequência qualquer ser funcional não é a quantidade relevante no processo evolutivo. A probabilidade que desempenha papel central é a de que um mutante seja melhor ou igualmente eficiente a seu predecessor, isto é, aquele do qual foi gerado, dentro do conjunto particular de todas os mutantes acessíveis a este ancestral (SMITH, 1970). Para entender como essa probabilidade condicional é suficiente pra suportar o processo evolutivo é preciso explorar o significado do termo “mutantes acessíveis” e definir precisamente esse conjunto. Dessa necessidade surge o conceito de espaço de sequências.

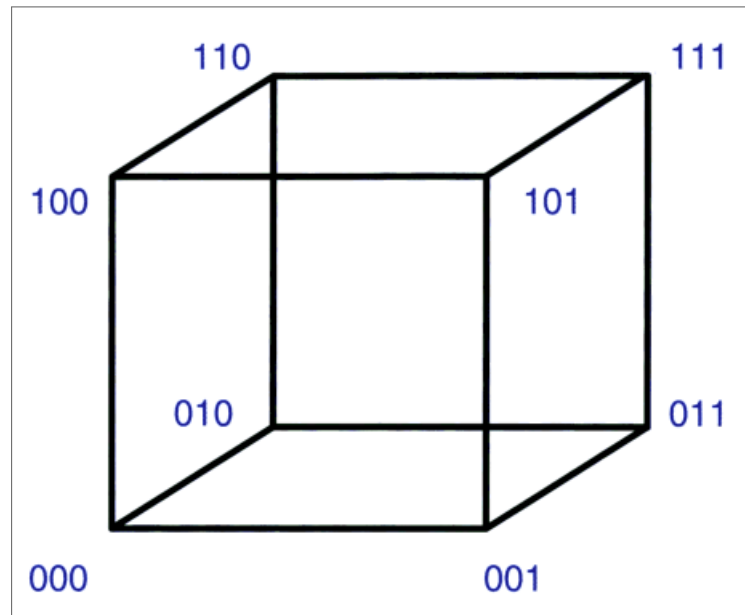
2.2 ESPAÇO DE SEQUÊNCIAS

Para uma dada proteína pode-se distinguir entre duas classes de mutações: (i) as que provocam pequenas mudanças na proteína, como substituição, acréscimo ou deleção de um único aminoácido e (ii) aquelas relacionadas a grandes mudanças. Em muitos aspectos é plausível assumir que mutações envolvendo mais de um aminoácido não contribuem em larga escala pois a probabilidade de que resultem em um produto eficiente deve ser muito baixa (SMITH, 1970).

Assumindo isso, o arquétipo da dinâmica evolutiva que leva uma sequência selvagem a uma sequência mutante favorável torna-se o de um processo que ocorre por passos de mutação unitários, através de sequências intermediárias também favoráveis. Tal processo deve formar, portanto, uma rede contínua de sequências funcionais imersa no espaço de todas as sequências possíveis. Continuidade aqui se refere à passagem exclusivamente por sequências vizinhas, e vizinhança, por sua vez, remete a sequências que diferem entre si pela mutação de um único locus (SMITH, 1970).

Temos assim o espaço de sequências como a abstração do conjunto completo de sequências possíveis de mesmo comprimento dotado do conceito de vizinhança definida pela distância de Hamming, medida pelo número de sítios pelo qual diferem dois objetos (NOWAK, 2006). A Figura 1 apresenta o exemplo de um espaço de sequências binárias, onde vértices representam os genomas e arestas conectam primeiros vizinhos.

Figura 1 – Espaço de sequências binárias de comprimento $L = 3$. Para um alfabeto binário e comprimento arbitrário, o espaço de sequências constitui um hipercubo.



Fonte: NOWAK (2006)

Está claro agora que “mutantes acessíveis” remete ao conjunto de primeiros vizinhos de uma determinada sequência e agora estamos aptos a esclarecer o problema levantado na última seção. Pode-se dizer que o processo evolutivo segue um algoritmo a partir de sequências já existentes e, portanto, necessariamente funcionais. Ocorre que existe uma forte correlação entre a eficiência de uma sequência e suas vizinhas, traduzida numa maior probabilidade de encontrar um mutante eficiente na vizinhança de uma sequência também eficiente do que sortear uma sequência eficiente dentre todas por puro acaso. Ilustrativamente, é mais fácil achar um pico quando se está em uma cordilheira do que ao escolher um lugar qualquer na superfície da Terra (SMITH, 1970).

Embora caricaturas como a anterior sejam bastante úteis, elas possuem suas limitações, e devem ser consideradas com cuidado. Os espaços de sequências são lugares peculiares caracterizados por alta dimensionalidade e uma enorme quantidade de vizinhos próximos, assim como de rotas entre dois pontos. As rotas diretas, em particular, contabilizam $d!$, onde d é a distância de Hamming entre as sequências inicial e final (NOWAK, 1992; EIGEN; MCCASKILL; SCHUSTER, 1989). Tomando novamente a estimativa de prótons do Universo, um espaço com 10^{80} pontos teria um diâmetro de 133 unidades de comprimento, e claramente um pequeno número de mutações levaria a regiões totalmente diferentes. Disso podemos inferir que qual-

quer caminhada em espaços dessa natureza precisa de um guia muito efetivo. Na evolução, a seleção natural desempenha esse papel (NOWAK, 1992).

2.3 RELEVO DE ADAPTAÇÃO

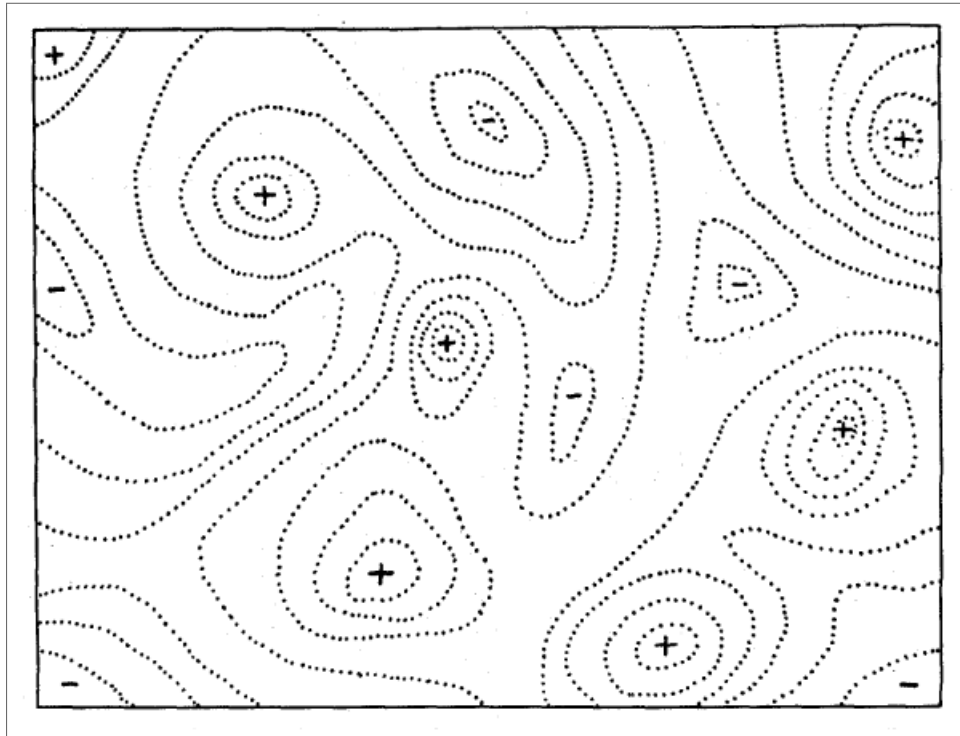
Sem dúvida os conceitos de aptidão e caminhada evolutiva motivaram a formulação do espaço de sequências. Entretanto, a inclusão do aspecto de eficiência e aptidão, conhecido como “fitness”, é feito a posteriori, e dá origem a um outro espaço conhecido como Relevo de Adaptação. O relevo de adaptação foi introduzido por Sewall Wright em 1932 e posteriormente associado ao espaço de sequências por Manfred Eigen (NOWAK, 2006; NOWAK, 1992). WRIGHT propõe o conceito por meio da seguinte pergunta: se todas as combinações possíveis de genes puderem ser graduadas de alguma forma com respeito ao seu valor adaptativo, qual seria a natureza dessa entidade?

Afim de imaginar tal espaço, começemos pela combinação de maior valor adaptativo. A partir dela a aptidão das demais sequências deve decair de forma mais ou menos regular de acordo com o número de substituições pelo qual diferem. Uma distribuição de indivíduos concentrada em uma destas sequências vizinhas seria então capaz de “sentir” o gradiente adaptativo e “caminhar” até se estabilizar no topo, isto é, aquelas sequências próximas da mais apta tendem a dominar na população e mutações direcionadas a ela se fixam com maior frequência (WRIGHT, 1932). Devido à dificuldade de representação gráfica desses espaços hiperdimensionais, as ilustrações em 2D ou 3D são muitas vezes caricaturas impróprias e costumam se assemelhar a representações de curvas de nível como na Figura 2, apresentada por WRIGHT utilizando o termo “campo de combinações de genes”.

É claro que, em geral, esses relevos possuem mais de um pico. Dada uma condição inicial arbitrária, a força seletiva leva a população para o pico mais próximo. Embora existam vários outros picos adaptativos, estes são separados por vales, o que torna o gradiente insuficiente no processo de otimização depois que um pico foi alcançado, ou seja, a longo prazo. O problema evolutivo abstraído sobre um relevo acidentado como este se caracteriza então pela necessidade de um mecanismo de busca que permita à população explorar a vizinhança da pequena região que ocupa à espera de experimentar um gradiente pertencente a outro pico, tendo assim a chance de ser levada continuamente de picos mais baixos para mais altos. O mecanismo em questão é de tentativa e erro, movido pela ocorrência de mutações completamente novas, ainda não presentes na diversidade da população (WRIGHT, 1932).

A intensidade da força seletiva e da taxa de mutação são fatores determinantes para a distribuição da população em torno de um pico. Quanto maior a força seletiva e menor a taxa de mutação, mais centrada é a população e, portanto, menos diversa. Uma seleção mais branda ou uma taxa de mutação alta provocam, por sua vez, uma distribuição mais dispersa (WRIGHT, 1932). O estudo detalhado da dinâmica de distribuição de genomas em uma população é tão importante quanto o relevo de adaptação sobre o qual ocorre e dá origem ao conceito de

Figura 2 – Representação bidimensional de relevo de adaptação fictício. Sinais de + e - representam máximos e mínimos de fitness, respectivamente.



Fonte: WRIGHT (1932)

quasi-espécie, explorado mais a frente.

No que diz respeito aos relevos em si, o conceito se resume à forma como o genótipo e ambiente determinam o fenótipo, enquanto este último determina a taxa de reprodução, ou fitness. O relevo de adaptação é então a convolução dos dois mapeamentos, uma relação direta entre genoma e fitness (NOWAK, 2006). Note que relevos adaptativos não são estáticos tanto quanto o ambiente não é, estando sujeitos a uma mudança gradual e contínua. Estas mudanças abrangem escalas de tempo diversas, que devem ser consideradas cuidadosamente ao se estudar tais relevos (WRIGHT, 1932).

2.4 RELEVOS EXPERIMENTAIS

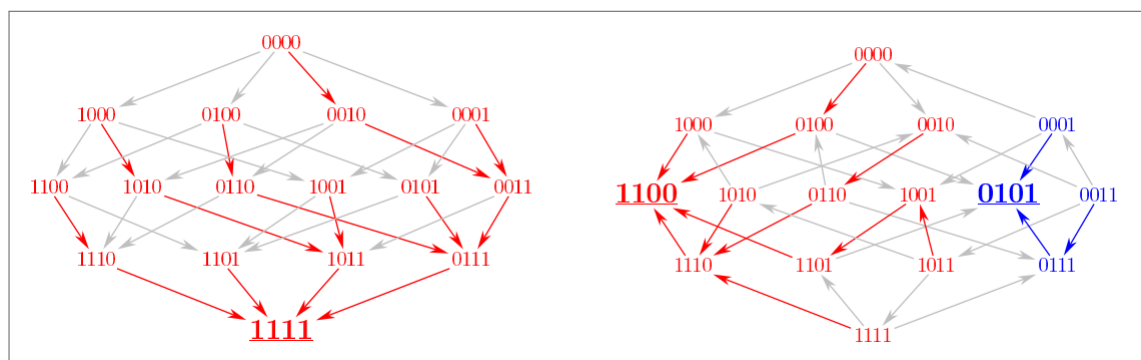
A relação entre genótipo, fenótipo e fitness é uma questão central na biologia e desempenha papel fundamental no entendimento das causas e vínculos do processo evolutivo assim como em teorias envolvendo divergência e especiação, sexo, robustez genética e previsibilidade evolutiva (SZENDRO et al., 2013b; VISSER; KRUG, 2014). Apesar de sua importância conceitual, o impacto do relevo de adaptação na biologia evolucionária tem sido limitado devido à falta de dados empíricos sobre a topografia destes relevos, cenário que tem mudado de maneira significativa apenas recentemente, no século XXI. Trabalhos experimentais tem conseguido construir e analisar o conjunto completo de combinações para pequenos grupos de mutações, fornecendo

vislumbres dos relevos e impulsionando os trabalhos teóricos (VISSER; KRUG, 2014).

Tal mudança se deve principalmente a um avanço em particular. Estudos do fenômeno epistático, isto é, a forma como a expressão de certos genes depende de outros genes, em algumas situações de maneira intensa o suficiente para levar uma mutação de benéfica a deletéria apenas em função do background genômico, sugeriram que a evolução tinha um forte fator determinístico, e portanto, predizível. No entanto, dado um par de genótipos ancestral e seu descendente, e as mutações pelas quais diferem, o estudo da previsibilidade requer informação sobre todos os genomas possíveis envolvendo combinações dessas mutações. Dessa forma, o esforço empírico de construção de pequenos relevos através da criação de mutantes, a mutagênese, e aferição de seu fitness, se colocou como uma necessidade (VISSER; KRUG, 2014).

Limitada a um conjunto pequeno de mutações, esta técnica oferece informação de uma parte ínfima do espaço genotípico. No entanto, a informação obtida é completa dentro desse conjunto, todas as caminhadas evolutivas podem ser exploradas e comparadas. Técnicas alternativas podem oferecer informações sobre o espaço como um todo, porém ao custo de serem majoritariamente de caráter qualitativo (VISSER; KRUG, 2014). Atualmente o número de estudos sistemáticos em relevos empíricos se aproxima de 20 e continua a crescer, envolvendo conjuntos de 3 a 9 genes. Em geral, utiliza-se um representante do fitness, características como resistência a antibióticos ou taxa máxima de reprodução. Na Figura 3 são ilustrados dois relevos experimentais construídos dessa forma. Infelizmente a quantidade de dados envolvendo essa classe de experimentos ainda é insuficiente para permitir conclusões firmes a respeito dos relevos reais, devido principalmente à variabilidade dos métodos, sistemas e tipos de mutação utilizados (VISSER; KRUG, 2014).

Figura 3 – Dois relevos experimentais são representados por sequências binárias onde 0 e 1 representam, respectivamente, ausência e ocorrência de uma determinada mutação. Setas coloridas apontam para o vizinho de maior fitness e um conjunto de setas da mesma cor indica uma bacia de atração para um máximo local. À esquerda, o relevo de mutações benéficas para a bactéria *Methylbacterium extorquens* e, à direita, o relevo de mutações associadas à resistência a medicamentos contra malária.



Fonte: VISSER; KRUG (2014)

É importante ressaltar que tais estudos sofrem de alguns vieses, e destacamos aqui os principais. O conjunto de mutações explorado pertence a um grupo específico de mutações que

foi selecionado individual ou coletivamente, e cujo fenótipo é observável. Entretanto, populações sofrem um número de mutações muito além daquelas fixadas, e mutações de fenótipo desconhecido também contribuem no processo evolutivo. Este é um problema de escala, cuja solução requer o estudo de relevos de dimensão real. Na prática, porém, trata-se de uma tarefa limitada pelo crescimento exponencial de genótipos possíveis com o comprimento. Os primeiros estudos em relevos completos de sequências pequenas já estão ocorrendo. No entanto, a análise é basicamente impraticável nessa escala e a informação possivelmente enviesada, devido principalmente ao uso de amostras muito pequenas de cada genótipo (VISSER; KRUG, 2014).

2.5 RELEVOS TEÓRICOS

Paralelamente ao lento progresso dos estudos experimentais, a área teórica foi bastante explorada através de diversos modelos. Tais modelos foram construídos sobre uma grande variedade de hipóteses, algumas até contraditórias entre si (SZENDRO et al., 2013b). Eles tem um importante papel complementar aos avanços experimentais, sendo úteis, por exemplo, para quantificar os desvios associados à inferência de relevos experimentais em larga escala a partir de amostras pequenas (VISSER; KRUG, 2014).

Uma classe de modelos bastante explorada é a dos modelos de campo aleatório. Neles atribui-se o fitness diretamente a cada genótipo por meio de algum algoritmo probabilístico. São especialmente úteis na avaliação de como os relevos variam com a dimensionalidade. O modelo mais simples dessa classe é conhecido como *House of Cards*, no qual o fitness é atribuído independentemente a cada sequência a partir de uma distribuição de probabilidade fixa (VISSER; KRUG, 2014).

Outro exemplo é o *Rough Monti Fuji*, que parte de um modelo aditivo, isto é, onde cada locus contribui aditivamente para o fitness, e então aplica-se um ruído a essa contribuição individual. Através da intensidade do ruído é possível ajustar o desvio da aditividade e, portanto, a rugosidade do relevo (VISSER; KRUG, 2014).

Seguindo essa capacidade de ajuste do *Rough Monti Fuji* temos o Modelo NK. Ele é caracterizado pelo comprimento N dos genomas e pelo número K de interações epistáticas. Um dado gene forma um conjunto de $K + 1$ genes interagentes, a este conjunto atribui-se aleatoriamente o fitness das 2^{K+1} combinações possíveis (VISSER; KRUG, 2014). Para $K = 0$ a contribuição de cada gene é independente dos demais e a sequência de maior valor adaptativo é obtida otimizando todos os genes simultânea e individualmente, dando origem ao único máximo do relevo. Com $K = N - 1$, no entanto, cada gene depende de todos os outros e, portanto, não é possível otimizá-los simultaneamente, caracterizando o fenômeno de frustração e resultando num relevo totalmente acidentado. Entre esses dois extremos, níveis intermediários de rugosidade podem ser obtidos. O Modelo NK incorpora a capacidade de simular as estatísticas associadas às interações genéticas que levam a diferentes rugosidades

no relevo sem a necessidade de formulação dos mecanismos bioquímicos envolvidos (SZENDRO et al., 2013b).

Um outro tipo de abordagem conceitual na construção de relevos teóricos envolve a hipótese de que a funcionalidade ou não de uma mutação precede a importância das diferenças no valor adaptativo, enfatizadas nos modelos citados anteriormente. Nos modelos neutros, como são chamados, existe uma probabilidade p de uma sequência possuir fitness unitário e uma probabilidade $1 - p$ de possuir fitness nulo. Esses modelos também podem ser incorporados aos primeiros através de uma taxa de letalidade que torna inviável uma fração das sequências do relevo (SZENDRO et al., 2013b).

Todos os relevos citados até agora compartilham uma semelhança. Nesses modelos a formulação das origens estruturais de propriedades como rugosidade são evitadas em favor da aplicação direta de seus efeitos imediatos. Construções a priori, baseadas nas bases químicas e físicas de proteínas e sequências genéticas, como afinidade de ligação, estabilidade e interações, fornecem uma classe alternativa de modelos feitos a partir de primeiros princípios. Devido à sua maior complexidade costumam requerer grande capacidade de simulação e análise, o que pode representar uma dificuldade desnecessária dependendo do objeto de estudo (SZENDRO et al., 2013b; VISSER; KRUG, 2014).

2.6 QUASI-ESPÉCIE

Nas últimas duas seções analisamos os aspectos teóricos e experimentais envolvendo a topografia dos relevos de adaptação. Entretanto, igualmente importante à construção de tais relevos é o entendimento da dinâmica adaptativa que ocorre neles. Em química, o termo espécie designa um conjunto de moléculas idênticas, e não tem nenhuma relação com a definição biológica. Entretanto, as proteínas assim como o DNA e RNA não constituem um conjunto de moléculas iguais, revelando a necessidade de um novo termo para esse grupo (NOWAK, 2006). O termo “quasi-espécie” foi então introduzido por Manfred Eigen e Peter Schuster para descrever o conjunto de moléculas estruturalmente parecidas fruto de erros no processo de replicação do DNA (NOWAK, 1992).

Mais precisamente, a quasi-espécie é definida como a distribuição estacionária de genomas em um processo de reprodução sujeito a mutação e seleção (NOWAK, 1992; DOMINGO; SCHUSTER, 2015). Tal dinâmica pode ser descrita por um sistema de equações diferenciais que apresentamos agora. Consideremos uma população infinita de indivíduos cada um carregando um dos n possíveis genomas de comprimento fixo. A frequência relativa do genoma i nesta população é dada por x_i , onde $x_i \in [0, 1]$ e $\sum_{i=1}^n x_i = 1$. A estrutura da população em um dado instante é completamente descrita pelo vetor $\vec{x} = (x_1, x_2, \dots, x_n)$. Cada indivíduo i se reproduz a uma taxa f_i , o relevo de fitness fica determinado pelo vetor $\vec{f} = (f_1, f_2, \dots, f_n)$, e o fitness médio da população é $\phi = \sum_{i=1}^n x_i f_i$, fruto do produto interno entre $\vec{x}$ e $\vec{f}$. Por último, entra em cena a mutação. A chance de um indivíduo do tipo i produzir um do tipo j

é dada pelo elemento q_{ij} da matriz de mutação $Q = [q_{ij}]$, uma matriz estocástica. Definidos os termos, o sistema de equações diferenciais que determina a evolução temporal do ensemble é apresentado na Equação 2.1 (NOWAK, 2006).

$$\dot{x}_i = \sum_{j=0}^n x_j f_j q_{ji} - \phi x_i \quad i = 0, 1, \dots, n. \quad (2.1)$$

De acordo com o primeiro termo da Equação 2.1 o incremento na frequência do genoma i recebe uma contribuição de cada genoma j proporcionalmente às taxas de reprodução f_j , e mutação q_{ji} de j para i , incluindo a chance de reprodução sem erros q_{ii} . Através do segundo termo cada genótipo é removido proporcionalmente a sua frequência afim de manter a população constante. O fator de proporcionalidade que garante o desconto preciso é justamente o fitness médio ϕ (NOWAK, 2006). É claro que o fitness e a matriz de mutação podem ser incorporados numa única matriz $W = [f_i q_{ij}]$ que carrega toda a informação sobre a dinâmica do problema. A Equação 2.1 pode então ser escrita em forma vetorial como apresentado na Equação 2.2, onde I representa a matriz identidade (NOWAK, 1992).

$$\dot{\vec{x}} = (W - \phi I) \vec{x} \quad (2.2)$$

Soluções estacionárias das Equações 2.1 e 2.2, se existirem, são precisamente a quasi-espécie. Podemos abordar agora algumas propriedades e termos gerais. Primeiro, a quasi-espécie está distribuída em torno da sequência mestra, aquela de maior frequência na população, não necessariamente a mais apta, embora seja o caso mais comum (DOMINGO; SCHUSTER, 2015). A sequência consensual é a sequência média da distribuição e tende a coincidir com a sequência mestra, a menos que haja uma alta taxa de mutação (NOWAK, 1992; DOMINGO; SCHUSTER, 2015).

A frequência de um dado mutante na quasi-espécie é essencialmente determinada por dois fatores: (i) sua distância de Hamming em relação à sequência mestra e (ii) a diferença de fitness com respeito a mesma. Esses fortes vínculos tornam a quasi-espécie mais que um conjunto arbitrário de mutantes, mas sim uma distribuição muito bem definida centrada na sequência mestra (DOMINGO; SCHUSTER, 2015).

O comportamento do fitness médio em função do tempo até que se atinja o equilíbrio depende da condição inicial e, portanto, não tem um comportamento geral. Entretanto, sabe-se que a região do espaço que leva a um comportamento monotonicamente crescente cobre a grande maioria do mesmo, condizente com o arquétipo evolutivo (DOMINGO; SCHUSTER, 2015).

Entre as observações importantes podemos citar que o número astronômico de possibilidades no espaço de sequências impõe um limite à realidade do modelo de quasi-espécie: devido à sua finitude, populações reais são apenas locais no espaço de sequências, de forma que soluções globais não existem. Entretanto, sabe-se que mutações importantes ocorrem dentro

de uma vizinhança pequena da sequência mestra, de tal forma que soluções locais coincidem com soluções globais em geral (DOMINGO; SCHUSTER, 2015).

A quasi-espécie se comporta como um reservatório de mutantes, provendo um conjunto de variantes potencialmente úteis à adaptação (DOMINGO; SCHUSTER, 2015) e, por isso, interpreta-se que ela é o verdadeiro objeto de ação da evolução (NOWAK, 1992; EIGEN; MCCASKILL; SCHUSTER, 1989). A seleção atua estabilizando a distribuição da quasi-espécie enquanto a evolução é a desestabilização da quasi-espécie frente a uma nova mutação vantajosa (NOWAK, 1992).

Como já foi dito antes, uma população é capaz de enxergar gradientes no relevo de adaptação e caminhar até se estabilizar em torno de um pico. Diz-se que a população está localizada, sinônimo de adaptada. Este processo de localização no relevo é sensível não só à topologia do mesmo como também à taxa de mutação. Quanto menor a taxa de mutação mais centrada a distribuição vai estar, quanto maior, mais dispersa. É plausível supor, portanto, que essa dispersão pode ser levada ao limite, e para uma taxa de mutação crítica a população perderia a referência do pico e com ela a adaptação. Tal taxa é conhecida como limite de erro e tem uma importante relação com a estabilidade da quasi-espécie, embora seja um tema explorado essencialmente no campo teórico (NOWAK, 2006).

2.7 PREVISIBILIDADE EVOLUTIVA

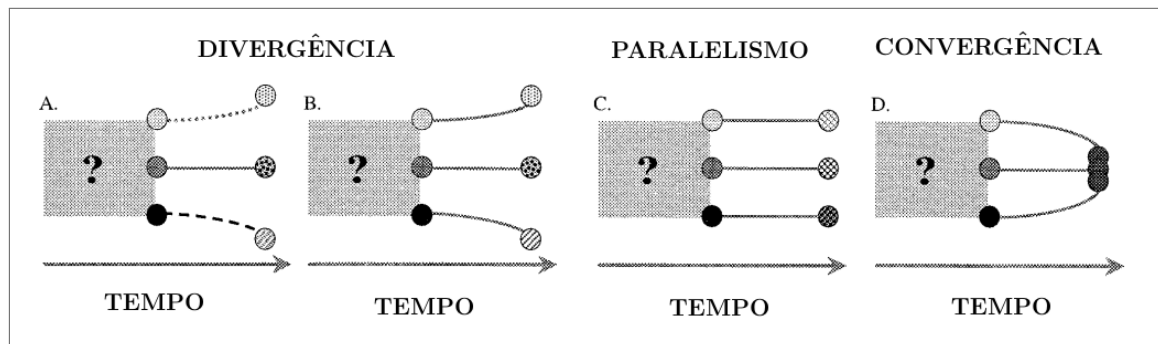
Existe atualmente uma visão praticamente indissociável da evolução na qual seus resultados e rumos são aleatórios, frutos da mais pura improvisação. Entretanto, algumas alternativas devem ser consideradas e listamos aqui perguntas pertinentes nesse sentido, de acordo com MORRIS:

- Seria a evolução predizível de alguma maneira?
- É possível identificar e quantificar a viabilidade de formas alternativas de organização biológica?
- Quão fortuitas são as grandes transições na história da vida?
- Quão vinculada pelo passado é a diversificação evolutiva?
- Soluções evolutivas se aproximam mais de uma reposta meramente suficiente ou uma altamente eficiente?

Essas perguntas, longe de representarem uma objeção à existência da evolução e à consistência do mecanismo de seleção, trazem à tona apenas o fato de que muito provavelmente não temos ainda uma explicação completa do fenômeno (MORRIS, 2009). Não há dúvidas de que a natureza ostenta uma grande divergência de populações e espécies e, mesmo na hipótese de

que a divergência seja inerentemente adaptativa, ainda nos resta outra questão: esta divergência é uma resposta adaptativa a diferentes ambientes ou uma adaptação diferente ao mesmo ambiente? Em outras palavras, a diversidade de soluções representa a diversidade do problema ou a contingência do processo (DYKHUIZEN, 1992)? Na prática, duas linhas podem ser diretamente exploradas: (i) populações idênticas evoluem em paralelo e (ii) populações diferentes convergem para a mesma solução (COLEGRAVE; BUCKLING, 2005)? Paralelismo, convergência e divergência são ilustradas na Figura 4.

Figura 4 – Populações sofrem diferenciação a partir de um ancestral desconhecido sujeitas a determinadas condições ambientais (linhas). C e D apresentam repetibilidade, embora com padrões diferentes de evolução, paralela e convergente, respectivamente. A e B mostram divergência, onde diferenças iniciais são amplificadas por condições ambientais diferentes em A, e por condições semelhantes em B. Em C tais diferenças são mantidas ao longo do tempo.



Fonte: TEOTÔNIO; ROSE (2001)

O processo adaptativo é fundamentalmente guiado por dois tipos de fatores, forças seletivas determinísticas e eventos fortuitos de reprodução e mutação, cuja influência no resultado evolutivo é causa de intensa discussão e motivo de grande esforço experimental e teórico. Experimentalmente, o estudo da previsibilidade tem sido abordado em escala microevolutiva através da repetibilidade de mudanças adaptativas em populações replicadas de micróbios. Em um desses estudos observou-se, por exemplo, fortes indícios de paralelismo na aquisição de resistência a antibióticos por certos patógenos, um resultado altamente relevante no contexto do desenvolvimento farmacêutico. É importante notar que a repetibilidade em experimentos de replicação fornece uma medida a posteriori, e portanto fraca, de previsibilidade, uma vez que não permite qualquer predição além do observado (SZENDRO et al., 2013a). Enquanto os primeiros estudos experimentais de previsibilidade limitavam-se à inferência a partir de seu resultado final, atualmente o uso de microrganismos oferece um procedimento muito mais rico. Devido ao curto tempo entre gerações, microrganismos podem ser usados para acompanhar a adaptação em tempo real e sob condições ambientais controladas. Além disso, tais seres podem ser mantidos em estado inanimado, o que permite comparação direta entre ancestrais e descendentes (DYKHUIZEN, 1992; COLEGRAVE; BUCKLING, 2005).

Uma das pesquisas mais antigas em evolução microbiana envolve a bactéria *Escherichia coli*, explorada desde 1988, e levanta alguns resultados interessantes. Em um dos experimentos (LENSKI et al., 1991), doze populações idênticas foram sujeitas a um meio limitado em glicose

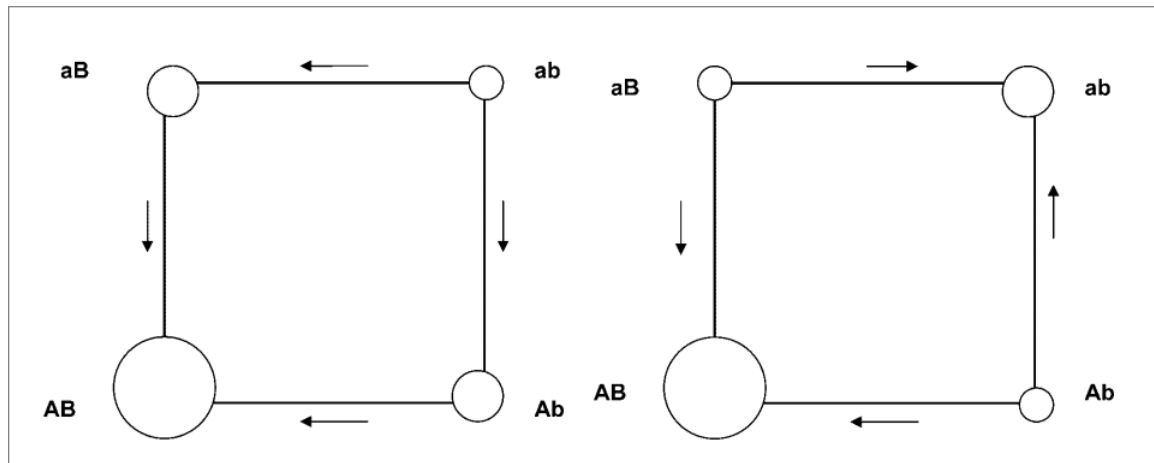
e seu fitness medido através de competição direta com o ancestral. Todas apresentaram um mesmo padrão de crescimento do fitness: rápido no início e lento por volta da milésima geração, sugerindo a aproximação de um pico. Por outro lado, uma diferença pequena mas significativa no fitness foi mantida entre as populações. Quando submetidas a outras condições ambientais, no entanto, as mesmas populações apresentaram grande variação em fitness, indicando uma variabilidade igualmente intensa a nível genético. Este resultado mostra que fenótipos similares podem ser obtidos de maneiras diversas, por meio de mutações em diferentes genes. Nesse mesmo estudo explorou-se ainda o papel das condições iniciais. Embora populações diferentes tenham apresentado divergências no fitness, essa divergência não dependeu do ponto de partida (COLEGRAVE; BUCKLING, 2005).

As diferenças inter-populacionais observadas no experimento não são conclusivas e sugerem três possíveis cenários. O primeiro deles assume que efeitos estocásticos influenciam a ordem e o tempo de ocorrência das mutações benéficas provocando a divergência. Essa divergência, no entanto, seria transiente. No segundo cenário, também por efeitos estocásticos populações atingem diferentes picos, e a divergência é mantida indefinidamente. Essa situação é suportada pela existência de um relevo acidentado e abundante em máximos locais. A última possibilidade é de que o relevo seja consideravelmente sensível à distribuição de genótipos na população, isto é, que os indivíduos o influenciem significativamente, tornando-o instável (DYKHUIZEN, 1992).

Muitos fatores podem influenciar a repetibilidade de trajetórias adaptativas, os principais são tamanho de população, taxa de mutação e topologia do relevo, este último intimamente relacionado às interações epistáticas (SZENDRO et al., 2013a). No que diz respeito à relação entre topologia e previsibilidade, um aspecto notadamente importante é a rugosidade do relevo. Em um relevo com um único pico, por exemplo, espera-se que independentemente do genótipo inicial e da ordem em que ocorram as mutações, ao final de qualquer caminhada a população terá alcançado o máximo global. Em um relevo acidentado, no entanto, mesmo fixada a condição inicial cada passo pode ser decisivo ao levar na direção da bacia de atração de um dos vários máximos locais. Assim, a distinção entre relevos suaves e rugosos caracteriza uma das formas mais gerais e comuns de abordagem do problema de previsibilidade evolutiva. A rugosidade e abundância de máximos locais tem ainda um outro efeito imediato: frequentemente a população não atingirá o melhor resultado possível (COLEGRAVE; BUCKLING, 2005; KORONA et al., 1994). A grande variabilidade dos relevos é tão marcante que mesmo o menor relevo possível, um de dimensão dois com apenas um tipo de mutação em cada gene (Figura 5), não é um caso trivial no que diz respeito à aspectos como rugosidade e número de picos, e conseqüentemente, caminhadas possíveis. Este relevo é capaz de apresentar duas situações notadamente diferentes que, em última instância, são decisivas na capacidade de previsão (COLEGRAVE; BUCKLING, 2005).

De acordo com a intuição dos processos estocásticos espera-se naturalmente que a dependência da previsibilidade evolutiva com o tamanho de população seja muito simples: quanto

Figura 5 – Relevo adaptativo de dois genes com possibilidade de apenas uma única mutação em cada. O tamanho das circunferências representa o fitness do genótipo e mudanças adaptativas favoráveis são indicadas pelas setas. À esquerda, um relevo suave, onde qualquer genótipo inicial inevitavelmente leva ao único máximo “AB”. À direita, um relevo rugoso, com dois máximos “AB” e “ab”. Populações centradas em “Ab” ou “aB” podem atingir ambos os picos, dependendo da ordem em que sofrem as mutações benéficas. Uma população iniciada em “ab” não é capaz de alcançar o máximo global.



Fonte: COLEGRAVE; BUCKLING (2005)

maior a população, mais determinístico o sistema, tornando-se total para uma população infinita. Uma análise mais detalhada do problema, no entanto, mostra que trata-se de uma relação muito mais rica. População e taxa de mutação juntas determinam regimes de criação de mutantes cujo efeito sobre a repetibilidade de trajetórias não é trivial, a citar: SSWM (*strong selection weak mutation*), interferência clonal, e tunelamento estocástico (SZENDRO et al., 2013a). O primeiro regime é mais simples, SSWM, é caracterizado por uma taxa de mutação μ baixa o suficiente tal que o surgimento e fixação de um determinado mutante ocorre isoladamente, isto é, raramente existe a presença de duas ou mais mutações diferentes. Dessa forma, a população se comporta como uma entidade homogênea, um caminhante forçado a realizar passos de mutação unitários ascendentes em fitness. Em geral, tais caminhos monotonicamente crescentes não são abundantes nos relevos, o que implica uma considerável previsibilidade no SSWM (SZENDRO et al., 2013a).

À medida que a população N aumenta, a existência simultânea de diferentes mutações provoca o fenômeno da interferência clonal. Nessa situação, o conflito entre mutações individuais favorece aquelas de maior fitness, tornando os passos mais tendenciosos e, portanto, mais determinístico o processo. Esse comportamento aproxima a caminhada do limite “greedy” (ganancioso), onde a mutação mais eficiente é fixada com certeza, situação maximamente previsível. Com incrementos ainda maiores em N um outro evento se torna proeminente, a abundância de indivíduos carregando múltiplas mutações. De fato, com uma população maior mutações deletérias sobrevivem tempo o suficiente para permitir novas mutações no mesmo indivíduo, permitindo assim o salto sobre vales de fitness, efeito conhecido como tunelamento estocástico. Com isso caminhos totalmente novos são permitidos, cuja quantidade é proporci-

onal ao número de segundos vizinhos, consideravelmente superior ao de primeiros. O resultado é uma inesperada diminuição da repetibilidade com o aumento da população. População e taxa de mutação determinam esses regimes através dos parâmetros $N\mu$ e $N\mu^2$, que correspondem ao suprimento de mutações individuais e duplas, respectivamente. Em resumo, para N e μ tais que $N\mu \ll 1$ tem-se o regime SSWM, para $N\mu^2 < 1 < N\mu$ ou próximo desse intervalo, tanto interferência quanto tunelamento ocorrem, com predominância do primeiro, enquanto que a situação se inverte para $1 < N\mu^2$. É importante notar que trata-se apenas de um cenário aproximado, e que tais conclusões podem depender sensivelmente do relevo entre outros aspectos (SZENDRO et al., 2013a).

Apesar de sua solidez ostensiva, a teoria Darwinista não deve ser considerada uma teoria completa. Existe ainda uma grande incompreensão de aspectos fundamentais, a citar, a maneira como os organismos tomam formas cada vez mais complexas e porque é tão recorrente a convergência evolutiva. Subjacente a estas dúvidas está uma crescente evidencia de que estes processos são predizíveis em algum nível (MORRIS, 2010).

3 FUNDAMENTAÇÃO ESPECÍFICA

3.1 MEDIDAS DE PREVISIBILIDADE

Uma das principais perguntas que guiam o estudo da previsibilidade evolutiva é a seguinte: o processo adaptativo que leva a um fenótipo previamente conhecido segue uma rota repetidamente de maneira quantificável? De outra maneira, conhecidos os fenótipos inicial e final, quão similares quantitativamente são as trajetórias que levam de um ao outro (LOBKOVSKY; KOONIN, 2012)? Uma primeira análise do problema de mensurar a previsibilidade deve fazer menção, sem dúvida, à importância de calcular a fração de trajetórias acessíveis à evolução, dados dois pontos fixos em um relevo (LOBKOVSKY; WOLF; KOONIN, 2011). Outra noção fundamental é de que qualquer medida de previsibilidade deve ser definida sobre um ensemble de trajetórias, cada qual com sua probabilidade de ocorrência sob determinada dinâmica (LOBKOVSKY; KOONIN, 2012). Nesta seção apresentamos algumas das medidas mais importantes nesse contexto, como são construídas, e seu significado.

Embora o entendimento das contribuições de determinismo e estocasticidade na evolução sempre tenha sido um problema extremamente interessante, sua abordagem direta não foi possível até recentemente. Enquanto resultados como a dominância das mutações neutras ou mesmo a forma como genes independentes influenciam uma mesma função biológica já apontavam para uma presença considerável de aleatoriedade, a imprevisibilidade permaneceu ainda um aspecto qualitativo, e uma definição precisa baseada em trajetórias se fazia necessária (LOBKOVSKY; KOONIN, 2012). Os avanços tecnológicos da última década, a citar o sequenciamento genético e a mutagênese, abriram as portas da área experimental em evolução e permitiram o tratamento direto do problema da previsibilidade. A predição de trajetórias evolutivas ainda é, no entanto, uma tarefa desafiadora que requer um entendimento completo e detalhado dos efeitos das mutações no fenótipo (LOBKOVSKY; KOONIN, 2012; LOBKOVSKY; WOLF; KOONIN, 2011). Mesmo com as técnicas mais avançadas, o crescimento exponencial de combinações em função do número de mutações limita a resolução do problema em nível de nucleotídeos a situações envolvendo menos de uma dezena de mutações (LOBKOVSKY; KOONIN, 2012).

A primeira forma direta de abordagem quantitativa da previsibilidade evolutiva foram os experimentos de evolução paralela. Embora apresentem resultados promissores, tais experimentos envolvem uma série de dificuldades técnicas. Afim de se obter estatísticas significativas é necessária a caracterização de inúmeros clones a partir de uma grande quantidade de populações, assim como uma uniformidade estrita ao longo do tempo e entre populações uma vez que condições ambientais flutuantes levam a uma maior diversidade. Em situações com populações muito grandes existe ainda a necessidade de uma análise profunda da dinâmica de interferência clonal. Em razão destas dificuldades inerentes, as conclusões obtidas sob evolução paralela são majoritariamente qualitativas. Uma alternativa experimental favorável ao tratamento quantitativo da repetibilidade evolutiva é baseada na medição e análise de relevos

de adaptação. Dados oriundos dessa linha já mostram avanços significativos no entendimento da epistasia, por exemplo. No que diz respeito à previsibilidade, alguns aspectos são de especial interesse como a relação da rugosidade do relevo com a repetibilidade de trajetórias e a estatística de acessibilidade dos máximos locais (LOBKOVSKY; KOONIN, 2012).

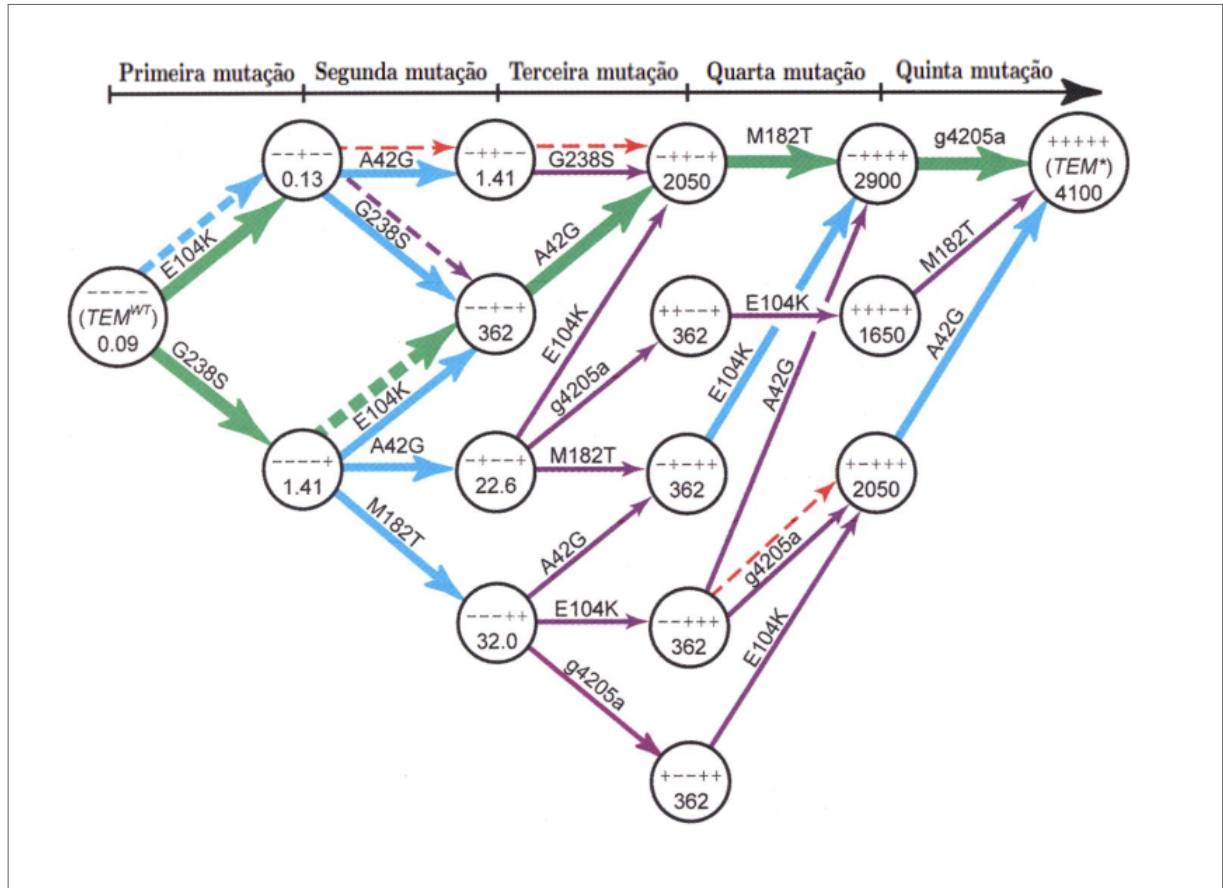
O limite SSWM é particularmente interessante para se entender os efeitos da topografia do relevo na previsibilidade sem precisar considerar dependências com a dinâmica, dado que a dinâmica populacional real pode envolver interferência entre mutações múltiplas além de outras complexidades (MANHART; MOROZOV, 2014). Uma vez que todas as trajetórias nesse regime são monotônicas no relevo, a estatística desses caminhos é uma quantificação importante da acessibilidade dos picos e, em algum grau, da previsibilidade das caminhadas adaptativas (LOBKOVSKY; KOONIN, 2012).

Um trabalho notável assumindo a dinâmica SSWM foi realizado por WEINREICH et al.. Nele mediu-se a acessibilidade das 120 ($5!$) trajetórias possíveis envolvendo 5 mutações que conferiam resistência a um determinado antibiótico na bactéria *Escherichia coli*. O estudo foi baseado no relevo de fitness construído através da medição da resistência antibiótica de todas as 32 (2^5) combinações possíveis das mutações disponíveis (Figura 6). Um resultado imediato foi a descoberta de que apenas uma das 5 mutações sempre conferia acréscimo na resistência, independentemente da presença ou não de mutação nos demais loci. Além disso, notou-se também que três das mutações podiam até mesmo diminuir esta habilidade dependendo das mutações nos outros sítios. Logo, não foi surpreendente a constatação de que apenas 18 das 120 trajetórias possíveis eram monotonicamente crescentes em fitness e, portanto, acessíveis no regime SSWM. Vale ressaltar que, no contexto da análise de trajetórias há uma importante distinção entre caminho possível e acessível, o primeiro se referindo às combinações permitidas pelo tipo de mutação envolvida, como mutações simples ou duplas, enquanto o último está relacionado aos caminhos permitidos pela dinâmica de seleção em questão, por exemplo, a adaptação por mutações benéficas ou neutras (ROY, 2009).

É intuitiva a noção de que quanto maior o número de trajetórias direcionadas a um determinado máximo, menor a previsibilidade do processo. Embora tal suposição de uma correlação negativa entre número de trajetórias e previsibilidade seja uma consideração razoável, ela pode ser muito mais precisa em dois aspectos fundamentais: (i) em relação aos detalhes da distribuição, isto é, o peso de cada trajetória no ensemble e (ii) o grau de similaridade entre os caminhos (LOBKOVSKY; KOONIN, 2012). Seguindo o primeiro ponto, ROY introduziu uma medida que chamaremos de *predictability*, visando explorar o efeito de mudanças da dinâmica sobre a previsibilidade no mesmo relevo experimental obtido por WEINREICH et al..

Considere a probabilidade de que duas tentativas de adaptação independentes tomem exatamente o mesmo trajeto. Enquanto a repetibilidade de um dado caminho medida assim é simplesmente o quadrado da sua probabilidade de ocorrência, a repetibilidade total, ou

Figura 6 – Diagrama adaptativo das 10 trajetórias mais prováveis entre o genótipo original TEM^{wt} e o mutante resistente TEM*. Circunferências representam o genótipo, cuja identidade é descrita por uma sequência de símbolos + e – que correspondem respectivamente à presença ou ausência das mutações g4205a, A42G, E104K, M182T e G238S, nesta ordem. Números representam a resistência ao antibiótico cefotaxima em mg/ml. Setas representam as mutações nelas rotuladas. A probabilidade de ocorrência destas mutações é exibida em escala logarítmica tanto em cor quanto largura da seguinte forma: verde/largo, 0,316 a 1,0; azul/médio, 0,1 a 0,316; roxo/estreito, 0,0316 a 0,1; vermelho/muito estreito, menor que 0,0316. Onde há setas duplas, contínuas e segmentadas representam as probabilidades sob os modelos de fixação correlacionada e descorrelacionada com o fitness, respectivamente.



Fonte: WEINREICH et al. (2006)

predictability, é a soma dos quadrados da probabilidade de cada caminho acessível

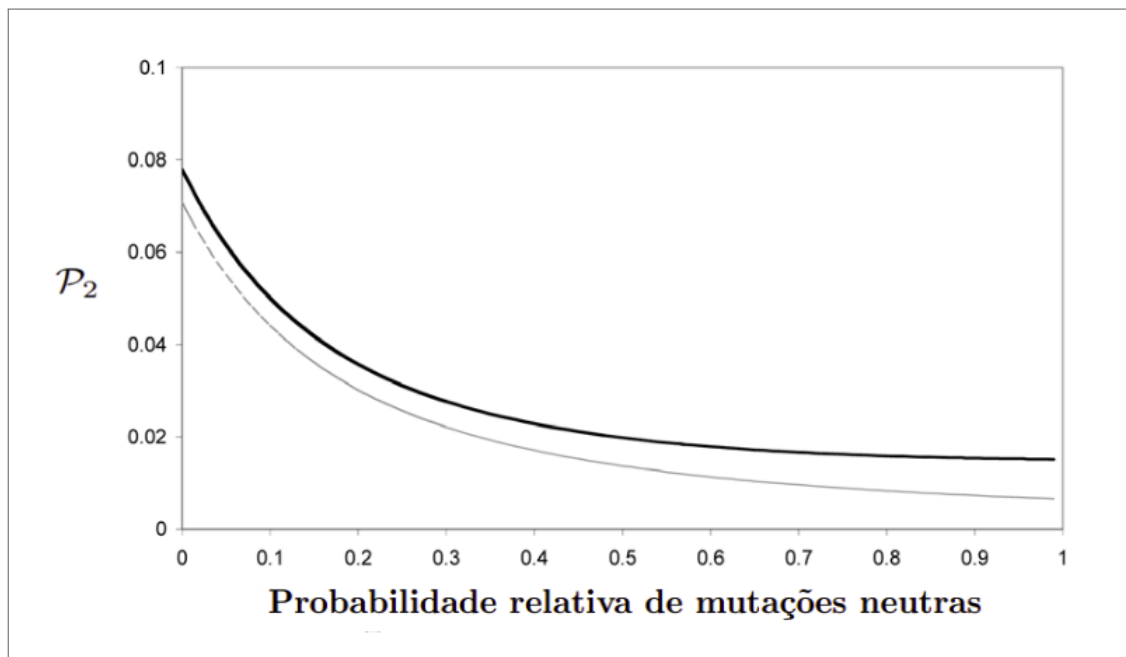
$$\mathcal{P}_2 = \sum_i^n p^2(\phi_i), \quad (3.1)$$

onde ϕ_i representa o i -ésimo caminho, $p(\phi_i)$ é a probabilidade de ocorrência deste caminho, e a soma é feita sobre o ensemble de n trajetórias (ROY, 2009). Trata-se de uma medida simples de repetibilidade que varia de $\mathcal{P}_2 = 1$ na ocorrência de um único caminho a $\mathcal{P}_2 = 1/n$ para n caminhos igualmente prováveis, situação maximamente imprevisível. Dessa forma, o inverso de $\mathcal{P}_2$ pode ser interpretado como o número efetivo de caminhos que contribuem no processo evolutivo. Uma generalização imediata desta quantidade é a probabilidade de observar um mesmo caminho em um número $k > 2$ de tentativas aleatórias $\mathcal{P}_k = \sum_i^n p^k(\phi_i)$.

Analogamente, tal medida estará compreendida entre $\mathcal{P}_k = 1$ e $\mathcal{P}_k = 1/(n^{k-1})$ (VISSER; KRUG, 2014). Neste trabalho, no entanto, aplicaremos apenas $\mathcal{P}_2$. A *predictability* é assim uma medida de entropia no que diz respeito às caminhadas.

Munido desta medida, ROY pôde inferir a previsibilidade de duas maneiras distintas, usando o número absoluto de caminhos acessíveis, como WEINREICH et al. e, adicionalmente, a distribuição de probabilidades desses caminhos. A segunda maneira de olhar para o problema revela a noção intuitiva de que mesmo um grande aumento no número total de caminhos pode representar pouquíssima mudança na previsibilidade, caso as probabilidades destes novos trajetos sejam suficientemente baixas. A *predictability* é a medida ideal na captura desse aspecto. Em seu trabalho, ROY faz uma análise comparativa da estatística de caminhos sob duas dinâmicas de forma a evidenciar a complementaridade desta nova medida com a acessibilidade. No primeiro caso, mutações necessariamente diminuem a distância ao máximo enquanto no segundo é permitida também uma única mudança que aumente tal distância (mutação retrógrada). Ele observou que, embora a passagem da primeira para a segunda situação provoque um aumento de 50% na fração de caminhos acessíveis, a *predictability* praticamente não foi afetada, sofrendo decréscimo de 9%. Ainda nesse mesmo estudo comparou-se, de maneira similar, o efeito da permissão de mutações neutras na repetibilidade em função da sua probabilidade de ocorrência (ROY, 2009), como mostrado na Figura 7.

Figura 7 – *Predictability* em função da probabilidade de mutações neutras relativamente à probabilidade de mutações benéficas. Linhas preta e cinza são obtidas excluindo e admitindo a possibilidade de mutações não unitárias, respectivamente. Quando as mutações neutras são muito menos prováveis que mutações benéficas (1%), $\mathcal{P}_2$ sofre pouca alteração (5%). Quando as probabilidades são aproximadamente iguais, no entanto, $\mathcal{P}_2$ diminui significativamente, em torno de quatro vezes.



Fonte: ROY (2009)

Embora a estatística de caminhos monotônicos revele informação considerável sobre o

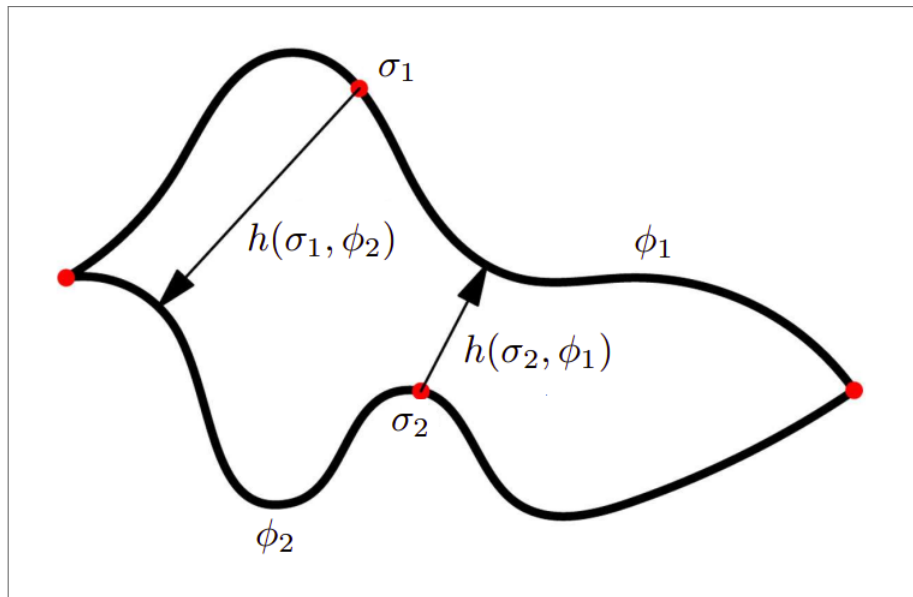
relevo, uma retrospectiva confiável do passado evolutivo só é possível se tais caminhos forem parecidos de uma forma quantificável (LOBKOVSKY; WOLF; KOONIN, 2011). Dessa necessidade surge outra medida importante, a divergência média de caminhos, denotada por $\mathcal{D}$. Esta medida é construída sobre uma ideia muito simples acerca da relação entre previsibilidade e proximidade dos caminhos: a um número alto de caminhos similares e agrupados em uma região deve estar associado um maior grau de repetibilidade do que a um conjunto pequeno de caminhos significativamente diferentes (LOBKOVSKY; KOONIN, 2012).

Sejam ϕ_1 e ϕ_2 dois caminhos acessíveis e arbitrários cujos comprimentos são $l(\phi_1)$ e $l(\phi_2)$ respectivamente, σ_1 um ponto (sequência) pertencente a ϕ_1 , e $h(\sigma_1, \phi_2)$ a distância de Hamming entre σ_1 e o ponto mais próximo pertencente ao caminho ϕ_2 , definimos então a divergência $d(\phi_1, \phi_2)$ entre o par de caminhos ϕ_1 e ϕ_2 :

$$d(\phi_1, \phi_2) = \frac{1}{l(\phi_1) + l(\phi_2)} \left(\sum_{\sigma_1 \in \phi_1} h(\sigma_1, \phi_2) + \sum_{\sigma_2 \in \phi_2} h(\sigma_2, \phi_1) \right). \quad (3.2)$$

Isto é, $d(\phi_1, \phi_2)$ é a média sobre as menores distâncias de Hamming entre cada ponto $\sigma_1 \in \phi_1$ e todos os pontos $\sigma_2 \in \phi_2$, e vice-versa (MANHART; MOROZOV, 2014). Claramente, a divergência assim definida é simétrica entre ϕ_1 e ϕ_2 , isto é, $d(\phi_1, \phi_2) = d(\phi_2, \phi_1)$. O procedimento de tomada das menores distâncias de Hamming é ilustrado na Figura 8.

Figura 8 – Ilustração do cálculo da divergência $d(\phi_1, \phi_2)$ entre um par de caminhos ϕ_1 e ϕ_2 . Para cada ponto $\sigma_1 \in \phi_1$ procura-se o ponto mais próximo pertencente a ϕ_2 e determina-se a distância $h(\sigma_1, \phi_2)$. O mesmo é feito sobre o caminho ϕ_2 e a divergência $d(\phi_1, \phi_2)$ é calculada como a média de $h(\sigma_1, \phi_2)$ e $h(\sigma_2, \phi_1)$ conjuntamente.



Fonte: LOBKOVSKY; KOONIN (2012)

A partir disso, dado um ensemble de caminhadas que compartilham os mesmos pontos inicial e final, e seja $p(\phi_i)$ a probabilidade de ocorrência do i -ésimo caminho, podemos então

definir a divergência média de caminhos:

$$\mathcal{D} = \sum_{\phi_1 \neq \phi_2} d(\phi_1, \phi_2) p(\phi_1) p(\phi_2). \quad (3.3)$$

Uma vez que a divergência de um caminho em relação a ele mesmo é necessariamente nula, a soma é feita sobre todos os pares de caminhos distintos pertencentes ao ensemble (LOBKOVSKY; WOLF; KOONIN, 2011). A divergência média representa o valor esperado da divergência entre dois caminhos quaisquer do ensemble tomados ao acaso. Trata-se de uma medida refinada de previsibilidade, ou melhor, imprevisibilidade (LOBKOVSKY; WOLF; KOONIN, 2011), avaliando não só a acessibilidade e distribuição dos caminhos, mas também sua proximidade genotípica (MANHART; MOROZOV, 2014). Nesse sentido, torna-se um critério fundamental para evitar conclusões equivocadas em situações onde a fração de trajetórias acessíveis é escassa, dado que tais caminhos podem estar fortemente espalhados sobre o relevo (LOBKOVSKY; WOLF; KOONIN, 2011).

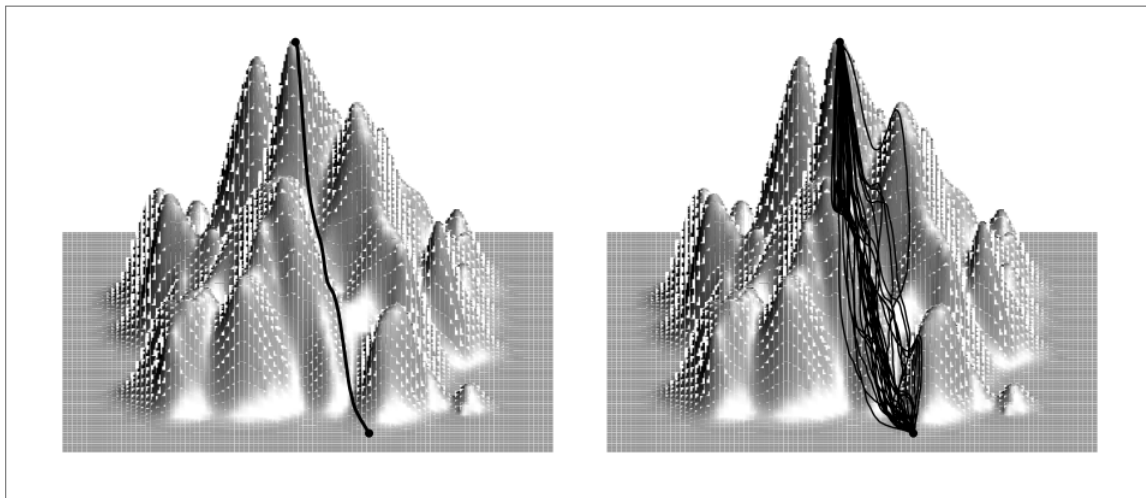
A divergência média de caminhos só possui significado quando aplicada a ensembles de trajetórias que compartilham os mesmos pontos inicial e final, de forma a medir o grau com que esses pontos determinam o trajeto adaptativo (LOBKOVSKY; KOONIN, 2012; LOBKOVSKY; WOLF; KOONIN, 2011). Uma medida de divergência mais abrangente e representativa do relevo pode ser obtida ainda calculando-se a divergência para cada par de pontos inicial e final possível no relevo, e tomando-se em seguida a média apenas sobre divergências médias cujas trajetórias possuem a mesma distância entre as sequências final e inicial (LOBKOVSKY; WOLF; KOONIN, 2011). Note que alguns pontos podem não sustentar o início ou fim de caminhadas e que tal possibilidade depende fortemente da dinâmica. No regime SSWM, por exemplo, máximos locais são pontos absolutamente estáveis e não há caminhadas partindo deles, enquanto que para populações finitas tal caminhada é possível, a princípio, já que tais picos são apenas localmente estáveis.

3.2 DINÂMICA ADAPTATIVA

A dinâmica evolutiva de uma população sujeita à seleção, mutação e deriva genética sobre um relevo de adaptação é, sem dúvida, um problema bastante complexo. A literatura voltada a este problema apresenta frequentemente dois importantes limites em termos do tamanho da população N , e da taxa de mutação por geração μ (Figura 9). Um deles ocorre quando o valor esperado de mutações por geração $N\mu$ é extremamente pequeno, de tal forma que a população deve consistir de um único genótipo durante a maior parte do processo. A população se comporta efetivamente como um simples caminhante adaptativo se movendo por passos unitários de mutação e guiado exclusivamente por gradientes positivos de fitness, destinado a parar definitivamente sobre o primeiro máximo local que alcançar. Dada a natureza aleatória das mutações (passos), trata-se de um processo completamente estocástico. O segundo limite

em questão representa uma situação fundamentalmente oposta, isto é, determinística, presente quando o tamanho da população tende ao infinito. Desta vez, o processo é descrito pelas equações de quasi-espécie 2.1, todos os genótipos estão presentes na população e, portanto, a estocasticidade fruto da seleção dos indivíduos a serem reproduzidos deixa de ser expressada (JAIN; KRUG, 2007).

Figura 9 – À esquerda, sob uma baixa taxa de suprimento de mutantes por geração ($N\mu \ll 1$) a população se comporta de maneira homogênea em praticamente todos os instantes, e mutações são fixadas sequencialmente como passos. À direita, com o aumento do número de mutações por geração diversos genótipos passam a coexistir, cada um seguindo sua própria trajetória, de forma que a evolução da população é melhor representada por um conjunto de caminhos que se espalha a partir do genótipo inicial e volta a convergir no máximo. A cada geração a estrutura da população pode ser descrita por uma distribuição de probabilidade no espaço de sequências.



Fonte: LOBKOVSKY; KOONIN (2012)

Nos dois limites apresentados a população ou está sujeita a seguir um caminho que é uma variável estocástica, cuja distribuição está completamente determinada pelo relevo, ou está fadada a seguir a mesma trajetória evolutiva em direção ao máximo global em qualquer realização do processo, desde que fixada a condição inicial. Surge imediatamente a seguinte pergunta: como se comportaria tal população para parâmetros entre esses dois extremos? A primeira afirmação concreta que se pode fazer sobre um cenário intermediário de populações finitas e número esperado de mutantes por geração $N\mu > 1$ é que, do ponto de vista da dinâmica de quasi-espécies o sistema não mais se comporta exatamente como as médias, e flutuações fruto de mutações raras devem ser levadas em consideração. Comparando com o extremo estocástico, por outro lado, fica claro que efeitos totalmente novos devem surgir do caráter coletivo (JAIN; KRUG, 2007). De fato, para valores finitos de população e número esperado de mutantes por geração, a análise é muito mais complexa, devido principalmente à interferência clonal. Muitos dos resultados analíticos se tornam impraticáveis, exceto através de aproximações ou simplificação de parâmetros, como tamanho do relevo. Dessa forma, nos propomos então a abordar a questão simulacionalmente (JAIN; KRUG; PARK, 2011).

Na verdade, a importância da interferência clonal em dinâmicas evolutivas não é absoluta,

mas irá ocupar um importante papel neste estudo por tratarmos unicamente de indivíduos assexuados, onde seus efeitos são pronunciados. Populações assexuadas se adaptam ao ambiente através da ocorrência e fixação de mutações benéficas. Privadas de recombinação, devem acumular estas mutações de maneira sequencial. Em geral, a fixação de uma mutação leva um longo período, que aumenta com o tamanho da população, de tal forma que ela é mantida em baixas frequências durante a maior parte do tempo, oferecendo assim a chance de surgimento de uma nova mutação benéfica na população ancestral. Enquanto em populações sexuadas duas mutações benéficas em diferentes linhagens estão sujeitas a recombinação, capazes de gerar um mutante duplo ainda mais apto, em populações assexuadas por outro lado, estas linhagens devem competir entre si. Esta competição pode interferir intensamente na fixação da primeira mutação, diminuindo sua velocidade de espalhamento, ou até mesmo eliminando-a (GERRISH; LENSKI, 1998).

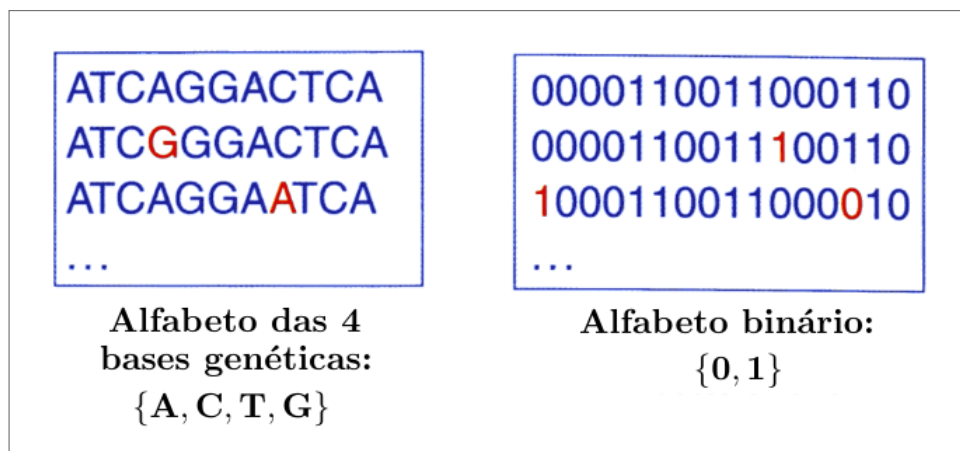
Antes de nos estendermos sobre as dinâmicas de adaptação é importante, naturalmente, esclarecer em que circunstâncias surge tal processo. O principal cenário inicial de um experimento evolutivo é, sem dúvida, a ocorrência de uma mudança ambiental. Ocorrendo isto, a população irá encontrar-se mal adaptada a esta nova condição ambiental e a subsequente adaptação requer o suprimento de mutações benéficas (JAIN; KRUG; PARK, 2011). O objetivo da adaptação é então levar a população ao novo genótipo mais apto, uma tarefa sujeita a muitas dificuldades, como a dependência em mutações imprevisíveis. Além disso, mesmo havendo o suprimento de mutações, estas são constantemente perdidas ao longo do processo de reprodução, inclusive as benéficas. Note que, caso as mudanças ambientais ocorressem numa escala de tempo mais rápida que a da evolução, estaríamos lidando com um relevo dinâmico, inconstante, situação que não exploramos aqui.

Em nosso estudo estaremos interessados unicamente em caminhadas que atinjam o máximo global. Além disso, utilizaremos constantemente um modelo de relevo aleatório, de tal forma que será regra a aplicação de médias sobre relevos. Afim de evitar ao máximo fontes de variação nas caminhadas, de relevo para relevo, buscamos uma condição inicial que segundo alguma característica conveniente seja minimamente ambígua com relação ao máximo global. A priori, duas características surgem como fortes candidatas, o fitness e a distância de Hamming. Neste ponto podemos nos basear na literatura, onde se observa que o fitness da sequência inicial tem muito menos influência nas propriedades de caminhadas evolutivas do que sua distância de Hamming à sequência final. Portanto, optamos por tomar uma distância fixa da sequência mais apta em todos os relevos. Salvo menção contrária, o genótipo escolhido é o antípoda, aquela sequência que difere do máximo global em todos os L loci e é, portanto, diametralmente oposta a esta (máxima distância de Hamming). Uma consequência imediata desta escolha é a observação das caminhadas mais longas ao máximo global, e o comprimento médio delas deve representar um limite superior (JAIN; KRUG, 2007; FRANKE et al., 2011).

Como já foi esclarecido, nosso foco é o estudo da previsibilidade evolutiva e sua relação com parâmetros populacionais e topográficos, fatores que não devem depender sensivelmente

de detalhes específicos do espaço de sequências. Dessa forma optamos por explorar um dos cenários mais simples, capaz de sustentar um processo adaptativo como o descrito até aqui. A dinâmica de adaptação será simulada sobre indivíduos assexuais e haploides (que possuem um único conjunto completo de cromossomos), cujo genótipo é representado por sequências binárias $\sigma \in \{0, 1\}^L$ de comprimento L , como mostrado na Figura 10 em comparação ao alfabeto das quatro bases genéticas $\{A, C, T, G\}$. Consequentemente, qualquer relevo construído sobre esse espaço terá a estrutura topológica de um hipercubo de dimensão L , ilustrado na Figura 11 para valores crescentes de comprimento das sequências. Este tipo de caricatura dos genomas se aproxima bem de uma série de situações, a exemplo de um único gene ou um genoma bialélico, onde as unidades básicas em consideração seriam loci e genes, respectivamente. Outras aplicações desta mesma modelagem do espaço genético são comuns e podem ser encontradas, por exemplo, nos trabalhos de JAIN; KRUG; JAIN; KRUG; PARK; FRANKE et al..

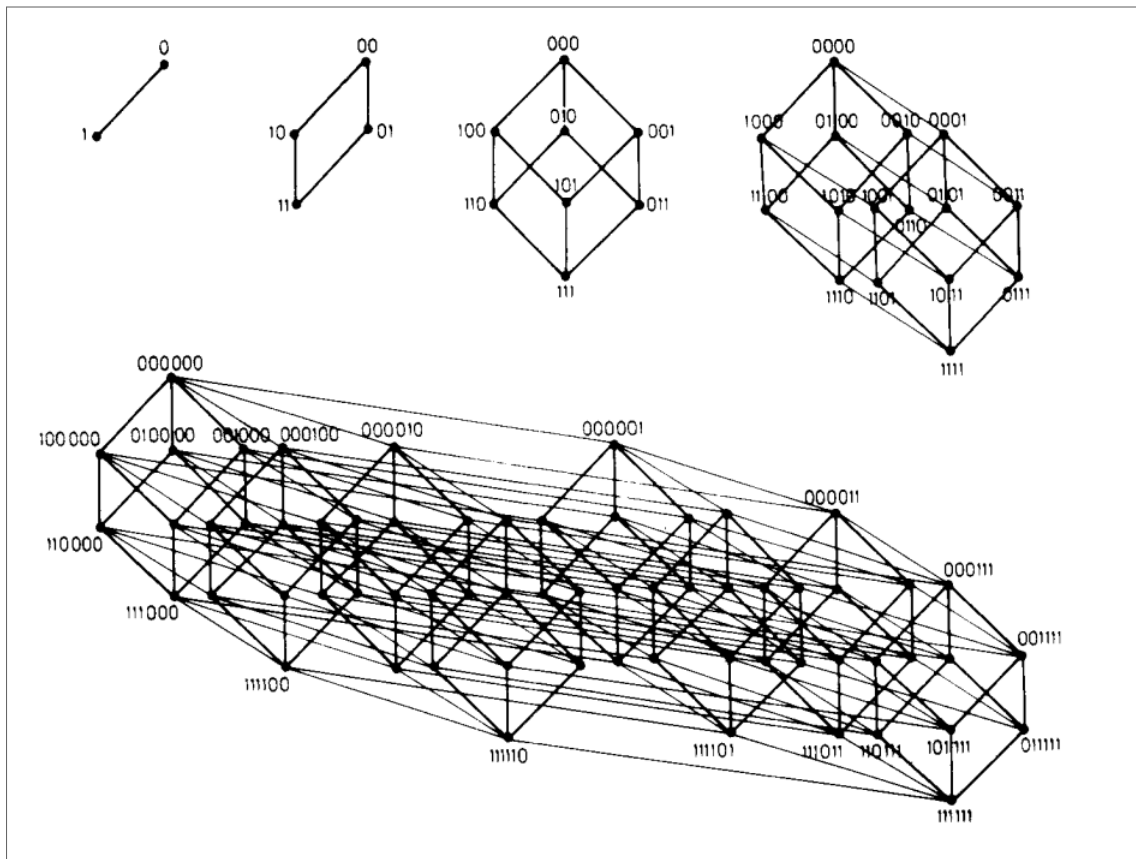
Figura 10 – No mundo natural toda a informação fenotípica é armazenada no DNA, escrita em sequências de um alfabeto de quatro bases genéticas $\{A, C, T, G\}$. Em experimentos microbiais, no entanto, é comum o uso de uma base binária $\{0, 1\}$ na análise do processo de adaptação, dado que na maioria das vezes a unidade de interesse são mutações específicas em cada gene, de um alelo para outro, situação melhor representada pela ausência (0) ou presença (1) da mutação. No alfabeto natural os genótipos tem três vizinhos para cada locus, isto é, acessíveis por uma única mutação (destacadas em vermelho), enquanto sob o alfabeto binário só há um.



Fonte: NOWAK (2006)

A partir daqui vamos descrever detalhadamente os dois tipos de dinâmicas adaptativas explorados nestes espaços, deixando as questões de atribuição do fitness para a próxima sessão. No momento, é suficiente saber que o fitness de cada sequência assume um valor real positivo no intervalo fechado $[0, 1]$. A primeira dinâmica apresentada é a caminhada adaptativa, efetuada por um único indivíduo representante da população inteira, situação associada ao regime de mutação fraca. Afim de explorar situações com taxas de mutação maiores, devido a seus efeitos na heterogeneidade da população durante a evolução, simulamos na segunda dinâmica a adaptação de uma população finita de tamanho N , seguindo a amostragem de Wright-Fisher, descrita em detalhes na Seção 3.2.2.

Figura 11 – Para seqüências binárias o espaço de seqüências de dimensão L é construído através da duplicação do espaço de dimensão $L - 1$, como ilustrado da esquerda para direita em ordem crescente de dimensionalidade.



Fonte: EIGEN; MCCASKILL; SCHUSTER (1989)

3.2.1 Caminhante adaptativo

Na situação limite de uma população cuja taxa $N\mu$ de novos mutantes por geração é absolutamente pequena, o tempo esperado de ocorrência de um mutante, da ordem de $(N\mu)^{-1}$, é tão alto que não há interferência clonal, e o genótipo dominante sempre alcança fração unitária na população (JAIN; KRUG, 2007). Portanto, quando as mutações são raras ($N\mu \ll 1$) a população se comporta praticamente como um único genótipo a qualquer instante do processo, de tal forma que, como uma boa aproximação, apenas as seqüências que estão a um passo mutacional da seqüência focal são alcançáveis. Diz-se que a adaptação tem um horizonte curto. Dada uma seqüência, se qualquer passo unitário a partir desta é deletério trata-se então de um máximo local, e a existência de um mutante mais apto a dois passos deste é irrelevante. Embora cada um dos primeiros vizinhos mais aptos seja potencialmente acessível, apenas um irá se fixar, configurando um passo. O genótipo alcançado se comporta como a nova seqüência focal, e o processo se repete até que se atinja um seqüência desprovida de vizinhos mais aptos, um máximo local. Em poucas palavras, a evolução por seleção natural sempre corre em direção a configurações mais aptas, enquanto houver uma disponível (ORR,

2002).

De fato, neste cenário de mutação fraca o processo pode ser comparado de certa maneira com uma cadeia de Markov, como argumenta GILLESPIE, no sentido de que a longo prazo podemos rastrear a adaptação em termos de eventos de substituição, sem fazer considerações a respeito da estrutura interna da população durante a fixação de uma mutação. O processo de fixação de um alelo mutante é efetivamente instantâneo comparado com a escala de tempo da evolução, caracterizada pelo tempo de escape de um alelo da chamada “camada limite”, associada à aparição deste alelo e sua subsequente sobrevivência à deriva genética. É bem verdade que além disto há poucas semelhanças com o modelo de GILLESPIE, uma vez que este trata de um espaço de alelos mutuamente acessíveis (ORR, 2002).

De acordo com essas considerações, ao utilizar um caminhante adaptativo em lugar de uma população finita estaremos abordando indiretamente o âmbito das mutações fracas. Embora seja importante esclarecer essa motivação vale lembrar também que o caminhante adaptativo é notório de maneira geral como modelo de estudo de processos de otimização e, no contexto do espaço de sequências especificamente, da otimização combinatorial. Neste regime um caminho é considerado seletivamente acessível se os valores de fitness ao longo dessa trajetória são monotonicamente crescentes em fitness. Em outras palavras, genótipos separados por vales de fitness são inacessíveis entre si. Já dentro do conjunto dos caminhos acessíveis é possível distinguir entre caminhos diretos e indiretos, onde os primeiros representam as caminhadas mais curtas entre duas sequências, e o segundo caso representa as demais. Seja h a distância de Hamming entre duas sequências, o número de caminhos (monotônicos ou não em fitness) entre elas é simplesmente $h!$. É importante notar que para um caminhante adaptativo cada mutação a mais do que o mínimo necessário h representa uma forte diminuição na probabilidade de ocorrência de uma trajetória, da ordem da taxa de mutação, de forma que é esperada pouca influência dos caminhos indiretos (FRANKE et al., 2011).

Enquanto populações finitas percorrem caminhos cujas probabilidades dependem de uma maneira não trivial com a taxa $N\mu$ de suprimento de mutantes por geração, caminhantes adaptativos realizam trajetos cujas probabilidades dependem unicamente das probabilidades individuais de cada passo, que são, por sua vez, determinadas somente pelo relevo e modelo de seleção escolhido (FRANKE et al., 2011). Gillespie se referiu a este cenário como o já mencionado SSWM (seleção forte e mutação fraca). Em seu trabalho ele define o que chama de seleção forte através de um coeficiente de seleção s , onde $Ns > 1$. Ele também esclarece que tal seleção só é considerada forte comparativamente ao tamanho da população, mas não em termos absolutos do parâmetro s , diretamente relacionado às diferenças de fitness no relevo (GILLESPIE, 1983; GILLESPIE, 1984; ORR, 2002). Aqui não utilizaremos um parâmetro de regulação da intensidade de seleção. Aplicaremos a distinção entre seleção forte e fraca de maneira dicotômica diretamente nas probabilidades da caminhada, de forma que teremos dois tipos de caminhada. Primeiro, sob seleção fraca, a hipótese é de que qualquer mutação benéfica é igualmente provável. Considere o caminhante situado no genótipo $\sigma^{(0)}$ cercado por L vizinhos

(de acordo com nosso modelo de genoma binário), dos quais apenas k são mais aptos. A probabilidade $p = p(\sigma^{(0)}, \sigma^{(i)})$ de que o caminhante realize um passo em direção ao vizinho $\sigma^{(i)}$, que difere de $\sigma^{(0)}$ apenas pelo locus i , é dada simplesmente por seu valor neutro (JAIN; KRUG; PARK, 2011)

$$p(\sigma^{(0)}, \sigma^{(i)}) = \begin{cases} 0, & \text{se } F(\sigma^{(i)}) < F(\sigma^{(0)}) \\ 1/k, & \text{se } F(\sigma^{(i)}) > F(\sigma^{(0)}) \end{cases} \quad i = 1, 2, \dots, L, \quad (3.4)$$

onde F representa o fitness. Claramente, o inteiro $k = k(\sigma^{(0)})$ depende da vizinhança da sequência inicial $\sigma^{(0)}$ e deve satisfazer a equação $k = \sum_{F(\sigma^{(i)}) > F(\sigma^{(0)})} 1$, onde a soma é feita sobre o conjunto de vizinhos mais aptos. Note que, embora o fitness seja o critério de distinção entre mutações benéficas e deletérias, o caminhante sob seleção fraca não “enxerga” seu valor absoluto. Para uma seleção forte, no entanto, vamos supor que quanto maior o fitness de um determinado vizinho maior sua probabilidade de escapar à eliminação por deriva genética e, portanto, maior sua probabilidade de fixação. Neste caso postularemos então que a probabilidade de realização de um passo benéfico é diretamente proporcional à diferença de fitness entre as sequências final e inicial. Seguindo as mesmas convenções do caso anterior, e definindo $\Delta F(\sigma^{(0)}, \sigma^{(i)}) = F(\sigma^{(i)}) - F(\sigma^{(0)})$ como a diferença de fitness entre o vizinho no i -ésimo locus e $\sigma^{(0)}$, então a probabilidade $p(\sigma^{(0)}, \sigma^{(i)})$ sob o regime de seleção forte será

$$p(\sigma^{(0)}, \sigma^{(i)}) = \begin{cases} 0, & \text{se } F(\sigma^{(i)}) < F(\sigma^{(0)}) \\ \Delta F(\sigma^{(0)}, \sigma^{(i)}) / \sum_j \Delta F(\sigma^{(0)}, \sigma^{(j)}), & \text{se } F(\sigma^{(i)}) > F(\sigma^{(0)}) \end{cases} \quad i = 1, 2, \dots, L, \quad (3.5)$$

onde, novamente, a soma em j é feita sobre o conjunto de vizinhos mais aptos.

3.2.2 População finita

No intuito de abordar explicitamente os efeitos da abundância de mutantes, e em diferentes graus, simularemos uma população finita para diferentes tamanhos N e taxas de mutação μ tal que $N\mu > 1$. Em nosso modelo de dinâmica populacional a reprodução ocorre em passos discretos de tempo, não havendo convivência de indivíduos de gerações diferentes, e o tamanho da população permanece constante. A estrutura da população em uma determinada geração depende somente da geração imediatamente anterior, de acordo com a amostragem padrão de Wright-Fisher. Dada a frequência absoluta $X(\sigma, t)$ de cada genótipo σ na geração t , onde $t \in \mathbb{N}$, cada um dos N descendentes “escolhe” seu ancestral com uma probabilidade que deve ser proporcional tanto à frequência $X(\sigma, t)$ de cada genótipo na geração anterior quanto ao valor adaptativo $F(\sigma)$ do mesmo. Inserido dessa maneira, o fitness representa o valor esperado de descendentes de um único indivíduo de determinado genótipo, conhecido como fitness Malthusiano. Seguindo este algoritmo, o conjunto de N genótipos presentes na geração

seguinte é uma variável estocástica de distribuição multinomial, cujas probabilidades pra cada genótipo são

$$p(\sigma^i) = \frac{X(\sigma^i, t)F(\sigma^i)}{\sum_{j=1}^{2^L} X(\sigma^j, t)F(\sigma^j)} \quad i = 1, 2, 3, \dots, 2^L, \quad (3.6)$$

onde o índice i (sem parênteses para diferenciar do contexto de 3.4 e 3.5) varre agora todas as 2^L sequências possíveis, podendo inclusive ser associado diretamente à representação decimal i de cada sequência binária (SZENDRO et al., 2013a; JAIN; KRUG; PARK, 2011). Escolhidos os ancestrais, no entanto, entra ainda em cena a chance de mutação, a ser incluída na Equação 3.6. Vamos nos limitar à possibilidade de mutações em no máximo um único locus por indivíduo, descartando assim a chance de mutações múltiplas. Cada individuo reproduzido está sujeito a esta mutação com uma probabilidade dada por μ , a taxa de mutação por indivíduo por geração, e cada locus tem probabilidade igual de ser acometido. A probabilidade $p(\sigma^i)$ deve depender agora não só de sua própria frequência, mas também da frequência de todos os seus L primeiros vizinhos. Seja $h(\sigma^i, \sigma^j)$ a distância de Hamming entre os genótipos i e j , então de acordo com nosso modelo a chance de mutação da sequência i para a sequência j , denotada por $w(\sigma^i, \sigma^j)$, deve ser

$$w(\sigma^i, \sigma^j) = \begin{cases} 0, & \text{se } h(\sigma^i, \sigma^j) > 1 \\ \mu/L & \text{se } h(\sigma^i, \sigma^j) = 1 \\ 1 - \mu & \text{se } h(\sigma^i, \sigma^j) = 0 \end{cases} \quad i, j = 1, 2, \dots, L, \quad (3.7)$$

Portanto, a probabilidade de $p(\sigma^i)$ de que um passo de reprodução arbitrário gere um descendente do genótipo σ^i é

$$p(\sigma^i) = \frac{\sum_{j=1}^{2^L} w(\sigma^j, \sigma^i) X(\sigma^j, t) F(\sigma^j)}{\sum_{k=1}^{2^L} X(\sigma^k, t) F(\sigma^k)} \quad i = 1, 2, 3, \dots, 2^L, \quad (3.8)$$

onde a soma no numerador também é feita sobre todas as 2^L , mas apenas σ^i e seus L primeiros vizinhos tem contribuição efetiva (não nula). A cada geração a população e as probabilidades da Equação 3.8 são redefinidas, determinando a nova distribuição multinomial para próxima descendência. Tal processo deve ser iterado até que surja o primeiro indivíduo cujo genótipo é o máximo global do relevo. Na verdade, sabendo que a caminhada do antípoda ao máximo global pode ser muito custosa em gerações e tempo computacional, principalmente em relevos com rugosidade moderada e alta, definimos um tempo limite de 10^9 gerações para que estas caminhadas terminem. As dinâmicas que não encontraram o máximo global neste tempo foram então descartadas.

3.3 MODELO NK

O Modelo NK é um modelo de relevo de adaptação proposto por Stuart Kauffman. Em sua publicação original (KAUFFMAN; WEINBERGER, 1989), Kauffman argumenta que o modelo captura características importantes do fenômeno de maturação da resposta imune, o processo de adaptação de anticorpos em direção a uma resposta cada vez mais eficiente a um determinado antígeno, eficiência esta medida pela afinidade do anticorpo com o agente invasor. Na verdade, o trabalho se estende para além de uma análise específica desse fenômeno e propõe uma abordagem consistente da evolução proteica de maneira geral. A propriedade que torna o modelo interessante é a capacidade de regular a rugosidade do relevo, num espectro que vai de uma topografia do tipo “Monte Fuji”, com um único pico, até relevos altamente acidentados. Esta tunabilidade é feita através do parâmetro K , que determina com quantos outros genes um determinado gene interage, de maneira que a contribuição individual de cada um para o fitness depende de K outros. No contexto de genes e alelos, K mede assim a intensidade da epistasia. O outro parâmetro fundamental é o número de loci das sequências, originalmente denotado por N , justificando o nome do modelo. No entanto, utilizaremos L em seu lugar, mantendo nossa convenção para o comprimento de sequências, enquanto N continuará reservado ao tamanho da população (KAUFFMAN; WEINBERGER, 1989).

Inicialmente proposto para o estudo da evolução de genomas, ao longo do tempo o modelo foi aplicado a diversas áreas, desde física e biologia teórica à economia, onde quer que fosse possível modelar sistemas complexos dotados de partes interagentes, sempre visando a maximização de alguma medida de performance (KAUL; JACOBSON, 2006). Este modelo é fortemente inspirado em problemas do tipo vidro de spins e se baseia num paradigma comum a uma série de sistemas como este, isto é, que envolvem processos de otimização combinatorial: sistemas constituídos de muitas partes, quantificadas por L , onde cada parte pode assumir um dentre A estados possíveis, cuja contribuição para uma determinada medida global de interesse depende em média de K outras partes. Ao tratar de genótipos, interpreta-se L como o número de genes, A como o número de alelos possíveis àquele gene, e K o número de relações epistáticas entre alelos. Para proteínas, por outro lado, tem-se uma sequência de L aminoácidos, A igual a 20 de acordo com o “alfabeto” de aminoácidos naturais, e K novamente um parâmetro associado à dependência entre aminoácidos na determinação de propriedades funcionais da proteína, em função das interações químicas subjacentes. Embora, de acordo com o que já foi dito, não haja motivo aparente pra considerar K tendo um valor fixo, sendo provavelmente diferente de sítio pra sítio num cenário geral, ele é de fato tomado como constante no modelo, comum a todos os loci (KAUFFMAN; WEINBERGER, 1989). Note que os termos “sistema”, “partes”, e “estados” configuram definições fundamentais do modelo, e é imprescindível identificar precisamente esses elementos qualquer que seja a aplicação de interesse. No primeiro exemplo temos o genoma na posição de sistema, cujos componentes são os genes, e o conjunto de alelos de um dado gene representa seus possíveis estados (KAUL; JACOBSON, 2006). Fixados

os valores L , K e A , é necessário também especificar em seguida quais sítios afetam quais. Nesse ponto existem várias possibilidades e podemos citar as mais importantes como sendo a escolha (i) de vizinhos mais próximos, (ii) totalmente aleatória, (iii) ou baseada em alguma distribuição de probabilidade. Seguiremos com o segundo caso, onde genes interagentes são escolhidos ao acaso. Além disso, como temos utilizado sequências binárias, tem-se $A = 2$ (KAUFFMAN; WEINBERGER, 1989).

Esclarecida a ideia geral do modelo, podemos estabelecê-lo formalmente. Já adaptando ao nosso contexto, os sistemas serão genomas e a medida de performance o fitness. Considere $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_L)$ uma sequência binária de comprimento L , isto é, $\sigma_i \in \{0, 1\}$ onde o índice inferior denota o locus. Nosso sistema pode assumir então 2^L estados, representados por um vetor $\sigma \in \{0, 1\}^L$. Então o fitness $F(\sigma)$ da sequência σ é simplesmente a média

$$F = \frac{1}{N} \sum_{i=1}^L f_i \quad (3.9)$$

das contribuições individuais f_i de cada locus (REIA; CAMPOS, 2020; CAMPOS; ADAMI; WILKE, 2002). Ocorre que no Modelo NK o valor de cada parcela f_i é função dos estados do sítio i e do conjunto $\Pi(\sigma_i)$ de K outros loci escolhidos aleatoriamente dentre os $L - 1$ restantes ($0 \leq K \leq L - 1$), isto é $f_i = f_i(\sigma_i, \Pi(\sigma_i))$, de tal forma que cada f_i constitui uma tabela de 2^{K+1} valores (REIA; CAMPOS, 2020).

Por fim, resta saber como gerar as tabelas f_i , isto é, como atribuir as possíveis contribuições individuais de cada sítio e estado. Neste ponto surge uma hipótese central do modelo (KAUFFMAN; WEINBERGER, 1989). Em geral, computar medidas de performance em sistemas reais é uma tarefa desafiadora, marcada pela dificuldade de se definir precisamente tal medida assim como quantificá-la no contexto das complexas interações entre componentes. Portanto, em situações em que não é possível definir uma medida de performance deterministicamente, isto é, em primeiros princípios, a abordagem universal é de atribuir as contribuições individuais f_i a partir de uma distribuição de probabilidade (KAUL; JACOBSON, 2006). Espera-se assim que a presença de interações complexas seja suficiente para que a atribuição aleatória dos valores f_i capture as características estatísticas fundamentais do relevo (KAUFFMAN; WEINBERGER, 1989). Utiliza-se então a distribuição uniforme no intervalo $(0,1)$, e como consequência da Equação 3.9 o fitness das sequências também está limitado ao mesmo intervalo. Fixados todos esses parâmetros, fica automaticamente determinado um ensemble de relevos de adaptação, associado a todas as configurações possíveis dos conjuntos $\Pi(\sigma_i)$ e das tabelas f_i , sendo o primeiro um conjunto discreto e o segundo, contínuo (REIA; CAMPOS, 2020). Definido desta maneira, onde tanto f_i como $\Pi(\sigma_i)$ são determinados aleatoriamente, estamos lidando com a versão totalmente aleatória do Modelo NK (CAMPOS; ADAMI; WILKE, 2002). Note que, do Teorema Central do Limite, a distribuição dos valores de fitness F será uma gaussiana, e isto ocorrerá independentemente da distribuição escolhida para $\{f_i\}$.

A dependência (ajustabilidade) da rugosidade do relevo com o parâmetro K é um importante aspecto do modelo, cuja ilustração se torna clara ao analisarmos os dois extremos:

nenhuma interação ($K = 0$) e interação máxima ($K = L - 1$). Para $K = 0$, todos os sítios são independentes e não há ambiguidade, no sentido de que apenas um dos valores binários 0 ou 1 é o melhor em cada sítio, não importando o estado dos demais. Seguindo esse raciocínio, o máximo global é obtido tomando-se o melhor estado em cada sítio individualmente. Qualquer sequência que difere do máximo global possui pelo menos um sítio que não assumiu seu melhor estado e tem, portanto, pelo menos um vizinho mais apto. Logo, fica claro que o máximo global é na verdade o único máximo do relevo. Por este mesmo argumento nota-se que qualquer sequência subótima está conectada a este máximo por um conjunto de mutações unitárias ascendentes em fitness. O comprimento de uma caminhada adaptativa será simplesmente a distância Hamming do genótipo inicial ao máximo, e as mutações envolvidas podem ser realizadas em qualquer ordem. Quando $K = L - 1$, tem-se uma situação absolutamente oposta. A contribuição em fitness de cada locus depende de todos os demais, de tal forma que a mudança de estado de qualquer um altera a contribuição de todos para um novo valor completamente aleatório. O relevo associado é rico em máximos locais, com um número esperado de $\frac{2^L}{L+1}$ por relevo, e as caminhadas tem um comprimento da ordem de $\ln L$. Qualquer que seja a condição inicial, apenas uma pequena fração dos máximos são acessíveis. Pelo mesmo argumento, segue que o máximo global é dificilmente alcançável. Entre os dois limites o que fica claro é que o crescimento de K cria cada vez mais vínculos, e com eles mais conflitos de otimização dos sítios individualmente, fenômeno conhecido como frustração em modelos de sistemas de spins (KAUFFMAN; WEINBERGER, 1989).

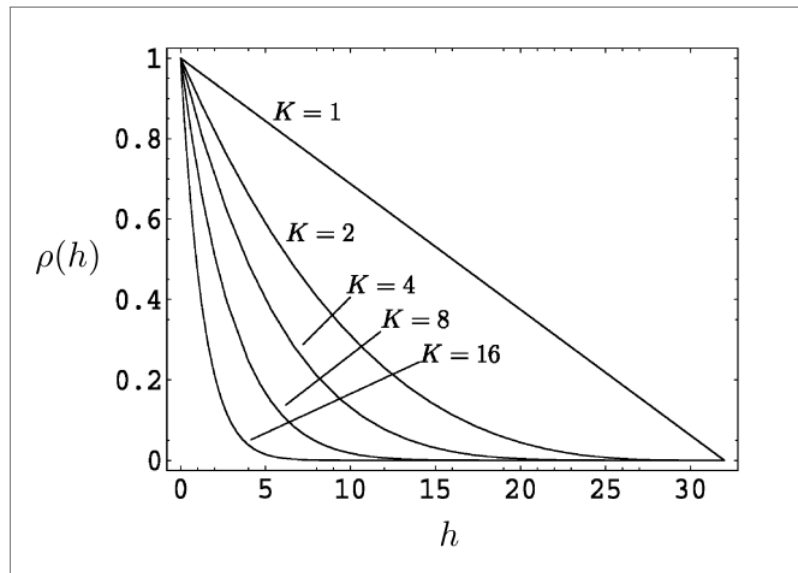
Uma importante característica dos relevos de adaptação é sua correlação, ou melhor, auto-correlação. Nesse contexto correlação se refere ao grau de similaridade do fitness de sequências vizinhas, e está intimamente relacionada com o parâmetro K . No caso $K = 0$, por exemplo, a mudança de estado de um único sítio provoca no máximo uma diferença de $1/L$ no fitness, apresentando assim alta correlação. Já no limite $K = L - 1$, tem-se um relevo totalmente aleatório e, portanto, correlação nula (KAUFFMAN; WEINBERGER, 1989; REIA; CAMPOS, 2020).

É possível definir uma medida de auto-correlação de várias formas. Uma particularmente simples, $\rho(d)$, é definida em função da distância de Hamming h , e segue da ideia de que em média a similaridade entre duas sequências quaisquer do relevo pode ser quantificada pela probabilidade de que a contribuição de fitness de um determinado gene permaneça inalterada após as h mutações que levam de uma sequência à outra. Dado que essa contribuição depende de K outros genes, então procuramos a probabilidade de que as h mutações ocorram nos $L - K - 1$ genes que não interagem com o gene em questão e que excluem também o próprio gene. Trata-se basicamente de um problema de contagem sem reposição. Para $h = 1$, a probabilidade é simplesmente $(L - K - 1)/(L)$. Para $h = 2$, tem-se $(L - K - 1)(L - K - 2)/[L(L - 1)]$, já que a segunda mutação deve ocorrer em um sítio diferente da primeira. Por indução, a auto-correlação $\rho(h)$ como função de h deve ser:

$$\rho(h) = \frac{(L - K - 1)!(L - h)!}{L!(L - K - h - 1)!}. \quad (3.10)$$

Note que para $h \geq L - K - 1$ não é possível evitar a mudança de valor de f_i , e $\rho(h)$ deve ser nulo. A auto-correlação é um importante parâmetro de rugosidade do relevo. Sua relação com K é apresentada na Figura 12. Para um decaimento lento de $\rho(h)$, observado para valores baixos de K , a informação é preservada por longas distâncias, o que implica num relevo suave, isto é, sem mudanças bruscas na topografia. Um decaimento agudo como o observado para valores altos de K , por outro lado, revela um relevo acidentado, uma vez que sequências próximas carregam pouquíssima informação umas sobre as outras (CAMPOS; ADAMI; WILKE, 2002).

Figura 12 – Função de auto-correlação no Modelo NK para $L = 32$. A função $\rho(h)$ no gráfico difere sutilmente da Equação 3.10 e é dada na forma $(L-K)!(L-h)!/[L!(L-K-h)]$, onde aparentemente o autor da publicação original desconsidera que uma mutação no próprio gene i também altera o valor de f_i . O comportamento de ambas, no entanto, é basicamente o mesmo, onde valores baixos de K correspondem a decaimentos lentos de $\rho(h)$, enquanto valores altos apresentam decaimentos acentuados, culminando no caso extremo $K = L - 1$, onde $\rho(h)$ vai a zero já para $h = 1$ (não ilustrado).



Fonte: CAMPOS; ADAMI; WILKE (2002)

A família NK de relevos é notável, portanto, pela possibilidade de explorar as implicações do grau de epistasia na topografia e, conseqüentemente, na dinâmica adaptativa (KAUFFMAN; WEINBERGER, 1989). De acordo com tudo que foi dito fica claro que trata-se de um modelo simples de descrever, dotado porém de grande riqueza. Se por um lado pode-se considerar o Modelo NK como não sendo realístico o suficiente, por outro não há dúvida de que representa um importante passo inicial que não deve ser ignorado antes de qualquer rebuscamento, e espera-se naturalmente que aspectos gerais deste modelo representem em algum grau o mundo real (REIA; CAMPOS, 2020).

4 RESULTADOS

O objetivo deste trabalho é o estudo da previsibilidade evolutiva no contexto do importante conceito do relevo de adaptação, explorado simulacionalmente através do Modelo NK. Nossa análise se concentra nos dois principais fatores influentes na previsibilidade: topografia do relevo e dinâmica populacional. Na primeira seção exploramos a topografia, mais precisamente a relação entre os parâmetros L e K , que determinam o nível de rugosidade, e medidas de acessibilidade e previsibilidade. Neste cenário nos restringimos à aplicação de caminhadas adaptativas, uma vez que estas não envolvem os parâmetros de tamanho de população e taxa de mutação, favorecendo assim uma análise focada apenas nas características do relevo. A segunda seção, por outro lado, explora justamente essas duas variáveis, assim como a relação da previsibilidade com os regimes de produção de mutantes. Aqui as caminhadas também são realizadas sobre relevos de adaptação do Modelo NK. Uma terceira e última seção avalia boa parte das quantidades anteriores porém no âmbito de um relevo experimental de dimensão 8, do fungo *Aspergillus niger*, permitindo assim a comparação entre um modelo teórico e um relevo experimental.

4.1 CAMINHADA ADAPTATIVA E RUGOSIDADE

As caminhadas adaptativas, associadas ao regime de baixa mutação, são processos consideravelmente mais simples quando comparados à complexidade, fruto de efeitos estocásticos, de dinâmicas de população finita. Portanto, são especialmente úteis ao explorar características inerentes ao relevo de fitness. Para este fim dispomos de dois tipos de caminhadas, a aleatória e a probabilística, associadas às Equações 3.4 e 3.5, respectivamente. Na primeira parte desse estudo exploramos o comportamento de diversas medidas de acessibilidade e repetibilidade de trajetórias como função dos parâmetros L e K . O comportamento geral em termos de L e K não depende do tipo de caminhada, e optamos então por utilizar apenas a caminhada aleatória. Num segundo momento analisamos relevos individualmente, através de medidas equivalentes porém aplicadas por máximo local, isto é, separadamente sobre ensembles de caminhadas que alcançam um mesmo máximo, prestando especial atenção na relação destas medidas com a distância do pico ao antípoda, nosso ponto de partida. Nessa subseção são utilizados e comparados os dois tipos de caminhadas.

4.1.1 Valores médios em função de L e K

Para decifrar o comportamento da rugosidade com L e K , assim como sua influência na previsibilidade evolutiva, foram simuladas por relevo 10.000 caminhadas aleatórias armazenando-se cada uma das seguintes medidas:

- Quantidade de máximos locais
- Acessibilidade do máximo global
- Fração de máximos locais acessíveis
- Quantidade de caminhos monotonicamente crescentes
- *Predictability* com relação aos pontos finais
- *Predictability* com relação aos caminhos evolutivos
- Comprimento médio de caminhos

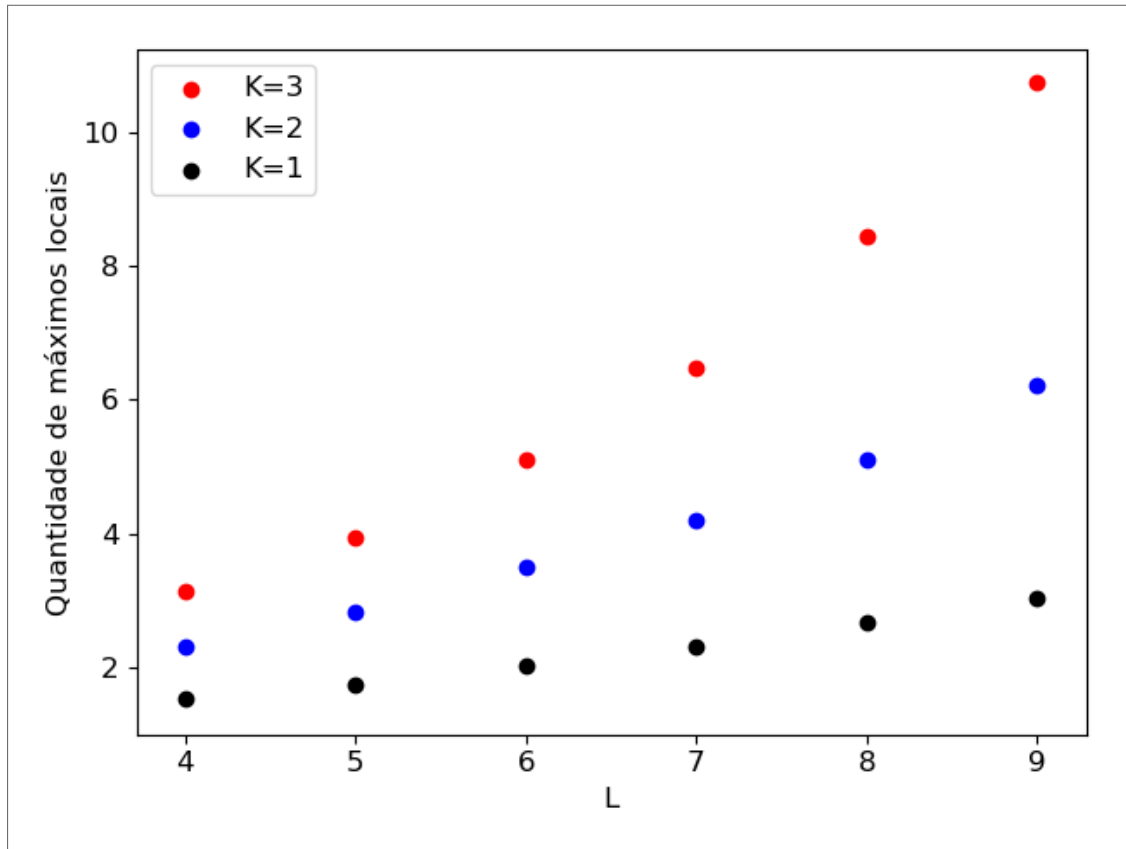
Levando em consideração a variância entre realizações do modelo, repetiu-se então o processo sobre 1.000 relevos distintos para cada valor de L e K , fornecendo assim uma média sobre relevos. O fato de caminhadas adaptativas serem direcionadas a máximos locais diversos tem duas consequências imediatas a serem ressaltadas. Primeiro, devem ser descartados aqueles relevos cujo antípoda é um máximo local, uma vez que o caminhante já se encontra inicialmente estável e não há dinâmica. Segundo, não é possível utilizar a divergência média de caminhos (Equação 3.3) nesta análise, pois só é aplicável a ensembles de trajetórias que compartilham os mesmos pontos inicial e final, limitação que não ocorre para as demais quantidades listadas acima.

No topo da lista está a quantidade de máximos locais, um indicador direto do nível de rugosidade, e trata-se da única medida acima intrínseca do relevo, isto é, independente do tipo de caminhada ou da condição inicial. É uma medida bastante simples e em análises mais profundas da topografia costuma ser apenas complementar a um conjunto de outras medidas mais elaboradas, a exemplo do desvio de aditividade apresentado por AITA; IWAKURA; HUSIMI. Por outro lado, é uma medida suficientemente capaz de evidenciar a relação básica da rugosidade com L e K , assim como os aspectos mais notáveis da relação com a previsibilidade, de maneira que outras medidas não se fazem necessárias nesse contexto.

Na Figura 13 exibimos a quantidade média de picos como função de L para valores de K de 1 a 3. Como foi mencionado na Seção 3.3, num relevo totalmente descorrelacionado, quando $K = L - 1$, o valor esperado do número de picos é exatamente $\frac{2^L}{L+1}$ (KAUFFMAN; WEINBERGER, 1989). Na Figura 13 apenas o ponto ($L = 4, K = 3$) representa um conjunto de relevos totalmente aleatórios, para os quais a aplicação da fórmula acima fornece um valor de $\frac{16}{5} = 3,2$, em contraste com o valor de aproximadamente 3,13 obtido nas simulações, um erro de uma parte em 50.

De acordo com o mesmo gráfico a quantidade de máximos locais é monotonicamente crescente tanto em L quanto em K . Embora o número $\frac{2^L}{L+1}$ seja válido apenas para o caso $K = L - 1$, o comportamento do número de picos para qualquer $K \neq 0$ segue basicamente a mesma relação. A probabilidade de ocorrência de um pico aumenta com o número de sequências no relevo e diminui com o número de vizinhos. Ocorre que, enquanto o número de sequências

Figura 13 – Quantidade de máximos locais em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.



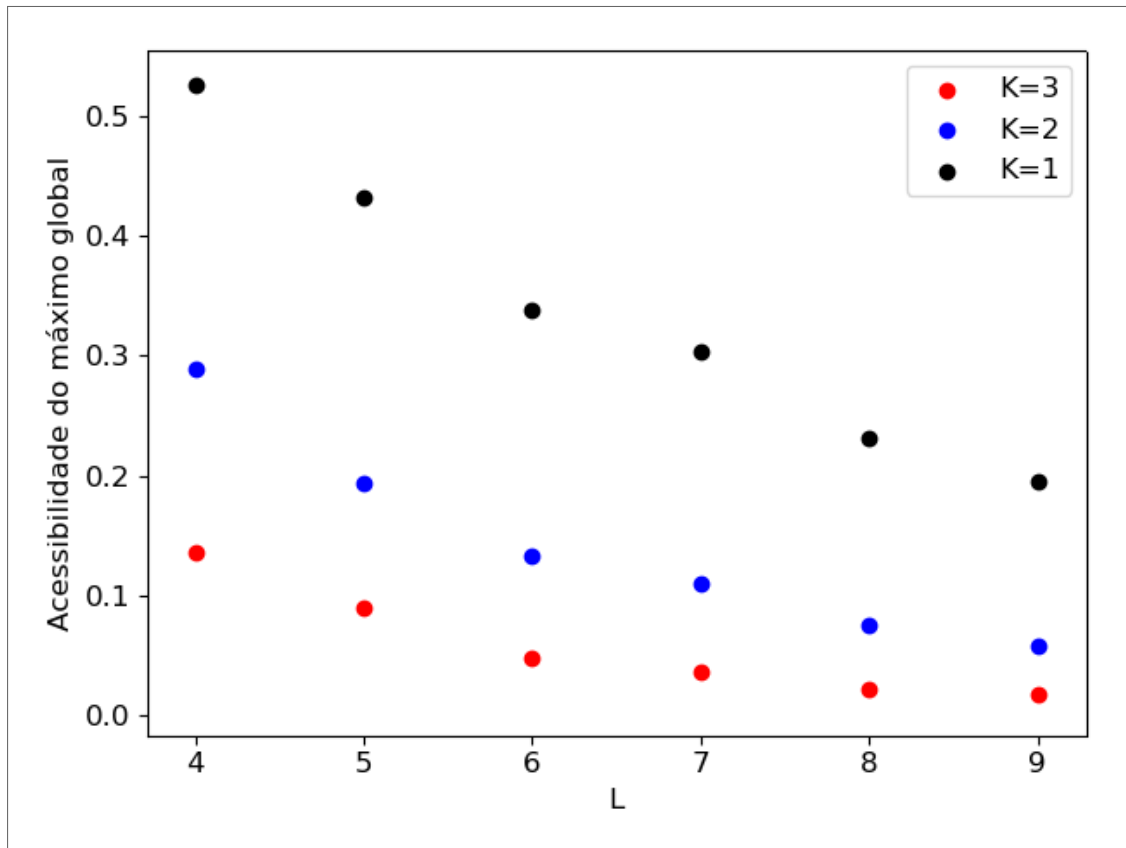
Fonte: Elaborado pelo autor (2020)

crece exponencialmente com L , o número de vizinhos cresce apenas linearmente. Pelo mesmo argumento, fica claro também que a razão do número de picos pelo número total de sequências é da ordem de $\frac{1}{L}$ e, portanto, em termos de densidade a abundância de picos diminui com L , ao contrário de seu valor absoluto. Seguindo a análise feita na Seção 3.3, a relação da quantidade de picos com K se explica pelo fenômeno da frustração. Para $K = 0$, cada sítio tem um único valor ótimo e só existe uma sequência onde este estado ocorre simultaneamente para todos, o máximo global. Todas as outras sequências possuem pelo menos um locus não otimizado e assim pelo menos um vizinho mais apto. À medida que K cresce, cada sítio depende cada vez mais dos outros. Aumenta então a probabilidade de que uma única mutação provoque a otimização simultânea de vários sítios e com isso sequências cada vez mais distantes do máximo global ganham a chance de se tornarem um máximo local. Por fim, o efeito de L é amplificado com K , e vice-versa, simplesmente pela maior disponibilidade de sequências.

A segunda medida da lista é a acessibilidade do máximo global, isto é, a fração do número de caminhadas em que este máximo é alcançado dentre todas. Note que ao utilizar o antípoda como condição inicial as trajetórias partem não só de uma distância de Hamming fixa do máximo global em todos os relevos, mas da maior distância possível. Estes dois aspectos provocam uma maior robustez da medida e a obtenção de um limite inferior, respectivamente.

O resultados são mostrados na Figura 14.

Figura 14 – Acessibilidade do máximo global em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.



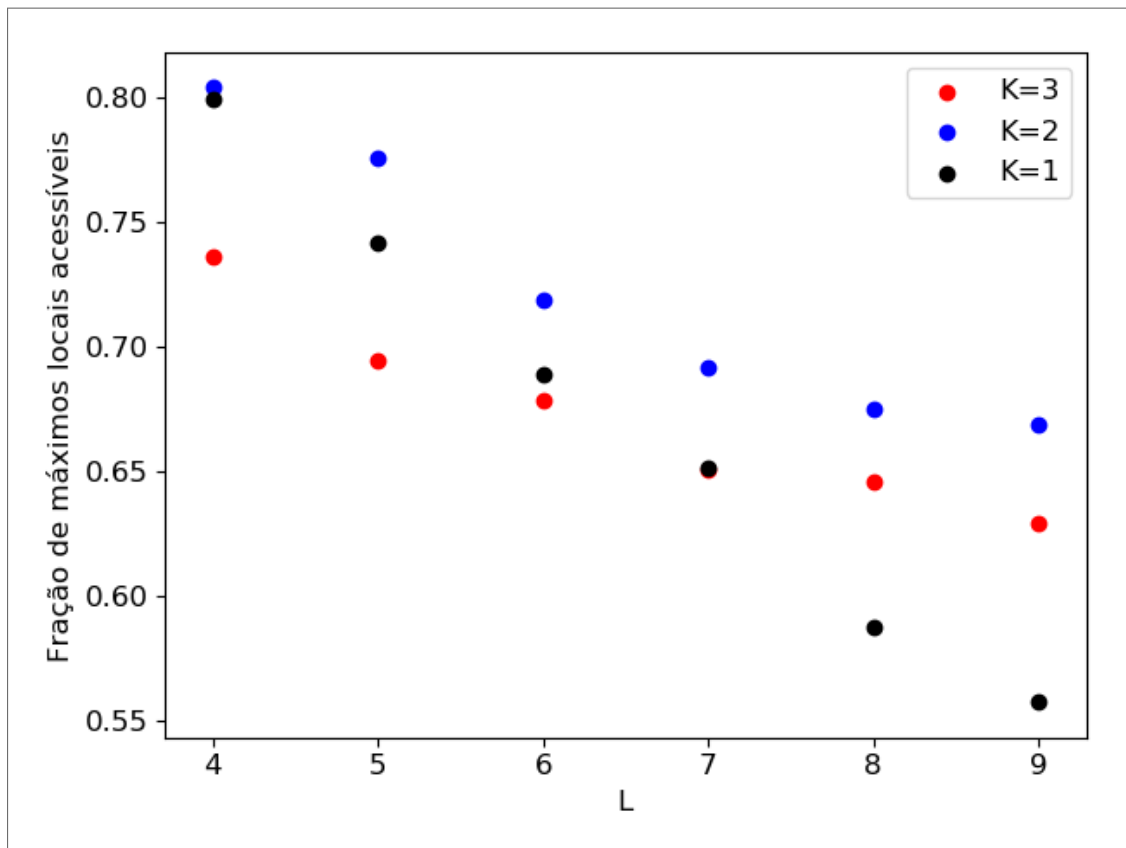
Fonte: Elaborado pelo autor (2020)

A acessibilidade do máximo global está intimamente relacionada ao número absoluto de picos, uma vez que ao partir do antípoda cada máximo local acrescentado ao relevo surge necessariamente entre o ponto de partida e o máximo global, e representa assim mais um obstáculo ao acesso deste último. Portanto, trata-se de uma medida inversa à quantidade de picos no relevo, e pela mesma argumentação também apresenta comportamento monotônico em L e K . Por ser decrescente com o número de picos, pode-se dizer que a acessibilidade do global mede então o nível de suavidade do relevo.

Voltamos agora para a fração de máximos acessíveis, isto é, aqueles conectados ao ponto de partida por pelo menos uma trajetória monotônica em fitness. Observe que tal quantidade não depende do tipo de caminhada, dado que ambos os tipos envolvem somente passos unitários e ascendentes em fitness, e diferem apenas pela probabilidade de ocorrência. Por outro lado, esta quantidade depende sensivelmente da condição inicial. O resultado das simulações é apresentado na Figura 15. Neste gráfico fica claro um comportamento da fração de máximos acessíveis não tão simples quanto o do número de picos, apesar da relação direta com este. Embora haja uma dependência monotônica em L , o mesmo não é observado em K .

Na verdade, o próprio ordenamento em K varia com L . A complexidade dessa dependência reside na intrincada relação entre a abundância de picos, distância entre picos, e distância dos picos ao ponto de partida.

Figura 15 – Fração de máximos locais acessíveis em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.



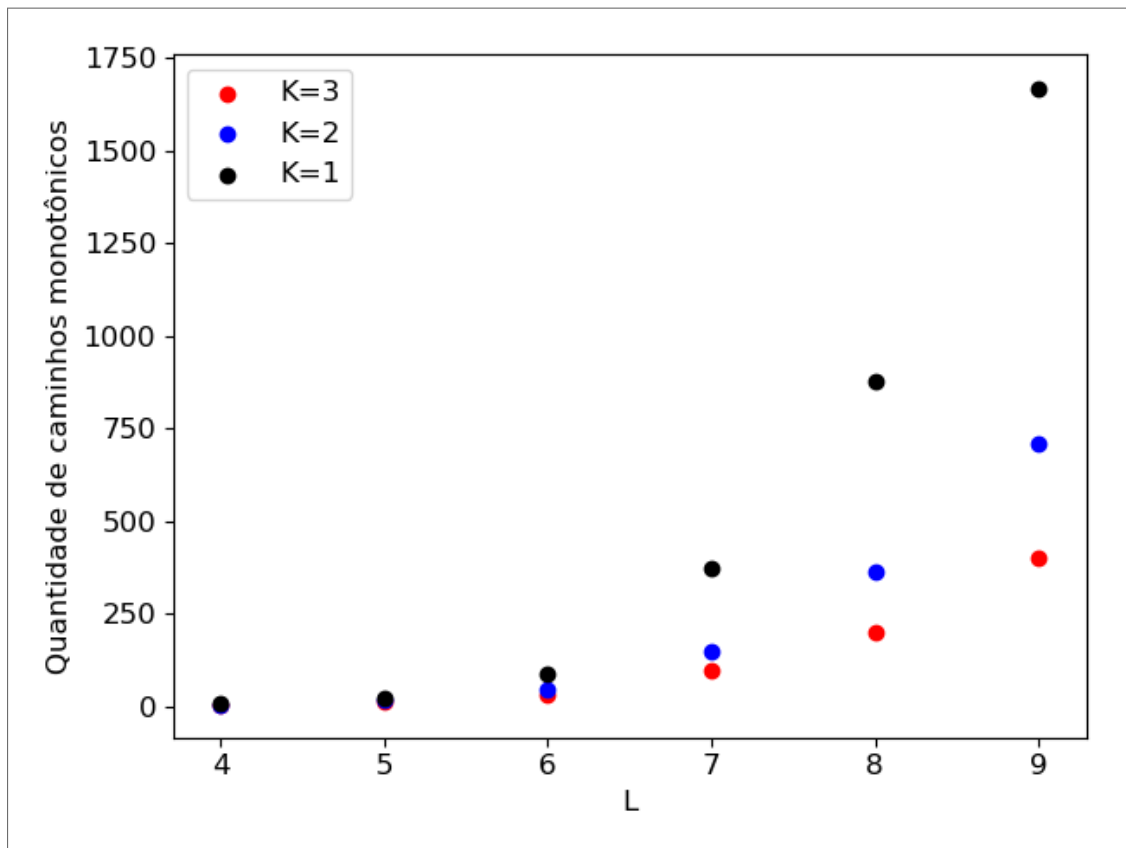
Fonte: Elaborado pelo autor (2020)

Vale ressaltar que o número total de máximos acessíveis não compartilha a mesma riqueza de comportamento da fração. De fato, em valores absolutos a abundância de máximos acessíveis acompanha o comportamento simples da quantidade de picos no relevo (Figura 13). Em termos de previsibilidade, é intuitivo que um menor valor tanto da fração quanto do número absoluto de máximos acessíveis implique num maior determinismo. Porém, ambos os casos tendem a configurar uma medida pobre, e só devem ser consideradas de maneira complementar a quantidades que envolvam a frequência com que cada máximo ocorre, como a *predictability* com relação aos pontos finais.

Na Figura 16 apresentamos a quantidade média de caminhos monotônicos em função de L . Os caminhos monotonicamente crescentes são os únicos permitidos a caminhantes adaptativos e mesmo em dinâmicas populacionais tem grande presença no regime de baixas mutações, justificando sua relevância. Assim como a fração de máximos acessíveis, o número de caminhos monotônicos depende apenas do ponto de partida mas não do tipo de caminhada.

Sua dependência é monotônica tanto em L quanto em K , seguindo comportamento oposto ao da densidade de picos. A dependência em L é trivial, e a dependência em K ressalta o papel da epistasia. Esta cria vínculos adicionais, limitando o acesso ao máximo global.

Figura 16 – Quantidade de caminhos monotonicamente crescentes em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.



Fonte: Elaborado pelo autor (2020)

Uma maneira particularmente interessante de esclarecer essa relação é através da introdução do conceito de “*tree component*”, como apresentado por LOBKOVSKY; WOLF; KOONIN. O *tree component* é simplesmente o conjunto de todos os pontos do relevo que possuem no máximo um vizinho mais apto. O nome se refere ao fato de pontos pertencentes a este conjunto formarem uma ou várias estruturas disjuntas ramificadas e sem *loops*. Para um caminhante, estar sobre um ponto pertencente ao *tree component* significa ter sua trajetória completamente determinada. Ocorre que o número de pontos contidos no *tree component* é diretamente proporcional à densidade de picos no relevo, isto porque cada pico cobre um subconjunto diferente do *tree component*. Por outro lado, quanto maior a área coberta pelo *tree component*, menor a distância do ponto de partida a esta estrutura e menor, portanto, a diversidade de passos e trajetórias permitidos ao caminhante. O resultado final é o aumento do número de caminhos distintos com a diminuição da densidade, que é fruto por sua vez do aumento do comprimento L das sequências ou diminuição da intensidade K das interações

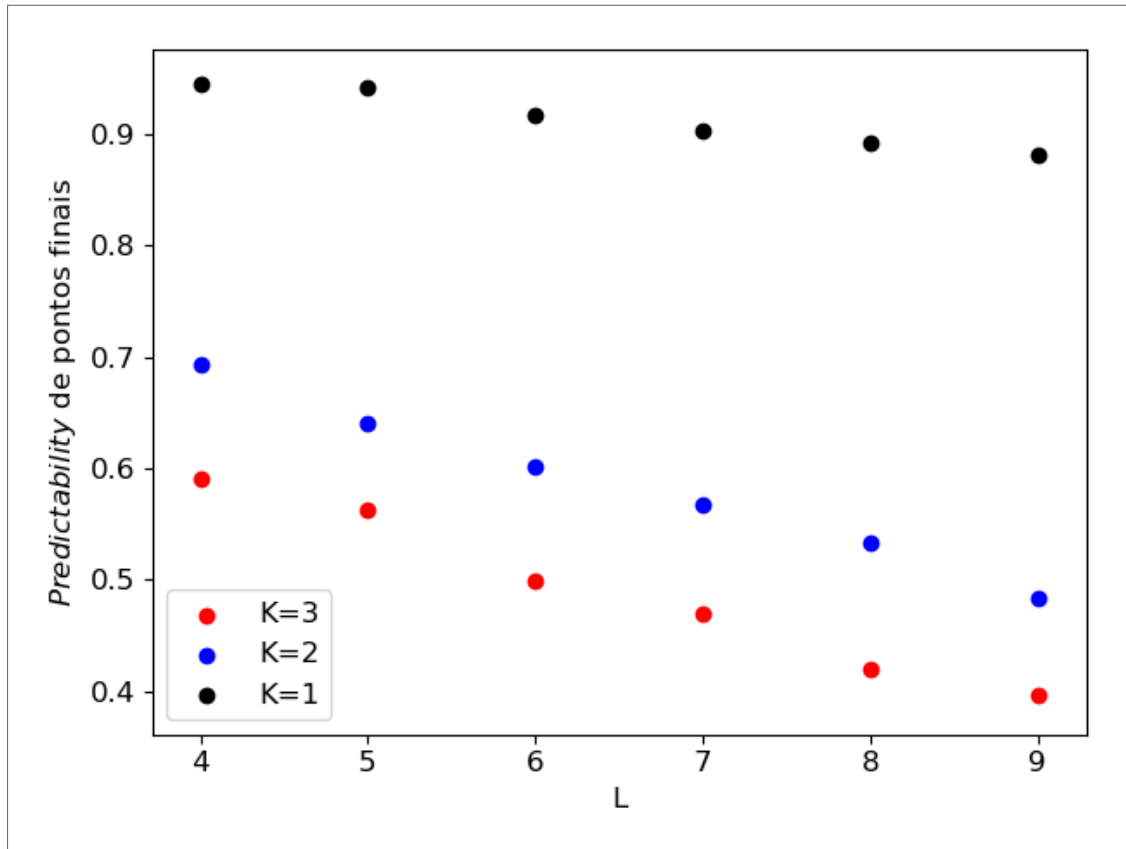
epistáticas.

Assim como a fração ou o número de máximos acessíveis, o número de caminhos monotônicos é uma medida limitada de previsibilidade. Sem dúvida, espera-se que a previsibilidade aumente com a diminuição de trajetórias possíveis, desde que essa menor quantidade de caminhos não seja compensada por uma distribuição mais uniforme das probabilidades de ocorrência.

As duas próximas medidas a serem discutidas são a *predictability* com relação aos pontos finais e aos caminhos evolutivos. Ambas envolvem a aplicação da Equação 3.1, no primeiro caso à distribuição de probabilidade de acesso dos máximos locais e, no segundo, à distribuição de probabilidade de ocorrência dos diferentes caminhos. Antes de qualquer análise, há uma observação pertinente a ser feita sobre a *predictability* de caminhos. Diferentemente da divergência média de caminhos, a *predictability* é aplicável a ensembles de caminhos que terminam sobre sequências diferentes. Ao aplicá-la dessa maneira, no entanto, ela inclui duas fontes de entropia, a informação sobre o máximo alcançado e a informação sobre a trajetória escolhida a esse máximo. A *predictability* de pontos finais, por outro lado, só inclui a primeira fonte. Dessa forma, é interessante analisar primeiro a *predictability* aplicada aos pontos finais e só então interpretar seu uso na distribuição de caminhos já dispondo do comportamento de uma das fontes de entropia. Dito isso, a Figura 17 apresenta o gráfico da dependência média da *predictability* de pontos finais com L e K .

De acordo com o que foi exposto na Seção 3.1 sobre a *predictability*, o valor resultante da Equação 3.1 quando aplicada à frequência de acesso dos diferentes máximos pode ser interpretada como o inverso do número efetivo de máximos acessados. Seguindo este raciocínio, os resultados da Figura 17 indicam que o número efetivo de máximos acessados cresce tanto com L quanto com K , simplesmente porque o incremento de ambos aumenta o número de máximos no relevo, aumentando também o número dos acessíveis, independentemente de sua quantidade relativa. Assim, a *predictability* com relação aos pontos finais tem comportamento simplesmente oposto ao do número de máximos locais com respeito a L e K , sugerindo que a distribuição de probabilidade em si é menos influente que o número absoluto de máximos na previsibilidade. A distribuição, no entanto, tem participação central na forma como a variação da *predictability* em função de L se intensifica com o valor de K . Para $K = 1$, por exemplo, a *predictability* varia de 0,95 a 0,88, intervalo associado a uma variação entre 1,05 e 1,14 do número efetivo de máximos acessados, isto é, praticamente constante e igual a 1 para todo o intervalo de L analisado. No caso $K = 1$, independentemente do acréscimo do número de máximos provido pela incremento de L , a grande dispersão destes máximos garante a dominância exclusiva do pico mais próximo ao antípoda. À medida que K cresce, aumenta não só o número de máximos locais mas também sua densidade, provocando uma uniformização da concorrência dos picos na vizinhança do ponto inicial uma vez que a proximidade é determinante na dominância, e quanto mais uniforme a distribuição de probabilidade menor o valor fornecido pela Equação 3.1. Passamos agora para a distribuição de caminhos, ilustrada

Figura 17 – *Predictability* de pontos finais em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.



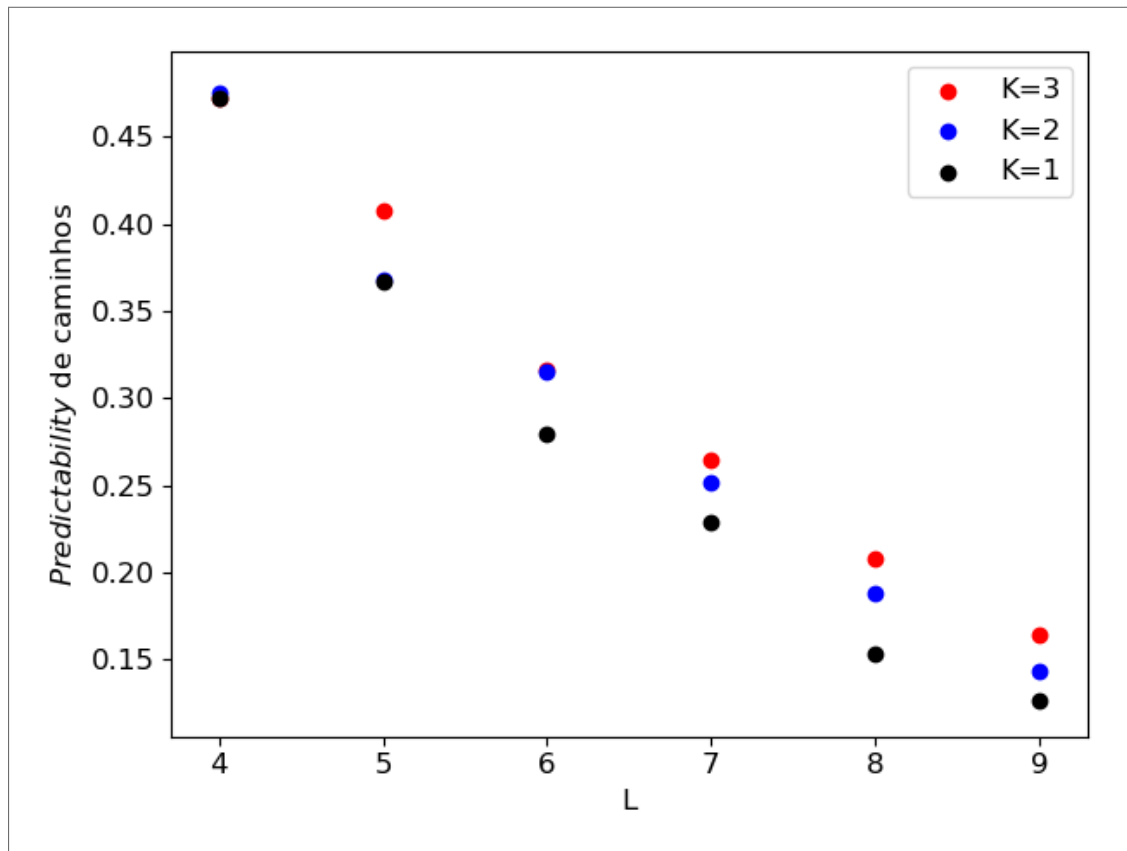
Fonte: Elaborado pelo autor (2020)

na Figura 18.

Apesar da já mencionada complexidade de fatores envolvendo a *predictability* de caminhos, esta medida apresenta um comportamento simples e robusto de decrescimento em termos de L . O principal fator por trás desse comportamento é o aumento do número de caminhos monotônicos com o tamanho do relevo. Além disso, dado que o número efetivo de máximos acessíveis também aumenta, sabemos que as frequências dos caminhos tendem a ficar cada vez mais bem distribuídas, isto é, uniformes. Em resumo, ambas as fontes de entropia confluem para uma diminuição da repetibilidade de trajetórias com L . O mesmo não ocorre com K . De fato, o aumento de K está associado tanto ao aumento do número de máximos acessíveis quanto à diminuição do número de caminhos monotônicos. Se por um lado o primeiro efeito provoca uma distribuição mais uniforme dos caminhos, diminuindo a repetibilidade, o segundo causa uma diminuição do número efetivo de caminhos, aumentando a repetibilidade. Aparentemente, o segundo efeito predomina, embora a variação seja pequena. No entanto, é difícil fazer tal afirmação. Basicamente, como a aplicação da *predictability* desta maneira não é usual, não sabemos precisamente como a Equação 3.1 confere pesos a estes dois fatores.

A última medida a ser considerada é o comprimento médio das caminhadas. Trata-se de mais uma medida cujo comportamento é totalmente determinado pela densidade de picos.

Figura 18 – *Predictability* de caminhos em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.

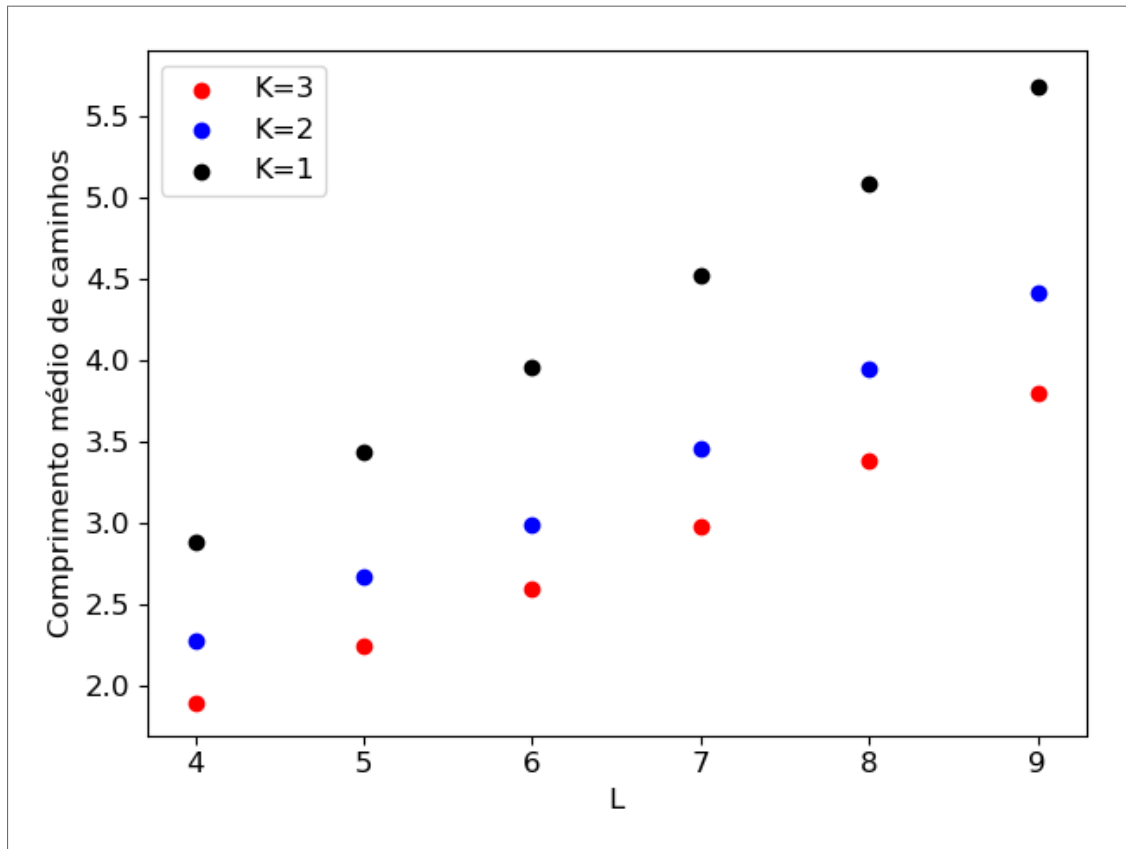


Fonte: Elaborado pelo autor (2020)

Quanto menor a densidade de picos, mais distante do antípoda estão os máximos em média, e mais longas as caminhadas. Um K menor também significa uma menor correlação do relevo e, com isto, de fato a probabilidade de que os máximos locais estejam distantes é elevada. A dependência aproximadamente linear em L para todos os valores de K chama atenção e está associada provavelmente ao fato de a densidade de picos ser uma função inversamente proporcional da quantidade L de vizinhos. Para o comprimento médio das caminhadas, por outro lado, isso se reflete numa dependência diretamente proporcional com esta mesma quantidade.

Sem dúvida, a discussão sobre o conjunto de medidas envolvendo propriedades dos relevos e de caminhadas sobre eles resultou numa série de conclusões que tornaram mais claro o arquétipo básico das relações entre rugosidade, acessibilidade e previsibilidade. Se tivermos que elencar as mais importantes, no entanto, duas merecem destaque. Primeiro, é notável a presença dos conceitos de número e densidade de picos ao longo de todo o texto, muito embora os parâmetros de controle fossem o comprimento L das sequências e a intensidade K das interações epistáticas. Isso indica que o número e a densidade de máximos são as quantidades que governam majoritariamente o comportamento das medidas evolutivas em relevos de adaptação, e que os diferentes parâmetros e formas de controle de rugosidade de cada modelo teórico ocupam um papel muito mais sutil nesse cenário. A segunda observação

Figura 19 – Comprimento médio de caminhos em função de L para diferentes valores de K . Cada ponto representa o valor médio sobre 1.000 relevos gerados pelo Modelo NK e caminhadas partindo do antípoda.



Fonte: Elaborado pelo autor (2020)

é mais específica e envolve a previsibilidade. A aplicação das duas formas de *predictability* mostrou uma concorrência interessante entre o número de máximos e o número de caminhos na determinação da previsibilidade de trajetórias como função de K . Observou-se que um aumento de K é acompanhado de uma diminuição da previsibilidade do máximo alcançado pelo caminhante, ao mesmo tempo em que essa maior rugosidade do relevo provoca uma restrição das trajetórias permitidas ao caminhante. O surgimento de tal concorrência num modelo com poucos parâmetros como o NK evidencia que o problema da previsibilidade pode ser complexo mesmo em cenários simplificados.

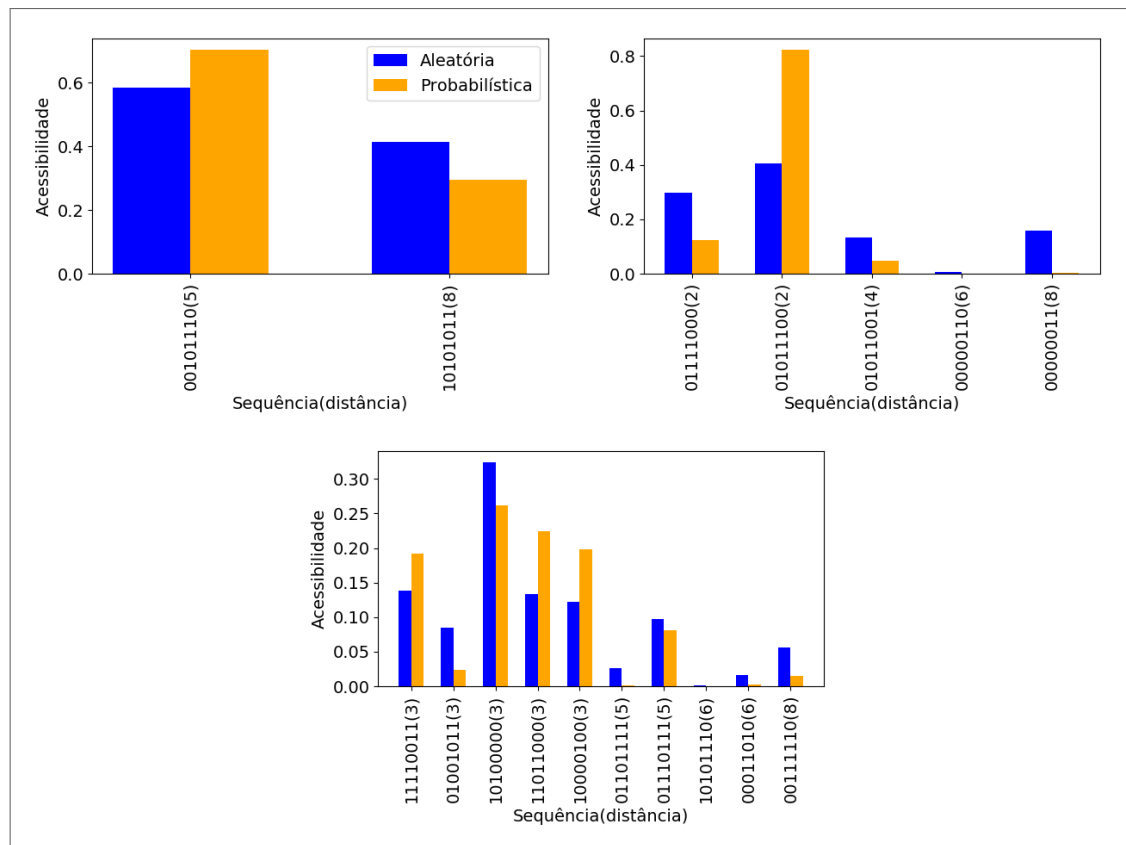
4.1.2 Medidas por máximo local

Ao contrário da subseção anterior, onde as medidas representavam o relevo como um todo, aqui voltamos nossa atenção para os máximos locais, isto é, extraíndo quantidades sobre ensembles de caminhadas que terminam sobre um mesmo pico. Por conta dessa diferença fundamental podemos agora fazer uso da divergência média de caminhos (Equação 3.3). As medidas aplicadas por máximo são listadas a seguir:

- Acessibilidade
- Quantidade de caminhos monotonicamente crescentes
- *Predictability*
- Divergência média de caminhos
- Comprimento médio de caminhos

Uma vez que a quantidade e as características dos máximos locais são únicas de cada relevo, estaremos limitados a analisá-los individualmente, em lugar das médias sobre relevos. Dessa forma, torna-se mais difícil destacar relações gerais, e há sempre a ressalva da possibilidade de manifestação de particularidades do relevo em questão. Foram gerados 3 relevos do Modelo NK com $L = 8$, um para cada valor de $K \in \{1, 2, 3\}$. Dentre as infinitas realizações possíveis do Modelo NK, escolhemos os relevos analisados aqui seguindo como único critério a acessibilidade de todos os máximos, afim de potencializar a capacidade de comparação entre eles. Foram simuladas 10.000 caminhadas por relevo partindo do antípoda.

Figura 20 – Acessibilidade por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.



Fonte: Elaborado pelo autor (2020)

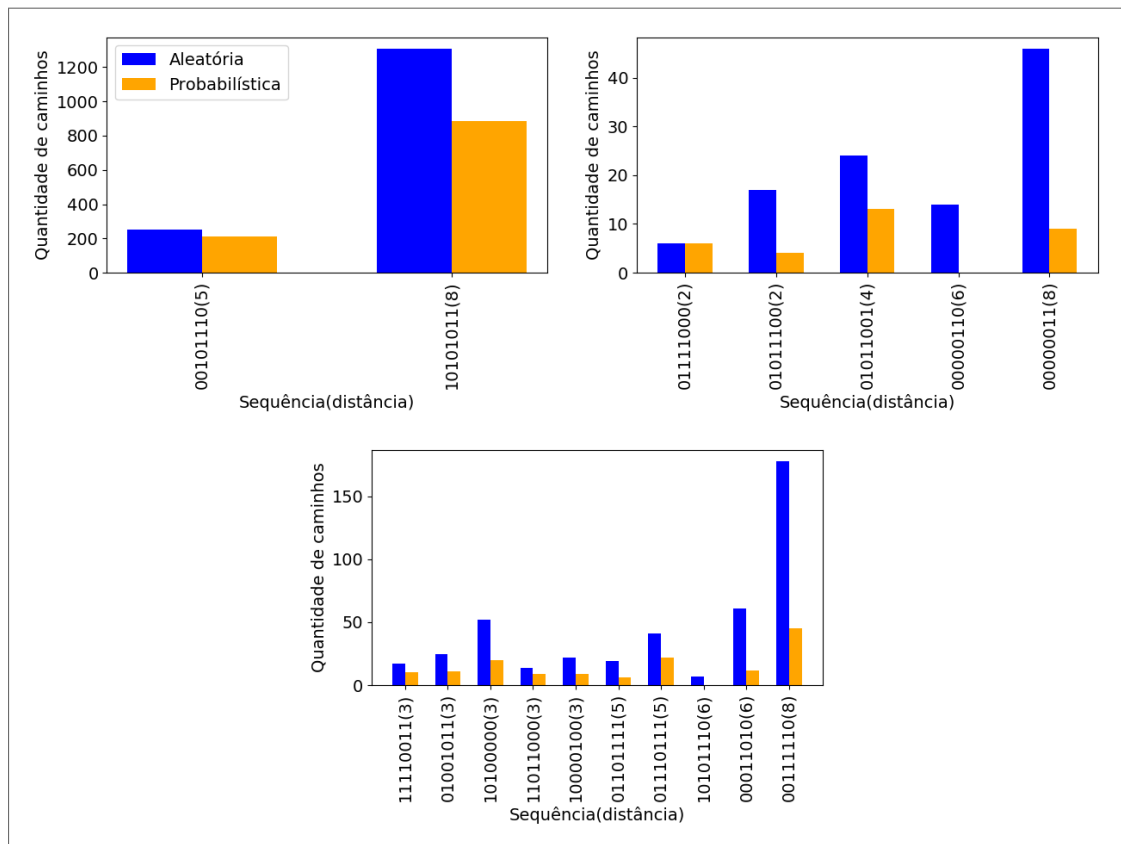
Começando pela acessibilidade, apresentamos na Figura 20 para cada relevo um gráfico de barras da probabilidade de ocorrência de cada máximo sob as caminhadas aleatória e probabilística, onde as sequências estão em ordem crescente de distância ao antípoda, o ponto de partida. Quando degeneradas neste critério, estão em ordem crescente de fitness.

Nos três gráficos é possível notar uma correlação negativa apreciável da acessibilidade com a distância de Hamming ao ponto de partida. Tal resultado se deve ao fato de a bacia de atração de um máximo suficientemente próximo da sequência inicial cobrir uma grande parte de sua vizinhança. No caso $K = 1$, por exemplo, é inesperada a alta acessibilidade do máximo global, dado que este se encontra diametralmente oposto ao ponto de partida. Seu único concorrente, no entanto, se encontra mais próximo dele que do antípoda, permitindo assim essa acessibilidade incomum. Olhando agora para os três relevos, é possível perceber também que a acessibilidade do máximo global sempre foi menor sob caminhadas probabilísticas. Isto ocorre porque, embora este seja o máximo de maior fitness, sua grande distância ao antípoda torna menos íngremes os passos que levam até ele. Note que o ordenamento em fitness dos grupos de sequências igualmente distantes ao antípoda deveria apresentar, segundo o critério anterior de inclinação dos passos, um comportamento monotônico. Não fica aparente, no entanto, qualquer correlação entre acessibilidade e fitness, sugerindo que no caso de máximos na mesma “esfera” de Hamming outros fatores possuem uma influência maior na acessibilidade, provavelmente a distribuição espacial dos máximos e bacias de atração, aspectos essencialmente estocásticos e praticamente descorrelacionados com o fitness.

A medida a seguir é a quantidade de caminhos monotônicos ou, equivalentemente, o número de caminhos distintos, exibido na Figura 21. É possível notar uma correlação positiva entre o número de caminhos e a distância ao ponto de partida, apesar de menos intensa no caso $K = 3$. Esse resultado deriva diretamente do aumento do número de possibilidades com a duração da caminhada (REIA; CAMPOS, 2020), como evidenciado pela superioridade do número de trajetórias ao máximo global em todos os relevos. Ademais, vale ressaltar que pelos seus comportamentos opostos com a distância ao antípoda, acessibilidade e número de caminhos apresentam uma correlação negativa intuitivamente inesperada. Nos gráficos isso se reflete em uma inversão sutil das barras. O fato de que o número de caminhos distintos depende da versão da caminhada adaptativa indica apenas que a aleatória permite uma maior exploração do relevo enquanto que a probabilística adota os caminhos mais íngremes. Além disso, um maior número de caminhos também significa uma maior sobreposição com as bacias de atração de outros máximos.

Passamos agora para as duas medidas de caracterização da previsibilidade que levam em consideração as distribuições de probabilidade das caminhadas. A primeira e mais simples, a *predictability* (Equação 3.1), é exposta na Figura 22. Primeiramente, é possível notar a oposição da *predictability* com o número de caminhos, porém bem menos clara para $K = 3$. Seguindo a interpretação da *predictability* como o inverso do número efetivo de caminhos, isto sugere que a distribuição de trajetórias se torna cada vez mais informativa que o número

Figura 21 – Quantidade de caminhos monotônicos por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.

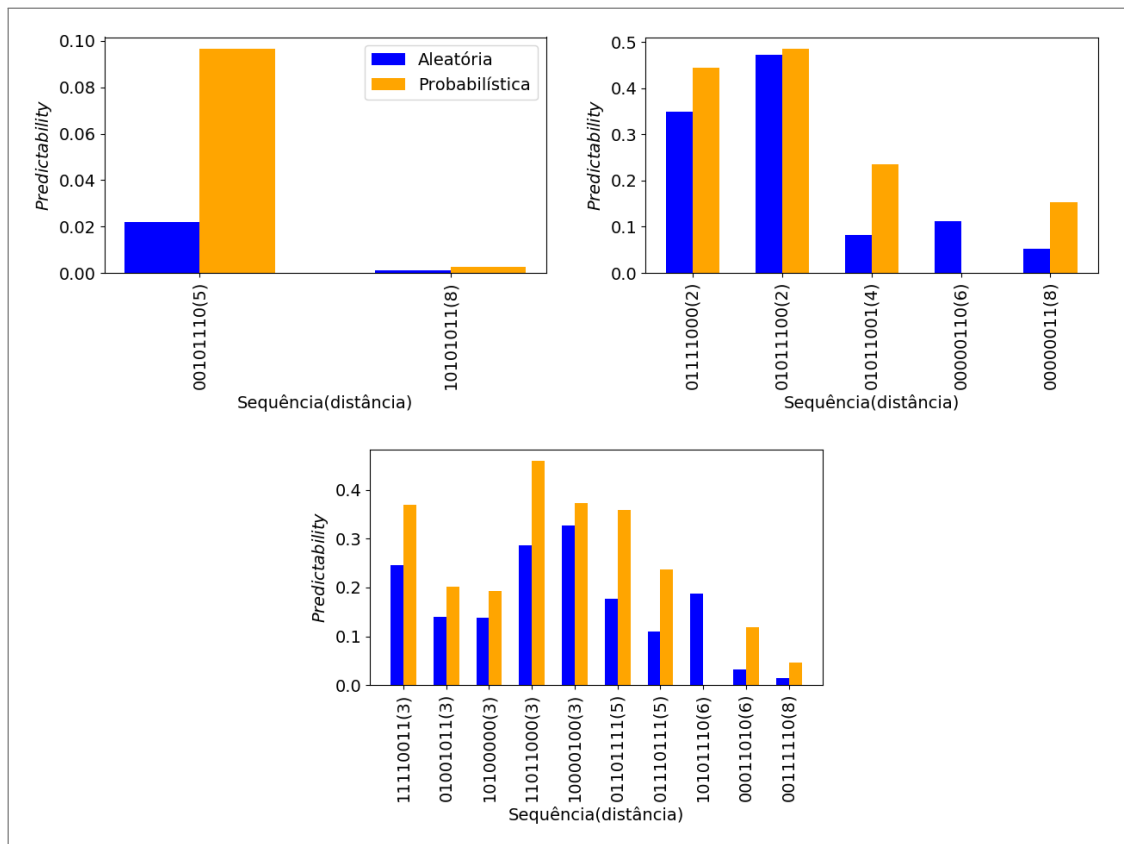


Fonte: Elaborado pelo autor (2020)

absoluto delas, à medida que a rugosidade aumenta. A relação com a acessibilidade é semelhante, porém a correlação é positiva, pois de forma geral a acessibilidade não está associada a um maior número de caminhos. No que diz respeito aos tipos de caminhadas, espera-se um maior valor da *predictability* no caso probabilístico, motivado pelo menor grau de aleatoriedade dos passos, comportamento de fato observado para todos os máximos onde o acesso via caminhada probabilística foi não nulo.

No caso da divergência média de caminhos, apresentada na Figura 23, não é possível firmar qualquer relação com a acessibilidade ou número de caminhos monotônicos nos relevos explorados. Sua relação com a distância de cada máximo ao antípoda também não é óbvia, indicando correlação positiva para $K = 1$, negativa para $K = 2$, e nenhuma correlação para $K = 3$. Com a *predictability*, no entanto, fica aparente uma correlação negativa entre estas quantidades, que diminui de intensidade com o aumento de K . De fato, para $K = 1$ e $K = 2$ observa-se que as barras mais altas no gráfico da Figura 22 tendem a diminuir quando levadas ao gráfico da Figura 23, enquanto valores mais baixos tendem a aumentar, corroborando a definição da *predictability* e divergência como medidas de previsibilidade e imprevisibilidade, respectivamente. Para $K = 3$, no entanto, a oposição não é mais tão marcante, indicando que a similaridade entre caminhos, quantificada apenas pela divergência, passa a desempenhar

Figura 22 – *Predictability* por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.

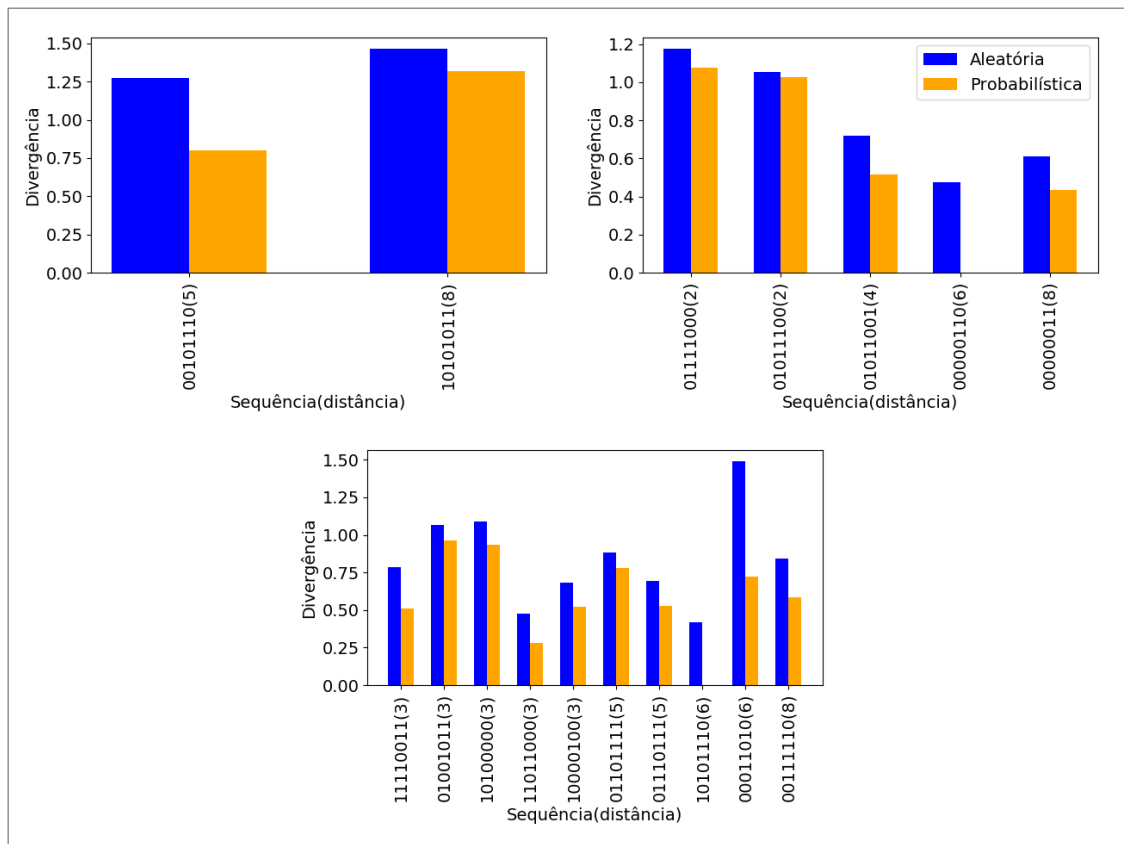


Fonte: Elaborado pelo autor (2020)

um papel importante em relevos moderadamente rugosos. Como esperado, a divergência foi menor sob caminhadas probabilísticas para todos os máximos.

Por último, o comprimento médio de caminhadas é mostrado na Figura 24. Sem dúvidas há uma grande correlação entre comprimento médio e distância de Hamming ao antípoda. Um resultado menos óbvio, mas também esperado, é a ocorrência de menores valores do comprimento médio sob caminhadas probabilísticas em relação às aleatórias, observado amplamente nos três relevos. Resumidamente, caminhantes probabilísticos enxergam melhor os gradientes do relevo e, conseqüentemente, são mais eficientes na escalada aos máximos. Comparando o valor do eixo vertical com a distância ao antípoda, mostrada no eixo horizontal, é possível inferir a presença de caminhadas indiretas sempre que o primeiro valor for maior que o segundo. Nesse sentido, vale ressaltar ainda que a diferença entre o comprimento de caminhadas diretas e indiretas é sempre um múltiplo de dois, uma vez que passos que aumentam a distância de Hamming ao pico de destino devem necessariamente ser revertidos em algum momento. No relevo para $K = 1$, por exemplo, o comprimento médio das caminhadas que levam ao máximo que dista 5 mutações do antípoda está mais próximo de 6 que 4 em ambas as caminhadas, indicando uma maior contribuição de caminhadas indiretas na acessibilidade deste. Em relação ao fitness, não se observa qualquer correlação entre a altura dos picos e o comprimento das

Figura 23 – Divergência média de caminhos por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.

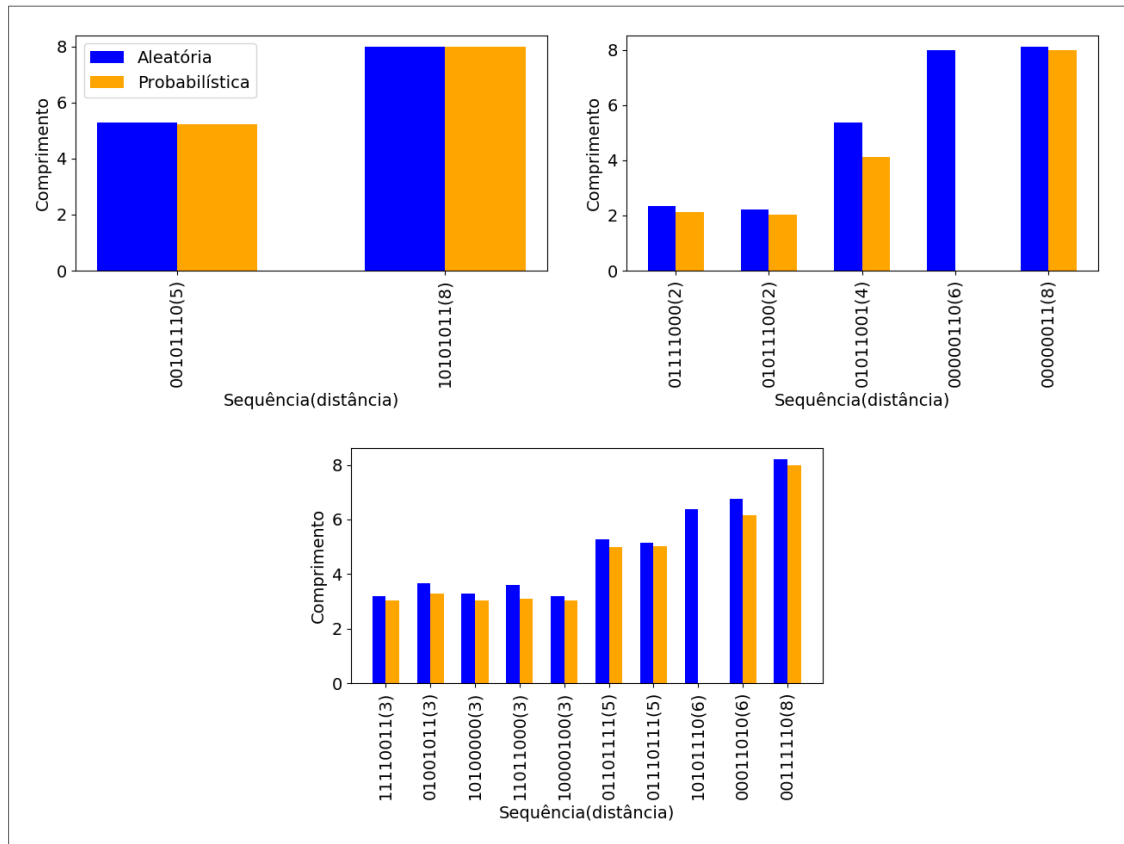


Fonte: Elaborado pelo autor (2020)

caminhadas.

Através da análise comparativa da acessibilidade e previsibilidade sobre diferentes máximos em um mesmo relevo podemos concluir tanto a presença de comportamentos gerais quanto da unicidade de cada relevo e, portanto, da variabilidade entre eles. Entre as relações gerais, estas sempre envolvem de uma maneira ou de outra dois fatores topográficos: a distância do pico em questão e quão íngreme são os passos envolvidos. No que diz respeito às medidas em si, o resultado mais interessante é a diminuição da correlação entre número de caminhos monotônicos e *predictability*, assim como entre *predictability* e divergência, com o aumento do parâmetro K . Segundo essa observação, podemos inferir que, sob baixa rugosidade tanto a inclusão das distribuições quanto a inclusão de uma medida de similaridade de caminhos acrescenta pouca informação sobre a previsibilidade das trajetórias. À medida que a rugosidade aumenta, no entanto, essas diferenças começam a passar de redundantes a complementares, indicando que em relevos acidentados a previsibilidade se torna um fenômeno mais complexo, embora não seja claro como isso acontece.

Figura 24 – Comprimento médio de caminhadas por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 10.000 caminhadas em três relevos de tamanho $L = 8$ para valores de $K = 1$ e $K = 2$, da esquerda para direita, e $K = 3$ abaixo.



Fonte: Elaborado pelo autor (2020)

4.2 DINÂMICA POPULACIONAL

Nesta seção exploramos a intrincada dependência da repetibilidade de trajetórias evolutivas com os parâmetros de tamanho de população e taxa de mutação, nesta ordem. Os resultados envolvem a simulação de dinâmicas populacionais de acordo com o algoritmo de reprodução assexuada do tipo Wright-Fisher descrito na Subseção 3.2.2. Foram replicadas 1.000 populações por relevo, onde cada passo é gravado como a sequência mais apta presente na geração, e a caminhada continua até o surgimento da sequência mais apta do relevo, o máximo global. Extraíu-se do ensemble de caminhadas o seguinte conjunto de medidas:

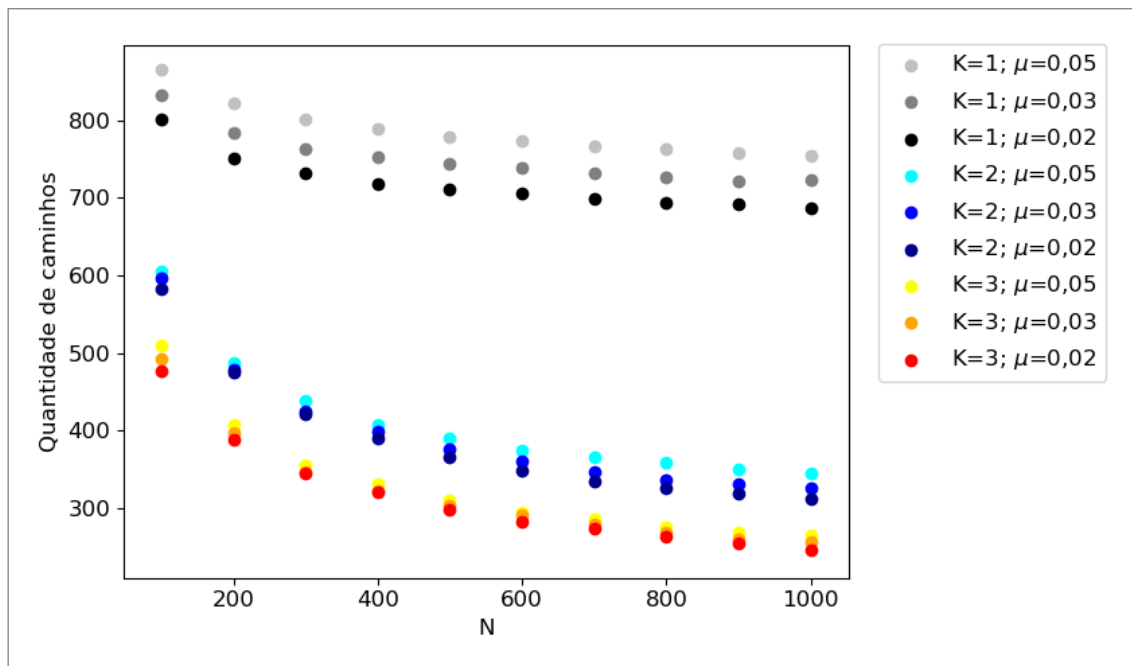
- Quantidade de caminhos
- *Predictability*
- Divergência média de caminhos
- Comprimento médio de caminhos

Foi feita então a média das medidas sobre 50 relevos de tamanho $L = 8$ para cada valor de $K \in \{1, 2, 3\}$ e para cada par de valores N e μ de tamanho da população e taxa de mutação escolhidos.

4.2.1 Tamanho da população

Começamos a análise pelo comportamento das medidas listadas em função do tamanho N de população, variando de 100 a 1.000 indivíduos por geração, em intervalos de 100, para três valores diferentes da taxa de mutação $\mu \in \{0,02; 0,03; 0,05\}$. A Figura 25 traz os resultados para a quantidade de caminhos distintos.

Figura 25 – Quantidade de caminhos em função do tamanho N de população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.



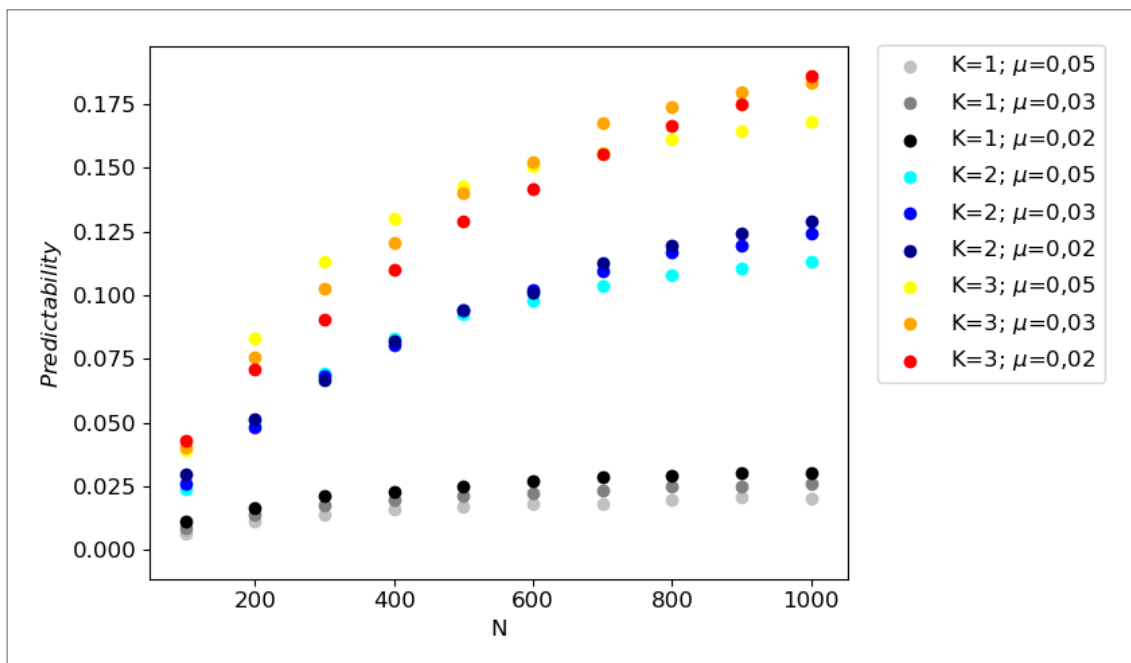
Fonte: Elaborado pelo autor (2020)

Na Figura 25 é possível notar que para todos os valores de μ e K simulados a quantidade de caminhos é monotonicamente decrescente com N . Basicamente, quanto maior a população maior a quantidade média de mutantes por geração. À medida que a diversidade de genótipos cresce, a identidade do indivíduo mais apto se torna menos sujeita a fatores estocásticos e os passos se tornam cada vez mais determinísticos. No limite de uma população infinita teria-se uma única trajetória. Comparativamente às caminhadas adaptativas, isto é equivalente a aumentar o peso da diferença de fitness na tomada dos passos, tendendo por fim ao limite “greedy” onde o vizinho mais apto é escolhido deterministicamente. Outra observação pertinente é o aumento do número de caminhos com a taxa de mutação. Este resultado soa contraditório, uma vez que μ tem o mesmo efeito que N no número absoluto $N\mu$ de mutan-

tes esperados por geração. Ocorre que, com população fixa, μ provoca também um aumento relativo desta quantidade. Esse aumento relativo carrega um forte fator estocástico, dado que mesmo a sequência mestra, isto é, a mais populada, está constantemente perdendo indivíduos via mutação. Além disso, μ contribui quadraticamente para o valor esperado de mutações duplas $N\mu^2$ enquanto N o faz apenas linearmente. Dada a imensa quantidade de mutantes de segunda ordem em diante, a possibilidade de cruzamento de vales via passos não unitários de mutação se torna expressiva e a identificação do indivíduo mais apto por geração se torna mais caótica, de maneira que o aumento de μ acaba produzindo um efeito oposto ao de N .

A *predictability*, mostrada na Figura 26, parece seguir as mesmas considerações anteriores, apresentando também dependência monotônica em N , porém crescente, claro. Existe, no entanto, uma diferença marcante em relação à quantidade de caminhos: a variação da *predictability* com o tamanho da população depende fortemente do valor de K , apresentando uma mudança modesta para $K = 1$ e uma diferença ostensiva para $K = 3$. Assim, conclui-se que relevos mais acidentados potencializam o fator determinístico do aumento da população. Ainda em comparação à quantidade de caminhos, a *predictability* apresenta um comportamento sutilmente mais complexo com μ , que se torna proeminente com o aumento da rugosidade. Uma análise mais detalhada será feita na Subseção 4.2.2 a seguir, onde tratamos diretamente da taxa de mutação. Por hora, podemos concluir que embora a taxa de mutação não tenha influência significativa na dependência do número absoluto de caminhos com o tamanho da população, o mesmo não é verdade para o número de caminhos efetivos, isto é, de trajetos com participação expressiva no ensemble.

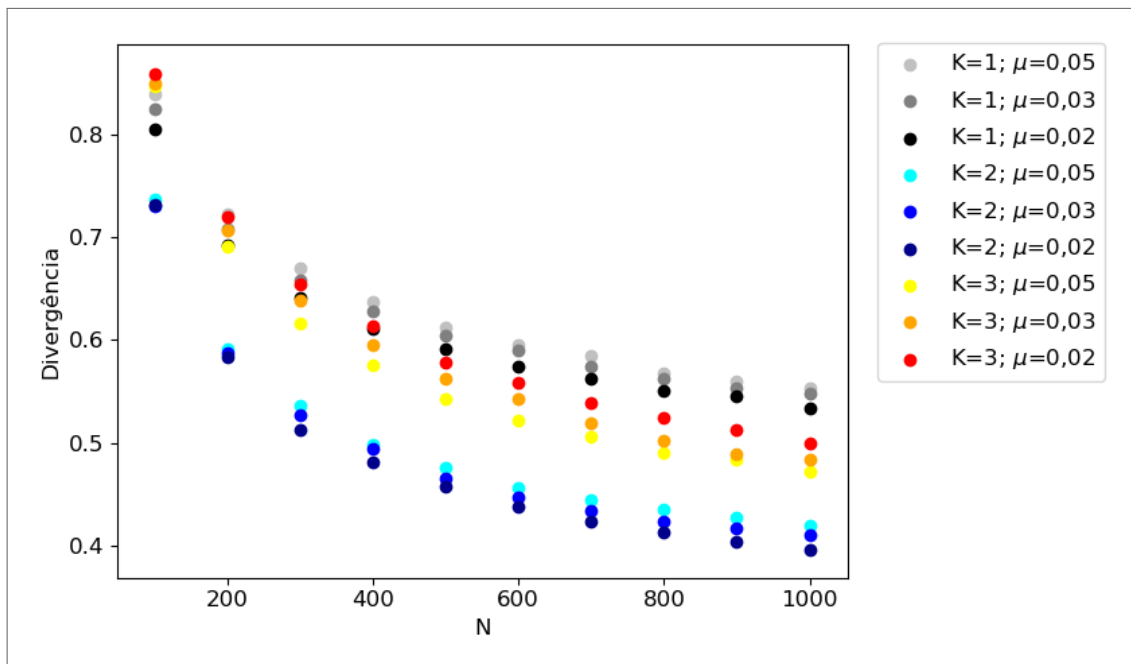
Figura 26 – *Predictability* em função do tamanho N de população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.



Fonte: Elaborado pelo autor (2020)

Na Figura 27 mostramos a divergência média de caminhos como função de N . Tratando-se de uma medida de imprevisibilidade, é esperado o comportamento monotonicamente decrescente com N , observado para todos os valores de μ e K . O fato que realmente chama atenção é a minimização da divergência para curva de $K = 2$, valor intermediário entres os três analisados. Como já discutimos, tanto o número real quanto efetivo de caminhos diminui com K , o que explica a diminuição da divergência (imprevisibilidade) de $K = 1$ para $K = 2$, mas não o aumento de $K = 2$ para $K = 3$. Este resultado inesperado deve estar associado, portanto, à uma característica própria da divergência, a contabilização da similaridade entre caminhos. Fica evidente então que o aumento da rugosidade produz dois efeitos conflitantes na previsibilidade de trajetórias, a restrição da quantidade de caminhos acessíveis e o aumento da dissimilaridade entre estes.

Figura 27 – Divergência média de caminhos em função do tamanho N de população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.

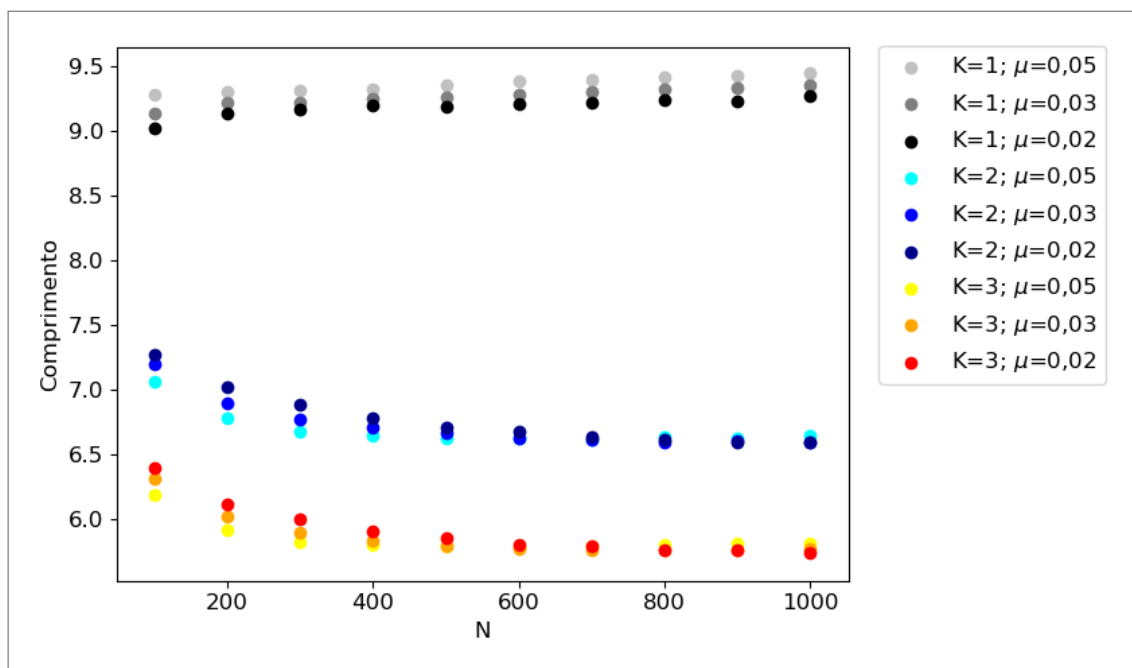


Fonte: Elaborado pelo autor (2020)

A quarta e última medida é o comprimento médio das caminhadas, apresentada na Figura 28. Nesse contexto há uma importante ressalva a ser feita sobre dinâmicas populacionais. Diferentemente do que acontece com caminhantes adaptativos, onde um trajeto do antípoda ao máximo global deve necessariamente conter L ou mais mutações, numa dinâmica de população finita podem haver caminhadas mais curtas que L , como ocorre para $K = 2$ e $K = 3$. É importante notar também que, enquanto comprimento e número de gerações são equivalentes em caminhadas adaptativas, em populações finitas esses conceitos são claramente distintos, e o comprimento está mais para um representante da diversidade de genótipos pelo qual a população passou. De fato, estamos armazenando a sequência mais apta de cada geração, o

que significa que se dois mutantes que tendem a se fixar consecutivamente passem a coexistir e competir devido ao aumento da população, ocorrerá então o armazenamento de apenas um deles em vez dos dois seguidamente, provocando assim a diminuição do comprimento médio com N . Curiosamente, nas simulações este resultado não foi observado no caso $K = 1$, onde se observa uma dependência crescente do comprimento médio com o tamanho da população. Podemos conjecturar, portanto, que tal fenômeno é particular de relevos suaves. Como vimos anteriormente, tais relevos possuem grande abundância de caminhos alternativos e, consequentemente um número elevado de linhagens deve ocorrer paralelamente numa mesma população. Assim, uma provável explicação para este comportamento peculiar é que o incremento de N aumenta o número de linhagens exploradas, o que implica numa maior varredura de sequências e num maior armazenamento destas. A conclusão que fica é que devido a nossa escolha de caracterização de uma caminhada, o comprimento médio se torna uma quantidade ambígua e pouco informativa.

Figura 28 – Comprimento médio de caminhos em função do tamanho N da população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.



Fonte: Elaborado pelo autor (2020)

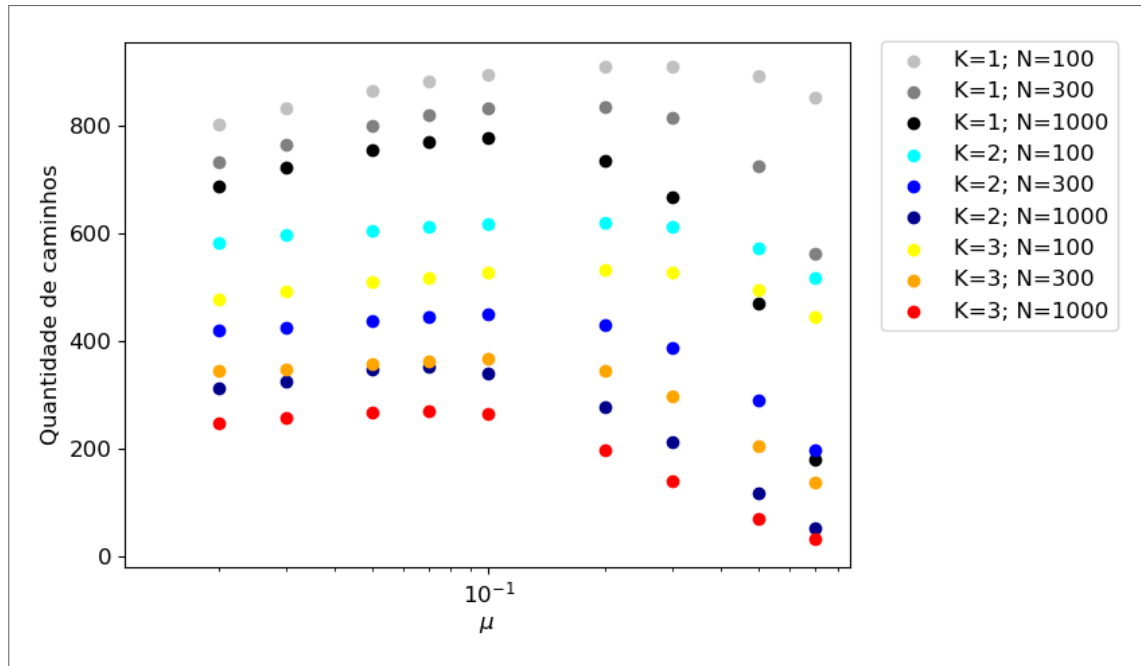
De maneira geral os resultados acima confirmaram o esperado efeito determinístico do aumento da população. Para todas as combinações possíveis de μ e K , as medidas sempre se comportaram monotonicamente e indicando um aumento da previsibilidade com a população. Entretanto, é possível notar também que todas as curvas parecem saturar na direção de N crescente. Na verdade, a presença de um valor mínimo ou máximo e a subsequente inversão da monotonicidade para valores de N imediatamente além dos simulados também é uma possibilidade assim como a assíntota. De fato, este seria o resultado mais esperado em vista do

que foi discutido na Seção 3.2.2, onde esclarecemos como a transição de $N\mu = 1$ para $N\mu^2 = 1$ leva de um aumento a uma diminuição da previsibilidade com o tamanho da população. Para os nossos parâmetros, no entanto, nenhuma população alcançou $N\mu^2 = 1$. A investigação desse fenômeno exige claramente a simulação de populações maiores (SZENDRO et al., 2013a).

4.2.2 Taxa de mutação

Procuramos agora explorar a dependência das mesmas medidas com a taxa de mutação. Utilizamos o conjunto $\mu \in \{0,02; 0,03; 0,05; 0,07; 0,10; 0,20; 0,30; 0,50; 0,70\}$ com nove valores da taxa para três valores diferentes $N \in \{100, 300, 1000\}$ de tamanho de população.

Figura 29 – Quantidade de caminhos em função da taxa μ de mutação para diferentes valores de tamanho N de população e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.

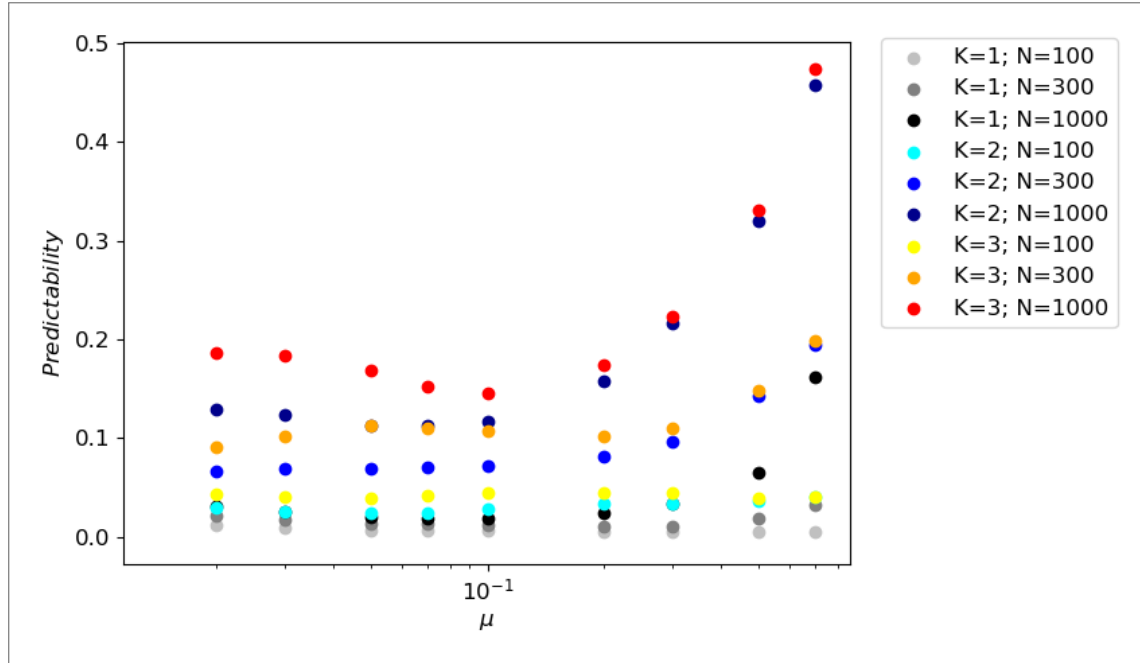


Fonte: Elaborado pelo autor (2020)

Começando pela quantidade de caminhos distintos, retratada na Figura 29, observa-se um interessante comportamento não monotônico em μ compartilhado por todos os valores de N e K . Na discussão anterior sobre esta mesma medida já havíamos identificado na Figura 25 um aumento com a taxa de mutação para os valores $\{0,02; 0,03; 0,05\}$. Tal aumento foi associado à instabilidade e diversidade causada pela introdução de mais mutantes com população fixa. No entanto, para valores maiores que 0,1 a quantidade de caminhos passa a diminuir com a taxa de mutação, muito embora o fator caótico continue aumentando. A causa deste fenômeno é a aceleração da busca pelo máximo global através do aumento do número de saltos, isto é, passos não unitários. A tendência é que por puro acaso a presença generalizada e constante de mutações provoque uma busca mais efetiva do máximo global,

com caminhadas menores e, portanto, menos possibilidades de trajetórias. É importante notar nesse ponto que essa efetividade tem pouco significado, uma vez que com taxas de mutação altas assim fica cada vez mais difícil manter o máximo global ou qualquer outra sequência por um tempo significativo, um fenômeno conhecido como “limite de erro”.

Figura 30 – *Predictability* em função da taxa μ de mutação para diferentes valores de tamanho N de população e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.

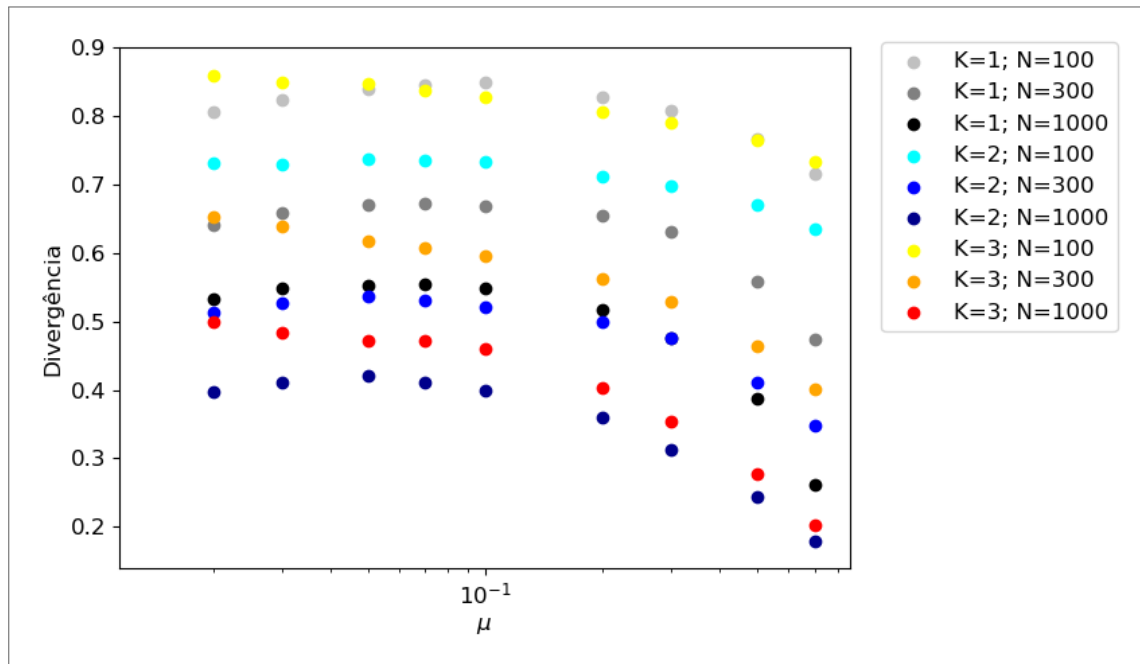


Fonte: Elaborado pelo autor (2020)

A Figura 30 explora a dependência da *predictability* com μ . Da mesma forma que a quantidade de caminhos, a *predictability* apresenta comportamento não monotônico. No caso da *predictability*, no entanto, há o agravante de diferentes valores de K e N apresentarem dependências diferentes com a taxa de mutação. O aumento da *predictability* para valores de μ acima de 0,1 parece unanimidade, acompanhando a diminuição dos números absoluto e efetivo de caminhos. Na região anterior a 0,1 algumas curvas tendem a apresentar um primeiro aumento e subsequente diminuição, enquanto outras apresentam apenas o decrescimento. É difícil identificar com precisão a origem destas diferenças. O que se pode afirmar é que, para todas as três populações simuladas o menor valor de μ simulado (0,02) está associado a um produto $N\mu$ igual a 2, 6 e 20, respectivamente. Por outro lado, todas atingem $N\mu^2$ em algum momento. Nesse cenário já havíamos explicado na Seção 3.2.2 como o aumento de $N\mu$ em torno de 1 está associado à produção de primeiros vizinhos e aumento do determinismo, enquanto a aproximação de $N\mu^2$ diminui o determinismo pelo aumento caótico da diversidade. Além disso, vimos que o posterior aumento leva à otimização da busca pelo máximo global e um novo aumento da previsibilidade. Em resumo, se cada um destes 3 regimes ocorresse com pouca superposição, o comportamento esperado se aproximaria do observado para a curva

laranja, com $K = 3$ e $N = 300$. Entretanto, a possibilidade de conflito e supressão entre esses regimes é bastante plausível e, de acordo com nossos dados, não fica claro o que determina a prevalência de cada um.

Figura 31 – Divergência média de caminhos em função do tamanho N de população para diferentes valores da taxa μ de mutação e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.



Fonte: Elaborado pelo autor (2020)

Com a divergência, mostrada na Figura 31, a situação é igualmente confusa. Enquanto as curvas para $K = 1$ e $K = 2$ apresentam um comportamento equivalente ao da quantidade de caminhos, isto é, um crescimento seguido de um decrescimento com μ , nos relevos mais acidentados de $K = 3$ a divergência é curiosamente monotônica, a despeito de todos os efeitos e regimes citados. Tal diferença entre os valores de K é marcante, pois ocorre para taxas de mutação baixas, aquelas de maior interesse biológico. A simulação de menores ordens de grandeza em μ se mostra fundamental para um melhor entendimento do problema.

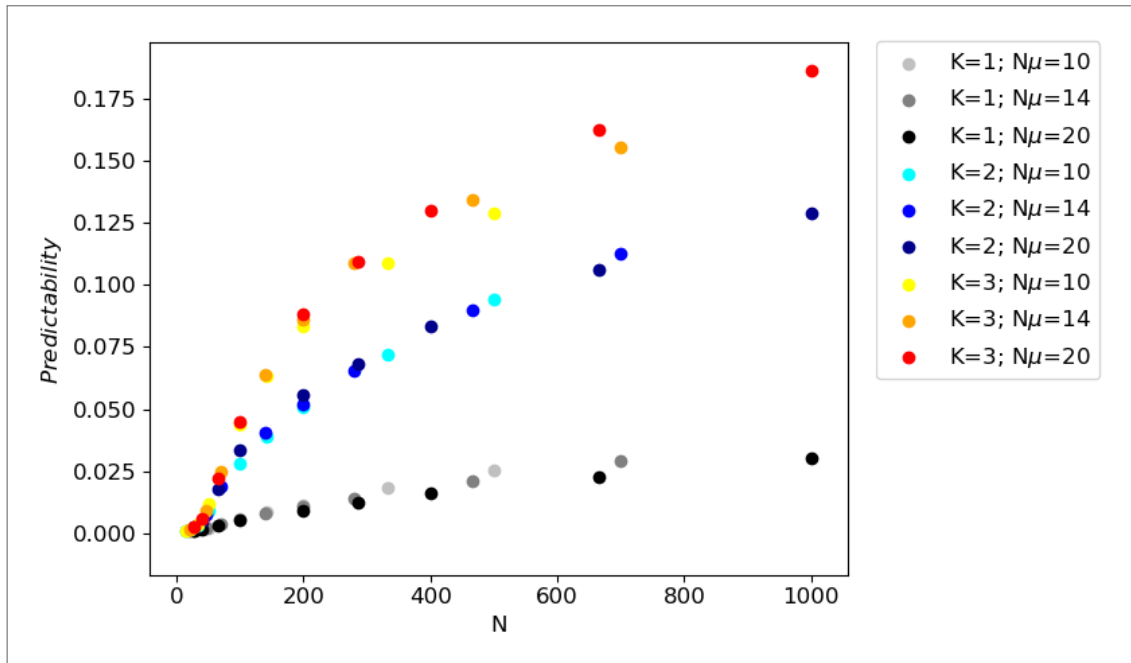
Se por um lado a dependência de medidas de previsibilidade com o tamanho da população teve comportamentos e causas aparentemente claros, o mesmo não se observou em relação à taxa de mutação. De fato, N esteve sempre associado a um aumento da previsibilidade para todos os valores e intervalos de parâmetros, condizente com a diminuição do fator estocástico. Com a taxa de mutação por outro lado houve um comportamento muito mais complexo. Como já dissemos, as causas desse comportamento não são claras, mas é evidente que a diferença por trás das relações com N ou μ reside em como ambos aumentam o número de mutantes, porém N em valor absoluto e μ em valor absoluto e relativo. Qualquer explicação para a dependência da previsibilidade com μ deve então ser explorada sob esta ótica. A implementação de medidas para distinção dos diferentes regimes também se faz necessária. Mas fica evidente também a

existência de uma mutação crítica delimitando a ocorrência do aumento de determinismo com o fluxo de mutantes.

4.2.3 Fluxo de mutantes fixo

Apresentamos brevemente os resultados das mesmas dinâmicas populacionais anteriores, tanto em função de N quanto μ , porém desta vez mantemos o número esperado de mutantes fixo. Nos concentramos na *predictability* e divergência. Analisamos três valores de $N\mu \in \{10, 14, 20\}$. Mantendo a ordem das duas subseções anteriores, as Figuras 32 e 33 trazem *predictability* e divergência em função de N .

Figura 32 – *Predictability* em função do tamanho N de população para diferentes valores do número esperado $N\mu$ de mutantes e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.

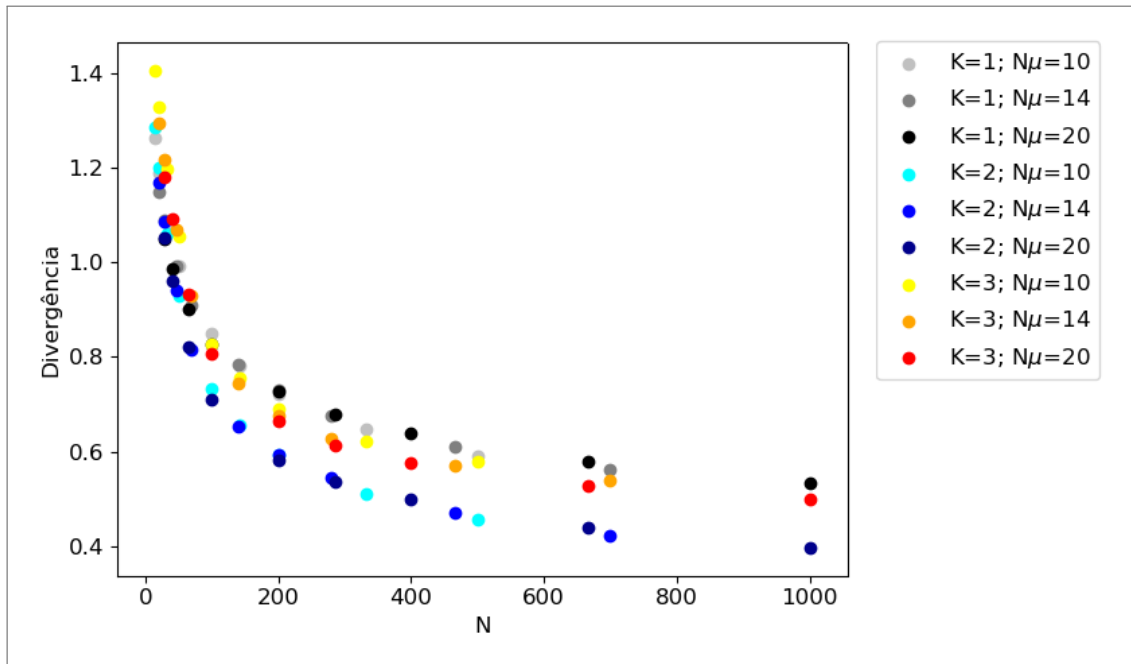


Fonte: Elaborado pelo autor (2020)

Ambas as medidas apresentam comportamento monotônico em N , sendo a *predictability* crescente e a divergência decrescente. Assim como nas Figuras 26 e 27 podemos observar que a *predictability* é crescente com K enquanto a divergência cai de $K = 1$ para $K = 2$, porém aumenta de $K = 2$ para $K = 3$. Isso evidencia mais uma vez que embora o número efetivo de caminhos diminua com a intensificação da epistasia, isso não implica necessariamente numa maior previsibilidade, uma vez que os caminhos podem ter se tornado mais divergentes, como de fato observamos.

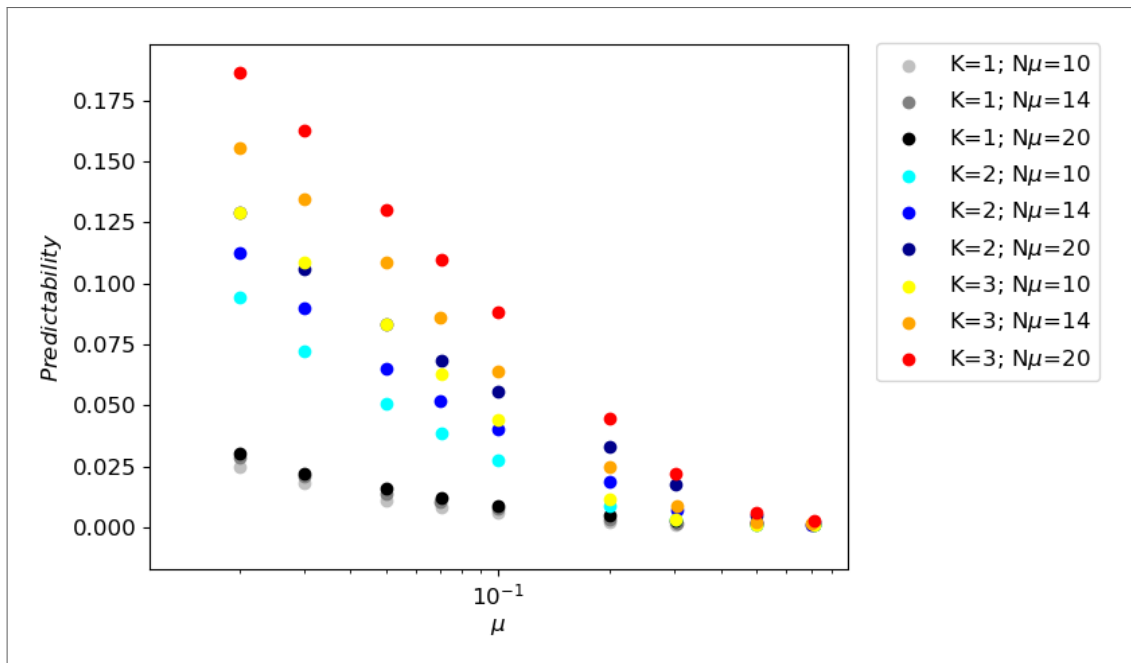
Nas Figuras 34 e 35, as duas medidas estão em função da taxa de mutação para valores fixos de $N\mu$, e mais uma vez nota-se a relação anterior com K . Além disso, o comportamento

Figura 33 – Divergência média de caminhos em função do tamanho N de população para diferentes valores do número esperado $N\mu$ de mutantes e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.



Fonte: Elaborado pelo autor (2020)

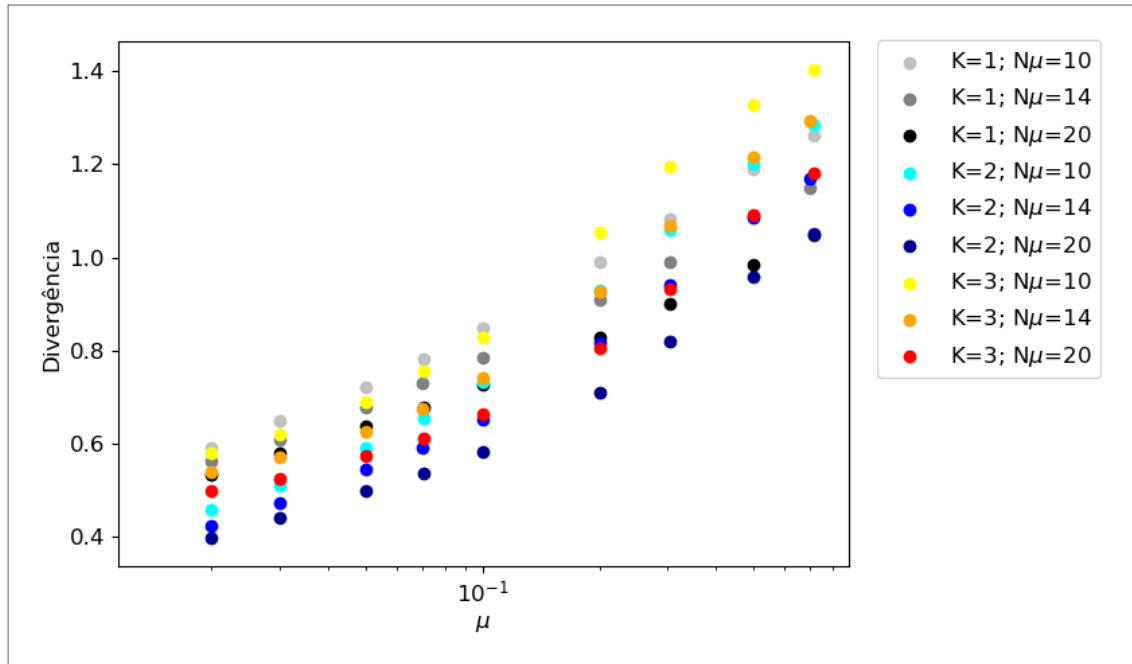
Figura 34 – *Predictability* em função da taxa μ de mutação para diferentes valores do número esperado $N\mu$ de mutantes e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.



Fonte: Elaborado pelo autor (2020)

de ambas as quantidades é monotônico com μ , diferentemente das Figuras 30 e 31, também em função de μ , porém com N fixo.

Figura 35 – Divergência média de caminhos em função da taxa μ de mutação para diferentes valores do número esperado $N\mu$ de mutantes e do parâmetro K , de acordo com a legenda. Cada ponto representa o valor médio sobre 50 relevos de tamanho $L = 8$.



Fonte: Elaborado pelo autor (2020)

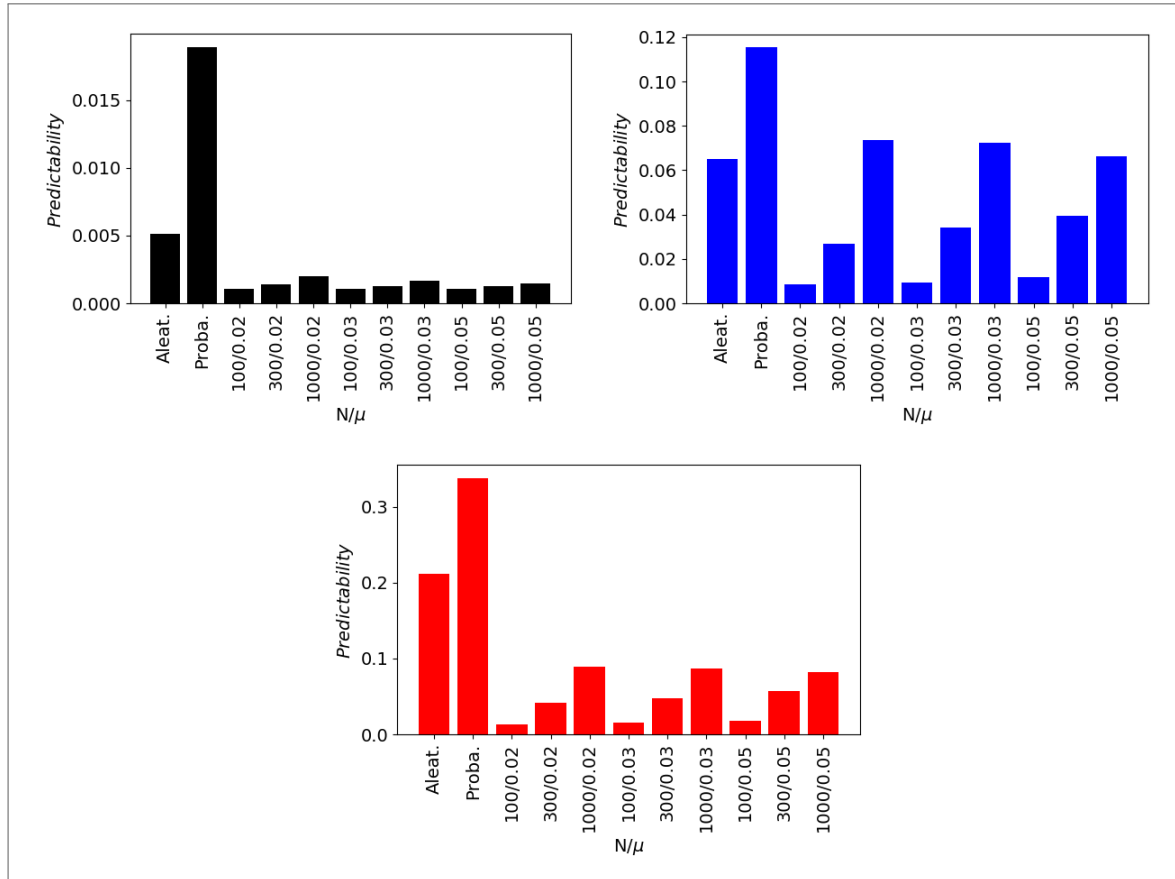
4.2.4 Caminhadas adaptativas VS dinâmicas populacionais

Aqui buscamos comparar medidas de previsibilidade entre a dinâmica de população finita e caminhada adaptativa, lembrando que a última corresponde ao limite de baixas mutações da primeira. O limite de baixa mutação pode ocorrer ainda sob seleção fraca ou forte, associadas respectivamente aos tipos aleatório (Equação 3.4) e probabilístico (Equação 3.5) de caminhada.

Embora estejam relacionadas por um limite, caminhadas adaptativas e dinâmicas de população finita são fundamentalmente diferentes. Além disso, nas nossas simulações em particular, as populações concluem suas caminhadas exclusivamente sobre o máximo global. Assim, a comparação com caminhadas adaptativas deve se restringir à parcela daquelas que atingem esse mesmo máximo. Por esse motivo, as simulações desta seção foram obtidas sob um conjunto particular de relevos, aqueles onde o máximo global é acessível sob caminhadas adaptativas. Em seguida selecionamos apenas o conjunto de caminhadas que acessam este máximo, enquanto as caminhadas adaptativas direcionadas a máximos locais são irrelevantes e foram descartadas. Nos concentramos essencialmente na *predictability* e divergência média de caminhos, apresentadas nas Figuras 36 e 37, respectivamente. Os gráficos trazem os resultados de médias sobre 10 relevos para $L = 8$ e $K \in \{1, 2, 3\}$, nas cores preta, azul e vermelha, em ordem crescente.

Na Figura 36 a característica mais marcante é a superioridade absoluta das caminhadas

Figura 36 – *Predictability* sob diferentes dinâmicas adaptativas. As duas primeiras barras à esquerda de cada gráfico mostram os valores referentes às caminhadas adaptativas aleatória e probabilística, nesta ordem. O restante das barras está associado às dinâmicas de população finita para diversos valores de N e μ indicados no eixo horizontal. Preto, azul e vermelho representam K igual a 1, 2 e 3, respectivamente. Cada barra representa uma média sobre 10 relevos de tamanho $L = 8$. Todas as dinâmicas partem do antípoda.

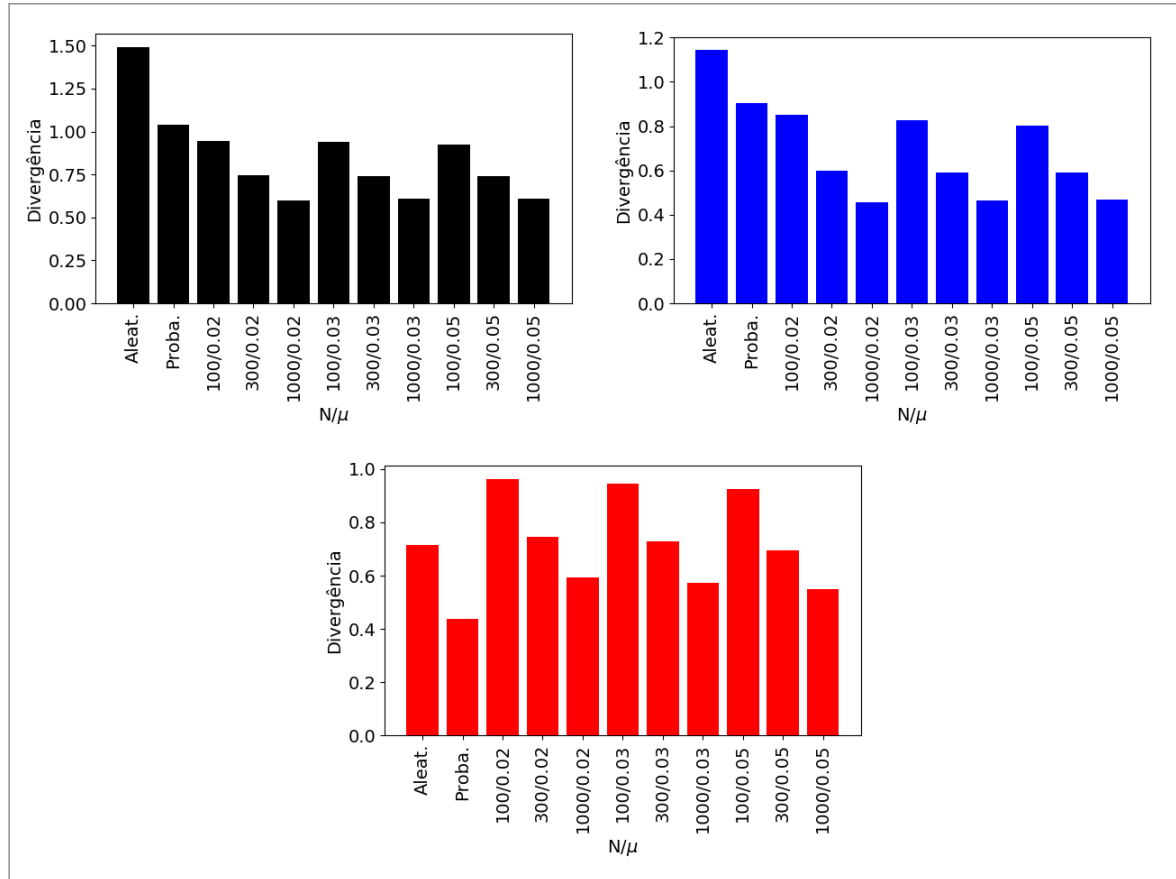


Fonte: Elaborado pelo autor (2020)

probabilísticas, sugerindo assim que o limite SSWM de seleção forte e mutação fraca é a situação mais previsível dentre todas. Sem dúvida, o fato de caminhadas adaptativas experimentarem apenas caminhos monotonicamente crescentes em fitness representa uma grande vantagem em termos de previsibilidade. A gama de possibilidades extra permitidas às populações finitas tanto pela adoção de trajetórias não monotônicas como de passos não unitários tendem a torná-las mais diversas e imprevisíveis. Dessa forma, até mesmo a caminhada aleatória foi superior às populações para $K = 1$ e $K = 3$. No caso $K = 2$, entretanto, o efeito imprevisível de uma seleção fraca sobressaiu esse fator, e a caminhada adaptativa aleatória teve uma *predictability* comparável às das populações com $N = 1000$. Não está claro porque isto ocorreu apenas para $K = 2$.

Na Figura 37 pode-se observar que para populações finitas a divergência se opõe à *predictability*, como esperado. De fato, é possível notar que valores mais altos de *predictability* estão associados a valores mais baixos de divergência, e vice-versa. O comportamento das caminhadas, no entanto, é bem mais intrigante. Mesmo possuindo os maiores valores de *predictability*

Figura 37 – Divergência média de caminhos sob diferentes dinâmicas adaptativas. As duas primeiras barras à esquerda de cada gráfico mostram os valores referentes às caminhadas adaptativas aleatória e probabilística, nesta ordem. O restante das barras está associado às dinâmicas de população finita para diversos valores de N e μ indicados no eixo horizontal. Preto, azul e vermelho representam K igual a 1, 2 e 3, respectivamente. Cada barra representa uma média sobre 10 relevos de tamanho $L = 8$. Todas as dinâmicas partem do antípoda.



Fonte: Elaborado pelo autor (2020)

para todos os graus de epistasia considerados, as caminhadas adaptativas probabilísticas também apresentaram valores altos de divergência para $K = 1$ e $K = 2$, perdendo apenas para as caminhadas aleatórias. Em suma, nestes dois casos a caminhada probabilística percorre efetivamente poucos caminhos, ao mesmo tempo em que estes caminhos compartilham pouca similaridade. No caso $K = 3$ a situação é mais equilibrada. A caminha probabilística tem a menor divergência, fortalecendo conclusões de seções anteriores sobre o efeito restritivo da rugosidade sob os tipos de caminhadas, isto é, em baixa quantidade e confinadas a regiões específicas do relevo. Os maiores valores foram obtidos pelas populações finitas de tamanho $N = 100$. Pelas habilidades de salto e passo não monotônicos já mencionadas, as populações finitas não sofrem uma restrição tão forte de caminhos permitidos quanto as caminhadas adaptativas.

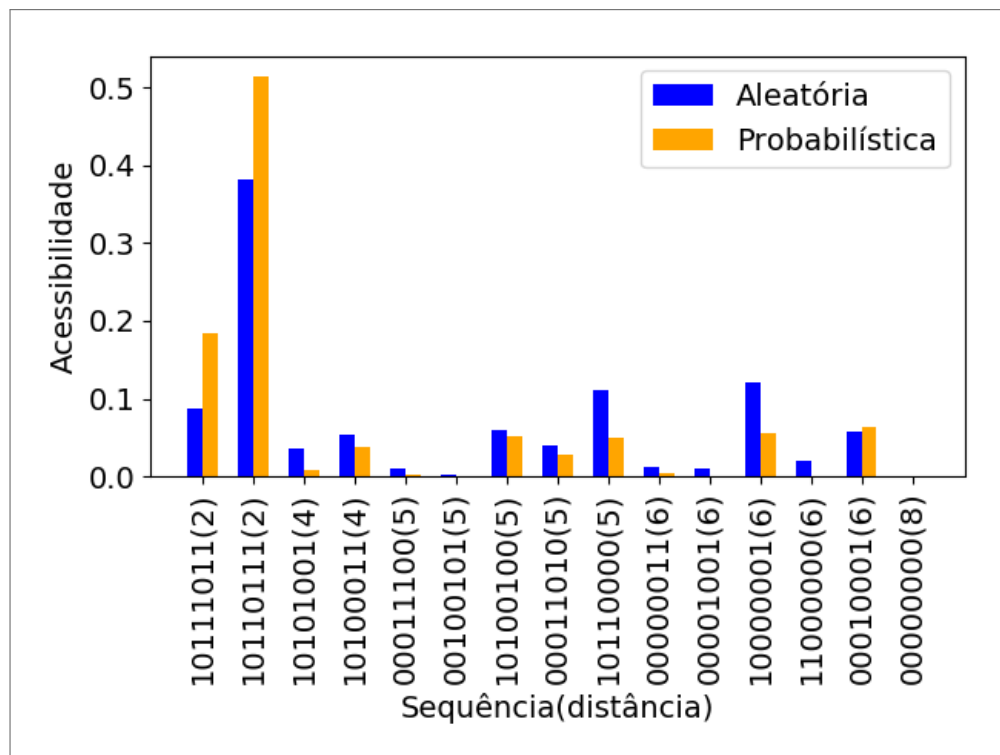
4.3 RELEVO EXPERIMENTAL: *ASPERGILLUS NIGER*

Nesta seção abandonamos o Modelo NK de relevos teóricos e passamos a explorar as mesmas medidas e relações das seções anteriores no contexto do relevo experimental do fungo filamentoso *Aspergillus niger*. Trata-se de um relevo de dimensão $L = 8$ onde a sequência selvagem (00000000) é a mais apta, e cujo método de obtenção é detalhado por VISSER; HOEKSTRA; ENDE. Das 256 combinações possíveis apenas 186 são viáveis, isto é, possuem fitness não nulo (FRANKE et al., 2011). O antípoda, em particular, também pertence ao conjunto das sequências inviáveis. Nas simulações, no entanto, não há grandes problemas em usá-lo como condição inicial. De fato, tanto nas caminhadas adaptativas como nas dinâmicas populacionais as primeiras sequências vizinhas e viáveis do antípoda tem necessariamente de surgir por mutação, mesmo que o antípoda tenha fitness não nulo.

4.3.1 Caminhadas adaptativas

Começamos pela análise comparativa de ensembles de 1.000.000 caminhadas adaptativas aleatórias e probabilísticas que atingem diferentes máximos do relevo. A primeira medida é a acessibilidade (ver Figura 38) onde os máximos são ordenados por distância ao antípoda e, quando degenerados neste critério, ordenados em fitness.

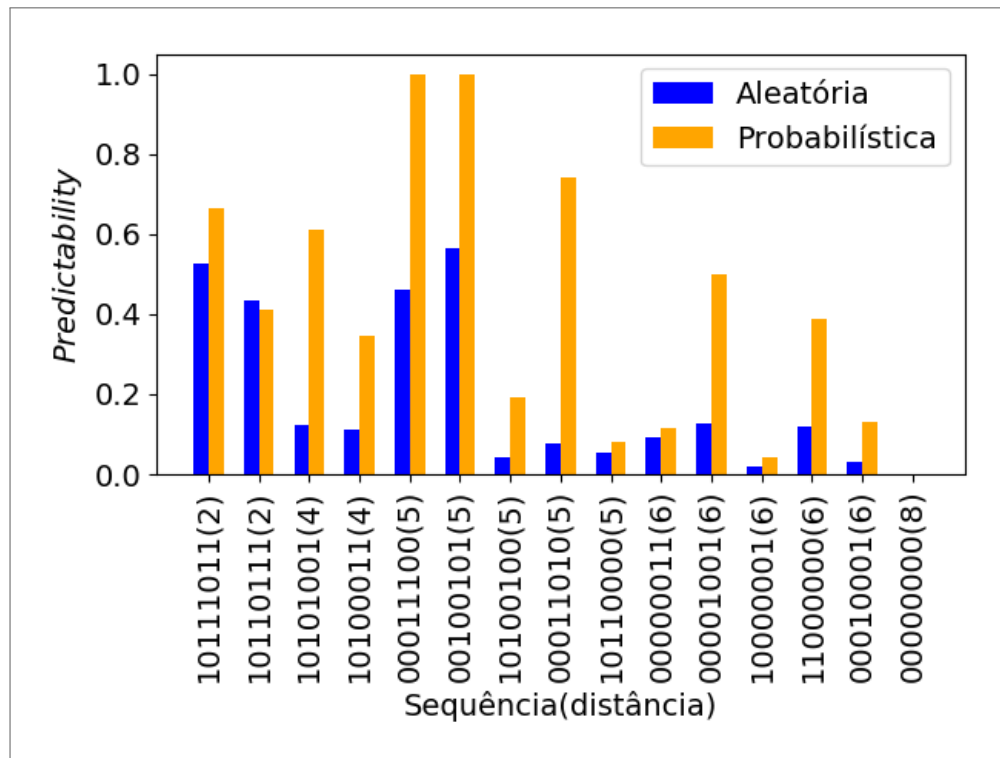
Figura 38 – Acessibilidade por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 1.000.000 caminhadas no relevo experimental do fungo *Aspergillus niger*.



Fonte: Elaborado pelo autor (2020)

À primeira vista a Figura 38 já evidencia a inacessibilidade do máximo global sob qualquer um dos tipos de caminhadas adaptativas. Todos os 15 máximos locais são acessíveis por ambas as caminhadas. Apesar de um dos dois máximos locais mais próximos do antípoda acumular em torno de 40% a 50% da acessibilidade total do relevo, a acessibilidade dos demais máximos parece não estar correlacionada com a distância ao ponto de partida, ao contrário do que observamos para o Modelo NK na Figura 20. Entre máximos equidistantes do antípoda, o ordenamento crescente em fitness sugere uma correlação positiva entre acessibilidade e aptidão, fato não observado no Modelo NK. Vale notar também que a quantidade de máximos deste relevo sugere um alto grau de epistasia, uma vez que podemos observar na Figura 16 que o maior grau de epistasia simulado ($K = 3$) para um relevo do Modelo NK de tamanho $L = 8$ tem um valor esperado de máximos locais em torno de 8 picos.

Figura 39 – *Predictability* por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 1.000.000 caminhadas no relevo experimental do fungo *Aspergillus niger*.

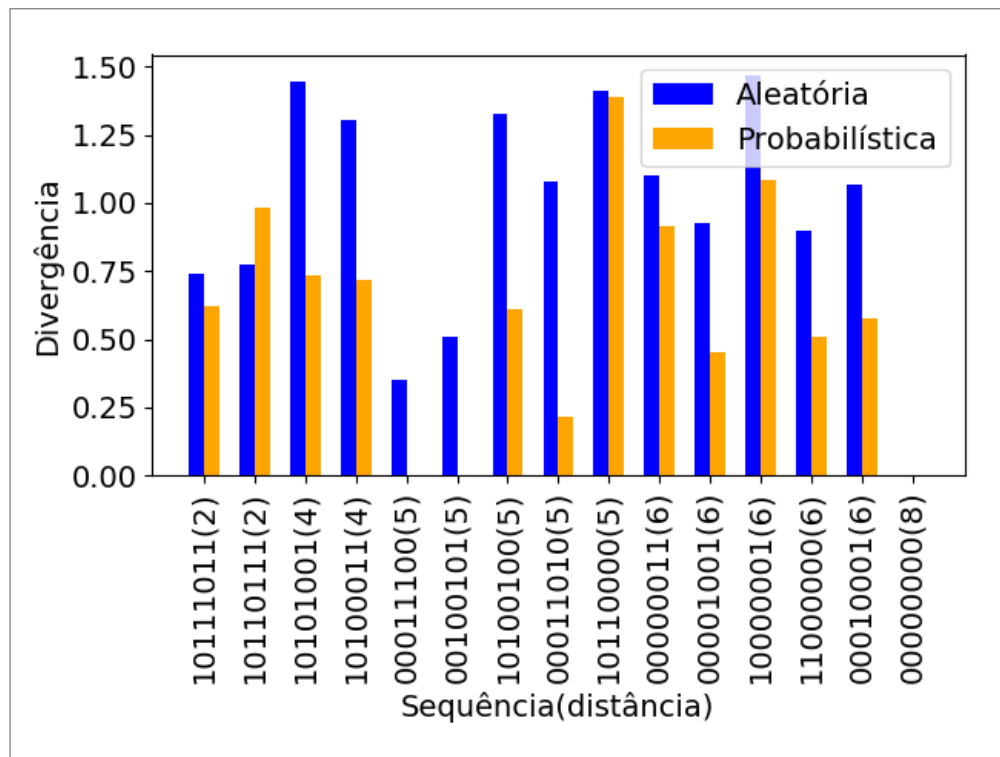


Fonte: Elaborado pelo autor (2020)

A *predictability* (ver Figura 39) é absolutamente maior sob caminhadas probabilísticas, como esperado. Não há correlação aparente com acessibilidade ou fitness, enquanto a distância ao antípoda está correlacionada negativamente de maneira sutil, sendo mais expressiva sob caminhadas aleatórias. De fato, há uma discrepância marcante entre a *predictability* sob os dois tipos de caminhadas para os máximos mais distantes. Esta observação pode ser melhor interpretada considerando-se os resultados anteriores, quando mostramos como o número de caminhos possíveis cresce com a distância ao antípoda. Enquanto sob caminhadas aleatórias estes trajetos tendem a ser igualmente prováveis e, devido à sua abundância, resultar numa

predictability baixa, sob passos probabilísticos, por outro lado, a distribuição de probabilidade destes mesmos trajetos tende a ficar mais tendenciosa a cada passo, de forma que poucos caminhos são efetivamente acessados, causando uma alta *predictability* para máximos longínquos. Dois dos caminhos apresentam *predictability* igual a 1, fruto da existência de um único caminho. Nesses casos a divergência é necessariamente nula, como observado na Figura 40.

Figura 40 – Divergência média de caminhos por máximo local em ordem crescente de distância ao antípoda, ponto de partida. Foram simuladas 1.000.000 caminhadas no relevo experimental do fungo *Aspergillus niger*.



Fonte: Elaborado pelo autor (2020)

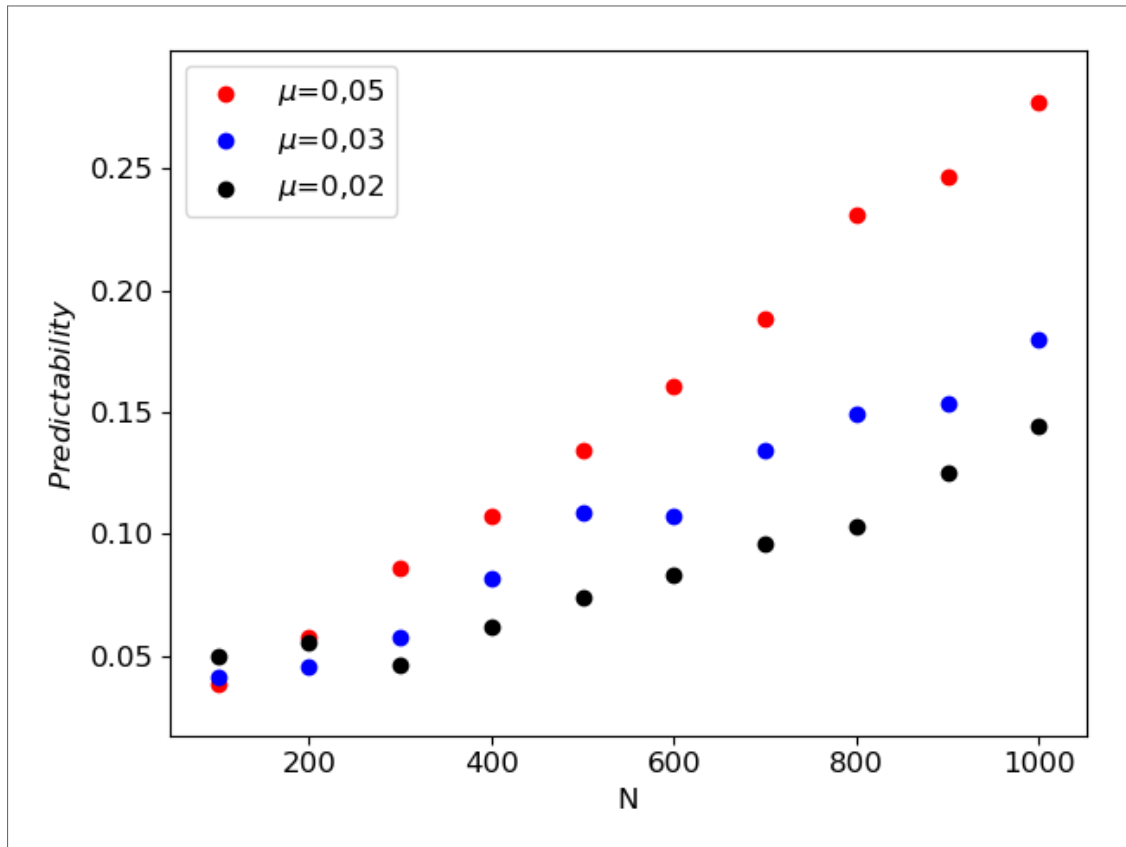
A divergência média de caminhos apresenta valores semelhantes ao caso $K = 3$ do Modelo NK, porém com um comportamento bastante complexo, sem indicação de correlação com qualquer uma das quantidades já mencionadas, até mesmo com a *predictability*, em relação a qual se espera uma oposição. Uma provável causa desse comportamento é a alta letalidade do relevo, isto é, a inviabilidade de 70 dos 256 genótipos, dado que esta medida é sensivelmente afetada pela estrutura dos caminhos. A sugestão é que a presença de genótipos de fitness nulo torne as trajetórias nesse relevo experimental particularmente únicas em relação a relevos completos como os do Modelo NK.

4.3.2 Dinâmica populacional

Uma vez exploradas as caminhadas adaptativas, nos resta agora entender aspectos da dinâmica populacional sobre o relevo experimental em questão. Os resultados envolvem a

simulação de 1.000 caminhadas independentes. Novamente, focamos nas duas principais medidas de previsibilidade aplicadas ao longo do trabalho, *predictability* e divergência média de caminhos. Começamos pela análise da dependência destas quantidades com o tamanho de população N e, em seguida, com a taxa de mutação μ .

Figura 41 – *Predictability* em função do tamanho de população N para diferentes valores da taxa de mutação μ . Foram simuladas 1.000 caminhadas partindo do antípoda no relevo experimental do fungo *Aspergillus niger*.

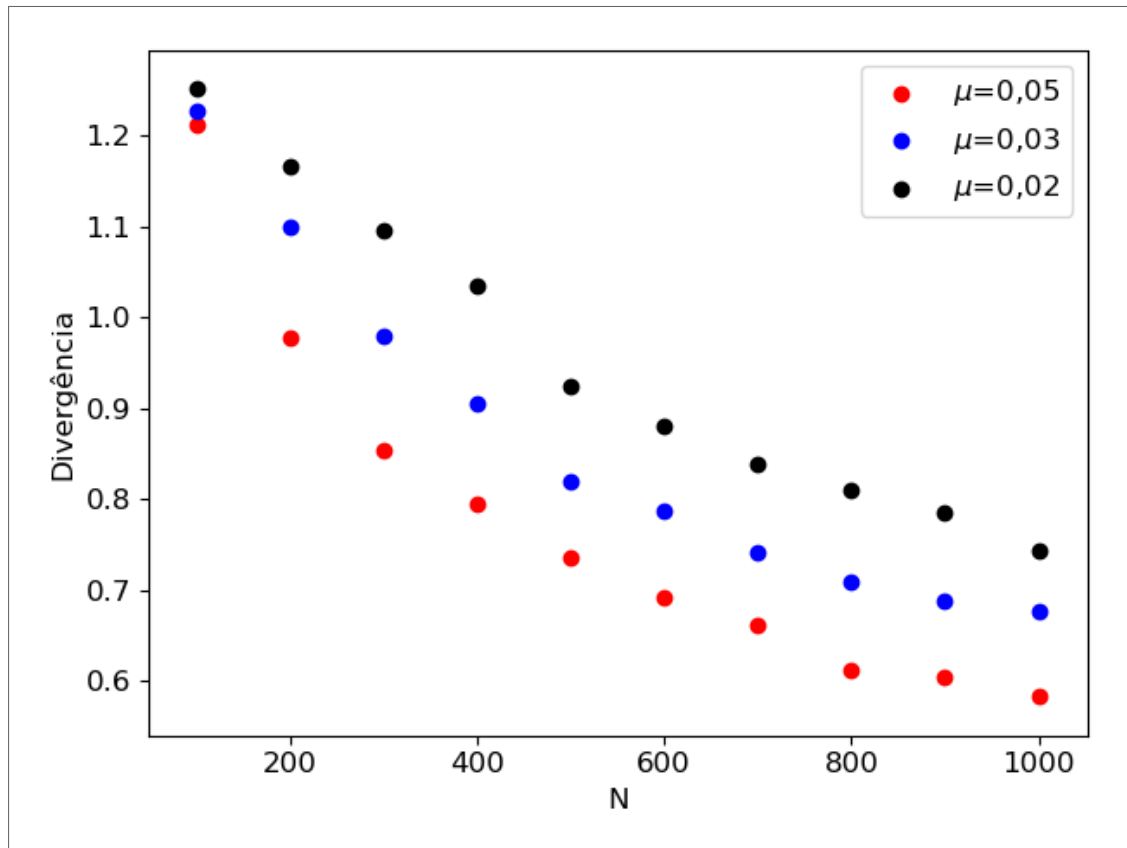


Fonte: Elaborado pelo autor (2020)

Na Figura 41 a *predictability* está em função de N . Quando comparada ao gráfico da Figura 26, o intervalo de valores da *predictability* no relevo experimental coincide com o caso $K = 3$ do Modelo NK, e também é monotonicamente crescente. A diferença que chama atenção é que enquanto no relevo teórico a *predictability* apresenta concavidade negativa para os três valores de μ escolhidos, isto é, crescendo desaceleradamente com N , para o relevo real o crescimento é acelerado no intervalo de 100 a 300 indivíduos, e aproximadamente linear a partir de então. De fato, na Figura 26 as curvas parecem atingir um mínimo ou assíntota horizontal em torno de $N = 1000$. Como discutimos antes, a expectativa é de que seja de fato um mínimo e a *predictability* passe a diminuir com N , fruto da aproximação do parâmetro $N\mu^2$ de 1, associado ao regime de produção de segundos vizinhos e uma consequente perda de previsibilidade. No relevo experimental, no entanto, não há qualquer sinal de que a *predictability* deve atingir um valor máximo. Esse resultado alerta para o fato de que a superposição e concorrência dos

diferentes regimes em dinâmicas populacionais pode depender sensivelmente do tipo de relevo.

Figura 42 – Divergência média de caminhos em função do tamanho de população N para diferentes valores da taxa de mutação μ . Foram simuladas 1.000 caminhadas partindo do antípoda no relevo experimental do fungo *Aspergillus niger*.



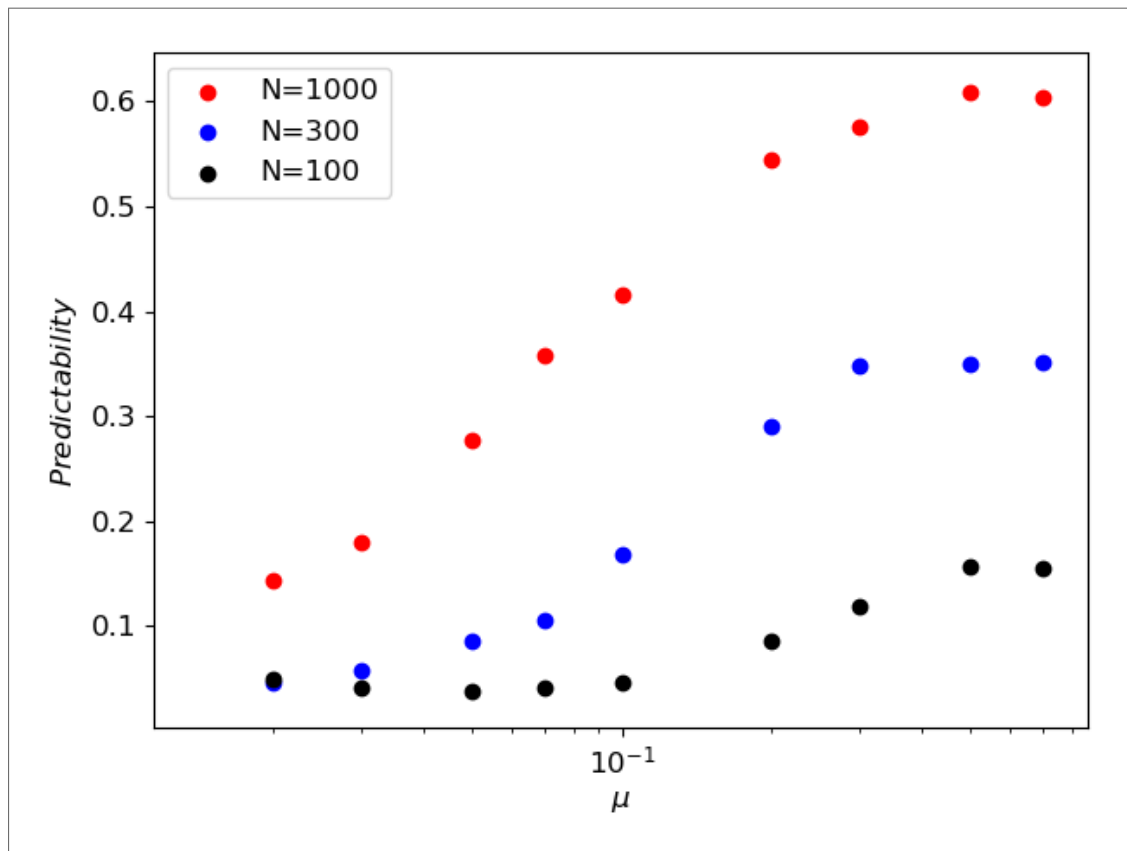
Fonte: Elaborado pelo autor (2020)

A divergência média de caminhos apresenta comportamento muito similar tanto no modelo teórico quanto experimental, como mostrado na Figura 42. Porém, assim como na *predictability*, o relevo experimental parece não atingir um valor extremo para $N = 1000$. Mais uma vez as curvas se aproximam mais do caso $K = 3$, único no qual as curvas caem com μ (Figura 27). Os valores, no entanto, são ligeiramente maiores no relevo experimental.

A Figura 43 traz novamente a *predictability*, desta vez em função da taxa de mutação. Em relação aos resultados do Modelo NK (Figura 30), a única curva que apresenta comportamento similar é aquela para $N = 100$. Esta mesma curva também é a única entre as três que mostra um decrescimento inicial da previsibilidade. A medida que μ aumenta a *predictability* passa a crescer, assim como nas demais curvas que não experimentaram o efeito anterior. Mais uma vez fazemos a ressalva de que taxas de mutação altas demais tem pouco significado biológico, mesmo para microrganismos e vírus, assim como essa efetividade é falsa, uma vez que a população é incapaz de manter qualquer sequência mestra de maneira estável.

Por último, exibimos na Figura 44 a divergência média de caminhos como função da taxa de mutação. Mais uma vez as curvas se comportam como aquelas para $K = 3$ no gráfico

Figura 43 – *Predictability* em função da taxa de mutação μ para diferentes valores de tamanho de população N . Foram simuladas 1.000 caminhadas partindo do antípoda no relevo experimental do fungo *Aspergillus niger*.

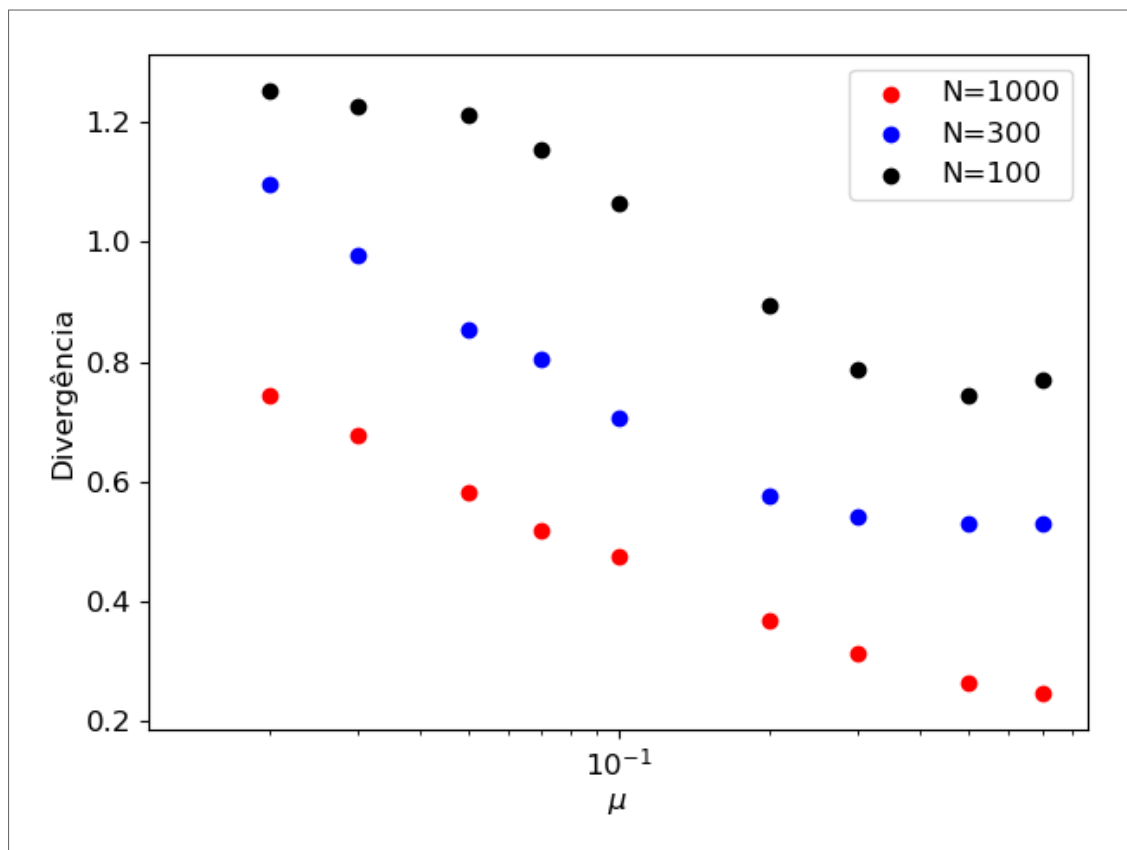


Fonte: Elaborado pelo autor (2020)

equivalente do Modelo NK (Figura 31), as únicas monotonicamente decrescentes em fitness. O decaimento inicial, no entanto, é bem mais intenso no relevo experimental. Assim como a *predictability*, os valores da divergência são maiores que o valor médio do relevos teóricos com $K = 3$.

A série de resultados acima sobre o relevo experimental do fungo *Aspergillus niger* quando comparada aos respectivos resultados sob Modelo NK sugerem uma maior similaridade do primeiro com o nível mais alto de epistasia simulado neste trabalho, $K = 3$. Obviamente é importante investigar níveis maiores de epistasia, principalmente levando em conta que o número de picos foi muito maior do que o valor esperado para $K = 3$. Outra importante consideração é que uma comparação justa e efetiva com o Modelo NK exige a implementação de sequências inviáveis no mesmo, uma espécie de taxa de letalidade, uma vez que no contexto explorado aqui é difícil identificar se as diferenças entre Modelo NK e relevo experimental são fruto das particularidades de um relevo real ou simplesmente da presença de sequências de fitness nulo (SZENDRO et al., 2013b).

Figura 44 – Divergência média de caminhos em função da taxa de mutação μ para diferentes valores de tamanho de população N . Foram simuladas 1.000 caminhadas partindo do antípoda no relevo experimental do fungo *Aspergillus niger*.



Fonte: Elaborado pelo autor (2020)

5 CONCLUSÃO

Os primeiros resultados analisados neste trabalho envolvem o comportamento de diversas medidas de acessibilidade, previsibilidade e topografia no Modelo NK. Dado o caráter estocástico, exploramos um comportamento médio sobre diversas realizações do modelo como função do comprimento L das sequências e do parâmetro K de intensidade das interações epistáticas. Embora as medidas em questão tenham diferentes significados, invariavelmente a explicação de seus comportamentos remete em última instância a duas quantidades: número e densidade de picos. Tais quantidades, por sua vez, estão intimamente relacionadas com o nível de frustração do relevo, isto é, a epistasia. Assim, o fenômeno epistático apresenta caráter fundamental na determinação dos caminhos tomados pela evolução. O outro resultado relevante da Seção 4.1 foi a observação de que o aumento de K provoca o acesso de um maior número de máximos ao mesmo tempo em que diminui o número de caminhos. Em outras palavras, o aumento da frustração no relevo leva, contraditoriamente, a uma imprevisibilidade do resultado final e uma previsibilidade do caminho tomado até este resultado. A análise de relevos individuais nesta mesma seção apresentou pela primeira vez no trabalho a aplicação da divergência média de caminhos. Comparando número de caminhos, *predictability* e divergência, observou-se que tais medidas tendem a ser redundantes em relevos suaves, enquanto o aumento da rugosidade mostra a necessidade de contabilização tanto da distribuição de probabilidade dos caminhos quanto de aspectos estruturais destes. Este mesmo resultado também foi observado nas dinâmicas populacionais da Seção 4.2, e apontam para uma maior complexidade de efeitos na previsibilidade evolutiva com a intensificação da epistasia.

Ainda na Seção 4.2 sobre dinâmicas de população finita, a dependência das medidas de repetibilidade apresentam comportamento monotônico com o tamanho de população, enquanto a dependência com a taxa de mutação se mostrou muito mais intrincada. A saturação das curvas em função de N indica um provável extremo das medidas para valores de tamanho de população mais altos, condizente com a esperada diminuição de previsibilidade provocada pela abundância de mutações duplas. A investigação desse fenômeno exige a simulação de populações maiores, pelo menos da ordem de 10.000. Similarmente, resultados realmente relevantes no contexto biológico estão associados a taxas de mutações menores que as exploradas aqui, pelo menos da ordem de 10^{-3} . Um outro importante resultado associado às populações finitas é fruto da comparação com caminhadas adaptativas, onde observou-se uma maior previsibilidade das caminhadas adaptativas probabilísticas em todos os cenários, indicando que o regime de mutação fraca e seleção forte é a situação mais previsível de todas. A observação de um conflito entre diminuição do número de caminhos e aumento da dissimilaridade entre eles em função de K foi, sem dúvida, o dado mais interessante. Uma investigação de valores de K maiores que 3 deve revelar mais informações.

Por fim, a análise de um relevo experimental apresentou grandes semelhanças com os resultados do Modelo NK, tanto em comportamento quanto em valor absoluto, muito embora

não compartilhem a presença da taxa de letalidade. De fato, a inclusão de indivíduos inviáveis no modelo teórico se faz necessária para uma comparação mais precisa.

REFERÊNCIAS

- AITA, T.; IWAKURA, M.; HUSIMI, Y. A cross-section of the fitness landscape of dihydrofolate reductase. *Protein engineering*, Oxford University Press, v. 14, n. 9, p. 633–638, 2001.
- CAMPOS, P. R.; ADAMI, C.; WILKE, C. O. Optimal adaptive performance and delocalization in nk fitness landscapes. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 304, n. 3-4, p. 495–506, 2002.
- COLEGRAVE, N.; BUCKLING, A. Microbial experiments on adaptive landscapes. *Bioessays*, Wiley Online Library, v. 27, n. 11, p. 1167–1173, 2005.
- DOMINGO, E.; SCHUSTER, P. What is a quasispecies? historical origins and current scope. In: *Quasispecies: From Theory to Experimental Systems*. [S.l.]: Springer, 2015. p. 1–22.
- DYKHUIZEN, D. E. Experimental evolution: replicating history. *Trends in ecology & evolution*, Elsevier Current Trends, v. 7, n. 8, p. 250–252, 1992.
- EIGEN, M.; MCCASKILL, J.; SCHUSTER, P. The molecular quasi-species. *Adv. Chem. Phys.*, v. 75, p. 149–263, 1989.
- FRANKE, J.; KLÖZER, A.; VISSER, J. A. G. de; KRUG, J. Evolutionary accessibility of mutational pathways. *PLoS Comput Biol*, Public Library of Science, v. 7, n. 8, p. e1002134, 2011.
- GERRISH, P. J.; LENSKI, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica*, Springer, v. 102, p. 127, 1998.
- GILLESPIE, J. H. A simple stochastic gene substitution model. *Theoretical population biology*, Elsevier, v. 23, n. 2, p. 202–215, 1983.
- GILLESPIE, J. H. Molecular evolution over the mutational landscape. *Evolution*, JSTOR, p. 1116–1129, 1984.
- GOULD, S. J. *Wonderful life: the Burgess Shale and the nature of history*. [S.l.]: WW Norton & Company, 1990.
- JAIN, K.; KRUG, J. Deterministic and stochastic regimes of asexual evolution on rugged fitness landscapes. *Genetics*, Genetics Soc America, v. 175, n. 3, p. 1275–1288, 2007.
- JAIN, K.; KRUG, J.; PARK, S.-C. Evolutionary advantage of small populations on complex fitness landscapes. *Evolution: International Journal of Organic Evolution*, Wiley Online Library, v. 65, n. 7, p. 1945–1955, 2011.
- KAUFFMAN, S. A.; WEINBERGER, E. D. The nk model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of theoretical biology*, Elsevier, v. 141, n. 2, p. 211–245, 1989.
- KAUL, H.; JACOBSON, S. H. Global optima results for the kauffman nk model. *Mathematical Programming*, Springer, v. 106, n. 2, p. 319–338, 2006.
- KORONA, R.; NAKATSU, C. H.; FORNEY, L. J.; LENSKI, R. E. Evidence for multiple adaptive peaks from populations of bacteria evolving in a structured habitat. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 91, n. 19, p. 9037–9041, 1994.

- LENSKI, R. E.; ROSE, M. R.; SIMPSON, S. C.; TADLER, S. C. Long-term experimental evolution in *Escherichia coli*. i. adaptation and divergence during 2,000 generations. *The American Naturalist*, University of Chicago Press, v. 138, n. 6, p. 1315–1341, 1991.
- LOBKOVSKY, A. E.; KOONIN, E. V. Replaying the tape of life: quantification of the predictability of evolution. *Frontiers in genetics*, Frontiers, v. 3, p. 246, 2012.
- LOBKOVSKY, A. E.; WOLF, Y. I.; KOONIN, E. V. Predictability of evolutionary trajectories in fitness landscapes. *PLoS Comput Biol*, Public Library of Science, v. 7, n. 12, p. e1002302, 2011.
- MANHART, M.; MOROZOV, A. V. Statistical physics of evolutionary trajectories on fitness landscapes. In: *First-passage phenomena and their applications*. [S.l.]: World Scientific, 2014. p. 416–446.
- MORRIS, S. C. The predictability of evolution: glimpses into a post-darwinian world. *Naturwissenschaften*, Springer, v. 96, n. 11, p. 1313–1337, 2009.
- MORRIS, S. C. Evolution: like any other science it is predictable. *Philosophical Transactions of the Royal Society B: Biological Sciences*, The Royal Society, v. 365, n. 1537, p. 133–145, 2010.
- NOWAK, M. A. What is a quasispecies? *Trends in ecology & evolution*, Elsevier, v. 7, n. 4, p. 118–121, 1992.
- NOWAK, M. A. *Evolutionary dynamics: exploring the equations of life*. [S.l.]: Harvard University Press, 2006.
- ORR, H. A. The population genetics of adaptation: the adaptation of dna sequences. *Evolution*, Wiley Online Library, v. 56, n. 7, p. 1317–1330, 2002.
- REIA, S. M.; CAMPOS, P. R. Analysis of statistical correlations between properties of adaptive walks in fitness landscapes. *Royal Society Open Science*, The Royal Society, v. 7, n. 1, p. 192118, 2020.
- ROY, S. W. Probing evolutionary repeatability: neutral and double changes and the predictability of evolutionary adaptation. *PLoS One*, Public Library of Science, v. 4, n. 2, p. e4500, 2009.
- SALISBURY, F. B. Natural selection and the complexity of the gene. *Nature*, Springer, v. 224, n. 5217, p. 342–343, 1969.
- SMITH, J. M. Natural selection and the concept of a protein space. *Nature*, Nature Publishing Group, v. 225, n. 5232, p. 563–564, 1970.
- SZENDRO, I. G.; FRANKE, J.; VISSER, J. A. G. de; KRUG, J. Predictability of evolution depends nonmonotonically on population size. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 110, n. 2, p. 571–576, 2013.
- SZENDRO, I. G.; SCHENK, M. F.; FRANKE, J.; KRUG, J.; VISSER, J. A. G. D. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2013, n. 01, p. P01005, 2013.

TEOTÓNIO, H.; ROSE, M. R. Perspective: reverse evolution. *Evolution*, Wiley Online Library, v. 55, n. 4, p. 653–660, 2001.

VISSER, J. A. G. D.; KRUG, J. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, Nature Publishing Group, v. 15, n. 7, p. 480–490, 2014.

VISSER, J. D.; HOEKSTRA, R. F.; ENDE, H. van den. Test of interaction between genetic markers that affect fitness in *aspergillus niger*. *Evolution*, Wiley Online Library, v. 51, n. 5, p. 1499–1505, 1997.

WEINREICH, D. M.; DELANEY, N. F.; DEPRISTO, M. A.; HARTL, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *science*, American Association for the Advancement of Science, v. 312, n. 5770, p. 111–114, 2006.

WRIGHT, S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In: *Genetics: Proceedings of the Sixth International Congress on*. [S.l.: s.n.], 1932. p. 355–366.